Nouvelles technologies, nouvelles méthodes ? Création et analyse de corpus au service de l'enseignement des langues : le cas du hongrois

Szilvia Szita

Institut National des Langues et Civilisations Orientales

École doctorale n°265

Langues, littératures et sociétés du monde

PLIDAM

THÈSE

présentée par

Szilvia SZITA

soutenue le 8 avril 2022

pour obtenir le grade de **Docteur de l'INALCO**

en Sciences du langage : linguistique et didactique des langues

Nouvelles technologies, nouvelles méthodes ?

Création et analyse de corpus au service de l'enseignement des langues : le cas du hongrois

Thèse dirigée par :

M Thomas SZENDE Professeur des universités, INALCO

RAPPORTEURS :

M Alex BOULTON Professeur des universités, CNRS et Université de Lorraine
Mme Mojca PECMAN Professeur des universités, Université de Paris

---

MEMBRES DU JURY :

M Alex BOULTON Professeur des universités, CNRS et Université de Lorraine, France
Mme Mojca PECMAN Professeur des universités, Université de Paris
Mme Andrea NAGY Maître de conférences - HDR, Université de Debrecen (Hongrie)
M Thomas SZENDE Professeur des universités, INALCO

Remerciements

L'engagement dans une thèse est autant un exercice de réflexion individuelle qu'une aventure collective. Le directeur de thèse en est le premier guide. Plus qu'un directeur qui a su mettre sur la voie et conseiller, au-delà des échanges enrichissants et d'une ouverture intellectuelle rare, le professeur Thomas Szende m'a fait l'honneur de sa confiance en m'accompagnant tout au long de ce cheminement. Je tiens à lui exprimer toute ma gratitude.

Je tiens aussi à remercier les membres du jury qui m'ont fait l'honneur et le plaisir d'accepter d'analyser ce travail. Ce jury est à l'image d'un parcours personnel : international et réunissant autour d'une même table des experts passionnés par le même domaine scientifique. Je n'aurais pu souhaiter de meilleurs interlocuteurs qu'Alex Boulton et Andrea Nagy pour en être rapporteurs et Mojca Pecman pour l'examiner et échanger sur ce travail.

Des rencontres, parfois décisives, sont venues tout au long de cette thèse augmenter, influencer, étendre le champ de recherche, nourrir la réflexion. Les membres du laboratoire de PLIDAM et ceux du groupe de recherche « KorSzak » pour l'enseignement du hongrois ainsi que Mónika Szirmai, Anne O'Keeffe, Susan Hunston, Peter Crosthwaite, Maaike Beliën, Jan Hulstijn et Jenny Audring m'ont fait l'honneur de leurs échanges. Je les en remercie sincèrement.

Une mention toute particulière va ici à Anne Buscha qui m'a généreusement proposé de devenir co-auteure des manuels d'allemand et des grammaires pédagogiques et dont la créativité, la profondeur de pensée et l'amitié m'ont accompagnée et sont une source d'inspiration. Je tiens également à remercier Katalin Pelcz, co-auteure de nos manuels et d'autres matériels pédagogiques pour l'enseignement de hongrois. Mme Buscha et Mme Pelcz étaient sur ce chemin plus que des co-auteures et des partenaires de réflexion : elles sont devenues des amies et je les considère comme exemples à bien des égards.

Ma reconnaissance émue va à Michael Hoey, décédé malheureusement trop tôt, dont la théorie du « Lexical priming » m'a ouvert les yeux sur le domaine de la linguistique de corpus et est déterminante aujourd'hui pour mon approche pédagogique. Sans son ouvrage et sans son encouragement, je ne me serais probablement pas engagée dans ce travail de recherche.

Cette thèse a été inspirée, avant tout, par des questions des apprenants auxquelles j'ai été confrontée aux cours de langues. Dans une volonté d'améliorer ma propre pratique pédagogique, j'ai cherché à y intégrer

de nouvelles méthodes ainsi que les résultats actuels de la recherche linguistique. Mes étudiants m'ont significativement aidée dans ce travail et je tiens également à les en remercier.

Enfin, le soutien des plus proches, amis et famille est un élément décisif dans l'accomplissement de cette aventure. Mon mari Jean-Marc, mes parents, ma belle-famille, mon frère et ma belle-sœur, mes amies Lucy Krul et Nadia Wijers en sont les piliers.

À vous toutes et tous,

Merci.

Table des matières

Introduction générale

Domaine relativement nouveau de la science du langage, la linguistique de corpus repose sur l'exploration systématique de vastes bases de données langagières à l'aide des logiciels d'analyse. Les questions essentielles auxquelles ce domaine s'intéresse sont les suivantes : Comment les natifs[1] utilisent-ils leur propre langue dans les différentes situations de communication ? Quelle sont les caractéristiques principales de leur usage langagier ? Peut-on, en outre, déduire de ces observations des renseignements concernant le fonctionnement du langage en général ? C'est par une approche empirique propre à ce domaine (en étudiant un très grand nombre d'énoncés réels et en identifiant des schémas d'usage) que l'on peut déceler des informations précieuses, pertinentes et ce, non seulement pour une meilleure description de la langue concernée mais aussi pour un enseignement des langues plus efficace. Les résultats de ces analyses peuvent d'une part influer sur le contenu des manuels et la nature des outils pédagogiques (enseigner *quoi*) et d'autre part les approches didactiques (enseigner *comment*).

Depuis sa naissance dans les années 1950, le domaine de la linguistique de corpus a rapidement progressé. De ce fait, nous disposons d'outils de plus en plus performants aptes à gérer de très grandes quantités de textes de toutes natures et à révéler un large spectre de caractéristiques de l'usage langagier. Ces évolutions ont également entraîné des changements dans l'apprentissage des langues (notamment pour les langues les plus étudiées) : de nouveaux types de dictionnaires, de grammaires et de manuels ont ainsi vu le jour. De nombreuses collections de textes offrent par ailleurs des outils d'analyse de corpus pour les apprenants et les enseignants (cf. Charles 2005, Chambers 2019). En dépit de ces avancées significatives, *la plupart des développements autour des corpus pour l'enseignement des langues semblent concerner avant tout les logiciels plutôt que la méthodologie et les questions liées à la conception des corpus à des fins pédagogiques*. C'est invariablement la même recommandation qui est donnée aux créateurs de corpus pédagogiques : à l'instar des corpus à des fins linguistiques, il suffirait de compiler des collections de textes authentiques organisés par sujet, genre et/ou registre pour remplir les fonctions d'un corpus pédagogique. Cette recommandation peut convenir à l'enseignement des langues dans le cadre de l'enseignement supérieur ou avec les apprenants maîtrisant la langue cible à un haut niveau. Ces étudiants ne sont cependant que le « sommet de

[1] Nous utiliserons le terme 'natifs' pour désigner les utilisateurs d'une langue avec des compétences comparables aux natifs ainsi que ceux qui identifient la langue en question comme leur langue maternelle ou leur première langue.

l'iceberg ». La population à laquelle nous nous intéressons dans le cadre de ce projet de recherche sont les apprenants qui constituent « la base de l'iceberg ».

Problématique

Ce projet de recherche cherche à construire un pont entre la linguistique de corpus et les pratiques pédagogiques au service de l'enseignement des langues. Au cœur de notre réflexion se place le constat que *les acteurs principaux de l'enseignement des langues (les auteurs d'ouvrages à fin pédagogiques, les enseignants et les apprenants) n'utilisent pas les corpus aux mêmes fins que les linguistes*. Alors qu'ils peuvent manifester un intérêt pour la linguistique, ces groupes sont d'abord caractérisés par des attentes et des besoins spécifiques. Les auteurs des ouvrages pédagogiques explorent ainsi les corpus dans le but principal d'identifier les éléments linguistiques susceptible d'être inclus dans le matériel à un niveau de compétence donné. Les professeurs de langues utilisent par ailleurs les corpus comme ils utiliseraient d'autres outils pédagogiques : dans l'objectif de faire progresser leurs étudiants et de les aider à devenir des utilisateurs compétents de la langue cible. Enfin, le but principal des apprenants est de maîtriser la langue et ils sont prêts à cette fin à faire usage de toutes sortes d'outils mis à leur disposition dans l'objectif d'obtenir des réponses à des questions bien précises concernant l'usage de certains éléments langagiers. Or, les corpus existants n'ont pas été nécessairement conçus en intégrant les exigences spécifiques de l'ensemble de ces acteurs. Face à ces attentes et ces besoins, les questions principales que nous explorerons dans ce travail de recherche sont les suivantes :

- Comment utiliser les corpus conçus à fins non pédagogiques et à fins pédagogiques pour améliorer la qualité des matériels pour l'enseignement des langues ?
- Comment construire des corpus à fins pédagogiques ?
- Comment l'utilisation des corpus peut-elle contribuer à l'amélioration des pratiques pédagogiques existantes et à l'implémentation de nouvelles méthodologies efficaces ?
- Comment les apprenants des niveaux de compétences linguistiques inférieurs (A1, A2 et B1) – constituant la majorité de la population des apprenants – peuvent bénéficier de l'exploration des corpus ? Comment l'utilisation des corpus peut servir à améliorer leurs compétences linguistiques, i.e. la qualité de leurs produits oraux et écrits ?

Notre travail s'articule autour de deux grands axes. Il a pour ambition, dans un premier temps, d'aborder la question de l'usage des corpus existants pour les explorations linguistiques à des fins pédagogiques. *Notre recherche portera plus précisément sur la relation entre le lexique et la grammaire* – un

domaine de recherche d'intérêt particulier pour les linguistes de corpus, sous-exploré pour la langue hongroise. Ces deux domaines sont en l'occurrence présentés séparément au sein des matériels pédagogiques dédiés à cette langue, suggérant qu'il s'agit de deux entités qui ne se rapprochent que par moments. Nous explorerons en quelle mesure les analyses des textes concrets, authentiques confirment la validité de cette séparation de principe pour une langue morphologiquement complexe comme le hongrois. *Notre première hypothèse est qu'il est possible de démontrer une interconnexion étroite entre ces deux aspects de la langue susceptible d'être exploitée au service du développement de nouveaux contenus pédagogiques ainsi que de nouvelles approches méthodologiques pour l'enseignement des langues.*

Notre étude porte, dans un deuxième temps, sur les questions concernant *la constitution d'un type de corpus particulier : les corpus à fins pédagogiques*. Les corpus appartenant à cette catégorie tiennent compte des besoins des apprenants et proposent du matériel linguistique qui leur est pertinent et accessible. Idéalement, de tels corpus représentent l'usage langagier des natifs et contiennent des énoncés liés à des situations diverses que les apprenants peuvent comprendre. *Notre deuxième hypothèse est donc que, pour parvenir aux besoins des apprenants aux niveaux de compétences linguistiques inférieurs, il est nécessaire de construire de nouveaux corpus pour les niveaux de compétences linguistiques inférieurs (A1–B1), selon des principes différents de ceux définis pour les corpus à des fins linguistiques.* Plus précisément, nous chercherons à répondre aux questions suivantes :

- Dans quelle mesure les critères utilisés pour la construction des corpus non pédagogiques peuvent s'appliquer à celle des corpus pédagogiques ?
- À quels critères ces corpus doivent-ils répondre pour permettre un usage effectif dans le cadre de l'enseignement des langues ?
- Comment ces corpus peuvent-ils être explorés par l'apprenant dès le début de son parcours, aux niveaux A1 à B1 ?
- L'utilisation de ces corpus, effectuée dès le début de l'apprentissage, peut-elle enrichir l'expérience de l'apprentissage et de fournir des opportunités d'observation de l'usage langagier que d'autres ressources (manuels, dictionnaires...) ne peuvent pas offrir ? Si oui, quels avantages offre-t-elle ?

Orientations théoriques et approche méthodologique

Ce projet de recherche s'inscrit dans un cadre interdisciplinaire : il se situe au croisement de la linguistique appliquée – plus précisément, de la linguistique de corpus – et de la didactique de

l'enseignement des langues étrangères. Ainsi, nos explorations linguistiques seront motivées par des considérations pédagogiques tout le long de ce travail.

La linguistique de corpus est un domaine de la science du langage qui n'est pas dédié à l'étude d'un aspect particulier de la langue choisie (McEnery 2011). Sa recherche a pour objet les manifestations langagières concrètes, produites dans des situations réelles et elle analyse, comme évoqué précédemment, des textes authentiques. Elle utilise des méthodes empiriques afin de cartographier les caractéristiques de l'usage de la langue ou une partie de la langue étudiée. *En appliquant ces méthodes, nous explorerons quelques caractéristiques de la langue hongroise à partir des questions concernant le lexique et la grammaire.* Notre outil le plus important sera le Concordancier permettant l'analyse de tous les exemples avec l'élément linguistique choisi. L'étude systématique de ces occurrences sert à faire émerger des schémas linguistiques (*patterns*, voir le chapitre 3, A.8) et à mettre en évidence des répétitions lexicales, grammaticales, sémantiques et pragmatiques. Nous nous appuierons sur une méthodologie définie par Sinclair (1991) et modifiée par Hoey (2005) pour identifier ces schémas et créer ainsi le « profil » de l'élément choisi.

Notre intérêt tout au long de ces explorations, reste cependant celui du didacticien, notre but ultime étant d'accroître l'efficacité de l'enseignement et de l'apprentissage des langues. À cette fin, *les explorations quantitatives et qualitatives à l'aide des outils statistiques serviront à identifier les éléments linguistiques méritant d'être inclus dans les matériels pédagogiques pour les niveaux inférieurs*. Nous nous référerons aux descripteurs du Cadre européen commun des références pour les langues (CECRL) afin de déterminer les sujets et les types de textes ainsi que les sujets pertinents pour chaque niveau. *Nous chercherons également à identifier les principes généraux d'une présentation et d'une pratique fondées sur l'analyse de corpus.* Les réflexions sur la dimension pédagogique de l'analyse de corpus ont été inspirées par le travail des linguistes-didacticiens préconisant l'« Apprentissage sur corpus » (« Data-driven learning ») notamment par les travaux de Johns, Boulton et Tyne. Le travail de nombreux chercheurs portant sur la construction de corpus pédagogiques, sur la didactisation des résultats d'analyse de corpus et sur leur intégration dans les matériels pour l'enseignement des langues, formera le deuxième pôle des considérations didactiques.

Plan de la thèse

Le présent travail se divise en trois grandes parties, chacune exposant un aspect différent de la constitution et de l'utilisation des corpus à des fins pédagogiques.

La Partie I (Chapitres 1 à 6) sera dédiée au recensement de la littérature concernant les avancées de la linguistique de corpus. Nous présenterons les méthodes de la linguistique de corpus et ses résultats pertinents ainsi que les principes de construction des corpus et des ouvrages pédagogiques se fondant sur une « approche corpus ». Ces cinq chapitres fournissent le fondement pour le travail original sur la langue hongroise, présentés dans les deuxième et troisième parties de cette thèse.

Le chapitre 1 définira les objectifs de la linguistique de corpus et présentera les différents types de corpus non pédagogiques ainsi que les paramètres principaux qui jouent un rôle dans leur construction. Le chapitre 2 sera consacré aux critères pour la constitution des corpus pédagogiques. Notre principal objectif est, dans cette partie, d'identifier dans quelle mesure les critères définis pour les corpus non pédagogiques s'appliquent aux corpus à des fins pédagogiques pour lesquels le principe de la présentation d'un langage à caractère naturel (« natural-sounding language ») est aussi important que l'accessibilité des textes aux niveaux de compétences linguistiques inférieurs. Ce chapitre sera suivi par la présentation des outils d'analyse de corpus les plus importants ainsi que par quelques exemples pertinents qui démontrent l'intérêt de leur utilisation (Chapitre 3). Le chapitre 4 recensera les résultats les plus pertinents du domaine pour l'enseignement et l'apprentissage des langues. En connaissant la multitude des études dans ce domaine, nous n'avons bien évidemment pas la prétention de proposer une description exhaustive des résultats obtenus. Nous nous concentrerons sur ceux qui sont susceptibles d'influencer la présentation et la pratique de la langue cible dans les matériels pédagogiques. Nous nous intéresserons avant tout aux caractéristiques de l'utilisation de la langue par les locuteurs natifs ou experts, telles qu'elles émergent des énoncés réels. Une attention particulière sera prêtée à la question du sens des mots (que la linguistique de corpus ne sépare pas de leur utilisation) et à l'importance de l'environnement textuel. Nous traiterons également de la connexion entre lexique et grammaire, de l'importance de la fréquence d'occurrences des éléments linguistiques dans le corpus, de l'importance du registre et celle du langage interactionnel et transactionnel. Nous conclurons cette première partie par la présentation de quelques ouvrages pédagogiques (dictionnaires, grammaires et manuels) qui pourront servir d'exemples pour la réalisation d'une « approche corpus » (Chapitres 5 et 6).

La Partie II du présent travail sera dédiée à l'exploration des corpus non pédagogiques au service de l'enseignement du hongrois aux niveaux de compétences linguistiques inférieurs. Nous nous intéresserons avant tout au contenu et à la manière de présentation des résultats de l'analyse de corpus dans les matériels pédagogiques. À cette fin, deux aspects lexicaux (les mots fréquents à usages multiples et les

synonymes) et deux aspects grammaticaux (l'utilisation des préfixes et les deux conjugaisons) seront analysés en détail. Aucun de ces aspects n'a jamais encore été étudié à l'aide des outils de la linguistique de corpus. Nous chercherons à répondre aux questions suivantes (Chapitres 7 à 12) :

- Dans quelle mesure l'analyse des corpus peut-elle contribuer à une meilleure compréhension d'une langue morphologiquement complexe comme le hongrois ?
- L'analyse de corpus est-elle utile pour complémenter les outils existants et pour améliorer les matériels pour l'enseignement des langues (les grammaires pédagogiques, les dictionnaires pour apprenants et les manuels pour les niveaux de langue inférieurs) ?
- La présentation de certains aspects problématiques de cette langue peut-elle être améliorée par des exemples tirés des corpus ?
- Quelles sont les cas dans lesquels l'utilisation de corpus s'avère particulièrement utile ?

La discussion autour de ces questions sera suivie par une réflexion concernant la manière de présenter les résultats dans le matériel pédagogique et sur la nature des activités pour le cours de langue qui aident à pratiquer et à consolider les phénomènes observés. Le but ultime de ces explorations sera de déterminer en s'appuyant sur les résultats révélés à partir du corpus si et en quelle mesure les aspects grammaticaux et les aspects lexicaux sont interdépendants.

La Partie III de cette thèse (Chapitres 13 à 15) concerne la création et l'utilisation des corpus pédagogiques pour l'observation et la pratique langagières. Nous explorerons leur potentiel et leurs limites à travers l'exemple concret des corpus écrits et oraux compilés pour la série de manuels « MagyarOK » dont nous sommes auteure. Ces corpus pour les niveaux de compétence inférieurs contiennent des données de différents types (textes descriptifs, opinions, évaluations, etc.) qui permettent *d'observer les caractéristiques de l'usage langagier des natifs d'un côté et d'utiliser les énoncés enregistrés dans le corpus comme modèles pour les produits linguistiques de l'apprenant de l'autre*. Nous exposerons une approche en trois étapes pour la construction d'un corpus à contenu mixte (données authentiques, semi-authentiques et créées à partir du corpus) et les considérations autour de l'équilibre fragile entre authenticité (des énoncés réels) et accessibilité (des énoncés compréhensibles au niveau linguistique de l'apprenant). Nous aborderons les questions suivantes, liées à la collection et à la modification des données authentiques :

- Où collecter pour le corpus pédagogique pour les niveaux inférieurs ?
- Quel type de matériel se prête à être inclus dans un tel corpus ?

- Comment peut-on générer du langage à caractère naturel aux niveaux linguistiques des apprenants ?
- Comment peut-on modifier des énoncés authentiques pour les rendre accessibles aux niveaux inférieurs ?

Nous proposerons également des activités pour l'observation et la pratique des éléments linguistiques que contiennent les corpus et éluciderons (toujours « dans l'esprit corpus ») l'intérêt de l'analyse manuelle des textes complets inclus dans le corpus pédagogique.

Nous conclurons ce projet de recherche par une synthèse (Chapitre 16) évoquant les nombreuses voies de recherche possibles qui peuvent prolonger ou compléter les résultats obtenus, notamment concernant une méthodologie cohérente pour la constitution et l'utilisation des corpus pour l'apprentissage des langues.

À la fin de cette introduction, nous souhaitons faire deux remarques à propos de l'usage langagier de cette thèse :
1) Les mots « natif », « locuteur expert » et « utilisateur expert » seront utilisés comme des synonymes.
2) Si nous faisons dans cette thèse référence à des personnes en utilisant systématiquement la forme masculine, c'est seulement aux fins d'une meilleure lisibilité. Le terme comprendra, bien évidemment, tous les différents genres.

PARTIE I : Les avancées dans la linguistique de corpus

Introduction à la Partie I

La première partie de ce travail de recherche est dédiée à la présentation de l'état de l'art en linguistique de corpus, avec un focus particulier sur son application à l'apprentissage des langues. Les premiers chapitres s'articuleront autour de la méthodologie de recherche ainsi que sur les principes de création des corpus à fins linguistiques et à fins pédagogiques. Nous argumenterons en faveur d'une séparation de ces deux types de corpus et nous recenserons les critères pour la construction des corpus à fins pédagogiques.

Toujours dans une approche axée sur la didactique, nous ferons par la suite l'inventaire des résultats les plus pertinents de la linguistique de corpus susceptibles d'augmenter l'efficacité de l'enseignement des langues. Comme la linguistique de corpus est un domaine de recherche qui se veut avant tout empirique, ses résultats nous informent sur l'usage langagier des natifs tel qu'il apparaît dans des énoncés réels, produits dans une variété de situations. Une réflexion sur les avancées dans ce domaine nous permettra d'identifier les éléments-clés à intégrer dans la pratique pédagogique.

Les deux derniers chapitres de cette partie seront consacrés à une analyse d'ouvrages pédagogiques conçus dans une « approche corpus ». Il s'agit d'ouvrages dont les auteurs ont consciemment intégré des connaissances sur l'usage langagier issues de la linguistique de corpus. Ces résultats ont un impact sur leur démarche théorique (la manière dont les informations linguistiques sont systématisées dans l'ouvrage) ainsi que sur leur approche pratique (le type d'exercices et la manière de présenter la langue).

Le recensement de la littérature nous permettra de mieux appréhender l'exploration de la langue hongroise dans les Parties II et III de cette thèse.

Chapitre 1 : Le domaine de la linguistique de corpus et les principes de construction de corpus non pédagogiques

Dans ce premier chapitre, il convient de définir les sujets de recherche ainsi que les considérations méthodologiques de la linguistique de corpus afin de la différencier des autres domaines de la science du langage. La première partie du chapitre offre, dans cette optique, une présentation de la linguistique de corpus en tant que méthode scientifique permettant l'exploration de nombreux aspects de la langue. La deuxième partie du chapitre sera dédiée à la présentation des différentes catégorisations possibles des corpus et la dernière partie présentera les principes les plus importants de la construction de corpus non pédagogiques. Au cours de l'exploration de ces sujets, nous nous concentrerons avant tout sur les aspects liés à l'enseignement des langues ; la dimension pédagogique reste ainsi au premier plan tout au long du chapitre.

A) Le domaine de la linguistique de corpus

La linguistique de corpus est un domaine relativement nouveau de la science du langage. Elle doit sa naissance à l'émergence des outils informatiques qui permettent d'analyser de très grandes quantités de données linguistiques. Dans l'ouvrage « L'histoire de la linguistique de corpus », McEnery (2013 : 745–746) définit le sujet de recherche de ce domaine comme suit :

> « Contrairement à des domaines tels que la phonologie, l'étude de la variation sociale du langage ou l'analyse critique du discours, la linguistique de corpus ne concerne pas directement l'étude d'un aspect langagier particulier. Elle s'intéresse avant tout à l'ensemble de procédures ou de méthodes qui permettent d'étudier la langue. La linguistique de corpus est donc une méthodologie qui peut être appliquée […] à plusieurs domaines d'études différents de la linguistique. » (notre traduction)

1) Les caractéristiques méthodologiques

La recherche en linguistique de corpus a pour objet les manifestations concrètes et réelles de la langue, que les outils – principalement numériques – permettent d'explorer de manière empirique. Brazil (1995 : 24) définit le corpus comme une « collection du langage utilisé » (notre traduction), expliquant que le « langage utilisé » est « la langue qui se produit dans des circonstances dans lesquelles les locuteurs faisaient clairement quelque chose d'autre que démontrer le fonctionnement du système [langagier] » (notre traduction). Cette définition est pertinente en ce

qu'elle souligne le fait que les textes[2] réunis dans le corpus n'ont pas été générés dans le but de démontrer ou réfuter une hypothèse linguistique. Un corpus contient, en effet, des textes que les locuteurs ont prononcés ou écrits à un moment donné, dans une situation de communication réelle. Tout l'intérêt de l'enregistrement et de l'analyse de tels énoncés repose sur le fait qu'ils nous aident à obtenir des connaissances plus approfondies sur l'usage langagier et, par conséquent, d'en fournir une meilleure description. Ce domaine de la science du langage adopte donc une position descriptive ; les linguistes de corpus observent l'utilisation de la langue et décrivent les phénomènes sélectionnés (ou la langue dans son intégralité) à partir de leurs observations. Ils ne se soucient pas de prescrire comment les natifs doivent parler et écrire, ce qui est correct et incorrect, acceptable et inacceptable, mais cherchent à rassembler des informations à propos de l'usage langagier en tant que tel. À cette fin, le linguiste doit tout d'abord choisir un corpus approprié ou définir le contenu du corpus à construire s'il n'en existe pas encore un. En fonction de la variété langagière qu'il souhaite étudier – par exemple celle d'un écrivain, d'une région ou d'une époque –, il doit sélectionner des textes appropriés. Comme souligné par Biber et al. : « Un corpus cherche à représenter une langue ou une partie d'une langue » (notre traduction) (1998 : 246) ou encore par Sinclair (1991 : 13) : « [Un corpus est] un ensemble de textes naturellement produits, collectionnés suivant certains principes » (notre traduction). Ces grands corpus peuvent être explorés de façon efficace à l'aide des outils numériques rendant possibles les analyses statistiques pertinentes. Ces études concernent la fréquence de mots et d'unités multi-lexicales[3]. Leur but est d'identifier des « schémas » (*patterns*, pour une définition, voir le chapitre 3, A.8) d'usage et de définir ce qu'on entend par « norme ou utilisation standard » et « utilisation créative » de la langue (cf. Carter 2004 ; Hanks 2013 ; Jones and Waller 2015 ; Taylor 2012). Les méthodes appliquées ont incité Brezina (2018) et d'autres chercheurs (McEnery et Hardie 2012) à qualifier la linguistique de corpus de « méthodologie quantitative *par essence* » (Brezina 2018 : 3, notre traduction).

Après avoir identifié le but principal et l'approche méthodologique de la linguistique de corpus, il convient maintenant de situer l'émergence de ce domaine dans un contexte historique pour apprécier sa différence par rapport à d'autres grands courants parallèles. L'analyse du langage assisté par ordinateur devient plus largement accessible dans les années 1950. C'est aussi l'époque des premiers travaux de Chomsky sur la grammaire générative. Ceux-ci ont révolutionné la recherche en syntaxe et ont, en conséquence, influencé notre façon de penser la langue. Cette branche de la linguistique théorique s'oppose fondamentalement aux méthodes de la linguistique

[2] Nous entendons par « textes » les énoncés écrits et oraux.
[3] Unités linguistiques composées de plusieurs mots, voir le chapitre 3 pour plus de détails.

de corpus et Chomsky (1968) lui-même a fortement réfuté l'intérêt de ce domaine dès le début. Selon lui, les données empiriques ne présentent aucun intérêt pour la linguistique car le langage externe (*E-language* ou *performance*) n'est qu'une approximation du langage interne (*I-language* ou *compétence*) et ne fournit pas d'explication sur notre capacité à comprendre des énoncés que nous n'avons jamais entendus. Il soutient que la compétence ne s'exprime pas pleinement dans le langage extérieur car ce dernier est de « qualité dégénérée » (Chomsky 1968 : 8, notre traduction) qui ne dévoilerait que très partiellement le potentiel du langage humain. De plus, le langage observé (toujours un ensemble fini de données linguistiques) n'est d'aucune utilité lors de l'étude de la compétence. Il est intéressant de noter que les remarques de Chomsky ont eu un effet plutôt inattendu : elles ont incité les linguistes de corpus à définir plus précisément les principes de base pour construire des corpus représentatifs et équilibrés ainsi qu'à établir des théories linguistiques fondées sur l'étude du langage produit dans des circonstances naturelles. Ces principes seront exposés dans la section C) de ce chapitre et les théories linguistiques les plus pertinentes pour l'enseignement des langues seront brièvement présentées au Chapitre 4.

Des principes rigoureux employés lors de la collecte et de l'organisation des données ainsi que les descriptions plus détaillées concernant le contenu des corpus (Egbert et al. 2020) ont rapproché ce domaine des sciences naturelles dont Brezina (2018) résume ainsi les trois critères essentiels : (1) la reproductibilité des résultats, (2) la divulgation des preuves empiriques et (3) le fondement des résultats sur des analyses statistiques. Ces principes prévoient que le choix de corpus et les méthodes d'analyse doivent être décrits de façon claire et transparente afin que d'autres chercheurs puissent reproduire la même étude et en vérifier ses résultats. Il est également recommandé de « mettre les corpus à la disposition d'autres chercheurs pour leur permettre de les explorer davantage et ainsi de faire progresser les connaissances dans ce domaine » (notre traduction) (Brezina 2018 : 2).

Une autre caractéristique notable qui distingue l'approche de la linguistique de corpus de plusieurs autres écoles, est *sa manière de vérifier les hypothèses scientifiques*. Alors que d'autres domaines travaillent avec des exemples inventés, soumis à des locuteurs natifs pour des jugements d'acceptabilité, la linguistique de corpus oblige le linguiste à fournir des preuves empiriques à l'appui de toute déclaration faite sur le langage. Ces preuves empiriques sont formées de données linguistiques tirées de corpus, comme évoqué précédemment. L'utilisation des *mesures statistiques* rend l'analyse et l'interprétation des éléments choisis plus sûres et transparentes et permet de quantifier et de comparer les résultats de plusieurs études (Brezina 2018).

Le tableau suivant résume les objectifs et les caractéristiques méthodologiques de la linguistique de corpus dont les éléments-clés sont les suivants :

- La linguistique de corpus étudie l'usage langagier en appliquant des méthodes empiriques.
- Les linguistes de corpus analysent avec les logiciels une collection de textes réels et authentiques de taille significative. Le but de ces analyses outillées est d'identifier des schémas dans l'usage langagier et d'explorer l'environnement textuel des éléments choisis.
- Les chercheurs effectuent des analyses qualitatives (études d'exemples concrets dans le corpus) et quantitatives (identification des informations statistiques pertinentes sur l'élément étudié) pour fournir une meilleure description, et éventuellement une nouvelle théorie de la langue.

2) Les branches de la linguistique de corpus

Selon que l'on accorde ou non un statut théorique au domaine, nous distinguons deux grandes écoles au sein de la linguistique de corpus. Les linguistes appartenant à la première école différencient la « linguistique fondée sur le corpus » (corpus-based linguistics) et la « linguistique guidée par le corpus » (corpus-driven linguistics) (cf. Tognini-Bonelli 2004, 2010 pour une analyse approfondie). Les études fondées sur le corpus utilisent des données de corpus pour explorer une théorie ou une hypothèse généralement établie dans la littérature, ceci dans le but de la valider, de la réfuter ou de l'affiner. Il s'agit donc d'une méthode au service d'une approche théorique. Comme relevé par Jones et Waller (2015 : 8–9, 30), la première application d'un corpus est de tester et de remettre en question nos intuitions sur le langage : « [Un corpus] peut souligner ou réfuter une idée que nous avons de l'utilisation du langage. » (notre traduction). *Il est également possible de découvrir des schémas d'usage concernant l'élément langagier étudié, que le linguiste pourrait négliger ou juger moins saillant en s'appuyant seulement sur son intuition d'expert* (Jones et Waller 2015 : 13). En revanche, les chercheurs en linguistique « guidée par le corpus » affirment que le corpus lui-même devrait être la seule source de nos hypothèses sur le langage. Ces linguistes forment le groupe des « néo-firthiens » dont les théories s'appuient sur l'observation et sur l'interprétation de l'usage langagier. Dans cette tradition s'inscrivent les travaux de Sinclair, Hoey, Hunston, Stubbs, Hanks, Teubert et Tognini-Bonelli – pour ne nommer que les chercheurs les plus éminents. Au chapitre 4, nous fournirons une présentation succincte des théories néo-firthiennes suivantes : le Principe de l'idiomaticité (*Idiom*

Principle) et le Principe du libre choix (*Open Choice Principle*) par Sinclair (1991, 2004b), la théorie du Priming ou Amorçage lexical (*Lexical Priming*) par Hoey (2005) et la grammaire des schémas (*Pattern Grammar*) par Hunston et Gill. Ces théories ont été choisies parce qu'elles peuvent fournir, comme nous le verrons par la suite, des bases solides pour des méthodologies cohérentes pour l'enseignement des langues.

La deuxième école regroupe des chercheurs qui n'acceptent pas cette approche binaire et n'accordent pas un statut théorique à la linguistique de corpus. Ils rejettent également l'affirmation que le corpus puisse être un objet apte à la théorisation (cf. McEnery 2013, Brezina 2018). Ces chercheurs ne préconisent pas la distinction entre linguistique de corpus comme méthode scientifique d'exploration de la langue, et linguistique de corpus comme méthode scientifique pour la construction de théories linguistiques. Ils utilisent les méthodes et les outils d'analyse de corpus à des fins diverses : pour l'analyse de discours (par exemple, Ädel et Reppen 2008 ; McEnery et Hardie 2012 ; Rühlemann 2007, 2018), pour l'étude de différents registres (par exemple, Biber 2012 ; Biber et Conrad 2009 ; Biber et Egbert 2018 ; Nini 2019 ; Frérot et Pecman 2021; Spina et Tanganelli 2012), comme aide à l'interprétation des œuvres littéraires (par exemple, Mahlberg 2017 ; Mahlberg et Stockwell 2015), comme outil pour la traduction (Kübler 2014a et b, Kübler et al. 2018) et en tant qu'élément constitutif d'une approche pédagogique (Burnhard et McEnery 2000 ; Cavalla 2019 ; Chambers 2019 ; Charles 2007, 2014, 2015 ; Cobb et Boulton 2015 ; Frankenberg-García et al. 2011 ; Friginal 2018 ; Kramer 2011 ; Landure et Boulton 2010 ; Leńko-Szymańska 2017 ; Meunier et Reppen 2015 ; O'Keeffe et al. 2007 ; Poole 2018 ; Pérez-Paredes 2021 ; Pérez-Paredes et Mark 2021, Pérez-Paredes et al. 2020 et d'autres). Dans le cadre de ce projet de recherche, nous nous intéresserons, avant tout, à cette dernière application liée à la construction et à l'exploration des corpus pédagogiques.

B) Les différents types de corpus

> *Il n'y a pas de bon et mauvais corpus – celui qui est approprié dépendra des besoins de l'utilisateur.* (Anne O'Keeffe 2007 : 34, notre traduction)

Les corpus se catégorisent de différentes manières selon les critères choisis pour la description de leur contenu. Ces critères détermineront également le processus de sélection de textes qui feront partie de la base de données. Genre, sujet, période du temps et lieu de la collection, nombre de langues incluses dans le corpus – chacun de ces critères peut fournir la base d'une catégorisation valide (Tognini-Bonelli 2010). Nous présenterons dans ce qui suit les catégories établies par la

littérature en mettant l'emphase sur leur intérêt pour la didactique des langues. Au cours de cette présentation, nous évoquerons également quelques problèmes relatifs à la catégorisation des corpus. D'après Davies (2015 : 15) nous distinguons :

1. Selon la nature des données : des corpus écrits et des corpus oraux ;
2. Selon le profil du corpus : des corpus généraux et des corpus spécialisés ;
3. Selon la nature du corpus par groupes d'utilisateurs : des corpus généraux ou spécialisés non-pédagogiques et des corpus pédagogiques ;
4. Selon les fournisseurs de données : des corpus d'utilisateurs experts et des corpus d'apprenants ;
5. Selon la ou les langues ou variétés linguistiques : des corpus monolingues et des corpus parallèles/multilingues ;
6. Selon la période de temps choisie : des corpus historiques (diachroniques) et des corpus synchrones.

Il est important de considérer que les échantillons de langage pour un corpus doivent, autant que possible, être constitués de textes entiers plutôt que d'énoncés arbitraires et sans aucun lien logique. Une telle démarche permet non seulement l'étude du phénomène linguistique choisi dans des environnements textuels différents mais elle autorise également l'analyse structurelle des textes inclus dans la collection.

1) Les types de corpus en fonction de la nature des données : corpus écrits et corpus oraux

Selon la nature des textes, nous distinguons des corpus écrits et des corpus oraux. Un corpus écrit contient des textes enregistrés dans un format écrit, provenant de livres, de pages Web ou d'autres sources. Un critère supplémentaire pour qu'une collection de textes puisse être qualifiée de corpus est qu'elle ait été publiée dans un format lisible par les outils numériques (McEnery 2013). La majorité de ces textes sont rassemblés aujourd'hui sur Internet car un grand nombre d'ouvrages littéraires ou autres ainsi que des journaux sont numérisés. Une partie de la communication interpersonnelle se déroulant à présent sur les réseaux sociaux ; il est donc également possible de réunir des données linguistiques reflétant les propriétés de ces interactions. L'accès à une multitude de textes variés nous permet de construire des corpus généraux de taille significative ainsi que de créer des collections plus ciblées qui explorent, par exemple, un sujet ou un genre spécifique.

À la différence des corpus écrits, les bases de données contenant des énoncés produits à l'oral sont généralement de taille plus restreinte. Idéalement, ces corpus oraux devraient contenir des enregistrements audio et/ou audio-visuels ainsi que leurs transcriptions, ce qui n'est pas le cas actuellement en raison des difficultés techniques liées à l'exploitation des données multimédia, d'une part, et des problèmes relatifs à la protection des données personnelles, de l'autre. Néanmoins, le format multimédia serait particulièrement bénéfique pour l'enseignement des langues car il conserve les informations textuelles et contextuelles (André 2017, 2018, 2019, 2020 ; André et Ciekanski 2018 ; Bisson et al. 2014 ; Braun 2006, 2010 ; Chambers 2019 ; Fortanet-Gómez et Querol-Julián 2010 ; Montenero et Rogers 2019). À présent, ce sont surtout les transcriptions qui se prêtent à l'analyse avec des outils numériques[4]. Or, la transcription des textes oraux implique toujours la perte inévitable des données audio-visuelles, c'est-à-dire de la voix et de l'image propres à la situation dans laquelle les énoncés ont été produits (Thompson 2004)[5]. Au sein des corpus écrits, les transcriptions de ces interactions orales représentent ainsi une catégorie intermédiaire.

Il est important de noter que les termes « corpus écrits » et « corpus oraux » sont des termes généraux car cette catégorisation ne garantit guère l'homogénéité des données linguistiques. Un corpus écrit peut contenir des entrées de forum, des articles scientifiques et/ou des textes littéraires – tous reflétant des usages langagiers très différents les uns des autres. De même, un corpus oral peut inclure des cours universitaires ou des conférences scientifiques représentant des usages formels aussi bien que des interactions du quotidien, des conversations dans un magasin ou dans un café par exemple. La typologie de caractérisation que nous présenterons dans des sections suivantes permettra de fournir plus de précisions quant à la nature du corpus.

2) Des types de corpus en fonction de leur profil : corpus généraux et corpus spécialisés

Les corpus se divisent en deux groupes majeurs selon leur profil : des corpus généraux et des corpus spécialisés. Ces deux types de corpus remplissent des fonctions différentes.

[4] Il existe quelques tentatives isolées d'intégrer les informations audio-visuelles dans les corpus. Par exemple, le corpus multimédia de Sketch Engine (sketchengine.eu) propose l'enregistrement audio des éléments sélectionnés par l'utilisateur mais il n'est pas possible d'écouter plus de quelques mots par énoncé. Le *Corpus of Interactional Data* (Blanche et al. 2017) offre huit heures de vidéos annotées avec transcription pour le français. Le BNC propose également des enregistrements d'interactions en format audio et en tant que transcriptions. Les corpus pédagogiques présentés au chapitre 2 incluent du matériel multimédia mais seules les transcriptions des textes sont analysables avec les outils numériques.

[5] Nous aborderons plus en détail la question de la transcription au chapitre 14.

Les corpus généraux sont « construits pour refléter l'utilisation du langage par des groupes très vastes et variés » (Friginal 2017 : 12, notre traduction) et contiennent des milliards de mots. Ils comprennent généralement de nombreux types de textes différents – littérature, articles de journaux, entrées de forum et de blog, etc. – ainsi que des textes provenant de plusieurs régions et périodes. Leur objectif n'est pas de présenter « ce qui est correct » […], mais uniquement « ce qui est souvent dit dans un contexte particulier », comme l'explique Hunston (2009 : 145, notre traduction). Parmi les exemples de tels corpus, citons le British National Corpus (BNC) et le Corpus of Contemporary American (COCA) pour l'anglais, le corpus COSMA pour l'allemand, le National Corpus of Contemporary Welsh (CorCenCC) pour le gallois, le Corpus national du tchèque (CNC), le Corpus national hongrois (MNSZ), Ortolang pour le français et bien d'autres.

Les corpus spécialisés sont généralement de taille limitée et comprennent des textes liés à des sujets et/ou à des genres spécifiques (cf. Biber 1993, 2009, 2012 ; Biber et Conrad 2009 ; Biber et Egbert 2018). Ils sont généralement conçus pour permettre l'étude d'une variété linguistique particulière. Les corpus thématiques contiennent ainsi des textes sur le même sujet tels que des récits de voyage, des recettes, des discussions sur le weekend. Les corpus par genre peuvent quant à eux comprendre, par exemple, des e-mails formels, des articles scientifiques, des discours parlementaires ou des conversations entre médecin et patient. Un exemple pertinent de tels corpus est le « sms4science.org » des clavardage (en anglais : *chats*) en français.

Un corpus mixte est compilé en tenant compte du sujet *et* du genre des textes. Ces corpus permettent aux utilisateurs d'identifier simultanément les caractéristiques langagières liées au sujet et au genre étudiés. Ils peuvent contenir des articles scientifiques publiés sur un sujet bien précis, des textes parus dans la même rubrique d'un ou plusieurs journaux, etc. À ce groupe de corpus appartiennent, par exemple, la collection d'articles scientifiques dans le domaine des études forestières par Friginal (2017) ou les corpus multilingues multithématiques pour la traduction par Kübler (2014a et b).

3) Les types de corpus en fonction de groupes d'utilisateurs : corpus à fins linguistiques et corpus à fins pédagogiques

De nombreux corpus ont pour objectif de faciliter la recherche linguistique – en effet, la majorité des corpus entre dans cette catégorie. Le critère le plus important pour la construction d'un corpus à fins linguistiques est qu'il doit comprendre des données authentiques collectées dans des situations du quotidien dans lesquelles les locuteurs utilisent la langue pour accomplir une action

(Sinclair 1991, 1997). Une fois la condition d'authenticité remplie, le profil du corpus sera défini par les besoins du créateur de corpus. Ces corpus sont construits dans le but de fournir des informations sur les habitudes langagières des locuteurs natifs et des utilisateurs experts ainsi que des exemples réels pour l'usage de la langue dans différentes situations.

Le corpus pédagogique est un autre type de corpus ou, pour reprendre le terme de Leech (1997), le « corpus pour enseignement » (*teaching-oriented corpus*). Ce type de corpus est conçu spécifiquement à des fins pédagogiques et tient compte, par conséquent, des besoins des apprenants (Braun 2005, 2010). Contrairement aux corpus à des fins linguistiques, il n'existe pas à l'heure actuelle de consignes bien définies concernant la manière de réaliser un corpus pédagogique. Le chapitre 2 sera dédié à la présentation détaillée des tentatives de définitions de critères de construction d'un corpus pédagogique efficace.

Il est important de noter que les corpus à fins linguistiques peuvent être également exploités dans l'objectif de créer du matériel pédagogique. Ils peuvent fournir des renseignements utiles sur l'usage langagier pour les ouvrages pédagogiques et pour les cours de langues. Ils peuvent, par exemple, permettre d'identifier les usages typiques d'une forme grammaticale dans un contexte donné (par exemple, explorer l'utilisation du passé en utilisant des récits sur le weekend) ou du vocabulaire-clé lié à un sujet (par exemple, cartographier les mots et les expressions utiles pour l'achat de vêtements en s'appuyant sur des interactions dans des magasins de vêtements et/ou sur des textes dans des catalogues de mode et sur des blogs, etc.). Ils peuvent également apporter des réponses aux questions qui échappent à des règles claires, et cela même aux niveaux de compétences linguistiques inférieurs, comme cela sera démontré dans la Partie II de cette thèse.

4) Les types de corpus en fonction des locuteurs : corpus de l'usage langagier expert et corpus d'apprenants

Toutes les bases de données appartenant à la catégorie du « corpus de l'usage langagier expert » contiennent en principe des énoncés produits par des natifs ou des locuteurs à compétences linguistiques proches des natifs. Ce critère est aisément rempli quand il s'agit de la littérature, d'articles de journaux et d'autres sources « validées » par la communauté pour leur qualité. Pour autant, si le corpus est compilé à partir des pages Web, il est difficile d'obtenir des renseignements sur le niveau exact de compétences linguistiques des locuteurs. Dans le cas des langues moins enseignées, nous pouvons supposer que la majorité des énoncés a été produite par des natifs mais dans le cas des langues largement parlées, la réponse est moins claire : les énoncés peuvent provenir

des utilisateurs experts mais aussi des apprenants. En outre, ces corpus (même ceux compilés strictement à partir des manifestations langagières des natifs ou des utilisateurs experts) ne sont homogènes ni dans leur contenu ni quant à leur qualité langagière : un corpus peut ainsi comporter différents types de textes aux sujets multiples comme des textes plus ou moins bâclés. Par exemple, les contributions sur les réseaux sociaux (blogs, forums, commentaires, manifestations importantes de l'utilisation langagière informelle) contiennent souvent des fautes d'orthographe et des phrases maladroitement construites. Quelle valeur attribuer à ces textes ? Quelle est alors leur place dans l'usage linguistique « expert » ? Le chapitre 12 explorera ces questions dans le contexte pédagogique.

Dans les années 1980 une nouvelle catégorie de corpus a émergé, comprenant des produits linguistiques écrits et/ou oraux des apprenants (*learner corpora*). Ces corpus ont été créés afin de fournir des informations sur l'usage langagier des apprenants (Gilquin et Granger 2010, 2015). La possibilité de compiler des collections de taille significative a permis d'explorer ces textes avec les outils d'analyse de corpus et de révéler les difficultés particulières relatives à la langue-cible (par exemple, les difficultés les plus fréquemment identifiées chez les apprenants du français) ; celles liées au niveau des compétences linguistiques choisi (par exemple, les difficultés typiques des apprenants au niveau A2) ou les problèmes d'une population particulière (par exemple, celles des apprenants de hongrois dont la première langue est le chinois). Ces résultats peuvent être confrontés à ceux obtenus par des méthodes traditionnelles, qualitatives sur un plus petit nombre de textes pour mieux décrire les caractéristiques de l'usage langagier des apprenants. Comparer les résultats a d'autant plus d'intérêt que les deux méthodes mettent l'accent sur l'étude de phénomènes différents. Les études traditionnelles priorisent la morphologie et la grammaire, les études de corpus d'apprenants, en revanche, « se caractérisent par une forte concentration sur le lexique, la lexico-grammaire et une gamme de phénomènes de discours » (Gilquin et Granger 2015 : 420, notre traduction). Ces corpus peuvent être mis à disposition des utilisateurs sans annotation, sous la forme d'une collection simple de textes analysables avec les outils numériques ou, de plus en plus fréquemment, avec une annotation d'erreurs.

Alors que ses objectifs sont clairs, le contenu d'un corpus d'apprenants peut varier selon les critères définis par le chercheur. Ce contenu peut rassembler des textes d'étudiants dans la langue-cible dans un contexte similaire ou dans des contextes différents ; les données peuvent être collectées suivant une approche longitudinale ou de façon ponctuelle, au même niveau de compétences

linguistiques ou à différents niveaux, d'une ou de différentes nationalités en incluant une ou plusieurs langues (Gilquin et Granger 2015 : 419).

L'étude de ces corpus peut également servir à améliorer la qualité des matériels pédagogiques (voir les chapitres 5 et 6). L'analyse de ces textes peut indiquer des tendances d'utilisation d'un élément lexical, grammatical ou pragmatique par les apprenants ou, de façon plus large, l'efficacité d'une approche pédagogique (cf. Ellis et al. 2015 ; Ellis et Ogden 2017 ; Guilquin et Granger 2015 ; Nesselhauf 2005 ; O'Sullivan and Chambers 2006).

5) Les types de corpus en fonction de la période de temps choisie : corpus historiques (diachroniques) et corpus synchrones

Les corpus permettent d'étudier la langue choisie en se concentrant sur une ou plusieurs périodes. En fonction de ce choix, il convient de consulter un corpus diachronique ou un corpus synchrone. De nombreuses études au sein de la linguistique de corpus sont diachroniques, c'est-à-dire que le chercheur étudie l'utilisation de la langue à travers deux ou plusieurs périodes pour mesurer la présence éventuelle des différences dans le temps (Kytö et Smitterberg 2015 ; Hilpert et Mair 2015). Les *corpus diachroniques* sont donc « délibérément construits pour contenir du matériel linguistique contrastif » (Wynne 2004 : 7, notre traduction) dans le but de rendre possible l'observation de l'apparition et de la disparition des mots ainsi que celle des changements de fréquence dans l'utilisation de l'élément langagier étudié. Par conséquent, un critère impératif pour la composition des corpus diachroniques est la comparabilité. Selon Leech (2007 : 141), deux corpus sont comparables s'ils diffèrent en termes d'un seul paramètre ; dans le cas des corpus diachroniques, ce paramètre est le temps. Pour cette raison, les corpus diachroniques contiennent avant tout des genres textuels attestés pendant toutes les périodes échantillonnées.

Les *corpus synchrones* renferment des textes provenant de la même période. Ils illustrent l'usage langagier de la période choisie à travers un ou plusieurs genres[6] de textes. Néanmoins, la définition de « la même période » peut être problématique lorsque la société connaît des changements rapides et les mots apparaissent et disparaissent rapidement, car deux corpus enregistrés avec un écart de quelques années peuvent révéler des différences significatives. Le corpus de « frTenTen17 » en est un exemple pertinent. Ce corpus contient des textes publiés en français sur Internet avant 2017.

[6] Nous utiliserons ce terme comme défini par Biber et Conrad (2009 : 18): le registre décrit « les caractéristiques linguistiques typiques liées à un contexte situationnel spécifique » (*typical linguistic features associated with the situational context)*.

Dans ce corpus, le mot « Corona » fait référence à une bière légère, à des hôtels et à des personnes portant ce nom. Si l'on inclut dans ce corpus des textes de journaux publiés à partir de 2019, il est fort probable que ce mot sera avant tout associé à la pandémie provoquée par le virus Covid-19. D'autres éléments langagiers comme « Big Data » ont émergé et pratiquement disparu en quelques années.

6) Les types de corpus en fonction des langues incluses : corpus monolingues et multilingues

Les corpus peuvent être catégorisés en fonction du nombre de langues qu'ils contiennent. Une grande partie des corpus existants sont des « corpus monolingues » contenant des énoncés dans une seule langue. Un autre groupe représente les « corpus multilingues ou parallèles » qui comprennent les mêmes textes en plusieurs langues – en l'occurrence un texte original et sa traduction en une ou plusieurs autres langues. Les segments (phrases, paragraphes) correspondants sont alignés, l'utilisateur peut alors rechercher tous les exemples d'un mot ou d'une phrase dans une langue et les résultats seront affichés avec les phrases correspondantes dans l'autre langue. Ce mode de présentation permet de comparer les différentes langues. Les corpus multilingues facilitent considérablement l'étude des phénomènes interlinguistiques et peuvent fournir des données d'apprentissage pour les systèmes de traduction automatique. Ils sont également utiles pour la traduction et la linguistique comparative (cf. Kübler 2014a et b).

C) Considérations générales concernant la construction des corpus non pédagogiques

La construction d'un corpus est précédée par la définition des paramètres qui le caractériseront. Ces paramètres sont largement déterminés par les objectifs que le corpus doit remplir. Les questions les plus importantes à considérer sont la taille du corpus, l'authenticité, la représentativité et l'équilibre (Habert 2001). Les pages suivantes sont consacrées à la présentation succincte de ces critères essentiels sur lesquels nous reviendrons dans le chapitre 2 pour aborder les principes de construction des corpus pédagogiques.

1) La taille du corpus

La taille d'« [u]n corpus est en l'occurrence mesurée en termes de nombre de mots qu'il contient (…) [i.e. en terme du] nombre de tokens » (Jones et Waller 2015 : 5). Selon Webster et Kit (1992), le token est une entité en elle-même qui, lors des traitements ultérieurs, ne sera pas découpée en

unités plus petites. Ainsi, un token pourra être constitué de plusieurs mots, comme c'est le cas par exemple de certaines expressions idiomatiques (donner suite à quelque chose), des mots composés (pomme de terre) ou de certains chiffres (10 000). À l'inverse, un mot peut être décomposé en deux unités (par exemple *au* décomposé en *à le*) » (Actes du DiLiTal 2017). Davies (2015 : 11) divise les corpus en six catégories selon leurs tailles :

1. Petits corpus de première génération de 1 à 5 millions de tokens comme le Brown Corpus ;
2. Corpus de deuxième génération de taille moyenne, équilibrés par genre, tels que le British National Corpus de 100 millions de mots ;
3. Des corpus plus grands et actualisés régulièrement (toujours équilibrés par genre), comme le Corpus de 450 millions de mots de l'anglais américain contemporain (COCA) ;
4. Archives de textes volumineuses telles que Lexis-Nexis ;
5. Archives de textes extrêmement volumineuses telles que Google Livres ;
6. Le Web en tant que corpus exploré par un moteur de recherche comme Google[7].

Alors que l'anglais et, dans une plus petite mesure le français, donnent accès à une multitude de corpus, les langues moins parlées et enseignées se trouvent dans une situation bien moins avantageuse. Le hongrois, sujet d'exploration de cette thèse, dispose seulement de trois corpus de grande taille accessibles au public : le corpus « huTenTen12 », le Corpus national du hongrois et le corpus hongrois de la « Leipziger Datenbank ». Ces trois corpus font partie de la catégorie 2, c'est-à-dire des corpus de taille moyenne sans mise à jour régulière. Il existe également quelques petits corpus spécialisés comme « Childes » pour le langage des enfants, le corpus hongrois de textes juridiques dans le corpus multilingue « EUR-Lex » ou le corpus pédagogique accompagnant les manuels de la série « MagyarOK ». Nous disposons en outre de quelques grandes collections de la catégorie 4 comme le « Magyar Elektronikus Könyvtár » (Bibliothèque digitale hongroise) contenant des textes littéraires et non littéraires digitalisés et nous pouvons bien évidemment consulter les textes hongrois publiés sur le Web (catégorie 6). Un inconvénient majeur dans le cas de ces derniers corpus est cependant le manque d'outils pour une exploration systématique[8].

Il est important de noter que si les grands corpus tels que le British National Corpus (BNC) pour l'anglais, le Corpus national du hongrois (« Magyar Nemzeti Szövegtár » ou MNSZ) pour le

[7] Il convient de noter que les trois dernières collections ne sont pas équipées d'outils d'analyse de corpus mais sont seulement constituées de bases de données de textes authentiques de taille très significative.
[8] Pour une présentation détaillée des corpus utilisés pour notre étude du hongrois, voir le chapitre 7.

hongrois, le corpus « frTenTen17 » pour le français ou le corpus COSMA pour l'allemand, pour ne nommer que quelques exemples, se composent de centaines de millions de mots, la taille n'était qu'un des critères pour leur conception. Comme le soulignent O'Keeffe et al. (2007 : 4), la nature et l'objectif du corpus sont des aspects aussi importants que sa taille :

> « Pour les corpus de langue parlée, tout ce qui dépasse un million de mots est considéré comme important ; pour les corpus écrits, tout ce qui est inférieur à cinq millions est considéré comme assez petit. En termes de pertinence, cependant, c'est souvent le contenu et la structure d'un corpus, par opposition à sa taille, qui sont les facteurs déterminants. » (notre traduction)

Ainsi, la taille optimale dépend, avant tout, des objectifs scientifiques du créateur de corpus. Aujourd'hui nous disposons de ressources abondantes qui nous permettent de compiler des corpus (écrits) rapidement, tout en incluant une grande variété de textes. Il peut être difficile de résister à la tentation de compiler des corpus en fixant la taille comme seul critère, d'après la devise : « Plus c'est grand, mieux c'est ». Or, un très grand corpus n'est pas toujours le plus adapté aux objectifs de l'utilisateur : un corpus à contenu plus limité mais plus ciblé, peut offrir des avantages quand il s'agit d'identifier les caractéristiques d'un élément lexical dans un contexte particulier. Par exemple, pour identifier les caractéristiques des lettres de motivation françaises, il suffit d'un corpus plus limité qui ne contient rien d'autre que des lettres de motivation en cette langue, aucun argument ne justifiant la création d'un corpus plus grand et plus général (par exemple l'ajout des CV, d'offres d'emploi et d'autres textes).

La constitution d'un corpus implique d'autres critères comme la nature des données linguistiques à inclure, la question de l'authenticité et de la représentativité ainsi que celle de l'équilibre. Ces critères seront explorés dans les pages suivantes.

2) L'authenticité du corpus

Un des principes de base de la construction des corpus est que les textes doivent être rassemblés à partir des communications authentiques des personnes « ordinaires » occupées à leurs activités normales au quotidien (Sinclair 1997). Cela implique que les énoncés inclus dans la collection n'ont pas été générés à la demande du linguiste mais produits spontanément, sans consignes spécifiques. Ce principe sert à éliminer le biais potentiel qui serait de produire des énoncés dont le seul but est d'illustrer l'hypothèse du linguiste. Le principe de l'authenticité stipule également que les textes

contiennent tous les éléments nécessaires à leur compréhension et à la reconstruction du contexte dans lequel ils ont été produits (Brezina 2018 ; Stefanowitsch 2020). Un corpus écrit comprend, en principe, des textes destinés à être lus (œuvres littéraires, articles de journaux, etc.) qui définissent donc leur propre contexte. Par exemple dans un roman, les personnages et les lieux sont décrits avec précision pour permettre au lecteur de situer les événements. Les dialogues sont accompagnés par la description de la mimique, des gestes et des circonstances dans lesquelles les phrases sont prononcées. Le texte englobe ainsi tous les éléments qui sont nécessaires à son interprétation.

Il existe cependant d'autres types de textes écrits dont l'interprétation correcte nécessite des connaissances qui ne sont pas incluses dans le texte. Par exemple, une interaction écrite (un « *chat* ») entre amis ne contiendra pas les informations connues de tous les locuteurs si le message est clair sans cela et si le texte est parfaitement interprétable. En revanche, si un tel texte est intégré dans un corpus, le contexte et les connaissances implicites y manqueront et cela rendra la recontextualisation problématique. En suivant cette ligne de pensée, Widdowson (1978 : 80) distingue deux types d'authenticité : dans le premier cas, il s'agit de textes originaux, non modifiés, remplissant le critère du « genuineness ». Ce critère caractérise le texte lui-même et est une qualité absolue. La deuxième catégorie, que Widdowson nomme « authenticity », décrit la relation entre texte et lecteur/locuteur et elle concerne la recontextualisation correcte du message.

Le critère de l'authenticité en tant que possibilité de recontextualiser les énoncés, s'avère particulièrement problématique dans le cas des corpus contenant des transcriptions de textes oraux. Ici, l'environnement textuel ou le « co-texte », pour utiliser le terme de Sinclair (1991, 2004a), est présent mais le contexte (tout ce qui est extérieur au langage) manque entièrement. Dans quelle mesure peut-on donc qualifier les données linguistiques dans ce type de corpus comme « authentiques » et recontextualisables ? En reprenant une analogie botanique émise par Sinclair (1991 : 6), O'Keeffe (2000, cité dans Timmis 2015 : 139) note aussi :

> « [D]ans le cas de l'étude des interactions orales transcrites sur papier, il s'agit de fleurs séchées, dépourvues de leurs contextes et de leurs environnements naturels. Et puisque ces interactions sont « fanées », […] nous sommes confrontés au danger implicite de leur mauvaise interprétation sémantique et pragmatique. » (notre traduction)

Suivant la logique de cette argumentation, plusieurs chercheurs partagent le point de vue que les données dans un corpus de transcriptions peuvent être considérées comme originales et non modifiées (*genuine*) mais non authentiques (*authentic*) (Flowerdew 2015 ; O'Keeffe et al. 2007 ; Timmis 2015 ; Widdowson 1978, 2003).

3) Représentativité et équilibre

Une définition largement admise de la représentativité est fournie par Manning et Schütze (1999 : 119) qui déclarent qu'un corpus est considéré comme représentatif si les résultats obtenus lors de son analyse sont généralisables à l'ensemble de la langue étudiée. Sans représentativité, comme l'exprime Leech (2007 : 135), « tout ce qui est jugé vrai d'un corpus est simplement vrai de ce corpus – et ne peut être étendu à rien d'autre » (notre traduction). Egbert et al. (2020 : 5) nous rappellent par ailleurs qu'un corpus sert « d'indicateur pour un domaine linguistique d'intérêt », que nous l'utilisons « avec l'espoir de pouvoir identifier (…) des informations que l'on peut étendre à l'utilisation de la langue dans le domaine concerné » (notre traduction). Bien que la représentativité soit un critère important pour la construction de corpus, elle reste néanmoins une notion vague. Cela n'a rien de surprenant face à la difficulté de fournir une définition des termes comme « l'ensemble du langage étudié » ou « la généralisabilité des informations ». Quant à la construction de corpus représentatifs, deux approches principales se dégagent liées au processus de collecte des données que nous présenterons brièvement dans les paragraphes suivants.

La première approche préconise la construction des « corpus monitorés » (*monitor corpus*) (Sinclair 1991 : 24–26). Ces corpus prennent de plus en plus d'ampleur au fil du temps en incluant de plus en plus de textes. L'un des exemples connus de ce type de corpus est la « Bank of English » créée dans les années 1980 par l'Université de Birmingham. Ce corpus a été continuellement enrichi depuis et comprenait plus de 500 000 000 de mots au début des années 2000 (Hunston 2002 : 15) et renferment actuellement 4 500 000 000 de mots. Une caractéristique des corpus monitorés comme celui-ci est que les proportions relatives des différents types de données linguistiques (articles scientifiques, textes littéraires, communications personnelles, etc.) peuvent varier avec le temps car l'équilibre des genres dans le corpus n'est pas un principe essentiel de leur construction.

Contrairement aux corpus monitorés, les « corpus équilibrés » (*balanced corpora*) ou les « corpus d'échantillons » (*sample corpora*) tentent de représenter un type de langage particulier sur une période de temps donnée (Biber 1993 ; Leech 2007). Les créateurs de tels corpus ne prévoient pas d'augmenter le matériel linguistique en continu, mais préconisent un échantillonage soigneux et

spécifique dès le départ, reflétant la langue ou une partie de la langue telle qu'elle existe à un moment donné (McEnery 2013 : 6). Un exemple d'une telle approche est donné par le corpus Lancaster-Oslo/Bergen, qui représente un « instantané » du langage écrit standard de l'anglais britannique moderne au début des années 60 en s'appuyant sur des échantillons de 2 000 mots.

Ces deux approches ont chacune l'ambition d'assurer que les bases de données ne sont pas biaisées, mais l'une et l'autre choisissent deux chemins différents. Le créateur du corpus monitoré cherche à ne pas interférer avec le contenu de sa collection : il n'effectue aucun échantillonnage qui serait, pour lui, l'équivalent d'une « sélection subjective » (Váradi 2002), donc source d'un biais. En revanche, le créateur du corpus équilibré maintient qu'un jugement d'expert est nécessaire pour décider quels textes sont aptes à faire partie de la collection. Selon lui, ce n'est qu'ainsi que l'on peut éviter le déséquilibre dans le corpus, ce qui serait aussi une forme de biais.

Une combinaison des deux approches est représentée par le Corpus de l'anglais américain contemporain (« Corpus of Contemporary American English », Davies 2009, 2015). Le COCA s'enrichit au fil du temps (cela le rend identique aux corpus monitorés), mais il le fait selon une conception explicite (ce qui l'apparente aux corpus équilibrés). Chaque section ajoutée au COCA est conforme à la même répartition d'ensemble des variétés de textes. Ce corpus est donc conçu selon les principes d'échantillonnage mais il est également évolutif dans le temps, ce qui semble être une manière efficace de composer des corpus représentatifs et équilibrés.

Sinclair (1991 : 25) distingue un troisième type de corpus qu'il appelle des « corpus opportunistes » (*opportunistic corpora*). Ces corpus n'adhèrent pas à des principes d'échantillonnage rigoureux et renferment principalement des données collectées pour une tâche spécifique. Aujourd'hui, il est souvent nécessaire d'appliquer cette approche quand il s'agit des corpus oraux dont la construction repose sur un nombre de compromis, comme nous l'exposerons dans les chapitres 2, 13 et 14 de cette thèse.

4) Le corpus : une collection d'expériences linguistiques

Avant de résumer les arguments principaux en faveur de l'utilisation des corpus, il convient de se rappeler que tout corpus contient d'abord une chose : ce que ses créateurs ont décidé d'y inclure. Ce constat peut sembler tautologique, mais il nous avertit du fait que la construction d'un corpus repose sur des décisions humaines. Par conséquent, chaque corpus a ses caractéristiques, son profil

et sa « personnalité » et aucun corpus ne peut intégrer la totalité du langage, quel que soit le nombre de mots qu'il comprend. Comme Cook (1998 : 59) le souligne :

> « Les corpus ne sont que des autorités partielles. L'expérience cumulative d'un individu, bien que moins accessible à un accès systématique, reste beaucoup plus vaste et plus riche. Même un corpus de trois cent millions de mots ne représente que trois mille livres environ, ou ne peut-être que l'expérience linguistique d'un adolescent » (notre traduction).

Bien qu'aujourd'hui certains corpus comptent bien plus de trois cent millions de mots, l'argument de Cook est toujours valide : un grand corpus reste « un instantané partiel de la langue capturée et prise à une époque » (Jones et Waller 2015 : 15). Concernant la comparabilité des corpus, Egbert et al. constatent également (2020 : 4-5) qu'« [e]n pratique, nous sommes rarement confrontés à une décision entre deux corpus aux conceptions identiques (...) « toutes les caractéristiques ne sont presque jamais égales » (notre traduction) et que l'analyse de deux corpus à paramètres similaires ne fournira jamais exactement les mêmes résultats[9]. En outre, même un corpus homogène à première vue peut contenir des textes qui illustrent des usages langagiers variés. Un corpus compilé à partir du même journal, par exemple, ne sera pas uniforme du simple fait que tous les articles proviennent de la même source : ses différentes sections (sports, nouvelles internationales, économie) refléteront des usages et, à l'occasion, des styles différents (McEnery et al. 2006).

Face à ces considérations, il convient de nous interroger sur les gains qu'offre la consultation de corpus. Nous partageons le point de vue de Hoey (2009 : 37) qui déclare que *les corpus contiennent des échantillons langagiers liés à différentes situations auxquelles les locuteurs sont exposés au cours de leur vie*. Il est vrai, dit-il, qu'aucun corpus ne peut jamais être complet car même le plus grand ne présentera pas toutes les phrases jamais dites et écrites dans une langue ; ces ensembles de données peuvent tout de même offrir *des informations précieuses sur* « *les expériences linguistiques probables des locuteurs* » (notre traduction). Ainsi, les analyses statistiques d'un corpus suffisamment grand et bien construit permettent d'observer les caractéristiques de l'usage langagier des natifs. Il est clair que les corpus ne seront jamais des autorités absolues, néanmoins ils peuvent fournir des indications précieuses

[9] Dans la Partie II de cette thèse, nous comparerons les résultats issus de deux grands corpus de hongrois. Nous constaterons que ces résultats sont très proches. Ceci indique que les corpus compilés de façon similaire mais contenant des textes différents sont susceptibles de révéler des informations similaires et que, dans certains cas au moins, le deuxième corpus peut être utilisé pour vérifier les résultats obtenus à partir du premier.

sur le comportement linguistique des locuteurs. De plus, les résultats d'analyse seront en termes de leur quantité et de leur qualité, largement supérieurs à ceux obtenus sans étude des énoncés réels et authentiques. « Un corpus ne nous fournit pas simplement des réponses sur la langue utilisée, mais il nous donne *une base d'indices qui peuvent solliciter de meilleures descriptions de la langue* », comme le formulent Jones et Waller (2015 : 16, notre traduction, nous soulignons). Il est cependant important de noter que le corpus *ne contient pas de renseignements concernant les raisons d'une utilisation particulière* ; il incombe donc au linguiste de trouver des interprétations valides soutenant les résultats de son analyse (cf. Brezina 2018 ; Egbert et al. 2020 ; Stefanowitsch 2020).

Les avantages principaux que l'utilisation des corpus peut offrir se résument ainsi aux points suivants :

- L'analyse de corpus *fournit des informations statistiques sur la langue ou sur une variété particulière langagière d'une part et des manières de présentation de la langue se prêtant à l'analyse qualitative de l'autre.*
- L'intérêt principal de ces explorations est de parvenir à une description plus précise de la langue ainsi que de développer de nouvelles théories du langage fondées sur les résultats de l'analyse.
- L'analyse outillée offre des avantages significatifs : certains phénomènes langagiers peuvent échapper à l'observation simple et à l'intuition mais ils deviennent visibles quand la langue est présentée à l'aide des logiciels, par exemple en utilisant des lignes de concordance (Barlow 2004).
- Cette présentation de la langue fait émerger des schémas (*patterns*), c'est-à-dire des répétitions lexicales et/ou structurales et donne des renseignements sur leurs environnements textuels ainsi que sur leur fréquence (Biber et al. 1998 ; Hunston 2010 ; Stefanowitsch 2020).

Ce chapitre a exposé les principes de base de la linguistique de corpus et présenté, de façon succincte, ses différentes écoles. Il a également mis en évidence les considérations liées à la catégorisation des corpus ainsi que les principes les plus importants de leur construction. Cependant, les corpus ne servent pas seulement la communauté des linguistes. Connaître « les expériences linguistiques probables des locuteurs » a aussi une valeur inestimable pour les professionnels de l'enseignement des langues où la présentation et la pratique de l'usage langagier des natifs jouent un rôle déterminant (cf. Boers 2021 ; Robinson et Ellis 2008 ; Hoey 2005 ; Nation

2015 ; Pérez-Paredes 2021 ; Pérez-Paredes et al. 2020 ; Taylor 2012). Les résultats d'analyses peuvent être ainsi intégrés dans des ouvrages pédagogiques (dictionnaires, grammaires, manuels de cours) pour augmenter leur efficacité et les corpus pertinents eux-mêmes peuvent être exploités dans le cadre pédagogique. Les « corpus pédagogiques », catégorie présentée brièvement plus haut, jouent un rôle crucial dans cette approche. Dans le chapitre suivant, nous explorerons les besoins auxquels ceux-ci doivent répondre ainsi que les critères spécifiques liés à leur construction.

Chapitre 2 : Construire le corpus pédagogique

Ce chapitre présentera les principales considérations inhérentes à la conception de corpus pédagogiques pour les niveaux de compétences linguistiques inférieurs (niveaux A1, A2 et B1). Nous explorerons plus précisément les questions suivantes :

- Pourquoi les corpus non pédagogiques existants ne sont-ils que partiellement adaptés à l'enseignement des langues ?
- Comment construire des corpus pédagogiques qui tiennent compte des besoins des apprenants ?
- Comment collecter, où collecter et que collecter pour ces corpus ?

Ce chapitre se divise en deux parties. Dans un premier temps, nous discuterons de l'intérêt des corpus pédagogiques en général ; nous exposerons par la suite les principes de base de leur construction.

A) L'intérêt des corpus pédagogiques

1) Les problèmes des corpus non pédagogiques

Le domaine de la linguistique de corpus a rapidement évolué au cours des dernières décennies. En conséquence, de nouvelles méthodes de recherche ainsi que de nouvelles manières de collecter des données pour les dictionnaires et les descriptions grammaticales sont apparues. Ces développements ont également eu un impact sur l'apprentissage des langues : les résultats de recherche concernant l'utilisation langagière dans des situations authentiques ont été intégrés dans certains manuels[10] ; de nombreuses activités ont été créées pour permettre aux enseignants et aux apprenants d'explorer la langue-cible en consultant des bases de données langagières avec des outils d'analyse de corpus. En dépit de ces avancées, *les considérations principales dans le domaine de la pédagogie de corpus semblent être avant tout liées aux solutions techniques et non à la conception d'une méthodologie ou à la création de corpus spécifiques pour l'enseignement des langues*. Invariablement, la même recommandation est donnée à ceux qui souhaitent construire de tels corpus : à l'instar des linguistes, ils devraient compiler des collections de textes (1) authentiques, (2) organisés par sujet, genre et/ou registre (cf. Bernardini 2004 ; Charles 2007, 2014 ; Hunston 2002, 2009).

[10] Voir le chapitre 6 pour deux exemples.

Or, auteurs de manuels, apprenants et enseignants consultent les ressources linguistiques avec des objectifs différents de ceux des linguistes : *plutôt que d'étudier la langue pour mieux en décrire ses caractéristiques, ils souhaitent trouver des réponses fiables aux questions qu'ils se posent sur son utilisation ainsi que des exemples authentiques illustrant ses spécificités* (Flowerdew J. 2009 ; Flowerdew L. 2009, 2015 ; Charles 2007, 2014). Par conséquent, le groupe d'utilisateurs est loin d'être homogène quant à ce qu'ils attendent d'un corpus. Les objectifs respectifs des différents groupes sont résumés dans le tableau ci-dessous (tableau 1) (basé sur Boulton et Thomas 2012 : 7 et Römer 2006) :

Tableau 1 : Les objectifs des différents groupes d'utilisateurs.

Sachant que les objectifs pour lesquels ces groupes consultent les corpus divergent, il est logique de supposer que *le type de corpus répondant à leurs besoins spécifiques ne peut pas être le même*. Comme le note Reppen (2016 : 410), dans un contexte pédagogique :

> « Le principe directeur de l'utilisation de corpus ou de ressources de corpus dans un contexte d'apprentissage des langues est d'améliorer l'apprentissage. Ceci peut être accompli grâce à des activités qui sensibilisent l'apprenant aux différents contextes d'utilisation de la langue et également à travers des activités qui favorisent l'autonomie de l'apprenant. La simple existence d'un corpus n'en fait pas nécessairement un outil approprié pour l'apprentissage des langues » (notre traduction).

Braun (2007 : 308) remarque également que de nombreux corpus largement accessibles et recommandés pour l'enseignement des langues « ont été créés dans une perspective de recherche linguistique et non avec des objectifs pédagogiques, de sorte que *leur contenu et leur conception ne répondent pas nécessairement aux besoins pédagogiques* » (notre traduction, nous soulignons). Le problème principal des corpus non pédagogiques est qu'ils contiennent une *multitude de sujets et de registres ainsi qu'un langage complexe, difficilement accessibles aux apprenants,* surtout aux niveaux inférieurs. Les considérations sur le style et sur le niveau langagier ne font pas partie des principes de construction de ces corpus, comme nous l'avons vu au chapitre 1. Par exemple, ces collections ne font pas la distinction entre langage familier et langage formel, standard ou délibérément non standard. De nombreuses occurrences peuvent mettre au défi la compréhension des apprenants face à la complexité, l'idiomaticité, le manque d'environnement textuel plus large et de contexte situationnel des textes présentés dans de tels corpus (cf. Aston 2001 ; Braun 2006, 2007 ; Meunier 2012). En outre, les contenus de certains énoncés peuvent être problématiques : ils peuvent exprimer des opinions discutables du point de vue éthique, ce qui les rend inutilisables en tant qu'exemples langagiers dans un cadre pédagogique. Ces facteurs peuvent expliquer pourquoi les corpus non pédagogiques existants ont été accueillis avec un enthousiasme modéré par les enseignants et les apprenants en langues (cf. Ädel et Reppen 2008 ; Bernardini 2004 ; Cavalla et Loiseau 2013 ; Chambers 2019 ; Römer 2006).

2) Le point de vue de l'apprenant comme point de départ

Les inconvénients évoqués ci-dessus ont conduit certains linguistes-didacticiens à repenser le concept et le contenu du « corpus pédagogique » (terme crée par Willis 2003). Une partie des experts incite les enseignants et les apprenants à compiler de nouvelles bases de données en fonction de leurs besoins, en définissant eux-mêmes les paramètres adéquats (cf. Aston, 1997, 2002 ; Charles 2005 ; Tribble et Jones 1997). La création de telles bases de données présente plusieurs avantages. D'une part, la collecte et l'analyse des données sont toutes deux guidées par des

considérations pédagogiques dès le départ. D'autre part, les créateurs ont davantage de contrôle sur la pertinence et l'accessibilité du contenu.

Une autre manière de réaliser des corpus pédagogiques, également suggérée par plusieurs chercheurs, est de sélectionner un sous-ensemble du matériel linguistique à partir d'un grand corpus non pédagogique et de mettre les textes dans leur intégralité à la disposition des apprenants (cf. Aston 2002 ; Braun 2006 ; 2007 ; Chambers 2019 ; Flowerdew L. 2009 ; Kennedy et Miceli 2017). Ces corpus spécifiques, de taille réduite, peuvent être organisés par genre et/ou sujet (cf. Henry & Roseberry 2001 ; Tribble 1997, 2001). « L'avantage de créer des corpus écrits et oraux spécifiquement destinés à l'enseignement est que ces corpus peuvent être conçus avec un objectif clair », comme le fait remarquer Friginal (2018 : 24, notre traduction). Ils peuvent ainsi offrir un apport linguistique pertinent pour les apprenants. Ce processus permet également une certaine liberté et flexibilité dans le choix des textes. De telles collections ont été créées à partir de lettres commerciales (Flowerdew 2015), de publications dans divers domaines académiques (Charles 2015 ; Friginal 2018 et d'autres) ou de récits reprenant des sujets identifiés dans le CECRL (Kennedy et Miceli 2010, 2017), pour ne citer que quelques exemples.

Ces deux approches se ressemblent en ce qu'elles proposent une réduction de la taille du corpus tout en conservant la forme originale, non modifiée des textes sélectionnés, en accord avec le principe d'authenticité tel que défini dans le chapitre 1. Or, comme nous l'avons évoqué précédemment, le principal problème concernant les corpus existants, non pédagogiques est que leur langage est bien plus complexe que ce que les apprenants, avant tout aux niveaux inférieurs, doivent (et sont capables de) comprendre. Les blogs sur la gastronomie, par exemple, peuvent sembler des sources appropriées en termes de contenu car les « habitudes alimentaires » font partie des thèmes définis par le CECRL, mais ils sont néanmoins susceptibles d'inclure des éléments allant bien au-delà des compétences linguistiques des apprenants. Ces obstacles ont incité certains chercheurs à repenser les principes de construction de corpus. Braun (2010 : 82) suggère, par exemple, que les corpus répondant aux besoins des apprenants doivent être construits par des chercheurs qui « prennent le point de vue de l'apprenant comme point de départ » (notre traduction).

Comment alors respecter ce point de vue ? Pour répondre à cette question, il convient tout d'abord de considérer la fonction du corpus pédagogique en le comparant avec un corpus construit à fins linguistiques. Leur relation est identique à celle existant entre les livres dits « normaux » et les

manuels de cours. Tout comme les manuels pour les niveaux inférieurs avec leur vocabulaire limité et leurs textes simples représentent une catégorie particulière de livres, les corpus conçus pour l'enseignement des langues peuvent être considérés comme une catégorie spéciale de corpus. Les manuels ont pour objectif d'introduire petit à petit les éléments de la nouvelle langue pour augmenter graduellement les compétences linguistiques des apprenants. Aucun utilisateur n'attend des manuels une présentation complète de la langue-cible dès le départ. Ce constat s'applique également aux corpus pédagogiques : *pour qu'ils puissent remplir leur fonction, ils ne peuvent contenir des textes écrits dans la langue-cible sans qu'aucune présélection n'ait été faite.*

Quel est donc le rôle de ces corpus pour l'enseignement des langues aux niveaux de compétence inférieurs ? Nous sommes d'avis qu'*il est d'une importance cruciale que les apprenants se familiarisent avec les corpus et les outils d'analyse dès les premières étapes de leur apprentissage si nous souhaitons que la consultation de ces bases de données leur vienne aussi naturellement que l'utilisation des applications, des vidéos et d'autres ressources*. Ils doivent « grandir » avec des corpus car ce n'est que par une pratique continue que les bénéfices de ce travail deviennent perceptibles et incitent les apprenants à continuer à utiliser ces ressources tout au long de leur parcours.

B) Construire des corpus pédagogiques pour les niveaux de compétences linguistiques inférieurs

Dans cette section, nous étudierons en détail les principes de construction pour le corpus pédagogique aux niveaux de compétences linguistiques inférieurs. Dans un premier temps, nous proposerons une définition des niveaux de compétences linguistiques inférieurs selon le Cadre européen commun de référence pour les langues (CECRL). Cette partie sera suivie d'une présentation des thèmes que le corpus doit contenir ainsi que des questions concernant les principes de construction de corpus.

1) Définition des niveaux de compétences linguistiques inférieurs

Avant d'exposer les principes recommandés pour la conception de corpus pédagogique, nous devons définir ce que l'on entend par « niveaux de compétences linguistiques inférieurs ». À cette fin, nous utiliserons les descripteurs de niveaux du CECRL. Ces descripteurs définissent à la fois la gamme de sujets et le niveau de complexité langagière que les apprenants sont censés maîtriser, deux axes qui doivent déterminer le contenu du corpus. Le tableau suivant expose brièvement les descriptions des trois niveaux.

Le candidat de niveau A1 :

- Peut comprendre et utiliser des expressions familières et quotidiennes ainsi que des énoncés très simples qui visent à satisfaire des besoins concrets.
- Peut se présenter ou présenter quelqu'un et poser à une personne des questions la concernant – par exemple sur son lieu d'habitation, ses relations, ce qui lui appartient, etc. - et peut répondre aux mêmes types de questions.
- Peut communiquer de façon simple si l'interlocuteur parle lentement et distinctement et se montre coopératif.

Le candidat de niveau A2 :

- Peut comprendre des phrases isolées et des expressions fréquemment utilisées en relation avec des domaines de priorité immédiats (par exemple, informations personnelles et familiales simples, achats, environnement proche, travail).
- Peut communiquer lors de tâches simples et habituelles ne demandant qu'un échange d'informations simple et direct sur des sujets familiers et habituels.
- Peut décrire avec des moyens simples sa formation, son environnement immédiat et évoquer des sujets qui correspondent à des besoins immédiats.

Le candidat de niveau B1 :

- Peut comprendre les points essentiels quand un langage clair et standard est utilisé et quand il s'agit de situations familières dans le travail, à l'école, dans les loisirs, etc.
- Est à même de répondre à la plupart des situations rencontrées en voyage dans une région où la langue-cible est parlée.
- Peut produire un discours simple et cohérent sur des sujets familiers et dans ses domaines d'intérêt.
- Peut raconter un événement, une expérience ou un rêve, décrire un espoir ou un but et exposer brièvement des raisons ou explications pour un projet ou une idée.

Il ressort de la description ci-dessus que le trait commun des apprenants aux niveaux A1, A2 et B1, soit aux niveaux de compétences linguistiques inférieurs, est qu'*ils ne sont pas des utilisateurs autonomes de la langue-cible*. Le CECRL propose également un certain nombre de descripteurs liés aux tâches communicatives et aux situations que les apprenants sont censés maîtriser au terme de chaque niveau. Selon ces descripteurs, les apprenants des niveaux A1, A2 et B1 doivent être capables de participer à des « interactions de la vie quotidienne simples et routinières » (Conseil de

l'Europe, mis à jour en juillet 2021). Afin de fournir du matériel linguistique *pertinent, authentique et accessible* pour ces niveaux, le corpus doit donc traiter des « situations simples du quotidien » liées « aux tâches habituelles » et aux « sujets familiers et courants » en utilisant un langage « clair et standard ».

Comment composer des corpus susceptibles de remplir ces critères ? Nous avons noté dans le chapitre 1 que l'un des critères les plus importants de la construction des corpus non pédagogiques est l'authenticité : ils doivent donc contenir des énoncés non modifiés et collectés sans que le linguiste n'interfère avec leur production (le langage produit ne doit pas être sollicité dans le but d'illustrer un phénomène linguistique). Respecter ce principe est primordial dans le cas des corpus à fins linguistiques, mais sa stricte implémentation fait émerger des questions dans le cas des corpus pédagogiques. Dans l'intérêt de l'accessibilité, les corpus pour les niveaux inférieurs doivent comprendre un langage assez simple lié aux sujets déterminés par le CECRL. Il est cependant impossible d'enregistrer des récits et des interactions authentiques qui correspondent *exactement* à ce que les apprenants sont censés comprendre à un niveau donné. Ainsi, pour présenter un niveau langagier approprié, il apparaît nécessaire d'appliquer un contrôle rigoureux du processus de collecte des données et des modifications langagières faisant suite à l'enregistrement[11].

Dans la partie suivante, des corpus pédagogiques existants seront utilisés pour illustrer les principes généraux de construction de base de données linguistiques pour les niveaux inférieurs. Les corpus pour l'enseignement des langues, fondés sur une réflexion méthodologique sont plutôt rares (Ädel 2010 ; Braun 2007 ; Breyer 2011 ; Chang 2014) : les exemples auxquels nous nous référerons par la suite incluent le corpus multilingue « Sacodeyl » (Chambers 2019), le corpus « ELISA » (Braun 2006, 2007) et le corpus « Backbone » pour l'anglais (Chambers 2019), le corpus « CWIC » pour les apprenants de l'italien (Kennedy et Miceli 2010, 2017) et « FLEURON », le corpus de l'Université de Nancy pour les apprenants du français (André 2017, 2019, 2020a et 2020b ; André et Ciekanski 2018).

2) Considérations concernant la construction du corpus pédagogique

Dans cette section, nous étudierons le contenu du corpus pédagogique ainsi que l'applicabilité des principes définis pour la construction des corpus à fins pédagogiques. Les différentes sous-sections exploreront les questions suivantes : (1) le lieu et le mode de la collecte ; (2) l'importance de la

[11] Ce point sera traité plus en détail aux chapitres 13 et 14 de la Partie III.

cohérence intertextuelle ; (3) l'authenticité et l'accessibilité langagières dans le corpus pédagogique ; (4) la taille et la représentativité du corpus et (5) la manière de présenter le contenu du corpus.

2.1) Le contenu du corpus : le lieu et le mode de la collecte

Le CECRL donne non seulement une description générale des différents niveaux de compétences linguistiques mais indique aussi les types de textes et les thèmes que les apprenants sont censés maîtriser. Ceux-ci peuvent aider à organiser le matériel du corpus pédagogique et à déterminer la nature des textes à inclure. Le tableau 2 présente les thèmes pour les niveaux A1, A2 et B1.

Niveau A1	Niveau A2	Niveau B1
Informations personnelles	Informations personnelles	Informations personnelles
Vie familiale	Vie familiale	Activités dans une ville et à la campagne
Vie quotidienne	Vie quotidienne	Vie familiale
Activités dans une ville et à la campagne	Activités dans une ville et à la campagne	Loisirs
Habitudes alimentaires	Habitudes alimentaires	Études et vie professionnelle
Voyage et transport	Voyage et transport	Consommation
Loisirs	Loisirs	Voyage et transport
Vacances et fêtes	Vacances et fêtes	Films et littérature
Consommation	Consommation	Média, réseaux sociaux
Études et vie professionnelle	Études et vie professionnelle	Environnement

Tableau 2 : Sujets pour les niveaux A1, A2 et B1 du CECRL.

Il ressort de ce tableau que la majorité des sujets (les informations personnelles, les loisirs, la vie quotidienne, la vie professionnelle, l'éducation et les voyages) sont abordés aux trois niveaux. Cette répétition thématique offre une liberté et une flexibilité aux auteurs de manuels ainsi qu'aux créateurs de corpus pour décider du contenu à introduire à chaque niveau.

Ces sujets sont, en effet, abordés dans la plupart des manuels de niveaux A1, A2 et B1. Pourquoi donc construire des corpus supplémentaires ? La raison est simple : même si les situations et les éléments linguistiques présentés dans ces livres sont pertinents, *le nombre d'occurrences d'un élément langagier choisi restera toujours limité en raison de la longueur du manuel. Des corpus pédagogiques peuvent compléter*

le contenu du manuel et augmenter ainsi le nombre de rencontres avec des éléments-clés de la langue-cible, liés à des situations précises.

Où collecter du matériel approprié pour ces corpus ? Tous les récits publiés sur Internet peuvent, en effet, prétendre y trouver leur place. Ces documents serviront alors de « matière brute » pour le corpus pédagogique écrit. Des textes narratifs (textes dans les blogs, recensions et avis sur les restaurants, les films ou les logements, etc.) sont relativement faciles à collecter et leur volume permet une étape de sélection. En revanche, ces textes ne conviennent pas sous leur forme originale pour un corpus pédagogique et doivent être modifiés pour garantir leur accessibilité (voir 2.3 ci-dessous et les chapitres 13 et 14 dans la Partie III de cette thèse).

Les situations de communication pertinentes pour les niveaux A1–B1 comprennent des « interactions de la vie quotidienne simples et routinières » majoritairement informelles et, dans une plus petite mesure, formelles. Or, la collecte de telles interactions soulève un certain nombre de questions pratiques et théoriques. La première question concerne les lieux de collecte, car ces types de textes sont rares dans les corpus existants et, plus généralement, sur Internet. La plupart de ces communications sont de nature privée (échanges de mails, « *chats* », messages audio ou audio-visuels) et ne sont accessibles qu'aux participants. De plus, certains types d'interactions du quotidien sont rarement réalisés sur Internet ; par exemple, il est difficile d'y trouver des conversations dans lesquelles quelqu'un achète de la nourriture ou explique quelle coupe de cheveux il souhaite avoir – pour la simple raison que ces interactions se déroulent majoritairement à l'oral, dans des lieux bien précis comme dans un magasin ou chez le coiffeur.

À première vue, les contributions sur les réseaux sociaux librement accessibles (commentaires, publications, blogs, forums) peuvent sembler une alternative acceptable permettant au chercheur de récupérer rapidement une grande quantité de contributions qui évoquent, au moins en partie, les interactions de la vie quotidienne. En réalité, la communication sur les réseaux sociaux suit des règles particulières, différentes de la communication « normale », en dehors de la communication par Internet (Cardon 2013 ; Domonkosi 2018a, 2018b). Par exemple, les locuteurs ont tendance à prendre plus de liberté lorsqu'ils communiquent avec des inconnus, les critiques peuvent être plus sévères et les compliments plus exagérés que dans la vraie vie où les interlocuteurs se trouvent face à face. Il est indubitable que ces sources sont judicieuses pour les apprenants d'aujourd'hui et, par conséquent, méritent de former un sous-ensemble dans le corpus pédagogique mais un corpus entièrement fondé sur des sources issues de réseaux sociaux ne peut remplacer une base de

données de conversations réelles car la nature des interactions n'est que partiellement comparable (Page et al. 2014, Chapter 2). Ces interactions doivent être enregistrées sur place, en présentiel, démarche qui est loin d'être simple.

Les interactions entre client et fournisseur de services – un type de dialogues extrêmement important pour les niveaux inférieurs – sont particulièrement difficiles à collecter. Tout d'abord, ils impliquent des problèmes de confidentialité. Par exemple, lors de la collecte des interactions dans des lieux de service, les commerçants doivent être convaincus que les créateurs de corpus « ne les espionnent pas » et qu'ils utiliseront les données uniquement à des fins de recherche. Les clients qui entrent dans la boutique pour effectuer un achat ne sont pas toujours ravis d'être enregistrés. Beaucoup d'entre eux ne connaissent pas la notion de corpus et l'explication peut prendre un temps considérable et est susceptible d'affecter l'humeur du client, surtout s'il est pressé. Ainsi, le succès de la collecte de données et le temps que le créateur du corpus doit y investir dépendent largement de la bonne volonté des participants et le caractère naturel et authentique peut en être affecté.

Ces constats s'appliquent également aux communications personnelles, écrites ou parlées. Il est difficile d'obtenir des contributions écrites des personnes que le créateur de corpus ne connaît que de façon superficielle car il s'agit de documents susceptibles de contenir des informations privées. Ainsi, lors de la collecte des textes écrits, les constructeurs de corpus doivent fréquemment solliciter leurs proches. Cependant, même eux sont souvent réticents à mettre leurs communications personnelles à la disposition du chercheur, bien que les échanges soient anonymisés dans le corpus (Kennedy et Miceli 2010). Enfin, un autre type de problème se pose lors des interactions orales : les locuteurs sont en effet souvent embarrassés et manquent de spontanéité quand ils se savent enregistrés. Quelques solutions à ces problèmes, comme l'improvisation d'acteurs, seront proposées dans le chapitre 14.

2.2) La cohérence intertextuelle

De même que pour certains corpus spécialisés, il existe un principe pertinent pour l'organisation du corpus pédagogique : *celui de la cohérence intertextuelle*. Ce type de cohérence « est généré par *un thème général commun dont relèvent tous les textes du [sous-]corpus* » (Braun 2006 : 53, notre traduction, nous soulignons). Dans les corpus présentés, la cohérence intertextuelle est garantie du fait que tous les participants parlent des mêmes sujets. Cette démarche thématique a l'avantage d'être transparente, de faciliter la recontextualisation et de faire émerger le vocabulaire-clé lié aux sujets

d'intérêt ainsi que des expressions à fonctions communicatives (hésiter, donner une réponse négative polie) grâce aux répétitions présentes au travers de plusieurs textes. Les rencontres répétées contribuent également à une meilleure compréhension du sens du mot ou de l'expression en question (plusieurs exemples montrent le fonctionnement du mot dans des contextes similaires) ainsi qu'à la mémorisation (Boers 2021 ; Ellis N. C. 2002, 2006, 2008 ; Ellis et Ogden 2017 ; Frankenberg-García 2012 ; Hoey 2005 ; Nation 2013 ; Schmid 2016 ; Taylor 2012 ; Trofimovich et McDonough 2011).

Dans un souci de transparence et de contrôle du contenu, les créateurs des corpus « Backbone », « Sacodeyl » et « ELISA » ont opté pour des entretiens structurés (Braun 2006, 2007 ; Chambers 2019). Les personnes interrogées parlent de leurs expériences et abordent successivement quelques sujets de la vie quotidienne. Mais si le langage dans ces corpus apparaît spontané et non scripté, la situation n'est cependant pas naturelle. Le corpus « CWIC » par Kennedy et Miceli (2010), seul corpus pédagogique écrit pour les niveaux inférieurs, est quant à lui catégorisé par sujet et comprend des textes édités, légèrement adaptés et/ou raccourcis. Les erreurs grammaticales, les fautes d'orthographe et les omissions involontaires ont également été supprimées avant de présenter les textes aux apprenants. Enfin, le corpus « FLEURON » est organisé par situation et contient des dialogues entre étudiants et personnel universitaire (André 2017, 2018, 2019, 2020a, 2020b).

Ces corpus remplissent ainsi parfaitement le critère de l'accessibilité et, partiellement, celui de la pertinence. Une de leurs principales limites vient cependant du fait que ces enregistrements sont majoritairement construits à partir d'entretiens (à savoir, des monologues avec des questions occasionnelles) et de textes narratifs alors que selon le CECRL, *les apprenants doivent être avant tout capables de participer à des interactions* du quotidien. Or, ces corpus (à l'exception du corpus « FLEURON ») ne donnent que peu d'occasions d'observer l'utilisation de la langue-cible dans le cadre de telles interactions.

L'authenticité situationnelle est un autre critère qui n'est que partiellement rempli dans ces corpus. Il est rare que les communications authentiques (orales ou écrites) soient exhaustives : elles ne traitent pas systématiquement de tous les aspects du sujet. Par exemple, les informations personnelles liées au sujet du CECRL intitulé « À propos de moi » sont généralement réparties dans des discussions réelles ; les locuteurs ne peuvent fournir toutes les informations à propos de leur vie, de leurs loisirs, de leur famille en une seule conversation. Ils contribuent aux dialogues en évoquant des expériences particulières, selon le contexte : ils évoquent de bons souvenirs lors d'une

fête d'anniversaire, par exemple. Leur discours n'est pas organisé par thème comme une interview, une présentation ou encore un examen de langue. De plus, les conversations spontanées sont plutôt imprévisibles car les interlocuteurs ont tendance à passer d'un sujet à l'autre au lieu de suivre un fil narratif bien défini (cf. De Fornel 1990 ; McCarten et McCarthy 2010 ; McCarthy 2002, 2003 ; Rühlemann 2007, 2018 ; Warren 2006). Les sujets et les éléments linguistiques associés émergent en pratique de manière progressive au long des échanges et les locuteurs construisent le discours ensemble, dans un processus de participation spontanée : ils réagissent à ce qui a été dit et expriment leur intérêt et leur attitude affective par des actions verbales et non verbales. Plutôt que de demander des informations spécifiques (comme cela se produit généralement lors d'un entretien ou d'une présentation), ils laissent leurs partenaires décider de ce qu'ils souhaitent partager (cf. Carter 2004 ; McCarthy 2002, 2003 ; Mondada 2002). Les mêmes observations s'appliquent aux contributions écrites : seul un nombre limité de genres écrits traite systématiquement d'un sujet donné. Un blog sur la nourriture ou sur la mode portera évidemment sur ces sujets-là, alors que d'autres types de textes (nouvelles, e-mails, SMS, entrées de forum, sites Web d'entreprise) n'exploreront aucun sujet en détail. Or, l'apprentissage des langues nécessite une systématisation même si celle-ci est artificielle : *une certaine médiation pédagogique semble ainsi être nécessaire pour que le contenu du corpus soit en accord avec les besoins et les compétences linguistiques des apprenants.*

2.3) L'authenticité et l'accessibilité langagières dans le corpus pédagogique

De nombreuses publications soutiennent que les grands corpus à fins linguistiques fournissent « des instantanés fiables du langage réel » (Salazar 2014 : 31, notre traduction). Ce constat, évident à première vue, mérite une réflexion plus approfondie dans le cadre pédagogique.

Il est incontestable que les énoncés authentiques illustrent l'usage langagier des locuteurs (cf. Hunston 2002 ; Widdowson 2003 ; McCarten 2010 ; Timmis 2015), mais dès que ces exemples réels sont enregistrés dans un corpus, les contextes situationnels et communicatifs plus larges qui permettent de les interpréter facilement, sont perdus (Sinclair 1997 : 34), rendant leur recontextualisation problématique, comme évoqué au chapitre 1 (cf. Braun 2007 ; Kaltenböck et Mehlmauer-Larcher 2005 ; Prodromou 1997 ; Widdowson 1978, 2003). Alors qu'elle s'applique à tous les corpus, cette caractéristique pose plus de difficultés dans le cadre de l'apprentissage des langues que pour les corpus à seules fins linguistiques. Ce fait a incité Widdowson (1978, 1998, 2003) à souligner l'importance de distinguer « langage sous forme originale ou langage non manipulé » (*genuine language*) et « langage authentique » (*authentic language*) pour le corpus pédagogique. « Langage non manipulé » se réfère à la nature des énoncés alors que « langage

authentique » désigne la relation entre énoncé et locuteur, indiquant ainsi la capacité de ce dernier à interpréter le message correctement. Cette distinction est capitale car le simple fait de présenter un langage non manipulé ne garantit pas que les apprenants soient capables de comprendre le sens des énoncés dans la mesure où certains éléments essentiels pour leur interprétation ne sont pas inclus dans le corpus. Comme l'observe Widdowson (1998 : 709) :

> « Un dialogue [authentique] peut se rapprocher de l'utilisation d'un langage réel, mais, pour cette raison même, il est tout à fait inutile pour l'apprentissage des langues. Non seulement ces énoncés sont structurellement incomplets, mais leur signification est mystérieuse […] car une grande partie du sens […] n'est pas du tout contenue dans la langue mais dans le contexte » (notre traduction).

Même si affirmer que les dialogues de la vie réelle sont « inutiles pour l'apprentissage des langues » peut faire l'objet de discussions, force est de constater que les énoncés réels se produisent localement, à un moment donné, ce qui les rend difficilement interprétables en dehors de ce contexte particulier. Les énoncés contiennent en effet de nombreuses références déictiques et des constructions elliptiques. Les locuteurs n'évoquent pas non plus les faits implicites, connus de tous les participants. Par ailleurs, toute réaction ne se manifeste pas verbalement lors de l'interaction, une partie du message n'est en l'occurrence exprimée que par des gestes ou par des mimiques : les locuteurs pointent du doigt une chose ou utilisent des expressions faciales quand cette solution leur semble plus simple, plus efficace et plus claire (McCarthy 2000, 2001, 2003 ; Rühlemann 2007, 2018 ; Thompson 2004). Cette composante non verbale de la communication a des conséquences pour l'interprétabilité *a posteriori* du message. Comme le constate Widdowson : « Les étrangers, qui ne possèdent pas ces informations, ne peuvent pas établir le lien contextuel nécessaire pour donner [aux énoncés] un sens approprié » (1998 : 709, notre traduction).

Ces particularités du langage réel soulèvent également quelques questions pour la construction du corpus pédagogique.

- Les conversations enregistrées et transcrites, doivent-elles être modifiées dans un souci de clarté et de compréhensibilité ?
- Doit-on compléter les parties manquantes (en ajoutant des explications) pour que toute information apparaisse dans l'enregistrement et dans sa transcription ?
- Les résultats de ce procédé, peuvent-ils être encore considérés comme du « langage authentique, non modifié » ?

Ajouter des phrases explicatives comme si elles faisaient partie du dialogue original est susceptible de transformer ces interactions en dialogues artificiels. Il serait sans doute peu naturel d'entendre dans un dialogue pendant le petit déjeuner des phrases telles que « Pourrais-tu me passer la marmelade ? Elle se trouve sur ta gauche à côté du verre à eau d'Annie. Annie est ta maman assise à côté de toi » ou de lire dans un e-mail « Ça te dirait de se voir aujourd'hui, vendredi 10 au Sam's, le pub irlandais sur la place Gutenberg où nous allons souvent prendre une bière après le travail ? » au lieu de juste dire : « RDV chez Sam ce soir ? »

Un enregistrement multimédia peut aider à conserver des éléments contextuels, extérieurs au discours. Néanmoins, certains aspects de l'interaction nécessiteront toujours quelques explications. Par exemple, il conviendra de clarifier les questions suivantes : Quelle est la relation entre les locuteurs ? Sont-ils des collègues, des amis, des membres de la même famille ou des étrangers ? S'agit-il d'une rencontre fortuite, d'un rendez-vous ou peut-être d'une rencontre régulière (pause de midi, repas de famille...) ? Où la situation se déroule-t-elle : dans la rue, à la maison, à l'université, au bureau ou dans un autre lieu ? Le créateur de corpus peut décider s'il souhaite rajouter de telles informations ou non, par exemple, sous forme d'une brève présentation de la situation précédant l'interaction filmée.

Dans le cas des transcriptions, un bref résumé peut précéder le dialogue[12] et quelques explications supplémentaires peuvent préciser le contenu des énoncés qui ne sont pas clairs. Par exemple, lorsque le locuteur dit « Peux-tu me le passer ? » et pointe du doigt le pot de confiture, son geste crée le contexte clarifiant le sens du message. L'information « pointe vers le pot de confiture » peut être incluse dans la transcription entre parenthèses afin de rendre la phrase recontextualisable. Ces modifications rendent les textes originaux plus accessibles car elles fournissent des éléments explicatifs à propos du contexte[13].

Dans le cas des apprenants aux niveaux inférieurs, une autre difficulté émerge fréquemment : non seulement le manque de contexte peut être source d'incompréhension, mais le langage lui-même peut poser problème. Les énoncés allant bien au-delà du niveau de l'apprenant parce que trop

[12] Puisque les dialogues aux niveaux inférieurs ont tendance à être courts, cette description est susceptible d'apparaître lors de la consultation du texte dans le Concordancier.
[13] Tout au long de ce projet de recherche, nous utiliserons la distinction de Sinclair (1991, 1997) entre « co-texte » et « contexte » selon laquelle « co-texte » se réfère à l'environnement textuel de l'élément linguistique étudié et « contexte » aux circonstances dans lesquelles le discours se déroule, i.e. tout ce qui est à l'extérieur des manifestations langagières (voir aussi la section 1 au chapitre 3).

complexes et/ou trop spécifiques, ne pourront pas faire partie du corpus pédagogique, même si les sujets qu'ils abordent sont pertinents. *Un critère auquel nous devons donc accorder une attention particulière dans le cadre de la construction de corpus pédagogiques est donc l'accessibilité langagière du contenu*. Que cette composante ne fasse pas partie des principes de construction de corpus à fins linguistiques[14] n'a rien d'étonnant : une sélection de textes sur la base de leur complexité linguistique irait contre les principes d'authenticité et de représentativité et impliquerait l'introduction d'un biais important. Alors qu'un tel procédé n'est pas acceptable dans le domaine de la recherche linguistique, il peut être utile (voire inévitable) pour les corpus conçus dans le cadre de l'apprentissage des langues. De telles collections doivent en effet garantir que la complexité linguistique, le contenu informationnel ainsi que le contexte situationnel des énoncés correspondent à la fois aux besoins des apprenants et à leur niveau de compétences. Cette contrainte oblige à une sélection très soigneuse du contenu aux niveaux A1, A2 et B1. *Ainsi, au-delà de la sélection, la modification de certains textes semble être une composante nécessaire pour en favoriser l'accessibilité* (alors même que ce procédé violerait le principe d'authenticité, tel qu'évoqué précédemment).

Notons enfin que lors de l'étude des interactions, *les apprenants ne participent pas à la construction du discours mais ils en sont des observateurs appliqués*. Nous pourrions les visualiser comme un public à l'écoute des autres. Dans cette position, ils n'ont pas besoin de se concentrer sur leur éventuelle contribution à la discussion et ils peuvent accorder toute leur attention à observer le comportement verbal et non verbal des participants. Cette activité constitue un élément tout aussi précieux de l'apprentissage car elle prépare les apprenants à s'engager par la suite dans des discussions similaires.

Comment est-il donc possible de concilier le critère d'authenticité avec celui d'accessibilité ? Tout bien considéré, un compromis semble être inévitable, nécessitant une redéfinition du contenu linguistique du corpus pédagogique pour les niveaux inférieurs. Pour refléter plus précisément la nature des données dans un tel corpus, *nous proposons donc d'introduire les dénominations « langage authentique modifié » ou « langage à caractère naturel » à la place du terme « langage authentique »*.

[14] Voir le chapitre 1 pour les critères de construction des corpus à fins linguistiques.

2.4) La taille et la représentativité du corpus pédagogique

La question de la taille s'avère par ailleurs des plus cruciales dans le cadre de la construction d'un corpus pédagogique : Comment déterminer le nombre de mots que de tels corpus doivent contenir ? Quels sont les facteurs à considérer pour choisir la taille des corpus pour les niveaux inférieurs ?

Une des principales considérations liées à la taille du corpus est sa finalité. Les corpus pédagogiques ont une double fonction : ils sont censés permettre aux apprenants d'observer des modèles linguistiques typiques et habituels, i.e. « des modes d'expression préférés des natifs » (*native speakers' preferred ways of saying things*) (Lewis 1997 : 12) ou « la façon usuelle de dire des choses » (*the normal ways of saying things*) (Langacker 2008 : 84) ainsi que l'utilisation des éléments choisis (mots, expressions, structures grammaticales). La taille d'un corpus pédagogique résulte d'un compromis. D'une part, pour que les apprenants puissent consulter les textes non seulement à l'aide des outils d'analyse, phrase par phrase, mais aussi dans leur intégralité (ceci dans le but de faciliter la recontextualisation des énoncés), ces corpus doivent avoir une taille limitée[15]. D'autre part, ils doivent être suffisamment volumineux pour que les apprenants puissent les questionner sur l'environnement linguistique typique d'un élément donné et obtenir des résultats statistiquement fiables (Brezina 2018 ; Egbert et al. 2020 ; Stefanowitsch 2020).

Dans leurs articles sur la conception de corpus écrits pour les niveaux intermédiaires, Kennedy et Miceli (2010, 2017) soulignent qu'il est plus avantageux de compiler une collection de textes soigneusement choisis que de construire une grande base de données sans présélection. Pour cette raison, elles préconisent la construction d'un petit échantillon de langage soumis à un contrôle de qualité rigoureux (2017 : 105). Leur corpus s'adresse aux apprenants des niveaux intermédiaires d'italien (niveaux B1 et B2) et comprend environ 500 000 mots. Les corpus « Sacodeyl », « ELISA » et « Backbone », destinés à donner aux étudiants des exemples d'usage oral des natifs pour les faire progresser dans l'apprentissage, sont compris également entre 200 000 et 500 000 mots (Chambers 2019 : 465). Braun qualifie son corpus « ELISA » d'un « petit corpus expérimental » (a small experimental corpus) qui se compose de 25 entretiens audio-visuels avec des locuteurs natifs d'anglais de différents pays et de différents parcours. Ces personnes parlent de leur parcours professionnel et des ressources culturelles et naturelles de leur pays. Tous les entretiens suivent un schéma similaire et couvrent une gamme identique de sujets, dont la plupart sont pertinents pour les contextes éducatifs (Braun 2007 : 310). En revanche, « FLEURON » est continuellement

[15] Voir aussi 2.5.

enrichi même si sa taille reste comparable à celle des autres corpus cités. Ce corpus comprend des situations de communication universitaires et se veut un « échantillon représentatif des interactions auxquelles les étudiants doivent ou peuvent participer lors de leurs séjours universitaires » (André 2017 : 302). En raison de leur faible taille, ces corpus ne contiennent qu'un nombre limité d'énoncés et, par conséquent, ne peuvent offrir que relativement peu d'exemples contenant les éléments langagiers que l'utilisateur souhaite étudier. Malgré cela, certains schémas d'utilisation langagière peuvent émerger car ces corpus suivent tous le principe de cohérence intertextuelle, tel que défini dans la section 2.2.

Face à ces compromis, quelle taille fixer pour le corpus pédagogique ? Selon Aston (1997) un corpus comprenant entre 20 000 et 200 000 mots serait déjà suffisant. Ces petits corpus conçus pour l'apprentissage des langues « peuvent être spécifiquement ciblés sur les connaissances et les préoccupations de l'apprenant » (Aston 2002 : 9, notre traduction). O'Sullivan et Chambers (2006 : 53), suivant la ligne de pensée d'Aston, soulignent ainsi que

> « [l]es corpus de taille limitée présentent des avantages par rapport à leurs homologues plus grands quand il s'agit de les utiliser avec des étudiants peu expérimentés, car ils sont plus faciles à gérer, plus faciles à apprivoiser, plus faciles à interpréter, plus faciles à construire et plus faciles à reconstruire. En outre, ils ont tendance à être plus clairement structurés et leurs limites sont également plus claires » (notre traduction).

Il est cependant tout à fait possible de construire des corpus de plus grande taille qui répondent aux besoins des apprenants dans la mesure où leur contenu est soigneusement sélectionné. Pour cette raison, *il nous semble plus approprié de distinguer les corpus conçus à fins pédagogiques et les corpus conçus avec un objectif de recherche linguistique, indépendamment de leur taille.* La différence fondamentale entre les deux types de collections est que la première est fondée sur une sélection déterminée par plusieurs paramètres liés à l'utilisateur, alors que la deuxième ne l'est pas. *Au lieu de poser la question de la taille en soi, il faut donc poser la question de la taille en combinaison avec le contenu de la collection*. D'autant plus qu'avoir à disposition un corpus pédagogique de taille plus importante offre plusieurs avantages. Élargir le corpus en y ajoutant des textes en accord avec les descripteurs du CECRL augmente le nombre de rencontres avec les éléments langagiers pertinents et fournit plus d'exemples qui illustrent leur fonctionnement. Suite à des expositions multiples aux éléments-clés, les apprenants « sont capables de généraliser à partir des expressions qu'ils rencontrent et de créer de nouvelles

expressions conformes à ces généralisations » (Taylor 2012 : 173, notre traduction). L'observation d'un nombre suffisant de modèles langagiers constitue donc un attribut déterminant afin de permettre aux apprenants de produire des énoncés fondés sur ces exemples et de rapprocher leur usage langagier de celui des natifs (Nesselhauf 2005 ; O'Sullivan et Chambers 2006 ; Schaeffer-Lacroix 2012).

Augmenter la quantité de textes en incluant davantage d'exemples correspondant à la même situation communicative et/ou sur une thématique donnée peut également contribuer à *une représentativité accrue* des corpus pédagogiques (Forti et Spina 2019). Idéalement, en effet, *le corpus pédagogique contient suffisamment de matériel linguistique pour fournir des informations statistiques sur la fréquence des mots*, *des collocations, des schémas grammaticaux et d'autres phénomènes*. Un corpus de faible taille ne peut pas contenir suffisamment d'exemples pour permettre aux apprenants de différencier l'utilisation langagière idiosyncratique (correspondant au cas où seulement une personne ou un nombre limité de personnes utilisent un certain élément) du langage de « tout le monde » (l'utilisation langagière typique de toute la communauté des locuteurs). Afin de pouvoir séparer ces deux catégories, il est ainsi nécessaire de s'appuyer sur des corpus pédagogiques plus larges.

Que les corpus pédagogiques existants soient cependant de taille relativement faible n'est guère surprenant. Leur compilation – la sélection et la modification des textes – est une activité extrêmement chronophage comme nous le verrons aux chapitres 12, 13 et 14. De plus, les textes pour les niveaux de compétences linguistiques inférieurs ne peuvent être ni très longs (le niveau A1 prévoit environ 100 mots par texte), ni trop complexes (les descripteurs des niveaux A1–B1 parlent du « langage simple ») (Conseil d'Europe, mis à jour en juillet 2021), ce qui signifie que de nombreux textes sont nécessaires pour compiler une collection d'aussi faible taille que de 20 000 mots.

2.5) Accès au contenu linguistique

Un consensus semble avoir émergé parmi les créateurs de corpus pédagogiques quant au fait que *l'accès à des textes individuels dans leur intégralité ainsi que l'accès à la collection via des outils d'analyse sont nécessaires pour que les apprenants bénéficient de manière optimale du corpus* (cf. Braun 2006, 2007 ; Kennedy et Miceli 2010, 2017 ; André 2017, 2020 ; Chambers 2019). Les approches didactiques présentées dans les paragraphes suivants intègrent ces deux composantes (lecture et analyse manuelle des textes ainsi que leur exploration avec des outils numériques) à différents degrés.

« L'Apprentissage basé sur les données » ou « l'Apprentissage sur corpus » (*Data-driven learning* ou *DDL*) est une méthode créée par Johns (1991), puis reprise et perfectionnée par plusieurs chercheurs[16]. D'après Johns, « l'apprenant en langue est aussi, pour l'essentiel, un chercheur dont l'apprentissage doit être guidé par l'accès à des données linguistiques – d'où le terme apprentissage sur corpus » (1991 : 2, notre traduction). Cette approche présuppose que l'apprenant consulte le corpus de la même façon que le linguiste et qu'il effectue un « travail de détective » pour trouver des réponses à ses questions (Bernardini 2000 ; Boulton 2010, 2017 ; Boulton et Tyne 2014 ; Johns 1991 ; Landure et Boulton 2010 ; Leńko-Szymańska 2017 ; Tyne 2012). Les questions posées par l'apprenant peuvent concerner différents aspects de la langue : la différence d'utilisation de plusieurs synonymes, les schémas d'utilisation d'un mot ou d'une expression et autres. Ainsi, l'Apprentissage sur corpus prévoit l'exploration des bases de données linguistiques (non pédagogiques) par des outils numériques.

Cette approche dans sa forme originale présente pourtant un inconvénient majeur : alors que l'exploration peut bien fonctionner avec les apprenants de niveaux avancés, elle est difficilement réalisable avec les étudiants de niveaux inférieurs car elle exige une capacité critique à classer et à interpréter les résultats avec assurance. Or, même aux étudiants de niveaux intermédiaires (B1, B2), l'aptitude à explorer le corpus de façon autonome semble faire défaut (Kennedy Miceli 2010 : 29–30)[17].

Dans une approche plus adaptée aux niveaux inférieurs, Braun (2005, 2006) recommande de mettre tous les textes du corpus dans leur intégralité à la disposition des apprenants et de commencer l'exploration du corpus par la lecture de ces textes. Cela permet aux apprenants de les explorer non seulement avec des outils numériques mais aussi en tant que textes cohérents, dans leur intégralité. L'argument principal des partisans de cette approche est que l'apprentissage des langues est principalement « concerné par le discours, c'est-à-dire par l'utilisation du langage dans des situations de communication concrètes » (Braun 2005 : 52). Contrairement à une consultation de corpus motivée par la recherche linguistique, les explorations langagières en classe commencent le plus souvent par l'étude de textes entiers car c'est en les lisant et/ou en les écoutant que les apprenants se familiarisent avec leur contenu (vocabulaire, style, registre, etc.). Les textes sont également des points de départ pour une pratique au cours de laquelle les éléments linguistiques pertinents sont consolidés. Cette phase aide ainsi les apprenants à comprendre et à interpréter les données explorées avec les outils d'analyse de corpus. Par la lecture de plusieurs textes, ils

[16] Pour une revue de publications sur le sujet, voir Boulton 2017.
[17] Comment réaliser des requêtes aux niveaux inférieurs sera le sujet de la Partie III de cette thèse.

formeront une idée du vocabulaire-clé (lié à un thème ou à une situation) et de son utilisation, ce qui les rendra capables d'effectuer ultérieurement des requêtes pertinentes à l'aide des outils de corpus et d'analyser les résultats de manière compétente et efficace. Certains chercheurs s'accordent également à dire qu'il est plus facile d'interpréter les résultats des requêtes lorsque les apprenants sont familiarisés avec les textes-sources (cf. Aston 2001 ; Gavioli 1997 ; Hunston 2009, 2010).

De la même façon, la troisième approche suggérée par Kennedy et Miceli (2010) prévoit la mise à disposition des textes entiers ; elle leur accorde néanmoins une autre fonction. Le contenu du corpus, disent-elles, sert avant tout *d'ensembles de modèles langagiers. L'apprenant analyse les textes manuellement et à l'aide des logiciels dans l'objectif d'enrichir ses propres récits avec des éléments langagiers observés chez les natifs et d'améliorer ainsi ses compétences linguistiques.* Quant au processus d'implémentation, Kennedy et Miceli expliquent qu'elles ont intégré le travail sur leur corpus dans un projet d'écriture créative. Leurs apprenants d'italien ont été invités à créer des entrées autobiographiques plus longues sur un nombre de thèmes. Cette tâche a fait réaliser aux apprenants l'intérêt du travail avec des corpus intégrant une multitude de textes sur le même sujet. En plus de leur permettre d'identifier des expressions utiles pour leurs propres textes, cette approche « leur a fait prendre conscience de l'importance de la phraséologie et les a encouragées à s'aventurer au-delà de leur répertoire linguistique existant » (p. 29-30, notre traduction).

Le tableau 3 résume les différences entre les critères de construction pertinents entre les corpus à fins linguistiques et pédagogiques.

Les principaux critères de construction des corpus à fins linguistiques	**Les principaux critères de construction des corpus à fins pédagogiques**
Textes sans interconnexion dans les grands corpus, textes avec cohérence intertextuelle dans certains corpus spécialisés	Cohérence intertextuelle
Aucun contrôle d'accessibilité langagière (dans l'intérêt de l'authenticité)	Contrôle d'accessibilité langagière
Authenticité : langage authentique, non modifié	Authenticité limitée : textes sélectionnés et modifiés, langage à caractère naturel
Taille (grande taille pour les corpus généraux, taille plus faible pour les corpus spécialisés)	Taille *en combinaison avec la* qualité de données
Représentativité et équilibre	Un nombre suffisant de textes sur le même sujet et/ou liés à la même situation communicative
Présentation des textes dans les outils d'analyse	Présentation des textes dans les outils d'analyse *et* dans leur intégralité

Tableau 3 : Critères de construction pertinents des corpus linguistiques et des corpus pédagogiques.

Comme le montre le tableau 3, la construction des corpus pédagogiques se distingue en de nombreux points de celle des corpus linguistiques. Or, construire des corpus pédagogiques offre plusieurs avantages : ils permettent aux apprenants d'étudier l'utilisation de la langue-cible dans des situations de communication sélectionnées, adaptées à leur niveau, et fournissent des exemples susceptibles d'émerger dans ces situations. Les textes peuvent également servir de modèles pour les produits linguistiques des apprenants.

Ce chapitre a exploré les considérations principales liées à la construction des corpus pédagogiques pour les niveaux de compétences linguistiques inférieurs. Nous avons discuté de l'intérêt de ces corpus ainsi que du terme « niveaux de compétences inférieurs » comme défini par le CECRL. Nous avons noté que les corpus pédagogiques pour les niveaux inférieurs ne peuvent (et ne doivent) pas remplir tous les critères définis pour les corpus à des fins linguistiques. En revanche, d'autres paramètres – non applicables aux corpus non pédagogiques – s'imposent, tels que la cohérence intertextuelle et l'accessibilité du langage. L'écart s'explique par les différentes fonctions de ces deux groupes de corpus : tandis que les corpus à fins linguistiques servent de bases de données pour la recherche, leurs développeurs doivent éviter, autant que possible, toute forme de biais. L'objectif des créateurs de corpus pédagogiques est de guider les apprenants pas à pas dans le processus d'apprentissage en leur proposant des textes qu'ils sont capables d'interpréter et d'en intégrer des éléments dans leur usage langagier. Ces corpus *doivent ainsi concilier, autant que possible, les trois principes suivants* :

- Leurs contenus doivent être *authentiques*, ils doivent donc contenir du langage « réel », produit dans des circonstances naturelles.
- Ces contenus doivent être *pertinents*, c'est-à-dire que les thèmes et les types de textes abordés dans le corpus doivent être utiles aux apprenants.
- Le langage dans ces corpus doit être *accessible* aux niveaux de compétences linguistiques inférieurs pour que les apprenants puissent les consulter sans difficulté.

Nous avons souligné tout au long de ce chapitre à quel point cette tâche était loin d'être simple et impliquait un certain nombre de compromis. Dans l'intérêt de l'accessibilité et de la pertinence, les textes initiaux doivent notamment être présélectionnés et soumis à un contrôle de qualité, ce qui impacte l'authenticité du contenu du corpus. Le contraire est également vrai : si la conservation

des textes sous leur forme originale respecterait le principe d'authenticité, elle nuirait en revanche à l'accessibilité linguistique et, par conséquent, à la pertinence des données.

Prendre des textes authentiques comme point de départ et les modifier quand l'accessibilité le nécessite, apparaît comme un compromis acceptable. Le résultat de ce mode de construction n'est plus alors un langage authentique mais « *un langage produit dans des situations authentiques, modifié* » ou « *un langage à caractère naturel* » (natural-sounding language) (Boulton et Cobb 2017 : 380). Dans les chapitres suivants, nous utiliserons le terme « langage à caractère naturel » pour qualifier la nature du matériel linguistique dans le corpus pédagogique.

Le chapitre 3 présentera les outils d'analyse de corpus ainsi que leur utilité pour l'exploration du corpus pédagogique.

Chapitre 3 : Termes, outils et mesures de la linguistique de corpus et leur utilité pour l'enseignement des langues

Ce chapitre est dédié à trois aspects pratiques de la linguistique de corpus : (1) la présentation des termes-clés et de leurs définitions ; (2) la présentation des outils d'analyse et (3) des mesures statistiques habituellement utilisées pour l'analyse des données.

La section 1 définira les termes suivants : mot-clé, co-texte et contexte, mot-clé en contexte, collocation et unité multi-lexicale, N-gram, colligation, schéma (pattern), type et token. La section 2 présentera les outils d'analyse qui peuvent être utiles dans le cadre pédagogique, notamment : (1) la fonction « Liste de mots » (Wordlist) pour l'identification des mots fréquemment utilisés dans le corpus choisi, (2) l'extracteur de collocation, (3) le Concordancier pour des exemples avec l'élément choisi et (4) l'extracteur de N-grams. Nous illustrerons ces fonctions à l'aide de Sketch Engine, le logiciel que nous utiliserons tout au long de ce projet de recherche pour l'exploration de la langue hongroise dont nous exposerons les résultats dans la Partie 2 de cette thèse. Les mesures statistiques de base comme la fréquence absolue et relative ou le rapport type/token ainsi que celles plus complexes comme l'Information mutuelle, le score T et le LogDice seront brièvement présentées dans la section 3 de ce chapitre.

A) Quelques termes-clés de la linguistique de corpus

Dans cette section, nous définirons les termes-clés de la linguistique de corpus que nous utiliserons tout au long de ce projet de recherche. Pour certains de ces termes, il existe une multitude de définitions (par exemple pour le mot « collocation »), d'autres sont des termes spécifiques de la linguistique de corpus avec une définition établie dans le domaine (« token », « N-gram »). Le but n'est pas ici d'engager la discussion autour des termes « épineux » ; nous souhaitons simplement fournir une définition pertinente pour notre projet de recherche, fondée sur les méthodes de l'analyse de corpus.

1) Mot-clé en contexte

Le terme « mot-clé en contexte » (*keyword in context* ou *KWIC*) fait référence au mot que l'utilisateur souhaite étudier à l'aide des outils d'analyse de corpus. Il peut consulter le Concordancier (voir B.3) pour trouver des exemples d'utilisation. Le tableau 4 présente quelques exemples à partir du nom « travail » en tant que mot-clé en contexte :

sable, est alors interprétable. Ce principe est l'objet, encore à l'heure actuelle, de	travaux	et de controverses. Certains linguistes le trouvent insuffisant pour rendre compte
uons brièvement ici la "logique propositionnelle" qui est directement issue de ses	travaux	. La logique propositionnelle est un langage formel dont les "formules bien formée
.e système de Montague est extrêmement séduisant et a fait l'objet de nombreux	travaux	, pour l'étendre et l'affiner tant d'un point de vue syntaxique que sémantique. C'es
·obien" et a privatif, "absence de") est une méthode de soins (mais également de	travail	en dehors des soins) qui consiste à accomplir une tâche donnée sans apporter d
édération hospitalière de France ou FHF (hôpitaux publics) a lancé un groupe de	travail	pour lutter contre les opérations chirurgicales "inutiles", face au déficit de la Sécu
igeants à un outil accessible au middle management. " Lire l'article "Le Stress au	travail	", interview de Fabrice Guez fondateur de la Société française de prévention et d
;s, co-auteur, avec Anne-Carole Delhommeau, de l'ouvrage "Agir sur le stress au	travail	", éditions Nathan / Les échos.fr " Lire l'interview Commentaire de l'auteur : Je sui
uctueux, notamment sur la prévention des risques psychosociaux et du stress au	travail	. Il est le créateur du logo d'iTELE entré en vigueur en septembre 2002. Etienne I
n coup de coeur chaque semaine. Ce n'est que la surface (la plus) visible de son	travail	. Stéphane est aussi le créateur de LA MUSICALE, la nouvelle émission de CAN/

Tableau 4 : Exemples de mot-clé en contexte pour le mot « travail ».

Le mode de présentation (le « mot-clé en contexte » apparaissant dans différents contextes) reflète lui-même un principe essentiel de l'approche de la linguistique de corpus. L'étude de l'élément choisi doit toujours inclure celle de son environnement textuel car les mots « ne se produisent jamais en isolation » (Manca 2012 : 5). C'est à l'aide d'une analyse tenant compte de l'environnement textuel du mot choisi que les schémas (*patterns*, voir A.8) peuvent émerger et les usages typiques deviennent visibles. L'accès au contexte d'usage d'un élément nous permet également d'obtenir des renseignements concernant son/ses sens[18].

2) Co-texte et contexte

Dans le cadre d'une approche empirique, il convient de distinguer l'environnement *linguistique* et l'environnement *situationnel* de l'élément étudié. Dans le terme « mot-clé en contexte » (voir 1.1), « contexte » fait référence à l'environnement linguistique du mot-clé pour lequel certains linguistes de corpus préfèrent utiliser le terme « co-texte » (Sinclair 2004a). Le co-texte s'étend de l'entourage immédiat du mot (quelques éléments à sa droite et à sa gauche) aux phrases qui l'incluent, précèdent et suivent, jusqu'au texte et à l'ensemble des textes dans lesquels il apparaît. Le terme « contexte » (ou « contexte situationnel » dans la terminologie de Halliday (1985) et de Firth (1957)) fait référence à tous les éléments qui entoure le texte et qui peuvent jouer un rôle dans son interprétation. Il s'agit donc de son environnement non-verbal et spatio-temporel, comme la relation entre les locuteurs, leurs gestes et mimiques, le lieu de l'interaction et les circonstances ainsi que les objets dans leur entourage (cf. Firth 1957). Ces éléments sont des parties constitutives du discours et des éléments essentiels pour son interprétation correcte, car « une situation particulière attire un éventail de mots et de phrases que nous sommes susceptibles d'utiliser » (Marca 2012 : 13). Cependant, ces éléments n'apparaissent pas dans le corpus écrit (à moins que son créateur ne rajoute des indications à cette fin), ce qui peut poser problème pour la recontextualisation des énoncés, comme évoqué dans le chapitre 2.

[18] Voir aussi le chapitre 4 sur les résultats pertinents de l'analyse de corpus pour l'enseignement des langues.

Dans ce travail, nous utiliserons les termes « co-texte » ou « environnement textuel » d'une part et « environnement non-textuel », « environnement situationnel » ou « contexte » de l'autre, de façon interchangeable.

3) Mot-clé

Il ne faut pas confondre le « mot-clé en contexte » avec le « mot-clé » tout court. Ce dernier signifie :

> « [D]es mots individuels (« tokens ») qui apparaissent plus fréquemment dans le corpus étudié que dans le corpus de référence. Tout mot individuel peut être un mot-clé s'il est utilisé plus fréquemment dans le corpus étudié que dans le corpus de référence. En réalité, le résultat comprendra principalement des noms et des adjectifs car les fréquences des autres parties du discours ont tendance à être similaires dans tous les textes » (Sketch Engine 2021, notre traduction).

La liste des mots-clés est donc le résultat d'une comparaison entre deux corpus. L'intérêt d'établir une telle liste est que les mots dominants dans le corpus étudié donnent des indications sur la nature de son contenu. En outre, une telle liste peut faire émerger les particularités lexicales de ce corpus en l'opposant au corpus de référence, constitué dans ce cas du corpus général le plus grand que l'utilisateur a à sa disposition (Geluso et Hirch 2019). Par exemple, dans le corpus « EurLex français » contenant des textes juridiques de l'Union européenne, les mots-clés sont : *article, règlement, paragraphe, commission, membre, produit, mesure, État.* En revanche, dans le corpus de *CHILDES* du langage des enfants, les mots-clés sont : *maman*, *papa, maison, bébé, monsieur, attention, voiture*. Ces mots-clés sont ainsi spécifiques au registre et/ou au thème auquel le corpus est dédié.

4) Collocation et unité multi-lexicale

Recenser les écrits autour du terme de « collocation » dépasserait largement l'objectif de cette étude. Le terme connait une multitude de définitions – chacune saisissant la relation entre les éléments constitutifs de ces cooccurrences privilégiées ou associations habituelles de façon différente. Les relations sémantiques et fonctionnelles entre ces derniers ont été étudiées, parmi d'autres, par Mel'čuk (2003, 2018), Halliday (1985) et Krishnamurty (2006), les facteurs psycholinguistiques par Wray (2002, 2008), par Langacker (1987, 1991b), par Goldberg (2005) et par Hilpert (2014), pour ne nommer que quelques auteurs éminents.

Dans le cadre du présent travail, nous nous intéressons avant tout à la définition du terme de « collocation » telle que formulée par les linguistes de corpus fondée sur des mesures statistiques. Kilgarriff (2006) note que chaque mot a une forte tendance à apparaître en compagnie de certains mots particuliers au lieu de former des associations de façon aléatoire, même si cette dernière possibilité serait envisageable du point de vue grammatical. L'omniprésence des collocations est largement reconnue en linguistique de corpus : la majorité des chercheurs de ce domaine partagent le point de vue selon lequel tous les éléments lexicaux sont aptes à former des collocations (cf. Sinclair 1991 ; Stubbs 1996).

La notion de « collocation » est généralement attribuée à Firth (1957) pour qui la collocation est une abstraction au niveau syntagmatique, plutôt qu'au niveau conceptuel (Fellbaum 2007 : 9). D'après Firth, le sens d'une collocation n'est pas considéré comme une disposition mentale des locuteurs qui existe indépendamment de l'usage. Comme le formule Firth lui-même : « Les collocations sont les mots concrets avec leur environnement lexical habituel » (Collocations are actual words in habitual company.) (1957 : 14)[19].

Selon Sinclair (1991), la collocation est *la relation de cooccurrence entre un mot ou une phrase servant de « base » et ses « collocatifs »*. Ces derniers peuvent correspondre à des formes individuelles des mots concrets (« je *prends* une décision ») ou des lemmes (« *prendre* une décision »), les exemples étant directement observables et dénombrables dans le corpus. Sinclair recommande l'inventorisation des formes individuelles plutôt que celle des lemmes car, comme nous le verrons par la suite, toutes les formes du même lemme ne s'associent pas, ou pas avec la même fréquence, à des collocatifs différents. Hoey (2005) a par ailleurs avancé l'idée que chaque mot est « amorcé » (*primed*) à être accompagnés par d'autres mots particuliers avec lesquels il forme des collocations[20]. Il précise cette définition en rajoutant la notion supplémentaire de la mesurabilité : la collocation, dit-il, est « la relation qu'un élément lexical entretient avec d'autres éléments *qui apparaissent avec une probabilité supérieure à la probabilité aléatoire dans son environnement textuel* » (Hoey 1991 : 5-6, notre traduction). Stubbs constate également que « l'attraction entre certaines unités linguistiques est bien plus forte qu'elle n'est généralement supposée, et cette attraction peut être mesurée de différentes manières » (Stubbs 2009 : 132, notre traduction ; v. aussi Xiao 2015). Selon ces définitions, les combinaisons

[19] Firth donne l'exemple du mot « nuit » (*night*) et affirme qu'un de ses sens dérive de son association fréquente avec le mot « sombre » (*dark*), et un des sens de « sombre » provient, par conséquent, sa collocation avec « nuit » (1957 : 196).
[20] Voir aussi le chapitre 4 pour la présentation succincte de la théorie de l' « Amorçage (Priming) lexical ».

suivantes de mots sont des collocations : *en hiver, prendre une décision, bon week-end*, *passer un bon weekend, faire preuve, faire mal*[21].

Une caractéristique des collocations est que leurs composants ne se doivent pas nécessairement se suivre dans le texte pour former une unité de sens. Une autre particularité des collocations, telles que définies par la linguistique de corpus[22], est qu'elles peuvent être opaques (leur sens n'étant pas la somme du sens de leurs constituants, comme dans les expressions « chambre froide » ou « chambre double ») ou parfaitement claires (le sens des éléments s'additionne comme dans « chambre confortable » ou « chambre parentale »). Des combinaisons fréquentes de mots qui ne forment pas ensemble d'unité de sens, ne sont pas considérées comme des collocations. Par exemple, même si les mots « et » et « je » se suivent très souvent dans des textes, nous ne les qualifions pas de collocations puisque leur ensemble ne forme pas de sens interprétable[23] et il n'y a pas de relation logique entre les deux mots. Nous nous référons à ce type de séquence de mots consécutifs comme des « N-grams » (voir ci-après la section A.5).

Certains linguistes de corpus préfèrent utiliser le terme « unités multi-lexicales » au lieu de « collocation ». Les unités multi-lexicales sont alors décrites comme « des séquences de mots – opaques ou non, du point de vue sémantique – susceptibles d'être lexicalisées en tant qu'unité et de forme plus ou moins stable » (voir p. ex. Nation 2013 : 479, notre traduction).

La littérature distingue quatre grands types d'unités multi-lexicales : (1) une unité multi-lexicale peut être un groupe de mots qui se produisent généralement ensemble, comme « saisir l'opportunité » ; (2) ou un ensemble de mots dont le sens n'est pas la somme du sens des parties, comme « en gros » ou « rouler quelqu'un dans la farine » (duper quelqu'un) ; (3) le terme peut également faire référence à toutes les combinaisons d'un mot particulier et des mots qui l'accompagnent fréquemment (« tu sais », « joue un rôle important/significatif/décisif ») ; et (4) il peut désigner des groupes de mots qui sont intuitivement considérés comme des formules, c'est-à-dire des éléments stockés sous forme de choix uniques, par exemple « je ne sais pas », « que puis-je pour vous ? » (Nation 2013 : 479). Ces catégories sont fondées sur une variété de critères

[21] Exemples tirés du corpus de « frTenTen17 » par Sketch Engine.
[22] Il convient de noter qu'il existe quelques différences entre les définitions de chercheurs individuels dans le domaine de la linguistique de corpus. Cette thèse utilise celle la plus largement acceptée.
[23] Certains linguistes de corpus ne font pas de distinction entre collocations, N-grams et unités multi-lexicales. Pour eux, si l'argument statistique (une association mesurablement forte) est rempli, les partenariats de mots – formant une unité de sens ou non – sont qualifiés de collocations (cf. Brezina 2018 ; McEnery et Hardie 2012).

de nature différente tels que la fréquence de cooccurrence, la compositionnalité, la forme et le stockage ; il n'est donc pas vraiment surprenant qu'il existe une liste longue et croissante de termes pour désigner les unités multi-lexicales (pour une discussion plus approfondie, voir Wray 2000)[24]. Dans ce projet de recherche, le terme d'« unités multi-lexicales » fait référence à tous les types de séquences de mots mentionnés précédemment.

Une découverte significative de la linguistique de corpus est que chaque registre langagier possède ses unités multi-lexicales « préférées », apparaissant avec une fréquence supérieure à celle d'autres registres. Il est donc possible d'identifier certaines particularités des registres en répertoriant leurs unités multi-lexicales typiques (Biber et Gray 2010 ; Biber et Egbert 2018 ; Ellis et Ogdan 2017 ; Grieves et Warren 2010). Au-delà d'une description linguistique des registres, ces unités multi-lexicales ont également leur importance dans l'apprentissage des langues. Par exemple, la présentation et la pratique des unités multi-lexicales caractérisant les interactions informelles peuvent augmenter l'efficacité de l'enseignement aux niveaux inférieurs et rendre l'usage langagier des apprenants plus naturel, plus proche des natifs (cf. Cowie 1992 ; Ellis 2002 ; Schmitt 2004).

Dans cette thèse, nous utiliserons de préférence le terme « unité multi-lexicale » et nous ne nous référerons au terme « collocation » qu'occasionnellement, avant tout dans les citations.

5) N-grams

Le « N-gram » est un terme générique pour *décrire des séquences continues de N mots* qui apparaissent avec une fréquence notable dans le corpus. Il est important de souligner que les composantes d'un N-gram doivent obligatoirement former une séquence ininterrompue ; ils comprennent donc tout assemblage de mots consécutifs qui, sans être fixe, n'est pour autant pas fortuit. Les unités multi-lexicales dont les éléments se suivent dans les énoncés observés, sont, de ce fait, elles aussi des N-grams. En revanche, celles dont les éléments ne se suivent pas (par exemple, en présence d'un ou plusieurs éléments intercalés ou lorsque l'ordre des composantes est inversé) ne sont pas qualifiés de N-grams.

Les N-grams sont très utiles en ce qu'ils mettent en évidence les associations fréquentes de mots consécutifs, formant ou non, ensemble une unité de sens (Sketch Engine 2021). Une question liée à leur analyse est la longueur optimale de séquences qui méritent d'être étudiées. Alors que pour

[24] Une partie de ces unités multi-lexicales sont également des collocations étendues (« faire face à », « c'est-à-dire », « passe un bon weekend », etc.).

les langues isolantes, l'étude des 2-grams (ou bi-grams) est censée avoir une utilité limitée – dans la mesure où la majorité des 2-grams particulièrement fréquents est formée des combinaisons de préposition et d'un article ou d'autres mots fonctionnels (cf. Biber 2009) –, l'analyse des 2-grams est tout à fait justifiée dans le cas des langues morphologiquement complexes (voir, par exemple, Jantunen et Brunni (2013) pour le finnois ou Katinskaia et Sharoff (2015) pour le russe).

Des exemples de N-grams à trois composantes (3-grams) avec le mot « faire » dans le corpus de French Web 2017 par Sketch Engine sont : « de faire un », « faire une nouvelle », « faire un nouveau », « faire face à », « de le faire ». Certaines de ces séquences forment des unités multi-lexicales avec un sens bien identifiable (« faire face à »). D'autres séquences contiennent des mots sans relation logique entre eux ; à cette catégorie appartiennent « de faire un » et « faire une nouvelle ». Bien que ces associations ne soient pas en tant que telles porteuses de sens, elles indiquent certaines caractéristiques de l'usage langagier, par exemple l'ordre typique de leurs éléments. Les séquences « faire une nouvelle » et « faire un nouveau » indiquent que l'adjectif précède le nom – phénomène plutôt rare en français en général mais qui semble être typique dans ce cas particulier, avec l'adjectif « nouvelle/nouveau ». Ainsi, ces séquences peuvent faire émerger des usages langagiers plus ou moins surprenants (ou inattendus avec la fréquence révélée par l'analyse du corpus) méritant une étude approfondie. Ces renseignements sont, par ailleurs, particulièrement pertinents pour la langue hongroise dans laquelle les règles relatives à l'ordre des mots sont complexes.

6) Colligation

Les termes présentés dans les paragraphes précédents concernent les relations lexicales entre les composantes d'un énoncé. Cependant, il est également possible d'identifier des relations grammaticales (colligations) entre ces éléments. Sinclair (2003 : 145) définit le terme « colligation » de la façon suivante :

> « La colligation est similaire à la collocation dans la mesure où elle concerne toutes deux la cooccurrence de caractéristiques linguistiques dans un texte. La colligation est l'occurrence d'une classe grammaticale ou d'un modèle structurel avec un autre, ou avec un mot ou une phrase. « Négatif », « possessif » et « modal » sont les types de catégories essentiellement grammaticales qui figurent dans la colligation. Le terme a été utilisé pour la première fois par J. R. Firth et a été un peu élargi pour le travail de corpus » (notre traduction).

Cette définition a été enrichie par Hoey (2005) pour qui la colligation correspond non seulement aux associations grammaticales qu'un mot ou une séquence de mots privilégie ou évite, mais représente aussi sa position préférée dans une séquence (phrase, paragraphe, texte) (p. 43-44). Nous utiliserons le terme de « colligation » dans ce sens et inclurons systématiquement la position textuelle des éléments étudiés dans notre analyse[25]. Cet aspect est d'une grande importance pour le hongrois qui connait des règles complexes concernant l'ordre des mots, comme évoqué précédemment.

7) Schéma[26] (*pattern*)

La linguistique de corpus est concernée par plusieurs types d'éléments fréquemment co-sélectionnés dans les énoncés de différents locuteurs, dans différents textes. Ces structures répétées sont observables à quatre niveaux : nous distinguons (1) des schémas lexicaux, (2) des schémas grammaticaux, (3) des schémas sémantiques et (4) des schémas pragmatiques (cf. Sinclair 1991, Stubbs 2009, Hoey 2005). Les quatre catégories ont été établies par Sinclair qui conserve les définitions de Firth (1957) concernant les schémas lexicaux et grammaticaux et les complète avec les deux dernières catégories. Bien que Firth et Sinclair se réfèrent aux schémas lexicaux en tant que « collocations », nous préférons conserver dans cette thèse le terme « schémas lexicaux » pour éviter toute confusion avec la notion de « collocation » telle que définie précédemment. Les collocations, les N-grams ou, au sein des N-grams, les unités multi-lexicales représentent des exemples de « schémas lexicaux ». Les relations grammaticales typiques seront décrites par les termes « schémas grammaticaux » et « colligations », de façon interchangeable.

La troisième catégorie, i.e. la notion des schémas sémantiques est associée au nom de Sinclair (1991) qui la dénomme « préférence sémantique » et la définit comme la relation de cooccurrence entre les mots associés et leurs champs lexicaux caractéristiques. Hoey (2005) utilise le terme d'« associations sémantiques », O'Keeffe et al. (2007) proposent le terme « composantes sémantiques » dans un sens comparable. Plusieurs chercheurs soulignent qu'afin d'arriver à une définition plus précise de son sens et de rendre ce schéma visible, il est nécessaire de cartographier l'environnement textuel typique du mot choisi (cf. Biber et Reppen 2006 ; Sánchez-Cárdenas 2010 ; Stubbs 2009 ; Xiao et McEnery 2006). Dans cette thèse, nous utiliserons le terme « composantes sémantiques » pour désigner ce type de schéma.

[25] Voir la Partie II de cette thèse.
[26] Certains chercheurs français comme Legallois (2006) conservent le terme « pattern » dans leurs publications, avec le même sens.

Le quatrième type de schéma est la « prosodie sémantique », pour utiliser le terme de Sinclair (1991, 2003). Cette catégorie indique l'objectif communicatif général de l'unité observée, c'est-à-dire sa force illocutoire ou encore la raison motivant le locuteur dans son choix de mode d'expression. Le terme original de Sinclair est aujourd'hui souvent remplacé par les termes « associations pragmatiques » (Hoey 2005) ou « composantes pragmatiques » (O'Keeffe et al. 2007) puisqu'elle désigne les fonctions pragmatiques particulières de l'énoncé (Hoey 2005 : 26). Par exemple, le verbe « provoquer » signifie selon Larousse « être l'instigateur de quelque chose, l'amener, être la cause de quelque chose, l'entraîner » ; c'est donc un verbe d'action sémantiquement neutre. Or, une analyse de ses occurrences dans le corpus révèle que ce mot est loin d'être neutre comme la définition pourrait le suggérer. Il s'associe de préférence à des mots décrivant un événement ou un état plutôt négatif comme « accident », « colère », « crise » et « mort ». Cette étape est aussi cruciale pour l'enseignement des langues pour que l'apprenant puisse se forger une idée du fonctionnement de l'élément dans le discours. Nous utiliserons le terme « composantes pragmatiques » pour désigner ce type de schéma.

Ces quatre types de schémas offrent un cadre pour visualiser et pour interpréter les résultats de l'analyse de corpus. Puisque la systématisation des connaissances est une partie intégrale de l'apprentissage, ces schémas permettent de cartographier les caractéristiques d'un élément choisi de façon uniforme. Inclure dans l'enseignement des langues des tâches qui incitent les apprenants à découvrir et à pratiquer des schémas apparaît donc être des plus utiles[27].

8) Token

Pour décrire le contenu et la complexité d'un corpus, les termes de « token » et de « type » sont utilisés. Les « tokens » sont les éléments individuels et la taille du corpus est toujours définie par le nombre de ses « tokens ». Certains corpus – comme celui présenté dans le tableau 5 – différencient « tokens » et « mots », dans quel cas les « tokens » incluent également les chiffres ou des symboles comme © ou ↔ (v. tableau 5).

Tokens	6,845,630,573
Words	5,752,261,039
Sentences	272,578,993
Paragraphs	123,281,964
Documents	14,088,683

Tableau 5 : Le contenu du corpus « frTenTen17 ».

[27] Pour des exemples concrets, voir les chapitres 12 et 14.

Le terme « token » désigne le nombre total de mots dans un texte, un corpus, etc., indépendamment de leur fréquence de répétition. Il est important de noter que le nombre de « tokens » indique seulement la taille du corpus mais ne donne aucun renseignement sur sa complexité linguistique car un grand corpus n'est pas nécessairement un corpus varié. Le terme « type » fait en revanche référence au nombre de mots *distincts* dans un texte, un corpus, etc. Pour mesurer la complexité (ou la diversité lexicale) d'un corpus, le rapport type/token est utilisé (voir section C1.2 ci-dessous).

B) Les outils d'analyse de corpus

Les corpus sont généralement analysés avec des logiciels spéciaux (cf. Timmis 2015 ; McEnery et Wilson 1997 ; Stefanowitsch 2020) et la majorité des corpus dispose de boîtes à outils intégrées qui permettent d'explorer leur contenu de nombreuses façons. Il existe également des logiciels d'analyse de corpus avec lesquels l'utilisateur peut étudier un corpus compilé par lui-même, comme « AntConc », « WordSmith », « LancsBox » ou « Sketch Engine » ou « TXM ». La plupart de ces outils sont librement accessibles (AntConc, WordSmith et LancsBox, TXM), Sketch Engine est un logiciel dont certaines fonctions sont payantes[28].

Barlow (2004) résume les principales caractéristiques des outils de corpus en déclarant que leur fonction est de transformer les textes en les divisant en lignes de concordance et en listes de mots et de phrases. Ces fragments textuels sont réorganisés par ordre alphabétique ou par fréquence. *Ces « effets d'aliénation » forcent la distance nécessaire entre la manière dont la langue est présentée dans un texte normal et dans le corpus*. Selon Sinclair (1999, 2004), l'intérêt de cette technique d'observation est qu'elle peut fournir des renseignements sur une multitude d'aspects du texte et mettre en évidence de nombreux schémas d'usage qui échapperaient aux observateurs sans ce changement de focus dû à la façon dont l'information est présentée.

Les sections suivantes présenteront des outils standards d'analyse de données linguistiques. Leurs fonctions seront illustrées en utilisant « Sketch Engine » (Kilgarriff 2014), site Web d'hébergement de corpus muni d'un large spectre d'outils numériques d'analyse[29]. Notre choix pour Sketch Engine a été motivé par le fait que le site Web héberge de très grandes bases de données pour plus de 90 langues, dont le hongrois. Son interface est simple, les outils sont faciles à utiliser, avec la possibilité

[28] Pour une présentation détaillée des différents logiciels utilisables dans le cadre pédagogique, voir Pérez-Paredes (2021).
[29] Ces fonctions sont intégrées dans la majorité des logiciels, bien que leur interface puisse être différente.

de consulter plusieurs mesures statistiques intégrées. Les outils d'annotation fonctionnent bien pour le hongrois (tâche difficile pour les langues moins répandues, ayant une morphologie complexe comme le hongrois) assurant l'obtention de résultats fiables. Un autre avantage qu'offre Sketch Engine est la possibilité de créer son propre corpus et de le rendre accessible au public ; les corpus de hongrois présentés dans la Partie III de cette thèse sont également hébergés sur cette plateforme.

1) Wordlist (Liste de mots)

L'outil « Wordlist » présente les mots[30] ordonnés par leur fréquence d'occurrence dans le corpus choisi. La recherche peut être étendue à tous les mots ou limitée à une partie du discours (part-of-speech analysis ou POS analysis).

Comment ce programme fonctionne-t-il ? Tout d'abord, il parcourt l'ensemble des textes et réduit toutes les instances répétées en types ; c'est-à-dire que chaque instance (« token ») du mot est comptée mais la liste complète l'affiche une seule fois comme « type », suivi de l'indication de sa fréquence (Scott et Tribble 2006 : 12-13). Pour donner un exemple, le tableau 6 met en évidence les verbes les plus fréquents dans le Web français 2017, un très grand corpus général comprenant 6 845 630 573 tokens collectés sur des sites francophones publiés avant 2017.

1	être	127,742,200	14	savoir	4,660,480
2	avoir	77,036,301	15	consulter	4,458,970
3	pouvoir	23,451,230	16	suivre	4,221,307
4	faire	22,641,673	17	passer	4,160,629
5	devoir	7,690,740	18	venir	4,057,620
6	voir	7,089,696	19	vouloir	4,033,614

[30] Nous ne souhaitons pas aborder la discussion autour de la définition du terme « mot ». Pour ce travail, nous acceptons la définition du dictionnaire Larousse selon laquelle le mot est un « élément de la langue composé d'un ou de plusieurs phonèmes, susceptible d'une transcription écrite individualisée et participant au fonctionnement syntactico-sémantique d'un énoncé ».

7	prendre	7,036,318	20	utiliser	3,895,720
8	aller	7,022,419	21	proposer	3,532,077
9	dire	6,912,023	22	faillir	3,245,286
10	mettre	6,117,338	23	devenir	3,123,364
11	permettre	5,869,790	24	connaître	3,087,857
12	donner	5,448,613	25	rester	2,907,555
13	trouver	4,798,472	26	partir	2,906,279
27	comprendre	2,852,719	40	tenir	2,160,203
28	aider	2,627,958	41	porter	2,094,187
29	modifier	2,608,159	42	vivre	2,069,915
30	présenter	2,603,180	43	découvrir	2,030,571
31	créer	2,555,289	44	partager	1,974,712
32	rendre	2,526,978	45	situer	1,966,222
33	penser	2,456,080	46	offrir	1,956,197
34	parler	2,386,446	47	arriver	1,953,276
35	lire	2,363,212	48	retrouver	1,882,820
36	demander	2,338,094	49	aimer	1,871,115
37	réaliser	2,312,141	50	jouer	1,866,414
38	agir	2,245,597			
39	laisser	2,176,356			

Tableau 6 : Les 50 verbes les plus fréquents dans le corpus « frTenTen17 ».

Le corpus contient 127 742 200 occurrences avec toutes les formes de « être », verbe figurant à la tête de la liste, suivi d'« avoir », de « pouvoir » et de « faire ». Une première analyse superficielle indique que ces quatre verbes apparaissent avec une fréquence beaucoup plus importante que les autres verbes sur la liste et que le premier mot est deux fois plus présent dans le corpus que le deuxième. Les listes de mots-clés de n'importe quel corpus afficheraient la même tendance : nous y trouverions un nombre relativement faible de mots en tête de la liste apparaissant avec une fréquence très significative, suivis d'un grand nombre d'autres mots à faible fréquence et des « hapax legomena », c'est-à-dire des mots qui n'émergent qu'une seule fois dans le corpus entier

(Scott et Tribble 2006). Zipf a formalisé cette observation par une formule mathématique qui porte aujourd'hui son nom. La loi de Zipf exprime que le deuxième mot le plus utilisé dans le corpus s'affiche deux fois moins souvent que le premier, le troisième mot le plus utilisé trois fois moins souvent que le premier et ainsi de suite. (Zipf 1965 : 24 ; Oakes 1998 : 54–55 ; Pustet 2004 : 8). Scott et Tribble (2006 : 29) complètent la loi de Zipf avec la notion de la « cohérence » en proposant que certains mots ont tendance à émerger dans tous les types de textes avec une fréquence plus ou moins constante alors que d'autres sont plutôt limités à certains types de textes et/ou liés à des thèmes particuliers. Ces éléments lexicaux sont les mots-clés (comme définis dans la section A.3) relatifs au texte ou aux collections de textes choisis.

Les listes de mots organisées par fréquence peuvent être particulièrement utiles pour l'enseignement des langues : elles peuvent aider les concepteurs du curriculum à décider quels éléments lexicaux ils doivent inclure à des niveaux différents de compétences. Elles peuvent également être utiles aux auteurs de manuels et aux enseignants pour identifier les unités lexicales récurrentes dans des types différents de textes (lettres de motivation, entrées de blog, critiques de restaurants, etc.). Ces informations spécifiques peuvent être incluses dans le matériel pédagogique, augmentant ainsi son utilité.

2) Extracteur d'unités multi-lexicales : Word Sketch

Cet outil permet de résumer les informations concernant l'environnement typique des mots simples et des unités multi-lexicales. Le tableau ci-dessous (tableau 7) montre le profil du verbe « être » (127 742 200 occurrences dans le cirpus) avec ses modificateurs typiques, ses sujets pronominaux, ses compléments et ses objets, etc. Les colonnes sont organisées par fréquence d'occurrence. Ces combinaisons de mots peuvent être des N-grams incomplets (« c'est d'oublier », « est le premier ») ou des unités multi-lexicales (« vous êtes la bienvenue », « c'est le cas ») (voir les sections A.4 et A.5 pour les définitions de « N-gram » et « unité multi-lexicale »). En cliquant dans l'application sur les trois points à côté des mots, le co-texte apparaît sous la forme de lignes de concordance.

objects of "être"	subjects of "être"	modifiers of "être"	infinitive objects of "être"
bienvenue Votre aide est la bienvenue ! Comment	aide incomplète . Votre aide est la bienvenue	plus est plus	dire C' est dire
cas est le cas	section faire ? Cette section est vide , insuffisamment	très est très	hyper est hyper
ébauche Cet article est une ébauche concernant	article aide . Cet article est une ébauche concernant	bien est bien	conseiller est conseiller
question est question	objectif L' objectif est	aussi est aussi	faire c' est faire
chose est quelque chose	parce parce que c' est	donc est donc	donner c' est donner
parce c' est parce	but but est de	également est également	oublier C' est oublier
fois est la première fois	problème problème est	toujours est toujours	prendre c' est prendre
temps est temps	résultat résultat est	déjà est déjà	aller c' est aller
homme est un homme	question question est	encore est encore	mettre c' est mettre
membre est membre	chose chose est	peu est un peu	accepter c' est accepter
occasion est l' occasion	homme homme est	ainsi C' est ainsi	voir ai été voir
premier est le premier	idée L' idée est	là est là	créer c' est créer

Tableau 7 : Les collocatifs les plus fréquents du verbe « être » dans le corpus « frTenTen17 ».

Alors que dans le cas du verbe « être », nous trouvons un nombre significatif des partenariats de mots fréquents qui ne forment pas d'unités multi-lexicales interprétables, d'autres recherches peuvent donner une liste d'unités multi-lexicales compréhensibles à première vue, comme dans le cas du nom « chambre » (1 157 340 occurrences dans le corpus) (tableau 8) :

verbs with "chambre" as object	verbs with "chambre" as subject	modifiers of "chambre"	prepositions preceeding noun/nouns after preposition
climatiser chambres climatisées	dhôtes chambres dhôtes	spacieux chambres spacieuses	devant devant la chambre
décorer chambres décorées	disposer chambres disposent d'	double chambre double	dans dans la chambre
meubler chambre meublée	comprendre Les chambres comprennent	confortable chambres confortables	avec avec la Chambre
louer louer une chambre	posséder chambres possèdent	froid chambre froide	vers vers la chambre
réserver réserver une chambre	offrir chambres offrent une	criminel chambre criminelle de la Cour	de de la chambre
équiper chambres équipées	donner chambres donnent	consulaire chambres consulaires	par par la Chambre
aménager chambres aménagées	voter Chambre vote	funéraire la chambre funéraire	pour pour la chambre
séparer chambre séparée	bénéficier chambres bénéficient d'	civil chambre civile	jusqu'à jusqu'à la chambre
insonoriser des chambres insonorisées	ouvrer chambres s' ouvrent	individuel chambre individuelle	depuis depuis la chambre
rénover chambres rénovées	doubler chambre double	correctionnel chambre correctionnelle	en en chambre
ranger ranger sa chambre	allier chambres allient	parental chambre parentale avec	à à la Chambre des
doter des chambres dotées d' une	siéger chambres siègent	régional chambre régionale des comptes	entre entre les chambres

Tableau 8 : Les collocatifs les plus fréquents du nom « chambre » dans le corpus « frTenTen17 ».

Les listes générées par Word Sketch incluent deux types de partenariats de mots : des associations fréquentes et des associations fortes. Le premier groupe est formé par des unités multi-lexicales composées de deux éléments eux-mêmes fréquents, par exemple, « chambre + confortable » ou « ma + chambre ». Le deuxième groupe comprend des expressions telles que « chambre climatisée » et « chambre d'hôte » dans lesquelles « climatisé » et « d'hôte » s'associent plus fréquemment à « chambre » qu'à d'autres noms. Ces associations sont fortes, ce qui est exprimé, statistiquement parlant, par un haut score d'Information mutuelle[31].

Ces deux types d'associations sont tous deux intéressants pour l'enseignement des langues car – même si le sens de ces expressions peut être déduit de leurs composantes – ce sont des unités multi-lexicales extrêmement utiles à retenir aux niveaux inférieurs, en raison de leur fréquence. Word Sketch peut donc être un outil pour sélectionner les unités multi-lexicales qui doivent être présentées dans le matériel pédagogique. Il peut être également utilisé dans le cours de langues

[31] Voir section C.2.2 pour la description de cette mesure statistique.

quand il s'agit d'obtenir un aperçu général de l'environnement linguistique typique de l'élément choisi.

3) Le Concordancier (Concordancer)

L'outil le plus important de l'analyse de corpus est le Concordancier. Les lignes de concordance permettent de présenter les énoncés de manière non-linéaire, en offrant des exemples avec le mot choisi, extraits de plusieurs textes. Le concordancier *rassemble donc des énoncés* « *qui ont été produits à des moments différents par différents locuteurs ; il rend visibles des schémas récurrents et nous permet de les compter* » (Mauranen 2004 : 103, notre traduction, nous soulignons). L'analyse d'un grand nombre de textes offre des avantages clairs par rapport au travail avec un seul texte, car « dans un texte individuel, on ne peut observer ni de relations syntagmatiques répétées ni de relations paradigmatiques. Ce sont précisément ces deux relations que les lignes de concordance rendent visibles » (Tognini-Bonelli 2004 : 18, notre traduction).

Le Concordancier se présente sous forme d'un tableau dont les axes verticaux et horizontaux fournissent des informations complémentaires. L'axe vertical met en évidence les formes les plus usitées dans de nombreux exemples alors que l'axe horizontal fournit le sens, à la fois dans les lignes de concordance individuelles et dans l'ensemble des lignes. Néanmoins, et même si ces outils présentent la langue sous un format nouveau qui permet de révéler certaines de ses caractéristiques, ce sera le linguiste qui, en utilisant son expérience et son intuition, regroupera en fin de compte les instances observées en catégories sémantiques (Stubbs 2009). Le travail d'analyse *ex post* du linguiste reste donc indispensable.

Afin d'illustrer le fonctionnement du Concordancier, considérons à titre d'exemple l'adverbe « bien », collocatif fréquent du verbe « être ». Le corpus contient 1 141 796 occurrences dans lesquelles « bien » suit directement ce verbe. Les tableaux suivants présentent vingt de ces exemples de deux manières différentes. Dans le premier tableau, les occurrences sont organisées par phrase, garantissant une bonne lisibilité horizontale. Dans le second, le mot-clé est placé au centre, permettant au lecteur d'observer ses voisins de gauche et de droite et d'identifier les schémas d'utilisation (tableaux 9a et 9b).

proteos.info	Il **est bien** difficile de faire plus idéologique.
lefigaro.fr	Je **serais bien** curieuse de connaitre vos sources.
inria.fr	La chaîne de processus **est bien** illustrée en regardant la régénération du foie après une injection toxique et destructrice.
recette-dessert...	La crème **est bien** plus épaisse et onctueuse.
revue-signes.in...	Mais cet enthousiasme et cette euphorie ont **été bien** éphémères.
pratique.fr	Il **est bien** souvent difficile de faire son choix parmi les très nombreuses offres de tablettes tactiles.
tomshardware.fr	C' **est bien** un modem que tu utilises aussi ?
senat.fr	L'un des défis **est bien** de fédérer les gestionnaires du niveau national au niveau local.
terre-mere.fr	Elle avait de grande falaise blanches et la végétation **était bien** verte.
msh-paris.fr	Mais ces propriétés **seront bien** plus mises en avant dans des développements ultérieurs.
x-grosses.com	La chienne est à la limite de lui procurer un bon facesitting tellement ses fesses **sont bien** rondes.
libraires-ensem...	Ce roman **est bien** plus que le récit initiatique d'un jeune métis découvrant la guerre et la discrimination.
cchst.ca	Il faut toujours s'assurer que le contenant **est bien** assujetti à la palette ou à un dispositif similaire.
corpusetampois....	La neige **est bien** de la neige, et sa valeur sur le ciel est scrupuleusement exacte.
onisep.fr	Les premières semaines sont dures car la charge de travail **est bien** plus importante qu'au lycée.
free.fr	Mais parler c' **est bien** autre chose.
recreanice.fr	Et vous êtes attentifs à ce qu'ils **soient bien** traités?
ille-et-vilaine...	Globalement le rétablissement de la route départementale 106 par un viaduc, **est bien** accepté.
culinotests.fr	Chez moi elle reste très rosée, mais elle **est bien** chaude à coeur.
danslemonde.net	La prédiction de l'avenir par la voyance par exemple **est bien** appréciable dans certain cas .

Table 9a : Lignes de concordance arrangées par phrase.

uvelle contradiction avec la réduction des pertes en ligne – ou rationner. Il	**est**	**bien** difficile de faire plus idéologique. Et de toute évidence, les négocia
e soient plus qu'un bashkir ou qu'un oudmourte ou bien un tchétchène. Je	**serais**	**bien** curieuse de connaitre vos sources. Vos pauvres bougres seraient v
niveau histologique (étude de tissus biologiques). La chaîne de processus	**est**	**bien** illustrée en regardant la régénération du foie après une injection to
ochains cuisiniers de mettre 8 jaunes et seulement 3 à 4 blancs. La crème	**est**	**bien** plus épaisse et onctueuse. Enfin parfait pour les gourmants ! Je vo
e tel par les gens du pouvoir. Mais cet enthousiasme et cette euphorie ont	**été**	**bien** éphémères. Quelques mois après l'instauration au pouvoir de l'Allia
d à quel usage ? Voici des éléments de réponse pour guider votre choix. Il	**est**	**bien** souvent difficile de faire son choix parmi les très nombreuses offres
dit quel genre de connexion internet tu as ? C'est du Wifi ou Ethernet ? C'	**est**	**bien** un modem que tu utilises aussi ? Quel fournisseur d'accès à intern
d'enjeux restent encore éclatés entre plusieurs ministères. L'un des défis	**est**	**bien** de fédérer les gestionnaires du niveau national au niveau local. En
ma vie : la Pandarie. Elle avait de grande falaise blanches et la végétation	**était**	**bien** verte. On a amarré dans un petit village appeler Rosée-de-Miel et l
e, comme il était décrit en fin de chapitre préliminaire. Mais ces propriétés	**seront**	**bien** plus mises en avant dans des développements ultérieurs. Sur le pl
ienne est à la limite de lui procurer un bon facesitting tellement ses fesses	**sont**	**bien** rondes. Le mec lèche la chatte de la grosse vicieuse durant cet ext
enfance et l'innocence perdue, la guerre civile entre Hutus et ... Ce roman	**est**	**bien** plus que le récit initiatique d'un jeune métis découvrant la guerre et
de tout autre dispositif motorisé. Il faut toujours s'assurer que le contenant	**est**	**bien** assujetti à la palette ou à un dispositif similaire. Limiter l'accès aux
ont solides, jusqu'aux masses d'arbres aérées qui les entourent. La neige	**est**	**bien** de la neige, et sa valeur sur le ciel est scrupuleusement exacte. An
es difficultés ? Les premières semaines sont dures car la charge de travail	**est**	**bien** plus importante qu'au lycée. Manon peine un peu car "les cours du
n. Bien obligé. Malgré soi. Presque rien à redire. Il y a sens. Mais parler c'	**est**	**bien** autre chose. Quel quiproquo peut se glisser! quel fouillis... Non. L'é
ıts, vous aimez les animaux n'est-ce pas? Et vous êtes attentifs à ce qu'ils	**soient**	**bien** traités? Alors vous partagez la même passion que le Parc Animalie
balement le rétablissement de la route départementale 106 par un viaduc,	**est**	**bien** accepté. Il est demandé aussi : - D'apporter une attention particuliè
dans la viande sans la cuire trop. Chez moi elle reste très rosée, mais elle	**est**	**bien** chaude à coeur. Et je la retourne souvent dans la "seconde mi-tem
z utilisé leurs forces . La prédiction de l'avenir par la voyance par exemple	**est**	**bien** appréciable dans certain cas . Un horoscope quant à votre avenir ,

Tableau 9b : Lignes de concordance arrangées par mot-clé en contexte (en rouge).

Sinclair (1991, 2004), Kjellmer (1984), Stubbs (1996) et Biber et al. (1998) ainsi que de nombreux autres linguistes de corpus, observent que *ce sont les lignes de concordance qui rendent l'omniprésence des*

schémas langagiers visible. Hanks (2012 : 401-403), en accord avec Sinclair et Stubbs, souligne également que les concordances font bien souvent émerger des collocations qui auraient probablement échappé à l'intuition. Les occurrences présentées dans les tableaux ci-dessous viennent illustrer cette proposition.

Alors que l'organisation par phrase dans le tableau 9a présente l'environnement textuel du mot de manière plus transparente, le tableau 9b est plus apte à révéler des répétitions lexico-grammaticales[32]. Par exemple, nous pouvons identifier la colligation suivante : X *est bien* + ADJECTIF/PARTICIPE (*difficile, curieuse, illustré, accepté*). Nous pouvons nuancer cette observation en catégorisant les instances de cette colligation selon leurs associations sémantiques : *X est bien illustré (« bien » exprime l'opinion positive du locuteur) / C'est bien difficile (« bien » intensifie l'adjectif) / X est bien vert (« bien » sert à confirmer et à renforcer un constat).* Une analyse plus approfondie révélerait d'autres caractéristiques importantes de ces unités multi-lexicales mais notre but est seulement ici d'attirer l'attention sur la multitude d'informations que le Concordancier peut fournir. Certains de ces renseignements méritent certainement d'être intégrés dans les matériels pédagogiques pour mieux refléter l'usage langagier des natifs et les sens typiques du mot choisi, comme nous le verrons par la suite.

4) Générateur de N-grams

Le générateur de N-grams présente des séquences continues de mots dans le corpus par ordre de fréquence. Le programme peut afficher tous les N-grams dans un corpus ou seuls ceux associés à la partie du discours ou à l'élément concret choisis. L'utilité de la liste des N-grams est qu'elle peut révéler des particularités d'usage, par exemple, le tableau 10 fait apparaître que l'infinitif « être » s'utilise le plus souvent avec des verbes modaux[33]. En cliquant sur les trois points à côté du numéro, les outils « Concordancier » et « Word Sketch » peuvent être sélectionnés pour une exploration approfondie (v. tableau 10).

[32] Nous retrouvons bien évidemment aussi des phrases comme « X est bien » sans aucun complément. Néanmoins, il convient de noter que cette expression n'est pas la seule à être fréquemment utilisée (bien qu'elle serait probablement la première qui viendrait à l'esprit) comme l'indiquent les lignes de concordance.
[33] Une analyse plus approfondie prouve que ce schéma fonctionne aussi avec d'autres verbes.

	Word	↓ Count ?		Word	↓ Count ?
1	ne peut être	20,995 •••	18	qui doit être	6,013 •••
2	avant d' être	20,049 •••	19	peut pas être	5,879 •••
3	doit pas être	12,502 •••	20	ne peut pas être	5,785 •••
4	peut être enlevé	11,445 •••	21	être considéré comme	5,586 •••
5	être enlevé et	11,412 •••	22	a pu être	5,387 •••
6	peut être enlevé et	11,400 •••	23	besoin d' être	4,723 •••
7	être enlevé et l	11,398 •••	24	est loin d' être	4,635 •••
8	ne pas être	11,056 •••	25	Il peut être	4,513 •••
9	ne doit pas être	10,537 •••	26	et peut être	4,365 •••
10	qui peut être	10,124 •••	27	peut être un	4,340 •••
11	loin d' être	9,867 •••	28	être mis en	4,227 •••
12	d' être un	7,465 •••	29	peut aussi être	4,129 •••
13	qui peuvent être	6,804 •••	30	d' être le	4,049 •••
14	ne peuvent être	6,789 •••	31	devraient être mieux	4,044 •••
15	vient d' être	6,736 •••	32	être à l'	4,037 •••
16	de l' être	6,331 •••	33	être mieux reliées	4,009 •••
17	il peut être	6,224 •••	34	être mieux reliées aux	4,008 •••

35	devraient être mieux reliées	4,008 •••	51	être écrit en	3,266 •••
36	de ne pas être	4,006 •••	52	être confondu avec	3,264 •••
37	l' être humain	3,975 •••	53	d' être à	3,247 •••
38	section devraient être mieux	3,962 •••	54	il doit être	3,231 •••
39	section devraient être	3,962 •••	55	d' être en	3,208 •••
40	cette section devraient être	3,957 •••	56	pas être écrit	3,204 •••
41	peut être utilisé	3,822 •••	57	pas être écrit en	3,182 •••
42	peut également être	3,729 •••	58	doit pas être écrit	3,181 •••
43	qui pourrait être	3,487 •••	59	être écrit en capitales	3,180 •••
44	après s' être	3,472 •••	60	être considérée comme	3,157 •••
45	susceptibles d' être	3,405 •••	61	aurait pu être	3,074 •••
46	et d' être	3,355 •••	62	peut être une	3,062 •••
47	il faut être	3,335 •••	63	pas être confondu	3,027 •••
48	ont pu être	3,332 •••			
49	d' être une	3,330 •••			
50	être à la	3,302 •••			

Tableau 10 : Les 3-et 4-grams les plus fréquents avec « être » (extrait).

Si maintenant nous cherchons des N-grams contenant le verbe « être » à la troisième personne du singulier, les 3- et 4-grams affichés font apparaître que le mot est fréquemment suivi d'un article défini ou indéfini et précédé du pronom « ce » (v. tableau 11) :

	Word	↓ Count
1	n' est pas	378,928
2	Cet article est	358,100
3	est la bienvenue	313,515
4	aide est la	313,213
5	aide est la bienvenue	313,202
6	Votre aide est	313,121
7	Votre aide est la	313,110
8	Cette section est	260,039
9	section est vide	259,678
10	Cette section est vide	259,674
11	est indexé par	150,033
12	article est indexé	150,025
13	article est indexé par	150,021
14	Cet article est indexé	150,020
15	est indexé par les	147,367
16	article est une	115,957
17	Cet article est une	115,230

	Word	↓ Count
18	est une ébauche	115,195
19	article est une ébauche	115,004
20	est une ébauche concernant	114,944
21	ce n' est	85,575
22	article est partiellement	82,261
23	Cet article est partiellement	82,252
24	est partiellement ou	82,197
25	est partiellement ou en	82,182
26	article est partiellement ou	82,182
27	ce qui est	69,454
28	qu' il est	64,940
29	que c' est	63,756
30	cet article est	62,473
31	c' est le	61,932
32	ce n' est pas	61,179
33	c' est un	59,867
34	de cet article est	59,771

35	est issu de	59,307
36	article est issu	58,597
37	cet article est issu	58,491
38	article est issu de	56,792
39	est issu de la	56,659
40	est la bienvenue pour	53,637
41	C' est un	53,593
42	c' est la	47,853
43	C' est le	47,075
44	C' est une	44,875
45	c' est une	43,392
46	et c' est	42,963
47	il n' est	41,285
48	Ce n' est	39,513
49	est une espèce	38,933
50	est un astéroïde	38,096

51	qui n' est	36,683
52	est une commune	32,307
53	C' est la	32,231
54	qui s' est	31,034
55	à l' est	30,370
56	n' est plus	29,576
57	il n' est pas	28,959
58	est l' un	27,976
59	c' est l'	27,719
60	c' est que	27,280
61	mais c' est	26,931
62	est l' un des	26,806
63	Il est le	26,632

Tableaux 11 : Les 3-et 4-grams les plus fréquents avec « est » (extrait).

La responsabilité incombe au linguiste d'interpréter correctement les données et d'identifier la présence éventuelle de biais. Par exemple, la liste ci-dessus (tableau 12) contient un nombre surprenant de N-grams avec le mot « article », tels que « cet article est », « article est issu », « article est indexé » ou encore « article est indexé par ». Une étude superficielle de ces occurrences peut en fournir une première explication : une partie du corpus a été compilée à partir d'articles publiés sur les sites Web, le texte contient donc des informations sur leur lieu de publication ainsi que des consignes sur la façon de les citer. (Pour éliminer ce biais lors d'une deuxième recherche, l'utilisateur peut décider de ne pas afficher les N-grams avec « est » contenant le mot « article ».) Il serait intéressant d'analyser ce résultat en classe pour en répertorier les différences et les ressemblances entre l'usage de « ce n'est » et « il n'est », phénomène qui peut poser problème aux apprenants de français.

La recherche avec la première personne du singulier offre des résultats encore différents (tableau 12) :

	Word	↓ Count ?		Word	↓ Count ?		Word	↓ Count ?
1	ne suis pas	27,106	14	Je suis d'	4,051	27	me suis dit	2,428
2	je me suis	21,526	15	je suis d'	3,858	28	suis un peu	2,397
3	je ne suis	19,841	16	Je suis un	3,754	29	Je suis en	2,383
4	je ne suis pas	16,601	17	mais je suis	3,651	30	que je me suis	2,368
5	que je suis	14,891	18	Je suis d' accord	3,486	31	je suis à	2,221
6	Je ne suis	11,997	19	je suis d' accord	3,279	32	ne suis pas sûr	2,171
7	Je ne suis pas	10,103	20	suis d' accord avec	3,234	33	suis pas un	2,131
8	Je me suis	9,354	21	je suis une	2,914	34	suis en train	2,120
9	et je suis	8,502	22	je suis très	2,823	35	si je suis	2,050
10	suis d' accord	6,888	23	que je ne suis	2,807	36	suis pas d'	1,978
11	je suis un	5,084	24	Je suis une	2,786	37	ne suis pas un	1,940
12	je suis en	4,596	25	Je suis très	2,627	38	je suis pas	1,899
13	j' en suis	4,279	26	suis pas sûr	2,499	39	suis pas d' accord	1,879

	Word	↓ Count ?	
40	je me suis dit	1,868	•••
41	ne suis pas d'	1,851	•••
42	me suis permis	1,844	•••
43	car je suis	1,828	•••
44	suis en train de	1,784	•••
45	et je me suis	1,755	•••
46	mais je ne suis	1,707	•••
47	ne me suis	1,659	•••
48	je suis bien	1,592	•••
49	j' y suis	1,562	•••
50	suis à la	1,496	•••

	Word	↓ Count ?	
51	je suis le	1,491	•••
52	dont je suis	1,410	•••
53	je suis sur	1,402	•••
54	je suis allé	1,395	•••
55	Je suis à	1,367	•••
56	suis permis de	1,338	•••
57	me suis permis de	1,330	•••
58	je suis dans	1,328	•••
59	suis tout à	1,322	•••
60	me suis rendu	1,316	•••
61	je suis un peu	1,316	•••
62	suis tout à fait	1,289	•••
63	je suis en train	1,246	•••

	Word	↓ Count ?	
64	quand je suis	1,227	•••
65	me suis fait	1,223	•••
66	je suis tombé	1,216	•••
67	je suis de	1,198	•••
68	et je ne suis	1,197	•••
69	je suis toujours	1,177	•••
70	je ne me suis	1,170	•••
71	Je suis le	1,150	•••
72	Je suis donc	1,150	•••
73	où je suis	1,144	•••
74	me suis pas	1,143	•••
75	moi je suis	1,142	•••
76	Mais je suis	1,112	•••

Tableau 12 : Les 3-et 4-grams les plus fréquents avec « suis » (extrait).

Les N-grams présentés dans les tableaux 10-12 indiquent clairement que les différentes formes d'un même lemme s'associent à des mots et à des structures grammaticales différents[34]. Ainsi, ces séquences de mots attirent notre attention sur l'interconnexion du lexique et de la grammaire qui, dans la logique de la linguistique de corpus, devraient être présentés ensemble dans les grammaires, les dictionnaires et les ouvrages pédagogiques[35].

5) Les outils intégrés dans des corpus pédagogiques présentés au chapitre 2

Cette section présentera brièvement les outils inclus dans les corpus pédagogiques décrits au chapitre 2. Nous pouvons tout d'abord constater que l'interface de ces plateformes est relativement simple et contient un Concordancier intégré. Une autre caractéristique commune à ces plateformes est qu'ils contiennent non seulement des textes ou du matériel multimédia mais aussi des activités. Le corpus « ELISA » propose un moteur de recherche simple et des listes de mots ainsi que des activités d'apprentissage pour certaines parties du corpus (Braun 2006). Le corpus « CWIC » comporte un moteur de recherche et une interface d'utilisateur personnalisée (Kennedy et Miceli 2010), et les corpus « Backbone » et « Sacodeyl » offrent un « Concordancier incorporé, facile à

[34] Selon Stubbs (2009), cette observation remet en question la validité du « lemme » en tant qu'unité linguistique.
[35] Voir aussi les chapitres 5 et 6 ainsi que la Partie II de cette thèse sur l'interconnexion du lexique et de la grammaire.

utiliser et des exercices pour un apprentissage plus efficace » (Chambers 2019 : 465). Le corpus « FLEURON » propose également un Concordancier intégré ainsi qu'un glossaire (André 2017). Dans certains de ces corpus, les textes sont regroupés par thème car les requêtes pédagogiques peuvent être liées aux thèmes présentés[36] ou au genre[37].

Braun (2006) propose d'améliorer la recherche dans le corpus par une annotation thématique par sujet et genre. Les apprenants pourraient ainsi non seulement effectuer des requêtes pour étudier des mots choisis mais pourraient aussi explorer les parties du corpus qui traitent du même thème. Ce qui est entendu ici par « thème » ne correspond pas nécessairement à un sujet du CECRL ; le terme, tel que défini par Braun, comprend également les fonctions pragmatiques comme, par exemple, « s'excuser » ou « se présenter ». En effet, des parties plus longues des textes peuvent être annotées en fonction des sujets qui y sont abordés (voyages, habitudes alimentaires, etc.) et aussi selon la fonction communicative des énoncés (salutation, négociation d'une décision, désaccord, etc.). Ce type d'annotation étendue est bénéfique en ce qu'elle permet de chercher dans tous les textes du corpus (car les énoncés à fonction pragmatique peuvent se retrouver dans les différentes parties) et d'identifier les modes d'expression caractéristiques liés au thème choisi (Braun 2010 ; Ädel et Reppen 2008 ; Biber et Egbert 2018). Braun souligne cependant que « cela implique l'identification et l'annotation de passages plus volumineux du corpus, au-delà d'une ligne ou d'une phrase de concordance – tâche qui ne sera guère réalisable sur la seule base des techniques du corpus » (Braun 2010 : 54, notre traduction).

Ces considérations font ressortir à nouveau le fait que les corpus pédagogiques ne peuvent pas être conçus selon les mêmes principes que les corpus à fins linguistiques si nous souhaitons que leur utilisation soit bénéfique pour l'enseignement des langues.

C) Mesures statistiques

Cette partie résumera les mesures statistiques les plus fréquemment utilisées lors de l'analyse de corpus. Selon Leech (2015), l'avantage des méthodes quantitatives est « la capacité de décerner des schémas […] qui n'étaient autrefois qu'intuitifs, mais qui peuvent maintenant être démontrés par des mesures statistiques des liens entre des éléments linguistiques » (p. 160, notre traduction). Nous avons vu à plusieurs reprises que *la notion de fréquence d'occurrence joue un rôle crucial dans l'analyse de*

[36] Les apprenants sont susceptibles de chercher des réponses à des questions telles que « Comment puis-je dire ce que j'aime à propos de ma ville ? » « Que puis-je dire à propos de mes loisirs ? »
[37] Par exemple : « Comment puis-je dire au revoir à la fin d'un e-mail informel ? »

corpus. De fait, la majorité des mesures que nous présenterons ci-dessous concernent, la fréquence des mots ou des séquences de mots. Nous présenterons en premier lieu les mesures statistiques de base et traiterons par la suite quelques mesures plus complexes.

1) Mesures statistiques de base

1.1) Fréquence normalisée et absolue

Les statistiques de base ne visent pas à tester la signification des résultats obtenus. Au contraire, elles ne présentent les données que de manière descriptive. La mesure statistique la plus élémentaire est le comptage des occurrences, c'est-à-dire le simple décompte du nombre d'exemples d'un mot ou d'un autre élément linguistique. Le tableau suivant affiche ces informations concernant le mot « décision » dans le corpus « frTenTen17 » du Sketch Engine (tableau 13) :

RESULT DETAILS

simple décision

Number of hits	**928,918**
Number of hits per million tokens	**135.7**
Percent of whole corpus	**0.014%**
Corpus size (tokens)	**6,845,630,573**

Nombre d'occurrences
Occurrences sur un million de tokens
Pourcentage dans le corpus entier
Taille de corpus (tokens)

Tableau 13 : Les statistiques d'occurrences du mot « décision » dans le corpus « frTenTen17 ».

Ce corpus contient 6 845 630 573 mots courants. Le mot « décision » y apparaît 928 918 fois, représentant 0,014% du total des données. La fréquence normalisée (ou la fréquence relative) de ce mot est 135,7 : ce chiffre signale à quelle fréquence le mot « décision » est susceptible d'émerger sur un million de mots (cf. Brezina 2018 ; Stefanowitsch 2020).

1.2) Le rapport type/token (RTT)

Le rapport type/token (type-token ratio) est une mesure de la complexité ou encore la diversité lexicale d'un ou de plusieurs textes. Il divise les « types » (le nombre total de mots différents) apparaissant dans un texte par les « tokens » (le nombre total de mots). Un rapport type/token élevé indique un degré important de variation lexicale tandis qu'un rapport bas indique le contraire, la plage se situant entre un 0 théorique (répétition infinie d'un seul type) et 1 (la non-répétition complète dans la séquence analysée). Par exemple, la phrase « J'ai téléphoné à mon ami et je lui ai raconté ma journée » contient 13 tokens mais seulement 11 types puisque « je » et « ai » sont répétés deux fois, le RTT de la phrase est donc 11/13, ou 0.84. Plus le rapport RTT est proche de 1, plus

la richesse lexicale du segment est grande. Il est également possible de comparer la complexité de plusieurs corpus de taille différente ; dans ce cas, une version normalisante de la procédure, le « rapport type-token standardisé » est utilisée (pour une explication détaillée v. Brezina 2018 : 58).

2) Mesures plus complexes

Nous distinguons deux types majeurs de mesures statistiques plus complexes : des mesures de la taille d'effet et des tests de signification statistique. La première catégorie correspond à un nombre mesurant la force de la relation entre deux variables dans une population statistique alors que la deuxième nous renseigne sur la signification du résultat obtenu (Evert 2009). Étant donné que les résultats issus d'un corpus sont en général soumis à des fluctuations, il est nécessaire d'utiliser ces tests pour évaluer la probabilité qu'un résultat particulier soit une coïncidence, ou non. Par exemple, si un mot apparaît avec une certaine fréquence dans le corpus étudié, il est nécessaire de comparer cette observation avec un corpus de référence pour savoir si ce résultat est spécifique au corpus étudié (i.e. non significatif) ou s'il s'agit d'un phénomène plus global. En général, un résultat est considéré comme significatif s'il y a 95% de chances qu'il ne soit pas attribuable au hasard (Brezina 2018).

2.1) Les tests de signification

Les deux utilisations les plus courantes des tests de signification en linguistique de corpus sont *le calcul de mots-clés et le calcul d'associations de mots* (Brezina 2018). Pour extraire des mots-clés, il convient de tester la signification de chaque mot apparaissant dans un corpus, en comparant sa fréquence à celle du même mot dans un corpus de référence. Lorsqu'il s'agit d'identifier les collocatifs pertinents d'un mot (par exemple pour un dictionnaire d'apprenant ou pour un manuel), il faut tester la signification de la fréquence de cooccurrence de ces mots et la confronter à celle d'autres mots apparaissant dans l'environnement immédiat de l'élément étudié. Ces procédés impliquent la réalisation de plusieurs milliers de tests (automatisés), par les outils intégrés dans le logiciel : l'utilisateur n'est donc confronté qu'aux résultats de l'analyse, c'est-à-dire la liste des mots-clés et des associations de mots pertinentes (Brezina 2018).

Ces deux grands groupes de mesures statistiques mettent en évidence différents aspects de l'association. La mesure statistique la plus appropriée est déterminée par le sujet de recherche. En concernant le choix, Evert (2009 : 34) donne les recommandations suivantes :

> « Concernant la signification, la théorie mathématique […] identifie clairement la vraisemblance logarithmique (Log-likelihood) comme la mesure la plus appropriée et

> la plus pratique. Concernant la taille d'effet, aucune recommandation claire ne peut être faite, car les mesures ont tendance à se concentrer sur différents aspects de la collocativité. En particulier, la mesure de l'Information mutuelle (Mutual Information) semble appropriée pour les associations relativement faibles qui sont comparées à la ligne de base de l'indépendance, tandis que logDice identifie des combinaisons de mots fixes avec une association presque totale » (notre traduction).

Les mesures statistiques concernant la force d'attractivité entre les éléments d'associations de mots donnent des informations sur la typicité de l'association. Par exemple, une « belle maison » ou une « jeune femme » sont des collocations faibles car leurs deux éléments, très fréquents dans le corpus, peuvent former des unités multi-lexicales avec un grand nombre d'autres mots. Ces collocations ne peuvent donc pas être considérées typiques dans le sens statistique - ce qui n'implique pas qu'elles soient inintéressantes pour l'enseignement des langues. Les collocations « chambre d'hôte » ou « voix suave » sont des collocations fortes car « Y » (d'hôte, suave) ne se combine pas avec une grande variété de mots en dehors de « X » (chambre, voix). Les outils les plus couramment utilisés pour mesurer la force des collocations sont l'Information mutuelle, le score T et le LogDice. Il est cependant important de noter que ces statistiques ne donnent pas de renseignements concernant les relations grammaticales entre les éléments.

2.2) Mesures de la force d'attractivité entre les membres d'une association de mots (1) : L'Information mutuelle (Mutual Information)

Le score d'Information mutuelle (IM) exprime le degré d'interdépendance statistique de deux mots. Cette mesure compare la probabilité d'observer les deux mots ensemble avec la probabilité de les observer séparément. Plus le score IM est élevé, plus le lien entre deux éléments est fort : un score de 3,0 ou plus, est à interpréter comme preuve que deux éléments sont des collocatifs. Plus le score IM se rapproche de 0, plus il y a de chances que les deux éléments apparaissent ensemble de façon aléatoire (Brezina 2018 ; Sketch Engine 2021 ; Stefanowitsch 2020).

Le problème avec cette mesure est qu'elle est fortement affectée par la fréquence des mots individuels : les associations de mots à basse fréquence ont tendance à atteindre un score IM élevé qui peut être trompeur comme les noms propres, termes techniques, expressions idiosyncratiques (« post mortem », « linguistique de corpus », « score T »). C'est la raison pour laquelle certains logiciels permettent de définir une limite de fréquence qui exclut les mots de basse fréquence du calcul (cf. Sketch Engine 2021).

2.2) Mesures de la force d'attractivité entre les membres d'une association de mots (2) : Score T

Le score T peut complémenter le score IM de façon utile car cette mesure tient compte de la fréquence de l'association des mots dans le corpus. Le score T exprime la certitude avec laquelle nous pouvons affirmer qu'il existe une association entre les mots, c'est-à-dire que leur cooccurrence n'est pas aléatoire.

Ce score permettra d'identifier, d'une part, des paires de mots grammaticalement conditionnées comme « dépendre de » « inciter à », et de l'autre les combinaisons stéréotypées qui ne sont pas confinées à des domaines ou des textes particuliers comme « décision finale », « prendre une décision » ou « faire le point ». Dans ces deux cas, le degré d'association n'est pas nécessairement très fort mais le score confirme qu'il s'agit très probablement d'un partenariat non-aléatoire car le corpus en contient un nombre élevé d'exemples.

Un problème potentiel de cette mesure est que la valeur est affectée par la fréquence de la collocation dans le corpus, ce qui explique pourquoi les combinaisons de mots très fréquentes ont tendance à atteindre un score T élevé bien qu'elles ne soient pas des collocations « fortes ». Cette limite prise en compte, dans la plupart des cas, le score T est cependant considéré plus fiable que le score IM (Sketch Engine 2021).

2.3) Mesures de la force d'attractivité entre les membres d'une association de mots (3) : LogDice

Le « LogDice » est un autre outil statistique utilisé pour l'identification de la force d'association entre les mots. Il est uniquement fondé sur la fréquence de la base et celle du collocatif, et de la fréquence de la cooccurrence de la base avec le collocatif. Cette mesure n'est pas affectée par la taille de la collection et peut donc être utilisée pour comparer les scores entre différents corpus. Le « LogDice » est la mesure préférée pour l'analyse de la typicité des associations de mots dans un grand corpus contenant un nombre significatif d'instances avec les éléments choisis[38].

[38] Pour une discussion plus approfondie sur les mesures statistiques, voir Brezina 2018 ; Evert 2009 ; Gries 2015 ; Stefanowitsch 2020.

Ce chapitre a présenté les termes-clés de la linguistique de corpus, quelques outils d'analyse ainsi que des mesures statistiques standards. Ce tour d'horizon nous a permis de mieux comprendre comment interpréter les résultats de l'analyse de corpus. Le chapitre 4 exposera les résultats les plus pertinents du domaine, obtenus avec les outils et mesures présentés ci-dessus, susceptibles d'être utiles pour l'apprentissage des langues.

Chapitre 4 : Les résultats pertinents de la linguistique de corpus pour l'enseignement des langues

« *Si nous accédons à l'ensemble des noms qui sont les compléments d'objet direct usuels de tel ou tel verbe, ou tous les adjectifs qui modifient typiquement tel ou tel nom, nous aurons fait le premier pas vers la découverte d'un ensemble riche mais gérable de schémas et de variations linguistiques, intimement lié à la réalisation du sens.* »

(Hanks 2013 : 7, notre traduction)

Ce chapitre a pour but de résumer les résultats les plus importants de la linguistique de corpus pour l'enseignement des langues. Ces résultats ont contribué au développement des approches pédagogiques de « l'apprentissage basé sur l'usage » dont les principes fondamentaux seront également exposés dans ce chapitre. Ceux-ci s'articulent autour de six grands axes : (1) l'optimisation de la méthodologie de l'analyse du langage ; (2) les caractéristiques de la co-sélection des éléments langagiers ; (3) l'identification de différents types de schémas ; (4) l'interconnexion du lexique et de la grammaire ; (5) l'importance des registres pour l'usage langagier ; (6) l'importance égale du langage transactionnel et du langage interactionnel. Les différentes sections de ce chapitre exploreront ces aspects et présenteront leurs implications majeures dans le cadre pédagogique. Ces différents aspects ne sont pas séparables : ils se complémentent en représentant les différentes façades de la même construction de l'usage langagier, telle qu'elle se révèle au linguiste de corpus.

A) Aperçu général

Les considérations pédagogiques intégrant les résultats de l'analyse de corpus s'inscrivent dans le cadre de l'apprentissage des langues basé sur l'usage (*usage-based approaches to language acquisition*) (cf. Barlow et Kemmer 2000 ; Langacker 2008 ; Pérez-Paredes et al. 2020 ; Pérez-Paredes et Mark 2021 ; Schmid 2016 ; Taylor 2012 ; Tomasello 2003). L'idée centrale de ces modèles est que la langue des locuteurs émerge à la suite d'une exposition à de nombreuses situations d'utilisation, et l'utilisation (réception et production) de la langue a toujours pour but de transmettre une signification particulière dans une situation de communication bien spécifique. Ces approches soulignent que *l'utilisation réelle de la langue façonne la forme linguistique et doit servir à ce titre de fondement à l'apprentissage des langues*. Tyler (2010 : 271-273) résume de façon suivante les cinq principes-clés de l'approche fondée sur l'usage :

- Le but premier du langage est la communication et l'usage communicatif influence le développement du langage et les changements langagiers.
- Le langage se produit toujours dans un contexte et ce contexte lui-même est une construction complexe. Par exemple, toutes les approches fondées sur l'usage identifient comme composante majeure du contexte les participants à une interaction. Des changements subtils dans la relation entre interlocuteurs entraînent ainsi des changements dans leurs choix linguistiques.
- Le sens ne se situe pas uniquement dans les éléments lexicaux ; les schémas grammaticaux contribuent également à sa construction (par exemple, Givón 2001 ; Langacker 2008 ; Boers 2021).
- La langue est apprise. Les modèles d'utilisation, y compris les différents schémas et les informations sur la fréquence d'occurrences de certains termes, sont considérés comme essentiels à cet apprentissage. Les approches fondées sur l'usage se tournent ainsi vers la fréquence pour expliquer les « régularités » du langage, à savoir « des régularités structurelles émergeant de l'analyse des caractéristiques distributionnelles de l'apport linguistique » (Ellis 2002 : 144, notre traduction). Cette approche considère l'apprenant en langue comme un « statisticien intuitif » (Harrington et Dennis 2002) qui tient compte inconsciemment de ces régularités distributionnelles.
- Le langage peut être expliqué par des modèles monostrataux, c'est-à-dire par des modèles qui ne distinguent pas deux ou plusieurs niveaux tels que la structure profonde et la structure de surface (Zyzik 2009).

Dans les pages suivantes, nous explorerons les résultats de la recherche en linguistique de corpus dans l'optique de les intégrer dans l'enseignement fondé sur l'usage.

B) Les unités multi-lexicales sont le noyau du lexique.

1) Qu'est-ce que cela signifie ?

Le fait que *les unités multi-lexicales soient au cœur du lexique* (cf. Sinclair 1991 ; Hoey 2005 ; Stubbs 2009) est selon toute vraisemblance l'une des conclusions les plus pertinentes de la recherche en linguistique de corpus transférables à la création de méthodologies d'enseignement efficaces. O'Keeffe et al. (2007 : 60) constatent que « les corpus révèlent *qu'une grande partie de notre production linguistique se compose d'unités multi-lexicales répétées plutôt que de simples mots* » (notre traduction, nous soulignons).

Pawley et Syder (1983 : 191) notent que « la maîtrise de l'usage idiomatique d'une langue repose dans une large mesure sur la connaissance d'un ensemble de "radicaux de phrases" (*sentence stems*) qui sont "institutionnalisés" ou "lexicalisés" ». Un radical de phrase est une unité à plusieurs composantes « dont la forme grammaticale et le contenu lexical sont entièrement ou largement fixes ». Sinclair (1991 : 110) note qu'un « utilisateur expert de langue a à sa disposition un grand nombre de phrases partiellement préconstruites qui constituent des choix uniques, même si elles peuvent sembler analysables en segments » (notre traduction).

Davis et Kryszewska (2012 : 13) suggèrent dans leur livre sur l'enseignement des unités multi-lexicales, que ces « énoncés probables » servent à assurer la fluidité d'une conversation, à établir des terrains d'entente entre les locuteurs, à « soulager l'effort cognitif et à augmenter le caractère naturel du langage ». Un argument similaire est avancé par Girard et Sionis (2004), cité dans O'Keeffe et al. (2007 : 77) :

> « L'utilisation des unités multi-lexicales toutes faites (*chunks*) « allège » la charge cognitive en allouant l'effort mental à d'autres aspects de la production [du langage] tels que l'organisation du discours et la fluidité de l'interaction » (notre traduction).

Enfin, il a été suggéré qu'« il est impossible de s'exprimer à un niveau acceptable pour les utilisateurs natifs, à l'écrit ou à l'oral, sans contrôler un spectre approprié d'unités multi-lexicales » (Cowie 2002 : 10, notre traduction ; cf. McCarthy 2002, 2003 ; Meunier 2012 ; Meunier et Granger 2008 ; Wray 2007). Cela implique que les aspects importants du sens d'un mot ne sont pas contenus dans le mot lui-même, mais sont formés par des associations caractéristiques et fréquentes auxquelles le mot participe (Hoey 2005). Comme Salazar (2004 : 9), en reprenant Sinclair (2004), le souligne, « [l]es mots n'apparaissent pas en isolement, mais se combinent les uns avec les autres pour générer du sens » (notre traduction). Nous pourrions également citer la fameuse phrase de Firth (1957 : 11) : « Vous connaîtrez un mot par son entourage » (You shall know a word by the company it keeps).

Le travail systématique concernant l'interdépendance des mots et de leur environnement textuel a été initié par Sinclair qui, comme démontre la citation suivante, s'oppose à définir le sens des mots isolément (2004a : 20) :

> « La position généralement acceptée en lexicologie est que chaque mot réalise une ou plusieurs significations […] et que lorsqu'une de ses significations est retenue, le mot est choisi. Reliant cette position à l'analyse des corpus, nous la considérerions comme confirmée si le choix d'un mot paraissait sans rapport avec le co-texte, choisi indépendamment des mots qui l'entourent ; mais les preuves à l'heure actuelle indiquent le contraire. Les indices suggèrent non seulement que les mots sont co-sélectionnés avec d'autres mots pour former des structures lexicales complexes […], mais aussi qu'une structure lexicale dans sa forme complète ne réalise normalement qu'un seul sens : c'est le co-texte qui élimine toute ambiguïté » (notre traduction).

Sinclair (2004a : 25) observe que (1) les mots ont tendance à s'associer et à produire un sens non ambigu par leurs combinaisons ; (2) qu'ils entrent en relation avec les mots qui les entourent, ce qui peut « compromettre l'image du mot comme unité indépendante de la langue » (notre traduction). Il constate que le sens que le locuteur souhaite véhiculer est souvent, ou peut-être même toujours, exprimé par un ensemble de plusieurs mots et que les modalités de leur co-sélection ont un lien direct avec le sens (Sinclair 2004b : 133). Sinclair propose d'introduire la catégorie d'« élément lexical » (*lexical item*) en tant que catégorie abstraite, distincte du « mot » qu'il décrit comme une unité de sens conciliant les caractéristiques paradigmatiques et syntagmatiques, observables en étudiant les lignes de concordance (Sinclair 2004b : 133, 144)[39]. D'après Sinclair (1997 : 29), l'analyse d'une centaine d'exemples concernant l'élément choisi comme point de départ de l'exploration est suffisante pour décrire ces unités avec toutes ses composantes grammaticales, lexicales, sémantiques et pragmatiques. Stubbs (2009) utilise, entre autres, l'expression « strong point » (point fort) pour explorer et révéler le phénomène de la co-sélection tel que défini par Sinclair[40] (tableau 14) :

[39] Dans cette thèse, nous utiliserons, pour éviter la confusion, les termes « unité multi-lexicale » et « unité de sens ». Le terme « unité lexicale » ne sera pas utilisé.

[40] Sinclair utilise, entre autres, les mots « budge » et « way » pour illustrer son argument. L'usage de ces mots se compare cependant difficilement avec le hongrois et avec le français, d'où notre choix de « strong point ».

Concordance 1: Illustrative examples of *strong point* preceded by a negative.

01It was soon clear that rowing was not my strong point. At hockey there was a v
03ting as usual for arithmetic was not her strong point especially mental arithm
05nating between men is evidently not your strong point. Perhaps a few lessons m
06rope. Zoological accuracy was not Tulp's strong point. The animal was a chimpa
07knowledge that, cooking was not Stella's strong point, for it had turned out
08 n turmoil consultation is not the BMC's strong point when it comes to mountai
09sion or argument anyway. that wasn't her strong point. Her eyesight was her st
10 organisation of business er isn't their strong point at the moment, whether i
11re that thinking doesn't seem to be your strong point. So why don't you try li
12d. The original XR's gearbox was never a strong point. Clean changes were poss
13need improving? Electronics was never my strong point. They hadn't invented el
14 f things I shouldn't. Tact never was my strong point, as Maxim will tell you.
15onfesses that finance has never been his strong point, broadens his horizons o
16at the young characters, never O'Casey's strong point, were played with great
17 Economic analysis was never Trevelyan's strong point and the England of the
18nomic management has never been Labour's strong point. The opinion polls conti
19isation has never been the IT industry's strong point, and the answer is "prob
20r. Contemporary art was anything but the strong point of the Salon, with Paris

Tableau 14 : Lignes de concordance concernant l'expression « strong point » (point fort) (Stubbs 2009).

En effet, l'expression « strong point » elle-même est déjà le résultat d'une co-sélection entre un adjectif et un nom. Si nous ne tenons compte que de ces deux éléments, nous pourrons dire que cette unité à deux composantes introduit la présentation des qualités d'une personne ou d'un groupe de personnes. Or, l'étude de l'environnement textuel plus étendu (les voisins de gauche et de droite de l'unité) révèle d'autres exemples de co-sélection. Tout d'abord, nous constatons que « strong point » est généralement précédé d'une négation car l'expression indique, en fait, une ou plusieurs faiblesses de la personne en utilisant un euphémisme. Le terme négatif semble donc faire partie de l'unité multi-lexicale « NEG + strong point », quand le locuteur souhaite exprimer ce sens.

Dans l'environnement immédiat de l'expression, nous trouvons encore d'autres éléments lexicaux et grammaticaux co-sélectionnés. « Strong point » s'inscrit en général dans une construction possessive précédée d'un article possessif (« my », « your ») ou d'un possesseur (« Labour's », « O'Casey's »). Le verbe dans ces phrases est « be » (être), au présent ou au passé. Enfin, l'expression « strong point » est généralement placée à la fin d'une proposition ou d'une phrase.

Nous pouvons continuer l'analyse de Sinclair pour collecter d'autres informations sur cette unité multi-lexicale[41]. En parcourant les 699 occurrences de la base de données d'« enTenTen18 » sur Sketch Engine, nous trouvons 645 phrases (92%) possédant la structure identifiée ci-dessus. Le reste (54 occurrences) ne contient pas de mot négatif mais leur contenu sémantique diffère. Alors que les phrases négatives expriment une faiblesse en utilisant un euphémisme (« X n'est pas le point fort de Y » signifiant : « Y est vraiment mauvais en X »), les phrases affirmatives font référence à différentes situations de la vie ou à des compétitions (sportives ou autres) au cours desquelles le locuteur doit être conscient de ses propres points forts pour bien y performer. Les phrases sans mot négatif exposent donc une réflexion sincère avec un enjeu. Ainsi, l'analyse plus complète identifie deux structures avec « strong point » associés à deux sens différents.

Nous pourrions aller encore plus loin et étudier l'environnement textuel au-delà de la phrase pour trouver d'autres composantes de co-sélection (par exemple, la place typique du mot dans un paragraphe ou dans le texte). Nous pourrions également explorer toutes les occurrences (même celles peu typiques) dans le corpus dans l'intérêt d'une description plus précise. Sinclair, comme d'autres linguistes de corpus, attire par ailleurs notre attention sur la difficulté de produire une analyse exhaustive du comportement d'une unité multi-lexicale et d'en déceler les limites.

Il est également important de rappeler que les résultats obtenus pour une langue ne sont pas nécessairement les mêmes pour une autre : les mots peuvent entraîner des co-sélections différentes selon la langue considérée, un aspect particulièrement important dans le cadre de l'enseignement des langues. Néanmoins, en dépit des variations locales, le phénomène la co-sélection semble être universel (Hoey 2005 ; O'Keeffe et al. 2007 : 77 ; Sinclair 2004b ; Szudarski 2017). Nous illustrerons cet aspect en étudiant l'équivalent de « strong point » en hongrois et en français.

Une comparaison avec le hongrois révèle que cette expression est associée à un environnement textuel identique à l'anglais. On observe néanmoins certaines différences : en hongrois, « erősség* » (point fort) est plus souvent utilisé dans une phrase affirmative qu'en anglais. Seulement environ 25% des occurrences (11 867 sur 47 520) ne contiennent pas en effet de mots négatifs. Ces occurrences se divisent en deux grands groupes : nous trouvons des exemples dans lesquels le locuteur énonce le point fort de quelqu'un – ou même son propre « erősség* » (point fort). L'autre

[41] Le but de la démonstration de Sinclair est d'attirer l'attention sur la présence de la négation (composante grammaticale au sein de l'unité multi-lexicale) d'une part, et sur le continuum entre lexique et grammaire de l'autre.

utilisation de l'expression est une citation *verbatim* ou une paraphrase d'un passage de la Bible : « Le Seigneur est ma force » (Psaume 28 : 7)[42]. Le tableau 15 présente quelques exemples d'occurrences issus du corpus « huTenTen17 ».

ı valami mondani valója. </s><s> Mellesleg bocs hogy ilyen gagyin írtam le,nem	erősségem	a fogalmazás </s><s> Ayumu </s><s> Új tag </s><s> 8 hozzászólás </s><
gam a multikulturális közegben és maximálisan ügyfél-orientált vagyok. </s><s>	Erősségem	a pénzügyi szektor, tanácsadóként elsősorban ezen a területen tevékenyke
ő alatt leginkább mikro- és kiválallkozókat segítettem a munkámmal, ez az igazi	erősségem	. </s><s> Jelenleg két irodával működök. </s><s> Ügyfélkörüm lefedi szint
kezdünk mondatot... de na... </s><s> Bocsi mindenkinek, mostanában nem volt	erősségem	az írás. </s><s> De ígérem bepótolom, csak most olyan sokminden lett kör
ból kifolyólag nem nagyon értem meg az embereket (bár igaz a türelem nem az	erősségem	), és ők sem engem. </s><s> A Flippo egy általam kitalált történetnek a fős
. </s><s> Számomra ez az utóbbi nem volt idegen, hiszen a pályán is ez volt az	erősségem	. </s><s> Ha egy nevet kellene mondanom "mesteremül", akkor az minden
t kezd viselkedni olyankor vízvezeték szerelőként, kell bevetni magam, ami nem	erősségem	. </s><s> Igaziból, technikából csak azért nem buktattak meg soha, mert ak
></s> Hogyan kerüljek oda melléjük, vagy akár az egyik szülőhöz. </s><s> Nem	erősségem	a kommunikáció, úgy érzem nem tudnék velük beszélgetni, csak pár percig
d az estém másolással és tanulással töltöttem. </s><s> A geometria nem volt az	erősségem	és nagyon le voltam maradva az új osztályomhoz képest. </s><s> Még sze
ísér. </s><s> Korábban atlétikával kezdtem, ahol a futás, ugrás és dobás volt az	erősségem	, így mikor az atlétika edzések megszűntek, átpártoltam a kézilabdához, en
nogy a mélységélességre milyen hatással van – a fizika sajnos soha nem volt az	erősségem	. </s><s> Ha az lett volna, akkor ma nem verseket olvasnék hazafelé a hév
ák megteremtésében már eddig is igen fontos szerepet játszottak. </s><s> Nem	erősségem	a búcsúzkodás, meg még azért úgy is találkozunk, de szerettem volna pár

Tableau 15 : Occurrences avec le mot « erősség.* » dans le corpus « huTenTen15 » (extrait).

Le tableau 16 montre des lignes de concordance concernant l'expression « point fort » en français. Nous pouvons observer quelques différences significatives par rapport au hongrois et à l'anglais. L'expression française est utilisée majoritairement dans sa forme positive ; elle n'exprime donc pas une faiblesse par euphémisme mais doit être prise au pied de la lettre. Les locuteurs parlent vraiment de leurs « points forts » ou de ceux d'une situation, d'une organisation ou industrie. L'expression française ne s'applique donc pas particulièrement aux personnes : une entité inanimée peut avoir aussi bien ses points forts. L'usage ironique, dominant dans le hongrois et l'anglais, apparait nettement plus rarement : parmi les 46 616 occurrences dans le corpus de « frTenTen17 » seulement 1561 (soit 3,5 %) contiennent une négation avec un sens comparable aux phrases anglaises et hongroises. L'expression est en général précédée d'un pronom possessif ou suivie d'une construction avec la préposition « de », indiquant une relation de possession, identique aux expressions anglaise et hongroise.

[42] « Ma force » est traduite en hongrois par « erősségem ».

.OI SENIORS-RECRUT'SENIORS est un événement décliné autour de 4	points forts	: </s><s> Une exposition rassemblant les entreprises qui ont compris l'int
et se mèle agréablement au sang dégoulinant des zombis : le deuxième	point fort	du film à mon gout. </s><s> Rien de remarquable, la scène d'actions sont
e d'innovation et un design unique qui ont fait sa renommée. </s><s> Les	points forts	</s><s> Pour se différencier des autres marques du domaine de la puéric
faibles revenus se voient écartés de la prévoyance. néoliane affiche des	points forts	qui en font un leader de l'assurance dans l'Hexagone : réactivité, qualité c
et informatique et une gamme d'armoire forte. </s><s> Domeau Concept,	Point Fort	Fichet et Serrurier à Sèvres depuis 1972 90 Grande Rue - 92310 Sèvres
, démocratiques. </s><s> Tout, mais conceptualisé ce qui est à la fois un	point fort	et un risque de faiblesse. </s><s> Au fil de ma lecture j'ai relevé des obje
fédérations industrielles ont voulu mettre en avant dans un manifeste les	points forts	d'une l'industrie innovante, porteuse de solutions pour l'avenir du pays. </
t d'ailleurs d'un puissant mode de prise de vue connectée qui est l'un des	points forts	du logiciel. </s><s> Quant au prix, il a bien baissé, mais se situe encore t
nanière entièrement autonome pour eux. </s><s> Après une analyse des	points forts	et des faiblesses de l'ancienne version, Data Filiation a apporté ses précc
remier événement à consacrer une soirée au Crowdfunding. </s><s> Les	points forts	de cette édition : </s><s> Projection de plus de 20 films et documentaires

Tableau 16 : Occurrences avec l'expression « point fort » dans le corpus « frTenTen17 » (extrait).

Cette comparaison d'une même expression en trois langues différentes souligne un autre aspect important de l'analyse de lignes de concordance : *elle peut révéler des différences significatives concernant l'utilisation des mots et des expressions dans plusieurs langues alors que leur sens peut sembler le même à première vue*. Par conséquent, les résultats de l'analyse dans une langue ne sont pas automatiquement transposables dans une autre langue même si la traduction de l'élément choisi est en principe identique (« strong point », « erősség », « point fort »). Les exemples confirment également que *c'est l'environnement textuel plus large qui contribue à préciser le sens et à révéler des différences entre l'usage de l'expression dans les trois langues* : les exemples anglais pointent la faiblesse d'une personne, le hongrois permet plus de flexibilité (les phrases peuvent être ironiques mais aussi avoir un sens littéral) alors que le français utilise, à quelques exceptions près, l'expression « point fort » dans un sens affirmatif, en étendant son usage sur les entités inanimées. Dans tous les cas, même s'il s'exprime différemment, le phénomène de co-sélection existe dans les trois langues.

Un autre exemple largement cité de co-sélection est tiré de l'étude du mot anglais « *cause* » (générer, provoquer, occasionner, produire) par Stubbs (2009). L'analyse révèle que ce verbe apparaît généralement avec d'autres mots (noms et adjectifs) possédant une composante négative. La « négativité » n'est donc pas un attribut du verbe mais une caractéristique typique des unités multi-lexicales qui l'incluent (v. le tableau 17).

cause problems	180	causing unnecessary suffering	14
causing death	99	causing serious injury	11
cause trouble	60	cause serious injury	10
causes problems	45	cause mental handicap	9
causing damage	41	cause serious damage	8
causing problems	41	cause severe damage	8
cause damage	40	caused extensive damage	7
cause cancer	37	causing criminal damage	7
cause difficulties	36	caused considerable damage	6
cause injury	36	caused great concern	6

Tableau 17 : Collocations fréquentes avec « cause » (provoquer, générer) (Stubbs 2009).

Dans le dictionnaire en ligne de Merriam-Webster, les sens du verbe listés sont les suivants : (1) « to serve as a cause or occasion of : *cause* an accident » (servir de cause ou occasion de : *provoquer* un accident) (2) « to compel by command, authority, or force : *caused* him to resign » (contraindre par ordre, autorité ou force : l'a *fait* démissionner). Ces définitions suggèrent que le verbe a une charge sémantique plutôt neutre, alors que les exemples observés dans le corpus révèlent une charge négative. Hunston (2007 : 252-3) a approfondi l'analyse de Stubbs identifiant la présence d'occurrences provenant majoritairement des articles scientifiques pour lesquelles le résultat « provoqué » (*caused*) n'est ni désirable ni négatif, mais neutre[43]. Hunston conclut ainsi son analyse en nous rappelant qu'il faut également considérer *le genre du texte et la situation communicationnelle* en décrivant les caractéristiques des unités multi-lexicales car les résultats peuvent différer selon ces facteurs[44]. Cette dernière analyse souligne l'importance d'un examen critique d'une analyse de corpus, en particulier l'identification de biais.

Ces exemples démontrent la forte interconnexion du mot étudié avec son environnement textuel. Ils font apparaître deux caractéristiques importantes de l'usage langagier : (1) premièrement, que chacun des éléments co-sélectionnés contribue à réaliser le sens désiré, non ambigu. (2) Deuxièmement, l'ensemble des co-sélections produites par des natifs et véhiculant un même message dénotent de fortes similitudes.

2) Quelles implications possibles pour l'enseignement des langues ?

Que le sens non ambigu d'une unité multi-lexicale, d'une phrase ou même d'un paragraphe soit le résultat d'une co-sélection plutôt que l'alignement de mots individuels, a un impact majeur sur

[43] Les exemples analysés par Stubbs proviennent de journaux non scientifiques (quotidiens et hebdomadaires), alors que ceux de Hunston incluent plusieurs genres de textes, y compris, donc, des articles parus dans des journaux scientifiques.
[44] Voir aussi les sections E) et F) ci-dessous.

l'enseignement des langues, en premier lieu sur la présentation du vocabulaire (Di Vito 2013 ; Gablasova et al. 2017). Si les apprenants doivent être capables d'interpréter correctement les énoncés des natifs et de produire eux-mêmes des énoncés avec un message clair, il est ainsi essentiel que l'enseignement se concentre sur la présentation d'unités multi-lexicales plutôt que sur celle des mots isolés (comme c'est le cas, par exemple, en proposant des listes de vocabulaire). O'Keeffe et al. (2007 : 63) suggèrent en ce sens que *les manuels devraient présenter et faire pratiquer les unités multi-lexicales considérées pertinentes pour le niveau donné dans tout le matériel, de manière cyclique.*

Pour expliquer aux apprenants ce qu'un mot « veut dire », il est préconisé de présenter son environnement textuel dans un certain nombre d'exemples ainsi que les schémas typiques de son usage[45], tout en adaptant cette présentation au niveau de compétences linguistiques des apprenants. Ces schémas fournissent des informations sur l'environnement textuel ainsi que sur les caractéristiques grammaticales, sémantiques et pragmatiques en attirant l'attention de l'apprenant sur le fait que ces aspects sont inséparables[46]. Cependant, *le principal bénéfice de cet exercice n'est pas la systématisation des résultats mais l'observation des exemples et l'accumulation de l'expérience linguistique qui en résulte*. Les rencontres multiples avec le mot-clé contribuent à l'assimilation consciente et inconsciente des schémas, alors que l'étude d'un tableau ne fait que systématiser les observations (Nation 2013 ; Szudarski 2017).

Finalement, force est de constater que l'apprenant qui ignore l'environnement textuel plus étendu peut être facilement trompé par la similitude de certaines expressions dans sa langue maternelle et dans la langue-cible. Les exemples montrent que l'exploration de l'environnement textuel de l'élément linguistique sélectionné est particulièrement utile car, en cas de doute, les apprenants sont susceptibles de prendre leur langue maternelle comme point de référence pour une traduction mot à mot (cf. Nation 2013 ; Szudarski 2017 ; Boers 2021). Il est également important d'attirer l'attention de l'apprenant sur les *enjeux pragmatiques* d'un tel choix. Par exemple, l'utilisation affirmative de l'unité multi-lexicale « strong point » dans la langue anglaise peut être perçue comme signe d'une confiance en soi démesurée même si le locuteur n'avait aucune intention de véhiculer cette image de lui-même.

[45] Voir la section C) pour une présentation possible des schémas.
[46] Voir la section D) sur l'interconnexion du lexique et de la grammaire.

C) Les schémas (patterns) sont omniprésents dans l'usage langagier.

1) Qu'est-ce que cela signifie ?

Dans le chapitre 3 (A.7), nous avons brièvement évoqué le terme « schéma » dans la linguistique de corpus. Il est possible d'établir de tels schémas pour chaque mot, que ce soit un verbe, un nom, un adverbe ou une autre partie du discours (cf. Hunston et Francis 2000 ; Hunston 2009, 2010 ; Krishnamurty 2006, Legallois 2012 ; Manca 2012 ; O'Keeffe et al. 2007 ; Sinclair 1991 et autres). La présentation des schémas contribue de ce fait à la systématisation des résultats de l'analyse de corpus.

Dans cette section, nous présenterons les catégories de schémas telles qu'établies par Sinclair ainsi que la catégorisation de Hoey (2005) qui, pour une description plus précise, enrichit celle de Sinclair des dimensions textuelles. Sinclair (2004 : 141) distingue cinq niveaux de co-sélection dont le premier et le dernier sont obligatoires. Ce sont : (1) l'existence d'un noyau, i.e. la preuve de l'occurrence d'un ensemble de mots comme unité, (2) la collocation, (3) la colligation, (4) la préférence sémantique et (5) la prosodie sémantique ou le sens/message de l'ensemble en tant qu'unité.

Au chapitre 3, nous avons présenté les définitions des termes « collocation » et « colligation » telles qu'utilisées dans ce projet de recherche. Ces définitions s'appuient essentiellement sur celles de Sinclair (2004a : 141-142) qui décrit les « collocations » comme des co-occurrences lexicales qui apparaissent à une distance de maximum quatre mots à gauche et à droite du mot ou de l'expression en question, tout en soulignant qu'il s'agit d'une approximation et non d'une mesure précise. Le terme « colligation » désigne la co-sélection grammaticale (par exemple, partie du discours, temps du verbe, mode du verbe, ordre des mots typiques). Sinclair s'appuie en ce sens sur les définitions de Firth (1957), mais il établit également deux autres catégories, celle de « la préférence sémantique » et de la « prosodie sémantique ». Le premier terme fait référence à l'observation que les mots régulièrement co-sélectionnés partagent certaines caractéristiques sémantiques, alors que le deuxième signifie que l'unité véhicule un message non ambigu dans son intégralité.

Dans le cadre de l'enseignement, O'Keeffe et al. (2007) proposent une terminologie légèrement différente en raison de la proximité des termes « préférence sémantique » et « prosodie sémantique ». Ils recommandent l'utilisation du terme « composantes sémantiques » pour la « préférence sémantique » et celui de « composantes pragmatiques » pour la « prosodie

sémantique ». Cette recommandation nous semble pertinente au sein d'une approche pédagogique[47].

Reprenons l'exemple de « strong point », « erősség » et « point fort » pour illustrer ces catégories et leur présentation possible (tableaux 18 et 19) :

STRONG POINT

Groupe 1 : X is not Y's strong point

Collocations typiques	**Sujets typiques :** des noms se référant à des qualités humaines
Colligations typiques	X + be + NEG + ArtPoss/NPoss + strong point Plus rarement : « POSS + strong point »
Composantes sémantiques	Évoque la faiblesse d'une personne ou d'un groupe de personnes.
Composantes pragmatiques	Tournure euphémique, éventuellement avec un élément d'humour.

Groupe 2 : X is Y's strong point

Collocations typiques	**Sujets typiques :** des noms se référant à des qualités humaines
Colligations typiques	ArtPoss/NPoss + strong point + be + X (nom ou phrase)
Composantes sémantiques	Évoque une qualité de personnalité ou d'un groupe de personnes.
Composantes pragmatiques	Fait partie d'une réflexion avec une notion d'enjeu.

Tableaux 18 et 19 : Profil de l'expression « strong point ».

Concernant les schémas langagiers, Sinclair (1991 : 4) attire notre attention sur deux observations importantes : « *[P]remièrement que certaines séquences de mots coexistent étonnamment souvent,* dans la mesure où chaque énoncé ou phrase écrite produite spontanément est unique ; *deuxièmement,* et à l'opposé de cette observation, *que même les expressions dites fixes démontrent des quantités surprenantes de variabilité* » (cité par Hunston and Francis 2000 : 21, nous soulignons, notre traduction). Il s'agit donc de

[47] Nos analyses dans la Partie II de cette thèse reprendront ces termes.

l'émergence des répétitions d'une part et celle des variations de l'autre, deux phénomènes qui deviennent observables lors de l'analyse de corpus.

Ces catégories sont essentiellement des outils pratiques pour la description des schémas mais elles ne donnent aucune explication pour l'omniprésence de ceux-ci, les théories langagières étant en général plutôt rares dans le domaine de la linguistique de corpus[48]. Comme le note Pace-Sigge (2013b : 151) : « [a]lors que la linguistique de corpus peut démontrer des relations entre les mots, *elle ne fournit rien de plus qu'une représentation de ce qui peut être observé. Elle n'explique pas pourquoi cela se produit en premier lieu* » (nous soulignons, notre traduction).

2) Expliquer l'existence des schémas : le Principe de l'idiomaticité de Sinclair et la théorie de l'Amorçage lexical (Lexical Priming) de Hoey

Pour l'existence des schémas, Sinclair propose une explication intitulée le « Principe de l'idiomaticité » (Idiom Principle). Ce principe énonce « *qu'un utilisateur d'une langue a à sa disposition un grand nombre de phrases semi-préconstruites qui constituent des choix uniques*, même si elles peuvent sembler analysables en segments » (Sinclair 1991 : 110, nous soulignons, notre traduction ; cf. Wray 2002). Selon cette interprétation, « NEG + strong point » est une phrase « semi-préconstruite » avec une position ouverte, celle du mot négatif, enregistré dans le dictionnaire mental de l'utilisateur en tant qu'unité dont un élément (le mot négatif) permet l'adaptation de la phrase à la situation en garantissant une certaine flexibilité, c'est-à-dire la possibilité de variations. D'après Sinclair, de telles structures sont omniprésentes dans nos énoncés et constituent donc la norme, contrairement aux choix libres (open slots), qui seraient plutôt l'exception (cf. Legallois 2006 : 32).

L'« Amorçage lexical » (Lexical Priming) par Hoey (2005) est une théorie reliant les observations sur l'usage langagier à des découvertes dans le domaine de la psycholinguistique. Hoey propose que les unités multi-lexicales ainsi que les schémas langagiers sont omniprésents dans la langue et *opèrent à tous les niveaux d'un texte*. D'après lui, l'amorçage (priming) est le moteur de l'usage, de la structure et du changement langagiers. Hoey (2005 : 8) note ainsi que :

> « Nous ne pouvons expliquer l'existence du concept de la « collocation » que si nous supposons que chaque mot est mentalement préparé pour une utilisation collocationnelle. Au fur et à mesure qu'un mot s'acquiert en le rencontrant à l'oral et

[48] Pour une explication, voir la section A.2 au chapitre 1.

> à l'écrit, il devient cumulativement chargé des contextes et des co-textes dans lesquels il est croisé, et notre connaissance de celui-ci inclut le fait qu'il coexiste avec certains autres mots, dans certains types de contexte. La même chose s'applique aux unités multi-lexicales construites à partir de ces mots ; ceux-ci deviennent, à leur tour, chargés des contextes et des co-textes dans lesquels ils se produisent » (notre traduction).

Selon cette théorie, le « priming » n'est donc pas nécessairement lié au sens mais plutôt aux habitudes d'usage : ce sont les rencontres répétées avec les éléments langagiers qui permettent que leur utilisation s'enracine (entrench) et ces enracinements, forment eux-mêmes la base d'autres amorçages. Pace-Sigge et Patterson (2017 : ix-x) expliquent le processus du « priming » comme suit :

> « L'exposition répétée d'une personne à des instances contextualisées de séquences phonétiques ou graphiques très similaires, amène cette personne à associer ces séquences aux caractéristiques récurrentes de ces contextes ; cette affirmation est basée sur une recherche psycholinguistique approfondie sur l'amorçage (priming) […] L'effet du « priming » généré par une telle exposition est que lorsque cette personne utilise le mot (ou un autre élément langagier) en question, elle reproduit généralement les caractéristiques récurrentes du contexte, assurant ainsi la perpétuation de l'association du mot (ou un autre élément) avec ces fonctionnalités ». (notre traduction)

Selon la théorie du « Priming lexical », la « charge cumulative » associée aux rencontres avec les éléments langagiers inclut les éléments suivants (Hoey 2005 : 13) :

> (1) Chaque mot est amorcé (primed) par d'autres mots particuliers ; ce sont ses collocatifs.
> (2) Chaque mot est amorcé par des ensembles sémantiques particuliers ; ce sont ses associations sémantiques.
> (3) Chaque mot est amorcé par des fonctions pragmatiques particulières ; ce sont ses associations pragmatiques.
> (4) Chaque mot est amorcé par (ou évite) certaines positions grammaticales et certaines fonctions grammaticales ; ce sont ses colligations.
> (5) Les co-hyponymes et synonymes diffèrent en ce qui concerne leurs collocations, leurs associations sémantiques et leurs colligations.

(6) Lorsqu'un mot est polysémique, les collocations, associations sémantiques et colligations d'un sens du mot diffèrent de celles de ses autres sens.
(7) Chaque mot est amorcé par un ou plusieurs rôles grammaticaux ; ce sont ses catégories grammaticales.
(8) Chaque mot est amorcé par (ou évite) des types particuliers de relations cohésives dans un discours ; ce sont ses collocations textuelles.
(9) Chaque mot est amorcé par des relations sémantiques particulières dans le discours ; ce sont ses associations sémantiques textuelles.
(10) Chaque mot est amorcé par (ou évite) certaines positions dans le discours ; ce sont ses colligations textuelles.

Hoey (2005 : 158) considère que sa théorie contextualise théoriquement et psychologiquement les idées de Sinclair sur le lexique en précisant que les dimensions textuelles jouent un rôle essentiel pour la co-sélection et, par conséquent, pour l'amorçage lexical. La « collocation » et la « colligation » telles qu'utilisées par Hoey peuvent être considérées comme équivalentes aux principes proposés par Sinclair ; les termes « préférence sémantique » et « association sémantique » sont également interchangeables (Hoey 2005 : 23). En revanche, Hoey évite le terme « prosodie sémantique » et le remplace par « association pragmatique ». Bien que proche, ce terme n'est pas l'exact équivalent de « prosodie sémantique », les différences provenant du fait que la théorie de Hoey prend les mots individuels comme point de départ de ses observations et propose une description du langage sur cette base, alors que Sinclair se concentre quant à lui sur les unités multi-lexicales. L'association pragmatique telle que définie par Hoey (2005 : 26) « se produit lorsqu'un mot ou une séquence de mots est associé à un ensemble de caractéristiques qui remplissent toutes des fonctions pragmatiques identiques ou similaires » (notre traduction). Les schémas textuels (points 8 à 10) ne sont pas traités de façon aussi systématique chez Sinclair, bien qu'il souligne leur importance (cf. Sinclair 2004b).

Hoey postule qu'il est non seulement possible de décrire l'environnement typique d'un mot mais aussi que, en cas de polysémie, *nous pouvons identifier des éléments que le mot a tendance à éviter quand il réalise un sens spécifique*. Il choisit plusieurs mots, entre autres le nom « reason » (raison), pour démontrer la validité de cette hypothèse. Ce mot a deux sens : (1) « logique, capacité rationnelle » d'une part et (2) « raison, cause » de l'autre. Hoey (2005 : 88-113) démontre, entre autres, que lorsque le mot signifie « cause », il est généralement précédé par l'article défini « the », alors que lorsqu'il fait référence à la capacité rationnelle d'une personne, il ne s'associe pas avec cet article.

Pour les études présentées dans la Partie II de cette thèse et pour les activités pédagogiques proposées dans la Partie III, nous nous concentrerons avant tout sur les quatre premiers principes observables à propos de tous les éléments langagiers, et nous ne ferons référence aux autres, plus spécifiques, que dans certains cas[49].

3) Quelles implications possibles pour l'enseignement des langues ?

Les études en linguistique de corpus démontrent que les mots ne sont pas des éléments isolés de la langue, mais s'inscrivent dans de nombreux schémas dont l'établissement offre une vue d'ensemble de leurs caractéristiques. Les identifier contribue à une meilleure compréhension de leur utilisation, et leur intégration dans les énoncés de l'apprenant entraine un usage langagier à caractère plus naturel. Il est donc recommandé que les activités visant l'observation et la pratique active des schémas linguistiques soient intégrées dans le matériel pédagogique et dans les cours de langues. Pour cela, O'Keeffe (2007) recommande notamment d'utiliser la catégorisation de Sinclair pour construire le « profil du mot » (word profile) choisi. Cette catégorisation a été présentée à travers l'exemple de l'unité multi-lexicale « strong point » dans les tableaux 20 et 21 et elle sera utilisée dans les Parties II et III de cette thèse.

L'existence des schémas souligne l'importance de la fréquence d'occurrences des éléments pour l'usage langagier[50]. Les utilisateurs, dont les apprenants, semblent en effet sensibles à la fréquence avec laquelle ils sont exposés à un élément donné (Ellis 2002 ; Nation 2013 ; Robinson et Ellis 2008 ; Taylor 2012 ; Tomasello 2003). D'après Taylor (2012 : 283), les distributions de fréquence « ne peuvent avoir été apprises que sur la base d'une exposition à l'usage » (notre traduction). Toute intuition linguistique émerge de l'usage ; le locuteur construit un schéma généralisé des événements sociaux et linguistiques récurrents dans lesquels un élément donné est couramment utilisé. La possibilité d'établir des schémas en tenant compte de la distribution de fréquence d'usage de différents mots ou de différents sens du même mot, « permet aux locuteurs d'étendre de manière créative leurs performances linguistiques au-delà des expressions déjà rencontrées, en offrant un modèle grâce auquel les éléments peuvent être combinés de manière originale » (Taylor 2012 : 284,

[49] Par exemple, nous considérerons les principes 5 et 6 en présentant l'usage des synonymes dans les chapitres 8 et 9.

[50] Nous ne pouvons pas exposer au sein de cette thèse les résultats de la recherche en psycholinguistique. Les expériences semblent confirmer qu'un mot rencontré sera mémorisé selon de nombreuses dimensions, incluant, entre autres, son environnement lexical immédiat, la construction syntaxique qui le contient et sa valeur sémantico-pragmatique. Les rencontres régulières et systématiques entraînent un renforcement de la représentation mentale de l'élément en question. Pour une excellente discussion de ce sujet du point de vue psycholinguistique, voir Taylor (2012).

notre traduction). Toutefois, pour que le locuteur parvienne à une compréhension aussi précise d'un mot ou d'une séquence de mots, il doit être exposé à de multiples exemples d'utilisation d'un élément dans divers contextes (Ellis 2002 ; Siyanova-Chanturia et Spina 2015). Or, de nombreux apprenants ne reçoivent jamais un apport linguistique aussi riche et varié. Comme le notent Omidian et Siyanova-Chanturia (2020) : « Le défi pour les apprenants consiste ici à reconnaître les associations sémantiques d'un élément avec certains contextes alors qu'ils n'ont pas encore développé un schéma représentatif de l'usage de cet élément dans leur esprit » (notre traduction).

Dans le cadre pédagogique, il serait donc souhaitable de présenter à plusieurs reprises les éléments langagiers importants, pour que l'apprenant ait suffisamment d'occasions d'observer et de mémoriser leurs usages divers avec les schémas qui leur appartiennent. Même si la catégorisation des caractéristiques typiques et la construction du « profil de mot » n'ont pas lieu de façon systématique pendant le cours, les rencontres fréquentes et la pratique régulière pourront renforcer les connaissances de l'apprenant et accroître ainsi ses compétences linguistiques.

D) Le lexique et la grammaire ne sont pas séparables.

« *Est-il sage de diviser la structure du langage en grammaire et autre chose (que ce soit le lexique ou la sémantique ou les deux) avant d'envisager la possibilité qu'il s'agisse d'un choix coordonné ?* »

(Sinclair 1991 : 3, notre traduction)

1) Qu'est-ce que cela signifie ?

La séparation entre lexique et grammaire a été pendant longtemps l'une des pierres angulaires de l'étude linguistique. La linguistique de corpus, comme certaines autres écoles[51], propose une autre approche : les résultats des analyses de grandes bases de données linguistiques indiquent qu'il existe une forte interconnexion entre le lexique et la grammaire (cf. Gabrielatos 2005 ; Römer et Schulze 2009 ; Sinclair 2000). Dans ce système que l'on peut appeler « grammaire lexique » ou « lexico-grammaire » (terme introduit par Halliday 1985 : 15 ; cf. Berber Sardinha 2019 ; Hoey 2005 ; Hunston 2015 ; Poole 2018), « les mots et les structures grammaticales ne sont pas considérés comme indépendants, mais plutôt interdépendants, l'un en relation avec l'autre » (Berber Sardinha 2019 : 1, notre traduction).

[51] Nous pensons, par exemple, à la grammaire de constructions (cf. Goldberg 1995) et à la grammaire cognitive (cf. Langacker 2008).

Biber et al. (1998 : 84) proposent de définir la lexico-grammaire comme « des associations entre des mots et des structures grammaticales » (*associations between words and grammatical structures*), et soulignent que ces associations peuvent être utilisées pour différencier des mots que l'on considérerait comme synonymes ou comme structures grammaticales similaires. Par exemple, les auteurs démontrent que les synonymes « begin » et « start » (entamer, commencer) entraînent des schémas grammaticaux différents : « begin » est en général utilisé avec une « to-clause » (proposition commençant par le mot « to ») et dans des structures transitives, alors que « start » apparait généralement dans des structures intransitives.

Les conclusions de Biber rejoignent celles de Hoey (2005) concernant la relation entre lexique et grammaire. La théorie de Hoey propose, comme exposé plus haut, une explication selon laquelle certains mots préfèrent ou évitent des associations lexicales, grammaticales, sémantiques, pragmatiques et textuelles particulières. Cela s'applique également aux différents sens du même mot. Les analyses de Hoey portent[52] sur le verbe « ponder » (considérer, réfléchir). Hoey utilise le corpus du journal britannique « The Guardian » pour expliquer la raison pour laquelle la phrase « X might be pondered » (X peut être réfléchi) ne semble pas naturelle aux locuteurs natifs. Il identifie dans le corpus 1057 exemples concernant le lemme « ponder » (il s'agit donc d'un verbe plutôt rare), dont seulement 8 sont au passif. Le verbe semble donc éviter l'utilisation au passif mais, comme révèlent les étapes successives de l'analyse, également l'utilisation du passé : le corpus ne contient en effet au total que 166 occurrences au passé (tous les temps du passé confondus), la majorité des phrases étant au présent.

Sinclair (1991, Chapitre 3) se base, entre autres, sur l'observation du mot « decline » pour démontrer la continuité entre le domaine du lexique et celui de la grammaire. Le verbe « decline », dit-il, a deux sens principaux : (1) refuser quelque chose et (2) être en baisse. Son analyse démontre que les différents sens se manifestent dans différents environnements textuels et ont des préférences grammaticales différentes. Le participe passé « declined » exprime en l'occurrence le sens de « refuser » alors que le participe présent « declining » est associé au sens d'« être en baisse ». L'interconnexion de la grammaire et du vocabulaire se manifeste également quand il ne s'agit pas de deux sens différents du même mot mais de légères modifications du sens qui dégagent des expressions plus longues avec le mot choisi. Nous illustrons cela en étudiant le verbe français « refuser » dans les expressions affirmatives et négatives à la première personne du singulier (« je

[52] Mis à part l'analyse des occurrences concernant le nom « reason », évoquée dans la section C.2.

refuse » et « je ne refuse pas »)[53]. L'étude des occurrences dans le corpus « frTenTen17 » révèle ainsi que différentes structures syntaxiques typiques (en bleu) accompagnent les deux expressions (tableaux 20 et 21) :

Je refuse (4 584)

Distribution des occurrences :

Je refuse de + INF : 2874 (62% des occurrences)

Je refuse (…) que … : 689 (15%)

Je refuse + COD : 864 (19%)

D'autres occurrences (par exemple, *ce que je refuse, je le refuse*) : 162 (4%)

Je ne refuse pas (355)

Distribution des occurrences :

Je ne refuse pas de : 93 (26%)

Je ne refuse pas que : 4 (1%)

Je ne refuse pas + COD : 258 (73% de toutes les occurrences)

Tableaux 20 et 21 : Les colligations de « je refuse » (tableau 20) et de « je ne refuse pas » (tableau 21) dans le corpus « frTenTen17 ».

Il suffit une simple observation de ces statistiques élémentaires pour constater la présence de différences significatives entre l'usage de « je refuse » et « je ne refuse pas ». Tout d'abord, la différence de fréquence d'utilisation est frappante : « je refuse » apparaît dix fois plus souvent dans le corpus que « je ne refuse pas ». Les deux utilisations – affirmative et négative – se distinguent également par leurs colligations typiques : « je refuse » est suivi le plus souvent de la préposition « de » et d'un verbe alors que « je ne refuse pas » est suivi majoritairement d'un complément d'objet direct[54, 55].

[53] Nous pourrions bien évidemment étendre les explorations sur les phrases avec d'autres mots de négation comme « jamais », « point » ou « guère » mais il s'agit seulement ici d'une démonstration simple de la validité de l'approche statistique.

[54] Nous pourrions continuer l'analyse afin d'établir d'autres schémas pour les deux utilisations au plan sémantique, par exemple. La phrase « je refuse de » signale une aversion forte vers la proposition qui la suit, or, la phrase négative suivie par le COD suggère la volonté du locuteur d'accepter avec plaisir ce qui lui a été proposé. Enfin, la phrase « je ne refuse pas de » indique une concession et est souvent suivie d'une conjonction indiquant une opposition (« mais », « néanmoins ») : J'ai refusé de … (INF) 816 ; J'ai refusé que … 32 ; J'ai refusé + COD : le 133, la 128, les 61, ce 31, cette 26, ces 7, un 67, une 48, des 42 (un colis, les avances de mon voisin, beaucoup d'offres, l'accès, la montée en série, l'hospitalisation, cinq ventes).

[55] Sinclair (1991, 1997, 2003, 2004b) étudie d'autres verbes à plusieurs sens, comme « yield » (donner, céder, produire, rapporter), et en tire des conclusions similaires. Ces exemples cités ainsi que les N-grams avec les

Non seulement les mots individuels mais aussi de nombreuses unités multi-lexicales font état de préférences grammaticales évidentes. Par exemple, l'expression « Comment vas-tu ? » est majoritairement utilisée au présent alors que les questions « Comment étais-tu hier ? » ou « Comment seras-tu demain ? » seraient envisageables du point de vue grammatical mais ces questions ne sont utilisées que très rarement et sont atypiques.

L'école néo-firthienne (v. la section 2 au chapitre 1) prend le lexique comme point de départ pour l'exploration de l'usage langagier (Hanks 2015) et constate qu'il existe des restrictions dans le choix des caractéristiques grammaticales qui peuvent accompagner une unité multi-lexicale réalisant un sens précis. Hunston et Francis (2000) utilisent l'approche inverse : elles prennent la grammaire comme point de départ de leurs explorations et intègrent les caractéristiques lexicales des éléments dans la présentation de la grammaire. Francis soulignait déjà en 1993 les restrictions imposées sur le lexique par un certain nombre de structures grammaticales. Ce concept a été largement développé par Francis et al. (1996 ; 1998) avant d'aboutir à un ouvrage entièrement consacré à la présentation d'un nouveau type de grammaire, la « Grammaire des schémas » (Hunston et Francis 2000). Cet ouvrage explore l'interface de la grammaire et du lexique et propose un inventaire des schémas lexico-grammaticaux pour l'anglais et présente une méthodologie détaillée, transposable à d'autres langues. Hunston et Francis (2000) définissent ainsi le « schéma grammatical » comme « une phraséologie fréquemment associée à (un sens d')un mot, en particulier en termes de prépositions, de groupes [verbaux et nominaux] et de clauses qui suivent le mot » (p. 3, notre traduction). Selon cette approche, grammaire et vocabulaire sont « mutuellement dépendants, en ce que chaque schéma se produit en conjonction avec un ensemble restreint d'éléments lexicaux, et chaque élément lexical se produit avec un ensemble restreint de schémas » (2000 : 3, notre traduction)[56].

L'étude d'exemples concrets de l'usage langagier réel a conduit McEnery et Hardie (2012 : 168) à observer que « [*l*]*'argument selon lequel la grammaire d'une langue et son lexique ne sont pas des entités séparées est cohérent avec une vision fonctionnelle plutôt que formelle de la langue*, qui donne la priorité à l'observation de l'utilisation réelle de la langue plutôt qu'à la dépendance à des instances de langage idéalisé s'appuyant sur l'intuition d'un locuteur » (nous soulignons, notre traduction). Bien que l'on puisse relever des différences dans les catégorisations de certains éléments langagiers effectuées par les

différentes formes du verbe « être » (chapitre 3) et l'environnement textuel de « point fort » (chapitre 3) fournissent d'autres exemples de l'interconnexion du lexique et de la grammaire.

[56] Voir aussi le chapitre 5 sur l'implémentation de la « Grammaire des schémas » dans les ouvrages pédagogiques.

différents chercheurs, l'existence d'une relation forte entre grammaire et lexique apparaît incontestable dans le domaine de la linguistique de corpus[57]. Pour conclure, citons une nouvelle fois Sinclair (2004b : 139) qui résume le principe essentiel de la lexico-grammaire comme suit : « la forme d'une unité langagière et sa signification sont deux perspectives sur le même événement » (notre traduction).

2) Quelles implications possibles pour l'enseignement des langues ?

Les résultats exposés dans cette section ont des implications profondes pour l'enseignement des langues. Ils suggèrent tout d'abord que « c'est le comportement des mots individuels qui dicte la description grammaticale » (Hunston 2009 : 142, notre traduction) et que toute présentation de la grammaire « doit tenir compte du lexique » (Sinclair 2001 : 353). Il est fort probable que les informations lexicales et grammaticales soient enregistrés ensemble dans la mémoire des locuteurs natifs et retirées sous forme d'« unités lexico-grammaticales prêtes à l'emploi » (ready-to-use lexicogrammatical units) (O'Keeffe et al. 2007 : 60).

Si les éléments lexicaux et grammaticaux sont sélectionnés ensemble et certaines formes s'associent de préférence à certains éléments lexicaux, il est logique de présenter les phénomènes grammaticaux avec le vocabulaire typique. Les auteurs de matériel pédagogique et les enseignants visant à guider les apprenants dans leur compréhension de l'usage langagier des natifs devraient donc s'efforcer de présenter autant d'unités lexico-grammaticales que possible. Cela est, par ailleurs, déjà le cas pour certains aspects de la langue. Ainsi, la présentation des articles possessifs (ou, pour les langues comme le hongrois, les terminaisons du possessif) est souvent associée à des textes sur les différents membres de la famille, et la formation du passé aux activités pendant le week-end ou les dernières vacances ou aux souvenirs d'enfance. Or, malgré certaines tentatives occasionnelles, la mise en œuvre de l'interconnexion de la grammaire et du lexique est loin d'être systématiquement poursuivie dans les ouvrages pédagogiques (cf. Römer 2006 ; Jones et Durrant 2010 ; Frankenberg-García et al. 2011 ; Hughes 2010 ; McCarten 2010).

L'approche lexico-grammaticale a des conséquences particulières pour les langues morphologiquement complexes (Jantunen et Brunni 2013). Une des principales caractéristiques de ces langues est que le phénomène de la suffixation est susceptible d'apporter des changements morphologiques dans le radical du mot. Ce phénomène est à présent peu exploré dans le cadre de

[57] Pour une analyse plus détaillée, voir Hunston (2015) qui utilise la structure « verb + someone + into + doing something » pour illustrer les similitudes et les écarts les plus importants.

la linguistique de corpus pour le hongrois. Pour cette raison, la Partie II de cette thèse sera consacrée dans son intégralité à l'étude de la relation entre les phénomènes lexicaux et grammaticaux en hongrois.

E) Le contexte social (registre) est important.

1) Qu'est-ce que cela signifie ?

Pour la recherche dans le domaine de la pragmatique, l'une des contributions majeures de la linguistique de corpus est la démonstration qu'*un « grand nombre d'utilisateurs de la langue, séparés dans le temps et dans l'espace, s'orientent à plusieurs reprises vers les mêmes choix linguistiques lorsqu'ils sont impliqués dans des activités sociales comparables.* » (O'Keeffe et al. 2007 : 60, notre traduction, nous soulignons).

Nous savons tous, intuitivement, que la langue n'est pas un « bloc » : nos expériences quotidiennes le confirment. Nous n'utilisons pas notre langue maternelle de la même manière lorsque nous parlons à un ami intime, à un fournisseur de services ou à un éminent professeur d'université. Nous avons à notre disposition un certain nombre d'éléments linguistiques qui diffèrent par leur degré de formalité et qui sont plus ou moins fixes, qu'il s'agisse du choix du vocabulaire, de la grammaire ou même de la prononciation. En observant d'autres personnes que nous-mêmes, nous pouvons constater le même phénomène : un professeur d'université parlera différemment avec ses collègues de bureau, avec ses étudiants au cours magistral ou avec la serveuse au café. Il sait changer de langage en fonction de la situation. Ces changements peuvent être cartographiés systématiquement car l'analyse de différents registres permet d'identifier les éléments fréquemment produits dans des contextes sociaux similaires (cf. Biber et Reppen 2002 ; Biber et al. 2002 ; Coxhead 2000 ; Friginal 2018 ; Leblanc 2016 ; Meunier et Granger 2008 ; Née et al. 2016 ; Poole 2018 ; Rühlemann 2007, 2018 ; Warren 2006).

Il convient également de noter que les différentes langues ne font pas nécessairement les mêmes choix pour réaliser un énoncé dans un registre donné. Par exemple, les livres de cuisine anglais utilisent dans les recettes de préférence l'impératif, les livres de cuisine allemands favorisent « man », « Sie » ou l'infinitif, alors que les livres hongrois emploient l'impératif de la première personne du pluriel et les livres français l'infinitif du verbe. Il est vrai pour toutes ces langues que dans une interaction orale, le cuisinier expliquerait la recette face au public en choisissant d'autres éléments langagiers dans l'intérêt de la politesse. De même, il utiliserait probablement un langage moins formel en publiant ses recettes sur les réseaux sociaux.

Une première distinction s'impose entre les registres oraux et les registres écrits (cf. Biber et Egbert 2016). Au sein de ces deux grandes catégories, on marque la différence entre la langue académique orale et écrite, la littérature, le discours scientifique écrit et oral, le langage de la presse ainsi que les interactions informelles et formelles, des cours ou des présentations. Ces distinctions sont loin d'être précises (il suffit d'évoquer que les sections sur la météo, le sport et l'économie au sein du même numéro du même journal diffèrent dans leur usage langagier) mais ces catégories permettent de diviser les larges corpus généraux en sous-ensembles de moindre taille (McEnery 2013).

Tenir compte des registres permet de préciser les résultats obtenus lors de l'analyse de l'élément langagier choisi. Les études liées à l'exploration du contexte social associé à l'usage langagier montrent également que de nombreux mots et expressions possèdent des caractéristiques typiques spécifiques au registre (cf. Biber 2012 ; Biber et al. 1998 ; Biber et Gray 2010 ; Nation 2013 ; Partington 2004 ; Pérez-Paredes et Sánchez-Tornel 2019 ; Scott et Thompson 2000 ; Scott et Tribble 2006 ; Szudarski 2017 ; Weber 2001 ; Yan et al. 2018). Concernant les unités multi-lexicales les plus courantes du langage académique, Biber et Gray (2010) remarquent, par exemple, que certaines d'entre elles se produisent plus fréquemment dans les cours que dans les manuels ou dans la prose académique. De même, les « séquences discontinues récurrentes » varient également en fonction du contexte social dans lequel elles se produisent. Par exemple, certaines structures verbales telles que « must be * to » (l'astérisque représentant un élément variable) sont plus habituelles dans la conversation, alors que des structures nominales (par exemple, « on the * hand ») et des structures basées sur les mots fonctionnels (« the * of this ») sont plus fréquentes dans la prose académique. Au-delà des unités multi-lexicales avec des constructions grammaticales courantes, Biber (1988, 2012), Biber et Egbert (2018) et Biber et Conrad (2009) montrent également une corrélation entre les éléments lexicaux et grammaticaux avec les sous-registres. Ils font ainsi état de variations dans les articles scientifiques (tous appartenant à la prose académique) d'un domaine à l'autre, ainsi qu'entre essais et articles, pour ne nommer que quelques exemples.

Pour une illustration dans le cas du français, prenons, par exemple, le verbe « apprécier ». Ce verbe n'est pas utilisé de la même façon dans des messages privés, dans les revues de spectacles, dans les récits personnels de blogs ou de forums, ou encore dans les textes à caractère officiel. Nous tirerons quelques exemples du corpus « frTenTen2017 » et de notre propre collection de messages privés pour illustrer ce phénomène. Dans les messages privés, « apprécier » est ainsi utilisé majoritairement à la première personne du singulier pour exprimer la reconnaissance du locuteur : « J'apprécie beaucoup votre gentillesse. » « J'apprécie tout particulièrement le professionnalisme

de votre entreprise ». Dans les recensions d'événements, le mot indique que le public prend plaisir à regarder et à écouter le spectacle : « Le public apprécie (la prestation/le film). » « Bel effet que le public apprécie. » Dans les récits personnels, l'appréciation est souvent liée à des moments agréables : « J'apprécie le silence. » Dans un contexte à caractère officiel, « apprécier » devient synonyme de « mesurer » ou « estimer » : « Le contracteur apprécie les risques correspondants » ou « L'échelle de CARS apprécie la sévérité de l'autisme » (v. tableau 22).

Messages privés	**Revues, opinions écrites**	**Textes à caractères officiels**
J'apprécie + MOD (vraiment, beaucoup, particulièrement) + **votre Y** (qualité, geste, proposition)	**X** (public, client, consommateur) **apprécie Y** (spectacle, musique, nourriture, service, calme, silence)	**X** (instance, personne officielle) **apprécie Y** (risque, degré de qqch)

Tableau 22 : L'utilisation du verbe « apprécier » dans les registres différents dans le corpus « frTenTen17 ».

Biber (1988) a développé une taxonomie intitulée « analyse multidimensionnelle » (*multidimensional analysis* ou MDA) qui permet la caractérisation des éléments distinctifs de différents registres. Ses critères reposent sur l'extraction des dimensions latentes de la variation des schémas de co-occurrences dans les manifestations langagières par registre, et ils permettent d'identifier les principaux facteurs linguistiques et extra-linguistiques qui influencent le profil du registre. Biber propose un ensemble de six dimensions pouvant expliquer la variation linguistique des registres les plus importants. Ces dimensions concernent les dimensions suivantes (Nini 2019 : 71-72) :

(1) Discours impliqué versus informationnel : Des scores faibles indiquent un discours dense en informations (prose académique), tandis que les scores élevés indiquent que le texte est affectif et interactionnel (conversations).
(2) Discours narratif versus non narratif : Un score élevé indique la présence forte d'une trame narrative (comme dans le cas des œuvres de fiction).
(3) Discours dépendant versus non dépendant du contexte : Des scores faibles pour cette dimension indiquent une dépendance au contexte (émissions de sport), alors que les scores élevés indiquent une indépendance par rapport au contexte (prose académique).
(4) Expression manifeste de la persuasion : Plus le score est élevé, plus le texte marque explicitement le point de vue de l'auteur (lettres professionnelles).

(5) Information abstraite versus concrète, non abstraite : Plus le score est élevé, plus le degré d'information technique et abstraite est élevé (discours scientifique).
(6) Élaboration linéaire : Les scores élevés indiquent que les informations exprimées sont produites sous certaines contraintes de temps (discours improvisé).

Ces caractéristiques sont exprimées par des éléments langagiers spécifiques. Par exemple, Biber, en analysant des textes anglais pour la dimension 1 (discours impliqué versus informationnel), identifie les phénomènes suivants comme typiques pour le discours impliqué : l'utilisation de la première et la deuxième personnes, la dominance du présent, des contractions (« don't », « she's »), le pronom « it », l'omission fréquente de la conjonction « that », « be » et « do » comme verbes courants, l'utilisation des particules de discours (« well », « you know »), des phrases relatives (plutôt que des participes) et des modificateurs (« quite », « relatively »).

Stefanowitsch (2020) a récemment examiné l'utilisation de « a * of » dans différents registres de l'anglais. En démontrant une variabilité d'usage de ce schéma grammatical simple selon les registres. Il a conclu qu'intégrer l'analyse du registre dans l'analyse des collocations et des colligations permettait de produire des résultats plus précis qu'une étude qui ne tiendrait pas compte de ce facteur.

La recherche de Biber concerne avant tout l'anglais mais ses critères ont été utilisés et adaptés pour l'analyse d'autres langues comme, par exemple, le russe (Katinskaia et Sharoff 2015), pour le portugais du Brésil (Berber-Sardinha et al. 2014) ou l'espagnol (Asención-Delaney 2014). L'analyse multidimensionnelle a été adoptée dans un grand nombre d'autres études concernant divers registres de la langue académique (Biber 2003 ; Gray 2013), les registres Web (Gray et Biber 2011 ; Biber et Egbert 2016) ou encore l'usage oral (Pérez-Paredes et Sánchez-Tornel 2019), pour ne citer que quelques exemples.

2) Quelles implications possibles pour l'enseignement des langues ?

Nous avons vu dans la section B.1 du chapitre 2 que les apprenants des niveaux de compétences linguistiques inférieurs (A1-B1) sont censés maîtriser, à l'oral comme à l'écrit, les interactions informelles simples, ainsi que les situations formelles et semi-formelles du quotidien. Ils doivent être capables de participer à ces situations et de respecter les règles d'usage dont, en premier lieu, les normes de politesse. Puisque chaque contexte social a ses particularités, et ce qui peut être dit ou écrit dans un contexte, peut sembler inhabituel, impoli ou même choquant dans un autre, il est

crucial que les apprenants aient suffisamment d'opportunités d'acquérir les normes liées aux différents contextes sociaux. Les modes d'expression préférés, pertinents par contexte social devraient donc être fréquemment utilisés, révisés et pratiqués dans les manuels et en cours afin que les apprenants puissent les observer et les mettre en pratique.

La taxonomie de Biber peut permettre aux auteurs d'identifier les caractéristiques des textes qu'ils souhaitent inclure dans les manuels et de créer des activités pour l'observation et la pratique de ces caractéristiques. Lors d'un cours de langues, l'enseignant peut inviter les apprenants à noter les éléments langagiers observables dans plusieurs textes du même type. Ce genre d'exercices fait réaliser à l'apprenant qu'il existe des éléments identifiables, propres aux registres et, par conséquent, contribue à une meilleure compréhension du fonctionnement de la langue-cible (cf. McCarten 2010 ; McCarten et McCarthy 2010 ; Pérez-Paredes et Bedmar 2009 ; Pérez-Paredes et Mark 2021)[58].

F) Le langage interactionnel est aussi important que le langage transactionnel.

1) Qu'est-ce que cela signifie ?

Les interactions de la vie quotidienne servent à la fois à échanger des informations factuelles et à socialiser à travers le langage. La recherche concernant les interactions entre les locuteurs n'est pas réservée à la linguistique de corpus, il suffit d'évoquer, par exemple, les travaux de Malinowski (1923) ou de Halliday (1985). L'avantage offert par l'exploration des conversations réelles avec des outils numériques est qu'elle permet d'identifier non seulement leurs principales caractéristiques mais aussi les éléments langagiers concrets, utilisés dans une situation spécifique.

Dans son livre sur les caractéristiques des conversations, Warren (2006 : 94) note que « les conversations […] combinent de manière particulière à la fois les utilisations interactionnelles et transactionnelles du langage ». Brown et Yule (1983 : 1, notre traduction) définissent les deux termes de la manière suivante : « C'est la fonction qui sert à l'expression du contenu que nous décrirons comme transactionnelle, et c'est la fonction impliquée dans l'expression des relations sociales et des attitudes personnelles que nous décrirons comme interactionnelle » (notre traduction). Ils notent, cependant, que les deux fonctions se mélangent dans les énoncés réels : le

[58] Dans la Partie III, nous fournirons quelques exemples pour la construction des sous-ensembles par registre au sein des corpus pédagogiques et nous proposerons quelques activités pour leur exploration.

langage transactionnel contient aussi des éléments du langage interactionnel et vice versa. Comme le souligne McCarthy (2003 : 37) : « Le langage interactionnel semble être un fil continu dans le tissu du discours, fil dont l'analyse avec les outils numériques révèle des régularités et des schémas et fournit un lexique de base pour ces interactions » (notre traduction).

Chaque conversation contient des phases plus ou moins longues dans lesquelles l'échange d'informations n'est pas primordial. Ce sont les phases dans lesquelles les locuteurs échangent des banalités et il ne se passe rien d'important – du moins c'est ce qui semble être le cas vu de l'extérieur. Or, ces phases sont loin d'être inutiles car elles mettent en évidence l'importance de l'établissement et du maintien du lien social (Baraldi et Gavioli 2012 ; Cheepen 2014 : 289 ; McCarten et McCarthy 2010 ; Rühlemann 2007, 2018 ; Warren 2006). En analysant des interactions entre clients et fournisseurs de services, McCarthy (2003 : 37) constate que les phases interactionnelles de la conversation ne se situent pas seulement dans les espaces entre les épisodes transactionnels, mais les facilitent en améliorant leur efficacité. Brown et Levinson (1978) notent qu'une des fonctions des éléments interactionnels dans les conversations est l'établissement d'un terrain d'entente entre les locuteurs qui semblent démontrer des efforts déployés pour maintenir une apparence d'accord, ce qui peut être réalisé, par exemple, par la répétition des mots du ou des partenaires.

L'ouvrage sur le phénomène de la conversation superficielle (*small talk*) édité par Coupland (2014) illustre comment des épisodes, à première vue sans importance, tels que les échanges phatiques, les anecdotes personnelles et les commentaires évaluatifs des locuteurs représentent la partie centrale du tissu du discours et contribuent à l'efficacité de sa progression vers ses objectifs transactionnels. Cela s'applique à la majorité des interactions orales comme les études publiées dans cet ouvrage le démontrent. Holmes (p. 27–32) analyse, par exemple, le rôle de conversations superficielles entre collègues et constate qu'elles assurent la transition d'un thème transactionnel vers un autre, et garantit à la fois des pauses dans la négociation ainsi que l'établissement de la proximité nécessaire pour une collaboration réussie.

L'analyse des conversations réelles permet d'identifier l'ensemble des mots qui remplissent régulièrement certaines fonctions interactionnelles et contribue ainsi à la co-construction du discours. Ces fonctions incluent, entre autres : ouvrir et clôturer la conversation, prendre la parole, reprendre la parole, inviter le partenaire à prendre la parole, signaler son intérêt, demander des clarifications, réagir et évaluer, signaler que l'on écoute (cf. McCarthy 2000, 2002, 2003 ; Warren

2006 ; Wray 2008). La manière dont les locuteurs réalisent ces actions dépend de la langue – mais leur importance semble être observable dans toutes les langues, dans toute interaction humaine (cf. Watzlawik et al. 2011 ; voir aussi Berger et Roloff 2019 pour la présentation de la recherche actuelle sur la communication interpersonnelle).

2) Quelles implications possibles pour l'enseignement des langues ?

De nombreux manuels destinés à des niveaux de compétences linguistiques inférieurs insistent sur l'importance du langage transactionnel dans la communication au détriment du langage interactionnel. Mettre l'emphase sur le langage transactionnel (noms de lieux, de personnes et d'objets, ainsi que des phrases qui servent à réaliser une action (promesse, requête, conseil, etc.)), suggère que sa maitrise représente le noyau des compétences linguistiques, et tout ce qui est « autour de ces éléments », c'est-à-dire le langage interactionnel, mérite moins d'attention. Il s'agit cependant d'un message fallacieux qui ignore l'importance de la fonction sociale du langage et de la forte présence des éléments interactionnels dans toute communication (cf. Boronkai 2011, De Fornel et Verdier 2018 ; McCarthy 1999, 2000, 2002, 2003 ; Rühlemann 2007, 2018 ; Németh et al. 2018 ; McCarthy et McCarten 2019 ; Schirm 2014 ; Warren 2006). Toute conversation contient un grand nombre de dispositifs dont la fonction est d'organiser la conversation et de suivre sa progression. Ceux qui ne maitrisent pas ces éléments, ne possèdent pas tous les moyens langagiers pour participer au discours et « risquent de devenir des participants de second rang » (McCarthy 1999 : 243, notre traduction).

Comme les apprenants doivent être capables de participer aux interactions du quotidien à partir du niveau A1, la présentation et la pratique du langage interactionnel doivent former une partie essentielle du curriculum. Son intégration dans le manuel et dans le cours de langues peut contribuer au développement des routines interactionnelles essentielles pour la vie quotidienne et l'établissement de relations avec les locuteurs natifs.

Ce chapitre a résumé les résultats les plus importants de la linguistique de corpus susceptibles de contribuer à un enseignement de langues plus efficace. Les résultats soulignent l'importance de l'interrelation et de la co-sélection entre différents aspects de la langue. Ainsi, ils impliquent que les mots ne doivent pas être étudiés séparément mais qu'ils doivent, au contraire, être présentés en contexte car ils réalisent des sens non ambigus au sein de leur environnement textuel. En étudiant cet environnement, plusieurs types de co-sélection émergent, permettant l'établissement de

schémas. Ces schémas, dans leur sens plus limité, concernent les composantes lexicales, grammaticales, sémantiques et pragmatiques au niveau des unités multi-lexicales et, dans un sens plus large, incluent l'interrelation de différents éléments au niveau textuel. Les études montrent aussi le rôle crucial du contexte social et du langage interactionnel pour l'usage langagier. Ces résultats ont enfin permis des tentatives de théorisation comme l'Amorçage lexical, le Principe de l'idiomaticité et la Grammaire des schémas, décrites de façon succincte dans ce chapitre.

Ces observations ouvrent la voie vers de nouveaux modes de présentation de la langue dans le cadre de l'enseignement, essentiellement dans les domaines du lexique et de la grammaire. Les chapitres 5 et 6 présenteront l'analyse d'une sélection de quelques ouvrages pédagogiques implémentant certains de ces résultats.

Chapitre 5 : Intégrer les résultats de la linguistique de corpus aux matériels pédagogiques : les grammaires

L'élément et son environnement ne sont finalement pas séparables, ou en tout cas pas par les techniques actuelles.
(Sinclair 2004 : 19, notre traduction)

Ce chapitre présentera les différentes possibilités d'intégrer les résultats pertinents de la recherche linguistique dans les grammaires pédagogiques. Nous exposerons les considérations générales pour la création de tels matériels et identifierons leurs principales caractéristiques en utilisant quelques ouvrages sélectionnés comme références. Nous nous concentrerons sur deux ouvrages, provenant des pays anglo-saxons, pays dans lesquels nous constatons une volonté forte d'intégrer les avancées de la linguistique de corpus dans les ouvrages pédagogiques.

Dans la présentation, nous mettons l'accent sur les aspects suivants :

- Quels sont les résultats de la recherche linguistique intégrés dans la présentation de la langue dans les grammaires choisies ?
- Comment l'interconnexion du lexique et de la grammaire est-elle présentée ?
- Quels sont les types d'exercices proposés ? Trouve-t-on des exercices novateurs ?
- Peut-on identifier une méthodologie cohérente mise en place de façon systématique dans l'ouvrage ?

A) Les grammaires pédagogiques

1) Aperçu général

Germain et Seguin (1995 : 85) définissent les grammaires pédagogiques comme des ouvrages qui « décrivent la compétence grammaticale d'un certain usage de la langue en vue d'en faciliter l'apprentissage ». Que ces descriptions soient fondées, au moins en partie, sur l'analyse de corpus, présente de nombreux avantages. Conrad (2000 : 549) soutient que trois changements occasionnés par des études de grammaire basées sur des corpus ont le potentiel de révolutionner l'enseignement de la grammaire :

- Les descriptions monolithiques de la grammaire pourraient être remplacées par des descriptions spécifiques au registre.

- Cette démarche permettrait une intégration plus forte de l'enseignement de la grammaire dans celui du vocabulaire.
- Le déplacement de l'accent de la précision structurelle aux conditions d'utilisation appropriées des constructions grammaticales.

La validité de cette approche semble tomber sous le sens car, comme le constatent Biber et Reppen (2002), les ouvrages pédagogiques qui n'incorporent pas les résultats de l'analyse de corpus dans les phénomènes grammaticaux, donnent aux apprenants des fausses informations sur l'usage langagier[59].

Parmi les grammaires pédagogiques actuelles, nous avons sélectionné la « Grammaire de schéma » (Pattern Grammar) par Hunston et Francis (2000) et la « Grammaire réelle » (Real Grammar) par Conrad (2009) afin de présenter deux options alternatives pour la création de grammaires fondées sur les résultats de l'analyse de corpus. Le premier ouvrage est une grammaire de référence qui systématise la grammaire de manière inédite en soulignant son interrelation avec le lexique. Le texte est complété par des fiches pour l'enseignement de l'anglais. Le deuxième ouvrage est une grammaire pratique qui contient la présentation des phénomènes choisis ainsi que des exercices incitant les apprenants à observer, systématiser et pratiquer de façon autonome ou pendant le cours de langues.

2) La « Grammaire des schémas » de Hunston et Francis

2.1) Le concept

Un des premiers exemples d'une grammaire reposant sur l'analyse de corpus susceptible de trouver une application dans le cadre pédagogique est la « Grammaire des schémas » de Hunston et Francis, publiée en deux tomes aux éditions Harper et Collins en 1996. Cette grammaire est basée sur l'analyse de corpus de la « Bank of English » qui est une collection de textes en anglais, principalement d'origine britannique. Ce corpus représentatif est formé d'un sous-ensemble de 650 millions de tokens tirés du corpus COBUILD (4,5 milliards de mots). La majorité des textes sont tirés de l'anglais écrit, collectés sur des sites Web, des journaux, des magazines et des livres, mais le corpus comprend aussi une grande base de données orales provenant des émissions de radio et de télévision et des conversations informelles (Collins Corpus 2021). Le premier objectif

[59] Les auteurs prennent, comme exemple, la présentation du présent progressif et du présent simple dans quelques grammaires pratiques et constatent que le présent progressif est clairement surreprésenté dans ces ouvrages bien que le présent simple prédomine largement dans l'usage langagier.

de la « Grammaire des schémas » est de décrire l'environnement syntaxique des éléments lexicaux et de fournir un cadre qui permet de capturer la relation étroite entre les mots, les unités multi-lexicales et les schémas grammaticaux. Le premier volume de cette grammaire est consacré aux schémas associés aux verbes et le deuxième à ceux des noms et des adjectifs ; les deux tomes sont accessibles sur le site de la maison d'édition pour une consultation en ligne.

Les principes fondamentaux de la « Grammaire des schémas » sont les suivants :

- Les mots polysémiques possèdent plusieurs schémas grammaticaux correspondant à leurs différents sens (v. aussi chapitre 4 sur l'interconnexion entre lexique et grammaire).
- Les schémas identifiés ne s'appliquent pas à un seul verbe. Il est possible de définir des groupes de verbes avec des sens similaires selon les mêmes schémas.
- Le sens émerge de la combinaison du mot et du schéma. Le schéma renferme une partie du sens et contribue à sa réalisation.

Hunston et Francis identifient une cinquantaine de schémas pour les verbes, 80 pour les noms et une cinquantaine pour les adjectifs. Le tableau 23 illustre l'approche adoptée dans l'ouvrage sur le modèle des verbes suivis de la structure « to + infinitif ».

8 V to-inf

The verb is followed by a to-infinitive.

This pattern has three structures:

- Structure I: Verbs in phase
  The number of victims continues to rise.
- Structure II: Verb with Object
  He expects to fly to Beijing soon.
- Structure III: Verb with Adjunct
  He hurried to catch up with his friend.

Structure I: Verbs in phase

V to-inf

	Verb group	to-infinitive	
Subject	**Verb**		**Completive**
The arrangements	appeared	to be	satisfactory.
Prison officers	continued	to patrol	the grounds.
He	refused	to comment.	

Phrasal verbs: V P to-inf

	Verb group	Particle	to-infinitive	
Subject	**Verb**			**Completive**
Dr Carey	went	on	to spell out	his views.
These theories	may turn	out	to contain	elements of truth.

Verbs with this structure belong to the following meaning groups:

I.1 The `begin' group	I.4 The `manage' group	I.7 The `hasten' gro88
I.2 The `appear' group	I.5 The `fail' group	I.8 The `chance' group
I.3 The `try' group	I.6 The `regret to say' group	I.9 The `tend' group

I.1 The `begin' group

These verbs are concerned with starting, stopping, or continuing an action.

*Edgar **began to laugh** again.*
*The social activities patients enjoyed before they became sick **will continue to be enjoyed** during the course of their illness.*
*Phil **went on to enjoy** more success at cricket than he had at football.*
*He treated us okay but I never **got to like** him.*

The verb *come on* is always used with verbs indicating the weather, with the Subject *it*.

*It **was coming on to rain** when finally Mac's lorry arrived.*

begin cease come commence continue get grow proceed start
come on go on settle down

Tableau 23 : Les catégories de verbes suivis par la structure « to + infinitif » dans la « Grammaire des schémas » (Hunston et Francis 2000).

La présentation commence par la liste des trois grandes catégories de verbes susceptibles d'être suivis du schéma « to + infinitif » : (1) les verbes exprimant une action graduelle ; (2) les verbes avec un complément d'objet direct ; (3) les verbes avec des compléments optionnels. Ces grandes catégories sont par la suite subdivisées en plusieurs groupes selon les caractéristiques sémantiques des verbes concernés. Le premier schéma est, par exemple, divisé en neuf sous-ensembles dont le premier contient les verbes incluant l'aspect de la temporalité, signalant le début, la fin ou la continuité d'une action (« begin group »), au deuxième groupe appartiennent les verbes indiquant une impression (« appear group »), au troisième, les verbes exprimant l'idée que quelqu'un essaie de faire quelque chose pour produire un résultat (« try group ») et ainsi de suite. La définition de la sémantique des verbes est complétée par au moins cinq exemples authentiques illustrant leur signification et leur environnement textuel. Quand une phrase seule ne suffit pas pour désambiguïser le sens, une autre phrase est rajoutée : « She made to move past him. He placed himself in her way. » Dans les cases grises à la fin de la description de chaque groupe, les verbes les plus importants y appartenant sont listés. Ce mode de présentation est maintenu dans tout l'ouvrage, démontrant la validité de cette approche pour les différentes parties du discours.

Sur le site Web de la maison d'édition, nous trouvons aussi des vidéos explicatives. Elles apportent le fondement théorique en utilisant des exemples compréhensibles et offrent une mise en contexte de cette approche en soulignant son intérêt pour l'analyse linguistique ainsi que pour l'apprentissage des langues. En outre, la page Web propose des plans de leçons et des fiches d'activités pour l'enseignement de l'anglais. En fonction des différents niveaux de compétences linguistiques et de l'expertise des apprenants dans l'analyse de corpus (essentiellement leur capacité d'analyser les lignes de concordance), trois étapes sont proposées que nous présenterons par la suite.

2.2) La pratique pédagogique à travers une séquence d'activités

Toutes les fiches proposées suivent la même approche pédagogique consistant en trois phases : (1) présentation ; (2) pratique et (3) production. D'abord, le schéma concerné est présenté en contexte ; cette étape est suivie d'une phase de mise en pratique qui familiarise les apprenants avec la façon dont le schéma est construit et utilisé. Les unités finissent par une ou plusieurs tâches de production permettant aux apprenants de créer leurs propres énoncés en intégrant le schéma.

Le matériel didactique pour le niveau débutant invite l'apprenant à travailler sur les mots polysémiques qui possèdent plusieurs schémas différents en fonction de leur sens. Comme exemple concret de ce processus, nous présenterons les étapes du travail avec le verbe « make » (faire). La fiche pour l'enseignant précise que dans le « British National Corpus », « make » est le dixième verbe le plus courant et le quatrième verbe d'action le plus courant après « say » (dire), « go » (aller) et « get » (recevoir, obtenir) (Collins Resources 2021). La leçon est dédiée à deux schémas élémentaires avec « make », le schéma 1.1 correspondant avec le schéma « V + N » (Francis et al. 1996 : 28) et le schéma 1.2 à « V + N + infinitif » (p. 297).

Avant d'introduire les schémas, les deux significations principales de « make » sont présentées sous forme de deux questions : (1) Quel genre de personnes vous font faire des choses ? (2) Que pouvez-vous faire ? (What do other people make you to do ? What can you make ?) Au cours de l'exercice 1, les apprenants classent les phrases listées en deux groupes selon leur sens (Meaning 1 et Meaning 2). Cet exercice est donc, avant tout, une activité d'observation et de systématisation (tableau 24).

Exercise One:
What is the meaning of ***make*** in each of these sentences?
Put a tick in the correct column. The first one has been done for you.

		meaning 1	meaning 2
1.	On a normal day we ***make*** around 2,000 loaves of bread an hour.	✓	
2.	The essay was full of mistakes so my teacher ***made*** me write the whole thing out again.		
3.	The boss ***made*** her work for the first ten months without a single day's holiday.		

Tableau 24 : Extrait de l'exercice 1 sur le site Web « Collins Resources » complémentant la « Grammaire des schémas », introduisant deux schémas avec « make » (niveau débutant).

L'exercice 2 explicite les schémas et se compose de deux parties. Les apprenants sont tout d'abord invités à identifier quel schéma correspond à quel sens et doivent dans un second temps compléter le tableau avec d'autres phrases tirées de l'exercice 1 (tableau 25).

Exercise Two:
Look at these patterns and decide which pattern reflects which meaning.

	pronoun/ noun group	verb	noun group	meaning 1	meaning 2
Pattern 1.1	we	make	2000 loaves a day	☐	☐
	she	*made*	*tea in the morning*		

	pronoun/ noun group	verb	pronoun	verb	noun group	meaning 1	meaning 2
Pattern 1.2	my teacher	made	me	write	the essay again	☐	☐

Complete each table by writing in parts of another sentence from exercise one.
We have done one for you as an example (pattern 1.1).

Tableau 25 : Extrait de l'exercice 2 sur le site Web « Collins Resources » complémentant la « Grammaire des schémas », pratiquant deux schémas avec le verbe « make » (niveau débutant).

La catégorisation des phrases connues rencontrées dans les exercices précédents permet à l'apprenant de prendre conscience des schémas grammaticaux, c'est-à-dire des régularités structurelles répétées, associées à chacun des deux sens. Compléter les tableaux en suivant l'exemple, permet l'apprenant d'approfondir ses connaissances sur le sens et la structure des schémas.

Les exercices 3 et 4 ciblent l'interconnexion du sens, du vocabulaire et des schémas grammaticaux. L'exercice 3 invite les apprenants à trouver la fin appropriée de phrases exprimant le sens 1 ou le sens 2. La nécessité d'une analyse grammaticale et lexicale fait de cet exercice une activité complexe qui permet de consolider les deux schémas et de renforcer l'idée qu'ils jouent un rôle significatif dans la réalisation du sens choisi du verbe « make ». L'exercice présente différentes formes du verbe (verbe au présent et au passé, gérondif) qui peuvent être associées à l'un ou l'autre des deux schémas (tableau 26).

Unit 1 Overview:
Match the first half of the sentence from the left with the second half from the right.
The first one has been done for you.

1	Toyota and Nissan began *making*		a lot of mistakes. We know better nowadays.
2	While he was working for us we *made*		you pay for any repairs that are needed.
3	We were young then and we *made*		him work hard which prepared him for later life.

Tableau 26 : Exercice 3 sur le site Web « Collins Resources » complémentant la « Grammaire des schémas ». Observation et systématisation des schémas du verbe « make » (niveau débutant).

Lors de l'exercice 4, les apprenants identifient le sens de « make » dans les phrases et mettent les mots dans le bon ordre (tableau 27).

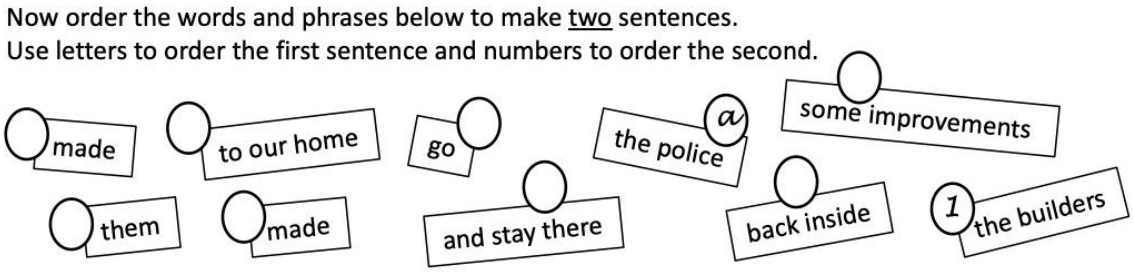


Tableau 27 : Exercice 4 sur le site Web « Collins Resources » complémentant la « Grammaire des schémas ». Créer des phrases avec l'élément étudié (niveau débutant).

Dans cet exercice, l'apprenant pratique les schémas grammaticaux avec de nouvelles phrases. C'est lui qui doit décider du schéma correct et de l'ordre des mots à partir du tableau qu'il a rempli dans l'exercice 2.

Dans l'exercice 5, l'apprenant produit ses propres phrases inspirées des images, tout en suivant les schémas déjà présentés (tableau 28).

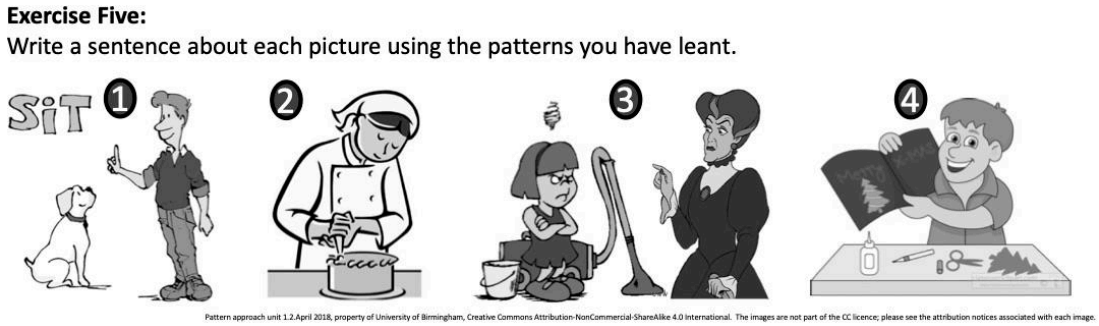


Tableau 28 : Exercice 5 sur le site Web « Collins Resources » complémentant la « Grammaire des schémas ». Produire des phrases avec les schémas observés (niveau débutant).

L'apport visuel donne un cadre aux énoncés à produire, tout en laissant place à la créativité de l'apprenant qui peut choisir, dans les limites du schéma, les mots qu'il souhaite utiliser. La fiche pour l'enseignant recommande de poursuivre l'activité avec des questions personnalisées connexes telles que : « Est-ce que vous avez fait quelque chose de spécial pour l'anniversaire d'un de vos proches ? » Cette dernière étape est bénéfique en ce qu'elle montre les possibilités de l'adaptation du schéma aux besoins de l'apprenant.

Tout au long de cette série d'exercices, nous pouvons observer l'approfondissement graduel du concept de « schéma » et des sens concrets des deux schémas proposés ainsi que l'élargissement des connaissances. L'utilité des schémas pour la production langagière devient également claire. Cette méthodologie cyclique, reprenant les connaissances acquises, les approfondissant et les élargissant, caractérise toutes les fiches proposées.

Le matériel didactique pour le niveau intermédiaire adopte *une approche plus complexe en intégrant l'observation et la pratique des schémas dans le travail avec des textes authentiques et semi-authentiques*. Ces textes peuvent être utilisés dans le cadre d'une leçon sur la compréhension écrite (activité familière aux apprenants et aux enseignants) mais, au-delà, chaque texte sert aussi d'apport linguistique pour identifier un schéma spécifique. Cette fois-ci, les points de départ ne sont pas les mots individuels (comme « make » dans la série d'activités au niveau débutant) mais les schémas grammaticaux. Nous utiliserons le premier texte comme exemple pour illustrer le procédé utilisé. Ce texte traite de bibliothèques publiques et contient de nombreux exemples d'utilisation de la structure possessive « N of N ». Les activités connexes comprennent quatre tâches : (1) la discussion, (2) la lecture, (3) l'identification du schéma et (4) la production du schéma. Les deux premiers exercices sont des exercices classiques (introduction du sujet à travers une discussion courte et lecture du texte), le troisième exercice reprend l'idée de l'exercice 2 présenté ci-dessus (compléter le tableau des schémas).

La production du schéma dans l'exercice suivant mérite plus d'attention car il s'agit d'un exercice novateur. Il s'appuie sur des données extraites du corpus de la « Bank of English » et présente les noms les plus saillants occupant la position du premier nom dans la structure « N of N ». Les apprenants sont invités à trouver des catégories de noms appropriés ainsi que des noms concrets qui peuvent être placés dans la case « unité nominale » (noun phrase) (tableau 29).

Pattern production task

These are common nouns that fit this pattern. Fill in the final section with a suitable noun phrase. The first one has been done for you

	the	NOUN	of	noun phrase
Pattern 1.1 N of n	the	custom	of	people
	-	role	-	
	-	feeling	-	
	-	dream	-	

Tableau 29 : Établir des schémas grammaticaux au niveau intermédiaire. Site Web « Collins Resources » complémentant la « Grammaire des schémas ».

Cette activité est plus ouverte qu'au niveau débutant : tandis que les débutants sont invités à systématiser des phrases listées sur la fiche, les apprenants au niveau intermédiaire sont encouragés à créer leurs propres énoncés correspondant à ce schéma. Pour cela, ils peuvent chercher des phrases adéquates dans un corpus ou s'appuyer sur leurs propres connaissances.

L'exercice suivant élargit l'environnement textuel du schéma et travaille au niveau de la phrase. Les exemples fournis proviennent toujours de la « Bank of English », l'apprenant peut les compléter avec des phrases construites par lui-même (tableau 30).

Example sentences:
Here are some example sentences from a reference corpus. Write your own sentence using the same **first noun**. We have provided an example for the first sentence.

the	NOUN	of	noun phrase	predicate
The	custom	of	using chopsticks	is growing in the West.
The	custom	of	Thanksgiving	is important to me.
The	custom	of		

Tableau 30 : Observer des phrases authentiques et créer ses propres phrases, site Web « Collins Resources » complémentant la « Grammaire des schémas », (niveau intermédiaire).

La fiche pédagogique conseille aux professeurs d'attirer l'attention des apprenants sur le fait que ce schéma est particulièrement utile pour la rédaction d'un essai car la nominalisation est une caractéristique de ce type de textes.

Le matériel didactique pour le niveau avancé offre des possibilités d'analyse de lignes de concordance et plusieurs exercices permettant l'observation des schémas. Les activités s'organisent autour de trois mots, l'adjectif « afraid », le verbe « manage » et le nom « time ». Les apprenants sont invités à explorer les différents schémas appartenant à ces mots et les sens correspondants. Les apprenants travaillent en pratique avec 50 lignes de concordance présélectionnées et regroupées par catégorie, leur tâche étant l'identification des caractéristiques communes des phrases appartenant au même groupe (tableau 31) – activité que Johns qualifie d'« apprentissage en douceur fondé sur corpus » (soft data-driven learning) (1991). Les questions suivantes sont proposées pour aider à l'exploration des schémas :

- Combien de sens a le mot « afraid » ?
- Identifiez les schémas grammaticaux avec cet adjectif.
- Les sens différents du mot s'associent-ils à des schémas grammaticaux différents ?

Here are 50 concordance lines taken from the Bank of English for the adjective ***afraid***.

Worksheet 1A

1. How many different meanings does *afraid* have?
2. Identify some of the most frequent grammatical patterns associated with *afraid*.
3. Are these different meanings associated with particular patterns?

```
1.       here, Lieutenant Andrews?" I'm afraid I can't make any comment at this
2.    Laughs) Literally. Gwendolen: I'm afraid I don't have experience in anything
3.       here, Lieutenant Andrews?" I'm afraid I can't make any comment at this
4.        can't tell you.  3rd MAN: I'm afraid I can't help you- except with
5.      light to encourage others.  I'm afraid I had to laugh when Mr Blunkett
6.    over. So on to the World Cup. I'm afraid I'm not all that convinced Colin

7   get him to change his behavior, I'm afraid we're going to have to show that we
8  indicating a fallen elm nearby, `I'm afraid we've just cut it down!" Oh, never
9      her an amicable grin. `Well, I'm afraid you are out of luck. This one is
10     living on his royalties. But I'm afraid Billy's got it wrong again. He
11     fan ended on an upbeat note. I'm afraid England were out for a hundred and
12    indeed. Belgium,deja vu? Yes, I'm afraid, folks, it's beginning to look that
```

Tableau 31 : Extrait de l'exercice 4 contenant 50 lignes de concordance catégorisées par schémas sur le site Web « Collins Resources » complémentant la « Grammaire des schémas »,. L'apprenant observe et identifie les schémas et sens différents (niveau avancé).

Le tableau 32 présente les trois schémas émergeant des phrases présentées dans l'exercice 4 (v. tableau 31) :

Schéma 1 : SUBJECT (sujet) + « be » + « afraid » + NOUN and VERB (nom et verbe)
« I'm afraid I can't make any comment » (Je suis désolé mais je ne peux faire aucun commentaire), « I'm afraid I don't have experience in anything » (Je suis désolé mais je n'ai aucune expérience.)
Schéma 2 : SUBJECT (sujet) + « be » + « afraid » + « that »
« I was afraid that I would have problems » (Je craignais d'avoir des problèmes), « She was afraid that it's not going to work. » (Elle avait peur que ça ne fonctionne pas.)
Schéma 3 : SUBJECT (sujet) + be + afraid + of -ing
« He was so afraid of hurting you » (Il avait tellement peur de te blesser), « He's afraid of feeling foolish in front of the crowd » (Il a peur de se sentir stupide face à la foule.)

Tableau 32 : Trois schémas basés sur les lignes de concordance catégorisées par schémas (v. tableau 31).

Les trois schémas se regroupent autour de deux sens principaux : les phrases dans le premier groupe appartiennent aux groupes « afraid = sorry » (X est désolé), les phrases dans les deux autres groupes sont plus proches du sens « anxious » (avoir peur, appréhender).

Une fois ces catégories établies, les apprenants étudient des phrases avec l'un des schémas observés dans le cas de l'adjectif « afraid », notamment le schéma de « X be + ADJ + that ». Ils sont invités à catégoriser 50 lignes de concordance selon le sens de l'adjectif, guidés dans leurs explorations par les questions suivantes : « Quels genres d'adjectifs émergent dans cette structure ? » « Y a-t-il une seule ou plusieurs significations associées à ce schéma ? » (tableau 33).

Let's take one of the patterns associated with the adjective *afraid* (Book 2, Pattern 84, page 400*)

Sheet 2A

N (personal pronoun)	v-link (verb to be)	adj (predicative adjective)	that (that clause)
I	*am*	***afraid***	*that*

Here are 50 concordance lines taken from the Bank of English which follow this pattern.

1. Put the adjectives together into 'meaning groups' (i.e. groups of adjectives with related meanings).
2. What are the main groups? What types of adjective are associated with this pattern?
3. Is this particular pattern associated with one or more meanings?

1 prior to--to that, but I was **afraid** that I would have problems, so I went to
2 afraid of the strike. She was **afraid** that it wasn't going to work.
3 of Patrick's shows. But he was **afraid** that she would be offended by close-to
4 owner. As it matures I am **afraid** that it is likely to get worse.
5 He told journalists: `I am **afraid** that the present British government is
6 talk to me." However, she is **adamant** that polygamy is not a good thing. `The
7 the England kit bag. He was **adamant** that he could never find satisfactory
8 substitute, for his pain. He is **angry** that the treatment was abandoned as too
9 after a restless night. He was **angry** that promised medical supplies had not
10 of the Treaty, but he was **anxious** that any new Hispano-Portuguese agreement
11 selling them. He was **anxious** that the birds should be provided with
12 questioned by police I am **ashamed** that I didn't know my rights. I didn't

Tableau 33 : Observer l'usage des adjectifs suivis de la conjonction « that » dans la « Bank of English ». Exercice sur le site « Collins Resources » complémentant la « Grammaire des schémas » (niveau avancé).

Des 50 lignes de concordance se dégagent les groupes suivants :

- « angry group » : indignant, sad, angry, livid, unhappy, heartbroken (groupe d'adjectifs exprimant la colère : indigné, triste, en colère, livide, malheureux, le cœur brisé)
- « anxious group » : afraid, anxious, fearful, ashamed (groupe d'adjectifs exprimant l'appréhension : avoir peur, être anxieux, être craintif, honteux)
- « aware group » : aware, conscious, unaware (groupe d'adjectifs exprimant que quelqu'un est conscient/inconscient de quelque chose)
- « sure group » : sure, certain, adamant, confident, positive (groupe d'adjectifs exprimant que quelqu'un est sûr de quelque chose : sûr, certain, catégorique, confiant, assuré)
- « thankful group » : fortunate, thankful, grateful, glad (groupe d'adjectifs exprimant le sentiment de reconnaissance : chanceux, reconnaissant, redevable, content)
- « hopeful group » : hopeful, optimistic (groupe d'adjectifs exprimant l'espoir: qui espère, optimiste)

À la fin de la série d'activités, les apprenants comparent les résultats de leurs observations.

2.3) Analyse succincte du matériel

La « Grammaire des schémas » propose une approche lexico-grammaticale qui met l'accent sur l'interrelation entre le sens du mot, les structures grammaticales et son environnement lexical. Initialement conçue comme une grammaire de référence, cet ouvrage peut être également très utile

pour l'enseignement de l'anglais. La présentation claire et concise des grandes catégories structurelles, suivie de celles des sous-groupes avec des exemples authentiques tirés de la « Bank of English » et de la liste des verbes les plus fréquemment utilisés avec le schéma en question, fournissent une description complète. Les vidéos explicatives viennent à l'appui du processus d'apprentissage et, en particulier, des compétences d'observation et d'analyse de l'apprenant, en donnant une image plus complète du fonctionnement de la grammaire des schémas. Bien évidemment, la « Grammaire des schémas » ne peut pas traiter de tous les phénomènes grammaticaux. Cet ouvrage doit donc être complété par une grammaire plus traditionnelle incluant des informations sur la morphologie, c'est-à-dire sur la manière dont les différents phénomènes grammaticaux sont formés.

Complétée par des fiches pédagogiques, l'utilisation de cet ouvrage peut faire partie des cours d'anglais. Les activités proposées introduisent et approfondissent graduellement une méthodologie connectant la grammaire, le vocabulaire et la dimension sémantique. L'exploration des schémas en plusieurs étapes, à trois niveaux différents, témoigne d'une approche didactique réfléchie et aide l'apprenant à développer ses compétences analytiques. Les exercices proposés pour les niveaux débutant et intermédiaire pratiquent l'utilisation des schémas avec des activités plutôt classiques (remplir les trous, compléter des tableaux, etc.) que les enseignants et les apprenants connaissent déjà. Cela évite une étape supplémentaire pendant laquelle les utilisateurs doivent se familiariser non seulement avec de nouveaux concepts mais aussi avec de nouveaux types d'exercices.

Un avantage majeur du travail avec l'ouvrage « Grammaire des schémas » est que *les schémas ont été identifiés par les linguistes, limitant ainsi la tâche de l'enseignant à la sélection d'exemples appropriés dans le corpus.* Cette démarche nous apparaît aussi pragmatique que réaliste car nous observons que les enseignants novices aux méthodes d'analyse de corpus sont en général peu sûrs d'eux-mêmes quand il s'agit d'interpréter les données linguistiques et de généraliser à la base d'exemples (Szita 2022a à paraître). L'enseignant peut par la suite décider de complémenter la liste des exemples présentés dans l'ouvrage par des phrases adaptées à partir d'un corpus choisi et développer, par ce biais, des « mini-corpus » pédagogiques pour l'enseignement. L'exposition de l'apprenant à un grand nombre de lignes de concordance respectant son niveau, lui permet de travailler de manière efficace sur le vocabulaire dans le contexte des schémas grammaticaux.

Un inconvénient majeur de cet ouvrage dans le contexte pédagogique est que les activités proposées ne sont que des exemples. Bien que les résultats de l'analyse (les schémas répertoriés)

soient déjà fournis dans le livre de grammaire et qu'il existe des corpus dans lesquels l'enseignant peut trouver des exemples, l'investissement exigé par cette approche reste significatif : l'enseignant doit étudier le schéma en question, identifier le corpus, chercher des phrases appropriées et les adapter au niveau de ses étudiants. S'il décide de compléter la liste des exemples par une série d'activités, sa charge de travail augmente encore davantage, ajoutant aux obstacles susceptibles de freiner l'utilisation du matériel.

Une autre question se pose quant aux exercices invitant l'apprenant à créer ses propres phrases en utilisant un schéma donné. Ces exercices ne proposent que peu d'éléments lexicaux pouvant être intégrés dans son récit. Or, « improviser » sans apport linguistique peut poser problème car la charge cognitive liée à ce genre d'exercices est importante : l'apprenant doit utiliser un schéma bien défini (dans un cadre de production guidé au niveau de grammaire), mais il a la liberté de créer ses propres phrases (dans un manque absolu de guide au niveau du vocabulaire). Le laisser s'appuyer sur ses connaissances lexicales et espérer qu'il construira des phrases à la fois correctes et à caractère naturel est une démarche peu réaliste, car ses phrases ne rempliront probablement pas simultanément les critères de l'exactitude *et* de l'usage naturel de la langue-cible. En rajoutant à la description de la tâche la phrase « Vous pouvez aussi chercher des exemples dans le corpus et les adapter » pourrait encourager l'apprenant à consulter le corpus pour en intégrer des éléments dans son récit, selon ses besoins.

Ce point nous mène au dernier constat. Pour que les apprenants puissent explorer les schémas dans les corpus par eux-mêmes, sans que l'enseignant soit obligé de fournir une présélection d'exemples pour chaque élément étudié, il faut avoir accès à des corpus appropriés pour les niveaux de compétences linguistiques inférieurs. Consulter la « Bank of English » au niveau A2 ou B1 constituerait une expérience peu gratifiante pour l'apprenant qui constaterait, avant tout, une chose : les limites de ses compétences. *Pour l'utilisation efficace de la « Grammaire des schémas » dans l'enseignement, il serait donc nécessaire de construire des corpus pédagogiques qui fournissent des exemples accessibles et permettent aux apprenants de travailler de façon autonome.*

3) La « Grammaire réelle » (Real Grammar) de Conrad

3.1) Le concept

Une grammaire pratique directement utilisable dans l'enseignement et l'apprentissage des langues, destinée aux apprenants des niveaux A2–B2 est la « Grammaire réelle » (Real Grammar) de Conrad. Cette grammaire pédagogique est parue en 2009 chez Pearson. La préface résume les éléments suivants innovants de l'ouvrage (Conrad 2009 : vi) :

- L'ouvrage est fondé sur corpus et le fait de l'authenticité est mis en exergue. Les auteurs déclarent que les informations présentées reflètent « les manières dont les gens écrivent et parlent véritablement ».
- Au lieu d'opérer avec les termes d'« usage correct » et d'« usage incorrect », cette grammaire met l'accent sur ce qui est typique et fréquent. Les auteurs soulignent que certaines structures bien que tout à fait correctes sont pourtant rarement utilisées et, pour cette raison, elles n'ont pas été incluses dans l'ouvrage. Au-delà de cette présélection, la présentation des aspects grammaticaux inclut systématiquement des informations de fréquence liées à l'usage.
- La sélection des phénomènes traités est également basée sur le corpus. Les auteurs présentent 50 aspects de la grammaire anglaise jugés particulièrement pertinents pour les niveaux A2–B2, sur la base de l'étude de corpus. Ces domaines comprennent, entre autres, le passif, les conditionnels, l'utilisation de clauses relatives et le traitement de l'aspectualité.
- La présentation de la grammaire fait la distinction entre la grammaire du langage écrit et celle du langage parlé. À l'intérieur de ces deux grandes catégories, les auteurs séparent l'usage langagier dans le cadre académique, dans la littérature (textes de fiction) et dans les conversations du quotidien.
- L'interconnexion de la grammaire et du lexique est abordée systématiquement : les phénomènes grammaticaux sont présentés avec des choix typiques au niveau du vocabulaire, dans des textes.
- Les exemples sont authentiques et ont été tirés du corpus de « Longman Corpus Network », ensemble de corpus comparable à la « Bank of English » présenté dans la section A.2.1.
- La pratique de la grammaire intègre des exercices développant les compétences de compréhension orale et écrite et de production orale et écrite.

Les activités se divisent en trois grandes parties :

- Le premier groupe d'activités invite l'apprenant à observer un phénomène donné.
- Les activités d'analyse permettent à l'apprenant d'identifier des structures et de découvrir par lui-même des schémas.
- Le dernier groupe d'activités offre des possibilités de pratique et d'approfondissement de l'aspect étudié par la production orale et écrite.

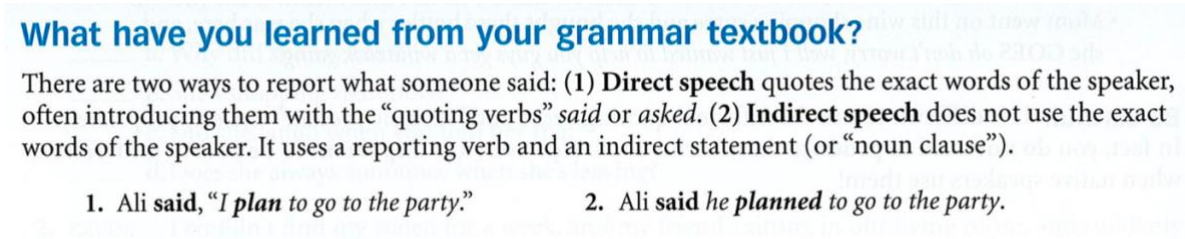
Dans la note à l'enseignant (Conrad 2009 : ix), l'auteure explique les modifications apportées à la matière linguistique dans le but de la rendre accessible aux apprenants. Par exemple, certains mots difficiles ont été remplacés par des mots plus simples et des phrases longues et/ou complexes ont été modifiées en supprimant certains éléments optionnels tels que les adverbiaux facultatifs. Dans les exemples de conversations, les faux départs, les hésitations et les éléments qui pourraient gêner la compréhension de la logique de la phrase, ont été supprimés.

À la base de cette description, nous pouvons constater que *l'ouvrage tient compte de la majorité de résultats pertinents de la linguistique de corpus exposés au chapitre 4. À savoir : l'importance de la fréquence statistique des occurrences, celle du registre et de l'environnement textuel, l'interconnexion de la grammaire et du vocabulaire, et le caractère réel des exemples*. Les simplifications apportées à la langue présentée servent l'accessibilité au niveau linguistique de l'apprenant.

3.2) Le contenu : présentation d'une séquence d'activités

Dans les pages suivantes, nous examinerons un chapitre en détail pour observer la manière dont les principes fondamentaux exposés ci-avant sont transposés dans la présentation des différents aspects grammaticaux ainsi que dans les exercices. Nous avons choisi l'unité 37 comme exemple dévolue aux différentes façons de rapporter le discours d'une autre personne.

La première partie de la présentation consiste en un résumé des explications relatives à la citation directe dans les conversations comme on peut le lire dans les grammaires traditionnelles (tableau 34) :



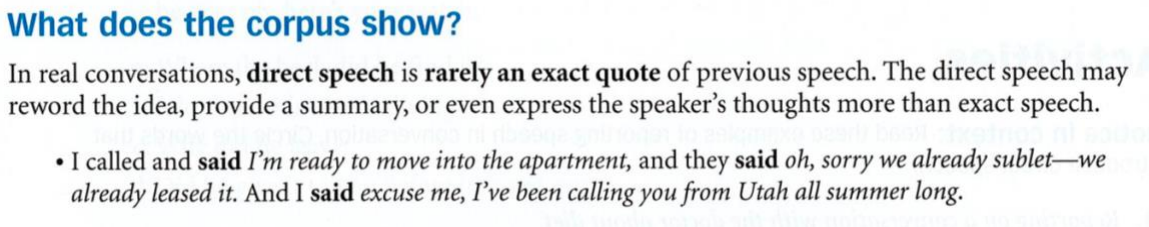
What have you learned from your grammar textbook?

There are two ways to report what someone said: (1) **Direct speech** quotes the exact words of the speaker, often introducing them with the "quoting verbs" *said* or *asked*. (2) **Indirect speech** does not use the exact words of the speaker. It uses a reporting verb and an indirect statement (or "noun clause").

1. Ali **said**, "*I **plan** to go to the party.*"
2. Ali **said** *he **planned** to go to the party.*

Tableau 34 : Présentation de la citation dans une grammaire traditionnelle (Conrad 2009).

La présentation explique que nous pouvons utiliser le discours direct ou indirect pour rapporter le discours d'une personne, le premier contenant un verbe de citation (dire, demander) et la citation mot à mot, le deuxième un verbe de citation et une proposition nominale. La Partie A complète ces informations par des observations reposant sur une analyse de corpus (tableau 35):



What does the corpus show?

In real conversations, **direct speech** is **rarely an exact quote** of previous speech. The direct speech may reword the idea, provide a summary, or even express the speaker's thoughts more than exact speech.

- I called and **said** *I'm ready to move into the apartment,* and they **said** *oh, sorry we already sublet—we already leased it.* And I **said** *excuse me, I've been calling you from Utah all summer long.*

Tableau 35 : Résumé des observations dans le « Longman Corpus Network » (Conrad 2009).

L'auteure explique que la citation mot à mot est plutôt rare dans le discours. Même quand nous prétendons fournir une citation directe, nous avons tendance à y rajouter notre impression ou notre opinion. Cette attitude évaluative semble être une tendance générale du comportement linguistique humain. Un discours neutre est rare, notre manière de rapporter ce qui a été dit inclut en général notre attitude personnelle vis à vis de l'objet cité.

La Partie B de la présentation est particulièrement intéressante car l'auteure offre des introductions typiques à des citations directes dont certaines sont, selon les grammaires traditionnelles, considérées comme stylistiquement problématiques. Le verbe le plus usité dans cette fonction est « say », un verbe neutre. Ce verbe est suivi de trois expressions courantes dans l'usage américain, « She goes », « I'm like » et « I'm all » (les trois signifiant « je dis ; il/elle dit »), associées à un niveau de langage familier, peu soigné. Conrad décrit ces trois expressions comme « très informelles », « utilisées avec des amis » et « utilisées majoritairement par les jeunes ». Ces trois indications concernant le registre et le groupe de locuteurs, aident l'apprenant à mieux comprendre celles qu'il peut utiliser et celles qu'il doit juste pouvoir comprendre à son niveau (v. tableau 36).

Say **is commonly used in direct speech in conversation.** *Ask* is rarely used (but *ask* is common for indirect speech). **Three other expressions** have become popular recently.

Verb/ Expression	Description of Use	Examples
1. ***say***	• most common verb	• And so I **said** *what are you doing?*
2. ***go***	• most often in simple present tense • used among friends • very informal • most common with younger adults but widely used	• He **goes** *I don't like to see girls in tight jeans.* • Jill said Annette called and Paul **goes** *well I didn't get the message.*
3. ***be like***	• can be used for thoughts (rather than speech) • used among friends • very informal • most commonly used by teenagers and young adults; also used by many older adults	• I'**m like** *are you from Idaho City* and she'**s like** *no do I look like it?* • Amy **was like** *uh, I think we should just buy some shelves.* • I spun around a couple of times, ran into a ditch and I'**m like** *what the heck just happened?* [describing a car accident]

Tableau 36 : Présentation des expressions typiques introduisant des citations directes (extrait) (Conrad 2009).

Le corpus révèle la présence des marqueurs de discours qui introduisent typiquement les citations (tableau 37). Ceux-ci ont une fonction pragmatique claire : ils rendent le discours plus naturel.

The **discourse markers *well, oh, look,*** and ***okay*** are sometimes used to mark the **beginning** of **direct speech.** (*See Unit 48 for more on discourse markers.*)

- And I said ***well*** *I'm gonna put it in, put a little five dollars in the thing and send it to the little children.*
- They brought the car over here and I said ***oh,*** *you made it over here with it.*
- He said ***look,*** *you guys have got to get together as a team and make a decision.*

Tableau 37 : Marqueurs de discours introduisant les citations directes (Conrad 2009).

Le tableau 38 clôture la phase de présentation qui, comme nous l'avons vu précédemment, met l'accent sur l'interface entre lexique et grammaire. La présentation est suivie d'activités de différents types dont la première consiste en un exercice d'observation. Les apprenants lisent les transcriptions de trois conversations authentiques et entourent les expressions introduisant les citations.

Notice in context: Read these examples of reporting speech in conversation. Circle the words that introduce direct speech.

1. *Reporting on a conversation with the doctor about diet.*

   He said to me the last time I was there, he says okay, well, you know what, let's add back two supplements. I go supplements? Pills? That's what you wanna add back, is just those pills? He goes wait, wait, wait. Okay, okay. I say how about a grain, you know. He goes okay corn tortillas. I said corn tortillas. You mean just the corn tortillas? How about corn bread, corn muffins? He's like no, no, no, just corn tortillas.

Tableau 38 : Extrait d'un exercice d'observation dans la « Grammaire réelle » (Conrad 2009).

L'exercice ci-dessus présente des textes contenant plusieurs citations successives. En élargissant l'environnement textuel des expressions et des phrases isolées à des textes plus longs, il montre aux apprenants la dynamique des récits ainsi que la fréquence des phrases introductives dans ces textes authentiques. Idéalement, l'enregistrement audio de la conversation serait inclus puisqu'il s'agit d'énoncés oraux. L'enregistrement peut faciliter la compréhension et l'observation de la prosodie des textes.

La deuxième activité implique l'analyse de quelques phrases contenant deux manières informelles d'introduire une citation, telles que présentées dans la partie B. Les apprenants sont invités à entourer les expressions « go » et « be like » et à souligner la citation directe (tableau 39).

Analyze discourse: Read Jennifer's reporting on a conversation with her colleague Suzi. Circle ***go*** and ***be like*** when they introduce direct speech and underline the direct speech.

1. Suzi asked me if I was gonna go to the seminar on Friday. She goes well, Jennifer, are you gonna go?
2. I go there won't be anyone there. She says well, Ken will be there.
3. I go he's teaching class from like eight thirty to noon.
4. She's like well, the seminar's for everybody.
5. I go oh, so the whole campus can close and everybody will go?
6. She's like well, yeah.

Tableau 39 : Exercice d'observation d'une citation informelle (extrait) (Conrad 2009).

Cette activité fait émerger le schéma typique de la citation directe par une répétition observable dans toutes les phrases : X (première ou troisième personne) « go/goes » / « am/is like » (+ marqueur de discours « well », « oh ») + citation directe.

La dernière activité présente de courts dialogues. Les apprenants sont invités à choisir la ou les réactions appropriées qui peuvent suivre l'énoncé contenant la citation directe (tableau 40).

Practice conversation: Read what your friend Kevin says. Then check (√) any response that is appropriate (there may be more than one possible response).

1. KEVIN: It was that night a couple weeks ago when they had all those sales. So my wife goes out to buy me a watch, and I'm all I'd like a Rolex*.
   _____ **a.** Why did she say that?
   _____ **b.** She's shopping right now?
   _____ **c.** Did she laugh when you told her that?
   _____ **d.** Does she always announce when she's leaving?
2. KEVIN: I couldn't find my watch for a week, and my friend's sitting in our living room, and suddenly he goes oh, here's your watch.
   _____ **a.** Where the heck did he find it?
   _____ **b.** I don't see him in the living room now.
   _____ **c.** Why did he leave?
   _____ **d.** Did he give it to you before he went out?

Tableau 40 : Dialogues incluant des exemples de citations (extrait) (Conrad 2009).

Cette activité place les citations directes dans leurs environnements textuels plus larges (ici, des interactions). Les dialogues présentés incluent la réaction de l'interlocuteur, la suite naturelle de la conversation. Si l'ouvrage est utilisé en cours d'anglais, les apprenants peuvent, par exemple, apprendre et jouer les dialogues, puis en modifier certains éléments[60].

Dans les exercices proposés, l'apprenant a donc l'opportunité d'observer non seulement le phénomène grammatical mais aussi tout ce qui l'entoure : des citations plus longues, des récits illustrant la citation d'échanges entre plusieurs personnes, ou encore les réactions possibles de l'interlocuteur.

Dans les paragraphes suivants, nous présenterons deux types d'exercices du même ouvrage pour montrer comment les productions orale et écrite s'intègrent dans les chapitres. (Le chapitre sur la citation n'en donne pas d'exemple.) L'aspect grammatical étudié concerne les adjectifs avec préposition. Les exercices d'observation et de pratique sont suivis d'un premier exercice oral qui incite les apprenants à débattre sur un ou plusieurs des sujets proposés en utilisant des adjectifs avec préposition (tableau 41). Dans le deuxième exercice, ils doivent rédiger un texte sur les sujets débattus (tableau 42).

[60] Voir aussi les activités proposées au chapitre 15 dans la Partie III de cette thèse.

3 Practice conversation: Debate the following topics with a partner. For each topic, choose one side of the argument and try to convince your partner that your point of view is the best one. Use each of the **adjective + preposition** combinations in the box at least once and any other combinations that you choose.

better for	*good for*	*happy with*	*right about*
different from	*great for*	*mad at*	*wrong with*

1. Driving a car versus riding a bike to work or school.
2. Reading the book versus watching the movie about the book.
3. Living in a city versus living in the country.
4. Sending a letter to a friend versus sending an email to a friend.

EXAMPLE

Riding your bike provides exercise, which is **good for** your health, and it is **better for** the environment than driving a car.

Tableau 41 : Activité de production orale dans la « Grammaire réelle » (Conrad 2009).

4 Practice writing: Choose one of the topics you debated above, and write a paragraph explaining your point of view. Use at least three **adjective + preposition** combinations in your paragraph. Share your paragraph with your partner.

EXAMPLE

I believe sending a letter can be **important for** communication between friends because it is more personal than an email. There are several other reasons, including . . .

Tableau 42 : Activité de production écrite dans la « Grammaire réelle » (Conrad 2009).

Ces deux exercices de production libre ont, à notre avis, un grand défaut : *ils ne proposent pas de vocabulaire-clé pour aider les apprenants à formuler leurs arguments*. Même si les sujets ne sont pas en soi difficiles (utiliser le vélo ou la voiture, lire et regarder un film, habiter en ville ou à la campagne, écrire une lettre ou un e-mail à un ami), afin de souligner l'interconnexion entre lexique et grammaire, il aurait été souhaitable d'inclure des expressions qui aident les apprenants (surtout aux niveaux A2 et B1) à rendre leurs énoncés plus naturels et authentiques (pour des exemples concrets voir la section B ci-dessous).

3.3) Analyse succincte de l'ouvrage

La « Grammaire réelle » de Conrad présente 50 phénomènes grammaticaux sélectionnés, fondés sur l'analyse d'un grand corpus général. Cet ouvrage incorpore non seulement des textes authentiques mais aussi la majorité des résultats pertinents de l'analyse de corpus ainsi qu'une manière originale et efficace de présentation des aspects grammaticaux choisis. Les phénomènes apparaissent dans leurs environnements textuels et l'apprenant trouve de nombreux renseignements concernant leur utilisation. Qu'il s'agisse d'informations sur la fréquence d'usage ou sur des points relatifs aux registres, aux langages parlé et écrit, ces informations supplémentaires donnent un aperçu plus précis du phénomène langagier tel qu'il est utilisé dans des textes

authentiques. Une place particulière revient à la « grammaire orale », variété de la grammaire largement négligée (car associée à un « mauvais » usage langagier, non standard) avant que la recherche en linguistique de corpus ne l'établisse comme une grammaire à part entière avec ses propres règles.

Au niveau méthodologique, les exercices accompagnant la présentation apparaissent en partie innovants par rapport aux grammaires traditionnelles. Nous pensons en particulier aux *exercices d'observation qui, plaçant le phénomène donné dans un environnement textuel plus large, authentique ou semi-authentique, permettent aux apprenants d'identifier des schémas d'usage et de voir comment l'aspect étudié s'intègre dans un discours plus complet.*

La partie qui, au niveau didactique, pose problème est celle consacrée à la production. Ici, les apprenants sont laissés « seuls » : *les exercices de l'expression orale ou écrite obligent les apprenants de s'appuyer sur leurs connaissances existantes qui, aux niveaux A2 et B1, ne suffisent probablement pas pour produire des énoncés à caractère naturel.* Ces exercices de rédaction libre sont d'autant plus redoutables que la recherche en linguistique de corpus souligne l'importance des unités multi-lexicales ainsi que celle de l'observation et de la pratique de l'usage langagier typique des natifs. Des exercices supplémentaires amenant progressivement l'apprenant, depuis l'analyse des textes jusqu'à la production libre intégrant des éléments multi-lexicaux, pourraient paver le chemin de l'observation vers les activités de plus en plus complexes de productions orale et écrite.

B) Avantages et limites des ouvrages présentés

Dans ce chapitre, nous avons exposé deux grammaires fondées sur l'analyse de corpus. Ces livres nous ont servi d'exemples pour illustrer les principales caractéristiques de ce type de grammaire, à savoir :

- La présentation des phénomènes grammaticaux dans un environnement textuel en utilisant des exemples authentiques ou semi-authentiques.
- La distinction entre les registres, entre l'usage langagier à l'oral et à l'écrit ainsi que l'emphase sur les schémas grammaticaux.

Cette approche est également observable dans d'autres ouvrages comme « The Longman student grammar of spoken and written English » par Biber et al. (2002), « The grammar of spoken and written English » par Carter et McCarthy (2006), « English Grammar Today : An A–Z of Spoken

and Written Grammar » par Carter et al. (2016) pour l'anglais, ou dans la nouvelle série « Bausteine einer Korpusgrammatik des Deutschen » pour la grammaire allemande (Konopka et al. 2020), pour ne citer que quelques exemples.

L'un des avantages de ces grammaires est de compléter la description des grammaires issues uniquement ou principalement de sources écrites. Nous avons vu à travers l'exemple de la citation directe que la description des phénomènes grammaticaux peut être enrichie et précisée par l'analyse de corpus[61]. Si le CECRL prescrit des compétences interactionnelles dès les niveaux débutants (compétences, en effet, essentielles dès le début de l'apprentissage), nous ne pouvons pas nous passer de la présentation des caractéristiques grammaticales du langage parlé.

Un autre trait de caractère commun de ces ouvrages est *qu'une partie des activités est dédiée à l'observation langagière, permettant aux apprenants l'identification des schémas et des environnements textuels typiques avant de passer à la pratique de l'aspect grammatical donné*. Ces exercices rendent l'apprenant conscient de la répétition structurelle entre des énoncés réalisant une même fonction. Il convient cependant de constater *qu'un travail plus approfondi sur les unités multi-lexicales pourrait augmenter l'efficacité des exercices et attirer encore davantage l'attention de l'apprenant encore plus sur l'interrelation entre les aspects grammaticaux et lexicaux de la langue étudiée*.

L'analyse de ces ouvrages nous a permis de repérer quelques critères caractérisant les grammaires pédagogiques existantes basées sur le corpus. Nous avons également identifié quelques suggestions qui pourraient améliorer l'efficacité de cette approche dans le cadre de l'enseignement des langues. Dans cette optique, les points suivants nous semblent être particulièrement pertinents :

- Les grammaires intégrant cette approche peuvent chercher à systématiser les phénomènes grammaticaux de la langue donnée ou se concentrer sur certains phénomènes qui échappent à des règles claires et méritent d'être explorés du point de vue de l'usage langagier.
- Dans l'intérêt d'un apprentissage efficace, il est nécessaire d'inclure des exercices qui amènent l'apprenant de l'observation vers la pratique guidée. Au lieu de passer après un seul exercice de la pratique à la production libre, l'apprenant pourrait être d'abord invité à compléter les unités multi-lexicales dans des textes authentiques ou semi-authentiques à

[61] Granger (2013) illustre ce constat par un autre exemple. Elle remarque que les grammaires non fondées sur corpus ne mentionnent pas le verbe « get » comme verbe utilisé dans des phrases passives, alors qu'il serait le premier choix des natifs dans le discours oral.

caractère naturel, par exemple, choisir la fin des phrases authentiques ou semi-authentiques à caractère naturel. Ces étapes successives présenteraient l'aspect grammatical donné dans un nombre de textes et une variété de contextes et contribueraient à son approfondissement.

- Quand l'apprenant doit rédiger un texte, il est souhaitable de fournir du vocabulaire typique, fondé sur un corpus, par exemple pour le débat concernant la voiture et le vélo : « when I am on my bike » (quand je fais du vélo), « good for the environment » (bon pour l'environnement), « less pollution in the air » (moins de pollution dans l'air) ; « exercice is good for you » (l'exercice physique est bien pour vous), « responsible for accidents » (responsable d'accidents), « safe for cyclists » (sûr pour les cyclistes), « implement protected bike lanes » (implémenter des pistes cyclables protégées). Ces unités multi-lexicales rajoutent à la complexité du débat en apportant des éléments à considérer et améliorent la qualité linguistique du texte de l'apprenant.
- Lorsque l'apprenant est censé produire un texte spécifique (débat, argumentation, récit), quelques expressions typiques de tels textes pourraient être présentés.
- Dans le cas de phénomènes du langage parlé, il est bénéfique d'inclure des enregistrement audio ou multimédia. Suivre les transcriptions n'est pas toujours aisé car les énoncés oraux sont notamment structurés par la prosodie, par les pauses et par d'autres caractéristiques phonologiques.
- Il est également recommandé d'intégrer des instructions telles que « Consultez le corpus X pour plus d'exemples » ou « Consultez des blogs et des forums traitant de ce sujet et réutilisez les unités multi-lexicales observées dans votre texte. » Cela permettrait aux apprenants de découvrir l'environnement textuel de l'aspect grammatical présenté et de l'utiliser activement par la suite.
- Ce dernier point soulève la question des corpus appropriés pour les niveaux de compétences linguistiques inférieurs. Ces corpus nous apparaissent nécessaires pour maximiser l'efficacité du travail avec ce type d'ouvrage et pour guider les étudiants vers un apprentissage autonome.

Dans ce chapitre, nous avons présenté deux grammaires fondées sur corpus pour démontrer l'intérêt de cette approche dans le cadre pédagogique. La « Grammaire des schémas » par Francis, Hunston et Manning propose une catégorisation possible des schémas grammaticaux, alors que la « Grammaire réelle » de Conrad met l'accent sur certains phénomènes langagiers. L'examen de ces

deux exemples nous montrent que les grammaires pédagogiques fondées sur les corpus ne devraient pas viser à l'exhaustivité. Elles peuvent se concentrer sur les points qui méritent l'exploration de corpus. Par exemple, il serait inutile d'examiner les aspects morphologiques des différentes formes du passé du verbe dans un corpus car leur description ne bénéficierait pas d'une telle démarche. En revanche, la présentation de la citation, phénomène « banal » à première vue, peut être significativement enrichie par l'approche corpus. Il apparaît donc souhaitable – voire essentiel – d'identifier les points grammaticaux méritant une étude dans le corpus et d'inclure les résultats de ces études dans les ouvrages pédagogiques.

Nous avons également montré que les phases d'observation, de pratique et de production pourraient être toutes trois enrichies par la consultation d'un corpus adapté au niveau et aux besoins des apprenants. Les opportunités d'observation et d'analyse, mais aussi la simple lecture d'un certain nombre de phrases (les rencontres répétées avec des énoncés de structure similaire), peuvent aider les apprenants à approfondir leurs connaissances concernant le point grammatical présenté et à accumuler de l'expérience linguistique. Ce prolongement n'est cependant pas envisageable sans corpus pédagogiques appropriés[62].

Finalement, les grammaires étudiées dans ce chapitre proposent deux manières différentes de sensibiliser l'apprenant à l'interconnexion des aspects grammaticaux, lexicaux, sémantiques et pragmatiques. Elles l'entrainent, chacune à sa façon, à faire l'expérience de la langue comme un tissu composé de ces aspects inséparables. Au chapitre suivant nous examinerons comment cette même démarche a été réalisée au sein de manuels.

[62] La Partie III de cette thèse sera dédiée à la création de tels corpus.

Chapitre 6 : Intégrer les résultats de la linguistique de corpus dans les matériels pédagogiques : les manuels de cours

Dans le chapitre 5, nous avons présenté des grammaires pédagogiques fondées sur une analyse de corpus. Nous avons montré que celles-ci tiennent compte de certains aspects de la langue qui n'apparaissent pas dans une approche plus traditionnelle et proposent de ce fait de nouvelles méthodes permettant de systématiser les connaissances de l'apprenant.

Ce chapitre exposera les diverses manières permettant d'intégrer les résultats de recherche en linguistique de corpus dans les manuels pour l'enseignement de langues. Notons de manière liminaire que, bien que l'on ait assisté à la publication d'un éventail croissant de grammaires pédagogiques inspirées par les résultats d'analyse de corpus, le nombre de manuels intégrant de tels résultats reste encore limité. Si certains ouvrages annoncent sur leurs couvertures que l'étudiant apprendra « la langue réelle », il apparaît à l'examen qu'il ne s'agit là que d'une stratégie de marketing. Hormis quelques expressions du langage parlé, l'usage langagier des natifs n'est pas reflété de façon systématique dans la majorité de ces ouvrages.

Dans ce chapitre, nous analyserons deux séries de manuels, un pour l'anglais et un autre pour le hongrois qui adoptent chacun deux approches didactiques cohérentes, intégrant l'utilisation des corpus. Nous analyserons la manière de présenter les informations provenant du corpus, quelques exercices innovants et l'adaptation des exercices classiques pour identifier, comme dans le cas des grammaires, les principales caractéristiques des manuels informés par le corpus.

A) Manuels de cours

1) Aperçu général

Concernant l'implémentation des résultats de la recherche linguistique dans les manuels de langues, Lee et McGarrell (2011 : 95) soulignent l'importance d'une collaboration étroite entre linguistes et auteurs de manuels. D'après eux, *les linguistes de corpus devraient fournir aux rédacteurs de manuels des directives claires sur le type d'informations qui méritent d'être incluses dans les manuels*.

Lors de la rédaction d'un manuel, les auteurs doivent prendre de nombreuses décisions pour garantir que l'approche pédagogique souhaitée se reflète véritablement tout au long de l'ouvrage. Meunier et Reppen (2015 : 498) proposent une liste de facteurs que les auteurs aspirant à créer un ouvrage avec le label « informé par le corpus » (*corpus-informed*) doivent considérer :

- Les auteurs doivent réfléchir sur la manière d'inclure les résultats issus de la recherche en linguistique de corpus ;
- Ils doivent considérer que doit être inclus précisément : vocabulaire, contextes d'utilisation, schémas grammaticaux, collocations typiques, informations sur la fréquence et ainsi de suite ;
- Ils doivent prendre des décisions liées à la présentation des informations provenant du corpus (textes, graphiques, lignes de concordance) ;
- Ils doivent effectuer une sélection de textes appropriés (oraux et écrits) et les adapter au niveau donné. Ils doivent également établir les critères de modifications qui garantissent dans ces textes un usage langagier à caractère naturel.

Dans les pages suivantes, nous étudierons deux ouvrages en particulier : la série de manuels pour l'anglais « Touchstone » et la série de manuels pour le hongrois « MagyarOK ». Ces ouvrages illustrent deux approches différentes et complémentaires : « Touchstone » fournit des informations explicites concernant l'usage langagier des natifs et propose des exercices de moins en moins guidés permettant à l'apprenant de les assimiler ; « MagyarOK » offre des séquences thématiques dans lesquelles les exercices approfondissent graduellement les caractéristiques choisies de la langue.

2) Manuels pour l'anglais : la série « Touchstone »

2.1) Le concept

« Touchstone » (McCarthy et al. 2005-2011) est une série pour adultes et jeunes adultes, qui amène les étudiants du niveau débutant au niveau intermédiaire (CECRL A1-B1)[63]. La série consiste en quatre volumes et s'adresse donc aux apprenants des niveaux de compétences linguistiques inférieurs. Le site Web de Cambridge University Press explique que la série s'appuie sur l'analyse du « Cambridge International Corpus », grande base de données linguistiques pour l'anglais, comprenant des conversations quotidiennes et des textes de journaux et de livres (https://www.cambridge.org/ms/cambridgeenglish/catalog/adult-courses/touchstone-2nd-edition/components 2021).

[63] Une deuxième série a été produite par les mêmes auteurs intitulée « Viewpoint » (2012–2013) en reprenant les mêmes principes. Dans notre présentation, nous n'utiliserons que « Touchstone » pour pouvoir montrer une séquence entière cohérente.

D'après cette description, l'une des principales caractéristiques de la série est *l'utilisation d'un langage à caractère naturel dans des situations réelles (même si certains textes ont été modifiés pour garantir leur accessibilité aux niveaux de compétences linguistiques inférieurs), y compris l'usage langagier à l'oral*. L'observation et la pratique régulières des interactions parlées accompagnent l'apprenant tout au long de la série en le préparant à participer à des situations du quotidien dès le début de son parcours, en accord avec les descriptions du CECRL. Une place particulière revient aux stratégies conversationnelles, présentées dans des dialogues semi-authentiques suivis d'activités orales. Ces séquences créent des opportunités pour des interactions et pour un discours personnalisé.

Tous les chapitres suivent la même structure : ils se construisent autour d'un thème et se divisent en quatre parties :

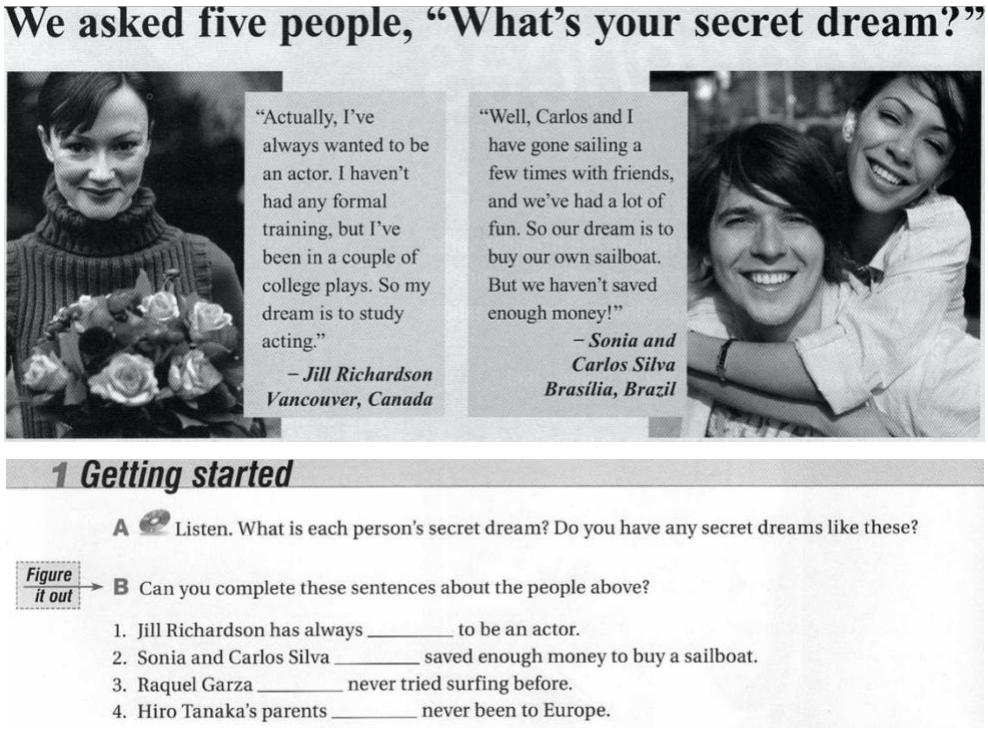
- La Partie A présente le thème en utilisant plusieurs textes courts, propose et applique un aspect grammatical issu des textes, et conclut sur un exercice oral.
- La Partie B travaille sur le vocabulaire lié à un aspect du thème et énonce également un aspect grammatical. Ces exercices sont suivis d'activités utiles pour enrichir le vocabulaire et pratiquer certains éléments de l'usage langagier typique des natifs.
- La Partie C est dédiée aux stratégies conversationnelles qui sont approfondies par des exercices d'écoute et des discussions guidées.
- La Partie D forme aux quatre compétences et contient des textes plus longs écrits et oraux ainsi que des exercices d'écriture et de discussion.

Dans les pages suivantes, nous exposerons les exercices de chaque partie qui présentent un intérêt particulier au vu de l'intégration des résultats d'analyse de corpus. Pour cela, nous avons choisi l'Unité 2 du « Touchstone 3 » (niveau A2 du CECRL), relative au thème des « Expériences ».

2.2) Le contenu : une séquence d'activités

Partie A, exercice 1 : Cinq personnes et leurs rêves

Le premier sujet dans la Partie A porte sur les rêves : les apprenants lisent et écoutent cinq textes de 2 à 3 phrases qui présentent les rêves de différentes personnes. *L'approche corpus se manifeste en ce que les textes sont adaptés de textes authentiques tout en gardant leur caractère naturel.* Deux exercices, un de vocabulaire dans lequel les apprenants doivent compléter des phrases sur les personnes et une question qui peut initier une courte discussion (« Avez-vous des rêves secrets similaires ? ») complètent le travail avec les textes (tableau 43).



We asked five people, "What's your secret dream?"

"Actually, I've always wanted to be an actor. I haven't had any formal training, but I've been in a couple of college plays. So my dream is to study acting."

– Jill Richardson
Vancouver, Canada

"Well, Carlos and I have gone sailing a few times with friends, and we've had a lot of fun. So our dream is to buy our own sailboat. But we haven't saved enough money!"

– Sonia and Carlos Silva
Brasília, Brazil

1 Getting started

A Listen. What is each person's secret dream? Do you have any secret dreams like these?

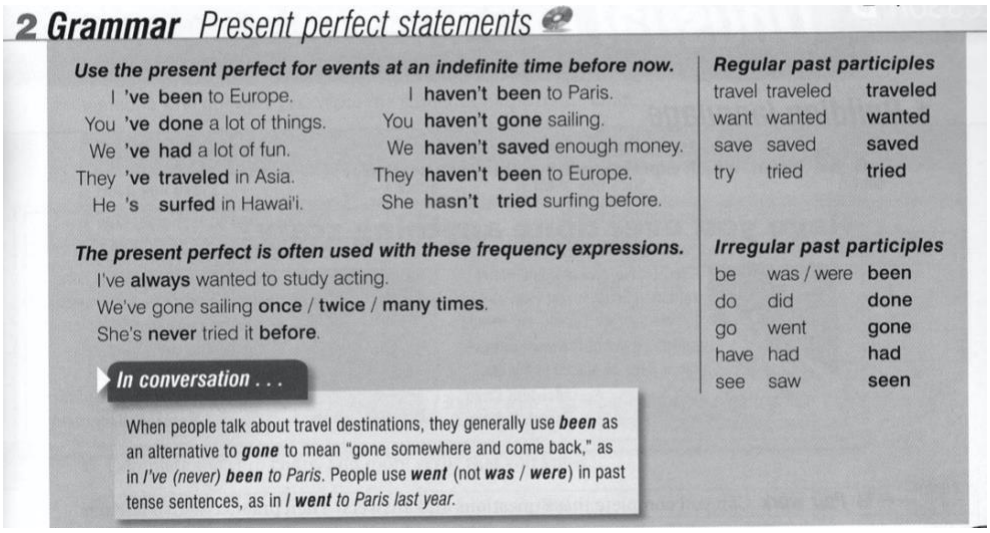
Figure it out → B Can you complete these sentences about the people above?

1. Jill Richardson has always _____ to be an actor.
2. Sonia and Carlos Silva _____ saved enough money to buy a sailboat.
3. Raquel Garza _____ never tried surfing before.
4. Hiro Tanaka's parents _____ never been to Europe.

Tableau 43 : Exercice introduisant le thème du chapitre avec des textes authentiques, modifiés (Touchstone A2).

Partie A, exercice 2 : affirmations au « present perfect »

L'exercice 2 présente l'utilisation du « present perfect » que les apprenants ont pu observer dans les textes, complétée par des notes sur la conversation. Nous apprenons que « I have been » (j'ai été) est utilisé plus fréquemment par les natifs que « I have gone » (je suis allé) ; en revanche le passé simple privilégie la structure « I went to » (je suis allé), plutôt que « I was ». Ces renseignements démontrent que la fréquence des éléments langagiers mérite d'être relevée car elle fournit des précisions utiles sur l'usage langagier (tableau 44).



2 Grammar *Present perfect statements*

Use the present perfect for events at an indefinite time before now.

I	**'ve been** to Europe.	I	**haven't been** to Paris.
You	**'ve done** a lot of things.	You	**haven't gone** sailing.
We	**'ve had** a lot of fun.	We	**haven't saved** enough money.
They	**'ve traveled** in Asia.	They	**haven't been** to Europe.
He	**'s surfed** in Hawai'i.	She	**hasn't tried** surfing before.

The present perfect is often used with these frequency expressions.

I've **always** wanted to study acting.
We've gone sailing **once** / **twice** / **many times**.
She's **never** tried it **before**.

In conversation . . .

When people talk about travel destinations, they generally use ***been*** as an alternative to ***gone*** to mean "gone somewhere and come back," as in *I've (never)* ***been*** *to Paris.* People use ***went*** (not ***was*** / ***were***) in past tense sentences, as in *I* ***went*** *to Paris last year.*

Regular past participles

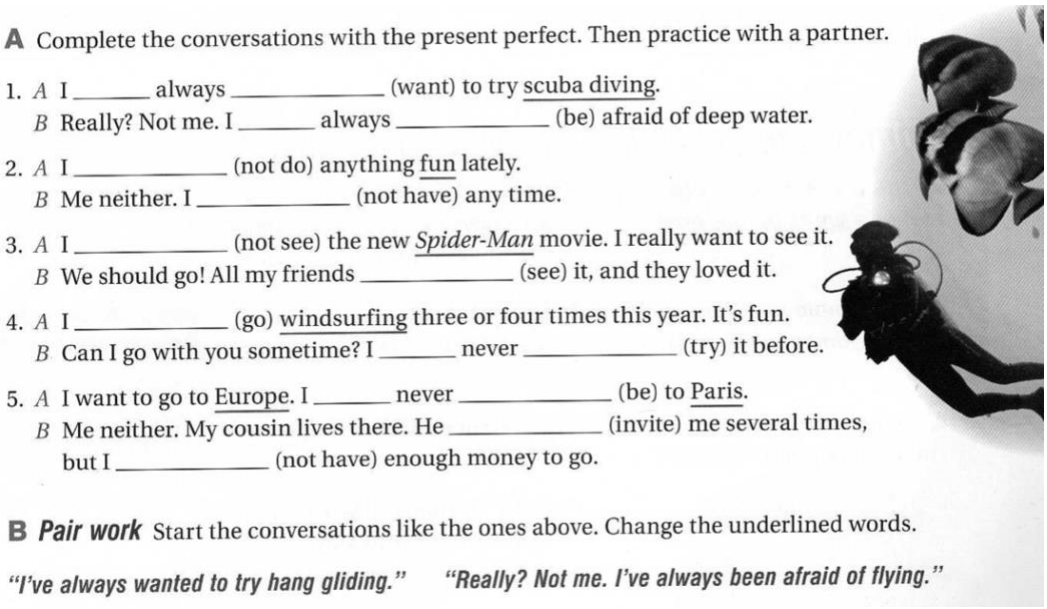
travel	traveled	**traveled**
want	wanted	**wanted**
save	saved	**saved**
try	tried	**tried**

Irregular past participles

be	was / were	**been**
do	did	**done**
go	went	**gone**
have	had	**had**
see	saw	**seen**

Tableau 44 : Présentation du « present perfect » intégrant quelques informations utiles pour la conversation (Touchstone A2).

L'exercice comporte des dialogues semi-authentiques (adaptés de dialogues authentiques), à caractère naturel. Ici, *les apprenants peuvent observer quelques réponses typiques permettant de réagir aux phrases du partenaire*. Ces brèves interactions fournissent aux apprenants des outils pour les interactions réelles du quotidien (tableau 45).



A Complete the conversations with the present perfect. Then practice with a partner.

1. *A* I _____ always _____ (want) to try scuba diving.
   B Really? Not me. I _____ always _____ (be) afraid of deep water.
2. *A* I _____ (not do) anything fun lately.
   B Me neither. I _____ (not have) any time.
3. *A* I _____ (not see) the new *Spider-Man* movie. I really want to see it.
   B We should go! All my friends _____ (see) it, and they loved it.
4. *A* I _____ (go) windsurfing three or four times this year. It's fun.
   B Can I go with you sometime? I _____ never _____ (try) it before.
5. *A* I want to go to Europe. I _____ never _____ (be) to Paris.
   B Me neither. My cousin lives there. He _____ (invite) me several times, but I _____ (not have) enough money to go.

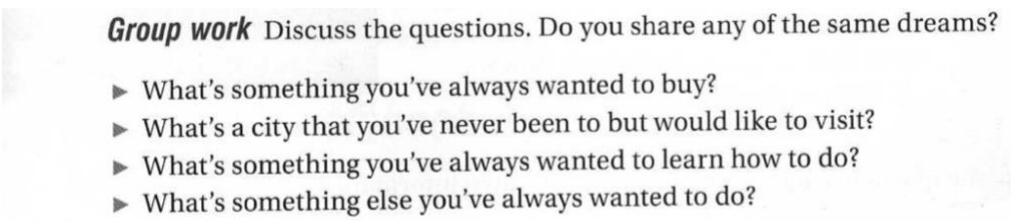
B ***Pair work*** Start the conversations like the ones above. Change the underlined words.

"I've always wanted to try hang gliding." ***"Really? Not me. I've always been afraid of flying."***

Tableau 45 : Observer et compléter des interactions contenant quelques réponses fréquentes (Touchstone A2).

La discussion de l'exercice B invite les apprenants à produire des dialogues semi-guidés : ils peuvent changer les mots soulignés et apporter des modifications aux dialogues de l'exercice A. Ainsi, *ces dialogues servent de modèles pour les interactions plus personnalisées*.

L'exercice 3 (non présenté ici) fait pratiquer par la suite ces éléments langagiers dans des conversations libres (tableau 46).



Group work Discuss the questions. Do you share any of the same dreams?

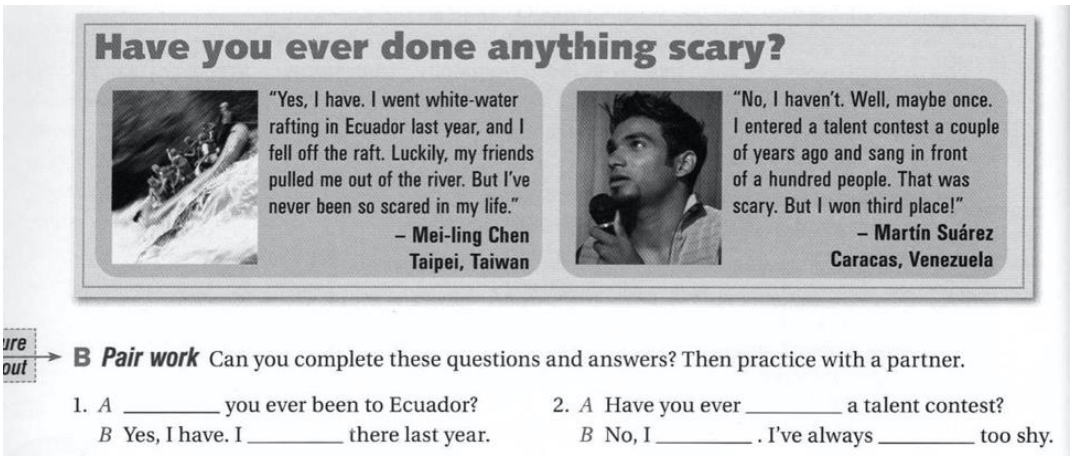
- What's something you've always wanted to buy?
- What's a city that you've never been to but would like to visit?
- What's something you've always wanted to learn how to do?
- What's something else you've always wanted to do?

Tableau 46 : Pratiquer les éléments langagiers présentés dans des conversations libres (Touchstone A2).

Au cours de cette série d'exercices, nous nous éloignons graduellement des textes-exemples (des monologues de l'exercice 1 et des dialogues de l'exercice 2), en personnalisant les énoncés. Après avoir assimilé suffisamment d'éléments langagiers, l'apprenant est invité, dans l'exercice 3, à mener à bien des discussions libres. L'aspect grammatical (les temps du verbe) émerge naturellement lors de ces discussions mais il n'est pas au centre de l'exercice : il est un des outils qui permettent aux apprenants de produire des phrases proches de celles des natifs. Ainsi, l'interconnexion entre grammaire et lexique se manifeste également dans les exercices de production relativement libre.

Partie B, exercice 1 : Avez-vous déjà fait quelque chose qui vous a fait peur ?

La partie B commence par un court exercice d'écoute intitulé « Building language » (Construire la langue) qui reprend le thème des expériences et fait aussi pratiquer quelques réactions fréquemment utilisées dans le quotidien (tableau 47).



Have you ever done anything scary?

"Yes, I have. I went white-water rafting in Ecuador last year, and I fell off the raft. Luckily, my friends pulled me out of the river. But I've never been so scared in my life."
– Mei-ling Chen
Taipei, Taiwan

"No, I haven't. Well, maybe once. I entered a talent contest a couple of years ago and sang in front of a hundred people. That was scary. But I won third place!"
– Martín Suárez
Caracas, Venezuela

B ***Pair work*** Can you complete these questions and answers? Then practice with a partner.

1. *A* _____ you ever been to Ecuador?
   B Yes, I have. I _____ there last year.

2. *A* Have you ever _____ a talent contest?
   B No, I _____. I've always _____ too shy.

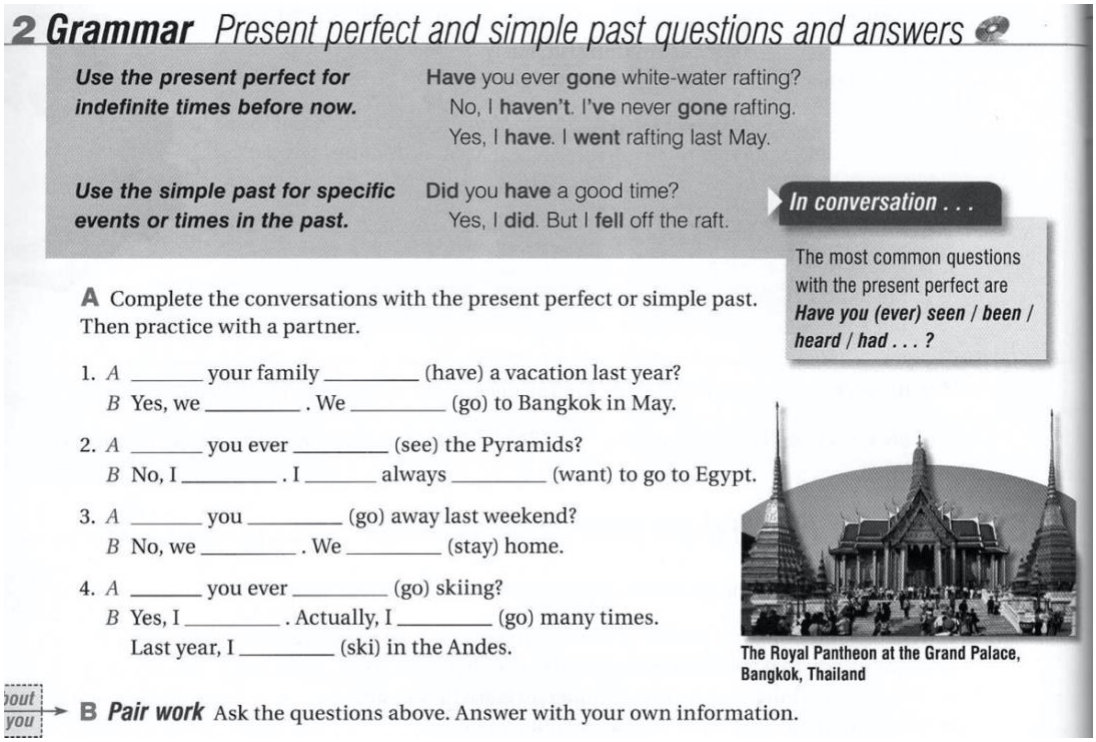
Tableau 47 : Questions et réactions fréquentes du quotidien (Have you ever ... ? – Yes, I have./No, I haven't.) (Touchstone A2).

Les textes et les réactions présentées dans l'exercice, contiennent de nombreuses phrases au « present perfect » et au « simple past ». Les différences dans l'utilisation de ces deux temps du verbe seront expliquées dans l'exercice 2, ici, la grammaire est intégrée dans des situations de communication. En outre, le texte comporte quelques unités multi-lexicales utiles car typiques telles que « I've never been so ADJ (*scared*) in my life » ou « That was ADJ (*scary*). » qui pourraient être présentées et pratiquées indépendamment afin d'en souligner l'importance.

Partie B, exercice 2: Le « present perfect » et le « simple past » dans des questions et des réponses

Dans l'exercice 2, la grammaire reprend les temps du verbe et oppose l'utilisation du « present perfect » à celle du « passé simple ». La note à côté du tableau nous renseigne sur les questions les

plus courantes avec le « present perfect » telles que « Have you ever seen / been / heard / had ...? » (As-tu déjà vu / été / entendu / eu ... ?). Comme dans la Partie A, la présentation de la grammaire est suivie de dialogues à compléter, adaptés de dialogues réels. Nous voyons de nouveau la pratique de la grammaire s'intégrer dans des situations (semi-)authentiques pour lesquelles le lexique typique est également proposé (tableau 48).



2 Grammar *Present perfect and simple past questions and answers*

Use the present perfect for indefinite times before now.

Have you ever **gone** white-water rafting?
No, I **haven't**. I've never **gone** rafting.
Yes, I **have**. I **went** rafting last May.

Use the simple past for specific events or times in the past.

Did you **have** a good time?
Yes, I **did**. But I **fell** off the raft.

In conversation . . .

The most common questions with the present perfect are ***Have you (ever) seen / been / heard / had . . . ?***

A Complete the conversations with the present perfect or simple past. Then practice with a partner.

1. *A* _____ your family _____ (have) a vacation last year?
   B Yes, we _____ . We _____ (go) to Bangkok in May.
2. *A* _____ you ever _____ (see) the Pyramids?
   B No, I _____ . I _____ always _____ (want) to go to Egypt.
3. *A* _____ you _____ (go) away last weekend?
   B No, we _____ . We _____ (stay) home.
4. *A* _____ you ever _____ (go) skiing?
   B Yes, I _____ . Actually, I _____ (go) many times. Last year, I _____ (ski) in the Andes.

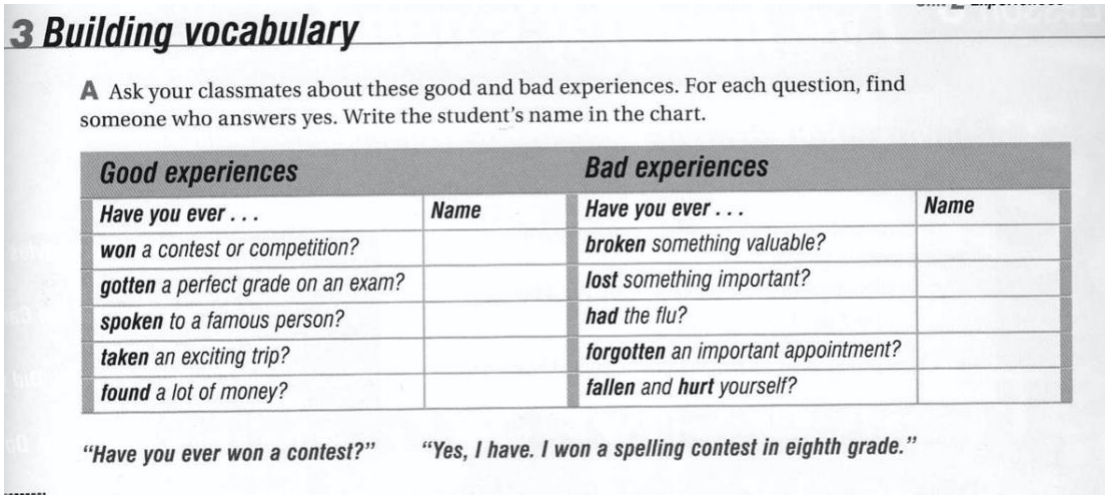
The Royal Pantheon at the Grand Palace, Bangkok, Thailand

B ***Pair work*** Ask the questions above. Answer with your own information.

Tableau 48 : Conversations à caractère naturel autour des expériences de voyage (Touchstone A2).

Partie B, exercice 3 : Construire le vocabulaire : bonnes et mauvaises expériences

Une autre activité qui met en exergue l'interconnexion entre lexique et grammaire est présenté dans l'exercice 3. Cet exercice propose quelques questions toutes faites, relatives aux bonnes et mauvaises expériences du quotidien et invite les apprenants à échanger sur ces thèmes. Les questions et l'exemple de réponse mettent en évidence la différence d'utilisation des deux temps du verbe. L'exercice présente également des unités multi-lexicales fréquentes, utiles telles que « win a contest » (gagner à une compétition), « get a good grade » (avoir une bonne note), « speak to a person » (parler à une personne), « take a trip » (partir en voyage), « forget an appointment » (oublier un rendez-vous), « hurt yourself » (se blesser) (tableau 49).



3 Building vocabulary

A Ask your classmates about these good and bad experiences. For each question, find someone who answers yes. Write the student's name in the chart.

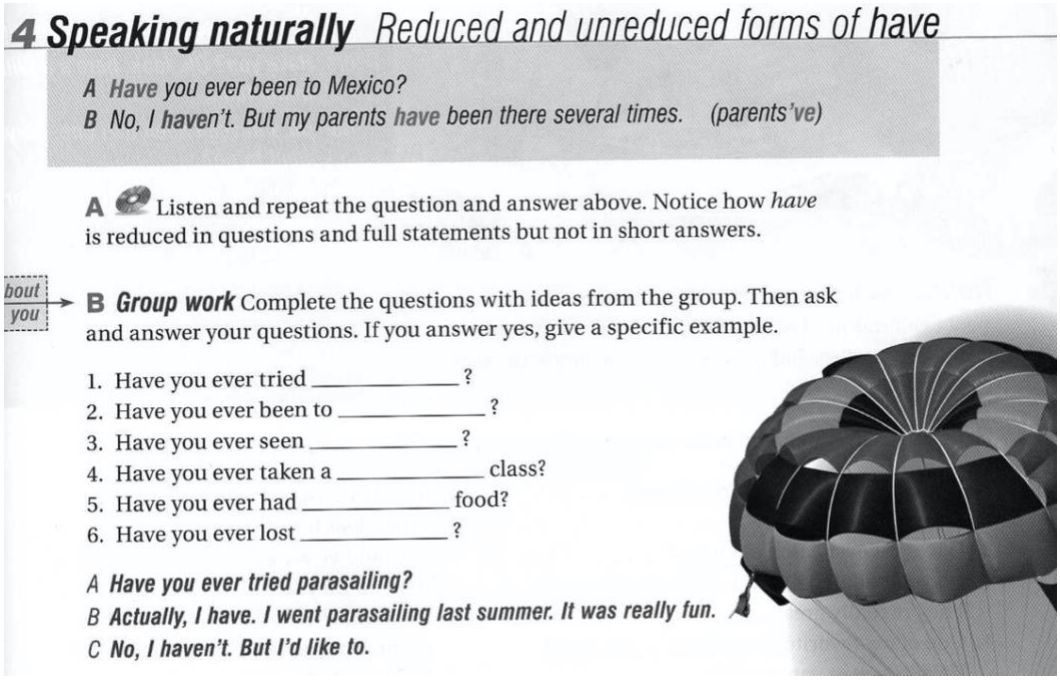
Good experiences		Bad experiences	
Have you ever . . .	**Name**	**Have you ever . . .**	**Name**
won a contest or competition?		**broken** something valuable?	
gotten a perfect grade on an exam?		**lost** something important?	
spoken to a famous person?		**had** the flu?	
taken an exciting trip?		**forgotten** an important appointment?	
found a lot of money?		**fallen** and **hurt** yourself?	

"Have you ever won a contest?" *"Yes, I have. I won a spelling contest in eighth grade."*

Tableau 49 : Pratiquer le « present perfect » et le vocabulaire sur des expériences vécues (Touchstone A2).

Partie B, exercice 4 : Parler de façon naturelle : La forme réduite de « have »

L'exercice 4 attire l'attention de l'apprenant sur l'importance du registre. Il est intitulé « Speaking naturally » (Parler de façon naturelle) et traite du phénomène de la contraction de « have » dans les interactions orales. Les apprenants complètent les questions avec des activités de leur choix et échangent à leur sujet. L'exercice fournit aussi une réaction positive et une réaction négative, les deux fréquemment utilisées, que les apprenants peuvent intégrer dans leurs réponses (tableau 50).



4 Speaking naturally *Reduced and unreduced forms of have*

A Have you ever been to Mexico?
B No, I haven't. But my parents have been there several times. (parents've)

A Listen and repeat the question and answer above. Notice how *have* is reduced in questions and full statements but not in short answers.

about you → **B Group work** Complete the questions with ideas from the group. Then ask and answer your questions. If you answer yes, give a specific example.

1. Have you ever tried _____?
2. Have you ever been to _____?
3. Have you ever seen _____?
4. Have you ever taken a _____ class?
5. Have you ever had _____ food?
6. Have you ever lost _____?

A ***Have you ever tried parasailing?***
B ***Actually, I have. I went parasailing last summer. It was really fun.***
C ***No, I haven't. But I'd like to.***

Tableau 50 : « Parler de façon naturelle » avec des questions et des réponses-modèles (Touchstone A2).

Dans la séquence de ces quatre activités, nous observons une fois de plus comment la présentation, la pratique et la production libre élargissent graduellement les connaissances linguistiques de

l'apprenant, tout en restant dans le cadre d'une approche lexico-grammaticale. Les deux temps du verbe sont introduits avec quelques fonctions typiques et avec un vocabulaire que l'apprenant peut rencontrer au quotidien et inversement : le vocabulaire autour des expériences est approfondi avec l'aspect grammatical que détermine ce thème.

Partie C : Stratégies conversationnelles

La Partie C met l'accent sur le registre et le langage interactionnel. Elle commence par un dialogue qui illustre quelques stratégies de conversation. Dans cette unité, il s'agit de phrases qui servent à exprimer son intérêt pour ce que l'autre personne dit et à faire avancer la conversation. Même si le dialogue lui-même n'est pas authentique, les éléments du langage interactionnel qui y sont intégrés ont été identifiés comme « typiques » dans la partie orale du Cambridge International Corpus (cambridge.org 2021). Les apprenants sont amenés à les observer et à les souligner dans le dialogue avant de les utiliser eux-mêmes. La note au bas du dialogue explique aux apprenants le « rituel » : les locuteurs apportent un commentaire et posent une question. Les apprenants sont invités à trouver dans le dialogue des exemples de cette dynamique (tableau 51).

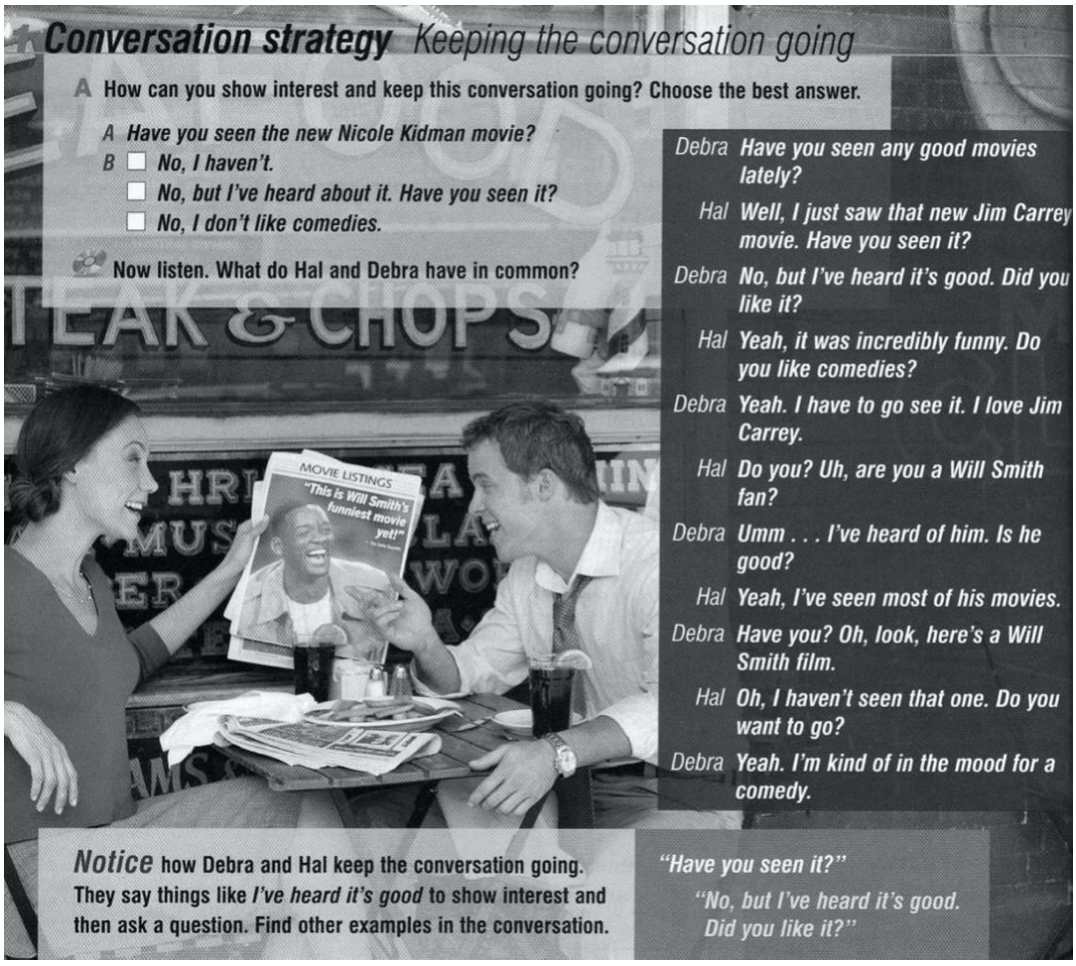


Tableau 51 : Observation du langage interactionnel dans un dialogue (Touchstone A2).

La note fournit des renseignements utiles concernant une manière possible d'aller plus loin dans la conversation et elle donne l'opportunité aux apprenants d'observer cette stratégie dans un dialogue informel.

2.3) Analyse succincte de l'ouvrage

L'unité choisie, issue de « Touchstone 3 », illustre la façon dont les résultats de l'analyse de corpus ont été intégrés dans cette série. Les activités incluses contiennent des précisions sur la fréquence d'usage (ici : les questions les plus courantes commençant par « Have you ... ? »). Elles relient la grammaire, le lexique et les connaissances pragmatiques de façon systématique, tout au long de l'ouvrage, et présentent la grammaire (ici : l'utilisation de deux temps du verbe) au sein d'une approche lexico-grammaticale. L'intégration consciente des éléments du langage oral (avant tout celle du langage interactionnel) ainsi que l'attention portée sur les registres et sur le degré de formalité des éléments présentés, caractérisent la série dans son ensemble.

Certains points auraient mérité, à notre avis, quelques approfondissements. De temps à autre, les apprenants sont invités à étudier des phénomènes langagiers importants, par exemple, ils observent des stratégies conversationnelles et l'utilisation des aspects grammaticaux. Ce genre d'exercice pourrait être introduit plus régulièrement dans l'ouvrage et, ce qui est encore plus important, appliqué au vocabulaire de façon systématique. Par exemple, la question « Avez-vous aussi des rêves secrets ? » (Partie A, exercice 1a), censée initier une brève discussion après la lecture de cinq textes courts, pourrait être *complétée par une liste d'unités multi-lexicales provenant des textes que les apprenants peuvent intégrer dans leurs réponses*. Une telle liste mettrait davantage l'emphase sur l'usage langagier et les débuts de phrases typiquement employées pour introduire la réponse à cette question (« My dream is to » (Mon rêve est de), « I've always wanted to » (J'ai toujours voulu), « I want to » (Je veux)) aideraient l'apprenant à construire des énoncés corrects.

Tous les textes de la série intègrent de nombreuses unités multi-lexicales qui mériteraient d'être présentées et mises en pratique séparément pour mettre en évidence leur importance et pour permettre aux apprenants de les mémoriser. Cette démarche serait d'autant plus justifiée que la recherche démontre que signaler explicitement les éléments méritant une attention particulière, contribue à un apprentissage plus efficace du vocabulaire (Nation 2015 ; Schmidt 1990 ; Schmid 2016 ; Taylor 2012 ; Trofimovich et McDonough 2011, Part II).

3) Des manuels pour le hongrois : la série « MagyarOK »

3.1) Le concept

La série de manuels de hongrois « MagyarOK » (Szita et Pelcz 2013-2019) a été publiée par l'Université de Pécs et comprend quatre volumes qui amènent l'apprenant du niveau débutant (A1) au niveau intermédiaire (B2). C'est une des rares séries conçues pour une langue autre que l'anglais fondées sur l'analyse de l'usage langagier des natifs et dont nous sommes co-auteure. Le contenu linguistique ainsi que l'approche didactique reposent sur les résultats de la linguistique de corpus (Szita 2014 ; Szita et Pelcz 2017 ; Szita et Pelcz à paraître). Les principes les plus importants de cette approche sont les suivants :

- Avant de produire des énoncés, les apprenants doivent toujours *observer des exemples (modèles) authentiques ou semi-authentiques afin qu'ils puissent observer comment les locuteurs natifs sont susceptibles de s'attaquer à la tâche*.
- Les énoncés illustrent des caractéristiques de la grammaire, du lexique, de la pragmatique, de la stylistique, de la prononciation ou de l'intonation, *tous les énoncés présentés dans le matériel doivent avoir un caractère naturel.*
- Le *langage transactionnel et le langage interactionnel* doivent être explorés et pratiqués tout au long de l'apprentissage.
- Enfin, les tâches et activités doivent *offrir aux apprenants la possibilité de reproduire le vocabulaire et les structures présentés dans des contextes significatifs*.

Ces principes servent d'une part à assurer la qualité de l'apport linguistique présenté dans le matériel et, d'autre part, à stimuler les apprenants à appliquer un langage à caractère naturel. La séquence ci-dessous tirée du manuel A2+ présente un exemple concret sur la manière dont les résultats de la linguistique de corpus ont été mis en œuvre.

Une difficulté rencontrée lors de l'écriture de la série a consisté en ce que, contrairement à l'anglais, nous ne disposions d'aucune description détaillée de la langue hongroise basée sur l'analyse de corpus (Szita 2014). Face à l'absence d'une telle description cohérente, détaillée et spécifique à la langue hongroise, les renseignements intégrés dans « MagyarOK » sont loin d'être exhaustifs et une approche lexico-grammaticale systématique n'a pas pu être mise en place dans les ouvrages. En outre, la construction des corpus ciblés (voir la Partie 3 pour plus de détails) ainsi que l'accès au grand corpus général de Sketch Engine et au Corpus national du hongrois nous ont permis d'identifier le vocabulaire-clé par sujet et par situation de communication.

Dans les pages suivantes, nous présenterons une séquence d'activités autour du thème du « week-end » qui illustre la démarche pédagogique appliquée dans les manuels. Comme dans le cas de « Touchstone », après le titre de chaque exercice, nous listerons le ou les résultats pertinents de la linguistique de corpus, présentés au chapitre 4.

3.2) Exemple de séquence informée par le corpus

L'ordre dans lequel cette séquence introduit les éléments linguistiques suit la structure des conversations réelles (voir aussi le chapitre 14 dans la Partie III pour des précisions concernant les conversations observées) :

- Pour qualifier d'une manière générale le week-end, les conversations réelles commencent par un adjectif. Pour cette raison, le premier exercice présente les adjectifs (« super », « bien », « chargé ») fréquemment utilisés pour décrire le week-end.
- Ce premier exercice est suivi par la présentation et la pratique d'un nombre significatif d'unités multi-lexicales décrivant diverses activités. Ces phrases les plus souvent utilisées serviront de modèles pour les énoncés des apprenants.
- Des phrases typiques signalant une réaction (« Vraiment ? » « Tant mieux. » « Pauvre de toi ») sont présentées et pratiquées par la suite. Ce sont les phrases avec lesquelles les locuteurs expriment leur réaction à ce qui a été dit.
- Tous les éléments présentés et pratiqués séparément au cours des exercices précédents, sont rassemblés à la fin de la séquence. Les apprenants sont invités à produire des interactions plus complexes, avec une dynamique analogue aux conversations naturelles qui ont inspiré cette série d'activités.

Exercice 1a : Comment était ton week-end ? Comment était votre week-end ?

Le premier exercice (tableau 52) présente douze adjectifs qui servent de réponses typiques à la question « Comment était ton week-end ? Comment était votre week-end ? ». Les adjectifs ont été sélectionnés en fonction de leur fréquence dans les réponses à cette question dans notre corpus de conversations et de messages sur les réseaux sociaux. Il convient de noter que de nombreux adjectifs font partie d'unités multi-lexicales, précédées d'un modificateur qui intensifie ou affaiblit leur signification. Ainsi, l'adjectif « jó » est inclus trois fois dans la liste, chaque fois avec un autre modificateur : « elég jó » (pas mal), « nagyon jó » (très bien) et « nem valami jó » (moyen). Le texte

dans la case jaune à côté de la photo indique que certains de ces adjectifs sont principalement utilisés dans des contextes informels.

1. Milyen volt a hétvégéd? Milyen volt a hétvégéje?

a) Mi pozitív, mi negatív? Csoportosítsa a szavakat és kifejezéseket!

~~*nagyon jó*~~ *– fárasztó – kellemes – nem valami jó – nem túl izgalmas – elég jó – egy kicsit stresszes – elég unalmas – átlagos – nagyon klassz – szuper – borzasztó – rettenetes – szörnyű*

szuper, nagyon klassz
borzasztó, szörnyű, rettenetes
→ **beszélt nyelv**

(1. Comment était ton week-end ? Comment était votre week-end ?

a) Qu'est-ce qui est positif et qu'est-ce qui est négatif ? Faites des groupes.

très bien - fatigant / chargé - agréable – moyen – pas trop excitant – pas mal - un peu stressant - assez ennuyeux - normal - vraiment cool - super - horrible - terrible – affreux)

Tableau 52 : Adjectifs (avec et sans modificateur) répondant à la question « Comment était ton week-end ? » « Comment était votre week-end ? » (MagyarOK A2+).

Dans cet exercice (tableau 52), les apprenants sont invités à répartir les mots et les phrases en deux groupes selon leur signification positive ou négative. Ils observent ensuite les adjectifs qui apparaissent avec un modificateur et la fonction du modificateur. Enfin, le schéma prosodique de ces expressions est observé : l'accent est toujours mis sur la première syllabe, les unités sont prononcées comme un mot.

Exercice 1b : Lire la liste des réponses possibles

Dans l'exercice suivant (tableau 53), dix-neuf phrases adaptées de notre base de données présentent des activités typiques du week-end. Voici la liste :

Nem csináltam semmi különöset.
Beszélgettem a barátnőmmel.
Telefonáltam a szüleimnek.
Az egész hétvégét a családommal töltöttem.
Találkoztam egy ismerősömmel.
Elmentem a barátaimhoz.
Végre kitakarítottam a lakást.
Bevásároltam a piacon.
Segítettem a fiamnak leckét írni.
Megírtam a magyar házi feladatot.
Megtanultam az új szavakat.
E-maileket írtam.
Végre kipihentem magamat.
Moziban / színházban / egy buliban / angolórán / rokonoknál / … voltam.
Úszni / táncolni / futni / … voltam.
Egyáltalán nem tudtam pihenni.
Sportolni is akartam, de nem volt rá időm.
Nem volt időm takarítani / vásárolni / …
Nem voltam sehol.

(Je n'ai rien fait de particulier. J'ai discuté avec mon amie. J'ai appelé mes parents. J'ai passé le week-end avec ma famille. J'ai rencontré un ami. Je suis allé voir mes amis. Finalement, j'ai réussi à faire le ménage. Je suis allé faire des courses au marché. J'ai aidé mon fils à faire ses devoirs. J'ai fait mes devoirs pour le cours de hongrois. J'ai appris les nouveaux mots. J'ai écrit des e-mails. J'ai enfin pu me reposer. J'étais au cinéma / au théâtre / à une fête / à un cours d'anglais. Je suis allé nager / danser / faire du jogging. Je ne pouvais pas me reposer du tout. Je voulais faire du sport mais je n'avais pas le temps. Je n'ai pas eu le temps de faire le ménage / faire des courses. Je ne suis allé nulle part.)

Tableau 53 : Réponses-modèles à la question « Comment était ton week-end ? » (MagyarOK A2+).

Tout d'abord, les apprenants étudient ces phrases : ils les lisent, écoutent et s'exercent à la prononciation et à l'intonation. *Cette liste sert également de « lignes de concordance » révélant des schémas structurels concernant la position typique du verbe conjugué*[64]. Comme le hongrois a des règles assez complexes pour l'ordre des mots, de telles observations peuvent être particulièrement utiles. Le tableau suivant montre la position du verbe (en rouge) et des mots qui le précèdent généralement (en violet) (tableau 54).

Tableau 54 : Lignes de concordance dans le manuel MagyarOK A2+. Les phrases autour des activités du week-end font émerger l'ordre de mots typiques dans des phrases répondant à la question : « Qu'est-ce que tu as fait ? Qu'est-ce que vous avez fait ? » (MagyarOK A2+).

Guidés par l'enseignant, les apprenants peuvent découvrir que le verbe conjugué est généralement placé au début de la phrase ou de la proposition et est ainsi mis en évidence. Cette position se

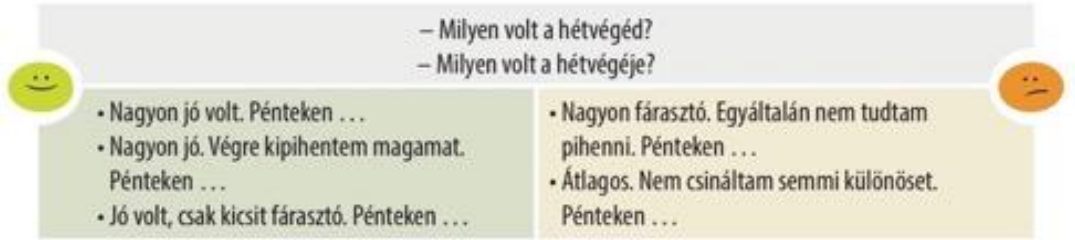
[64] Ce genre de présentation est préconisé, par exemple, par Boulton (2008).

justifie logiquement car dans une phrase répondant à la question « Qu'est-ce que tu as fait ? Qu'est-ce que vous avez fait ? », l'action (donc le verbe) est la partie la plus importante. Les verbes qui ont tendance à éviter l'emphase tels que le verbe « voltam » (j'étais) et « akartam » (je voulais) sont placés vers la fin de la phrase. Les apprenants peuvent également noter que seuls certains types de mots précèdent le verbe. Ce sont des modificateurs de phrases tels que « enfin » (végre) et des mots négatifs tels que « nem » (non, ne), « egyáltalán nem » (pas du tout). Une fois ces phrases analysées, les apprenants peuvent ajouter de nouvelles phrases à la liste et mémoriser celles qui s'appliquent à eux.

Des listes similaires qui attirent l'attention des apprenants sur les répétitions lexicales, structurelles et/ou prosodiques, font partie de nombreux exercices. En les étudiant, *les apprenants se familiarisent avec le concept de schémas et avec la lecture des lignes de concordance avant de travailler avec un véritable corpus.*

Exercice 1c : Parlez à deux partenaires. Utilisez les phrases de l'exercice b.

L'exercice suivant permet aux apprenants de réviser et de mettre en pratique le vocabulaire des exercices a) et b) en menant de brèves conversations sur leur week-end. L'accent est mis sur l'interconnexion du langage interactionnel et du langage transactionnel. L'exercice présente également quelques phrases typiques tirées de notre corpus oral qui peuvent servir à engager de telles conversations. Lors de la formulation de leurs propres énoncés, les apprenants peuvent s'inspirer des réponses-modèles fournies : ils peuvent les utiliser telles quelles ou les adapter selon leurs besoins (tableau 55).



(Comment était ton week-end ? Comment était votre week-end ?

☺ : *Génial. Vendredi, j'ai … / Très bien. Enfin, je me suis bien reposé(e). Vendredi, j'ai … / C'était bien, mais un peu fatigant. Vendredi, j'ai …*

☹ : *Très chargé. Je n'ai pas du tout eu le temps de me reposer. Vendredi, j'ai … / Normal. Je n'ai rien fait de spécial. Vendredi, j'ai …)*

Tableau 55 : Exemples de dialogues sur le week-end (MagyarOK A2+).

Puisque la fréquence avec laquelle les mots, les expressions ou les schémas grammaticaux apparaissent dans l'apport linguistique, semble avoir un impact sur leur acquisition (Ellis 2002 ;

Hoey 2005 ; Meunier 2012 ; Schmid 2016 ; Taylor 2012), il est primordial que les apprenants soient exposés plus d'une fois à des composants pertinents et aient de nombreuses occasions de les pratiquer. La répétition consciente dans le contexte d'une communication, comme le suggère Ellis (2002), peut être un moyen efficace de consolider les éléments langagiers importants. Néanmoins, la répétition ne signifie pas dans ce manuel que les apprenants sont censés produire les mêmes énoncés, exercice après exercice. En l'espèce, les tâches du manuel incluent des modifications et des variations des composants déjà connus, permettant à l'apprenant de consolider et d'élargir ses connaissances.

L'exercice 1c) est suivi de quelques activités permettant aux apprenants d'identifier les schémas de la conjugaison du passé (le hongrois n'en possède heureusement qu'un seul) sur la base d'un certain nombre d'exemples.

Exercice 2 : Tibor a fait une randonnée dans les montagnes

L'exercice 2 présente un dialogue plus long, adaptation d'une conversation enregistrée dans laquelle un jeune homme, Tibor raconte son dimanche à une amie. Le vocabulaire-clé (noms des activités) est introduit avec des dessins illustrant la journée de Tibor : il est parti en randonnée dans les bois, s'est perdu et a retrouvé son chemin, est arrivé chez lui épuisé et a appelé son amie pour lui raconter ses aventures. Cet exercice préparatoire est suivi d'une conversation informelle entre Tibor et son amie. Les apprenants peuvent d'abord écouter la conversation, puis compléter les verbes au passé dans le texte (tableau 56).

~~*kirándul*~~ *– elindul – pihen – megtalál – gyalogol – elkészül – gyalogol – megáll – bepakol – jár*

Lea: Már nagyon vártam, hogy hívsz. Mi újság? Mit csináltál ma?

Tibor: Kirándultam az erdőben. Későn keltem fel, de utána gyorsan .... a hátizsákomba a szendvicseket és a vizet, és ..... . Csak a térkép maradt itthon.

Lea: Hű, és nem tévedtél el?

Tibor: Dehogynem! Kétszer is! Szerencsére végül mindig ..... az utat. Egy kicsit többet ....., mint akartam, de nem baj. ..... néhány barlangban is. Ahol pedig ..... ebédelni, találtam egy-két ritka növényt.

Lea: Lefényképezted őket?

Tibor: Persze! Tudod, mennyire érdekel a botanika.

Lea: Igen, tudom. És mikor értél haza?

Tibor: Egy órával ezelőtt. Legalább 20 kilométert ..... .

Lea: Akkor biztosan nagyon elfáradtál.

Tibor: Igen. De ..... egy kicsit, miután hazaértem. Most már egyáltalán nem vagyok fáradt.

(Lea : Il me tardait que tu appelles. Comment vas-tu ? Qu'est-ce que tu as fait aujourd'hui ?

Tibor : J'ai fait de la randonnée dans les bois. Je me suis levé tard mais je me suis préparé très vite. J'ai mis des sandwichs et une bouteille d'eau dans mon sac à dos et je suis parti. Seulement, j'ai oublié de prendre une carte.
Lea : Aïe. Et tu ne t'es pas perdu ?
Tibor : Bien sûr que si. À deux reprises. Heureusement, je pouvais toujours savoir où j'étais. J'ai fait un peu plus de randonnée que prévu mais ça va. J'ai aussi trouvé quelques grottes. Et là où je me suis assis pour manger, j'ai trouvé des plantes rares.
Lea : Tu en a pris des photos ?
Tibor : Bien sûr. Vous savez à quel point je suis intéressé par la botanique.
Lea : Je sais… Quand es-tu rentré chez toi ?
Tibor : Il y a environ une heure. J'ai parcouru au moins 20 kilomètres.
Lea : Wow. Tu dois être bien fatigué maintenant.
Tibor : Ouais, je l'étais quand je suis rentré à la maison. Mais je me suis un peu reposé et maintenant je me sens bien.)

Tableau 56 : Langage transactionnel et interactionnel dans un dialogue sur le week-end (MagyarOK A2+).

Comme les apprenants ont déjà été introduits à l'histoire dans l'exercice a), ils peuvent désormais se concentrer sur le langage interactionnel qui donne un caractère naturel au dialogue. Ils peuvent observer les expressions qui assurent la fluidité de la conversation (« Tu dois être très fatigué. » « Aïe. » « Bien sûr. » « Tu sais ... », « Wow »), dont beaucoup ont été intégrées sans modification à partir de la conversation enregistrée. Cette tâche est suivie d'exercices de vocabulaire, de grammaire et de prononciation (non inclus ici) offrant aux apprenants la possibilité de consolider les connaissances nouvelles.

Exercice 5a : Randonneuse en bottes à talons hauts

Cet exercice est particulièrement intéressant du point de vue des registres : il présente une histoire semblable à celle de Tibor (présentée dans les exercices précédents), telle qu'elle a été publiée dans un journal. Avant de lire un article de journal court (80 mots), les apprenants étudient les éléments-clés du texte dans un diagramme Wordle (la taille de chacun indiquant sa fréquence relative). Il s'agit des mots simples et des unités multi-lexicales : « a perdu son chemin », « a demandé de l'aide », « le bon chemin », « 9 heures du soir », « la jeune femme », « en tenue de ville », « talons hauts », « dix kilomètres »[65] (tableau 57).

[65] Le hongrois étant une langue agglutinante, de nombreuses constructions multi-lexicales du français sont exprimées en mots uniques en ajoutant des suffixes. Par exemple, à la police / rendőrségre, sur la route / úton.

Tableau 57 : Mots-clés et expressions-clés dans le texte sur la randonneuse en bottes à talons hauts (MagyarOK A2+).

Ce type de visualisation met l'accent sur l'importance des unités multi-lexicales dans l'usage langagier naturel et attirent l'attention des apprenants sur celles-ci (Robinson et al. 2012). Après avoir étudié le diagramme, ils formulent des hypothèses à propos du contenu possible de l'article. Des phrases fréquemment utilisées pour discuter des suppositions (« Peut-être. » « Je dirais… » « je suppose… » « Tu as peut-être raison. » « Qu'est-ce que tu en penses ? ») peuvent être intégrées dans la discussion, donnant aux apprenants l'opportunité de pratiquer le langage interactionnel.

Les apprenants vérifient leurs hypothèses en lisant l'article : une jeune femme portant des bottes à talons hauts se dirigeait vers un ravin, s'est perdue dans les bois et a marché 10 kilomètres. Elle a finalement appelé la police mais elle n'a pas pu leur expliquer où elle se trouvait. Heureusement, le policier connaissait bien le bois et a finalement réussi à retrouver la jeune femme. Dans ce court article, les apprenants voient les mots et les expressions du diagramme Wordle avec leur co-texte ; ils peuvent observer leur position textuelle ainsi que la distance entre les répétitions d'éléments lexicaux utilisés plus d'une fois. En outre, ils peuvent remarquer quels éléments servent de dispositifs de cohésion garantissant que les différentes parties du texte sont liées les unes aux autres.

Exercice 5b : Interview avec la fille en bottes à talons hauts

Les exercices suivants offrent la possibilité de consolider le vocabulaire-clé du texte. Pourtant, ce n'est pas leur seul objectif : ils ajoutent du langage interactionnel au récit. Le texte est d'abord transformé en interview formelle, puis en conversation informelle avec la jeune femme, présentant et pratiquant ainsi deux registres différents. Alors que l'interview se concentre sur l'utilisation du

langage formel et sur les informations incluses dans le texte, le dialogue donne aux apprenants l'occasion d'utiliser des réactions informelles appropriées (« Pauvre de toi », « Oh non »). La répétition du texte selon des variations de registre offre à l'apprenant des opportunités de comparer l'usage langagier dans différents registres (tableau 58).

9. Riport a magas sarkú csizmás lánnyal

a) Vajon mit válaszol a lány a riporter kérdéseire? Használja az információkat a szövegből!

Riporter: Miért akart elmenni a szakadékhoz?
Lány: Nem is tudom. Spontán indultam el. Nem gondoltam, hogy útközben eltévedek.
Riporter: Nem volt térképe?
Lány: .....
Riporter: Hogyan találta meg végül a helyes utat?
Lány: .....
Riporter: Sokat gyalogolt?
Lány: .....
Riporter: Elfáradt?
Lány: .....

(a) Comment pensez-vous que la fille a répondu aux questions du journaliste ? Utilisez les informations du texte. Pourquoi vouliez-vous aller vers le ravin ? / Vous n'aviez pas de carte ? / Comment avez-vous finalement trouvé votre chemin ? / Avez-vous beaucoup marché ? / Étiez-vous fatiguée à la fin ?)

Tableau 58 : Transformation d'un texte narratif à un dialogue (MagyarOK A2+).

Ces exercices sont suivis de récits et de dialogues de différentes longueurs et de différents degrés de formalité. Les textes sont tous des adaptations issues de notre corpus et présentent le week-end de trois autres personnes. La logique sous-tendant le choix de proposer plusieurs textes sur le même sujet est que, *même s'il est impossible de prédire le déroulement d'interactions entières, les éléments-clés sont répétés et variés avec suffisamment de fréquence pour qu'ils puissent être notés et mémorisés*. De plus, le caractère naturel de l'utilisation de la langue par les apprenants peut être considérablement amélioré s'ils sont *exposés aux éléments-clés dans divers contextes avant de les adapter à leur propre situation.* Les apprenants sont, en pratique, plus disposés à prendre des risques pour exprimer leurs pensées s'ils savent que leurs énoncés seront, au moins en partie, corrects (Robinson et Ellis 2008).

La séquence présentée se termine par une tâche écrite. Les apprenants sont encouragés à « recycler » autant de mots et d'unités multi-lexicales issues des textes qu'ils le souhaitent pour évoquer un week-end de leur choix. L'intégration de ces éléments dans un récit personnel plus long leur offre l'occasion de reprendre le lexique et la grammaire abordés dans le chapitre tout en effectuant une

tâche pertinente. Une fois le texte corrigé, il peut être utilisé comme base pour des conversations plus longues et plus complexes : les apprenants peuvent engager des dialogues sur leur week-end en utilisant le contenu de leurs textes et en y ajoutant du langage interactif. Ils peuvent enfin suivre la dynamique des conversations enregistrées pour produire des conversations à caractère naturel.

3.3) La présentation de la grammaire

Nous étudierons à présent brièvement la manière dont la grammaire est présentée dans l'ouvrage. À la suite de certains exercices du livre de cours, des liens indiquent la page du cahier de grammaire qui leur est associée ; nous pouvons ainsi observer que la séparation de la grammaire ne correspond pas à l'approche lexico-grammaticale mais ce choix peut s'expliquer par le fait que le hongrois n'a pas encore de description lexico-grammaticale, comme évoqué précédemment. Le chapitre 4 systématise, entre autres, la forme du passé en présentant cet aspect de la grammaire de façon plutôt classique.

Le premier tableau (tableau 59a) montre les quatre groupes de verbes et donne quelques informations supplémentaires concernant la forme. Ce tableau est suivi d'un second (tableau 59b) montrant les terminaisons pour le premier groupe selon le type de conjugaison (conjugaison indéfinie et définie). La présentation souligne les aspects morphologiques en mettant en contexte l'utilisation des conjugaisons en rajoutant des compléments d'objet direct possibles.

FORMA
igető + *-t/-tt (-ott/-ett/-ött)* + személyrag

			
Ő Tibor.	Ő Ágota-Marietta.	Ő Ottó.	Ő pedig Ivett.
Tibor tegnap kirándult. Legalább húsz kilométert gyalogolt. Utána pihent.	Ágota-Marietta a reptéren megismerkedett egy kedves férfival. Amíg várakoztak, beszélgettek.	Ottó tortát sütött, és takarított. Elfelejtett jelentkezni az Ultrabalatonra…	Ivett tegnap egész nap otthon volt. Evett és ivott. Nem ment el bevásárolni.
-t	*-t/-tt (-ott/-ett/-ött)*	*-tt (-ott/-ett/-ött)*	**Rendhagyó igék**

- A magyarban csak egy múlt idő van. Jele: *-t/-tt (-ott/-ett/-ött)*.
- Mindegyik igecsoport ugyanazokat a személyragokat kapja.
- Az egyes szám első személy ragja mindig *-m*.

*(**Forme** : radical + -t/-tt (-ott/-ett/-ött) + terminasion du verbe*

1er groupe : C'est Tibor. Tibor a fait une randonnée hier. Il a marché au moins 20 kilomètres. Il s'est reposé ensuite. 2e groupe : C'est Ágota-Marietta. Ágota-Marietta a rencontré un homme sympatique à l'aéroport. Ils ont discuté en attendant l'avion. 3e groupe : C'est Ottó. Ottó a fait un gâteau et il a fait le ménage. Il a oublié de s'inscrire pour la course Ultrabalaton. 4e groupe : C'est Ivett. Ivett était à la maison toute la journée. Elle a mangé et elle a bu. Elle n'est pas allée faire les courses.)

Tableau 59a : La présentation du passé au niveau A2. (MagyarOK A2, cahier d'exercices)

Első csoport: *-t*

	Határozatlan ragozás		Határozott ragozás	
én	vártam	kértem	vártam	kértem
te	vártál	kértél	vártad	kérted
Ön ő	várt	kért	várta	kérte
mi	vártunk	kértünk	vártuk	kértük
ti	vártatok	kértetek	vártátok	kértétek
Önök ők	vártak	kértek	várták	kérték
	egy lányt	egy fagylaltot	Laurát	a jegyeket
én ⟶ téged, titeket	vártalak	kértelek		

Mikor *-t*?

- *-j, -l, -r, -m, -n* és *-ny* végű igék. Leggyakoribb az *-l, -r* és az *-n: ül, beszél, telefonál, vár, kér, akar, pihen, kíván* (Memoriter: *Jár a lányom.*)
- Sok *-ad/-ed* végű ige: *marad, elfárad, felébred*
- Néhány más ige: *alszik (aludni)* ⟶ *aludtam, aludtál, aludt; fekszik (feküdni)* ⟶ *feküdtem, feküdtél, feküdt*

(*Premier groupe : le tableau montre les deux conjugaisons. Texte à côté du tableau (sans la traduction des exemples) : Quand utilise-t-on un -t ? – Après j, l, r, m, n. Le l et le r sont les terminaisons les plus fréquentes. – Beaucoup de verbes finissant par -ad/-ed. – Quelques autres verbes.*)

Tableau 59b. La présentation de la conjugaison au passé (MagyarOK A2+, cahier d'exercices).

Ce sont là les exercices qui mettent en valeur le travail sur l'interconnexion du lexique et de la grammaire en reprenant les unités multi-lexicales importantes des textes du manuel, en mettant cette fois-ci l'emphase sur la forme. Par exemple, l'exercice 3 (tableau 60a) contient des expressions du texte sur Tibor telles que « későn kelt fel » (il s'est levé tard), « szkájpolt a barátnőjével » (il a parlé avec sa copine par Skype) ainsi que certaines phrases de l'exercice 1b du livre de cours (voir le tableau 53), et complète ce vocabulaire par d'autres unités multi-lexicales utiles comme « tanult a következő vizsgájára » (il a préparé son prochain examen), « kisétált a Tisza-partra » (il est allé /en se promenant/ sur le bord de la rivière Tisza).

3. Határozatlan vagy határozott ragozás? Írja be az igéket!

Gábor reggel elég későn kelt fel ..... *(felkel)*. Aztán ..... *(szkájpol)* a barátnőjével. ..... *(elmesél)* neki (azt), hogy mit ..... *(csinál)* előző nap. Utána ..... *(tanul)* a következő vizsgájára. Ebédre a kedvenc pizzáját ..... *(rendel)*. Délután ..... *(telefonál)* a barátjának. Fél négykor ..... *(kisétál)* a Tisza-partra, és a barátját ..... *(vár)*.

(Gábor s'est levé assez tard ce matin. Il a ensuite discuté avec sa copine par Skype. Il lui a raconté ce qu'il avait fait la veille. Puis, il a préparé son prochain examen. Pour le déjeuner, il a commandé sa pizza préférée. Il a appelé son ami dans l'après-midi. À 3h30, il est allé au bord de la rivière Tisza et a attendu son ami.)

Tableau 60a : Exercice de grammaire utilisant des unités multi-lexicales thématiques (quelques activités quotidiennes d'un étudiant) (MagyarOK A2+, cahier d'exercices).

L'exercice suivant (tableau 60b) reprend en partie le vocabulaire de l'article sur la randonneuse pour raconter une histoire légèrement humoristique que les apprenants peuvent lire et écouter. Leur tâche est d'observer et de souligner les verbes et d'identifier par la suite ceux qui sont au passé.

4. Melinda kirándult

a) Olvassa el a szöveget, és húzza alá az igéket! Melyik ige jelen idejű, és melyik múlt idejű? A szöveget meg is hallgathatja.

Egész életemben utáltam túrázni. De az egyik barátom tegnap egész este arról beszélt, hogy milyen szép a természet. Gondoltam, megnézem. Nehéz elhinni, de kirándultam hétvégén! Nem ültem a számítógép előtt egész szombaton, nem telefonáltam senkinek, és nem jártam a városban. Én csináltam a szendvicseket, én pakoltam be a hátizsákba! Rengeteget gyalogoltam az erdőben a sok egyforma fa között, ráadásul eltévedtem. Nagyon féltem… Soha többé nem megyek ki a városból! ☺

J'ai détesté les randonnées toute ma vie. Mais hier soir, un de mes amis n'arrêtait pas de répéter comme la nature était belle. Je me suis dit que je la regarderais moi-même. C'est difficile à croire mais j'ai fait une randonnée ce week-end. Je n'étais pas assise devant l'ordinateur tout samedi, je n'ai appelé personne et je ne suis pas allée en ville. C'est moi qui ai fait les sandwichs, c'est moi qui ai fait mon sac à dos ! J'ai marché et marché parmi les arbres qui avaient l'air tous pareils, et je me suis en plus perdue. J'avais très peur. Je ne quitterai plus jamais la ville ! ☺

Tableau 60b : Exercice d'observation des formes du verbe dans un texte thématique (MagyarOK A2+, cahier d'exercices).

Dans la partie b) (non inclus ici), l'apprenant est invité à poser des questions à partir des réponses fournies.

Ces réponses incluent des tournures typiques comme « Persze, hogy » (Bien sûr que), « Nem sokat: rengeteget! » (Pas juste beaucoup, mais énormément !) qui servent à mettre de l'émotion dans la réponse et à rendre la conversation plus naturelle. L'intégration du phénomène à étudier dans une situation de communication (au lieu de faire pratiquer des phrases isolées) caractérise la plupart des exercices et peut compenser en partie l'accent particulier porté sur la présentation plutôt classique de l'aspect grammatical dans les tableaux. Une analyse empirique, approfondie de la

grammaire hongroise pourrait fournir des précisions concernant l'utilisation de certains aspects langagiers[66].

3.4) Analyse succincte de l'ouvrage

L'approche adoptée dans les manuels de la série « MagyarOK » montre certaines similitudes avec celle appliquée pour la série « Touchstone » et se caractérise par les points forts suivants :

- Comme dans le cas des manuels d'anglais, le langage à caractère naturel constitue une pierre angulaire du matériel pédagogique.
- Les textes (dialogues et narratifs écrits et oraux) sont des adaptations des textes authentiques permettant aux apprenants d'observer les habitudes langagières liées à des situations communicationnelles diverses, pertinentes au niveau de compétences linguistiques de l'apprenant.
- L'emphase sur les unités multi-lexicales, le travail intensif sur le vocabulaire et la présentation cyclique des phénomènes linguistiques s'inscrivent dans le cadre d'une approche intégrant les résultats de la recherche linguistique.
- Les unités multi-lexicales tirées des textes sont présentées séparément (diagrammes Wordle, listes, modèles de dialogues) pour attirer l'attention de l'apprenant sur leur importance.
- Un rôle essentiel revient à la répétition et à la variation linguistiques : les apprenants rencontrent des textes à contenu similaire (il suffit de penser à l'article sur la randonneuse suivi d'une interview et d'une conversation informelle ou encore au narratif sur Tibor suivi d'un dialogue informel ou même aux textes et aux dialogues sur le week-end des apprenants). Ces textes répètent le même vocabulaire-clé relatif au sujet (langage transactionnel) tout en variant « ce qui l'entoure » selon la situation de communication (langage interactionnel). Cette approche offre, d'une part, des rencontres répétées avec les unités multi-lexicales fréquentes que les apprenants sont censés maîtriser au niveau donné et, d'autre part, des opportunités d'observation et de pratique du langage interactionnel.

À la différence de la série « Touchstone », la série « MagyarOK » ne contient pas d'informations explicites sur la fréquence d'usage des éléments inclus : c'est par leur répétition cyclique et par leur intégration dans plusieurs exercices successifs que l'importance de ces éléments devient visible.

[66] La Partie II de cette thèse sera consacrée dans son intégralité à quelques points lexico-grammaticaux concrets qui pourraient être présentés de façon plus efficace grâce à une analyse de corpus du hongrois.

Cette démarche remplit la fonction de sensibilisation du fait que certains éléments dans l'usage langagier des natifs sont plus usités que d'autres dans la situation donnée ; des informations sur la fréquence pourraient cependant aider les apprenants à hiérarchiser les éléments à apprendre.

Notons également que les manuels de « MagyarOK » donnent peu d'informations explicites sur les habitudes langagières orales alors que la recherche démontre que les apprenants ne les remarquent pas nécessairement à moins d'attirer directement leur attention sur ce point (cf. Ellis 2006, 2008 ; Schmidt 1990, 2010 ; Robinson 1995). L'intégration des expressions du langage parlé dans les exercices peut contribuer à sensibiliser l'apprenant sur leur importance ainsi qu'à faciliter leur mémorisation ; il pourrait cependant être utile de signaler explicitement la présence de ces éléments pour que l'apprenant puisse les identifier en tant que tels (et éviter, par exemple, leur utilisation à l'écrit).

Il convient de remarquer encore que, bien que les exercices suivant la présentation de la grammaire intègrent des unités multi-lexicales importantes, la présentation elle-même reste plutôt classique. Les tableaux se concentrent sur les aspects morphologiques (démarche par ailleurs compréhensible dans la mesure où le hongrois est une langue à cet égard bien plus complexe que l'anglais), et les informations supplémentaires concernent principalement la forme et, parfois, la fonction des éléments grammaticaux présentés. La raison en est, comme évoqué précédemment, que, pour l'heure, nous ne disposons pas d'une description détaillée de la grammaire hongroise fondée sur des corpus. Ce sont les exercices qui offrent l'opportunité d'intégrer des éléments lexicaux utiles à un aspect grammatical donné. Ces exercices sont par ailleurs identiques à ceux du livre de cours en ce qu'ils contiennent différents types de textes.

B) Caractéristiques principales des manuels informés par le corpus

Dans ce chapitre nous avons présenté deux manuels de cours intégrant les résultats en linguistique de corpus exposés au chapitre 4. Les principales caractéristiques de ces ouvrages sont les suivantes :

- Ils présentent du matériel linguistique qui s'appuie sur l'analyse de corpus et, lorsque cela est possible, sur une description grammaticale de la langue basée sur le corpus.
- Ces ouvrages proposent des textes semi-authentiques, à caractère naturel qui préparent l'étudiant à travailler avec des textes authentiques dans les phases ultérieures de son apprentissage.

- Ils mettent en évidence les unités multi-lexicales importantes (par exemple celles liées au sujet traité dans le chapitre et celles fréquemment utilisées indépendamment du sujet).
- Ils répètent et varient les unités multi-lexicales essentielles pour que l'apprenant ait la possibilité de les repérer, de les utiliser et de les mémoriser.
- Les manuels proposent des exercices permettant de pratiquer les aspects grammaticaux en intégrant le lexique pertinent et des exercices de vocabulaire qui incluent les aspects grammaticaux typiquement liés au sujet et/ou à la situation de communication en question.
- Ils exposent les différents aspects de la langue en mettant en évidence ceux présentant un intérêt particulier en raison de leur fréquence.
- Ils intègrent le langage interactionnel dans le matériel pédagogique et soulignent son importance (explicitement ou en guise d'exercices appropriés).
- Les ouvrages portent une attention particulière sur les spécificités de l'usage oral et sur les stratégies conversationnelles qui, pour les niveaux de compétences linguistiques inférieurs, sont du plus grand intérêt.
- Ils fournissent (1) des informations statistiques sur l'usage langagier pour guider l'apprenant vers les éléments particulièrement utiles à retenir, (2) des informations relatives au registre langagier en question et (3) des exercices qui invitent les apprenants à consulter des corpus pédagogiques.

En guise de conclusion, nous pouvons donc constater que *les manuels proposés intègrent les résultats de la recherche linguistique dans leur présentation du langage ainsi que dans leur méthodologie. Ils ne contiennent cependant pas d'exercices qui encourageraient les apprenants à consulter des corpus et à effectuer par eux-mêmes des observations*. Or, cette utilisation active par l'apprenant pourrait représenter un prolongement de la méthodologie en offrant d'autres occasions d'étudier le comportement des éléments lexicaux et grammaticaux importants. Il serait cependant nécessaire que les apprenants aient accès à cette fin à des corpus pédagogiques pour les niveaux de compétences linguistiques inférieurs. La lecture des lignes de concordance et l'utilisation d'autres outils numériques peuvent contribuer à l'approfondissement du ou des sens de l'élément lexical et de l'aspect grammatical choisis. L'intégration de ce travail pourrait augmenter le nombre de rencontres avec ces éléments dans des environnements textuels que l'apprenant est à même de comprendre. *Pour cela, comme dans le cas des grammaires pédagogiques, nous aurions besoin des corpus appropriés, accessibles aux apprenants*. Les manuels pourraient également bénéficier des *grammaires pédagogiques fondées sur corpus*, ouvrages qui manquent encore pour de nombreuses langues, dont le hongrois.

Résumé de la Partie I

La Partie I de cette thèse a été dédiée aux résultats de la linguistique de corpus et à leurs applications possibles pour la didactique des langues. Les chapitres 1 à 3 ont présenté le domaine de la linguistique de corpus ainsi que les outils de recherche dans ce domaine et les différents types de corpus. Le chapitre 4 a exposé un état des lieux des résultats pertinents pour le cadre pédagogique. Puisque recenser tous les résultats aurait largement dépassé le cadre de cette thèse, nous avons mis en exergue ceux qui ont une importance particulière pour l'enseignement des langues. Pour démontrer les possibilités de l'implémentation concrète dans ce cadre, les chapitres 5 et 6 ont examiné l'intégration de ces résultats dans des ouvrages pédagogiques (grammaires et manuels de cours) sélectionnés.

La linguistique de corpus repose sur une approche empirique de nature à la fois quantitative et qualitative. Ces études reposent sur la consultation des grandes bases de données linguistiques formées d'énoncés authentiques, prononcés dans des dialogues ou dans des narratifs réels. Les phénomènes linguistiques ne sont jamais étudiés de manière isolée mais toujours dans leur interrelation avec l'environnement textuel. Les outils numériques utilisés pour l'analyse peuvent fournir des renseignements statistiques pertinents sur la fréquence d'usage de l'élément choisi ainsi que de ces collocatifs.

L'avantage d'une telle approche pour l'enseignement consiste, avant tout, en une description plus précise de la langue-cible basée sur les énoncés réels, donc de l'usage langagier que l'apprenant doit maîtriser. L'intégration de ces résultats[67] est susceptible d'enrichir significativement les ouvrages à fins pédagogiques, comme nous l'avons vu aux chapitres 5 et 6.

Certes, « l'approche corpus », comme toutes les approches, a ses limites, même si actuellement, elle est la plus apte à témoigner de l'expérience langagière probable des natifs. Un corpus ne peut pas contenir d'évidence négative (il ne peut pas montrer ce qu'il ne contient pas) ni renfermer tous les énoncés imaginables que les locuteurs pourraient produire avec l'élément étudié. Ainsi, les résultats fondés sur une telle approche empirique doivent être systématiquement vérifiés, validés, précisés ou écartés par le chercheur.

[67] Par exemple, la présentation de l'environnement typique des éléments fréquemment utilisés, d'un grand nombre d'unités multi-lexicales, des spécificités des registres et des schémas langagiers.

Le chapitre 2 a été dévolu à la comparaison des corpus pédagogiques et linguistiques. Nous avons démontré que les corpus construits à des fins linguistiques ne conviennent pas nécessairement dans le cadre pédagogique, en particulier pour un usage aux niveaux de compétences linguistiques inférieurs. Nous avons argumenté que, pour bien servir les apprenants et les enseignants, il est utile de construire de nouveaux corpus qui tiennent compte de leurs besoins. Ces corpus doivent remplir trois critères essentiels : (1) ils doivent être constitués de données authentiques ; (2) leur contenu doit être pertinent pour les apprenants et (3) leur niveau langagier doit leur être accessible.

À la suite de cet état des lieux de la littérature et à l'analyse de quelques ouvrages pédagogiques, les Parties II et III de cette thèse se déclineront le long de trois axes. Ces parties portent sur les questions suivantes :

- Le recensement des connaissances sur l'usage langagier des natifs pour les intégrer dans la pratique pédagogique.
- Les méthodes adéquates de l'analyse langagière et la présentation des résultats dans le cadre pédagogique.
- La construction et l'exploration des corpus pédagogiques.

Nous prendrons pour exemple la langue hongroise pour laquelle la construction des corpus à fins pédagogiques est l'un des objets de ce travail. Néanmoins, les considérations exposées et les suggestions concrètes peuvent s'appliquer à l'enseignement et à l'apprentissage d'autres langues.

PARTIE II : Les corpus au service des enseignants

Introduction à la Partie II

Cette partie de la thèse est consacrée aux explorations de la langue hongroise fondées sur deux grands corpus – le corpus « huTenTen12 » de Sketch Engine et le Corpus national du hongrois – ainsi que sur des corpus à fins pédagogiques complétant la série de manuels de hongrois « MagyarOK ». Nous examinerons la manière dont l'utilisation des corpus peut enrichir la boîte-à-outils de l'enseignant et comment elle lui permet de trouver des réponses plus pertinentes aux questions des apprenants. La formation des professeurs de hongrois comme langue étrangère ne contient pas de module sur l'utilisation des corpus. Par conséquent, les corpus ne font pas partie des outils habituels des enseignants[68]. L'intérêt d'une approche corpus qui met en relation les différentes dimensions langagières (grammaire, lexique, sémantique et pragmatique) et sert de passerelle entre elles, a été annoncé au chapitre 4. Nous tâcherons de démontrer dans cette partie que la force descriptive d'une méthodologie qui relie intentionnellement les différentes dimensions de la langue (grammaire, lexique, sémantique et pragmatique) peut être, dans certains cas, supérieure à une description plus compartimentée pour une langue morphologiquement aussi complexe que le hongrois. Il s'agit, avant tout, des phénomènes langagiers considérés comme « difficiles » car échappant aux règles simples et évidentes.

Il convient de noter ici qu'aucun travail n'a jamais été effectué jusqu'à présent sur les possibilités qu'offrent les corpus pour l'enseignement du hongrois. L'approche corpus exposée dans cette partie sert, avant tout, à aborder des questions linguistiques concrètes que les apprenants posent souvent à l'enseignant. Dans cette optique, notre but est de proposer des réponses, au moins partielles, aux questions générales suivantes :

- La présentation de certains aspects « difficiles » du hongrois, peut-elle être améliorée par des exemples tirés des corpus ? Quelles sont les cas dans lesquels l'utilisation de corpus s'avère particulièrement utile ?
- Dans quelle mesure l'analyse des corpus peut-elle contribuer à fournir des réponses plus judicieuses aux questions des apprenants concernant les phénomènes langagiers échappant aux règles claires ? Comment l'enseignant peut-il procéder pour trouver des réponses pertinentes dans le corpus ?

[68] Il y a trois universités hongroises proposant une formation pour devenir enseignant de hongrois langue étrangère. Leurs programmes resepctifs (consultables en ligne) ne proposent pas de module sur l'utilisation des corpus à fins pédagogiques.

- L'analyse de corpus est-elle utile pour compléter les outils existants à disposition de l'enseignant et pour améliorer ses compétences professionnelles ?
- Comment adapter la présentation des résultats en cours de langues, aux niveaux de compétences linguistiques inférieurs ?

D'après notre expérience, ces questions concernent trois grands domaines en particulier : (1) le sens des mots (« Que veut dire exactement le mot « ... » ? »), (2) l'utilisation des synonymes (« Quelle est la différence entre « ... » et « ... » ? ») et (3) certaines descriptions grammaticales suivants s'articuleront autour de ces questions.

Après la présentation des corpus utilisés au chapitre 7, le chapitre 8 concernera l'utilisation des mots fréquents possédant des sens multiples. Dans cette section, nous étudierons en particulier l'usage de l'adjectif « nehéz » (difficile, lourd), un mot à usages multiples dont le sens est largement défini par l'environnement textuel.

Les chapitre 9 et 10 analyseront les différences dans l'utilisation des synonymes. Nous étudierons les occurrences de deux paires de verbes : « megjön/eljön » (qui se traduisent par « venir » ou « arriver ») et « tűnik/látszik » (comparables aux verbes français « sembler » et « paraître ») afin d'examiner dans quelle mesure les exemples authentiques explorés par les outils numériques peuvent mieux décrire leur usage.

Le chapitre 11 se concentrera sur la possibilité de préciser les règles d'usage de certains aspects de la langue considérés, avant tout, sous l'angle grammatical. Le chapitre sera consacré à une particularité grammaticale du hongrois, difficilement accessible aux apprenants : les deux conjugaisons. À partir du verbe « ad » (donner), nous analyserons comment aider à mieux appréhender l'utilisation des conjugaisons en incluant des aspects lexicaux, sémantiques et pragmatiques issus du corpus.

Après la présentation de la méthode d'analyse et de ses résultats, le dernier chapitre de la Partie II sera réservé aux questions liées à la didactique en tant que méthodologie permettant de présenter les résultats de l'observation du corpus de façon pertinente et compréhensible aux apprenants ainsi qu'à la sélection et la catégorisation des exemples dans le cadre pédagogique.

Pour explorer les phénomènes linguistiques identifiés, nous nous référerons aux méthodes de la linguistique de corpus. Ainsi, nous explorerons les corpus de hongrois avec les outils présentés au chapitre 3. Pour la présentation des résultats, nous utiliserons les critères établis par Hoey (2005), exposés au chapitre 4.

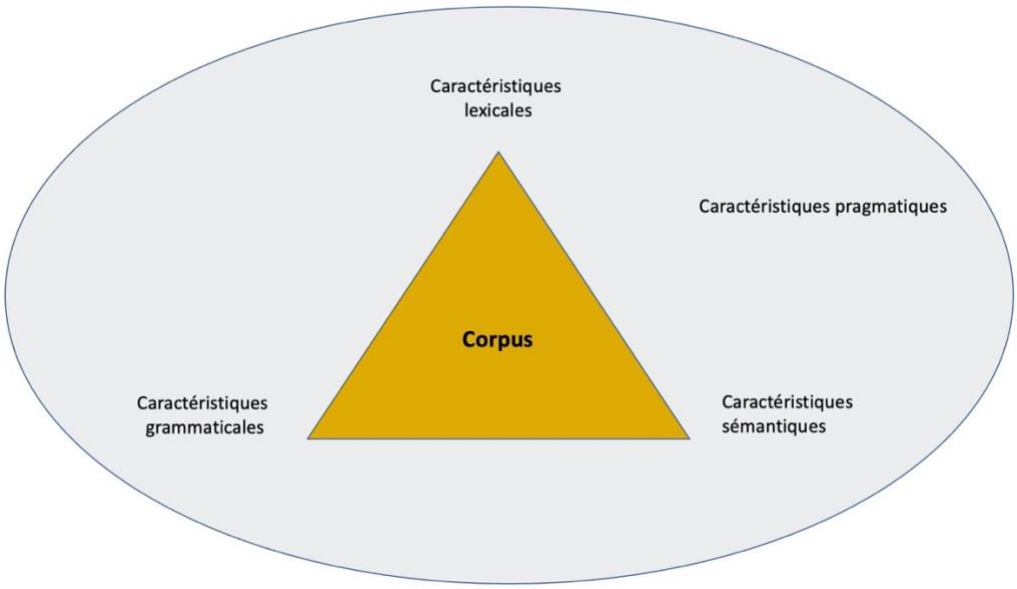


Tableau 61 : Mise en relation des domaines de la langue par l'analyse de corpus.

Nous tenons à préciser que *nous utiliserons les méthodes de la linguistique de corpus dans l'objectif d'augmenter l'efficacité de l'enseignement de certains points langagiers bien précis*. Nous nous sommes limitée à la présentation de quelques cas qui confirment l'intérêt de cette approche pour l'étude des mots à usages multiples, celle des synonymes et des deux conjugaisons. Cette démarche vise à démontrer l'efficacité d'un usage pédagogique des corpus tout en reconnaissant le fait qu'une description exhaustive des avantages et des limites de ces méthodes nécessiterait un travail plus approfondi, dépassant le cadre de cette thèse.

Chapitre 7 : Les corpus de hongrois

Ce chapitre présente les corpus écrits et oraux à partir desquels les explorations dans la Partie 2 de ce travail de recherche ont été effectuées. Il existe à l'heure actuelle un nombre très limité de corpus pour le hongrois. À l'exception de quelques corpus de taille plus petite et de profil plus spécifique tels que les corpus dédiés au langage des enfants ou aux textes juridiques européens, les deux grands corpus que nous exposerons dans les pages suivantes, sont les seuls susceptibles de pouvoir fournir des renseignements concernant l'usage langagier à l'écrit. Le choix des corpus oraux accessibles est encore plus limité et chacun de ces corpus met en évidence des problèmes particuliers que nous dévoilerons par la suite. De fait, les corpus dont nous disposons aujourd'hui ne permettent qu'une étude très partielle de l'usage du hongrois à l'oral. Ces corpus seront présentés dans ce chapitre mais leur étude ne fera pas partie de nos analyses.

Nous présenterons tout d'abord les corpus utilisés et fournirons une analyse succincte de leur contenu. Cette partie sera suivie de la description de leurs avantages et de leurs limites du point de vue de l'enseignant qui consulte ces corpus pour trouver des réponses à des questions bien précises.

A) Quelques mots à propos de la langue hongroise

Avant de présenter les corpus de hongrois, il convient de décrire, de façon succincte, la langue hongroise afin de permettre au lecteur de se forger une idée sur quelques-unes de ses particularités.

Le hongrois est une langue morphologiquement complexe qui appartient aux langues finno-ougriennes, apparentée donc, par exemple, au finnois et à l'estonien. Il s'agit d'une langue agglutinante dans laquelle plusieurs morphèmes peuvent se rajouter au radical pour en préciser certains aspects (la relation de possession ou le pluriel d'un nom ou le temps et le mode du verbe) et indiquer sa fonction dans la phrase. Szende et Kassai (2001 : 5) expliquent dans leur grammaire pédagogique l'agglutination de la façon suivante : « Plusieurs types d'élargissement peuvent ainsi se greffer sur une base pour produire des formes déclinées, conjuguées ou dérivées ». Par exemple, le nom « vonatokkal » (avec des trains) est composé de trois morphèmes : « vonat » (train), « -(o)k » (suffixe du pluriel précédé d'une voyelle de liaison) et « -kal » (suffixe correspondant à la préposition « avec » en français), ou encore le verbe « beszéltél » (tu as parlé) formé du radical

beszél » (il parle[69]), du -t (suffixe du passé) et de -(é)l (terminaison pour la deuxième personne du singulier, précédé d'une voyelle de liaison).

Au niveau sonorité, le hongrois connaît le phénomène de l'harmonie vocalique qui signifie que le type des voyelles dans les suffixes s'accorde avec le type des voyelles dans le radical. Ainsi, nous distinguons des mots à voyelles claires (e, é, i, í, ö, ő, ü, ű), à voyelles sombres (a, á, o, ó, u, ú) et à voyelles mixtes (contenant les deux types de voyelles). Ainsi, un mot contenant des voyelles sombres comme « vonat » (train) reçoit des suffixes contenant des voyelles sombres : « vonatokkal ». Le verbe « beszél » (il parle) contient des voyelles claires, les suffixes comporteront donc également des voyelles claires : « beszéltél ». Des mots à voyelles mixtes prennent des terminaisons à voyelle sombre.

Le hongrois est une langue phonétique qui utilise l'alphabet latin et la prononciation est relativement simple car cohérente. Les principales difficultés grammaticales concernent deux grands domaines : les verbes et l'ordre des mots (Szende et Kassai 2001 ; Nagyházi 2017). Le hongrois connaît notamment deux types de conjugaisons (conjugaison définie et conjugaison indéfinie), phénomène linguistique plutôt rare parmi les langues du monde (Szende et Kassai 2001). Le type de la conjugaison dépend du type de complément d'objet direct dans la phrase : si le complément d'objet direct est défini, nous utilisons la conjugaison définie, et si le complément d'objet direct est indéfini ou s'il n'y a pas de complément d'objet direct dans la phrase, nous utilisons la conjugaison indéfinie. La première difficulté est de décider du type du complément d'objet direct, la seconde, celle de la conjugaison correcte[70].

Un autre aspect de la langue qui peut poser problème est l'utilisation des préfixes qui sont « tributaires à la fois du sémantisme du verbe de base et des valeurs propres au préfixe » (Szende et Kassai 2001, voir aussi le chapitre 10). Ces préfixes peuvent indiquer la direction d'une action ainsi que son accomplissement mais ils peuvent également changer le sens du verbe (ad – donner, elad – vendre, felad – abandonner). Il est difficile pour l'apprenant de saisir quel est le préfixe approprié avec un verbe, si le préfixe est nécessaire et comment le préfixe impactera le sens du verbe.

[69] En hongrois, la troisième personne du singulier, conjugaison indéfinie est considérée comme le radical du verbe. C'est la forme que l'on trouve dans les dictionnaires.
[70] Voir aussi le chapitre 11 pour une description et une analyse plus détaillée.

Le troisième aspect problématique est lié à l'ordre des mots (Nagyházi 2017). Le hongrois est réputé pour être une langue pour laquelle l'ordre de mots est libre et flexible. Or, cette flexibilité a ses limites : la langue permet, en effet, une certaine variabilité dans l'ordre des mots mais cela ne veut pas dire que tout ordre de mots est acceptable dans toutes les situations. Les locuteurs peuvent organiser leurs énoncés en fonction de ce qu'ils considèrent être des éléments importants et moins importants dans leur phrase. L'ordre des mots est donc relatif à l'intention du locuteur mais il existe aussi d'autres facteurs déterminant l'ordre des mots, ce qui rend les règles véritablement complexes.

Dans la Partie II de cette thèse, nous explorerons, entre autres, dans quelle mesure les corpus peuvent contribuer à clarifier au moins certains de ces aspects identifiés comme problématiques car échappant aux règles claires et simples.

B) Les grands corpus et les corpus à fins pédagogiques

Comme évoqué précédemment, le nombre des corpus pour le hongrois est très restreint. Les plus importants sont le corpus « huTenTen12 » sur Sketch Engine et le « Magyar Nemzeti Szövegtár » (Corpus national du hongrois). Ces deux corpus ont été compilés à partir des données linguistiques accessibles sur Internet, le CNH contenant également un sous-corpus de transcriptions d'émissions radio. Il s'agit donc de deux très grands corpus mais, comme nous le verrons dans les pages suivantes, à deux profils différents.

Ces deux collections sont de nature « non pédagogique », i.e. elles n'ont pas été construites dans un objectif d'enseignement du hongrois mais pour permettre aux linguistes d'étudier la langue. Jusqu'à présent, ces corpus n'ont pas été véritablement explorés à fins pédagogiques ; ils pourraient cependant devenir – comme nous chercherons à démontrer par la suite – des aides précieuses pour les auteurs de matériels pédagogiques et pour les enseignants.

Les corpus pédagogiques présentés ci-dessous sont de plus petite taille. Ils ont l'avantage d'avoir été conçus en gardant à l'esprit le point de vue de l'apprenant (Braun 2006). Les énoncés dans ces corpus sont authentiques ou semi-authentiques, facilement recontextualisables car ils renferment du vocabulaire pertinent et accessible à l'apprenant. Leur inconvénient est qu'ils ne peuvent pas fournir de résultats statistiques fiables étant donné leur faible taille[71].

[71] Nous analyserons les corpus pédagogiques pour le hongrois dans la Partie III.

L'ensemble de ces corpus peut donner une image des caractéristiques du langage courant – la variété de la langue que les apprenants aux niveaux inférieurs sont censés maîtriser comme décrit au chapitre 1.

1) Corpus écrits

1.1) Le corpus « huTenTen12 » sur Sketch Engine

« HuTenTen12 » est un corpus hongrois composé de textes collectés sur Internet. Le corpus appartient à la famille des corpus « TenTen » sur Sketch Engine qui est un ensemble de corpus Web construits en utilisant la même méthode pour chaque langue. Les données ont été rassemblées par le programme « SpiderLing » en juin 2012 (Sketch Engine 2020). Ce corpus est actuellement la plus grande base de données pour le hongrois et il servira de point de départ principal pour nos explorations.

L'Internet peut être considéré comme la source la plus riche des manifestations du langage courant. Il contient une quantité très significative d'énoncés authentiques liés à une grande variété de sujets quotidiens. Plus précisément, le corpus « huTenTen12 » renferme 2 572 620 694 mots tirés de 6 447 178 documents accessibles sur Internet. Nous trouvons des sous-corpus du journal hebdomadaire HVG (11 150 575 mots), des pages de Wikipédia (7 994 455 mots) et des entrées d'un forum de management (6 349 641 mots). En plus de sélectionner ces grands sous-corpus, on peut également consulter des documents ou des types de pages Web (blog, forum, journal) en les choisissant manuellement. Bien que le corpus soit de taille très importante, le manque de précision dans la description de son contenu reste un problème pour l'utilisateur. Il est bien évidemment impossible de recenser et de lister tous les types de documents publiés sur le Web ou de cartographier leur distribution exacte mais une description plus longue pourrait orienter l'utilisateur vers les parties pertinentes pour sa recherche. Le tableau suivant (tableau 62) présente la composition du corpus « huTenTen12 ».

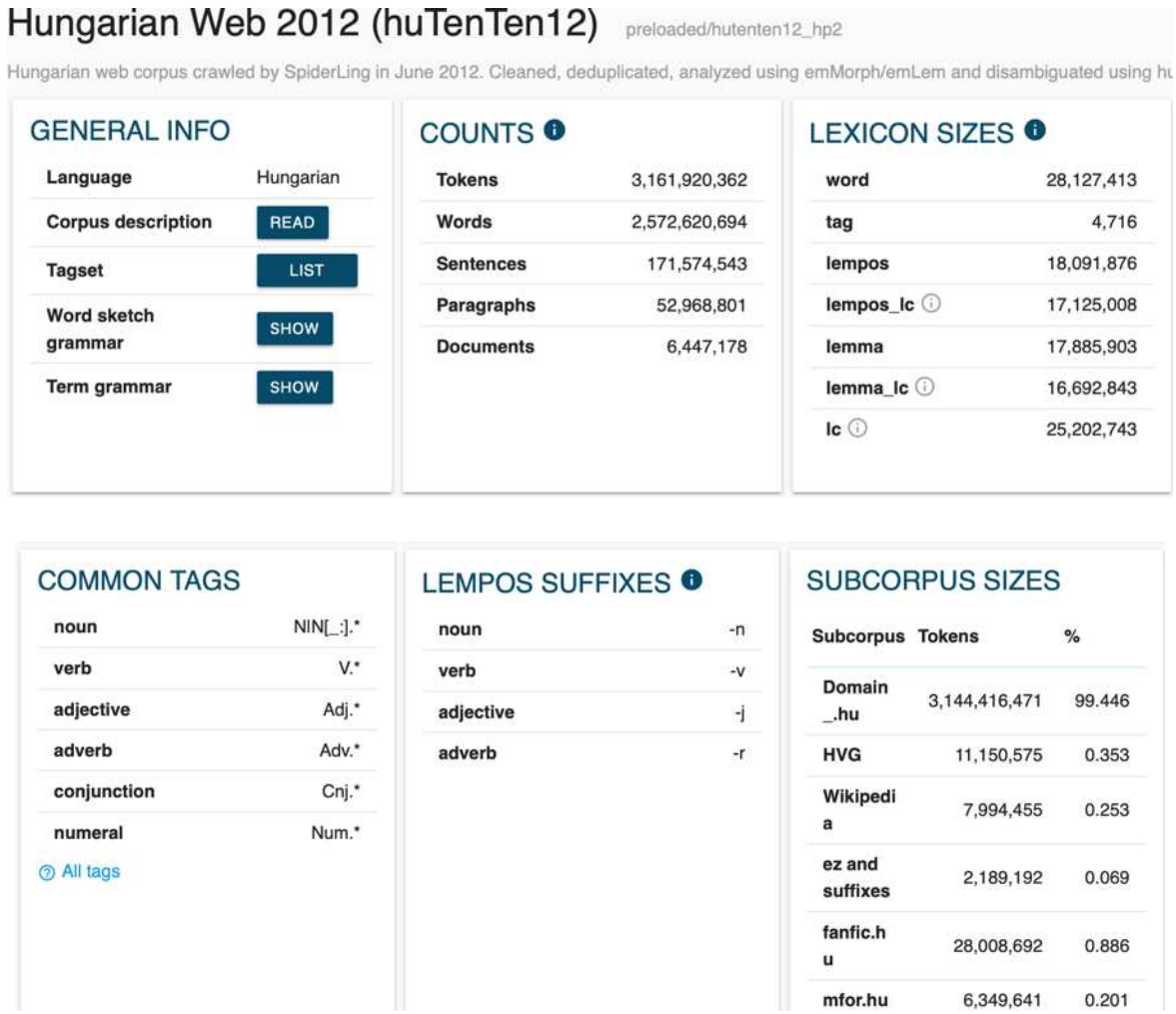


Tableau 62 : La composition du corpus « huTenTen12 ».

Afin de se forger une première idée sur la nature des données, nous avons utilisé l'outil « Wordlist »[72]. Pour cela, nous avons effectué une recherche des mille radicaux (lemmes) les plus communs dans le corpus. Par la forte présence des mots courants d'usage général et par l'absence des mots clairement associés à des domaines spécifiques et/ou à des usages de langue spécifique (juridique, journalistique, littéraire), cette liste confirme que le corpus contient majoritairement des exemples du langage courant. Les 100 lemmes les plus fréquents sont les suivants (tableaux 63 et 64) :

[72] Voir le chapitre 3 pour la description des outils numériques utilisés pour nos recherches.

	Lemma	Absolute Frequency ?
1	a	215,513,270
2	az	104,271,443
3	és	56,483,254
4	van	49,151,144
5	nem	39,531,197
6	is	36,030,738
7	ez	28,282,969
8	egy	28,015,834
9	ahogy	27,668,737
10	hogy	17,203,557
11	de	15,606,072
12	meg	13,983,577
13	ha	12,399,470
14	csak	11,784,250
15	már	11,301,209
16	ami	11,295,122
17	jó	10,781,292
18	még	10,000,321
19	kell	9,962,009
20	ő	9,844,353
21	sok	9,676,293
22	én	9,347,925
23	tud	9,167,540
24	mi	8,919,895
25	amely	8,489,933
26	aki	8,295,989
27	el	7,207,997
28	ki	6,766,316
29	mert	6,725,573
30	lesz	6,669,918
31	akkor	6,575,265
32	sem	6,443,862
33	minden	6,434,450
34	maga	6,165,813
35	nagy	6,150,454
36	év	5,795,060
37	olyan	5,653,200
38	így	5,588,373
39	mond	5,318,666
40	szerint	5,226,493
41	nagyon	5,102,312
42	pedig	4,866,841
43	úgy	4,819,478
44	ember	4,797,277
45	amint	4,641,472
46	amikor	4,547,019
47	új	4,536,995
48	majd	4,515,993
49	most	4,451,342
50	után	4,238,799
51	vagy	4,148,719
52	s	4,132,698
53	tesz	4,128,562
54	ők	4,095,295
55	két	3,938,855
56	más	3,804,109
57	idő	3,762,095
58	lát	3,746,803
59	nap	3,584,897
60	rész	3,509,171

1. *le/la*
2. *le/la*
3. *et*
4. *est, il y a*
5. *non, ne pas*
6. *aussi*
7. *ce, ça*
8. *un, une*
9. *comme*
10. *que*
11. *mais*
12. *et ou « meg » (préfixe)*
13. *si*
14. *seulement, ne que*
15. *déjà*
16. *que (pronom relatif)*
17. *bien, bon*
21. *beaucoup*
22. *je*
23. *peut, sait*
24. *quoi*
25. *lequel, laquelle*
26. *qui (pronom relatif)*
27. *« el » (préfixe)*
28. *qui (mot interrogatif)*
29. *parce que*
30. *sera*
31. *alors*
32. *non plus*
33. *tout*
34. *même*
35. *grand, large*
36. *année, an*
37. *tel, telle*
41. *très*
42. *mais, et*
43. *tellement*
44. *homme, on*
45. *comme*
46. *quand*
47. *neuf, neuve, nouveau, nouvelle*
48. *plus tard*
49. *maintenant*
50. *après*
51. *ou, (tu) es*
52. *et*
53. *met, fait*
54. *ils, elles*
55. *deux*
56. *autre*
57. *temps*

18. *encore*
19. *faut*
20. *il, elle*
38. *comme ça, ainsi*
39. *dit*
40. *selon*
58. *voit*
59. *jour, journée, soleil*
60. *part, partie*

61	fog	3,491,910	•••
62	magyar	3,478,978	•••
63	itt	3,450,305	•••
64	fel	3,439,666	•••
65	be	3,393,858	•••
66	ilyen	3,247,766	•••
67	mint	3,210,512	•••
68	szeret	3,149,017	•••
69	áll	3,146,028	•••
70	nincs	3,054,018	•••
71	akar	3,034,252	•••
72	egyik	2,989,906	•••
73	vesz	2,972,884	•••
74	azért	2,904,625	•••
75	hanem	2,874,696	•••
76	ahol	2,792,948	•••
77	ad	2,739,534	•••
78	megy	2,727,957	•••
79	te	2,718,585	•••
80	eset	2,694,957	•••

81	között	2,667,545	•••
82	tart	2,582,847	•••
83	dolog	2,582,716	•••
84	kerül	2,571,885	•••
85	szó	2,558,926	•••
86	ne	2,547,872	•••
87	él	2,517,724	•••
88	teljes	2,513,122	•••
89	azonban	2,491,961	•••
90	ír	2,480,021	•••
91	számára	2,446,190	•••
92	ott	2,430,383	•••
93	hely	2,424,721	•••
94	valami	2,400,988	•••
95	mindig	2,395,582	•••
96	kicsi	2,357,879	•••
97	jön	2,355,900	•••
98	élet	2,332,281	•••
99	sor	2,311,231	•••
100	ezért	2,238,166	•••

61. *auxiliaire du futur*
62. *hongrois, hongroise*
63. *ici*
64. *« fel » (préfixe)*
65. *« be » (préfixe)*
66. *tel, telle*
67. *comme*
68. *aime*
69. *debout, consiste*
70. *il n'y a pas*
81. *entre, parmi*
82. *dure*
83. *chose*
84. *coûte, vient*
85. *mot*
86. *ne pas*
87. *vit*
88. *complet, entier*
89. *cependant*
90. *écrit*
91. *pour*

71. veut	*92. là-bas*
72. l'un, l'une	*93. endroit, lieu, place*
73. prend, achète	*94. quelque chose*
74. « azért » (mot introduisant une explication)	*95. toujours*
75. mais	*96. petit, petite*
76. où (pronom relatif)	*97. vient*
77. donne, offre	*98. vie*
78. va	*99. ligne, queue*
79. tu	*100. pour cette raison, ainsi*
80. cas	

Tableaux 63 et 64 : Les cent lemmes les plus fréquents dans le corpus « huTenTen12 ».

Nous avons ensuite effectué une recherche des radicaux par partie du discours et nous avons analysé 50 concordances au hasard avec vingt verbes, noms, adjectifs, i.e. l'élément 1, 5, 10, etc. de la liste. Le critère principal était de savoir si les mots listés et/ou les occurrences choisies s'associaient clairement à l'usage de langue d'un domaine spécifique. Dans le cas des verbes et des adjectifs, nous ne trouvons aucun indice qui suggérerait que tel est le cas. Les lignes de concordance révèlent les origines mixtes des sources : des présentations des sites Web, des entrées de forum, des commentaires, des articles scientifiques populaires etc. La liste des noms donne une indication des sujets dominants dans le corpus (santé, religion, politique, travail), mais comme les occurrences proviennent majoritairement des contributions des sites Web à intérêt général, des forums et des articles scientifiques-populaires, le langage n'est pas vraiment caractéristique et associable à des usages particuliers.

Ce corpus peut être exploré en utilisant des codes de deux systèmes d'annotation complexes. Il est possible de chercher avec les tags des caractéristiques morphologiques de « emMorph » ou de se servir du système d'annotation développé pour le « Corpus national du hongrois » (voir la description dans la section 1.2, ci-dessous). En employant les deux systèmes, les taux d'erreurs se réduisent car nous sommes en mesure de vérifier les résultats de deux façons. Ce sera l'approche que nous effectuerons lors de nos recherches.

Les outils sont simples à appliquer et la présentation des résultats permet de les interpréter facilement, et également d'effectuer des analyses en combinant les outils disponibles. Par exemple, nous pouvons étendre la recherche à des collocations depuis la base et de l'un de ses collocatifs à

des unités multi-lexicales ou voir des exemples avec ces unités dans le Concordancier puisque le programme offre la capacité de passer d'un outil à l'autre (par exemple du Concordancier à Word Sketch et vice versa) et d'explorer les différentes facettes d'une question en utilisant plusieurs outils en combinaison.

Notons enfin que nous sommes parfois confrontés à des erreurs de POS-tagging et des erreurs d'identification du lemme. Bien que, par rapport à la quantité de données, ce genre d'erreurs soit plutôt rare, il est nécessaire de vérifier soigneusement les résultats.

1.2) Le « Magyar Nemzeti Szövegtár » (Corpus national du hongrois)

Afin de valider et de compléter les résultats tirés du corpus « hunTenTen12 » sur Sketch Engine, nous avons systématiquement consulté un autre très grand corpus, le Corpus national du hongrois.

La première version de ce corpus a été développée entre 1998 et 2001 par l'Institut de la linguistique de l'Académie hongroise des sciences dans l'objectif de devenir le corpus représentatif de l'utilisation de la langue dans la seconde moitié des années 1990 (Váradi, 2002). Ce premier corpus hongrois annoté contenait environ 187 millions de mots, couvrant les variétés linguistiques de Hongrie et des pays limitrophes (Oravecz et al. 2014).

Ce corpus renferme une collection de « données du langage véritablement utilisé à l'écrit ou à l'oral » (http://corpus.nytud.hu/mnsz/) « [L]es textes sont choisis et organisés selon différents critères (…) Ce n'est pas seulement une importante collection de textes mais il recèle aussi des informations bibliographiques sur ces textes, une annotation de la structure (…) et des parties du discours » (ibid.). En 2001, avec plus de 1,5 milliard de mots, cette base de données souhaitait devenir « le corpus représentatif du langage standard écrit (*írott köznyelv*) » (ibid). Plus tard, le concept du corpus a été modifié. Váradi (2002), un des créateurs de ce corpus, explique le changement d'objectif en exposant les problèmes liés au concept de la représentativité. Il note ainsi qu'il est impossible de concevoir un corpus capable d'illustrer de manière statistiquement fiable toutes les variétés de la langue. À partir de ce constat, les créateurs de corpus se sont engagés à construire *un corpus équilibré* plutôt qu'un corpus représentatif. Cette deuxième version a été créée en 2005. La taille du corpus original a été significativement augmentée : les sous-corpus ont été enrichis afin de présenter un corpus plus équilibré. Nous observons que l'augmentation la plus significative apparaît dans le cas des communications personnelles (de 18,6 millions à 338,6

millions de mots) ainsi que dans le rajout des transcriptions des communications orales (plus de 83 millions de mots).

Les sous-corpus sont organisés autour des genres différents : des textes de presse, des textes littéraires, scientifiques, officiels et personnels. Le tableau suivant (tableau 65) montre la composition exacte du corpus actuel (HGC) et celle du premier corpus (HNC) :

Register	HNC	HGC		Source
Journalism	84,500,000	643,257,776	(42%)	Daily/weekly newspapers
Literature	38,200,000	221,731,436	(14.5%)	Digital Literary Academy
(Popular) science	25,500,000	110,903,157	(7.2%)	Hungarian Electronic Library
Personal	18,600,000	338,600,000	(22.1%)	Social media
Official	20,900,000	135,401,305	(8.8%)	Documents from public admin.
(Transcribed) spoken	–	83,040,104	(5.4%)	Radio programs
	187,000,000	**1,532,933,778**		

Table 1: The composition of the HGC in number of tokens

Tableau 65 : La composition du CNH par nombre de mots (Oravecz et al. 2014).

Sur le site du Corpus national du hongrois, nous trouvons également quelques informations supplémentaires concernant l'origine des textes. Le sous-corpus de presse comporte principalement des articles politiques, économiques et sportifs, la partie « Textes littéraires » comprend la matière accessible sur l'Académie digitale de la littérature hongroise contemporaine. Les textes scientifiques incluent des textes provenant d'une trentaine de domaines. Les textes sont d'origine mixte : on y trouve aussi bien des textes de sciences populaires que des textes parus dans des journaux validés par les pairs, tirés de la « Magyar Elektronikus Könyvtár » (Bibliothèque digitale hongroise). La collection des communications officielles renferme les textes de lois, d'arrêts et de décrets et des débats parlementaires. Les communications personnelles sont essentiellement des entrées de forum.

Le contenu de ce grand corpus est quelque peu déséquilibré. Les constructeurs du corpus indiquent que presque la moitié des données provient de la presse, en rajoutant que ces textes « illustrent un large spectre des variétés langagières » (Oravecz et al. 2014). Ce déséquilibre devient évident dès que l'on étudie la liste des mots les plus fréquents dans le corpus (tableau 66) :

	szótő	szófaj	db	db / 1000 szó		szótő	szófaj	db	db / 1000 szó		szótő	szófaj	db	db / 1000 szó
1.	a	Det	11128421	72,40	34.	ki	Pre	305480	1,99	67.	között	NU	159583	1,04
2.	az	Det	3716414	24,18	35.	ami	Pro	287999	1,87	68.	első	Num	158569	1,03
3.	és	Con	2544751	16,56	36.	nagy	A	281134	1,83	69.	nap	N	157310	1,02
4.	hogy	Con	2166004	14,09	37.	mond	V	276868	1,80	70.	ad	V	154537	1,01
5.	A	Det	2103970	13,69	38.	mi	Pro	275076	1,79	71.	99	DIG	154526	1,01
6.	az	Pro	1803814	11,74	39.	maga	Pro	263983	1,72	72.	azonban	Con	154150	1,00
7.	nem	Adv	1693748	11,02	40.	mert	Con	258962	1,68	73.	sok	Num	152907	0,99
8.	is	Con	1677108	10,91	41.	én	Pro	245386	1,60	74.	ők	Pro	151718	0,99
9.	van	V	1418113	9,23	42.	-e	Clit	237612	1,55	75.	más	Pro	151698	0,99
10.	ez	Pro	1204269	7,84	43.	olyan	Pro	232947	1,52	76.	kérdés	N	151477	0,99
11.	egy	Num	899832	5,85	44.	jó	A	232826	1,51	77.	hanem	Con	150702	0,98
12.	Az	Det	730287	4,75	45.	több	Num	232803	1,51	78.	Ha	Con	147117	0,96
13.	meg	Pre	592986	3,86	46.	magyar	A	229934	1,50	79.	eset	N	146803	0,96
14.	kell	V	499659	3,25	47.	minden	Pro	225130	1,46	80.	elnök	N	146500	0,95
15.	csak	Adv	477956	3,11	48.	úgy	Adv	221524	1,44	81.	forint	N	144629	0,94
16.	lesz	V	469189	3,05	49.	pedig	Con	216513	1,41	82.	egyik	Pro	143627	0,93
17.	de	Con	462508	3,01	50.	új	A	215765	1,40	83.	kormány	N	139493	0,91
18.	már	Adv	452814	2,95	51.	tesz	V	211798	1,38	84.	akar	V	138696	0,90
19.	Ez	Pro	447310	2,91	52.	két	Num	211077	1,37	85.	ország	N	137225	0,89
20.	amely	Pro	417945	2,72	53.	00	DIG	205993	1,34	86.	kerül	V	135554	0,88
21.	ha	Con	402593	2,62	54.	ember	N	198039	1,29	87.	De	Con	135062	0,88
22.	még	Adv	396207	2,58	55.	Az	Pro	194263	1,26	88.	százalék	N	132780	0,86
23.	vagy	Con	381098	2,48	56.	után	NU	190805	1,24	89.	lát	V	131866	0,86
24.	mint	Con	370507	2,41	57.	Nem	Adv	185338	1,21	90.	törvény	N	129485	0,84
25.	szerint	NU	369481	2,40	58.	idő	N	178374	1,16	91.	98	DIG	128540	0,84
26.	el	Pre	362004	2,36	59.	majd	Adv	177497	1,15	92.	sor	N	128311	0,83
27.	tud	V	356833	2,32	60.	be	Pre	175615	1,14	93.	kap	V	127841	0,83
28.	s	Con	356453	2,32	61.	tart	V	173048	1,13	94.	fog	V	127768	0,83
29.	aki	Pro	350819	2,28	62.	rész	N	170894	1,11	95.	alap	N	127632	0,83
30.	év	N	338213	2,20	63.	most	Adv	168334	1,10	96.	2	DIG	127461	0,83
31.	sem	Adv	329570	2,14	64.	fel	Pre	164467	1,07	97.	itt	Adv	127399	0,83
32.	lehet	V	310500	2,02	65.	szó	N	162929	1,06	98.	hely	N	124262	0,81
33.	ő	Pro	306621	1,99	66.	1	DIG	162486	1,06	99.	vesz	V	123583	0,80

1. *le/la*
2. *le/la*
3. *et*
4. *que*
5. *Le/La*
6. *ce/cette ...-là*
7. *ne pas*
8. *aussi*
9. *est, il y a*
10. *ce/cette*
11. *un/une*
12. *Ce/Cette*
13. *meg (préfixe)*
14. *faut*
15. *seulement, juste, ne que*
16. *sera*
17. *mais*
34. *qui*
35. *que*
36. *grand*
37. *dit*
38. *quoi*
39. *même*
40. *parce que*
41. *je*
42. *-e (mot interrogatif)*
43. *tel/telle*
44. *bien, bon/bonne*
45. *plusieurs*
46. *hongrois*
47. *tout*
48. *de telle façon*
49. *et, mais*
50. *nouveau/nouvelle*
67. *entre*
68. *premier/première*
69. *jour, journée*
70. *donne*
71. *99*
72. *cependant*
73. *beaucoup*
74. *ils/elles*
75. *autre*
76. *question*
77. *(ne pas) mais*
78. *Si*
79. *cas*
80. ***président***
81. ***forint***
82. *l'un/l'une*
83. ***gouvernement***

18. déjà	*51. met, place*	*84. veut*
19. Ceci	*52. deux*	***85. pays***
20. lequel/laquelle	*53. 00*	*86. coûte, vient*
21. si	*54. homme*	*87. Mais*
22. encore	*55. Ce/Cette ...-là*	***88. pourcentage***
23. ou	*56. après*	*89. voit*
24. comme	*57. Non*	***90. loi***
25. selon, d'après	*58. temps*	*91. 98*
26. el (préfixe)	*59. d'ici peu*	*92. rangée*
27. peut, sait	*60. be (préfixe)*	*93. reçoit*
28. et	*61. fait, donne, effectue*	*94. fog (aux. du futur)*
29. qui	*62. part, partie*	*95. base*
30. an, année	*63. maintenant*	*96. 2*
31. non plus, ne pas	*64. fel (préfixe)*	*97. ici*
32. peut être, possible	*65. mot*	*98. endroit*
33. il/elle	*66. 1*	*99. achète*

Tableau 66 : Les 100 mots les plus fréquents dans le « Corpus national du hongrois ».

La liste inclut, en effet, des mots que l'on rencontre fréquemment dans tous les genres de textes. Ce sont les articles définis et indéfinis, le verbe « van » (être), les conjonctions « és » (et) et « de » (mais), des postpositions comme « szerint » (selon, d'après), des verbes modaux comme « kell » (il faut), lehet (il est possible), ou des pronoms personnels « én » (je), « mi » (nous), « ők » (ils/elles), etc. Cependant, nous trouvons également quelques résultats surprenants. Les mots comme « magyar » (hongrois) à la position 46, « elnök » (président) à la position 80, « kormány » (gouvernement) à la position 83, « ország » (pays) à la position 85, « százalék » (pourcentage) à la position 88 et « törvény » (loi) à la position 90 indiquent très clairement qu'il s'agit de textes liés à la vie politique. La majorité des lignes de concordance avec ces mots (72%) provient des sources de presse ou des sources littéraires (des romans historiques), leur langage portant également les caractéristiques de ces deux variétés langagières.

Dans leur publication « Doing linguistics with a corpus », Egbert et al. (2020) attirent notre attention sur un certain nombre de problèmes liés aux grands corpus. Ils mettent notamment l'accent sur la nécessité de comprendre leur nature et leur composition afin d'éviter de tirer des conclusions hâtives des résultats. Ils recommandent ainsi de vérifier le contenu exact du corpus

car la taille n'est qu'un indice parmi d'autres. Un grand corpus n'est en effet pas forcément un corpus équilibré (p. 22).

Nous avons cherché à contrebalancer ce déséquilibre en excluant certains sous-corpus (journaux, littérature, textes légaux) de nos explorations. Les autres sous-corpus (réseaux sociaux, transcriptions d'émissions radio) s'avérant en effet suffisamment larges pour fournir des résultats statistiquement pertinents.

Les publications concernant la construction de ce corpus (Oravecz et al. 2014 ; Váradi 2002) donnent aussi des informations liées à l'encodage – sujet très important pour les langues morphologiquement complexes comme le hongrois. Comme nous l'avons mentionné plus haut, le même système a été utilisé pour l'annotation du corpus sur Sketch Engine.

Dans le système de code MSD conçu pour l'annotation du hongrois (Erjavec 2004), les tags sont formés à partir d'une première lettre indiquant la catégorie principale (i. e. la partie du discours, par exemple : N = nom, V = verbe, etc.), puis d'une série d'autres lettres indiquant les caractéristiques du mot. Par exemple, le code pour le mot « asztalát » (« sa table » en fonction du COD) est Nc-sa—s3. « N » indique qu'il s'agit d'un nom, « c » indique un nom commun (au contraire d'un nom propre) au singulier à l'accusatif (il est le complément d'objet direct de la phrase) qui a une terminaison du possessif de la troisième personne (Recski 2014 : 469).

Cette annotation très détaillée permet de chercher de manière précise et de récupérer des informations statistiques concernant la fréquence des mots individuels mais aussi des catégories entières de mots (par exemple, tous les mots avec les mêmes caractéristiques morphologiques). Il est également possible de consulter des lignes de concordance et d'identifier des collocations avec le mot en question.

1.3) Les corpus écrits à fins pédagogiques

Ces corpus seront décrits en détail dans la Partie III de cette thèse. Nous nous restreignons donc ici à une présentation succincte de leur contenu. Le corpus écrit est constitué de trois parties : (1) des textes informés par le corpus, (2) des textes semi-authentiques/authentiques et (3) des textes entièrement authentiques. Toute la matière linguistique dans le corpus est basée sur des sujets et des situations de communication prescrites par le CECRL. La collection, contenant 711 000 mots au total, est structurée comme suit :

- Le sous-corpus 1 comprend toutes les données linguistiques des manuels « MagyarOK » et contient environ 150 000 mots.
- Le sous-corpus 2 comporte des récits semi-authentiques écrits par des locuteurs natifs sur des thèmes traités dans les manuels. Cette collection compte 61 000 mots.
- Le sous-corpus 3 se compose de textes authentiques de divers genres (articles de journaux, entrées de blogs et de forums) et de ressources personnelles telles que des courriels et des interactions sur les réseaux sociaux. Identique au sous-corpus 2, elle renferme des textes sur des sujets pertinents pour les niveaux de compétences inférieurs. Cet ensemble de données englobe environ 500 000 mots.

Il est clair que ces corpus ne peuvent pas fournir de résultats statistiques fiables en raison de leur taille. En revanche, ils peuvent servir de collection d'exemples qui correspondent le mieux au niveau de l'apprenant.

2) Les corpus oraux

2.1) Le corpus oral du Corpus national du hongrois

Le corpus oral du Corpus national du hongrois comprend 84 millions de tokens provenant des entretiens et des textes lus (des informations) à la radio, *le langage interactionnel ne fait donc partie de ce corpus que de manière limitée*. Dans la mesure où les apprenants des niveaux de compétences linguistiques inférieurs ont besoin d'acquérir, avant tout, des compétences interactionnelles, ce corpus seul ne peut pas fournir de données idoines informant sur l'usage langagier oral. Pour compléter cette lacune autant que possible, nous avons inclus notre corpus de langue parlée à fins pédagogiques qui sera présenté plus en détail dans la Partie III de cette thèse.

2.2) Les corpus oraux à fins pédagogiques

Nous avons évoqué à plusieurs reprises en Partie I que le discours oral possède ses propres caractéristiques. Il est donc nécessaire de l'inclure dans l'analyse, même si les bases de données de la langue parlée contiennent en l'occurrence significativement moins de données linguistiques, ce qui résulte en des valeurs statistiques moins fiables car moins représentatives. L'ensemble des données parlées de notre corpus compte environ 380 000 mots (plus de 40 heures de matériel enregistré mais toujours un corpus plutôt limité en raison de sa taille). Analogue au corpus écrit à fins pédagogiques, ce corpus comprend trois sous-parties qui sont les suivantes :

- Le sous-corpus 1 contient les transcriptions des dialogues du manuel. Ces conversations semi-authentiques – i. e. informées par le corpus mais éditées – présentent de nombreuses caractéristiques typiques des interactions orales.
- Le sous-corpus 2 inclut deux types d'interactions semi-authentiques/authentiques : (1) des improvisations d'acteurs et (2) de courtes interviews avec des locuteurs natifs. Ces interactions n'étaient pas scénarisées mais elles étaient en partie guidées.
- Le sous-corpus 3 comporte des données authentiques et est également divisé en deux parties : (1) des rencontres en des lieux de service différents et (2) des conversations informelles entre amis ou membres de la famille.

Contrairement à la collection des textes écrits, ce corpus fournit non seulement des exemples pour des usages identifiés dans le corpus « huTenTen12 » et dans le Corpus national du hongrois mais aussi des données linguistiques pour des situations de communication interpersonnelles et informelles qui, comme nous l'avons vu, ne font pas partie des grands corpus.

Il convient cependant de noter que *ce corpus est de taille très limitée et sert, avant tout, de source d'exemples et, de temps en temps, pour observer des usages différents du langage écrit.* Pour arriver à des résultats généralisables, nous aurions besoin de bases de données linguistiques nettement plus grandes, inexistantes à ce jour.

C) Contenu, avantages et limites des corpus utilisés

Le tableau suivant (tableau 67) fait apparaître les caractéristiques principales des corpus présentés ci-dessus. Nous avons inclus les avantages et les limites du point de vue de l'auteur des matériels pédagogiques et du professeur de langues :

	Corpus national du hongrois	**« huTenTen12 »**
Contenu	- Textes de presse - Textes littéraires - Articles scientifiques - Communications sur quelques forums choisis	- Textes provenant des sites Web publiés en hongrois sur Internet avant 2012

	- Transcription des entretiens de radio	
Nombre de mots (tokens)	1 532 933 778	2 572 620 694
Outils[73]	- Concordancier - Générateur de collocations	- Concordancier - Word Sketch - Word Sketch Difference - Wordlist - N-grams - Thésaurus - Keywords (outil non utilisé dans nos analyses)
Avantages	- Possibilité de sélectionner des sous-corpus larges - Image claire du contenu - Informations statistiques pertinentes	- Outils faciles à utiliser et à combiner pour de meilleurs résultats - Présentation claire des résultats - Informations statistiques pertinentes
Limites	- Forte prédominance du langage de la presse - Nombre plutôt limité d'outils - Présentation peu claire des collocations - Exemples demandant un travail de simplification et de recontextualisation par le professeur/l'auteur des matériels pédagogiques	- Difficulté de sélectionner des sous-corpus par genre et registre - Image floue du contenu - Exemples demandant un travail de simplification et de recontextualisation

Corpus « MagyarOK » à fins pédagogiques

[73] Pour une présentation des outils d'analyse de corpus listés, le lecteur est invité à revoir le chapitre 3.

Contenu	- Communications personnelles écrites et orales : e-mails et chats, conversations - Textes et entretiens semi-authentiques sur des sujets du CECRL - Dialogues et textes narratifs liés aux sujets du CECRL
Nombre de mots (tokens)	Corpus écrit : env. 711 000 mots Corpus oral : env. 380 000 mots
Outils*	- Concordancier - Word Sketch - Word Sketch difference - Wordlist - N-grams - Thésaurus et Keywords (outils non utilisés dans nos analyses)
Avantages	- Exemples accessibles aux apprenants - Situations de communication et genres de texte en accord avec le CECRL
Limites	- Peu d'informations statistiques à cause de la taille du corpus - Certaines collocations et unités multi-lexicales fréquentes ne sont pas incluses

Tableau 67 : Le contenu, les avantages et les limites des corpus de hongrois utilisé dans ce travail de recherche.

Pour nos analyses, nous utiliserons « huTenTen12 » comme corpus principal. Ce corpus offre plusieurs avantages dont les plus importants sont :

- La matière linguistique dans ce corpus est globalement plus proche de la variété de langue que les apprenants aux niveaux inférieurs doivent acquérir (langage courant).

- Il est possible de combiner des outils pour arriver facilement à des résultats précis et détaillés.
- La présentation des résultats est claire et relativement facile à interpréter.

Les sous-ensembles des journaux, des articles de vulgarisation, des messages personnels dans la partie écrite du Corpus national du hongrois nous serviront à vérifier les résultats obtenus à partir du corpus « huTenTen12 ». Une comparaison entre ces deux grands corpus à contenu similaire n'est pas inutile car elle sert à valider (ou à réfuter) les résultats. Le corpus oral de CHN ne fera cependant pas partie de nos analyses, car il ne contient que peu de données linguistiques bénéfiques pour les niveaux de compétences inférieurs. Les corpus oraux à fins pédagogiques sont de taille très limitée ne permettant aucune analyse statistique permettant une généralisation, ils seront donc également exclus de notre étude.

Ce chapitre a présenté les corpus actuels pour le hongrois dont les collections écrites seront utilisées pour nos analyses dans la Partie II de cette thèse. Les différentes collections ont toutes leurs avantages et leurs limites pour l'enseignement de la langue : les grands corpus permettent d'obtenir des résultats assez fiables sur l'usage langagier des natifs mais leur contenu n'est pas nécessairement accessible aux apprenants. Les corpus pédagogiques présentés remplissent la condition de l'accessibilité mais ne peuvent pas fournir suffisamment de données linguistiques pour en tirer des résultats fiables concernant l'utilisation de l'élément langagier étudié. Nous avons également constaté le manque de grands corpus oraux contenant des interactions du quotidien. Ces corpus dans leur ensemble permettent néanmoins une première analyse de l'usage du hongrois, en particulier celle des phénomènes linguistiques présentés aux chapitres 8 à 11.

Chapitre 8 : « Que veut dire ... ? » Mots à usages multiples : « nehéz » (lourd, difficile)

Les corpus présentés dans la partie précédente permettent d'observer et d'analyser les aspects sélectionnés du hongrois. Le chapitre 8 examinera le potentiel des corpus pour les questions considérées avant tout comme « lexicales » (bien que l'inséparabilité du lexique et de la grammaire devienne évidente pendant le processus), autour des mots à usages multiples. Pour illustrer le processus, nous utiliserons un exemple concret : l'adjectif « nehéz » (difficile, lourd) dont les sens possibles émergent, avant tout, d'une étude de ses environnements textuels.

A) L'adjectif « nehéz » (difficile, lourd)

Consulter un corpus n'est pas nécessairement utile pour définir le sens relativement clair des mots tels que « table », « femme » ou « chien », bien que les modes d'utilisation typiques de ces mots peuvent émerger d'une analyse de corpus. L'analyse peut être cependant profitable quand il s'agit des mots dont le sens dépend largement du contexte. D'un côté, l'identification des sens de tels mots sans environnement textuel est souvent problématique. De l'autre, l'ambiguïté est plutôt rare dans les énoncés concrets dans lesquels le sens du mot est élucidé par son environnement (cf. Hoey 2005 ; Sinclair 2004b ; Hanks 2013). Ces mots qualifiés de « difficiles » par les apprenants peuvent être catégorisés dans les groupes suivants :

(1) Des mots avec un ou plusieurs sens difficilement compréhensible(s) puisqu'ils n'ont pas d'équivalents exacts en français. Par exemple le mot « pedig » peut vouloir dire à la fois « et », « pourtant » et « cependant ».
(2) Des mots à sens multiples : « jár » (~ aller, aller régulièrement quelque part, circuler, sortir avec quelqu'un); « jó » (~ bon, qui fonctionne, qui convient, d'accord) ou « van » (~ il y a, exister, être, avoir).
(3) Des mots qui n'apparaissent pas dans la structure des phrases françaises et posent donc des problèmes de traduction : « azért », « akkor », « azt ». (Ce sont des éléments linguistiques qui se situent dans la première partie de la phrase mais qui font référence à la deuxième partie de la même phrase.)

Dans son ouvrage sur l'apprentissage des langues, Nation (2013) souligne l'importance du fait que les apprenants doivent être conscients de ce qu'implique connaître un mot. Nation propose que

cette prise de conscience doit être basée sur un système organisé pour qu'ils puissent facilement se souvenir des aspects à observer et puissent facilement vérifier les lacunes dans leurs connaissances (p. 584). Connaître un mot implique un large éventail de fonctionnalités listées dans le tableau (tableau 68) qui suit (voir en ce sens Nation 2013 : 49).

Forme :	
• Orale • Écrite • Composantes du mot	R Quelle est la sonorité du mot ? P Comment faut-il prononcer le mot ? R Comment se présente le mot à l'écrit ? P Comment faut-il épeler le mot ? R Quelles composantes forment le mot ? P Quelles parties du mot sont nécessaires pour exprimer le sens ?
Sens	
• Forme et sens • Concepts • Associations	R Quel est le sens de la forme donnée du mot ? P Quelle forme peut-on utiliser pour exprimer le sens donné ? R Que sait-on sur le mot au niveau conceptuel ? P À quels concepts peut-on faire référence en utilisant ce mot ? R Quels autres mots associe-t-on au mot en question ? P Quels autres mots peut-on utiliser à la place de celui-ci ?
Usage	
• Fonctions grammaticales • Collocations • Restriction d'usage (registre, fréquence)	R Quels schémas sont associés au mot ? P Avec quels schémas doit-on utiliser ce mot ? R Quels mots ou genres de mots accompagnent ce mot ? P Avec quels mots ou genres de mots doit-on utiliser ce mot ? R Où, quand et à quelle fréquence doit-on s'attendre à rencontrer ce mot ? P Où, quand et à quelle fréquence peut-on utiliser ce mot ?

Tableau 68 : Ce qu'implique connaître un mot (Nation 2013 : 49, notre traduction).

Les informations concernant la forme, bien qu'importantes dans l'absolu, ne présentent pas d'intérêt particulier dans le cadre de notre travail car elles peuvent être facilement identifiées sans utiliser de corpus. Nous nous concentrons sur les deux autres aspects, le sens et l'usage, domaines pour lesquels la consultation des corpus peut être très utile. Ils concernent le sens associé à la forme ainsi que les schémas utilisables pour aider l'apprenant à comprendre et à produire des

énoncés. Ces schémas existent à tous les niveaux : orthographique, morphologique, collocationnel, grammatical, sémantique et pragmatique (Nation 2013 : 50). Taylor (2012 : 111-119) attire notre attention sur l'importance de ces schémas en soulignant que les connaître permet aux apprenants de comprendre et de produire des énoncés qu'ils n'ont pas rencontrés auparavant sous cette forme exacte, ces schémas méritant ainsi plus d'attention de la part de l'enseignant et de l'apprenant.

Dans cette optique, nous chercherons à explorer les questions suivantes concernant l'adjectif « nehéz » :

- Peut-on observer des schémas dans l'environnement textuel de ce mot ? Quelles sont ses éléments lexicaux les plus fréquemment associés à des usages identifiés ?
- Quelles sont les particularités grammaticales associées à des sens/usages différents ?
- Y a-t-il des particularités sémantiques associées à des sens/usages différents ?
- Y a-t-il des particularités pragmatiques associées à des sens/usages différents ?

L'exemple concret de l'adjectif « nehéz » nous permettra de présenter une méthodologie d'exploration et une manière de synthétiser les résultats de la recherche[74].

B) Que dit le dictionnaire ?

« A magyar nyelv értelmező kéziszótára » (Encyclopédie de la langue hongroise) définit comme suit les différents sens de l'adjectif « nehéz »[75] :

nehéz

1. <Objet>, qui, en raison de son poids, nécessite un effort physique considérable pour être ramassé et transporté. *Nehéz csomag (colis lourd), nehéz zsák (sac lourd), nehéz táska (sacoche lourde).*
2. <Tissu> épais. *Nehéz szövet (tissu lourd), nehéz brokát (velours lourd).*
3. <Plat, boisson> difficile à digérer, qui alourdit. *Nehéz ételek (plats lourds), nehéz bor (vin lourd)*
4. <Tâche, devoir>, qui demande beaucoup d'efforts pour être réalisée. *Nehéz feladat (tâche difficile), nehéz munka (travail difficile).*
5. <Situation, circonstances> éprouvante(s) nécessitant force et effort. *Nehéz helyzet (situation difficile), nehéz körülmények (circonstances difficiles).*

[74] Pour un autre exemple avec « jó » (bon), un autre adjectif à usages multiples, voir Szita (2021).
[75] Dans la présentation ci-dessus nous nous restreignons à six sens sur quatorze car notre principal but est de présenter l'intérêt de la méthodologie proposée. En outre, les autres usages sont spécifiques (usage dans un domaine particulier, usage rare) et, en partie, archaïques.

6. <Période de temps ou situation> associée à de grandes difficultés et souffrances. *Nehéz időszak (période difficile), nehéz napok (journées difficiles), nehéz pillanatok (moments difficiles).*

À regarder cette catégorisation de plus près, l'interconnexion du sens et des éléments lexicaux entourant le mot-clé devient évidente. *Les catégories se basent avant tout sur le sens dérivant de l'ensemble de l'adjectif et de ses collocatifs* qui se complètent et se limitent pour fournir des significations non ambiguës.

C) Que dit le corpus ?

Dans ce qui suit, les usages de l'adjectif « nehéz » (difficile, lourd) seront présentés sur la base d'analyse de corpus du « huTenTen12 », complétée par celle du Corpus national du hongrois.

1) « nehéz » + infinitif

Le corpus « huTenTen12 » et le Corpus national du hongrois listent un grand nombre d'expressions avec « nehéz » + infinitif (il est difficile à/de + infinitif). La liste suivante présente les infinitifs accompagnant fréquemment cet adjectif. Les verbes apparaissent dans les deux corpus comme les plus fréquents, avec une occurrence supérieure à 500. Les chiffres font référence aux occurrences dans le corpus « huTenTen12 » mais les résultats dans le Corpus national du hongrois correspondent largement à ces valeurs[76] (v. tableau 69) :

« (nem) nehéz » + INF (216 332, 68/million) **Il (n') est (pas) difficile à + INF**			
elképzelni (6 898 dont 454 négatif)	*imaginer*	meghatározni (2 257 dont 12 négatif)	*déterminer*
megmondani (4 944, dont 23 négatif)	*dire*	belátni (1 337, dont 455 négatif)	*réaliser, voir*
eldönteni (4 316, dont 41 négatif)	*décider*	elmagyarázni (976, dont 12 négatif)	*expliquer*
megérteni (3 233, dont 209 négatif)	*comprendre*	megszokni (964, dont 18 négatif)	*s'habituer*
elhinni (3 120, dont 50 négatif)	*croire*	választani (1 027, dont 28 négatif)	*choisir*

[76] Les écarts s'expriment par les ordres différents de certains verbes mais toujours à l'intérieur de la liste et avec un écart toujours inférieur à 10% de fréquence relative.

megjósolni (2 240, dont 264 négatif)	*prédire*	megemészteni (554, dont 3 négatif)	*digérer*
elfogadni (2 145 dont 27 négatif)	*accepter*	elmondani (519, dont 1 négatif)	*en parler*

Tableau 69 : Les infinitifs les plus fréquemment associés à l'adjectif « nehéz » dans le corpus « huTenTen12 » et dans le « Corpus national du hongrois ».

Les exemples peuvent être divisés en trois groupes du point de vue sémantique :

- Le groupe 1 contient des verbes qui décrivent des activités cognitives (imaginer, croire, comprendre).
- Le groupe 2 rassemble des verbes qui font référence à une action engageant un travail d'analyse et un choix (décider, déterminer, choisir).
- Le groupe 3 est formé de verbes impliquant l'acceptation d'un fait accompli (accepter, digérer).

Ces trois groupes se distinguent également quant à leurs colligations : la forme négative apparaît plus fréquemment dans le cas des verbes qui font appel à la capacité cognitive de l'interlocuteur – le message (pragmatique) étant que les faits présentés sont tellement clairs et évidents que tout le monde est capable de les comprendre ou de les imaginer. Les expressions sont surtout utilisées au présent et au passé, principalement à l'indicatif. L'ordre typique des mots (NEG + « nehéz » + INF) est également reconnaissable et l'hypothèse selon laquelle les adjectifs précèdent généralement l'infinitif dans de telles constructions est confirmée par une analyse de corpus plus poussée. L'expression est généralement suivie d'une virgule et la conjonction « hogy » (que) ou un mot interrogatif[77].

2) « nehéz » + nom

L'adjectif « nehéz » est présent 920 507 fois dans « huTenTen12 » (394/million). Dans la liste des adjectifs les plus fréquents, il occupe la 30e place, il est donc extrêmement courant dans le corpus.

Dans la première étape de l'analyse, nous avons extrait les combinaisons de mots typiques constituées de deux composantes (bi-grams) : adjectif et nom. Pour identifier les collocations,

[77] Il convient de noter que cet usage fréquent n'apparaît pas dans l'encyclopédie (voir plus haut).

l'outil Word Sketch pour « huTenTen12 » et la liste des collocations dans le Corpus national du hongrois ont été utilisés. Le tableau 70 présente les bi-grams les plus répandus (plus de 500 occurrences, les chiffres indiquent le nombre d'occurrences dans « huTenTen12 »).

helyzet (53 629)	*situation*	idők (2 518)	*temps*
dolog (22 338)	*chose*	szívvel (2 262)	*au cœur lourd*
feladat (20 937)	*devoir, tâche*	ellenfél (2 100)	*adversaire*
időszak (9 881)	*période*	terep (1 776)	*terrain*
kérdés (7 619)	*question*	sors (1 372)	*destin, vie*
körülmények (6 107)	*circonstances*	ételek (1 187)	*plats*
munka (5 485)	*travail*	téma (992)	*sujet*
döntés (5 170)	*décision*	ember (953)	*homme*
ügy (3 438)	*affaire*	szülés (942)	*accouchement*
napok (2 505)	*jours*	pillanatok (625)	*moments*
nap (2 500)	*journée*		

Tableau 70 : Les noms les plus fréquents suivant l'adjectif « nehéz ».

Les collocations répertoriées peuvent être divisées en quatre groupes d'après leurs propriétés sémantiques :

(1) Le nom décrit une période de temps : « nehéz idők » (temps difficiles), « nehéz pillanatok » (moments difficiles), « nehéz nap » (journée difficile), « nehéz napok » (jours difficiles), « nehéz időszak » (période de temps difficile), « nehéz gyerekkor » (enfance difficile).

(2) Le nom décrit des circonstances ou une situation : « nehéz helyzet » (situation difficile), « nehéz ügy » (affaire difficile), nehéz körülmények » (circonstances difficiles).

(3) Le nom décrit une tâche qui demande un effort cognitif : « nehéz munka » (travail difficile), « nehéz feladat » (tâche difficile), « nehéz dolog » (chose difficile)

(4) Le nom décrit une situation où la réponse adéquate demande un effort cognitif : « nehéz ügy » (affaire difficile), « nehéz kérdés » (question difficile).

(5) Expressions idiomatiques : « nehéz ember » (personne difficile), « nehéz természet » (nature difficile), « nehéz eset » (cas / personne difficile), « nehéz ételek » (plats lourds), « nehéz szívvel » (au cœur lourd).

Au niveau grammatical, nous observons que certains noms sont utilisés au pluriel « nehéz körülmények » (situations difficiles), « nehéz pillanatok » (moments difficiles), d'autres optent pour le singulier. Nous trouvons également un mot, le nom « nap » (jour, journée) qui est utilisé au singulier ainsi qu'au pluriel. D'après Hoey (2005), les différentes formes s'associent à des usages différents, ces deux formes grammaticales devraient donc indiquer deux modes différents d'utilisation. À l'étape suivante nous examinerons l'environnement textuel plus large du mot au singulier et au pluriel pour voir si tel est le cas.

2.1) « Nehéz nap » (journée difficile) versus « nehéz napok » (jours difficiles) — seulement une différence grammaticale ?

La répartition du singulier et du pluriel dans le corpus de « huTenTen12 » est illustrée par le tableau suivant (tableau 71) :

nap* (au total : 2500)	napok* (au total : 2505)
nehéz nap (1560)	napok (847)
nehéz nap + possessif (677) (Sg1 : 279; Sg2 : 51; Sg3 : 209, Pl1 : 108, Pl2 : 6, Pl3 : 24)	napokat (500)
	napokon (360)
	napokban (352)
	napokra (129)
Avec d'autres terminaisons de cas : napot, napon stb. (266)	Avec d'autres terminaisons (24)
	napjai* (293) (~-m 76, ~-d 17, ~ 134, ~-nk 43, ~-tok 1, ~-ik 22)

Tableau 71 : La répartition du « nap* » et « napok* » dans le corpus de « huTenTen12 ».

Le nombre total des occurrences des deux unités multi-lexicales étant presque identique, nous ne pouvons pas séparer les deux utilisations par leur fréquence dans le corpus. Une étude de l'environnement textuel est nécessaire pour déterminer s'il s'agit alors d'une différence purement grammaticale qui n'affecte pas le sens/l'usage du mot. Nous examinerons donc si dans le premier cas, le locuteur fait simplement référence à *une* journée difficile et dans le deuxième cas à *plusieurs* journées.

En élargissant les collocations à deux composantes en unités multi-lexicales constituées de trois ou quatre composantes (tri-grams et quatre-grams), certaines particularités liées à l'usage du singulier et du pluriel émergent. Le tableau 72 présente les unités multi-lexicales les plus fréquentes

avec « nehéz nap » (journée difficile) dans le corpus de « huTenTen12 », organisées en fonction de leurs fréquences :

nehéz napom/-od/-ja stb. lesz/van/volt (677)	*qqn (je/tu/il etc.) aura/a/avait* ***une journée difficile***
egy nehéz nap után hazaérkezik / ki szeretne kapcsolódni / ... (531)	*il rentre/veut se détendre après* ***une journée difficile***
holnap **nehéz nap** vár Bencére/ránk/rám (61)	
nehéz nap áll/van mögötted (40)	***une journée difficile*** *nous attend demain*
nehéz nap elé nézünk (30)	*tu as* ***une journée difficile*** *derrière toi*
nehéz nap előtt állunk (27)	*nous aurons* ***une journée difficile*** *(demain)*
nehéz napon vagyok túl (106)	*on est devant* ***une journée difficile***
	j'ai passé ***une journée difficile***

Tableau 72 : Unités multi-lexicales fréquentes avec « nehéz nap » (journée difficile), organisées par suffixe.

L'analyse indique que les unités multi-lexicales les plus régulièrement utilisées sont *«* nehéz napom/-od/-ja stb. lesz/van/volt » (je/tu/il etc. + avoir (futur/présent/passé) + une journée difficile) et « egy nehéz nap után » (après une journée difficile). Dans toutes les constructions observées dans le corpus, la collocation fait référence à *une journée fatigante et longue*. Le vocabulaire des unités multi-lexicales n'est pas complexe ; nous y trouvons essentiellement des mots courants mais elles se caractérisent par une forte idiomaticité. Les unités multi-lexicales les plus usitées avec « nehéz napok » sont les suivantes (tableau 73) :

nehéz napok jöttek/jönnek (49)	***des jours difficiles*** *s'annoncent*
nehéz napok (lehettek/voltak) ezek nekik (36)	*ils ont traversé* ***des jours difficiles***
nehéz napjaim voltak (96)	*j'ai vécu* ***des journées difficiles***
nehéz napokat él át mostanában (249)	*il a vécu* ***des jours difficiles*** *ces temps-ci*
... segít átvészelni **a nehezebb napokat** (38)	*... aide à traverser* ***les jours plus difficiles***
Könnyebb elviselni **a nehéz napokat**, ha ... (11)	*Il est plus facile de supporter* ***des jours difficiles*** *si ...*
... a barát, aki szeretetével átsegít **a nehezebb napokon** (27)	*... l'ami qui t'aide à traverser* ***les jours plus difficiles***
... túljutottunk **a nehezebb napokon**. (5)	*...* ***les jours plus difficiles*** *sont derrière nous*

ezekben a nehéz napokban fontos, hogy / sokat segített az, hogy … (91)	***dans ces jours difficiles**, il est important / ça m'a beaucoup aidé que ...*

Tableau 73 : Unités multi-lexicales avec « nehéz napok » (journées difficiles), organisées par suffixe.

Le tableau 73 montre que les éléments lexicaux fréquents – avant tout les verbes – utilisés avec « nehéz napok » (pluriel) ne sont pas les mêmes que dans le cas de « nehéz nap » (singulier). L'usage de « napok » au pluriel indique une période plus longue, sombre, impliquant des difficultés psychologiques et de la souffrance, plutôt qu'une période de fatigue. La plupart des verbes qui l'accompagnent signalent un effort (affronter, supporter, aider à dépasser, survivre), alors que les verbes associés au singulier (« nap ») indiquent une attitude plutôt passive ou neutre (être, avoir, attendre).

Cette comparaison révèle ainsi que les deux collocations ne sont pas simplement des variations entre singulier et pluriel en fonction de la longueur de la période considérée comme étant difficile par le locuteur, mais le choix entre le singulier et le pluriel implique des environnements textuels différents. Dans le premier cas, le constat d'une journée difficile – passée ou à venir – est suivi d'une description des événements qui justifie cette caractérisation : une épreuve à l'université, une longue journée de réunions, par exemple. Dans le cas de « journées difficiles », les événements semblent avoir un impact plus profond. En effet, il ne s'agit pas d'un événement ponctuel à dépasser mais d'une chaîne d'événements ou d'un événement grave, par exemple : vivre des difficultés (divorce, situation d'échecs) et consulter un psychiatre, supporter quelque chose (un climat désagréable au travail), être soumis à un traitement médical lourd, et autres.

2.2) « Nehéz helyzet » (situation difficile), « nehéz körülmények » (circonstances difficiles) – deux synonymes ?

Dans cette partie, nous étudierons deux unités multi-lexicales fréquentes appartenant à la même catégorie sémantique : « nehéz helyzet » (situation difficile) et « nehéz körülmények » (circonstances difficiles). Selon le Thésaurus de Sketch Engine, « körülmények » et « helyzet » sont proches au niveau de l'usage (1 519 226 occurrences dans le corpus) et peuvent être considérés comme des synonymes. Nous chercherons à identifier dans ce qui suit en quoi leurs environnements textuels sont similaires (ou différents) et s'ils démontrent une proximité au niveau des unités multi-lexicales plus longues.

Les unités multi-lexicales précisent les possibilités d'usage et fournissent les éléments lexicaux qui sont nécessaires à leur réalisation. Les tri-grams et les quadri-grams les plus courants avec la collocation « nehéz helyzet » (situation difficile) sont listés dans le tableau 74 :

X nehéz helyzetbe hoz Y-t* (2964)	
X nehéz helyzetbe hozta Y-t	*X a mis Y dans une situation difficile*
X nehéz helyzetbe hozza Y-t	*X met Y dans une situation difficile*
X nehéz helyzetbe hozhatja Y-t	*X peut mettre Y dans une situation difficile*
*X nehéz helyzetbe kerül** (7312)	
X nehéz helyzetbe került	*X s'est retrouvé dans une situation difficile*
X nehéz helyzetbe kerül	*X se retrouve dans une situation difficile*
X nehéz helyzetbe kerülhet	*X peut se retrouver dans une situation difficile*
X nehéz helyzetben van/volt (7262)	
X nehéz helyzetben van	*X est dans une situation difficile*
X nehéz helyzetben volt	*X était dans une situation difficile*

Tableau 74 : 3- et 4-grams avec la collocation « nehéz helyzet » (situation difficile).

Le verbe est un élément crucial dans ces unités multi-lexicales car il détermine la structure de la phrase et la fonction grammaticale de « nehéz helyzet » (situation difficile). Observer ces unités offre des avantages à plusieurs niveaux. D'une part, inclure le verbe dans les observations contribue à préciser les modes d'utilisation de « nehéz helyzet » ainsi que ses propriétés sémantiques (quelque chose de désagréable arrive à quelqu'un ou quelqu'un fait l'expérience d'une difficulté). Des schémas grammaticaux tels que l'ordre des mots, les temps et les modes typiques du verbe émergent également : La plupart des énoncés sont au passé, suivis du présent. Au présent, le verbe apparaît souvent avec l'affixe « -hat/-het » (comparable au verbe modal ~ « peut » en français) et indique donc plutôt une supposition, une éventualité ou une conséquence possible qu'un fait.

Les deuxième et troisième groupes (« X nehéz helyzetbe került », « X nehéz helyzetben van ») sont des variantes par rapport au premier groupe d'exemples. Ici, le sujet subit la situation et il peut être, tout comme les compléments d'objet direct du premier groupe, une personne ou un groupe de personnes (le village, le parti, les entreprises, les écoles, le gouvernement). Il ne s'agit pas

forcément des conséquences financières (ce qui était un usage caractéristique dans le premier cas) mais de n'importe quelle situation qui implique une décision difficile ou un problème.

En examinant de près les unités multi-lexicales, nous obtenons d'autres informations pertinentes sur l'utilisation du mot choisi. On peut observer en effet non seulement l'ordre typique des trois éléments (« nehéz » + Nsuff + V), mais aussi la position de l'expression dans la phrase (sujet + « nehéz » + Nsuff + V /+ COD/). Puisque l'ordre des mots est un aspect complexe du hongrois, ces données peuvent être extrêmement utiles car les tendances observées ici s'appliqueront également à d'autres contextes avec d'autres unités multi-lexicales.

Le corpus permet également d'identifier les sujets et les compléments d'objet direct fréquents : deux groupes se dégagent. Les unités multi-lexicales du premier groupe montrent certaines similarités sémantiques : à part le pronom démonstratif « ez » (cela) qui a un sens très général, les autres sujets indiquent un fait légal ou économique (décision, régulation, crise) et les compléments d'objet direct sont les personnes ou les groupes de personnes affectés. La phrase entière implique le plus souvent une conséquence financière désavantageuse (dans 84 % des cas) (tableau 75).

Ez / a válság / a határozat / a rendelet / a rendelkezések / a döntés / az új törvény **nehéz helyzetbe hozhatja** X-et (személy) / a gazdaságot / a vállalatokat / az ország lakosait / a fogyasztókat / a lakosságot / a pártokat / az intézményeket / az expedíciót.	*Cela / la crise / le décret / la régulation / les mesures / la décision / la nouvelle loi* ***peut mettre*** *X (personne) / l'économie / les entreprises / les habitants du pays / les consommateurs / la population / les partis (politiques) / les institutions / l'expédition* ***dans une situation difficile.***

Tableau 75 : Les sujets et les CODs les plus fréquents de l'unité multi-lexicale « nehéz helyzetbe hozhatja » quand il s'agit d'une conséqauence financière désavantageuse.

Un plus petit groupe des occurrences (env. 12%) indique des difficultés dans le cadre d'une compétition (tableau 76) :

X sérülése/távozása/ visszalépése nehéz helyzetbe hozhatja a csapatot/a céget/a vállalatot.	La blessure/Le départ/La démission de X peut mettre l'équipe/la companie/l'entreprise dans une situation difficile.

Tableau 76 : Les CODs les plus fréquents de l'unité multi-lexicale « nehéz helyzetbe hozhatja » quand il s'agit d'une compétition.

L'unité multi-lexicale « nehéz körülmények » (circonstances difficiles), qui fait partie du même groupe sémantique, émerge dans des environnements textuels différents que « nehéz helyzet » (situation difficile). Nous le rencontrons avec une fréquence largement inférieure à celle de « nehéz helyzet ». « Nehéz helyzet » émerge 42 238 fois dans le corpus alors que « nehéz körülmények » n'est utilisé que dans 6 107 cas. Les unités multi-lexicales les plus usitées avec « nehéz körülmények » sont les suivantes (tableau 77) :

nehéz körülmények között + V (3839)	*dans des circonstances difficiles* + V
nehéz körülmények között él (1264)	*il vit*
nehéz körülmények között nőtt fel (39)	*il a grandi / passé son enfance*
nehéz körülmények között kellett *dolgoznia, helyt állnia, újrakezdenie az életét* (62)	*il a dû travailler/s'établir/recommencer sa vie* ***dans des circonstances difficiles***
nehéz körülmények között is + V (724)	*même dans des circonstances difficiles* + V
nehéz körülmények között is sikeres lehet	*il peut réussir* ***même dans des / malgré les circonstances difficiles***
nehéz körülmények között is kitartóan dolgozik	*il travaille assidûment* ***même dans des / malgré les circonstances difficiles***
nehéz körülmények között is lehet eredményt elérni	*on peut obtenir des résultats* ***même dans des / malgré les circonstances difficiles***

Tableau 77 : Unités multi-lexicales typiques avec « nehéz körülmények között ».

Le nombre d'unités multi-lexicales est plutôt limité dans le cas de « nehéz körülmények ». Seules deux ressortent avec une fréquence notable, toutes deux contenant la postposition « között » (ici : dans) correspondant à environ 75% des occurrences totales. Dans l'unité « nehéz körülmények között » + V, le verbe « vivre » prévaut à d'autres verbes (33%). Tous les autres verbes sont représentés avec nettement moins d'occurrences (v. deux exemples ci-dessus). L'unité « nehéz körülmények között is » + V peut s'associer à une grande variété de verbes et de constructions qui ont une propriété sémantique en commun : ils indiquent un résultat positif obtenu (progrès, motivation maintenue, réussite académique) malgré les circonstances difficiles.

Les exemples indiquent que les expressions « nehéz helyzet » et « nehéz körülmények » sont utilisées dans des environnements textuels différents avec des fréquences différentes, même s'il existe une proximité sémantique entre les noms « helyzet » et « körülmények ».

En suivant la même méthodologie, nous pouvons identifier les unités multi-lexicales plus complexes, contenant un verbe (ADJ + N + V) qui complètent celles à deux composantes (ADJ + N). Voici quelques exemples (tableau 78) :

nehéz idők, nehéz napok	*des temps difficiles, des jours difficiles*
nehéz időket/napokat élünk	*nous vivons des temps/jours difficiles (1)*
nehéz idők/napok járnak	*nous vivons des temps/jours difficiles (2)*
nehéz idők/napok jönnek	*des temps/jours difficiles s'annoncent*
X (dolog, tárgy) jól jön/jöhet még **a nehéz időkben.**	*X peut servir dans les temps difficiles*
nehéz ügy	*affaire difficile*
Ez **nehéz ügy**.	*C'est une affaire difficile.*
Ez **nehéz ügynek** tűnik.	*Cela semble être une affaire difficile.*
nehéz dolog	*chose difficile*
Nehéz dolgok ezek.	*Ce sont des choses difficiles.*
Xnek **nehéz dolga van/lesz**	*X a/aura une tâche difficile.*
Nem volt nehéz dolgom.	*Je n'avais pas de tâche difficile.*
nehéz természet	*caractère difficile*
nehéz természete van	*il a un caractère difficile*

Tableau 78 : Unités multi-lexicales typiques plus longues (ADJ + N + V).

Dans cette section, nous avons présenté une méthodologie capable d'explorer des corpus dans le cadre pédagogique dès qu'il s'agit d'observer les utilisations différentes des mots à usages multiples. L'analyse se déroule en plusieurs étapes, à partir des unités à deux composantes vers les unités multi-lexicales plus longues et plus complexes. Cette manière d'explorer les données peut apporter des éléments complémentaires aux informations proposées par l'encyclopédie, telles que des informations statistiques sur la fréquence, qui aident à réorganiser les sens listés, avec plus de collocations typiques et plus d'exemples d'usages typiques.

D) Que veut dire le mot « nehéz » ?

Dans les pages suivantes, à partir de l'exemple concret de l'usage de l'adjectif « nehéz », nous illustrerons une présentation possible des résultats d'une analyse de corpus. Nous pouvons visualiser les usages répertoriés plus haut de manière systématique en utilisant les catégories définies par Sinclair (1991) et Hoey (2005), présentées dans le chapitre 4.

La première catégorie concerne les éléments lexicaux s'associant fréquemment à l'adjectif « nehéz ». Nous utilisons ici le terme « collocation » tel que défini dans le chapitre 3, en incluant aussi bien des modificateurs que des noms et des verbes, précédés ou suivis par « nehéz ». Ces éléments lexicaux déterminent largement les « Composantes sémantiques » de l'unité multi-lexicale. Au-delà de la précision du sens de l'unité, les propriétés sémantiques peuvent également faire référence à l'attitude du locuteur et à un jugement de valeur (positif ou négatif) de sa part. La rubrique « colligations typiques » listent les propriétés grammaticales des unités : l'ordre des mots, les structures syntaxiques typiques dans lesquelles le mot choisi émerge, les parties de discours dans son entourage et d'autres caractéristiques. Les « composantes pragmatiques » résument la fonction des unités multi-lexicales dans le discours et indiquent, lorsque cela est possible, l'endroit où elles ont tendance à apparaître dans le texte (voir le chapitre 4 pour plus de détails). Le tableau 79 présente le profil de l'adjectif « nehéz ».

« nehéz » + INF (Il est difficile à/de + INF*)*

Groupe 1 : capacités cognitives

Collocations typiques	« megérteni » (comprendre), « felfogni » (saisir), « belátni » (réaliser), « elképzelni » (imaginer), « megjósolni » (prédire), « elhinni » (croire), « megmondani » (dire) et d'autres verbes relatifs aux capacités cognitives et à l'imagination.
Colligations typiques	« nehéz » + INF, hogy (Il est difficile de + INF que) « nem nehéz » + INF, hogy (Il n'est pas difficile à + INF que)
Composantes sémantiques	Exprime la difficulté ou la facilité de saisir quelque chose par les moyens cognitifs. La négation fait appel à la capacité de l'interlocuteur de comprendre ou d'imaginer quelque chose que le locuteur considère être simple.
Composantes pragmatiques	Introduit la description plus détaillée d'un fait ou d'une problématique.

Groupe 2 : décision

Collocations typiques	« eldönteni » (décider), « meghatározni » (déterminer), « választani » (choisir) et d'autres verbes décrivant des activités analytiques impliquant une décision
Colligations typiques	« nehéz » + INF + , hogy (Il est difficile de INF si/que)

Composantes sémantiques	Exprime la difficulté d'analyse précédant un choix. Les verbes décrivent des activités analytiques qui impliquent une décision.
Composantes pragmatiques	Introduit la description plus détaillée d'une problématique complexe.

Groupe 3 : acceptation

Collocations typiques	« elfogadni » (accepter), « megemészteni » (digérer) et d'autres verbes indiquant l'acceptation d'un fait.
Colligations typiques	« nehéz » + INF + , hogy (Il est difficile de INF que) « nehéz » + INF + , hogy …, de … (Il est difficile de INF que ..., mais)
Composantes sémantiques	Exprime la difficulté d'accepter un fait.
Composantes pragmatiques	Introduit la justification de la difficulté d'accepter qqc. Le reste du texte propose souvent des points de vue différents et des solutions.

nehéz + N

Groupe 1 : temps

Collocations typiques	**Noms typiques :** « pillanat » (moment), « idők » (temps), « időszak » (période), « napok » (jours) et d'autres noms indiquant un moment ou une période de temps.
Colligations typiques	ADJ + N
Composantes sémantiques	Les noms indiquent un moment éprouvant ou une période de temps éprouvante. Le verbe annonce l'arrivée ou la présence de cette période.
Composantes pragmatiques	Expression d'un constat négatif. Introduit ou clôt un récit qui justifie ce jugement.

Groupe 2 : circonstances

Collocations typiques	**Noms typiques** : « körülmények » (circonstances), « helyzet » (situation) et d'autres noms faisant référence à des circonstances.
Colligations typiques	ADJ + N

Composantes sémantiques	Fait référence aux circonstances difficiles, souvent d'un point de vue financier. « Nehéz helyzetben van » peut également indiquer la difficulté de décider.
Composantes pragmatiques	Expression d'un constat négatif, souvent suivi ou précédé d'une justification.

Groupe 3 : tâches

Collocations typiques	**Noms typiques :** « munka » (travail), « feladat » (tâche), « kihívás » (challenge) et d'autres noms indiquant une tâche, un travail
Colligations typiques	ADJ + N
Composantes sémantiques	Décrit une tâche ou un travail qui demande de l'effort (intellectuel ou physique).
Composantes pragmatiques	Expression d'un constat négatif. Peut être suivi de l'annonce du succès ou de l'échec lié à la tâche.

Groupe 4 : affaire, question

Collocations typiques	**Noms typiques :** « ügy » (affaire), « kérdés » (question)
Colligations typiques	ADJ + N
Composantes sémantiques	Implique un effort cognitif avant une décision ou une réponse complexe.
Composantes pragmatiques	Expression d'un constat négatif, souvent suivi d'une justification de ce constat.

Groupe 5 : personnes

Collocations typiques	**Noms typiques :** « ember » (homme), « természet » (nature), « eset » (cas) « nehéz ember/eset » (un homme/cas difficile) « nehéz természete van » (a un caractère difficile)
Colligations typiques	ADJ + N
Composantes sémantiques	Utilisé pour décrire une personne. Exprime que la personne est difficile à supporter ou à satisfaire.
Composantes pragmatiques	Expression d'un constat négatif, souvent suivi d'une justification de ce constat.

Tableau 79 : Le profil de l'adjectif « nehéz ».

L'analyse peut s'arrêter là si notre but est de créer un profil de base avec les informations les plus importantes concernant l'usage des unités multi-lexicales à deux et à trois composantes avec l'adjectif « nehéz ». Or, les outils nous permettent d'aller plus loin dans l'analyse pour les cas présentant un intérêt particulier dans le cadre pédagogique. Au cours des pages suivantes, nous proposerons une façon possible d'explorer les unités multi-lexicales plus étendues.

E) Étudier les unités multi-lexicales complexes

Certaines unités multi-lexicales méritent une exploration plus approfondie. À titre illustratif, nous avons choisi « nehéz helyzetbe hoz », « nehéz helyzetbe kerül », « nehéz helyzetben van », trois unités contenant les composantes « nehéz », « helyzet » et un verbe. Un examen de l'environnement textuel plus étendu permet d'identifier les sujets et les compléments d'objet direct fréquents. Lors de l'analyse, deux groupes émergent dont le premier indique une difficulté, financière ou autre, pour une personne ou un groupe de personnes, causée par une forte contrainte extérieure (règle, loi, mesure, crise). Le deuxième groupe se compose de phrases relatives à une situation de compétition sportive ou économique où l'absence ou la blessure d'une personne met une équipe en difficulté (tableau 80).

« nehéz helyzetbe hoz »

Groupe 1 : Sujets et CODs typiques de « nehéz helyzetbe hoz » (76 % des occurrences)

Sujets typiques : Mesure, loi, régulation, crise		**CODs typiques :** Personne ou groupe de personnes
ez	**nehéz helyzetbe hozta**	X-et (személy)
a válság		a gazdaságot
a határozat		a vállalatokat
a rendelet	**nehéz helyzetbe hozhatja**	az ország lakosait
a rendelkezések		a fogyasztókat
a döntés		a lakosságot
az új törvény		a pártokat
		az intézményeket
		az expedíciót
		a pályakezdőket

cela		*X (personne)*
la crise		*l'économie*
le décret		*les entreprises*
la régulation	***a mis dans une situation difficile / peut mettre dans une situation difficile***	*les habitants du pays*
les mesures		*les consommateurs*
la décision		*la population*
la nouvelle loi		*les partis (politiques)*
		les institutions
		l'expédition
		les jeunes diplômés

Groupe 2 : Sujets et CODs typiques de « nehéz helyzetbe hoz » (24 % des occurrences)

Sujets typiques : Blessure, départ de X (personne)		**CODs typiques :** Groupe de personnes
X sérülése	**nehéz helyzetbe hozta**	a csapatot.
X távozása	**nehéz helyzetbe hozhatja**	a céget.
X visszalépése		a vállalatot.
La blessure de X	***a mis dans une situation***	*l'équipe.*
Le départ de X	***difficile / peut mettre dans***	*la companie.*
La démission de X	***une situation difficile***	*l'entreprise.*

Tableau 80 : Les sujets et les CODs typiques dans « nehéz helyzetbe hoz » selon les deux sens de cette unité multi-lexicale.

Il est également possible d'identifier les sujets typiques de « nehéz helyzetben van » (est dans une situation difficile), « nehéz helyzetbe kerül » (se retrouve dans une situation difficile) (tableau 81) :

« nehéz helyzetben van », « nehéz helyzetbe kerül »

Sujets typiques :
Personne ou groupe de personnes

X (személy)	
a gazdaság	
a vállalatok	
az ország lakosai	
a fogyasztók	
a lakosság	
a pártok	**nehéz helyzetben van(nak)**
az intézmények	**nehéz helyzetbe kerül(tek)**
az expedíció	**nehéz helyzetbe kerülhet(nek)**
a pályakezdők	

X (personne)	
l'économie	
les entreprises	***est/sont dans une situation difficile***
les habitants du pays	***s'est retrouvé/se sont retrouvés dans***
les consommateurs	***une situation difficile***
la population	***peut/peuvent se retrouver dans une***
les partis (politiques)	***situation difficile***
les institutions	
l'expédition	
les jeunes diplômés	

Tableau 81 : Les sujets typiques de l'unité multi-lexicale « nehéz helyzetbe kerül ».

Il n'est pas surprenant de retrouver dans cette liste un certain nombre de noms qui sont les compléments d'objet direct de « nehéz helyzetbe hoz » car les phrases ci-dessus présentent le résultat des actions formulées avec ceux-ci.

Le résumé des explorations peut être présenté de la même façon que les unités multi-lexicales plus courtes (tableau 82) :

X nehéz helyzetbe hoz*Y-t (1)

Collocations typiques	**Sujets typiques :** « rendelet » (mesure), « törvény » (loi), « rendelkezés » (régulation), « válság » (crise) **CODs typiques :** personne ou groupe de personnes
Colligations typiques	**Formes de verbes typiques :** « hozza » (met), « hozta » (a mis), « hozhatja » (peut mettre) Majoritairement COD défini et conjugaison définie **Ordre de mots typique :** S + nehéz helyzetbe hoz* + COD
Composantes sémantiques	Met quelqu'un dans une difficulté (financière).
Composantes pragmatiques	Peut clôturer un narratif.

X nehéz helyzetbe hoz*Y-t (2)

Collocations typiques	**Sujets typiques :** « X sérülése » (blessure de X), « X távozása » (départ de X) **CODs typiques :** personne ou groupe de personnes
Colligations typiques	**Formes de verbes typiques :** « hozza » (met), « hozta » (a mis), « hozhatja » (peut mettre) Majoritairement COD défini et conjugaison définie **Sujet** : NPoss **Ordre de mots typique :** S + nehéz helyzetbe hoz* + COD
Composantes sémantiques	Met quelqu'un dans une difficulté.
Composantes pragmatiques	Peut introduire ou clôturer un narratif.

« nehéz helyzetbe kerül », « nehéz helyzetben van »

Collocations typiques	**Sujets typiques :** personne ou groupe de personnes
Colligations typiques	« Y nehéz helyzetben van, mert » (Y est dans une situation difficile parce que). De préférence au présent. « Y nehéz helyzetbe került, amikor » (Y a été mis dans une situation difficile quand). De préférence au passé. « Y nehéz helyzetbe kerülhet, ha » (Y pourrait être mis dans une situation difficile si). Au présent avec l'affixe « -hAt », suivi d'une condition.
Composantes sémantiques	Se retrouve dans une difficulté (financière)
Composantes pragmatiques	Introduit un narratif en indiquant une explication.

Tableau 82 : Profil des unités multi-lexicales « nehéz helyzetbe hoz » et « nehéz helyzetbe kerül ».

Ces explorations montrent, entre autres, la possibilité d'approfondir l'analyse et de découvrir des schémas plus longs et plus complexes. L'étude de ces unités multi-lexicales peut enrichir le vocabulaire de l'apprenant, le rendre capable d'identifier lui-même des schémas et de développer ses compétences linguistiques à travers la multitude d'exemples auxquels il a été exposé pendant ce travail.

Dans ce chapitre, nous avons livré une analyse des usages des adjectifs « nehéz ». Les explorations ont mis en évidence la relation entre la grammaire, le vocabulaire et le sens du mot dans des unités multi-lexicales différentes. Nous avons démontré la richesse d'informations qui peut découler de l'analyse de corpus : l'enrichissement des compétences linguistiques par le biais des collocatifs fréquents, des caractéristiques grammaticales, sémantiques et pragmatiques appartenant à des utilisations différentes. Au-delà des renseignements liés à la langue en question, le plus grand bénéfice de ce procédé est qu'en apprenant à connecter ces domaines, l'utilisateur développe un regard neuf sur cette langue et une meilleure compréhension de ce que veut dire connaître un mot.

Chapitre 9 : « Quelle est la différence ? » Les synonymes « tűnik » et « látszik »

« L'ambiguïté dans le langage réel est (...) plutôt rare »
(Michael Hoey 2005 : 81, notre traduction)

Les chapitres 9 et 10 examineront dans quelle mesure l'analyse de corpus peut aider l'enseignant à définir, à présenter et à illustrer la différence entre les synonymes. Dans ce chapitre, nous étudierons les verbes « tűnik » et « látszik » (comparables aux verbes français « sembler » et « paraître ») dont l'usage pose problème aux apprenants en raison de leur interchangeabilité dans certains cas et dans d'autres, la nette préférence des natifs pour l'un des deux. Dans le chapitre suivant, nous nous concentrerons sur deux verbes à préfixe, « megjön » et « eljön », traduisibles par « venir » ou « arriver », selon le contexte et l'environnement textuel et avec des règles d'usage plutôt floues.

Nous commençerons les explorations relatives aux verbes « tűnik » et « látszik » par les définitions et exemples tirés du « A magyar nyelv értelmező kéziszótára » (Dictionnaire monolingue de la langue hongroise) qui fournit une présentation riche des sens possibles de ces deux mots et l'intuition des natifs. Nous analyserons ensuite les données linguistiques dans les corpus sélectionnés (voir le chapitre 7). Nous terminerons les recherches en répertoriant les caractéristiques observées.

Notre principal objectif dans ce procédé est de démontrer que les mots à sens similaires apparaissent dans des environnements textuels différents et montrent des caractéristiques différentes. Leur analyse basée sur le corpus contribue ainsi à relever les ambiguïtés concernant leur usage. Une partie des résultats peut être présentée aux apprenants, illustrée avec un grand nombre d'exemples sélectionnés et adaptés, si besoin[78].

A) Les synonymes « tűnik » et « látszik » (~ *sembler* et *paraître*)

Le travail sur les synonymes est un point central de l'analyse de corpus. Les études dans ce domaine cherchent à démontrer que l'ambiguïté dans l'usage langagier est plutôt rare car l'environnement textuel clarifie le sens (Hoey 2005 ; Sánchez-Cárdenas 2010 ; Sinclair 1991, 2004b ; Stubbs 2001 ; Taylor 2012). Cet environnement semble être différent dans le cas de deux synonymes qui, à

[78] Voir aussi chapitre 12 pour les considérations pédagogiques.

première vue, « veulent dire la même chose », ce qui a pour conséquence que les synonymes sont rarement interchangeables. Les analyses de synonymes effectuées en utilisant des corpus se multiplient rapidement, en particulier pour l'anglais mais aussi pour d'autres langues. Liu (2010) examine les différences dans l'usage de « chief », « major », « major », « primary » et « principal » en combinaison avec le nom « concern » en démontrant que la plupart des collocations sont motivées, si elles sont considérées à la lumière des mappages sémantiques des éléments-clés impliqués. Fondé sur ces considérations, son article plaide pour la nécessité d'inclure une analyse cognitive dans l'apprentissage des collocations, en plus de l'observation, de la mémorisation et d'autres activités. Kamber (2011) utilise le verbe « regarder » et ses « (quasi-)synonymes » comme il les nomme pour démontrer le phénomène de la désambigüisation. En analysant l'environnement textuel des verbes « regarder », « contempler », « dévisager », « épier », « scruter » et « toiser », il finit par conclure que l'usage de ces synonymes est différent et ils ne sont que très rarement interchangeables.

Markova (2012) analyse plusieurs paires de synonymes allemands tels que « kalt » et « kühl », « nett » et « angenehm ». Dans son étude, elle démontre que les adjectifs considérés comme synonymes montrent des différences significatives quant à leur comportement dès lors qu'il s'agit des énoncés réellement produits. Elle utilise des antonymes (qui devraient correspondre à tous les synonymes s'ils étaient vraiment interchangeables) ainsi que des formes grammaticalement modifiées telles que les synonymes précédés du mot négatif « nicht » pour soutenir cet argument et souligner la fonction de l'environnement textuel pour enlever toute ambigüité. Hoey (2014) avance également l'argument que le locuteur n'est pas totalement libre dans la construction de son discours, c'est-à-dire dans le choix de ses mots, y compris ses choix entre des synonymes. Cette sélection est, en fait, factice car l'environnement textuel n'accepte, en général, qu'un seul mot possible.

Pour les natifs, ces choix peuvent être évidents et, par conséquent, limités à une option ou deux mais les apprenants sont confrontés à des décisions plus problématiques. Sans l'expérience linguistique accumulée que possèdent les natifs et les utilisateurs experts de la langue-cible, leurs choix ne sont pas plus limités mais plus libres. De ce fait, ils peuvent se tromper et ne pas communiquer avec précision en employant les synonymes avec une flexibilité (de façon interchangeable) qu'un natif trouverait inhabituel ou même linguistiquement inacceptable. (cf. Hoey 2014 ; Partington, 1998 : 30 ; Tognini-Bonelli, 2001 : 34).

Les verbes « tűnik » et « látszik », comme les verbes français « sembler » et « paraître » s'utilisent à première vue de façon similaire et expriment des concepts similaires. Or, les deux verbes ne sont pas systématiquement interchangeables. Les unités multi-lexicales dont ils font souvent partie et les schémas sémantiques et grammaticaux caractéristiques peuvent contribuer à éclairer leurs différences d'usage comme nous le verrons par la suite.

B) Que dit le dictionnaire ? Que révèle l'intuition des natifs ?

Nous examinerons tout d'abord les définitions dans le « Dictionnaire monolingue de la langue hongroise » (2021, en ligne), outil précieux lorsque nous voulons explorer les caractéristiques d'usage d'un élément lexical. Nous y trouvons les explications suivantes :

látszik

1. Ce qui devient visible en regardant quelque chose : *Látszik rajta, hogy jó ember.* *(Ça se voit qu'il est une bonne personne.)* *Látszik, hogy fáradt.* (*On voit qu'il est fatigué.) A leveléből látszik, hogy szomorú. (On peut lire dans sa lettre qu'il est triste.)*

2. Quelqu'un/quelque chose donne une certaine impression : *Bosszúsnak / fáradtnak / haragosnak látszik. (Il a l'air ennuyé / fatigué / en colère.)*

3. Donne une impression qui ne correspond pas à la réalité : *Fiatalabbnak látszik a koránál. (Il paraît plus jeune que son âge.) A hegyek messziről sötétkéknek látszanak. (Les montagnes semblent bleu foncé au loin.) Úgy látszik, mintha a Nap a Föld körül forogna. (On dirait/On a l'impression que le soleil tourne autour de la terre.)*

tűnik

Quelqu'un/quelque chose semble/apparaît à quelqu'un/quelque chose : *Úgy tűnik nekem, hogy ... (Il me semble que ...) Álomnak tűnik. (Il me semble que c'est un rêve.) Csodának tűnik. (Il me semble que c'est un miracle.) Jónak tűnik. (Il/Ça me semble bon.) Szépnek tűnik. (Il/Ça me semble beau.)*

En lisant les descriptions et les exemples, il est évident que ces définitions ne sont pas d'une grande utilité pour les apprenants et les enseignants de hongrois (qui ne sont d'ailleurs pas le premier public de l'ouvrage) si notre questionnement concerne la différence d'usage entre les deux verbes. La définition 2 de « látszik » est la même que la définition de « tűnik » et les exemples ne suffisent pas pour distinguer l'utilisation des deux mots. L'analyse des corpus démontrera aussi que les définitions et les exemples ci-dessus ne correspondent qu'à une partie des usages.

Pour avoir un aperçu (non représentatif) des manières dont les enseignants répondent à la question « Quelle est la différence entre les verbes « tűnik » et « látszik » ? » en s'appuyant sur leur intuition, nous avons également demandé à six professeurs expérimentés de hongrois de fournir une explication. Voici le résumé de leurs réponses :

1. « Tűnik » exprime une sorte de réalité supposée, une impression ; « látszik » reflète une opinion basée sur un fait concret, observable.
2. « Tűnik » a une certaine subjectivité. La première phrase qui me vient à l'esprit : *Nekem úgy tűnik, hogy ...* (*Il me semble que ...*) L'autre usage est l'expression prudente d'une opinion personnelle : *Egyértelműnek tűnik, hogy ... (Il semble clair que ...).* « Látszik » : par exemple *Világosan, egyértelműen látszik (On peut voir clairement, sans ambiguïté).* Il implique une connaissance plus certaine, quelque chose qui peut être reconnu de l'extérieur, pas une impression du locuteur.
3. « Látszik » suggère une visualité plus concrète : *Innen már látszik a vár.* (*D'ici on peut voir le château.) Már messziről látszik, hogy ki jön az úton. (On peut voir de loin qui arrive sur la route.) Batman nem látszik a sötétben, mert fekete. (On ne peut pas voir Batman dans le noir parce qu'il est noir.).* « Tűnik » est moins concret : *Sikeres embernek tűnik. (Il semble être un homme qui a du succès*).
4. « Látszik » est plus statique, tandis que « tűnik » capte l'impression du moment. Les deux peuvent être utilisés pour décrire une impression donnée, qu'elle soit concrète ou métaphorique, positive ou négative.
5. « Látszik » est utilisé dans les cas où quelque chose peut/pourrait être visible à l'œil nu, il y a donc des signes visibles. Par exemple : *füstnek látszik (on dirait de la fumée), 25 évesnek látszik (elle doit avoir 25 ans), nagynak látszik (il semble grand).* « Tűnik » a un sens plus figuré : *könnyebbnek tűnik az út (la route semble plus facile), igazságtalannak tűnik (cela/il semble injuste), feleslegesnek tűnik (cela semble inutile), békésnek tűnik (semble paisible).*
6. La première expression qui me vient à l'esprit est *Úgy tűnik/látszik, hogy ...* (*Il semble que ...).* Ici, les deux mots sont synonymes. Dans d'autres cas, « látszik » fait plutôt référence aux choses visibles à l'œil nu, alors que « tűnik » est une sorte d'impression.

En lisant ces réponses, nous pouvons constater que, bien qu'il s'agisse de professeurs expérimentés de langues dont le hongrois est la première langue, leurs réponses ne se recoupent pas parfaitement. Leurs affirmations, comme nous le verrons plus loin, sont correctes, mais elles mettent l'accent sur des aspects différents de l'usage. Cependant, ils ont tous un point commun : ils s'appuient sur

des connaissances profondes acquises par l'usage au fil des ans, expérience linguistique dont les locuteurs non natifs ne disposent pas (Ellis et al. 2015 ; Hanks 2013 ; Taylor 2012 ; Tyler 2010). Ces connaissances implicites, non réfléchies sont mises à notre disposition en tant que matière première dans un corpus, nous donnant l'occasion d'observer les synonymes « en action » et d'approfondir la connaissance que nous en avons, de manière empirique. Comme montre le sondage présenté, les locuteurs natifs (y compris les professeurs de langues) ne sont pas nécessairement capables de lister toutes les significations et toutes les utilisations d'un élément lexical donné, d'où l'avantage des corpus. Les grandes bases de données rendent accessibles une partie des énoncés qui sous-tendent des connaissances des natifs pour confirmer ou réfuter nos intuitions et nos hypothèses. Nous explorerons par la suite comment ils peuvent contribuer à éclairer la différence d'usage entre les synonymes « tűnik » et « látszik ».

C) Catégorisation des exemples avec « tűnik »

Une simple recherche de fréquence dans le corpus « huTenTen12 » montre que le verbe « tűnik » apparaît 855 635 fois, alors que « látszik » n'apparaît que 535 783 fois. En d'autres termes, « tűnik » apparaît une fois et demie plus souvent (env. 1,5 : 1) que « látszik ». Le Corpus national du hongrois révèle des occurrences comparables : « tűnik » apparaît 138 814 fois, alors que « látszik » apparaît 91 083 fois. Ces chiffres indiquent déjà une différence : celle de leurs alternances dans l'usage langagier. Les raisons de cette différence peuvent être liées à une différence du nombre d'usages possibles (plus de significations pour « tűnik » que pour « látszik », à première vue une contradiction à la classification du dictionnaire qui attribue trois sens à « látszik » et seulement un à « tűnik ») ou à un usage plus étendu dans le cas de « tűnik » (plus d'unités multi-lexicales possibles ou des usages très nombreux avec certaines unités multi-lexicales). Dans les pages suivantes, nous explorerons, entre autres, les raisons éventuelles de cette différence de fréquence.

Le verbe « tűn* »[79] est le plus récurrent des deux verbes étudiés, nous trouvons 855 635 exemples avec ce verbe dans le corpus « huTenTen12 ». Ces usages incluent la formulation d'une impression qui peut être, comme nous le verrons par la suite, justifiée ou refutée par le locuteur.

[79] L'astérisque (*) indique que toutes les occurrences du verbe aux différentes personnes, aux différents modes et aux différents temps ont été prises en compte.

1) Exprimer une impression

1.1) Úgy tűnik/tűnt (, hogy) (Il semble/semblait que) (292 913)

Le collocatif le plus fréquent dans les corpus choisis est l'adverbe « úgy » (~ comme)[80] représentant un tiers (34,25 %) de toutes les occurrences. En voici quelques exemples (tableau 83) :

Úgy **tűnik**, az előadás mindenkinek maradandó élmény volt.	*Le spectacle* ***semble*** *avoir été une expérience mémorable pour tous.*
Úgy **tűnik**, hogy lassan talán rend lesz ott is.	***Il semble*** *qu'il pourrait bientôt y avoir de l'ordre là aussi.*
Ráadásul úgy **tűnik**, hogy egyre veszélyesebbé válik a sport.	*En outre, ce sport* ***semble*** *devenir de plus en plus dangereux.*
A győzelemnek, úgy tűnik, túl nagy ára van.	*La victoire,* ***semble-t-il****, a un prix trop élevé.*
Nekem úgy **tűnik**, nem érted, miről van szó.	*Tu ne* ***sembles*** *pas comprendre de quoi il s'agit.*
Úgy **tűnik**, a külföldi minta Magyarországon is működik.	*L'exemple étranger* ***semble*** *fonctionner en Hongrie aussi.*
Egyre inkább úgy **tűnik**, hogy a csapat megtalálta saját stílusát.	***Il paraît que*** *l'équipe retrouve de plus en plus son style.*
Úgy **tűnik**, mások is értékelik a munkánkat.	***Il semble que*** *d'autres personnes apprécient aussi notre travail.*

Tableau 83 : Quelques exemples avec l'unité multi-lexicale « úgy tűnik » du corpus « huTenTen12 ».

L'unité multi-lexicale « úgy tűnik » est utilisée quand tous les signes, et l'interprétation de ces signes par le locuteur, indiquent que la proposition dans la deuxième partie de la phrase est vraie. Le mot « nekem » (il me semble) renforce le caractère subjectif de l'impression (5 963 occurrences au présent et 1 524 au passé). Une recherche dans les corpus oraux montre que la version avec la conjonction « hogy » semble être la version préférée à l'oral.

Une étude plus approfondie du temps du verbe révèle que cette unité multi-lexicale est majoritairement utilisée au présent : « úgy tűnik » apparaît 219 531 fois (75%) dans le corpus. Il s'agit donc d'une impression que le locuteur ressent au moment où il s'exprime. La suite de son récit fournit des raisons indiquant que cette impression est vraie ou, plus rarement (22 % des occurrences au présent), qu'elle est inexacte.

[80] Le mot n'apparaît pas dans la traduction française.

Le passé (« úgy tűnt », il semblait) est représenté par 59 872 occurrences (20,5%). L'unité multi-lexicale « úgy tűnt » est majoritairement suivie de la justification de cette impression mais un plus grand pourcentage des phrases est suivi d'une proposition montrant que l'impression était trompeuse. L'analyse de l'environnement textuel plus étendu des premières 500 occurrences avec « úgy tűnik » et « úgy tűnt » révèle l'explication de cette différence : les phrases au passé s'intègrent dans des récits chronologiques et font référence à des impressions et à des hypothèses formulées à un moment du passé qui depuis, ont été validées ou rejetées.

Le reste des occurrences est couvert par d'autres formes moins fréquentes : nous trouvons 4790 exemples avec « úgy tűnhet » (il pourrait sembler), 732 avec « úgy tűnhetett » (il pouvait sembler), 449 avec « úgy tűnne » (il semblerait), 89 avec « úgy tűnhetne » (il pourrait sembler) et d'autres formes. Dans le cadre pédagogique, ces occurrences peuvent être écartées de la présentation générale car négligeables au vu de leur faible fréquence[81].

1.2) ADJnAk + tűn, (ADJ + N)nAk + tűn* (sembl* être + ADJ, sembl* être + (ADJ + N))*

Ces unités multi-lexicales représentent un très grand groupe (50,7 % de toutes les occurrences) dont les membres sont reliés, avant tout, par des caractéristiques grammaticales (terminaison, type de partie de discours) alors que les éléments lexicaux concrets (adjectif ou nom + adjectif) peuvent varier.

ADJnAk + tűn (semble être + ADJ)*

Dans ces unités multi-lexicales, le verbe est précédé d'un adjectif (et, dans certaines phrases, d'un numéral) et l'ensemble des éléments décrit l'opinion du locuteur concernant une proposition. Le tableau 84 présente la liste des adjectifs et des numéraux les plus courants (plus de 200 occurrences) :

jónak (6 401)	*bon*	egyértelműnek (2 142)	*évident*
egyszerűnek (4 752)	*simple*	logikusnak (2 486)	*logique*
biztosnak (3 015)	*sûr*	érdekesnek (2 380)	*intéressant*
	pour beaucoup	természetesnek (1 994)	*naturel*

[81] Voir aussi le chapitre 12 sur les considérations pédagogiques.

soknak (3 823)	*approprié*	ígéretesnek (2 048)	*prometteur*
megfelelőnek (670)	*impossible*	reménytelennek (2 067)	*sans espoir*
lehetetlennek (3 388)	*simple*	kilátástalannak (1 477)	*sans issue*
egyszerűnek (4 752)	*probable*	értelmetlennek (1 263)	*insensé*
valószínűnek (1 412)	*bizarre*	meggyőzőnek (229)	*convaincant*
furcsának (3 522)	*incroyable*	elkerülhetetlennek (225)	*inévitable*
hihetetlennek (3 610)		nehéznek (1 626)	*difficile*

Tableau 84 : Les adjectifs les plus courants avec « tűnik ».

(ADJ + N)nAk + tűn*

Des noms peuvent également s'associer à ce verbe avec la même fonction que les adjectifs présentés ci-dessus. En général, ces noms sont précédés d'un adjectif (ce que nous avons indiqué par la parenthèse connectant ces deux entités dans le titre). La variété lexicale est grande et même les unités multi-lexicales les plus usitées ne représentent pas individuellement une haute fréquence (leur nombre varie entre 3 548 pour le premier mot et 500 pour le dernier de la liste). Le tableau 85 en montre quelques exemples :

jó ötletnek	*bonne idée*
jó megoldásnak	*bonne solution*
jó/tökéletes/logikus választásnak	*bon choix, choix parfait/logique*
túlzásnak	*exagération*
egy örökkévalóságnak	*éternité*
egyszerű/könnyű/nehéz/lehetetlen feladatnak	*tâche simple/facile/difficile/impossible*
apróságnak	*futilité*
képtelenségnek	*impossibilité, absurdité*
lehetetlen/reménytelen/merész vállalkozásnak	*entreprise impossible/vouée à l'échec/téméraire*
rendes/megbízható embernek	*homme décent/fiable*

Tableau 85 : Les unités ADJ + N les plus courantes avec « tűnik ».

En étudiant leur composante pragmatique, nous pouvons constater que les adjectifs et les expressions « ADJ + N » peuvent refléter une opinion soit positive, soit négative. Il ne s'agit donc

pas seulement d'une impression mais d'un point de vue ou d'un jugement que le locuteur porte sur le sujet en question. Le tableau 86 en montre quelques exemples courants.

érdekesnek	*intéressant*	TŰN*,	hogy ...
meglepőnek	*surprenant*	*sembl**	*que ...*
elvontnak	*abstrait*		
értelmetlennek	*insensé*		
jó ötletnek	*une bonne idée*		
tökéletes megoldásnak	*une solution parfaite*		

Tableau 86 : Quelques unités multi-lexicales avec une composante pragmatique positive ou négative.

1.3) azért tűn ADJ/N-nAk, mert (sembl* être ADJ/N parce que)*

Dans un faible pourcentage d'occurrences (2%), le locuteur explique la raison qui justifie son impression (tableau 87) :

Csak azért **tűnik** paródiának, mert nem értetted meg a lényeget.	*Ça **vous semble être une parodie parce que** vous n'avez pas compris l'essentiel.*
Joschka Fischer épp azért **tűnt** meggyőzőnek, mert tartózkodott az ígéretektől.	*Joschka Fischer **semblait convaincant** précisément **parce qu'**il s'abstenait de faire des promesses.*
Azért **tűnik** olyan nagynak a gondod, mert túl közel állsz hozzá.	***Ton souci te semble très grand parce que** tu n'arrives pas à prendre du recul.*
Lehet, hogy csak azért **tűnt** ilyen idegennek a város, mert már régen nem jártam itt hétköznapokon.	*Peut-être que la ville **me semblait si étrangère parce que** je n'y suis pas allé depuis longtemps en semaine.*
Skóciában tanulni azért **tűnt** jó ötletnek, mert az ember jobban össze tudja egyeztetni a munkát a tanulással.	*Étudier en Écosse **me semblait une bonne idée parce qu'**on peut trouver plus facilement un équilibre entre le travail et les études.*
A legtöbb dolog azért **tűnik** ijesztőnek, mert ismeretlen.	*La plupart des choses **semblent effrayantes parce qu'**elles sont inconnues.*

Tableau 87 : Phrases explicatives avec l'unité « azért tűn* ADJ/N-nAk, mert ».

2) Opposition entre impression et réalité

2.1) « első látásra XnAk tűnik/tűnhet, de ... » (semble/peut sembler X à première vue mais ...)

Une certaine partie des énoncés contenant « ADJ/N-nAk + tűn* » renferme un élément lexical indiquant qu'il s'agit d'une première impression trompeuse. Les éléments modificateurs listés ci-dessous sont séparés de 3 à 5 éléments du verbe et apparaissent dans environ 6% des phrases du corpus (tableau 88) :

elsőre	*la première fois*
először	*au début*
(első) ránézésre	*à première vue*
az elején	*au commencement*
első látásra	*au premier regard*
első hallásra	*quand on l'entend pour la première fois*
kívülről (nézve)	*(vu) de l'extérieur*
kezdetben	*initialement*
eleinte	*au départ, au début*
első pillantásra	*au premier coup d'œil*
a kívülálló számára	*pour quelqu'un de l'extérieur*

Tableau 88 : Éléments lexicaux indiquant une première impression, associés à la structure « ADJ/N-nAk tűn* ».

Le fait qu'il s'agisse d'une « opposition entre impression et réalité », est également renforcé par des conjonctions. Dans 7,7 % des phrases, nous trouvons ainsi une conjonction renvoyant à une opposition. Ces conjonctions ne font pas partie de l'unité multi-lexicale au sens strict du terme, puisqu'elles ne peuvent pas être systématiquement assignées au verbe, mais elles font partie de son environnement textuel plus large, elles appartiennent au verbe, puisqu'elles contribuent à clarifier son sens. La répartition des conjonctions est présentée dans le tableau 89 ci-dessous :

Bár ...	*Alors que, quoique* (*au début de la phrase dans 14 % des occurrences, peut se combiner avec les conjonctions listées*)

…, de	..., *mais*
..., mégis	..., *pourtant*
X ellenére	*malgré X*
pedig	..., *et/mais*
mégsem	..., *cependant*
ugyanakkor	..., *en même temps*

Tableau 89 : Conjonctions exprimant l'opposition entre l'impression et la réalité dans les phrases avec le verbe « tűnik ».

Le tableau 90 présente quelques exemples de phrases entières et des débuts de phrases du corpus pour que le lecteur ait une idée plus précise de cet usage :

Igenis értékeli az alkotásokat, még ha első pillantásra úgy is **tűnik**, hogy nem.	*De toute évidence, il apprécie les œuvres, même si, à première vue, il semble que ce n'est pas le cas.*
Csak első pillantásra **tűnik** nehéz feladatnak, ugyanakkor meg lehet oldani.	*Cela ne semble une tâche difficile qu'à première vue, mais on peut le résoudre.*
Bár első pillantásra korántsem **tűnik** nyilvánvalónak, mégis …	*Même si cela ne semble pas évident au premier abord, …*
Első pillantásra talán ellentmondásosnak **tűnik**, mégis érdemes beszélni róla.	*À première vue, cela peut sembler controversé, mais ça vaut la peine d'en parler.*
Nem is annyira rossz eredmény, mint első pillantásra **tűnik**.	*Ce résultat n'est pas aussi mauvais qu'il y paraît à première vue.*
Első pillantásra jó ötletnek **tűnik**, de nem tudom.	*À première vue, cela semble être une bonne idée, mais je ne sais pas.*
Az előbbiek miatt könnyen úgy **tűnhet**, hogy erre bárki képes, de ez nem így van.	*Ce qui précède **peut donner l'impression que** tout le monde en est capable, **mais** ce n'est pas le cas.*
Ez egy nagy csapda, bár első látásra úgy **tűnhet**, hogy beválik.	*C'est un gros piège, **même s'il peut sembler** fonctionner à première vue.*

Tableau 90 : Exemples de phrases du corpus « huTenTen12 » exprimant une opposition entre impression et réalité.

Le tableau 91 présente un usage proche mais distinct dans lequel la fonction de « tűnik » est d'exprimer une impression initiale résultant d'une observation superficielle ou partielle :

« első látásra XnAk tűnik/tűnhet » (semble/peut sembler X à la première vue)

Első látásra	meglepőnek	TŰN*,	de
Első pillantásra	furcsának	TŰNHET*,	pedig
Első ránézésre	érdekesnek		mégis
Első olvasatra	szokatlannak		ugyanakkor
Első hallásra	különösnek		
Elsőre	ellentmondásosnak		
Először	túlzásnak		
Eleinte / Az elején	jó ötletnek		
Kezdetben	úgy		
A première vue	*surprenant*	TŰN*,	*mais*
Au premier coup d'œil	*étrange*	TŰNHET*,	*cependant*
Au premier regard	*d'intérêt*	(semble*	*pourtant*
En première lecture	*inhabituel*	peu* sembler)	*en même temps*
Quand on l'entend pour la	*étrange*		
première fois	*controversé*		
Au tout début	*une exagération*		
D'abord	*une bonne idée*		
Au début	*comme*		
Au départ			

Tableau 91 : Éléments lexicaux exprimant que la première impression est fondée sur une observation superficielle.

Cette analyse de l'environnement textuel plus large à partir des données issues du corpus nous a permis d'établir un schéma d'usage qui n'est pas identifié dans le dictionnaire. Nous avons ainsi pu collecter un certain nombre d'éléments lexicaux (conjonctions, adjectifs) qui occupent la même position dans la phrase et qui, malgré leur variété au niveau lexical, sont liés par des aspects sémantiques et grammaticaux. Le nombre de ces variations étant limité, il est possible de les répertorier et, puisque nous explorons le corpus dans le cadre pédagogique, d'en proposer une présentation pertinente à l'apprenant.

2.2) « talán XnAk tűnik, de », « lehet, hogy XnAk tűnik, de » (semble peut-être X)

Un autre sous-groupe des phrases introductoires signale également que l'impression du lecteur peut être fausse. Le locuteur comprend que les faits présentés puissent provoquer cette impression mais il insiste sur le fait que cette impression n'est pas véridique. Pour cela, il utilise des modificateurs comme « peut-être » (talán) ; « il est possible que » (lehet, hogy), « il se peut » (meglehet) (tableau 92) :

« talán XnAk tűnik, de » (semble X à la première vue)

talán	XnAk	tűn*
lehet, hogy	(ADJ,	
meglehet	ADJ + N)	
peut-être	*ADJ*	*tűn**
il est possible que	*N + ADJ*	
il se peut que		

Tableau 92 : Éléments lexicaux permettant une première impression illusoire de la part de l'interlocuteur.

3) Faits momentanés, susceptibles de changer

Dans environ 3 % des occurrences, les phrases font référence à un état actuel de choses qui peut encore changer. Ces phrases contiennent dans ce cas les modificateurs qui indiquent la temporalité de la situation (tableaux 93 et 94) :

Minden, ami eddig szépnek **tűnt**, elveszítette szépségét.	*Tout ce qui **avait semblé beau jusque-là** avait perdu de sa beauté.*
Ami eddig egyetlen ösvénynek **tűnt**, az röviddel ezelőtt kettévált.	*Ce qui **semblait être un seul chemin** jusqu'ici a bifurqué récemment.*
Ezek az intelligens épületek eddig inkább utópisztikusnak **tűntek**, mintsem kézzel fogható rendszereknek.	*Ces bâtiments intelligents **ont jusqu'ici semblé plutôt utopiques** que des systèmes tangibles.*
Ha valakinek eddig túl puritánnak **tűnt** a Windows Phone7 rendszer felülete, annak érdemes egy pillantást vetnie az alábbi koncepcióképekre.	*Si l'interface Windows Phone7 **a semblé trop puritaine jusqu'à présent**, ça vaut la peine de jeter un coup d'œil aux images de concept ci-dessous.*
Ez a legújabb, nem feltétlenül a legjobb készülék, bár egyelőre úgy **tűnik**, sok	*C'est le dernier, pas nécessairement le meilleur appareil, même si **pour l'instant il semble être** au moins aussi bon à bien des égards que l'iPhone.*

mindenben legalább annyira jó, mint az iPhone.	
Egyelőre úgy **tűnik**, 3 tizedet javítottam a félévi átlagomhoz képest.	***Pour l'instant, il me semble*** *m'être amélioré de 3 dixièmes par rapport à ma moyenne semestrielle.*
Hát igen, biztos sok idő még, mert egyelőre úgy **tűnik**, hogy csak bennem bízik, de még bennem sem 100% -osan.	*Oui, cela prendra certainement beaucoup de temps car* ***on dirait que, pour l'instant,*** *il ne me fait pas confiance, mais il ne semble pas me faire confiance.*
Egy 3D stratégiai játék lesz, ha elkészülök vele, mert egyelőre nagy falatnak **tűnik**, mint hobbiprojekt.	*Ce sera un jeu de stratégie en 3D quand j'en aurai fini, car* ***pour le moment, cela me semble un gros défi*** *pour un projet de loisir.*
A globális felmelegedés egyelőre megállíthatatlannak **tűnik**.	***Pour le moment,*** *le réchauffement climatique* ***semble imparable****.*

Tableau 93 : Phrases avec des éléments indiquant la temporalité de la situation.

eddig	úgy	**TŰNT***
egyelőre	érthetetlennek	**TŰN***
jelenleg	valószínűnek	
továbbra is	stb.	
jusqu'à présent	*comme*	***semblai*t***
pour l'instant	*incompréhensible*	***sembl****
actuellement	*probable*	
comme avant	*etc.*	

Tableau 94 : Les éléments indiquant un état de fait momentané.

Le seul modificateur utilisé pour parler du passé est « eddig » (jusque-là), les autres modificateurs sont réservés à l'usage au présent (tableau 94). « Eddig » peut également être utilisé au présent.

D) Que veut dire le mot « tűnik » ?

En tenant compte de l'environnement textuel plus large du verbe « tűnik », nous pouvons donc à présent établir les profils relatifs à ses usages. Dans les tableaux suivants, nous utilisons les catégories employées pour les mots à usages multiples. Ceux-ci sont présentés selon l'ordre de leur fréquence (tableau 95).

1) « Úgy tűnik, hogy » (Il semble que)

Collocation	Úgy tűnik, (hogy) Nekem úgy tűnik, hogy / Úgy tűnik nekem, hogy
Colligation typique	Usage au présent et au passé Unité multi-lexicale placée au début de la phrase
Composantes sémantiques	Introduit une phrase sur l'impression (positive, négative ou neutre) du locuteur
Composantes pragmatiques	Pas de composantes pragmatiques particulières.

2) « XnAk tűnik » (semble X)

Collocation	**Adjectifs fréquents (1) :** « valószínűnek » (vraisemblable), « biztosnak » (sûr), « kétségesnek » (douteux) **Adjectifs fréquents (2) :** « érdekesnek » (intéressant), « furcsának » (bizarre), « jónak » (bien) **Adjectif + nom :** « jó ötletnek » (une bonne idée), « jó megoldásnak » (une bonne solution), « kis túlzásnak » (un peu exagéré)
Colligation typique	X = ADJ ou ADJ + N Phrase d'introduction Usage au présent et au passé
Composantes sémantiques	(1) Exprime la certitude du locuteur par rapport à la proposition dans la deuxième partie de la phrase. (2) Exprime le jugement du locuteur par rapport à la proposition dans la deuxième partie de la phrase.
Composantes pragmatiques	Pas de composantes pragmatiques particulières.

3) « Első látásra XnAk tűnik, de » (semble X à première vue mais)

Collocation	**Modificateurs fréquents :** « első látásra/pillantásra » (à première vue), « elsőre » (d'abord), « először » (au début) et d'autres mots indiquant qu'il s'agit d'une première impression superficielle

	Conjonctions fréquentes : « bár » (quoique), « de » (mais), « pedig » (mais, cependant), « mégis » (pourtant), « ugyanakkor » (en même temps)
Colligation typique	**MOD + XnAk + tűnik/tűnhet (semble/peut sembler) + CONJ** Usage au présent et au passé Verbe à l'indicatif, souvent avec l'affixe « -het » Se trouve au début de la phrase
Composantes sémantiques	Exprime une première impression (superficielle) opposée à la réalité. La réalité est décrite dans la deuxième partie de la phrase.
Composantes pragmatiques	Introduit la justification d'un fait qui peut être contraire aux impressions.

4) « talán XnAk tűnik, de » (peut sembler X mais)

Collocation	**Modificateurs fréquents :** « talán » (peut-être), « lehet, hogy » (il est possible que), « meglehet » (il se peut que) **Conjonctions fréquentes :** « de » (mais), « mégis » (pourtant), « ugyanakkor » (en même temps)
Colligation typique	**MOD + XnAk + tűnik (semble) + CONJ** Introduit une phrase
Composantes sémantiques	Exprime une impression dont le locuteur admet qu'elle peut être perçue de façon négative mais le locuteur insiste également sur le fait que cette impression est fausse.
Composantes pragmatiques	Introduit la justification d'un fait qui peut être contraire aux impressions.

5) « eddig úgy/XnAk tűnik » (jusqu'à présent il semble que)

Collocation	**Modificateurs fréquents :** « eddig » (jusqu'à présent), « egyelőre » (pour l'instant), « jelenleg » (en ce moment), « továbbra is » (comme avant)
Colligation typique	**MOD + XnAk + tűnik/tűnt (semble) + CONJ** Usage au présent et au passé (souvent au passé)

Composantes sémantiques	Présent : indique une impression momentanée, susceptible de changer avec le temps Passé : indique un changement concernant une impression
Composantes pragmatiques	Peut introduire un passage qui explique pourquoi l'impression peut changer (ou pas) à l'avenir.

Tableau 95 : Profil du verbe « tűnik ».

E) Látsz*

Comme indiqué précédemment, le verbe « látszik » est moins fréquent dans le corpus que « tűnik » : nous y trouvons 535 783 exemples avec ce verbe. La majorité des catégories identifiées montre l'implication d'un élément visuel, ce qui peut expliquer l'usage plus limité du mot par rapport à « tűnik », indiquant ainsi une impression pouvant être fondée sur plusieurs facteurs. D'autres groupes reflètent des usages spécifiques, relatifs à la formulation d'hypothèses.

1) Exprimer une impression : Úgy látsz*, (hogy) ...

Le verbe « látszik » s'associe également à « úgy » : ils apparaissent ensemble dans 35% des occurrences du corpus « huTenTen12 » et dans 37% des énoncés du Corpus national du hongrois. Le tableau 96 présente quelques exemples typiques d'usage du verbe « látszik » en conjonction avec « úgy ».

Ebben, úgy **látszik**, én tévedtem.	*Il semble que j'ai eu tort à ce sujet.*
Elnézést, úgy **látszik**, félreérthetően fogalmaztam.	*Désolé, il semble que je me suis mal exprimé.*
Úgy **látszik**, az érvek nem változtak.	*Les arguments ne semblent pas avoir changé.*
A nőknek, úgy **látszik**, itt is meg kell küzdeniük a jogaikért.	*Ici aussi, les femmes semblent devoir se battre pour leurs droits.*
Úgy **látszik**, aktívabbak lettek a parlamenti pártok.	*Les partis parlementaires semblent être devenus plus actifs.*
Én értem őket, de ők – úgy **látszik** – nem értenek engem.	*Je les comprends, mais ils ne semblent pas me comprendre.*
Be is telefonál pár ember, úgy **látszik**, tetszik nekik a műsor.	*Quelques personnes nous appellent, il semble qu'elles apprécient l'émission.*
Úgy **látszik**, nincs szerencsém ezekkel a madarakkal. Nem tudom lefotózni őket.	*Il semble que je n'ai pas de chance avec ces oiseaux. Je n'arrive pas à les prendre en photo.*

Tableau 96 : Quelques exemples avec l'unité multi-lexicale « úgy látszik » du corpus « huTenTen12 ».

Ces phrases expriment le fait qu'il existe des signes extérieurs indiquant la validité de l'impression du locuteur. Dans la majorité de ces phrases « látszik » pourrait être remplacé par « tűnik », ce dernier suggérant une certitude moins forte concernant la sincérité de l'impression, comme vu dans la première partie de ce chapitre.

2) Une chose est perceptible : látsz* + N (N se voi*)

Le deuxième groupe important (32% des occurrences dans « huTenTen12 » et 31% dans le CNH) est constitué de phrases qui impliquent que quelque chose/quelqu'un est perceptible ou visible (ou non) à l'œil nu.

2.1) (nem) látsz + N (N (ne) se voi* pas)*

Dans 34 % des occurrences au sein de ce second groupe, le verbe signifie que quelque chose est (non) perceptible, (non) visible à première vue. Sur la base du corpus, nous ne pouvons pas identifier de noms spécifiques en tant qu'éléments lexicaux fréquemment associés à ce sens. Cela peut certainement s'expliquer par le fait qu'il s'agit d'un cas où le sens du verbe est parfaitement clair en soi, sans autres éléments, et par le fait que le choix de sujets de la phrase, c'est-à-dire les choses ou les personnes qui peuvent être visibles, est presque illimité. Le tableau 97 présente quelques exemples de tels usages.

Az öklömet horzsoltam le, máig **látszik**.	*Je me suis écorché le poing, on peut encore le voir.*
Tisztán **látszik** a makett és az emberek.	*Vous pouvez clairement voir le modèle et les personnes.*
(...) a kamera persze nem talált olyan szöget, hogy mind a kettejük arca **látsszon**.	*(...) bien sûr, la caméra n'a pas pu trouver un angle qui montre leurs deux visages.*
Nem **látszik** a topicnyitó üzenet.	*Vous ne pouvez pas voir le message d'ouverture du sujet.*
A borítón a rohamozó törökök mellett **látszik** az operatőr is!	*Sur la couverture, à côté des Turcs qui chargent on voit le caméraman !*

Tableau 97 : Quelques phrases indiquant la visibilité ou la non visibilité de quelque chose.

Dans de nombreuses phrases tirées du corpus, nous pouvons cependant observer la présence d'un modificateur. Ces modificateurs ont la fonction de renforcer le sens en précisant le degré de visibilité. Le tableau 98 en présente les plus usités :

tisztán	*nettement*	innen	*d'ici*
világosan	*clairement*	rögtön	*tout de suite*
egyértelműen	*évidemment*	szépen	*très bien*
nem (is)	*ne pas*	kevésbé	*moins*
alig	*à peine*	azonnal	*immédiatement*
(már) messziről	*(déjà) de loin*	máris	*aussitôt*
jól	*bien*	nyilvánvalóan	*sans ambiguïté*

Tableau 98 : Les modificateurs les plus fréquents précisant le degré de visibilité.

La grande majorité de ces mots indique ainsi que la chose ou la personne donnée est immédiatement visible, reconnaissable. Une petite partie des modificateurs signale le manque de visibilité. Nous verrons que ces modificateurs émergent également dans les sens 2.2) et 2.3) avec les mêmes fonctions.

2.2) látsz N-On / rajt**

Le verbe « látszik » peut être aussi utilisé pour qualifier quelque chose d'immédiatement observable, une qualité reconnaissable à partir de signes extérieurs. Cette occurrence se produit dans ce cas à la troisième personne du singulier ou du pluriel. Le tableau 99 propose quelques phrases relevant d'un tel usage.

Látszott rajta, hogy valamikor fényes volt a felülete.	*On pouvait voir que sa surface était autrefois brillante.*
Látszott rajta, hogy jól érzi magát.	*On pouvait voir qu'il se sentait bien.*
Lázas volt, félrebeszélt, **látszott** rajta, hogy orvosi segítségre van szüksége.	*Il avait de la fièvre, délirait, c'était clair (en le voyant) qu'il avait besoin de soins médicaux.*
A mű nagyon gyenge. **Látszik** rajta, hogy 1997 végén írták.	*L'œuvre est très faible. On voit qu'il a été écrit à la fin de 1997.*
Néha (...) nagyon rossz volt a közérzete, (...) de a kapuban kihúzta magát, s nem **látszott** rajta semmi.	*Parfois (...) il se sentait très mal, (...) mais au portail, il s'est ressaisi et n'a rien laissé voir.*
Beleegyezett, de **látszott** az arcán, hogy nem örül.	*Il l'a accepté, mais son visage montrait qu'il n'était pas heureux.*
Szépen süt a nap Mugellóban, ahogy ezen a képen is **látszik**.	*Le soleil brille au Mugello, comme vous pouvez le voir sur cette photo.*
A felvételen az **látszik**, hogy a stáb egy hat-hét autóból álló konvojt követ.	*Les images montrent l'équipage suivant un convoi de six à sept voitures.*

Az ekkor készült fotókon tisztán **látszik** a gyűrűsujján viselt ékszer.	*Les photos prises à l'époque montrent nettement le bijou à son annulaire.*

Tableau 99 : Exemples avec des compléments de lieu.

Le tableau 100 présente les compléments de lieu les plus fréquents associés à cet usage.

a képen	*sur l'image*	a grafikonon	*sur le graphique*
az arcán	*sur son visage*	a mozdulatán	*à son geste*
a fotón	*sur la photo*	a teljesítményén	*à sa performance*
a videón	*dans la vidéo*	a térképen	*sur la carte*
a felvételen	*sur l'enregistrement*	az eredményen	*au résultat*
az ábrán	*au diagramme*	a fényképeken	*sur les photos*
a szemén	*dans ses yeux*	a viselkedésén	*à son comportement*

Tableau 100 : Les compléments de lieu les plus fréquents indiquant le lieu où quelque chose est visible.

De ces noms se dégagent deux usages différents. Ils peuvent (1) nommer l'apport visuel qui sert de base d'observation ou (2) les signaux corporels révélant l'état d'âme de la personne observée.

2.3) N-bÓl látsz (se voi* de N, ressort* de N)*

Dans certaines phrases, le locuteur indique la source (en général un apport visuel) sur laquelle s'appuie son observation (tableau 101) :

a táblázatból	*du tableau*
a hozzászólásokból	*des commentaires*
a statisztikákból	*des statistiques*
az írásodból	*de ton texte*
a számokból	*des chiffres*
mindebből	*de tout cela*
az adatokból	*des données*

Tableau 101 : Les compléments de lieu indiquant la source révélant des informations (souvent quantifiables).

En général, il s'agit des sources d'informations (souvent quantifiables) exposées dans la deuxième partie de la phrase. Le mot indiquant la source possède la terminaison « -ból/-ből » (*de*) (tableau 102) :

Az adatokból az is jól **látszik**, hogy a vásárlók elsősorban a hárommillió forintos autók között válogatnak.	*Il ressort également des données que les clients choisissent principalement entre des voitures d'une valeur de trois millions de forints.*
Az adatokból az **látszik**, hogy a fiatalok nagyobbik része tapasztalatszerzésből megy el külföldre.	*Les données montrent que la majorité des jeunes partent à l'étranger pour acquérir de l'expérience.*
A táblázatból is **látszik**, hogy a rendelkezésre álló összegek nagyobbik hányadát a bérek teszik ki.	*Le tableau montre également que la plupart des montants disponibles sont représentés par les salaires.*
A számokból is az **látszik**, hogy több százezer főről van szó.	*Les chiffres montrent également que des centaines de milliers de personnes sont concernées.*

Tableau 102 : Quelques exemples avec des compléments de lieu indiquant la source révélant des informations (souvent quantifiables).

Dans l'environnement textuel de « látszik », nous rencontrons dans quelques cas (6% des occurrences) des éléments lexicaux faisant référence à ce qui sera dit par la suite ou a été dit précédemment. La phrase « X abból látszik, hogy » (X est reconnaissable/évident du fait que) introduit ainsi la phrase qui justifie la validité d'une impression ou d'une inférence.

2.4) ebből/abból látsz (se voi* de cela, ressort* de cela)*

La même terminaison (-ből/-ból) est également utilisée à la fin des éléments anaphoriques « ebből/abból » (de cela, du fait que) qui se réfèrent à des impressions en expliquant d'où ces impressions proviennent. Il y a donc une différence d'usage de cette terminaison avec un nom qui indique une source d'informations tangible, fiable et avec le pronom démonstratif utilisé avant tout pour exprimer une opinion personnelle. « Abból (is) látszik » (X est évident du fait que) anticipe la dénomination de la source, « ebből (is) látszik » (cela montre) indique que l'impression ou la conclusion décrite dans la deuxième partie de la phrase est confirmée par ce qui a été dit auparavant (tableau 103) :

abból látsz*	**abból látsz***
Tudod mit jelent az a szó, hogy 'minden'? Amit leírtál, abból jól **látszik**, hogy nem. Ez a reklám telibe kapja a kamaszokat, tehát jó. Ez abból is **látszik**, hogy most vitázunk róla.	*Tu sais ce que signifie le mot 'tout' ? Ce que tu as écrit* ***montre clairement*** *que ce n'est pas le cas. Cette publicité accroche les adolescents, elle est donc bien. C'****est évident du fait*** *que nous en discutons maintenant.*

Amit leírtál, abból **látszik**, hogy érted, miről van szó.	*Ce que tu as écrit **montre** que tu comprends de quoi il s'agit.*
ebből látsz*	**ebből látsz***
Különben ebből is **látszik**, milyen rendes ember vagy.	*De plus, **cela montre** à quel point tu es une personne bien.*
Ebből is **látszik** amúgy, mennyire elavult ez a mű.	***Cela montre** également à quel point cette œuvre a mal vieilli.*
Haydn zseni, ami ebből a felismeréséből is **látszik**.	*Haydn est un génie, ce qui **ressort de cette reconnaissance**.*
A szakértők szerint ebből világosan **látszik**, hogy itt nagy baj lesz.	*Selon les experts, **cela montre clairement** qu'il y aura de gros problèmes.*

Tableau 103 : Exemples avec « abból » et « ebből », indiquant une impression et son explication.

Les modificateurs fréquents (« jól » (bien), « egyértelműen » (évidemment), « tisztán » (nettement), « világosan » (clairement)) renforcent la justesse de l'interprétation des données. On observe donc là une différence significative par rapport à l'usage du verbe « tűnik » qui introduit quant à lui, comme nous l'avons montré, une impression, en général remise en question par la suite.

3) Confirmer ou refuter une hypothèse : INF + látsz*

Un autre sens du « látszik » est réalisé quand le verbe s'accompagne d'un infinitif (15 % des occurrences). Cet usage apparaît principalement dans le langage châtié. La phrase peut contenir à la fois des verbes transitifs et intransitifs. Les verbes transitifs les plus courants sont présentés dans le tableau 104 :

X	alátámasztani (435) igazolni (340) megerősíteni (162) bizonyítani (98)	LÁTSZ-*	azt, (azt) az állítást, (azt) a feltételezést, (azt) a hipotézist, (azt) a tényt,	hogy ...
X	*soutenir (435)* *justifier (340)* *confirmer (162)*	*LÁTSZ-**	*--- (mot ne pas traduit),* *le constat,* *la supposition,*	*que ...*

prouver (98)	*l'hypothèse,*
	le fait,

Tableau 104 : Les verbes les plus courants confirmant ou réfutant une proposition en combinaison avec látsz*.

Ces unités multi-lexicales sont utilisées quand il s'agit de juger de la validité d'une hypothèse.

Les cinq infinitifs intransitifs les plus courants et leurs sujets typiques montrent également que cet usage est associé à un niveau recherché de langage (tableau 105) :

rendeződni **látszik** a válság / a helyzet / X helyzete / az ügy / Y ügye / a szomszédokkal való kapcsolat / Z sorsa	*la crise / la situation / la situation de X / le cas / le cas de Y / les relations avec les voisins / le sort de Z* semble se résoudre
megoldódni **látszik** *a probléma / a gond / a válság / X ügye / több nyitott kérdés / Y helyzete*	*le problème / la crise / la question X / plusieurs questions ouvertes / la situation Y* semble être sur la voie de la résolution
X-nek ellentmondani **látszik** *Y nyilatkozata / az (a körülmény), hogy*	*la déclaration de Y / la circonstance que* ... X semble contredire à X
stabilizálódni **látszik** (153) *V helyzete / X létszáma / Y állapota / Z (az élelmiszerek) árindexe / a (forint) árfolyama / a részvényárak / a hitelek nagysága*	semble se stabiliser la *situation de V les effectifs de X / la situation de Y / l'indice des prix (alimentaires) de Z / le taux de change (forint) / le cours des actions / l'importance des prêts*
megdőlni **látszik** (152) *az elméletem / az az elmélet, mely szerint / az a vád, hogy / az a közhely, hogy / ez a hipotézis is / X (a sivár vegakonyha) sztereotípiája / ez a dogma*	*ma théorie / la théorie que / l'accusation que / le truisme que / cette hypothèse est aussi / le stéréotype de / X (la cuisine végane insipide) / ce dogme* semble s'effondrer

Tableau 105 : Les sujets typiques des verbes les plus courants confirmant ou réfutant une proposition en combinaison avec látsz*.

4) Exprimer la validité d'une impression : ADJ-nAk látsz-*; ADJ + NnAk látsz-*

Cet usage se manifeste dans une proportion plus faible qu'avec le verbe « tűnik » : il ne ressort que dans environ 13 % des phrases. C'est dans ce cas que le sens des deux verbes apparaît le plus proche ; ils peuvent même être intervertis dans certains cas.

4.1) ADJ-nAk látsz (sembl* être ADJ)*

Le verbe « látszik » peut être précédé d'adjectifs dont les plus fréquents sont présentés dans le tableau 106 (présentant les occurrences supérieures à 50) :

biztosnak	*sûr*	LÁTSZ*, hogy ...

valószínűnek		*probable*	LÁTSZ * *que* ...
(X alapján)	bizonyosnak	*(basé sur X) certain*	
elkerülhetetlennek		*inévitable*	
reménytelennek		*sans espoir*	
egyértelműnek		*clair*	
szükségesnek		*nécessaire*	
fontosnak		*important*	

Tableau 106 : Les adjectifs les plus fréquents associés au verbe « látszik ».

Une partie des adjectifs décrit le degré de certitude concernant la validité de l'impression ; l'autre partie exprime un jugement de valeur ou l'attitude du locuteur envers la proposition qui suit cette phrase introductoire.

Dans certaines phrases dans lesquelles le locuteur essaie de prédire la probabilité d'un événement futur, « tűnik » et « látszik » sont interchangeables, « látszik » exprimant un degré de certitude légèrement plus grand que « tűnik » (tableau 107).

Az első tizenegy hónap eredményei után már biztosnak **tűnik/látszik**, hogy ...	*Après les onze premiers mois de résultats,* ***il semble certain*** *que ...*
Néhány posztra már most biztosnak **tűnik/látszik** a jelölt neve.	*Pour certains postes, les noms des candidats* ***semblent déjà acquis****.*
A hírek alapján a dolog elég egyértelműnek **tűnik/látszik**.	*D'après les nouvelles, cela* ***semble assez évident****.*
(...) elkerülhetetlennek **tűnik/látszik** az orosz csapatok offenzívája.	*(...) une offensive des troupes russes* ***semble/serait inévitable****.*
A vég már-már elkerülhetetlennek **tűnik/látszik.**	*La fin* ***semble presqu'inévitable****.*
Az első három közé jutás azonban mindenképpen szükségesnek **tűnt/látszott**.	*Cependant, une place dans les trois premiers* ***semblait, tout bien considéré, nécessaire****.*

Tableau 107 : Les adjectifs les plus fréquents associés au verbe « látszik », verbe interchangeable dans ces phrases avec « tűnik ».

Bien que les deux verbes puissent être utilisés dans ce cas comme synonymes, leur répartition en pourcentage diffère significativement. Ce genre d'unités multi-lexicales ne représente ainsi que 9,5% du total des occurrences de « látszik » alors qu'il forme plus de 50% du total des occurrences de « tűnik ». C'est de toute évidence l'usage qui prédomine dans le cas de « tűnik » et moins

fréquent dans le cas de « látszik ». La raison de cette distribution peut être que les phrases avec le verbe « tűnik » sont susceptibles d'exprimer des degrés de certitude plus ou moins grands, ce qui permet donc plus de variété d'usage.

Une autre différence importante peut être notée : alors que dans certaines phrases aux structures grammaticales similaires, les deux clauses sont en opposition, dans le cas de « tűnik », la deuxième proposition ne s'oppose que rarement à la première. La plupart du temps, il va plus loin, expliquant ce qui semble certain, probable ou clair.

4.2) ADJ + NnAk látsz (sembl* être ADJ+ N)*

Comme pour « tűnik », « látszik » peut être également accompagné d'un adjectif ou d'une expression formée d'un adjectif et d'un nom. Il peut être utilisé pour donner une première impression qui ne repose pas nécessairement sur une source visuelle. La chose ou la personne pour laquelle le locuteur donne son opinion peut être multiple, il n'est donc pas possible d'identifier des noms typiques (tableau 108) :

kisiskolásnak	*écolier*
65 évesnek	*65 ans*
csirkefogónak	*canaille*
becsületes embernek	*honnête homme*
a régi erkölcsök megsemmisítésének	*destruction des anciennes moeurs*
kedves fiúnak	*gentil garçon*
jó ötletnek	*bonne idée*
megfelelő megoldásnak	*la bonne solution*

Tableau 108 : Quelques exemples avec la structure « ADJ + NnAk látsz* ».

On peut observer que ces expressions sont beaucoup moins fréquemment complétées que dans le cas de « tűnik », par des constructions qui font référence au contraste entre la première impression et la réalité. Il y a 394 occurrences de « kívülről » (de l'extérieur), 182 de « elsőre » (au début), 104 de « első pillantásra » (au premier coup d'œil) et seulement 85 de « első ránézésre » (à première vue). Une explication possible est ici que le verbe « látszik » implique déjà la visualité (lát = voir), de sorte qu'un élément s'y référant pourrait sembler redondant. En outre, la justesse de l'impression n'est pas remise en cause. D'autres formes n'apparaissent également que 130 fois dans le corpus, ce qui semble également montrer que l'utilisation de « látszik » ne remet généralement pas en question la validité d'une impression.

5) Formuler une évidence : Látsz*, hogy ... (Il est évident que / ça se voi* que)

Les 1205 occurrences commençant par ce verbe véhiculent toutes l'idée que le locuteur est convaincu de la justesse de sa conclusion qu'il considère évidente. Les éléments amènent les locuteurs à ce jugement sont nommés dans la partie précédente du discours (v. tableau 109).

Látszik, hogy nem járok eleget moziba.	*Il est évident que je ne vais pas assez au cinéma.*
Látszik, hogy még nincs gyereked!	*Il est évident que tu n'as pas encore d'enfants !*
Látszik, hogy fogalmad sincs a realitásokról.	*Il est évident que tu n'as aucune idée de la réalité.*
Minden mozdulatuk kiszámított - **látszik,** hogy nem először csinálják.	*Leurs moindres gestes sont calculés, il est évident que ce n'est pas leur première fois.*

Tableau 109 : Exemples de l'usage de « látszik » pour formuler une évidence.

Nous notons que « tűnik » ne présente pas une telle utilisation, il s'agit donc là d'un sens qui est clairement réservé à « látszik ». Cela n'est guère surprenant à la lumière de notre étude concernant « tűnik » qui a révélé que ce verbe comprenait toujours un certain degré d'incertitude ; un usage exprimant une grande certitude ne serait donc pas compatible avec ses composantes sémantiques et pragmatiques identifiées dans le cas du verbe « látszik ».

F) Profil de « látsz*» (par ordre de fréquence)

À partir des études présentées dans la section E), nous pouvons établir le profil du verbe « látszik » dans le tableau 110 ci-dessous :

1) Úgy látsz(, hogy) (Il sembl* que)*

Collocation	« Úgy látszik, hogy » (Il semble que)
Colligation typique (1)	Formes typiques du verbe : « látszik » (troisième personne du singulier, indicatif présent), « látszott » (troisième personne du singulier, indicatif passé). L'usage de « hogy » est optionnel.
Colligation typique (2) : ordre des mots	Unité multi-lexicale placée au début d'une phrase Éléments toujours dans cet ordre : « Úgy látszik(, hogy) »
Composantes sémantiques	Introduit une phrase décrivant l'impression (positive, négative ou neutre) du locuteur.
Composantes pragmatiques	Pas de composantes pragmatiques particulières.

2.1) látsz (se voi*)*

Collocations typiques	**Modificateurs fréquents :** (1) « jól » (bien), « világosan » (clairement), « egyértelműen » (manifestement), « tisztán » (nettement) et d'autres adverbes qui soulignent la validité évidente de l'impression (2) « alig » (à peine), « nem » (ne pas), « egyáltalán nem » (pas du tout) et d'autres mots négatifs.
Colligations typiques (1)	Formes typiques du verbe : « látszik » (troisième personne du singulier, indicatif présent), « látszott » (troisième personne du singulier, indicatif passé)
Colligations typiques (2) : ordre des mots	MOD + látszik
Composantes sémantiques	Les adverbes de la catégorie (1) renforcent la justesse évidente de l'impression décrite. Les mots dans la catégorie (2) implique une négation (X ne se voit pas (du tout) ou se voit à peine)
Composantes pragmatiques	Pas de composantes pragmatiques particulières.

2.2) XOn, XbÓl + látsz* (Cela se voi* à/de)

Collocations typiques	(1) « az arcán » (sur son visage), « a tekintetén » (dans son regard), « a szemén » (dans ses yeux) et d'autres mots indiquant des signaux corporels. (2) « az adatokból » (des données), « a statisztikából » (de la statistique), « a diagramból » (du diagramme) et d'autres mots indiquant des apports visuels qui servent de base d'observation.
Colligations typiques	(1) Nom avec la terminaison du possessif Xn, XbÓl + látszik
Composantes sémantiques	(1) Les signaux corporels laissent deviner l'état d'âme d'une personne. (2) L'apport visuel est la base d'une observation.
Composantes pragmatiques	Pas de composantes pragmatiques particulières.

3) INF + látsz* (sembl* + INF)

Collocations typiques	« alátámasztani » (soutenir), « megdönteni » (réfuter), « igazolni » (confirmer) et d'autres verbes liés à la formulation d'une hypothèse
Colligations typiques (1)	Formes typiques du verbe : « látszik » (troisième personne du singulier, indicatif présent), « látszott » (troisième personne du singulier, indicatif passé)
Colligations typiques (2) : ordre des mots	INF + látszik
Composantes sémantiques	Confirmation d'une hypothèse avec un élément de précaution Réfutation polie d'une hypothèse
Composantes pragmatiques	Dans les phrases impliquant la réfutation d'une hypothèse, la politesse incite à l'usage de « látszik » qui modère l'avis négatif.

4) XnAk látsz* (parai*/sembl* X)

Collocations typiques	« biztosnak » (sûr), « valószínűnek » (vraisemblable), « bizonyosnak » (certain) et d'autres adjectifs qui indiquent le degré de justesse de l'impression « elkerülhetetlennek » (inévitable), « reménytelennek » (sans espoir) et d'autres adjectifs ou adjectifs + noms qui qualifient l'impression
Colligations typiques (1)	Formes typiques du verbe : « látszik » (troisième personne du singulier, indicatif présent), « látszott » (troisième personne du singulier, indicatif passé)
Colligations typiques (2) : ordre des mots	ADJ + látszik ADJ + N + látszik
Composantes sémantiques	Indique le degré de justesse d'une impression Qualifie l'impression
Composantes pragmatiques	Pas de composantes pragmatiques particulières

5) Látsz*, hogy (Ça se voi* que, il est évident que)

Collocations typiques	Pas de collocations particulières.
Colligations typiques	Pas de colligations particulières.
Composantes sémantiques	Le locuteur constate que les faits sont déductibles de ce que l'on voit.
Composantes pragmatiques	Précédé d'une description des faits qui donnent raison à la conclusion introduite par « látszik, hogy ... »

Tableau 110 : Le profil du verbe « látszik ».

G) « Quelle est la différence ? » Profil contrastif de « tűnik » et « látszik »

Le profil des deux verbes, tel que révélé par les deux grands corpus, « huTenTen12 » et le Corpus national du hongrois peut être résumé comme suit (ordre déterminé par la fréquence) (tableau 111) :

TŰN-*	**LÁTSZ-***
1. úgy tűnik, (hogy) il semble que ⇓ Certitude faible de la justesse de l'impression	**1. úgy látszik, (hogy)** il semble que ⇓ Certitude plus forte de la justesse de l'impression/observation que « tűnik »
2 ADJ-nAk tűnik ⇓ Impression, opinion tirée de l'observation/des faits	**2. INF + látszik** ⇓ Faits permettant de tirer une conclusion (utilisé avant tout dans le langage érudit, souvent associé à des hypothèses et à des observations)
3 ADJ-nAk tűnhet, de / **talán ADJ-nAk tűnik, de** **(peut sembler ADJ mais)** ⇓ Première impression peu fiable, opposée à la réalité	**3.1 látszik** ⇓ Quelque chose ou quelqu'un est visible, détectable à l'œil nu **3.2 X-n/XbÓl látszik, hogy...** ⇓ (1) Quelque chose ou quelqu'un est reconnaissable sur un apport visuel (photo, vidéo)

	(2) Un état d'âme est révélé par les signes corporels
4 egyelőre/pillanatnyilag (úgy / ADJ-nAk) tűnik	**4.1 ADJ-nAk látszik** **4.2 ADJ + N-nAk látszik**
⇓	⇓
Situation momentanée qui peut encore changer	Conclusion tirée de l'observation/des faits
5 azért tűnik / tűnhet X-nek, mert ...	**5. Látszik, hogy ...**
⇓	⇓
Explication de la raison d'une impression	La validité de l'impression est évidente pour le locuteur

Tableau 111 : Profil contrastif de « tűnik » et « látszik ».

« Tűnik » et « látszik » sont deux verbes dont l'usage s'explore de façon efficace à partir des unités multi-lexicales ; ces verbes peuvent donc être présentés dans leur environnement textuel selon leurs différents sens dès les premiers moments de leur introduction à l'apprenant[82].

Il ressort de nos explorations que « tűnik » et « látszik », bien qu'ils présentent certaines similitudes de sens, se comportent, en fait, différemment dans les énoncés réels. Les collocatifs des deux verbes sont parfois similaires, mais le plus souvent, ils diffèrent sur au moins un point qui peut être lexical, grammatical ou sémantique. En cas d'unités multi-lexicales identiques, il existe une différence stylistique ou sémantique dans l'utilisation des deux verbes. Ces informations, issues d'une analyse statistique systématique du corpus, permettent de révéler de manière quantitative ces différences d'usage significatives alors qu'elles sont absentes des dictionnaires existants. Nous voyons donc un bénéfice direct pour le linguiste autant que pour l'enseignant et à l'apprenant d'une observation à grande échelle des énoncés authentiques.

Dans ce chapitre, nous avons exploré les tendances de l'utilisation de deux synonymes, « tűnik » et « látszik » à partir du corpus. Le chapitre suivant analysera un autre binôme de synonymes « megjön » et « eljön », deux verbes à préfixes, avant de résumer nos résultats et d'en tirer des conclusions.

[82] Voir aussi le chapitre 12 sur les considérations pédagogiques.

Chapitre 10 : « Quelle est la différence ? » Les synonymes « eljön » et « megjön »

Ce chapitre explore l'environnement textuel de deux synonymes, « eljön » et « megjön », tous les deux traduisibles par « venir » ou « arriver », selon le contexte. Ces deux verbes sont intéressants car ils touchent à l'utilisation des préfixes, phénomène traité, selon les cas, sous l'angle grammatical ou lexical. Le hongrois n'est bien sûr pas la seule langue à appliquer des préfixes : c'est le cas aussi pour d'autres langues comme, par exemple, les langues slaves ou germaniques. Ce n'est donc pas sa singularité qui rend cet aspect de la langue problématique dans le cas du hongrois mais la difficulté de fournir pour cette langue des règles claires d'usage et de définir le sens des préfixes de façon exacte, exhaustive et transparente.

Nous choisirons dans ce chapitre un ordre légèrement différent du cas de « tűnik » et « látszik » (voir le chapitre 8) : nous classerons ici un certain nombre d'exemples selon des catégories sémantiques, et nous effectuerons, par la suite, une étude des unités multi-lexicales. Cette approche nous permettra une meilleure comparaison des deux verbes dont l'usage se distingue, comme nous le verrons, surtout par les collocatifs typiques et par certains modificateurs. Une autre particularité de ce duo de synonymes est que tous deux peuvent former des phrases à aux seuls : « Eljöttem. » (Je suis venu/arrivé.) et « Megjöttem. » (Je suis arrivé/rentré), mais le contexte détermine dans ce cas clairement quel verbe doit être utilisé.

Comme les deux verbes sont fréquents à l'écrit et à l'oral, notre analyse cherchera également à mettre en évidence les différences entre les résultats dans les corpus écrits et oraux. Malgré la taille limitée de nos corpus oraux, ils s'avèrent cependant suffisants cependant pour faire émerger les usages propres au langage parlé ou, plus précisément, au langage des interactions. Ils sont en revanche trop restreints pour en déduire des schémas clairs d'usage. L'analyse des corpus oraux reste ainsi limitée.

A) Que disent les grammaires pédagogiques ?

Comme nous l'avons indiqué, les verbes « megjön » et « eljön » se traduisent par « venir » ou « arriver », selon le contexte. Le verbe « jön » sans préfixe signifie « venir », les préfixes « meg- » et « el- » n'ont pas de significations facilement identifiables ; les grammaires pédagogiques existantes se restreignent à constater que les préfixes « meg- » et « el- » servent typiquement à indiquer la perfectivité (Szili 2001 ; Szende et Kassai 2001; Szita et Görbe 2009). Comme le résument Szende

et Kassai (2001 : 265) : « Une des fonctions essentielles du préfixe consiste à signaler l'accomplissement ou le résultat d'un processus ». Les grammaires restent précautionneuses quant au sens et à l'usage des préfixes et insistent sur la multitude de leurs fonctions : ils peuvent exprimer la perfectivité, le début ou la fin d'un mouvement (sens itératif et duratif), l'intensité, la totalité et l'exhaustivité de l'action mais ils peuvent aussi complètement changer le sens du verbe (Balogh et al. 2000 ; Hegedűs 2004 ; Keresztes 1995 ; Kiefer 2006). Les auteurs constatent également qu'il n'est pas possible d'attribuer un sens dominant aux préfixes car la majorité des fonctions listées ci-dessus peuvent être remplies par n'importe quel préfixe. La description de l'utilisation des préfixes reste dans les ouvrages pédagogiques plutôt succincte, en raison de la difficulté de donner des règles claires, valides dans un grand nombre de cas.

Le cas spécifique de « megjön » et « eljön » a été envisagé par Waseda dans son article intitulé « Quelle est la différence entre "Eljött a tavasz" et "Megjött a tavasz" (Le printemps est arrivé.) ? ». Waseda effectue une analyse de ces deux phrases à traduction identique, dans le cadre de la sémantique cognitive et conclut ainsi[83] : « Le sens prototypique de « megjön » est « retourner », il indique un événement cyclique. Le préfixe « meg- » attire l'attention sur l'arrivée, sur la fin du mouvement comme état. Le sens prototypique de « eljön » est « se rapprocher depuis le lointain vers un point ciblé ». » Waseda remarque aussi que « le sujet du verbe « eljön » est souvent un être humain, notamment une personne attendue ou encore un événement personnifié de façon métaphorique » (Waseda 2017 : 98). Cette étude ne repose cependant pas sur une exploration rigoureuse et systématiques de données empiriques. Nous proposons ici d'étendre ce travail en nous basant sur une analyse de corpus.

B) Que dit le dictionnaire ?

Nous commencerons nos explorations par l'étude des sens de « eljön » et de « megjön » listés dans le « Dictionnaire monolingue de la langue hongroise » (2021). La classification dans cet ouvrage fournira le point de départ pour l'analyse de corpus.

[83] « A *megjön* ige prototipikus jelentése 'visszaérkezés', azaz ciklikus eseményre utalhat. A *meg-* igekötő grammatikalizációval, a mozgás végpontba érésére való figyelemáthelyezéssel a megérkezésre mint perfektív állapotra vonatkozhat. Az *eljön* ige prototipikus jelentése viszont 'távolból közelítés a célponthoz'. (...) az *eljön* ige alanya inkább ember, mégpedig valamilyen várt személy vagy metaforikusan megszemélyesített esemény. »

« eljön »

1. vient et arrive près du locuteur, dans son appartement ou dans le lieu en question ; s'approche et arrive. *Eljössz hozzám? (Tu viens chez moi ?) Jöjjön el holnap újra. (Revenez demain.) Sokan ott voltak az ünnepélyen, kár, hogy ti nem jöttetek el. (Il y avait du monde à la célébration, c'est dommage que vous ne soyez pas venus.)*

Part ou prêt à partir là où va le locuteur : *Eljössz velünk kirándulni? (Tu viens avec nous faire une balade ?) Ki jön el velem? (Qui vient avec moi ?)*

2. Part d'un endroit vers le locuteur en quittant son lieu de résidence : *Már eljött hazulról. (Il est déjà parti de chez lui.) Mikor visszamentem, még ott volt, csak később jött el. (Il était encore là quand je suis retourné, il n'est parti que plus tard.) Eljött egész a sarokig. (Il est venu avec moi jusqu'au coin.)* 3. (sens figuré) <Le temps, l'heure de qqch> arrive. *Végre eljött az ő napja. (Enfin, sa journée est arrivée.) Eljött a cselekvés órája. (L'heure de l'action est arrivée.) Eljött az indulás ideje. (Le moment du départ est arrivé.)*

« megjön »

1. <Personne> est de retour, arrive chez soi où le locuteur se trouve. *Megjön az iskolából. (Il rentre de l'école.) Megjött a külföldi útjáról. (Il est de retour de son voyage à l'étranger.) Megjöttek a fecskék. (Les hirondelles sont de retour.)*

2. <Personne> arrive là où elle est attendue et où le locuteur se trouve. *Megjöttek a vendégek. (Les invités sont arrivés.) Csakhogy megjöttetek! (Enfin, vous êtes là !) Megjött a csomag. (Le colis est arrivé.)*

3. (sens figuré) <Temps, phénomène, situation récurrent(e)> est là à nouveau. *Megjött a tavasz. (Le printemps est là.)* 4. (sens figuré) <Chose attendue ou phénomène attendu> se réalise. *Végre megjött az eső. (Enfin, la pluie est arrivée.)* <idom. : une abilité, une condition psychique ou physique> se développe, se rétablit. *Megjön a bátorsága. (Il a retrouvé son courage.) Megjön az esze. (Il redevient raisonnable.) Megjött az étvágya. (Il a retrouvé son appétit.)*

Les explications et les exemples fournis par l'encyclopédie s'adressent avant tout au natif en lui permettant de systématiser les connaissances implicites qu'il possédait déjà avant de lire ces entrées. Cependant, l'apprenant à un niveau de compétences linguistiques inférieur qui doit *apprendre* les différents usages, aura du mal à interpréter cette présentation, car il a nettement moins d'expériences linguistiques qu'un locuteur natif.

Il est également difficile de comprendre le sens de ces verbes à partir d'une seule traduction car celle-ci risque d'être identique pour les deux verbes. Le grand dictionnaire hongrois-français

(Akadémiai Kiadó 2021) propose ainsi comme traduction pour « megjön » (1) « arriver, venir, parvenir », (2) « revenir, être de retour ». « Eljön » est traduit par (1) « venir, accourir, arriver » ou (2) « quitter définitivement qqch (par exemple, son travail) ». Il est probable que l'apprenant ait des difficultés à décider s'il doit employer « megjön » et « eljön » dans une phrase indiquant qu'un objet ou une personne arrive ou est arrivé. « Megjön » et « eljön » sont, cependant, loin d'être interchangeables même si nous avons des difficultés à expliciter leurs différences[84]. Comme pour l'analyse des synonymes « tűnik » et « látszik », nous utiliserons le corpus « huTenTen12 » et le Corpus national du hongrois. Nous étudierons tout d'abord la fréquence de ces deux verbes dans ces corpus, puis, à travers de nombreux exemples, leurs environnements typiques, leurs caractéristiques sémantiques, pragmatiques et grammaticales. Ce procédé nous permettra de cartographier des schémas d'usage de « megjön » et de « eljön ». Notre objectif principal est donc de *présenter un procédé pour l'analyse des verbes à préfixe fondée sur le corpus qui peut être utilisé dans le cadre pédagogique.*

C) Le verbe « eljön » : que dit le corpus écrit ?

Dans les sous-parties consultées du Corpus national du hongrois ainsi que dans « huTenTen12 », nous trouvons des milliers d'exemples avec les verbes « megjön » et « eljön ». Le « huTenTen12 » liste 179 084 instances (56,6/million) et le Corpus national du hongrois 31 105 instances (45/million), la distribution étant presqu'identique dans les deux corpus[85]. Nous avons choisi 1000 occurrences au hasard par corpus que nous avons ensuite catégorisées selon leurs sens.

1) Arrivée d'un moment : « eljön » + moment du temps comme sujet

L'usage le plus fréquent correspond à une sous-partie de la deuxième signification dans l'encyclopédie (Le temps, l'heure de qqch arrive) : le verbe « eljön » est utilisé dans presque 50 % des cas (88 542 occurrences ou 49%) dans le corpus « huTenTen12 » avec un sujet qui fait référence à un moment dans le temps, ce groupe est donc le plus important (tableau 112).

Végre **eljött** az idő, hogy jobban megmutassuk magunkat.	***Le moment est enfin venu*** *pour nous de devenir plus visibles.*
Eljött a pillanat, amire már rég vártam.	***Le moment*** *que j'attendais depuis longtemps* ***est venu****.*

[84] Nous avons demandé à six collègues de décrire les usages possibles des deux verbes. De même qu'au sondage concernant « tűnik » et « látszik », les réponses étaient partielles.
[85] Distribution dans le CNH : Groupe 1 48%, Groupe 2 47%, Groupe 3 7%.

Hamar **eljött** a túra vége, és fájó szívvel, de elkezdtünk pakolni.	*Bientôt,* ***la randonnée a pris fin*** *et, avec un cœur lourd, nous avons commencé à faire nos valises.*
És **eljött** a várva várt délután.	*Et* ***vint l'après-midi tant attendu****.*
Teljesen kimerült, mire **eljött** a nagy nap.	*Il était complètement épuisé quand* ***le grand jour est arrivé.***
Mindenki életében **eljön** az a pillanat, hogy költözne vagy költöznie kell.	*Il* ***vient un moment*** *dans la vie de chacun où il bougera ou devrait bouger.*
Eljött a kapcsolati marketing kora.	***L'ère du marketing relationnel est arrivée.***
Egyszer csak **eljön** az a pillanat, hogy a természet ébredezni kezd.	*Tout à coup* ***vient le moment*** *où la nature commence à se réveiller.*
Eljött a videótechnika ideje.	***Le temps de la technologie vidéo est arrivé****.*
Eljön az utolsó Rockmaraton!	***Le dernier Marathon de Rock arrive*** *!*
És **eljött** a nap, amikor már csak egy dologra lett volna szükség.	*Et* ***le jour est venu*** *où une seule chose aurait été nécessaire.*
Hamar **eljött** a december.	*Décembre* ***est venu*** *vite.*

Tableau 112 : Exemples indiquant l'arrivée d'un moment (« huTenTen12 »).

Les phrases indiquent l'arrivée d'un moment de changement que le locuteur a attendu ou non. Les phrases fournissent aussi quelques informations grammaticales, en particulier sur l'ordre des mots : en général, le verbe précède le sujet et devant le verbe ne se placent que certains modificateurs : « hamar » (vite), « egyszer csak » (soudain), « végre » (enfin), « mire » (quand). Ces modificateurs montrent des similarités sémantiques car ils indiquent à quelle rapidité le moment arrive ainsi que l'attitude du locuteur à l'événement (le temps « ressenti »). Si la phrase ne contient pas de modificateur, elle commence souvent par le verbe.

2) Joindre le locuteur à un endroit, participer à un événement où le locuteur est présent : « eljön + LOC »

Le deuxième groupe correspond au premier sens du verbe (vient et arrive près du locuteur, dans son appartement ou dans le lieu en question ; s'approche et arrive) : ces phrases expriment qu'une personne vient à un endroit ou événement où se trouve le locuteur. Il est représenté avec une fréquence légèrement inférieure à celle du groupe 1 (82 611 occurrences ou 46 %) dans le corpus « huTenTen12 ». Au sein de ce groupe, nous avons créé quatre sous-groupes selon les différents schémas grammaticaux (tableau 113) :

(1) eljön + LOC

Több cég is **eljött** Pápára.

Köszönjük, hogy **eljöttek** a bemutatóra.

Reméljük, sokan **eljönnek** majd ebbe a táborba.

Vönöczky-Schenk Jakab ornitológus is **eljött** Budapestről Tárnokra.

(1) eljön + LOC

Plusieurs entreprises ***sont venues à Pápa****.*

Merci d'être ***venus à la présentation****.*

Nous espérons que beaucoup d'entre vous ***participeront à ce camp****.*

L'ornitologue Jakab Vönöczky-Schenk ***est venu de Budapest à Tárnok****.*

(2) eljön (+ LOC implicite, connu des locuteurs)

Miért érdemes **eljönni**? Először is, mert eltölthetsz egy jó hangulatú délutánt.

Még Gyöngyösről, Mosonmagyaróvárról is **eljönnek** a barátaim.

Köszönjük a segítséget mindenkinek, aki **eljött**.

Először **eljött,** és tanácsokat kért a gyűjtéssel kapcsolatban.

Csináltunk egy bulit, és nagyon sokan **eljöttek**.

Nem is gondoltam volna, hogy már első napon ilyen sokan **eljönnek**.

A szomszédos községekből is sokan **eljöttek**.

(2) eljön (+ LOC implicite, connu des locuteurs)

Quel intérêt de ***venir*** *? Tout d'abord, tu peux passer un après-midi sympa.*

Mes amis ***viennent*** *même d('aussi loin que) Gyöngyös et Mosonmagyaróvár.*

Nous remercions tout le monde qui ***est venu****, pour leur aide.*

Il ***est*** *d'abord* ***venu*** *(me voir et) demander conseil concernant la collecte.*

On a fait une soirée et beaucoup de gens ***sont venus****.*

*Je ne pensais pas qu'****il y aurait*** *autant de monde dès le premier jour.*

Beaucoup de gens ***sont venus*** *des villages avoisinants.*

(3) nem tud eljönni : 4 538

Sajnos csak Gabi tudott **eljönni**, talán legközelebb többen lesznek.

Dr. Lengyel Gyöngyi már nem tudott **eljönni**, de levélben ezt írta: ...

A polgármesterre vártunk, de végül csak a felesége tudott **eljönni**.

Ha valaki nem tud **eljönni** az órára, délután 14 óráig lemondhatja.

(3) nem tud eljönni : 4 538

Malheureusement seule Gabi ***a pu venir****, il y aura peut-être plus de monde la prochaine fois.*

Dr. Gyöngyi Lengyel ***n'a pas plus venir*** *mais elle a écrit la lettre suivante : ...*

Ils attendaient l'arrivée du maire mais finalement, ***seule sa femme est venue****.*

Si vous ***ne pouvez pas venir en cours****, vous pouvez annuler avant 14 heures.*

(4) el tud jönni (333)	
Jó, ha mindenki ráér, és **el** tud **jönni**.	***(4) el tud jönni (333)***
Csak az jelezze a részvételét, aki biztosan **el** tud **jönni**.	*C'est bien si tout le monde est disponible et* ***peut venir****.*
Várunk mindenkit, aki **el** tud **jönni** aznap.	*Confirmez votre présence seulement si vous êtes sûr* ***de pouvoir venir****.*
Jövő héten anyu **el** tud **jönni** egy egész napra.	*Nous attendons tout le monde qui* ***peut venir*** *ce jour-là. (Vous êtes tous les bienvenus.)*
Annak örülök a legjobban, hogy Dávid **el** tud **jönni**.	*Maman* ***peut venir*** *pour une journée entière la semaine prochaine.*
Aki nagyon akar és ráér, az **el** tud **jönni**.	*Je suis très content que Dávid* ***puisse venir****.*
	Ceux qui le veulent vraiment et ont le temps, ***viendront****.*

Tableau 113 : Exemples indiquant que le sujet de la phrase est attendu à un endroit ou à un événement (« huTenTen12 »).

Dans un grand nombre de ces phrases, l'endroit ou l'événement où le sujet de la phrase est invité, a été indiqué dans une phrase précédente et n'est pas répété dans la phrase contenant le verbe « eljön ». Toutes les phrases avec « eljön » sous-entendent l'intérêt pour l'événement proposé ou lancent une invitation pour celui-ci. Une différence notable concerne l'ordre des mots : dans ce cas, à la différence du premier groupe, le sujet précède généralement le verbe.

3) Partir d'un lieu, quitter un lieu provisoirement ou pour toujours : « eljön valahonnan » (partir + LOC)

Les deux derniers groupes présentés ci-dessous sont plutôt rares. Le groupe 3 (5% des occurrences) comprend des phrases faisant partie des récits dans lesquels une personne a quitté un lieu – pour toujours ou, plus rarement, avec l'intention d'y retourner (tableau 114) :

Három hét telt el, amióta **eljöttem** otthonról.	*Trois semaines se sont passées depuis que* ***je suis parti de chez moi****.*
Az egész úgy kezdődött, hogy **eljöttünk** Budaörsről.	*Tout a commencé quand* ***nous sommes partis de Budaörs****.*
Négy éve **jöttem el** a színháztól.	***J'ai quitté le théâtre*** *il y a quatre ans.*

Négyéves volt, amikor én **eljöttem** Debrecenből.	*Il avait quatre ans quand* ***je suis parti de Debrecen****.*
Eljöttem Nürnbergből, és Stuttgartban éltem egy darabig.	***J'ai quitté Nurenberg*** *et vécu pendant un certain temps à Stuttgart.*
Megkönnyebbültem, mikor **eljöttem** a vendéglőből.	*J'étais soulagé quand* ***j'ai quitté le restaurant*** *(j'ai démissionné du restaurant).*
Év közben **jöttem el** a gimiből.	***Je suis parti du lycée*** *pendant l'année scolaire.*

Tableau 114 : Exemples pour « partit d'un lieu » (« huTenTen12 »).

Ces phrases évoquent en général un événement du passé et font partie d'un schéma narratif plus long évoquant la vie d'une personne.

4) Vient chercher quelque chose ou quelqu'un : eljön N-ért

1 468 occurrences témoignent de cet usage dans le corpus. Il se traduit facilement en français : « eljön » accompagné d'un nom ou d'un pronom suffixé de « -ért » signifie « venir chercher un objet ou une personne » (tableau 115).

(1) eljön valamiért	**(1) Vient chercher quelque chose**
A volt barátja **eljött** a lakáskulcsért.	*Son ex-copain* ***est venu chercher la clé de l'appartement****.*
Nem tudsz **eljönni** az eszközért, ezért szeretnéd, hogy házhoz vigyük?	*Tu ne peux pas* ***venir chercher l'outil*** *et voudrais qu'on te le livre à domicile ?*
Rengeteg teendője mellett időt szakított arra, hogy **eljöjjön** a tortájáért.	*Il a pris le temps malgré ces multiples tâches pour* ***venir chercher le gâteau****.*
A fiú, aki **eljött** a csomagért, roppant udvarias volt.	*Le garçon qui* ***est venu chercher le colis****, était très poli.*
(2) eljön valakiért	**(2) Vient chercher quelqu'un**
Zsanettért **eljöttek** a tévéből.	*La télé* ***est venue chercher Zsanett****.*
A halál újra **eljött** áldozataiért.	*La mort* ***est venue chercher ses nouvelles victimes****.*
Nincs messze az idő, amikor Krisztus **eljön** a keresztényekért.	*Proche est le temps où le Christ* ***viendra pour les chrétiens****.*

Tableau 115 : Exemples pour « venir chercher quelqu'un ou quelque chose (« huTenTen12 »).

Le sens « venir chercher quelqu'un quelque part » (eljön valakiért valahova) étend le domaine sémantique du verbe par rapport aux occurrences déjà étudiées. Ici, le verbe ne signifie pas l'arrivée à un endroit comme dans le cas des groupes 1 à 3, mais décrit une action de courte durée suivie d'un départ.

D) Le verbe « eljön » dans les corpus oraux

Les usages et leurs fréquences observés dans le corpus d'entretiens à la radio dans le Corpus national du hongrois correspondent aux résultats dans le corpus écrit. Cela s'explique par le fait que le corpus d'entretiens remplit une fonction similaire au corpus écrit étudié : il comporte des récits, des opinions et des informations liés aux actualités. Même si le corpus contient des interactions, celles-ci ne sont pas les interactions du quotidien dans lesquelles le langage interactionnel joue un rôle plus important, comme nous le verrons par la suite.

Le corpus oral des interactions du quotidien (discussions informelles et semi-formelles, échanges avec des fournisseurs de services[86]) révèle certaines différences par rapport au corpus écrit. Il s'agit en première ligne d'une distribution différente de fréquences des usages plutôt que de sens nouveaux.

1) Joindre le locuteur à un endroit, participer à un événement où le locuteur est présent : « eljön + LOC »

Le sens que nous retrouvons le plus souvent est celui de « venir quelque part (à un lieu ou à un événement) où se trouve le locuteur ». Il ne s'agit donc pas d'un nouvel usage mais d'un usage plus fréquent à l'oral qu'à l'écrit se manifestant dans 57 % des occurrences (tableau 116) :

Ugye, **eljössz**?	*Tu viendras, n'est pas ?*
Eljössz?	*Tu viendras ?*
Eljössz megnézni a meccset?	*Tu viendras regarder le match ?*
Hát **eljöttél**.	*Tu es venu alors.*
Tudtam, hogy **eljössz**!	*Je savais que tu allais venir.*
Remélem, **el tudtok jönni**.	*J'espère que vous pouvez venir.*
Köszönöm, hogy **eljöttél**.	*Merci d'être venu.*

[86] Voir le chapitre 14 pour plus de détails.

Kösz, hogy **eljöttetek**.	*Merci d'être venus.*
Jó, hogy **eljöttél**.	*C'est bien que tu sois venu.*
Jól tetted, hogy **eljöttél**.	*Tu as bien fait de venir.*
Örülök, hogy **eljöttél**.	*Je suis content que tu sois venu.*
Örülnénk, ha **eljönnél**.	*Je serais content si tu pouvais venir.*
Szeretném megköszönni, hogy **eljöttek** a dr. Meinheimer tiszteletére adott fogadásunkra.	*Je vous remercie d'être venus à notre réception en l'honneur du dr. Meinheimer.*
Ha te ott leszel, talán a többiek is **eljönnek**.	*Si tu es là, peut-être que les autres viendront aussi.*
	Tu es venu nous voir ?
Eljöttél meglátogatni bennünket?	*Tu penses qu'il viendra ?*
Szerinted **eljön**?	*Tu es venu voir tes amis ?*
Eljöttél meglátogatni a barátaidat?	*Il y avait beaucoup de monde à la soirée ?*
Sokan **eljöttek** a bulira?	*J'aimerais que vous participiez à ce voyage.*
Szeretném, ha **eljönne** erre az útra.	*Quand il est venu me voir, il savait déjà qu'il allait partir.*
Amikor **eljött** hozzám, már tudta, hogy el fog utazni.	*Malheureusement, je n'ai pas pu venir.*
Sajnos nem tudtam **eljönni**.	*Tu peux venir ?*
El tudsz **jönni**?	*Je ne peux pas venir.*
Nem tudok **eljönni**.	

Tableau 116 : Exemples avec le verbe « eljön » indiquant qu'une personne joint le locuteur à un endroit ou participe à un événement où le locuteur est présent (corpus oral).

En comparant ces phrases avec les occurrences dans le corpus écrit, nous constatons que la fonction d'un certain nombre de phrases est différente : il s'agit habituellement de remerciements, d'invitations ou d'interactions dans lesquelles l'interlocuteur voudrait savoir si le ou la partenaire de conversation est disponible. Le locuteur s'adresse donc directement à une personne ou à un groupe de personnes. Pour cette raison, la première et la deuxième personnes ont une présence forte et le présent est utilisé plus souvent que dans le corpus oral d'entretiens et dans le corpus écrit. Rappelons néanmoins que notre corpus oral est de taille limitée, et les conclusions qu'il permet de tirer n'ont ici qu'un caractère indicatif. Bien que notable, l'observation qualitative d'une différence entre les usages oraux interactionnels ceux relevant de récits mériteraient une étude plus approfondie, pour l'instant impossible en l'absence de corpus appropriés.

2) Arrivée d'un moment : « eljön » + moment du temps comme sujet

Cet usage, le plus fréquent dans le corpus écrit, est le deuxième plus grand groupe dans le corpus d'interactions. Il se manifeste dans 31% des occurrences (tableau 117) :

Eljön még a mi időnk!	*Notre temps viendra encore.*
Eljön a te időd.	*Ton temps viendra.*
Eljön mindennek az ideje.	*Tout a son temps.*
Tudtam, hogy egyszer **eljön** ez a nap.	*Je savais que ce jour allait arriver.*
Eljött az én időm.	*Mon temps est venu.*
Eljött a nagy nap.	*La grande journée est arrivée.*
Eljött a búcsú ideje.	*Le temps des adieux est arrivé.*
Eljött a döntő pillanat.	*Le moment décisif est venu.*
Eljött a karácsony.	*Noël est arrivé.*
Eljött a szülinapom.	*Mon anniversaire est arrivé.*
Eljött augusztus eleje, és még mindig semmi.	*Mi-août est arrivé et toujours rien.*

Tableau 117 : Exemples indiquant l'arrivée d'un moment dans le temps (corpus oral).

Comme cette utilisation fait référence à un moment dans le temps et fait partie d'un récit (en général au passé), nous n'observons pas de différences entre l'usage dans les corpus oraux et écrits. Notons cependant que les phrases commencent avec le verbe, l'accent est sur l'action.

3) Partir d'un lieu, quitter un lieu provisoirement ou pour toujours : « eljön valahonnan » (partir + LOC)

Le troisième groupe (correspondant à une utilisation plus rare que nous observons dans 7% des cas) est formé des énoncés indiquant que quelqu'un quitte un lieu, provisoirement ou pour toujours (tableau 118) :

Láttam, hogy nincs szükség rám, ezért **eljöttem**.	*J'ai vu qu'on n'avait pas besoin de moi, je suis donc parti.*
Felvettek Egerbe, de onnan egy év múlva **eljöttem**.	*J'ai été admis (à une école) à Eger mais j'en suis parti après un an.*
Én **eljöttem** Magyarországról 1951-ben.	*Je suis partie de Hongrie en 1951.*
Hát, már volt kétszer, hogy **eljöttem** otthonról.	*Ben, ça m'est déjà arrivé deux fois de partir de la maison.*

Szakközépiskolába jártam Tatabányán, utána onnan **eljöttem**.	*J'ai été dans un lycée professionnel à Tatabánya, puis je suis parti.*
Az igazság az, hogy majdnem **eljöttem** a konferenciáról.	*À vrai dire, j'ai failli partir de la conférence.*
Veszprémben laktam, meghalt a nevelőapám, és akkor **eljöttem**.	*J'habitais à Veszprém et quand mon beau-père est décédé, je suis parti.*

Tableau 118 : Exemples indiquant que quelqu'un quitte un lieu provisoirement ou pour toujours (corpus oral).

Ces phrases font partie des récits sur l'histoire d'une personne. Dans le corpus oral cette personne est en l'occurrence le locuteur qui partage des informations sur sa vie.

4) Accompagner quelqu'un quelque part ou inviter quelqu'un quelque part : « eljön XvAl + LOC » (venir avec X + LOC)

Le quatrième groupe contient des énoncés dans lesquels « eljön » est utilisé dans le sens d'« accompagner le locuteur quelque part ou de participer à un événement/une action avec lui ». Cet usage apparaît dans 4% des cas (tableau 119).

Eljössz velem Párizsba, vagy nem?	*Tu viens avec moi à Paris ou pas ?*
Eljössz velem moziba?	*Tu viens avec moi au cinéma ?*
Ágnes is **eljött** velünk túrázni.	*Agnès est venue avec nous à la randonnée.*
Eljössz velünk?	*Tu viens avec nous ?*

Tableau 119 : Exemples indiquant qu'une personne accompagne une autre à un lieu donné (corpus oral).

Cet usage est nouveau par rapport aux corpus écrits : il s'agit d'interactions spontanées, souvent liées à une situation ponctuelle dans laquelle un des locuteurs propose une activité ou invite l'autre personne à un événement ou à un lieu et le deuxième locuteur doit décider s'il accepte ou pas. Le lieu peut être implicite quand el est connu des locuteurs, tel est le cas dans le dernier exemple du tableau 119.

5) Venir chercher quelque chose ou quelqu'un : « eljön valamiért/valakiért »

Ce groupe est de taille très limitée (1%) avec un sens précis : les énoncés décrivent l'action de venir chercher quelque chose ou quelqu'un. Les phrases présentent la perspective de celui qui parle (tableau 120).

Eljött a gyűrűért.	*Il est venu chercher la bague.*

Eljött értem apukám, és hazavitt.	*Mon père est venu me chercher.*
El tudsz **jönni** értem?	*Peux-tu venir me chercher ?*
Eljövök érted.	*Je viens/viendrai te chercher.*
Eljövök érted, ha akarod.	*Je viens/viendrai te chercher si tu veux.*

Tableau 120 : Exemples du sens « venir chercher quelque chose ou quelqu'un » (corpus oral).

E) Observer l'environnement textuel de plus près : les unités multi-lexicales

Les outils numériques permettent de cartographier plus précisément l'environnement textuel du verbe et d'identifier les unités multi-lexicales plus longues, plus complexes associées aux différents usages du verbe choisi. Ils aident à répertorier les sujets (section 1), les compléments (section 2) et les modificateurs typiques (section 3).

1) Les sujets de « eljön/eljött » (quelque chose / quelqu'un vient/est venu)

L'usage dominant dans le corpus écrit (un moment dans le temps arrive) est, en effet, reflété par les sujets typiques associés au verbe « eljön/eljött » que montre le tableau 121 :

a pillanat	*le moment*	a perc, hogy	*le moment où*
az idő	*le temps*	a Karácsony	*Noel*
a nap	*le jour*	a Kánaán	*le Canaan*
a tavasz	*le printemps*	az alkalom, hogy	*l'occasion où*
az óra	*l'heure*	az a kor, hogy	*l'époque où*
a vég	*la fin*	a szombat	*le samedi*
Krisztus	*le Christ*	a péntek	*le vendredi*
a nyár	*l'été*	a tél	*l'hiver*
Jézus	*Jésus*	a másnap	*le lendemain*
az a pont, amikor	*le moment où*	a halál	*la mort*
Messiás	*le Messie*	a hajnal	*l'aube*
az az időszak, amikor	*la période/le temps où*	a hétfő	*le lundi*
		az ebédidő	*la pause de midi*

Tableau 121 : Les sujets les plus fréquents du verbe « eljött » (X est arrivé/venu) dans « huTenTen12 ».

Les sujets de ces phrases indiquent un moment particulier dans le temps, le plus souvent un moment du passé. Le verbe est donc majoritairement utilisé au passé quand il exprime ce sens. Certaines phrases font référence à des contextes religieux et parlent, par exemple, de l'arrivée du

Christ, du Messie, de Jésus ou du Canaan. Ces phrases utilisent le verbe au présent car il s'agit d'une prophétie[87].

2) Joindre le locuteur à un endroit, participer à un événement où le locuteur est présent : « eljön + LOC »

La liste dans le tableau 122 montre les collocations les plus fréquentes avec la structure « X eljön + LOC » :

X eljön/eljött	*X vient/est venu*	az esküvőre	*au mariage*
a koncertre	*au concert*	látogatóba	*visite*
otthonról	*de chez soi*	Budapestre	*à Budapest*
a rendezvényre	*à l'événement*	a bálra	*à la balle*
a bulira	*à la soirée*	az előadásra	*au spectacle*
a találkozóra	*à la rencontre*	Magyarországra	*en Hongrie*
a temetésre	*aux obsèques*	onnan	*de là-bas*
a megnyitóra	*au vernissage*	idáig	*jusqu'ici*
az ünnepségre	*à la fête*	ide	*ici*

Tableau 122 : Les sujets les plus fréquents du verbe « eljön » dans « huTenTen12 ».

Tous les compléments listés, à l'exception de « otthonról » (de chez soi) et « onnan » (de là-bas), s'associent au sens « X vient quelque part, X participe à un événement ». Le mot « otthonról » apparaît dans des phrases indiquant que quelqu'un est parti de chez soi (pour un temps limité avec l'intention d'y retourner ou pour toujours) ». Le sujet de ces phrases est ici une personne ou un groupe de personnes.

3) Les modificateurs de « eljön »

La liste suivante présente les modificateurs les plus fréquents avec « eljön » (tableau 123) :

végre	*enfin*	lassan	*doucement, lentement*
hamarosan	*d'ici peu*	végül	*à la fin*
egyszer	*une fois*	majd	*un jour*
hamar	*rapidement*	úgyis	*de toute façon*
nemsokára	*dans pas longtemps*	újra	*de nouveau*
aztán	*puis*	amióta	*depuis*

[87] Le hongrois utilise en général le présent pour faire référence aux événements qui se dérouleront dans l'avenir.

előbb-utóbb	*tôt ou tard*	este	*le/ce soir*
mielőtt	*avant*	megint	*de nouveau*
amikor	*quand*	holnap	*demain*
ismét	*à nouveau*		

Tableau 123 : Les modificateurs fréquents de « eljön ».

Comme le sens « un moment arrive » est le plus fréquent dans le corpus, il n'est pas surprenant que la majorité des modificateurs listés indiquent l'échéance ou la vitesse à laquelle ce moment arrive. Une plus petite partie de ces modificateurs indique qu'il s'agit du retour d'un événement (« újra », « ismét », « megint » (à nouveau)). Le reste des modificateurs s'associe aux sens « partir d'un lieu » (onnan) et « arriver à un lieu » (ide, idáig). L'utilisation de « úgyis » (de toute façon) peut signaler l'abandon de la résistance (quoi que le locuteur fasse, le moment viendra). Les modificateurs contiennent tous un élément de subjectivité faisant référence à l'attitude du locuteur : nous rencontrons des expressions indiquant le temps ressenti, l'attente, le soulagement ou l'abandon de résistance par rapport aux moments en question.

F) Que veut dire « eljön » ?

La dernière étape de cette analyse du corpus permet de créer un résumé des résultats d'observations. Comme dans les chapitres 8 et 9, nous présentons sous forme synthétique dans le tableau ci-dessous le profil des usages étudiés dans le corpus avec leurs caractéristiques d'usage par catégorie (tableau 124).

1) Un moment attendu dans le temps, arrive

Collocations typiques	**Sujets** : Mot ou expression indiquant un moment dans le temps, par exemple : « a nagy nap » (le grand jour), « a pillanat » (le moment), « az idő » (le temps), « a december » (décembre), « a tavasz » (le printemps). **Modificateurs** : « végre » (enfin), « hamarosan » (d'ici peu), « egyszer » (un jour, une fois), « nemsokára » (bientôt), « aztán » (puis), « amikor » (quand), « lassan » (lentement), « megint / ismét / újra » (à nouveau).
Colligations typiques (1)	(1) Verbe au passé, 3e personne : **(MOD « végre », « aztán » +) Eljött + X** **(MOD « végre », « aztán » +) Eljött + X + hogy** (2) Verbe au présent, 3e personne :

	MOD (« amikor », « lassan », « megint » etc.) + eljön + X **MOD (« amikor », « lassan », « megint » etc.) + eljön + X + hogy**
Colligations typiques (2) : ordre des mots	(1) Passé : le verbe reçoit de l'emphase et précède en général le sujet. **(MOD +) Eljött + X** (2) Présent : le verbe est précédé d'un modificateur (complément de temps). Le verbe reçoit de l'emphase et précède en général le sujet : **MOD + eljön + X**
Composantes sémantiques	Un moment (décisif, de changement), que le locuteur a attendu ou non, est là.
Composantes pragmatiques	Pas de composantes pragmatiques particulières.

2) Une personne vient à l'événement auquel elle a été invitée

Collocations typiques	Sujet : une personne ou un groupe de personnes
Colligations typiques (1)	Présent et passé, toutes les personnes grammaticales : **« eljön.*/eljött.* » + LOC (directionnel)** « el tud.* jönni / nem tud.* eljönni » + LOC (dir.) « gyere el » + LOC (dir.)
Colligations typiques (2) : ordre des mots	Le verbe reçoit de l'emphase et précède en général le complément de lieu. **« eljön.* » + LOC (directionnel)**
Composantes sémantiques	Participer ou anticiper la participation à un événement plaisant.
Composantes pragmatiques	Première et deuxième personnes : Phrases utilisées pour inviter quelqu'un à un événement, pour donner une réponse à une invitation ou pour remercier quelqu'un d'être venu à un événement.

3) Une personne quitte un endroit temporairement ou pour toujours

Collocations typiques	**Sujet** : Une personne ou un groupe de personnes
Colligations typiques (1)	Passé, toutes les personnes grammaticales : **(MOD +) « eljött.* » (+ LOC)**

Colligations typiques (2) : ordre des mots	Le verbe reçoit de l'emphase et précède en général le complément de lieu. **« eljön.* » + LOC (ablat.)**
Composantes sémantiques	Partir d'un lieu pour un temps limité avec l'intention d'y retourner ou quitter un endroit pour toujours.
Composantes pragmatiques	Pas de composantes pragmatiques particulières.

4) Venir chercher une personne ou une chose

Collocations typiques	Sujet : une personne ou un groupe de personnes
Colligations typiques (1)	Passé, toutes les personnes grammaticales : **(MOD +) « eljött* » + -ÉRT**
Colligations typiques (2) : ordre des mots	Le verbe reçoit de l'emphase et précède en général le complément de lieu. « eljött* » + -ÉRT + X
Composantes sémantiques	Vient chercher quelque chose ou quelqu'un. Séjour de courte durée au lieu donné.
Composantes pragmatiques	1e et 2e personne : Proposition ou demande

5) Accompagner quelqu'un quelque part

Collocations	Sujet typique : une personne ou un groupe de personnes Compléments typiques : « velem » (avec moi), « velünk » (avec nous) et des compléments de lieu
Colligations (1)	Présent ou passé, toutes les personnes grammaticales : **« eljö* » + velem/velünk + LOC**
Colligations (2) : ordre des mots	**« eljö* » + velem/velünk + LOC**
Composantes sémantiques	Proposition d'accompagner le locuteur quelque part.
Composantes pragmatiques	Le locuteur invite quelqu'un quelque part. Le locuteur récite une histoire.

Tableau 124 : Profil du verbe « eljön ».

Il est important de préciser que *ces tableaux ne dégagent pas de règles mais des tendances d'usage que l'on observe fréquemment dans le corpus*. La liste des caractéristiques n'est pas exhaustive mais elle fournit néanmoins une description suffisamment riche pour la sélection des éléments pertinents pour

l'enseignement qui, illustrés par de nombreux exemples peuvent aider l'apprenant à se forger une image fiable des différents usages.

G) Le verbe « megjön » dans les corpus écrits

Le verbe « megjön » est moins fréquent dans les corpus écrits étudiés que le verbe « eljön ». Il figure 56 247 fois dans le corpus « huTenTen12 » (versus « eljön » avec 179 084 occurrences) et une proportion comparable (1 : 3) et 15 392 fois dans le Corpus national du hongrois (versus « eljön » avec 46 435 occurrences). La proportion des deux verbes (1 : 3) et la distribution des sens de « megjön » ainsi que le pourcentage des usages au présent et au passé (40% et 60%) sont comparables dans les deux corpus.

Le corpus « huTenTen12 » et le Corpus national du hongrois nous permettent d'identifier trois usages principaux du verbe « megjön ». Les chiffres cités indiquent les nombres d'occurrences dans le corpus « huTenTen12 ». Pour les calculs de distribution des différents usages, nous avons utilisé 1 000 exemples tirés au hasard.

1) Une personne ou une chose que l'on a attendue, arrive

L'usage le plus courant (56% de l'ensemble des occurrences) indique qu'une personne ou une chose arrive là où elle est attendue et où le locuteur se trouve[88]. Voici quelques exemples du corpus « huTenTen12 » (tableau 125) :

Minden csomagunk **megjött**.	*Tous nos bagages sont arrivés.*
Másnap már **meg is jött** a válasz, hogy mehetek.	*La réponse que je peux y aller, est arrivée déjà le lendemain.*
A csekkek mindig rendesen **megjönnek**.	*Les chèques arrivent toujours en temps voulu.*
Épp befejeztük az ebédet, mikor **megjöttek** Ákosék.	*Nous venions de finir le déjeuner quand Ákos et les autres sont arrivés.*
Délutánra **szokott megjönni** a kenyér.	*Le pain arrive en général dans l'après-midi.*
És **a pénz** is megjön majd a jövő héten.	*Et l'argent arrivera aussi la semaine prochaine.*
Öt órát vártam, míg végre **megjött**.	*J'attendais depuis 5 heures quand il est enfin arrivé.*
A héten **megjött** a könyv, amit rendeltem.	*Le livre que j'ai commandé est arrivé cette semaine.*

[88] Comme l'environnement textuel n'indique aucune différence entre les cas dans lesquels le sujet est une personne et ceux dans lesquels le sujet est une chose, toutes ces occurrences se retrouvent dans le même groupe.

Múlt hét pénteken **megjött** e-mailben a nyilatkozat.	*Vendredi dernier, la déclaration est arrivée par mail.*
Végre **megjött** az orvos.	*Le médecin est enfin arrivé.*
Na, **megjött** a telefon, hogy felvettek.	*J'ai reçu un coup de fil disant que j'ai été admis.*
Tíz perc múlva **megjött** az edző is.	*Dix minutes plus tard, l'entraineur est arrivé.*

Tableau 125 : Exemples indiquant qu'une personne ou une chose attendue arrive ou est arrivée.

Dans la plupart de ces phrases (85 % des occurrences présentant cette signification), le sujet est une chose. En étudiant les exemples, nous pouvons constater qu'il s'agit souvent d'objets que le locuteur a commandés (il est donc logique qu'il les attende) ou de documents officiels (également attendus car annonçant une décision). Dans une partie relativement petite des exemples (15 %), le sujet est une personne dont le locuteur a attendu l'arrivée. Le verbe est le plus souvent au passé car le locuteur raconte une expérience vécue, l'usage au présent fait en général référence à un événement répétitif (par exemple, à l'arrivée régulière des chèques ou du pain ci-dessus).

2) Une personne ou un animal domestique est de retour (de quelque part)

Une autre grande catégorie (28 % de l'ensemble des occurrences) est composée de phrases reflétant le sens « quelqu'un est de retour, arrive chez soi où le locuteur/la personne (le sujet de la phrase) en question se trouve » (tableau 126).

Apám szeptemberben **jött meg** a frontról.	*Mon père est revenu de la guerre en septembre.*
Reménykedtek, hogy majdcsak **megjön** a rég elment családtag.	*Ils avaient bon espoir que le membre de la famille parti depuis longtemps serait de retour un jour.*
Van, amikor két hét, van, amikor egy hét múlva jön haza, de azért **megjön** mindig.	*Quelquefois cela prend deux semaines ou une semaine, mais il finit toujours par rentrer/ revenir.*
Elmegy sétálni, de már jóval hat előtt **megjön**.	*Il fait une promenade mais il sera de retour bien avant 6 heures.*
Kettőkor **jöttem meg** a suliból.	*Je suis rentré à deux heures de l'école.*

Tableau 126 : Exemples indiquant que quelqu'un est de retour.

Dans les phrases du groupe 2, le verbe « megjön » indique qu'une personne est de retour. 88% des exemples au passé indiquent le moment dans le temps où l'événement s'est produit (« kettőkor, két hét múlva, szeptemberben) ; 79% des exemples au présent contiennent des modificateurs indiquant l'attitude du locuteur (son attente, espoir ou anticipation).

3) Un phénomène météorologique arrive ou est de retour

Un plus petit groupe (7% de l'ensemble des occurrences) contient les phrases dont le sujet indique un temps ou un phénomène météorologique. Comme cet usage est bien spécifique, la variété de vocabulaire apparaissant dans ces phrases, est limitée. Ce groupe représente un usage entre le groupe 1 et le groupe 2, car certaines phrases annoncent que le phénomène météorologique donné a été attendu, d'autres ne signalent qu'un phénomène habituel est de retour (indépendamment du fait qu'il a été attendu ou non) (tableau 127).

Végre **megjött** a tavasz.	*Le printemps est enfin là.*
Megjött a fagy is.	*Il a gelé. (Le gel est arrivé.)*
Megjött az első hó.	*La première neige est tombée (arrivée).*
Napok múlva **megjött** az eső.	*Après quelques jours, la pluie est arrivée.*

Tableau 127 : Exemples indiquant qu'un phénomène météorologique arrive ou est de retour.

4) Retrouver, redécouvrir une qualité en soi (expressions idiomatiques)

Ces expressions idiomatiques n'apparaissent qu'occasionnellement dans le corpus (12% de l'ensemble des occurrences), avec un nombre limité de sujets (« envie », « appétit », « courage », « raison ») (tableau 128).

Délutánra **megjött** az étvágyam.	***Mon appétit est revenu*** *dans l'après-midi.*
A győzelem után **megjött** a csapat étvágya.	*Après la victoire,* ***l'équipe a eu faim****.*
A diétával, amit Dávid tart, **megjött** az étvágya is.	*Avec le régime qu'il suit, David* ***a retrouvé son appétit.***
Ezt a bejegyzést olvasva nekem is **megjött** a játékhoz a kedvem.	*En lisant ce commentaire, je* ***ressens l'envie de*** *jouer.*
Köszi szépen a sok pozitív hozzászólást, így egy kicsit jobban **megjött** hozzá a kedvem.	*Merci pour les commentaires positifs,* ***je me sens un peu plus motivé****.*
Átmenetileg még a tanuláshoz is **megjött** a kedve.	*Provisoirement,* ***il a*** *même* ***retrouvé le goût*** *de l'apprentissage.*
Végre **megjött** a bátorsága.	*Il* ***a*** *enfin* ***retrouvé son courage****.*
Végre **megjött** az eszed.	*Tu* ***es*** *enfin* ***raisonnable****.*

Tableau 128 : Exemples avec des expressions idiomatiques.

H) Le verbe « megjön » dans les corpus oraux

Comme dans le cas de « eljön », nous retrouvons des usages semblables à ceux dans les corpus écrits mais la distribution de la fréquence est légèrement différente. Nous trouvons plus d'occurrences à la première et à la deuxième personnes (en particulier pour le groupe 2) ainsi que plus d'énoncés s'intégrant dans une interaction, accompagnés d'une réaction. En voici quelques exemples (tableaux 129 et 130) :

1) Une chose ou une personne que l'on a attendue, arrive (enfin)

Itt van, **megjött** a válasz.	*Voici, la réponse est arrivée.*
Végre **megjött** az italunk.	*Enfin, nos boissons sont arrivées.*
Ha **megjönnek** az eredmények, kivel kell beszélnem?	*À qui dois-je m'adresser quand les résultats arrivent ?*
Megjött a pizza! – Kimész átvenni? Köszi.	*Enfin, la pizza est arrivée. – Tu peux aller la récupérer ? Merci.*
Na, jó, ülhetsz a fotelban, amíg **megjön** a taxi. – De kedves vagy!	*Tu peux rester assis dans le fauteuil jusqu'à ce que le taxi arrive. – Tu es sympa, toi.*
Ha **megjön** Peti, mit mondasz neki? – Még nem tudom.	*Que diras-tu à Peti quand il arrive ? – Je ne sais*
Megjött a postás. – Ilyen korán? Később szokott jönni.	*Le facteur est arrivé. – Déjà ? Normalement, il vient plus tard.*
Jó, hogy **megjöttél**.	*C'est bien que tu sois là/arrivé.*
Csak hogy **megjöttél**! – Miért ? Valami gond van ?	*Enfin, tu es arrivé ! – Pourquoi ? Y a-t-il un souci ?*
Beszélek a sráccal, ha **megjön**. – Rendben.	*Je parlerai au garçon dès qu'il arrivera. – Très bien.*

Tableau 129 : Exemples du corpus oral indiquant qu'une personne ou une chose attendue est arrivée.

2) Quelqu'un est de retour (de quelque part)

Nézzétek, gólyák! **Megjöttek** végre! – Tényleg!	*Regardez les cigognes ! Elles sont enfin de retour ! – C'est vrai.*
Megjöttem! – Jól van, örülök neki.	*Je suis là ! – J'en suis ravi.*
Megjöttem! – Szia!	*Tu es arrivé/rentré ? – Oui.*
Örülsz, hogy **megjöttem**? – Persze.	*Je suis de retour. – Salut !*
Megjöttél? – Aha.	*Es-tu content que je sois de retour ? – Bien sûr.*
Mindig olyan lehetetlen időpontokban tudsz **megjönni**! – Miért? Most mi a baj?	*Tu as le talent d'arriver aux moments impossibles ! – Pourquoi ? Qu'est-ce que j'ai fait de mal ?*

Azt mondta, várjunk, amíg **megjön**. – Jól van.	*Il nous a dit d'attendre jusqu'à ce qu'il soit de retour. – Ça marche.*
Mikor **jöttél meg** Amerikából? – Két hete.	*Quand est-ce que tu es revenu des Amériques ? – Il y a deux semaines.*
Na, **megjöttetek** a nyaralásból? – Igen.	*Alors, vous êtes rentrés de vacances ? – Oui.*

Tableau 130 : Exemples du corpus oral indiquant que quelqu'un est de retour.

Ces interactions reflètent les mêmes usages que ceux dans les corpus écrits. Leur intérêt principal est qu'elles fournissent du langage interactionnel. Les apprenants peuvent observer des annonces, des exclamations, des questions ainsi que des réactions liées à la situation donnée permettant d'améliorer leurs compétences interactionnelles.

I) Observer l'environnement textuel de plus près : identifier les unités multi-lexicales

Comme dans le cas de « eljön », des renseignements plus précis peuvent être obtenus concernant l'environnement textuel de « megjön » en établissant une liste des unités multi-lexicales associées habituellement au verbe. Les sections suivantes présentent la liste des sujets les plus fréquemment rencontrés en lien avec « megjön » (section 1), les compléments (section 2) et les modificateurs (section 3) les plus typiques.

1) Les sujets typiques

Les sujets les plus fréquents du verbe « megjön » sont liés aux usages 1 (qqn est de retour, arrive chez soi où le locuteur se trouve) et 2 (qqn ou qqch arrive là où il est attendu et où le locuteur se trouve.) Les saisons « tavasz » (printemps) et « tél » (hiver), associés à l'usage 3, ainsi que « appétit » et « envie » (usage 4) sont des noms qui se retrouvent haut dans la liste, probablement en raison de la variété bien plus limitée des noms qui peuvent s'associer à ces sens plutôt qu'aux sens 1 et 2 (tableau 131).

étvágy	*appétit*	vérzés	*règles*
tavasz	*printemps*	papír	*document*
válasz	*réponse*	eredmény	*résultat*
Jézuska	*petit Jésus*	menstruáció	*menstruation*
busz	*bus*	értesítés	*notification*
kedv	*envie*	SMS	*SMS*

Mikulás	*Père Noel*	erősítés	*renforcement*
levél	*lettre*	parancs	*commande*
tél	*hiver*	alkatrész	*pièce*
csomag	*colis*	apu	*Papa*
számla	*facture*	anyu	*Maman*
pizza	*pizza*	fizu	*salaire*
engedély	*autorisation*	csekk	*chèque*

Tableau 131 : Les sujets les plus fréquents du verbe « megjön » dans « huTenTen12 ».

2) Les compléments de lieu typiques

Le tableau suivant montre les compléments de lieu les plus fréquents dans le corpus, indiquant l'endroit d'où revient le sujet de la phrase (tableau 132).

a nyaralásból	*des vacances d'été*	Amerikából	*d'Amerique*
a melóból	*du boulot*	Rómából	*de Rome*
a sétából	*de la promenade*	Párizsból	*de Paris*
a suliból	*du bahut*	a táborból	*du camp de vacances*
a munkából	*du travail*	a boltból	*du magasin*
az iskolából	*de l'école*	Angliából	*d'Angleterre*
a dokitól	*de chez le toubib*	Londonból	*de Londres*
Svájcból	*de Suisse*	szabadságról	*des vacances*
az oviból	*de l'école maternelle*	a kórházból	*de l'hôpital*
Pestről	*de Pest*		

Tableau 132 : Les compléments de lieu les plus fréquents avec le verbe « megjön » dans « huTenTen12 ».

Il n'est guère surprenant que les endroits liés à des activités quotidiennes (école, travail, magasin) dominent dans le corpus. L'autre groupe est formé de noms propres géographiques (continents, villes, pays). Même si ces noms sont les plus fréquents dans le corpus, le nom le plus usité (nyaralás - vacances) figure 68 fois dans le corpus et le 20^{e} mot sur la liste (kórház - hôpital) 8 fois.

Il est intéressant d'observer que le corpus oral ne contient que trois exemples avec cette structure. Dans le reste des phrases, le lieu d'où un des participants de l'interaction est de retour n'est pas mentionné, implicite, connu du locuteur et des partenaires conversationnels. Dans ces dialogues, « én » (je) est le sujet dominant.

3) Les modificateurs typiques

Le tableau suivant liste les modificateurs du verbe « megjön » dans le corpus « huTenTen12 » (tableau 133) :

közben	*entretemps*	reggelre	*(avant) le matin (avant) le*
nemsokára	*bientôt*	estére	*soir*
hátha	*peut-être, on peut espérer que*	délben	*à midi*
		pénteken	*vendredi*
végre	*enfin*	délutánra	*(avant) l'après-midi*
időközben	*entretemps*	majdcsak	*un jour (peut-être)*
mindjárt	*d'ici peu*	talán	*peut-être*
amióta	*depuis*	egyszer majd	*un jour*

Tableau 133 : Les modificateurs fréquents du verbe « megjön » dans « huTenTen12 ».

Les modificateurs indiquant un moment dans le temps, s'associent aux deux usages les plus courants : ils précisent quand une chose ou une personne que le locuteur a attendue est arrivée ou une personne est de retour. Les modificateurs reflétant l'attitude du locuteur (« végre » (enfin), « hátha/talán » (peut-être), « majdcsak/egyszer majd » (un jour), s'associent à l'utilisation des quatre groupes au présent et expriment l'espoir ou l'anticipation de l'événement donné.

J) Que veut dire le verbe « megjön » ?

Comme pour « eljön », l'analyse des usages et des unités multi-lexicales permet de dresser un tableau synthétique des sens du verbe « megjön » émergeant du corpus. Le tableau 134 montre les schémas relatifs aux différents usages de « megjön ».

1) Personne ou chose attendue arrive

Collocations typiques	**Sujets** : Personne ou chose **Modificateurs fréquents pour le passé** : « végre » (enfin) « közben/időközben », « nemsokára/hamarosan », d'autres mots indiquant un moment dans le temps (par exemple, année, mois, jour, partie de la journée) ou la distance temporelle de l'événement **Modificateurs fréquents pour le présent** : « hátha/talán/majdcsak » (peut-être) et d'autres mots indiquant l'attitude du locuteur face à l'événement ;

	« nemsokára/hamarosan » (d'ici peu) et d'autres mots signalant la distance temporelle de l'événement
Colligations typiques (1)	(1) Verbe au passé (souvent à la 3e personne) : **(MOD « végre », « estére », « pénteken », « két hét múlva » +) Megjött + X** (2) Verbe au présent (souvent à la 3e personne) : **MOD (« hátha/talán », « lassan », etc.) + megjön + X**
Colligations typiques (2) : ordre des mots	(1) Passé : le verbe reçoit de l'emphase et précède en général le sujet. **(MOD +) megjött + X.** Vagy : **ComplT89 + jött meg/megjött + X90** (2) Présent : le verbe est précédé d'un modificateur (complément du temps). Le verbe reçoit de l'emphase et précède en général le sujet. **MOD + megjön + X**
Composantes sémantiques	Indique l'arrivée d'un moment singulier, décisif. Ce moment est en général perçu par le locuteur comme positif. Passé : peut indiquer un soulagement. Présent : peut indiquer l'espoir ou l'anticipation.
Composantes pragmatiques	Pas de composantes pragmatiques particulières.

2) Une personne est de retour

Collocations typiques	**Sujets** : personne ou groupe de personnes **Modificateurs** : « végre » (enfin), « amióta » (depuis), « amikor » (quand) et d'autres mots indiquant un moment dans le temps
Colligations typiques (1)	(1) Verbe au passé : **(MOD « végre », « amikor » +) megjött + X** (2) Verbe au présent : **MOD (« hatra », « amikor » etc.) + megjön + X**
Colligations typiques (2) : ordre des mots	(1) Passé : le verbe reçoit de l'emphase et précède en général le sujet : **(MOD +) Megjött** + X vagy : **ComplT + megjött + X**

[89] Complément de temps.
[90] Le complément de temps est l'information la plus importante de la phrase.

	(2) Présent : le verbe est précédé d'un modificateur (complément de temps). Le verbe reçoit de l'emphase et précède en général le sujet : **MOD + megjön + X**
Composantes sémantiques	Indique l'arrivée d'une personne, perçue par le locuteur comme un moment positif.
Composantes pragmatiques	1e et 2e personne : annonce l'arrivée du locuteur dans une interaction Sinon pas de composantes pragmatiques particulières

3) Retour d'un moment récurrent ou d'un phénomène météorologique

Collocations typiques	**Sujets** : et d'autres mots indiquant des moments récurrents dans le temps ou des phénomènes météorologiques (également récurrents) **Modificateurs** : « végre » (enfin), « amióta » (depuis), « amikor » (quand) et d'autres mots indiquant un moment dans le temps
Colligations typiques (1)	(1) Verbe au passé : **(MOD « végre », « amikor » +) megjött + X** (2) Verbe au présent : **MOD (« amikor » etc.) + megjön + X**
Colligations typiques (2) : ordre des mots	(1) Passé : le verbe reçoit de l'emphase et précède en général le sujet : **(MOD +) megjött + X** (2) Présent : le verbe est précédé d'un modificateur (complément de temps). Le verbe reçoit de l'emphase et précède en général le sujet : **MOD + megjön + X**
Composantes sémantiques	Indique le retour d'un phénomène météorologique (pluie, neige, froid) ou d'une période (printemps, avril). Les modificateurs peuvent indiquer l'attitude du locuteur envers l'événement (le plus souvent son soulagement).
Composantes pragmatiques	Le retour de l'événement est en général perçu par le locuteur de manière positive. La phrase peut introduire un récit relatif à ce moment.

4) Usage idiomatique, usage limité à certains éléments lexicaux

Collocations typiques	**Sujets** : « kedv » (envie), « étvágy » (appétit), « ész » (sens, raison), « hang » (voix), « bátorság » (courage) et quelques autres noms indiquant une qualité, sensation ou émotion. **Modificateurs** : passé : « végre » (enfin), « hogy » (que) ; présent : « remélem » (j'espère), « talán » (peut-être), « egyszer majd » (un jour), « majdcsak » (un jour) et d'autres mots exprimant l'attitude du locuteur face à l'événement.
Colligations typiques (1)	Verbe à la 3e personne, au présent ou au passé, sujet avec terminaison possessive.
Colligations typiques (2) : ordre des mots	(1) Passé : le verbe reçoit de l'emphase et précède en général le sujet. **MOD + megjött + XPoss** « Végre megjött az eszed. » (Tu redeviens enfin raisonnable.), « Megjött az étvágyam. » (J'ai envie de manger.) (2) Présent : le verbe reçoit de l'emphase et précède en général le sujet. **MOD + megjön + X**
Composantes sémantiques	(1) Quelqu'un retrouve une envie en soi qu'il juge positive. (2) Quelqu'un retrouve ou découvre une qualité en soi qu'il juge positive.
Composantes pragmatiques	Les phrases peuvent servir à exprimer un compliment ou un soulagement de la part du locuteur.

Tableau 134 : Profil du verbe « megjön ».

Ces tableaux montrent, une fois de plus, la richesse d'informations que nous pouvons déceler en utilisant de grandes bases de données pour la description de l'élément linguistique choisi. Ainsi, nous avons pu observer, par exemple, que les modificateurs fréquents fournissent non seulement des précisions par rapport à l'événement donné mais peuvent également signaler l'attitude du locuteur envers cet événement. Une telle information est absente des dictionnaires standards ; elle permet de fournir des indications essentielles pour préciser et différencier, par exemple dans le cas de « megjön » et « eljön », le sens de plusieurs synonymes.

K) Quelle est la différence ? « Eljön » et « megjön »

Le tableau 135 résume les résultats obtenus par l'analyse de corpus (catégories ordonnées par fréquence) :

eljön	**megjön**
1 « eljön » + moment dans le temps ⇓ Un moment décisif, de changement arrive	**1 (végre) megjött** **(hátha) megjön** ⇓ (1) Quelque chose ou quelqu'un que le locuteur a attendu, arrive (2) Espoir du locuteur que cette chose ou une personne arrive
2 « eljön/el tud jönni » + LOC (directionnel) ⇓ Participe à un événement, invite quelqu'un à la participation ou réagit à une invitation (langage interactionnel)	**2 megjön** ⇓ Quelqu'un est de retour, chez soi
3 « eljön » + LOC (d'un endroit) ⇓ Part d'un endroit, quitte un lieu provisoirement ou définitivement	**3 « megjön » + moment recurrent dans le temps, phénomène météorologique** ⇓ Un phénomène météorologique arrive ou est de retour
4 « eljön » + Xért ⇓ Vient chercher quelqu'un ou quelque chose	**4 « megjön » + qualité** ⇓ Quelqu'un retrouve une qualité ou une envie en soi

Tableau 135 : L'usage des verbes « eljön » et « megjön » comme observé dans les corpus écrits et oraux.

L'analyse de l'environnement textuel des synonymes « eljön » et « megjön » permet d'établir deux profils clairement différents pour ces deux verbes dont des traductions sont quasi identiques. Cette analyse révèle une forte interconnexion entre les caractéristiques grammaticales, lexicales et

sémantiques des éléments étudiés[91] qui contribuent, dans leur ensemble, à fournir aux mots des caractéristiques singulières et à les rendre non ambiguës. Il est, en effet, possible de déceler les collocatifs fréquents et leurs particularités sémantiques et pragmatiques, ainsi que des renseignements relatifs aux structures grammaticales utilisées avec un sens donné du verbe.

Dans une optique plus large, nous pouvons également constater que l'analyse des corpus oraux peut enrichir nos observations en révélant des usages typiques dans les interactions, d'où l'intérêt de leur intégration dans cette étude, malgré leur faible taille.

Ce chapitre a exploré l'usage de « megjön » et « eljön », deux verbes qui sont non seulement des synonymes mais aussi des mots à structure similaire, comportant le même verbe (« jön », venir) comme radical, précédé d'un préfixe. Comme dans le cas de « tűnik » et « látszik », nous avons pu observer l'usage dans des environnements textuels différents rendant ces mots interchangeables à première vue, mais distincts à l'usage. L'analyse de « megjön » et « eljön » dans ce chapitre et celle de « tűnik » et « látszik » au chapitre précédent montrent clairement l'intérêt et le potentiel de l'utilisation des corpus pour la description de l'usage des synonymes dans le cadre pédagogique. La manière de présenter les informations obtenues lors de l'analyse ainsi que d'autres considérations relatives à cette question seront exposées en détail au chapitre 12.

[91] Par ailleurs, l'interconnexion entre toutes ces dimensions semble être elle-même la source de la confusion, car certains éléments lexicaux fréquents peuvent être utilisés avec chacun des deux verbes.

Chapitre 11 : Les deux conjugaisons : le cas du verbe « ad » (donner)

Après avoir analysé deux questions lexicales (le sens du mot « nehéz », les synonymes « tűnik » et « látszik ») et une question à la frontière du lexique et de la grammaire (les synonymes à préfixe « megjön » et « eljön »), ce chapitre mettra en exergue un phénomène qui a été jusqu'à présent traité comme une question purement grammaticale. Il s'agit d'une particularité de la langue hongroise : les deux conjugaisons. Nous examinerons si une analyse à l'aide des outils numériques peut contribuer à une meilleure compréhension de cet aspect de la langue, susceptible de poser des difficultés à l'apprenant.

Après la présentation de l'approche des grammaires pédagogiques (section A), nous étudierons le verbe « ad » (donner) et ses collocatifs typiques dans les deux conjugaisons (section B), en nous concentrant sur la troisième personne (section C) et la première personne (section D) du singulier de l'indicatif.

Rappelons qu'il s'agit ici d'une première étude dont le but est de démontrer l'intérêt de ce genre d'explorations et de proposer un procédé qui permet l'examen systématique de ce phénomène. Pour valider ou réfuter certaines propositions formulées au sein de ce chapitre, des études à plus grande échelle seraient nécessaires.

A) Que disent les grammaires pédagogiques ?

Comme évoqué plus haut, une particularité de la langue hongroise est qu'elle possède deux conjugaisons : la conjugaison définie et la conjugaison indéfinie. La conjugaison indéfinie s'utilise quand la phrase ne contient pas de complément d'objet direct ou quand le complément d'objet direct est indéfini. La règle semble donc simple et pourtant, « l'acquisition parfaite des deux séries de conjugaison est l'une des difficultés majeures du hongrois. » (Nyéki 1988 : 172). D'où provient cette difficulté ?

Les grammaires pédagogiques (cf. Forgács 2007 ; Hegedűs 2005 ; Keresztes (1995) ; Rounds 2008, Szende et Kassai 2001 ; Szita et Görbe 2009), listent les types de compléments d'objet direct et la conjugaison qui les accompagne. Selon ces ouvrages, les CODs indéfinis sont les CODs précédés d'un article indéfini, d'un nombre ou d'une quantité ou des mots « ilyen » (comme ceci) ou « olyan » (comme cela), les mots interrogatifs « kit » (qui), « mit » (quoi), « mennyit » (quelle quantité) et

« hányat » (combien) ainsi que les pronoms finissant par ces mots. La conjugaison est définie quand le complément d'objet direct est défini. À ce groupe appartiennent les CODs précédés d'un article défini (dont un sous- groupe sont les noms avec une terminaison du possessif), les CODs précédés du pronom demonstratif « ez » (celui-ci, celle-ci) ou « az » (celui-là, celle-là), les noms propres et les mots interrogatifs « milyet » (quel type, quel genre), « melyiket » (lequel, laquelle), les pronoms finissant par ces mots ainsi que les pronoms « egymást » (l'un l'autre) et « magamat » (moi-même).

La grammaire fonctionnelle de Hegedűs (2005 : 39-56) présente des tableaux de conjugaisons dans les différents temps et personnes, mais elle ne fournit aucune explication quant à l'existence des deux conjugaisons. Cet aspect de la langue est associé à une seule reprise à des fonctions (2005 : 279) où l'auteure indique que les deux conjugaisons jouent un rôle dans la réalisation de la cohérence textuelle avec les pronoms personnels. Dans les autres chapitres du livre, la relation entre les deux conjugaisons et leurs fonctions n'est pas traitée en détail. Pour présenter les deux conjugaisons, certaines grammaires pédagogiques optent pour le même mot et la même structure comme complément d'objet direct pour montrer la différence d'usage ; par exemple, Szita et Görbe (2009) utilisent de manière systématique le verbe « écrire » et le nom « lettre » pour présenter les différents types de CODs. Dans d'autres grammaires, les verbes diffèrent phrase par phrase (par exemple, Szende et Kassai 2001) et certaines grammaires se restreignent essentiellement à des tableaux de conjugaisons complétés par quelques phrases-exemples (par exemple, Keresztes 1995).

La tradition de la présentation des deux conjugaisons comme phénomène purement grammatical, est-elle justifiée à la lumière des études fondées sur le corpus ? Quelles précisions les études empiriques peuvent-elles apporter en vue d'une présentation plus claire et plus exacte ? Dans les pages suivantes, nous chercherons à répondre à ces questions. Ainsi, nous tâcherons de prouver qu'une description plus adéquate des deux conjugaisons implique bien plus d'aspects langagiers que le seul type du complément d'objet direct et le type de la conjugaison. Nous prendrons à cette fin comme exemple le verbe « ad » (donner) et nous analyserons les points suivants :

- Nous établirons d'abord la liste des CODs les plus fréquents sans faire la différence entre les occurrences avec l'une et l'autre conjugaisons. Cette étape nous permettra de nous forger une idée générale de l'usage du verbe en étudiant la sémantique des noms qui s'y associent habituellement.
- Nous identifierons par la suite les CODs les plus fréquents par conjugaison et par personne grammaticale afin de cartographier leurs similitudes et différences.

- Nous examinerons un grand nombre d'exemples avec les deux conjugaisons pour observer leurs environnements textuels plus larges. Dans cette phase, nous étudierons également les collocatifs s'associant aux deux conjugaisons à une fréquence comparable pour voir s'ils sont utilisés de la même façon.
- Nous fournirons enfin un aperçu des schémas d'usage.

Comme évoqué précédemment, en raison de la nature exploratoire de cette approche, notre seul but est de prouver l'intérêt de ce type d'étude. Nos résultats seront, par conséquent, illustratifs, non exhaustifs, mais nous espérons démontrer par ce biais l'intérêt d'une analyse plus complète et étendue des deux conjugaisons à partir d'une étude quantitative du corpus.

B) Le verbe « ad » (donner) et ses collocatifs les plus fréquents

Le tableau 136 (ci-dessous) montre les collocatifs les plus usités du verbe « ad » (donner) dans le corpus « huTenTen12 » (les deux conjugaisons confondues)[92]. Comme expliqué au chapitre 3, la simple fréquence n'est qu'un des indicateurs utilisés pour signaler la force d'association entre les membres d'une unité multi-lexicale, car le nombre d'occurrences d'une unité multi-lexicale est toujours influencé par la fréquence de ses constituants dans le corpus. Ainsi, des mots « travail » ou « image » sont des mots très présents dans le corpus, aptes à former des unités multi-lexicales avec un grand nombre de verbes[93]. La relation entre « donner + travail » et « donner + image » est donc plus faible qu'entre, par exemple, « donner + chance » ou « donner + renseignements » parce que ces derniers noms forment avant tout des unités multi-lexicales avec le verbe « donner », la relation entre les éléments est donc plus forte. Ceci explique pourquoi certaines unités multi-lexicales se retrouvent plus bas dans la liste alors qu'elles devraient apparaître plus haut si nous avions considéré seulement leur fréquence absolue (voir aussi le chapitre 3 pour la description des mesures statistiques). Le tableau 136 ci-dessous montre les 20 collocatifs les plus fréquents du lemme « ad » dans le corpus écrit de « huTenTen12 » :

lehetőséget (58 243)	*possibilité, opportunité*
tanácsot (27 497)	*conseil*

[92] Le Corpus national du hongrois fournit des résultats similaires, mais étant donné que ce corpus renferme un grand nombre de textes de journaux, certaines unités multi-lexicales clairement associables à ce genre (par exemple, « donner une interview ») sont plus fréquentes.

[93] « Chercher, trouver, effectuer » sont d'autres verbes fréquemment utilisés avec le nom « travail », et « afficher, renvoyer, projeter » avec le nom « image ».

választ (28 025)	*réponse*
okot (18 063)	*raison*
hangot valaminek (17 635)	*voix (s'exprimer)*
erőt (17 219)	*force*
hírt (15 554)	*nouvelles*
pénzt (17 567)	*argent*
otthont (14 741)	*maison, i.e. « héberger » (événement, exposition)*
esélyt (13 955)	*chance*
képet (15 954)	*image*
tájékoztatást (12 685)	*renseignements*
hálát (12 441)	*merci (dire merci, remercier)*
életet (15 606)	*vie*
számot (11 900)	*compte (tenir compte)*
koncertet (10 921)	*concert*
magyarázatot (10 726)	*explication*
címet (11 230)	*titre*
munkát (12 843)	*travail*

Tableau 136 : Les 20 collocatifs les plus fréquents du lemme « ad » dans le corpus écrit de « huTenTen12 ».

Cette table offre un premier aperçu concernant les unités multi-lexicales les plus usitées avec « ad » dans les deux conjugaisons. Bien que cette présentation ne sépare pas ces unités par conjugaison, elle montre clairement que les collocatifs les plus nombreux du verbe sont des noms abstraits comme « opportunité », « conseil », « réponse » et « raison ».

Pour obtenir des renseignements plus précis, nous avons analysé des occurrences par type de conjugaison et par personne. *Nous avons restreint l'analyse détaillée à l'indicatif sans verbe modal, identifié comme l'usage le plus commun dans notre corpus, nous permettant de récupérer un très grand nombre d'exemples. Nous avons par la suite analysé un échantillon de 1000 phrases avec chaque unité multi-lexicale fréquente (voir les tableaux plus bas) pour trouver des réponses aux questions suivantes* :

- Peut-on identifier des CODs typiques avec chacune des conjugaisons ?
- Existe-t-il des différences notables de fréquence entre la même unité multi-lexicale avec la conjugaison indéfinie et la conjugaison définie ?

- Existe-t-il des différences notables d'usage et de fréquence entre la même unité multi-lexicale de la même conjugaison à la première et à la troisième personnes ?
- Peut-on observer des schémas généraux d'usage qui vont au-delà d'une seule unité multi-lexicale ?

Pour identifier les collocatifs typiques par conjugaison, nous avons utilisé l'outil Word Sketch Difference. Nous avons cherché les formes « ad » (il/elle donne, conjugaison indéfinie) et « adja » (il/elle donne, conjugaison définie). Le premier chiffre à côté de chaque nom (colonne de gauche) indique le nombre d'exemples avec « ad » et le second (colonne de droite) le nombre d'exemples avec « adja » (tableau 137).

otthon	11,092	31	•••
felvilágosítás	1,188	10	•••
ok	8,010	107	•••
tér	3,040	54	•••
lehetőség	28,927	676	•••
önbizalom	1,242	15	•••
útmutatás	1,278	22	•••
hely	4,950	257	•••
tájékoztatás	3,548	98	•••
alkalom	4,620	208	•••
munka	4,383	293	•••
koncert	3,881	166	•••
esély	3,133	146	•••
lendület	1,550	59	•••
segítség	2,612	200	•••
kérdés	2,621	198	•••
kép	7,436	573	•••
tanács	5,162	305	•••
információ	2,780	188	•••
ember	3,571	382	•••
erő	8,494	775	•••
biztonság	2,536	187	•••
támogatás	1,868	208	•••
áttekintés	1,772	122	•••
hang	4,940	612	•••
engedély	1,465	147	•••
mód	2,647	499	•••
szabadság	1,474	169	•••

energia	1,870	270	•••
remény	2,411	329	•••
magyarázat	2,960	419	•••
eredmény	2,165	512	•••
pénz	2,888	1,005	•••
élet	3,610	1,531	•••
íz	1,286	285	•••
keret	1,627	574	•••
utasítás	1,153	289	•••
alap	2,783	1,501	•••
élmény	1,278	427	•••
hír	1,633	604	•••
forma	1,065	685	•••
érték	1,334	732	•••
parancs	786	317	•••
hangulat	698	356	•••
kéz	2,765	2,388	•••
ajándék	866	469	•••
érzés	660	431	•••
leírás	764	422	•••
cím	635	635	•••
jel	816	745	•••
kulcs	310	246	•••
ár	327	540	•••
keresztmetszet	153	209	•••
alak	130	442	•••
név	249	1,468	•••

bér	343	1,090	•••
élő	76	269	•••
küldetés	123	397	•••
százalék	253	1,647	•••
áldás	206	960	•••
hozzájárulás	121	611	•••
lényeg	43	327	•••
gerinc	27	325	•••
értés	25	403	•••
fél	16	387	•••
fej	55	2,370	•••
tudta	71	1,688	•••
voks	0	224	•••
beleegyezés	0	410	•••

(maison (héberger)	*énergie*	*location (non COD)*
renseignements	*espoir*	*live, en direct (non COD)*
raison	*explication*	*mission*
espace	*résultat*	*pourcentage*
possibilité, opportunité	*argent*	*bénédiction*
confiance en soi	*vie*	*consentement*
conseils	*goût*	*essence*
place	*cadre*	*colonne vertébrale*
informations	*commande*	*moitié, demie*
occasion	*base*	*tête (expr. id. : commencer qqch)*
travail	*expérience*	*savoir (faire savoir qqch à qqn)*
concert	*nouvelles*	*vote*
chance	*forme*	*accord)*
élan	*valeur*	
aide	*ordre*	
question (non COD)	*ambiance*	
image	*main*	
conseil	*cadeau*	
information	*sentiment*	
homme (non COD)	*description*	
force	*adresse, titre*	
sécurité	*signe, signal*	
soutien	*clé*	
aperçu	*prix*	
voix (exprimer)	*coupe transversale*	
autorisation	*forme (non COD)*	
occasion (permettre qqch)	*nom*	
liberté		

Tableau 137 : Les collocatifs les plus fréquents de « ad » et « adja » dans « Word Sketch Difference ».

Les valeurs mesurées montrent sans ambiguïté que, *même si les noms listés pouvaient théoriquement s'associer aux deux conjugaisons (il n'y a aucune contrainte grammaticale qui l'empêcherait), ils ont tendance à émerger bien plus souvent avec l'une qu'avec l'autre*. En tête de la liste, nous trouvons des noms qui accompagnent de préférence la conjugaison indéfinie : « otthon » (maison/venue), « tájékoztatás » (renseignement), « ok » (raison), « tér » (espace), « lehetőség » (possibilité/opportunité). En

comparant cette liste avec la liste présentant les collocatifs des deux conjugaisons (tableau 136), nous pouvons observer qu'un certain nombre de mots se trouve en tête des deux listes, ce qui indique la forte présence des collocations fréquentes avec la conjugaison indéfinie. Les noms présentés au milieu de la liste dans le tableau 137 sont associés à une fréquence comparable aux deux conjugaisons. Leur nombre est limité : nous n'en observons qu'une dizaine au total, il n'existe donc que peu de collocatifs fréquemment utilisés avec les deux conjugaisons. La troisième partie de la liste présente les noms de préférence utilisés avec la conjugaison définie.

Dans ce qui suit, nous étudierons d'abord les collocatifs avec la conjugaison indéfinie, puis ceux avec la conjugaison définie. Deux unités multi-lexicales et un groupe d'unités suivant le même schéma seront ensuite examinés plus précisément : « lehetőség + ad/adja » (« possibilité + il/elle donne », conjugaison indéfinie), « jel + ad/adja » (exemple d'un collocatif à fréquence comparable par conjugaison) et plusieurs unités multi-lexicales de la fin de la liste avec lesquelles la conjugaison définie est privilégiée.

C) Usage à la troisième personne du singulier

1) Un aperçu des collocatifs de « ad » (il/elle donne, conjugaison indéfinie) et de « adja » (il/elle donne, conjugaison définie)

Dans un premier temps, nous nous concentrerons sur les occurrences à la troisième personne du singulier : Le corpus contient 467 905 énoncés au total avec la forme indéfinie (« ad ») (ou 148 par million, nombre significatif) dont les collocatifs les plus usités ont été listés dans le tableau 137.

En étudiant la liste des collocatifs, il devient évident que la majorité de ces unités multi-lexicales est idiomatique : le verbe « ad » est plus ou moins délexicalisé (son sens est essentiellement déterminé par son collocatif), ce que montre aussi la multitude de verbes dans leurs traductions : « *offrir* une possibilité », « *héberger* un événement », « *donner* un conseil », « *fournir* une raison », « *exprimer* une opinion ou un sentiment », « *donner* de la force », « *donner* lieu ». La signification concrète du mot, notamment « tendre un objet réel à quelqu'un, mettre un objet en la possession d'une personne » (Larousse online) n'apparaît dans aucune des expressions.

Que les occurrences faisant référence à des objets réels soient moins fréquentes peut s'expliquer, du moins en partie, par la nature du corpus étudié. Puisqu'il s'agit de textes écrits (récits et narratifs), nous ne trouverons pas en effet dans ce corpus d'interactions relatives à des situations particulières du quotidien. Par exemple, des phrases fréquentes dans les interactions entre

fournisseurs de services et clients comme « Adom a blokkot. » (« Je vous donne le ticket. ») n'apparaîtront guère dans notre corpus écrit. Pour se forger une image plus exacte des unités multi-lexicales typiques, il est donc conseillé de consulter aussi des corpus oraux, comme nous le verrons par la suite.

Nous trouvons 253 837 occurrences avec la forme définie (« adja ») dont plus de la moitié sont des verbes avec un préfixe. Après l'élimination manuelle de ces derniers, seules 101 534 occurrences restent. En comparant l'usage des deux conjugaisons, nous pouvons donc constater un écart important de fréquences : à la troisième personne du singulier du présent, la conjugaison indéfinie (« ad ») est 4,5 fois plus usitée que la conjugaison définie, dans le corpus étudié. Même si les CODs les plus courants ont des fréquences nettement inférieures à ceux avec la conjugaison indéfinie, leur étude fait émerger plusieurs caractéristiques communes. Nous voyons ainsi se détacher un schéma grammatical : l*es CODs de ces unités sont des noms avec une terminaison possessive*[94] *(« X donne le N de Y »).* Alors que l'usage sans cette terminaison serait théoriquement envisageable, 78% des occurrences avec les 100 noms les plus fréquents dans le corpus contiennent une structure possessive. 22% contiennent un adjectif, essentiellement au superlatif qui servent à qualifier le nom.

Nous pouvons également constater que *les CODs sont majoritairement des noms abstraits*. Ils peuvent être regroupés selon leurs sens positifs, négatifs ou neutres, comme le montre le tableau 138.

COD (NPoss) + « adja » (conjugaison définie)

(1) Sens positif : particularité, importance, valeur etc. (67%)	***(1) Sens positif : particularité, importance, valeur ...* (67%)**
(Avec terminaison possessive)	**(Avec terminaison possessive)**
X (az együttműködés) jelentőségét az adja, hogy ...	*L'importance de X (coopération) tient au fait que ...*
X vonzerejét éppen az adja, hogy ...	*L'attrait de X est précisément que ...*
X egyedi jellegét az adja, hogy ...	*Le caractère unique de X est que ...*
A dolog pikantériáját az adja, hogy ...	*Le piquant de la chose est que ...*
Cégünk erejét az adja, hogy ...	*La force de notre entreprise est que ...*
A projekt indokoltságát az adja, hogy ...	*La justification du projet est donnée par le fait que ...*
X egyedi ízét az adja, hogy ...	*Le goût unique de X est donné par le fait que ...*

[94] Dans la langue hongroise la terminaison du possessif est rajoutée à la possession, c'est-à-dire au COD dans notre cas.

X ételeinek sajátosságát az adja, hogy ...	*La particularité des plats de X est que ...*
X különlegességét az adja, hogy ...	*La particularité de X est que ...*
Létünk értelmét az adja, hogy ...	*Le sens de notre existence est que ...*
X misztériumát az adja, hogy ...	*Le mystère de X est donné par le fait que ...*
X (a hangszer) történeti értékét az adja, hogy ...	*La valeur historique de X (l'instrument) est donnée par le fait que ...*
A feladat szépségét az adja, hogy ...	*La beauté de la tâche est que ...*
X hitelességét pedig az adja, hogy ...	*Et la crédibilité de X est donnée par le fait que ...*
X sikerét az adja, hogy ...	*Le succès de X est dû à ...*
X (a gimnázium) rangját az adja, hogy ...	*Le rang de X (le lycée) est donné par le fait que ...*
X (a tánc) varázsát az adja, hogy ...	*Le charme de X (la danse) est donné par le fait que ...*
(Sans terminaison du possessif)	***(Sans terminaison du possessif)***
A legnagyobb biztonságot az adja, hogy ...	*La plus grande sécurité est assurée par ...*
X-hez a legjobb alapot az adja, hogy ...	*La meilleure base pour X est de ...*
A kellemes érzést az adja, hogy ...	*La sensation agréable est donnée par le fait que ...*
(2) Sens négatif (4%)	***Sens négatif* (4%)**
X társadalmi veszélyét elsősorban az adja, hogy ...	*Le danger social de X est principalement dû au fait que ...*
X (a dolog) nehézségét elsősorban az adja, hogy ...	*La difficulté de X (la chose) est principalement due au fait que ...*
Az ügy hátterét az adja, hogy ...	*Le contexte de l'affaire est que ...*
A helyzet tragikumát az adja, hogy ...	*La tragédie de la situation est que ...*
(3) Sens neutre (29%)	***Sens neutre***
X és Y különbségét az adja, hogy ...	*La différence entre X et Y est que ...*
X kiindulópontját az adja, hogy ...	*Le point de départ pour X est ...*
X (cikkünk) aktualitását pedig az adja, hogy ...	*L'actualité de X (notre article) est donnée par le fait que ...*
X (a kiállítás) apropóját az adja, hogy ...	*Le propos de X (l'exposition) est le fait que ...*
A szerep súlyát az adja, hogy ...	*Le poids du rôle est donné par le fait que ...*

Tableau 138 : CODs typiques avec la conjugaison définie.

Contrairement à la conjugaison indéfinie, il est difficile d'identifier des noms particuliers typiquement associés à la conjugaison définie parce que le nombre des occurrences avec chaque nom est faible. Néanmoins, un certain nombre de collocatifs remplissant la fonction de CODs avec « adja » manifestent certaines similarités sémantiques : *dans 67 % des occurrences, le COD désigne une qualité positive du possesseur* comme « la beauté de X », « le sens de X », « la valeur de X », « la particularité de X », « l'importance de X », « la force de X », « l'attraction de X », « l'authenticité de X » ou « la magie de X ». 29 % des CODs sont neutres comme « le propos de X », « la raison pour X », « l'actualité de X ». Seuls 2% des CODs évoquent la source de quelque chose de négatif comme « le danger de X », « la difficulté de X ». La conjugaison définie est donc elle aussi susceptible d'attirer certains éléments linguistiques.

À l'aide des lignes de concordance, nous avons étudié 1000 phrases par type de conjugaison dans le but de catégoriser les compléments d'objet direct selon leurs propriétés sémantiques. L'analyse a fait émerger le constat que *les deux conjugaisons attirent non seulement des noms individuels différents mais aussi des champs lexico-sémantiques différents, associables aux conjugaisons*. Le tableau suivant résume les résultats de cette analyse (tableau 139).

Conjugaison indéfinie **Usages fréquents avec « ad »**	**Conjugaison définie** **Usages fréquents avec « adja »**
Caractéristiques sémantiques	**Caractéristiques sémantiques**
(En majorité des noms abstraits)	*(En majorité des noms abstraits)*
1 Opportunité, chance, possibilité	1 Réaction, réponse, aide, soutien
2 Renseignements, informations, explication, réponse, description, aperçu, image	2 Actualité, propos, occasion d'un événement
3 Force, énergie, sécurité, espoir	3 Accord, consentement, bénédiction
4 Nouvelles, signe	4 Particularité, singularité d'un événement/d'une chose
5 Venue, cadre, espace, forum	5 La base, le point de départ de X
Caractéristiques grammaticales (colligation)	**Caractéristiques grammaticales (colligation)**
1 En général, le COD n'a pas d'article : COD (NullArt) + ad	1 En général, le COD a une terminaison possessive : COD (NPoss) + S + adja
2 Le nom a tendance de précéder le verbe.	2 Unité multi-lexicale typique, plus longue : (NPoss) az adja, hogy …
	3 Le nom a tendance à précéder le verbe.

Tableau 139 : Catégorisation possible des collocatifs selon leurs caractéristiques sémantiques et grammaticales.

Le fait que ces groupes se distinguent aussi nettement, témoigne déjà d'une interconnexion du lexique et de la grammaire que nous exploiterons par la suite.

2) Étude de cas (1) : le nom « lehetőség » (possibilité, opportunité) comme COD

2.1) Conjugaison indéfinie : « lehetőséget + ad » et « ad* + lehetőséget »*

Dans cette section, nous analyserons une unité multi-lexicale typique avec la conjugaison indéfinie : « lehetőséget + ad » (possibilité, opportunité + donner). Cette unité figure en tête de la liste des collocatifs avec les deux conjugaisons et elle est également l'un des collocatifs les plus usités de « ad » à la conjugaison indéfinie.

Nous avons tout d'abord déterminé avec l'outil « Complex Query Language » de Sketch Engine la distribution exacte des occurrences selon l'ordre des mots. Nous avons défini la distance maximale entre « ad » et « lehetőséget » à trois mots et manuellement éliminé les occurrences avec un préfixe puisque les préfixes sont susceptibles de changer le sens du verbe et de s'associer ainsi à d'autres éléments lexicaux que le verbe sans préfixe. Le tableau indique les deux commandes utilisées, le nombre des occurrences ainsi que la distribution des résultats selon l'ordre du COD et du verbe (tableau 140) :

[word="lehetőséget"] [] {0,3} [word="ad"]	21 664 occurrences	**65,45 % (COD + V)**
[word="ad"] [] {0,3} [word="lehetőséget"]	11 511 occurrences	**34,55 % (V + COD)**
AU TOTAL	33 175 occurrences	**100 %**

Tableau 140 : Nombres d'occurrences avec « ad » et « lehetőséget » dans Complex Query Language (CQL).

Les unités multi-lexicales à trois et quatre composantes sont : « X YrA + ad lehetőséget » (X offre la possibilité de Y), « X arra ad lehetőséget, hogy » (X offre la possibilité de INF, l'emphase est sur « arra », rendant la proposition qui suit, particulièrement importante), « X lehetőséget ad arra, hogy » (X offre la possibilité de INF, l'emphase est sur le mot « lehetőséget », l'information la plus importante est que la proposition qui suit, *est réalisable*). Il convient de remarquer que, contrairement à la logique hongroise qui requiert un COD indéfini, sans article, la traduction française contient un COD défini (« la possibilité », COD précédé d'un article défini). Si la première langue de l'apprenant est le français, il peut donc arriver à la conclusion (incorrecte) qu'il doit utiliser la conjugaison définie dans cette phrase. Puisque les différentes langues montrent de

grandes différences concernant l'usage (ou l'absence) des articles[95], identifier des expressions fréquemment utilisées avec un type d'article ou sans article peut être particulièrement utile dans le cadre pédagogique (tableau 141).

Szerződésünk **lehetőséget ad** egy újabb – közösen egyeztetett – időpont kérésére.	*Notre contrat* ***vous donne la possibilité*** *de demander un autre rendez-vous, d'un commun accord.*
A játék **lehetőséget ad** a megmérettetésre, szórakozásra és aktív kikapcsolódásra.	*Le jeu* ***vous donne la possibilité*** *de participer à des compétitions, de vous amuser et de vous détendre activement.*
Y kifinomult, profi munkaeszköz, ami **lehetőséget ad** rendkívül precíz, jól definiált beállításokra is.	*Y est un outil de travail sophistiqué et professionnel* ***qui permet*** *également des réglages extrêmement précis et bien définis.*
A többszintű karrier **lehetőséget ad** arra, hogy – ha úgy szeretné - csupán napi néhány órában tevékenykedhessen.	*La carrière à niveaux multiples* ***vous donne la possibilité*** *de ne travailler que quelques heures par jour, si vous le souhaitez.*
Termünk **lehetőséget ad** a családi és társasági összejövetelekre.	*Notre salle* ***offre la possibilité*** *de tenir des réunions familiales et sociales.*

Tableau 141 : Exemples avec « lehetőséget » et « ad » dans le corpus « huTenTen12 ».

Les exemples du tableau 141 montrent que le complément de l'expression prend la terminaison « -ra/-re » (l'équivalent des « de » soulignés dans les phrases françaises), information que l'apprenant peut retenir par l'observation. Il peut également noter que dans le cas de « lehetőséget », seulement 4% des CODs (1327 exemples) sont précédés d'un adjectif ou d'un numéral indéfini. Ces modificateurs exprimant des concepts plutôt généraux se regroupent en trois catégories :

(1) Modificateurs quantifiant les opportunités (le plus souvent de façon positive) : « rengeteg » (beaucoup), « számos » (nombreux), « korlátlan » (illimité).
(2) Modificateurs précisant la nature de l'opportunité : « technikai » (technique), « választási » (de choix), « variációs » (de variation)
(3) Modificateurs donnant un jugement de valeur (en général positif) : « jó » (bon), « kiváló » (excellent), « újabb » (nouveau).

[95] Cela peut, en partie, expliquer la difficulté que les apprenants ont à identifier la conjugaison correcte dans les phrases concrètes.

2.2) Conjugaison définie : a (ADJ) lehetőséget + adja, X lehetőségét + adja

Les deux formes du COD défini couvrant 99% des exemples dans le corpus sont (1) le COD avec un article défini (précédé ou non de modificateurs) et (2) le COD avec une terminaison possessive[96]. Nous analysons d'abord les occurrences avec l'article défini.

Même si cette forme est acceptable en théorie, la présence de « lehetőséget » comme COD défini avec article défini, sans terminaison possessive, n'apparaît que dans seulement 676 énoncés (versus 33 175 énoncés comme COD indéfini). De ces 676 énoncés, 183 sont à la troisième personne du singulier de l'indicatif. Après avoir éliminé les 80 occurrences avec le préfixe « meg- »[97], l'échantillon final est constitué de 103 occurrences.

Comme dans le cas du COD indéfini, nous avons utilisé l'outil « Complex Query Language » pour déterminer la distribution des deux ordres de mots possible « a lehetőséget adja » et « adja a lehetőséget ». En définissant une distance maximale de 3 mots entre l'article défini et le nom, nous avons laissé ouverte la possibilité d'inclure un ou plusieurs modificateurs entre l'article et le nom. Selon l'ordre du verbe et du COD, la distribution se présente comme suit (tableau 142) :

[word="a|az"] [] {0,3} [word="lehetőséget"] [] {0,3} [word="adja"]	78 occurrences	**77 % (COD + V)**
[word="adja"] [] {0,3} [word="a|az"] [] {0,3} [word="lehetőséget"]	25 occurrences	**23 % (V + COD)**
AU TOTAL	103 occurrences	**100 %**

Tableau 142 : Nombres d'occurrences avec « adja » et « lehetőséget » dans Complex Query Language (CQL).

Le tableau 143 montre quelques exemples du corpus qui révèlent quelques tendances d'usage :

A legnagyobb fejlődési **lehetőséget** a párkapcsolat **adja**.	***La plus grande opportunité de développement réside*** *dans la relation.*

[96] Dans la langue hongroise, les noms avec une terminaison possessive sont en général précédés de l'article défini. Ces occurrences ne font bien évidemment pas partie du premier groupe.
[97] Là encore comme pour le cas de la conjugaison indéfinie afin de ne conserver que les occurrences relatives au verbe sans préfixe.

Krisha azt a nagyszerű **lehetőséget adja** nekünk, hogy ...	*Krisha* ***nous donne l'opportunité mag****nifique de ...*
A másik felhasználási **lehetőséget** pedig az **adja**, hogy ...	***L'autre possibilité d'utilisation est donnée*** *par le fait que ...*
A téli napokon a sportolási **lehetőséget** a konditerem **adja**.	*Les jours d'hiver, la salle de sport* ***offre la possibilité*** *de faire du sport.*
Ezt a **lehetőséget adja** számotokra az Avatar.	*C'est* ***la possibilité*** *que vous* ***offre*** *Avatar.*
Ezt a **lehetőséget adja** a Stressz teszt.	***Cette possibilité est fournie*** *par le « Stress test ».*
A választási törvény azt a **lehetőséget adja**, hogy ...	*La loi électorale vous* ***donne la possibilité*** *de ...*
A Twitter magát az eszközt, a **lehetőséget adja** csak.	*Twitter* ***ne donne*** *que l'outil et* ***l'opportunité****.*

Tableau 143 : Exemples avec « adja » et « lehetőséget » dans « huTenTen12 ».

Nous observons ainsi que le COD n'est précédé du seul article défini que dans 13% des cas, alors même que cette structure est présentée dans toutes les grammaires citées plus haut comme le schéma prototype du COD défini. Dans 87% des cas, nous trouvons ainsi un pronom démonstratif « ezt » (cette) ou « azt » (cette ...-là) et/ou des adjectifs devant le nom, précisant la nature de la « possibilité ». Bien qu'elle n'entre pas dans la formulation classique des grammaires, cette tendance peut s'expliquer par le fait que les adjectifs ainsi que le pronom démonstratif décrivent le COD plus précisément en le rendant encore plus singulier et mieux défini[98].

Le deuxième type de COD défini est « lehetőségét » (la possibilité de X), c'est-à-dire le nom « lehetőség » avec la terminaison du possessif. Le corpus contient huit fois plus d'exemples avec ce type de COD qu'avec le COD précédé d'un article défini. 357 verbes sont à la troisième personne du singulier de l'indicatif (valeur trois fois et demie plus importante que les occurrences avec l'article défini mais toujours très en deçà de celle observée dans le cas du COD indéfini[99]), ce qui nous a permis d'étudier tous les énoncés du corpus. Le tableau 144 met en évidence des exemples d'usage.

[98] Il est généralement vrai que le locuteur fournit plus de précisions pour un COD défini que pour un COD indéfini, car un COD défini est *per definitionem* plus particulier qu'un COD indéfini.
[99] Nous verrons par la suite que le COD défini avec une terminaison possessive semble être un des CODs définis les plus typiques.

X a változtatás **lehetőségét** adja.	*X vous donne* ***la possibilité de changer***.
X az önfejlesztés **lehetőségét** adja kezünkbe.	*X nous donne* ***l'opportunité de nous améliorer***.
A honlapok szintén a kapcsolatfejlesztés **lehetőségét** adják.	*Les sites Web offrent également* ***une opportunité de développer/créer des relations***.
X a nyelvgyakorlás **lehetőségét** adja a tanulóknak.	*X donne aux étudiants* ***la possibilité de pratiquer une langue***.
X az értelmezések tág **lehetőségét** adja.	*X donne* ***un large éventail d'interprétations***.
A kifelé fordítás a babának a szabad nézelődés **lehetőségét** adja.	*Le tourner vers l'extérieur donne au bébé* ***la possibilité de regarder librement autour de lui***.
X a hatékonyabb munkavégzés **lehetőségét** adja.	*X vous donne* ***la possibilité de travailler plus efficacement***.
X az újrakezdés **lehetőségét** adja.	*X vous donne* ***la possibilité de recommencer***.
Ez a helyzet nagyon sok konfliktus kialakulásának a **lehetőségét** adja.	*Cette situation peut donner* ***lieu à de nombreux conflits***.
Y a megújulás **lehetőségét** adja.	*Y donne* ***la possibilité de renouvellement***.

Tableau 144 : Quelques exemples avec « X Y lehetőségét adja » (X donne/offre la possibilité de Y).

En comparant la structure française et la structure hongroise, nous constatons une différence importante : dans la phrase hongroise, un nom précédé d'un article défini accompagne le mot « lehetőségét » (la traduction littérale des trois premières expressions serait « la possibilité *du changement* », « la possibilité *du développement personnel* », « la possibilité/l'opportunité *du développement des relations* ») alors que le français opère avec l'infinitif du verbe (la possibilité *de changer*, la possibilité *de nous améliorer*, la possibilité/l'opportunité *de développer/créer des relations*). Là encore cette différence de forme telle que révélée par l'analyse empirique, peut être la source de confusion pour l'apprenant. L'analyse de corpus démontre ici son intérêt dans le cadre pédagogique en permettant d'indiquer d'éventuels points de vigilance à l'enseignant.

3) Études de cas (2) : CODs définis avec la conjugaison définie

La deuxième partie de notre liste de collocatifs les plus fréquents (tableau 137) a présenté des noms les plus souvent associés à la conjugaison définie. Nous ne pouvons pas identifier de noms qui seraient particulièrement usités, la variété lexicale étant plus grande et le nombre total des exemples plus faible que dans le cas de la conjugaison indéfinie. Le tableau 145 fait état de cette variété :

Sokaknak fontos, hogy szülei – főleg anyja –beleegyezését **adja** a kapcsolatához.

Pour beaucoup, il est important que leurs parents, en particulier leurs mères, donnent leur consentement à leur relation.

Ha a svédek igent mondanak s a Közösség is beleegyezését **adja**, Svédország már 1995-tól tizennegyedikként beléphet a szervezetbe.

Si les Suédois disent oui et que la Communauté donne son accord, la Suède sera le quatorzième membre de l'organisation à partir de 1995.

Aki fordításra **adja** a fejét, az azt vállalja, hogy közvetíti valaki másnak a gondolatait.

Celui qui ***s'investit*** *dans la traduction, s'engage à transmettre la pensée de quelqu'un d'autre.*

Majd ha Párizsban találkozunk, értésemre **adja**, hogy mit határozott.

Quand nous nous retrouvons à Paris, il ***me fera savoir*** *ce qu'il a décidé.*

A fickó a fejét rázza, és értésemre **adja**, hogy nem tud ilyen személyről.

Le gars secoue la tête et ***me fait comprend****re qu'il ne connaît pas une telle personne.*

X implicit módon értésünkre **adja**, hogy ő mindkét kérdésre ugyanazt válaszolná.

Il ***nous fait comprendre*** *implicitement qu'il répondrait aux deux questions de la même manière.*

4000 korszerű szélturbina **adja** az ország energiaszükségletének felét.

4 000 éoliennes modernes ***fournissent la moitié des besoins énergétiques du pays****.*

Ez a verskezdő négy sor **adja** a szimmetria egyik felét.

Ces quatre lignes commençant le verset ***donnent la moitié de la structure symétrique****.*

Megkértem, hogy **adja** nekem a vacsorája felét!

Je lui ai demandé de me ***donner la moitié de son dîner*** *!*

A tavaly megjelent New Maps Of Hell album

adja majd a műsor gerincét.

Le nouvel album de Maps Of Hell sorti l'année dernière ***donnera l'épine dorsale*** *du spectacle.*

A film egyébként egy háromperces alkotás, és Nagy László egyik költeménye **adja** a gerincét.

Par ailleurs, le film est une œuvre de trois minutes, et l'un des poèmes de László Nagy en ***constitue l'épine dorsale****.*

Az árbevétel jelentős százalékát az export **adja**.

Les exportations ***représentent un pourcentage important des ventes****.*

Az orosz cég egyébként ma a világ olajtermelésének 2 százalékát **adja**.

Par ailleurs, la société russe ***représente aujourd'hui 2% de la production mondiale de pétrole****.*

De hogy egy neves divatház sporteszközhöz **adja** a nevét, azt mi se gondoltuk volna.	*Nous n'aurions jamais cru qu'une maison de haute couture renommée* ***donne son nom*** *aux équipements sportifs.*

Tableau 145 : Exemples avec COD défini et conjugaison définie.

Comme nous l'avons déjà évoqué, la très large majorité des CODs définis ont une terminaison possessive (soulignée dans le tableau 145). Les seuls mots dans la liste avec lesquels nous trouvons des instances sans terminaison possessive sont « a gerincet » (« la colonne vertébrale », 11 sur 325 occurrences) et « a nevet » (« le nom », 6 sur 1403[100] occurrences).

Dans le cas de « név » (nom), nous pouvons observer que les deux structures grammaticales s'associent à deux sens différents. Quand il a la terminaison du possessif « a nevét » (son nom), l'unité multi-lexicale signifie « donner son nom à qqch », c'est-à-dire engager sa réputation comme garantie de qualité. Il s'agit donc d'une expression fixe. Le COD « a nevet » (le nom) sans terminaison possessive a en revanche une autre fonction : il fait référence au processus concret de donner un nom à une personne ou à un produit. Dans ce dernier cas, le COD est en général précédé d'un qualificatif : « a legszebb nevet adja » (il donne les plus beaux noms), « az egykor sikeres nevet adja az új autónak » (il donne le nom à la voiture qui avait jadis tant de succès). Ces exemples sont cependant rares et, par conséquent, non représentatifs de l'usage de cette unité multi-lexicale dans le corpus étudié.

Le mot « gerinc » (épine dorsale) apparaît dans le contexte de l'œnologie, lié aux arômes du vin : « 34% olaszrizling adja a gerincet, amit a két rizling és chardonnay támogat. » (34% de riesling italien donne l'épine dorsale [de ce vin] soutenu par deux [autres] rieslings et du chardonnay.) Dans les autres phrases, c'est une personne qui « constitue l'épine dorsale » d'un événement, d'un concert, c'est-à-dire que cette personne est l'acteur principal de l'événement. Nous retrouvons également ces sens avec la terminaison possessive, la différence entre les deux usages est la fréquence : l'usage avec la terminaison possessive prédomine largement (314 versus 11 occurrences).

Pour les autres noms, il est difficile d'établir des catégories. Nous y trouvons aussi bien des expressions idiomatiques (par exemple, donner son consentement, faire comprendre) que de

[100] Nous avons manuellement éliminé du nombre total (1468) les exemples dans lesquels le mot « név » n'est pas le COD ainsi que les occurrences avec un préfixe verbal.

nombreuses phrases dans lesquelles « adja » a un sens plutôt abstrait (par exemple, fournir de l'énergie, donner la moitié de la structure symétrique, représenter la moitié des ventes).

4) Étude de cas (3) : Même collocatif, fréquence comparable des deux conjugaisons : « jel » (signe)

Le corpus contient 816 occurrences avec la conjugaison indéfinie et le nom « jel » comme COD. « Jelt/jelet » (« signal, signe » au singulier avec la terminaison du COD) apparaît 645 fois et « jeleket » (« signaux, signes » au pluriel avec la terminaison du COD) 171 fois. L'analyse de ces énoncés révèle trois contextes différents qui sont présentés dans le tableau 146.

4.1) Conjugaison indéfinie : « jelt/jelet + ad, jeleket + ad » (donne un signal, des signaux/signes)

(1) Communication du quotidien	***(1) Communication du quotidien***
Előzni szabad, de csak ha a pályabíró **jelet ad** a zászlóval.	*Vous pouvez dépasser, mais seulement si l'arbitre* ***donne le signal*** *avec le drapeau.*
Szerintem ez általában úgy történik, hogy a lány **jelet ad,** illetve a fiú vár erre a jelre.	*Je pense que cela arrive généralement lorsque la fille* ***donne un signal*** *ou que le garçon attend ce signal.*
Ha visszamosolyog, vagy valami egyértelmű **jelet ad** (kacsint), akkor akkor nyert ügy... odamész, bemutatkozol.	*Si elle sourit en retour ou* ***donne un signal clair*** *(clin d'œil) à quelque chose, alors tu as gagné l'affaire ... tu y vas et tu te présentes.*
Egy nő inkább csak **jeleket ad**, hogy szimpatikus neki a férfi.	*Une femme préfère seulement* ***donner des signes*** *indiquant qu'elle trouve l'homme sympathique.*
Az ingatlanpiac nem tud beszélni, csak **jeleket ad**.	*Le marché immobilier ne peut pas parler, il* ***donne*** *juste* ***des signaux****.*
Tudat alatt olyan **jeleket ad** a kutyának, hogy neki itt mindent szabad.	*Il* ***donne*** *inconsciemment au chien* ***des signes*** *qu'il peut faire ce qu'il veut ici.*
(2) Contexte religieux	***(2) Contexte religieux***
Ahogy a próféta kijelentette: Isten maga **ad jelet**.	*Comme le prophète l'a déclaré : Dieu lui-même* ***donne un signal****.*
Az Úr maga **ad** majd nektek **jelet**.	*Le Seigneur lui-même vous* ***donnera un signal****.*

Isten nem büntet! **Jeleket ad**, mutatva, merre menj.	*Dieu ne punit pas ! Il **donne des signaux** indiquant où aller.*
Ha kell, majd **jelet ad.**	*Si nécessaire, il **donnera un signal**.*
Maga az Isten, aki **jeleket ad** ma is őt kereső fiainak.	*C'est Dieu lui-même qui **donne des signes** à ses fils qui le cherchent aujourd'hui.*
(3) Signal émis par un appareil technique	***(3) Signal émis par un appareil technique***
A kábeltévé pedig folyamatos, teljes, kevert **jelet ad**.	*Et la télévision par câble **émet un signal continu, complet et mixte**.*
Az eszköz bemérhető **jelet ad**, akkor még GPS koordináta nélkül.	*L'appareil **émet un signal mesurable**, même sans coordonnées GPS.*
Az antenna elvileg csak akkor **ad jelet**, amikor rádiót hallgatsz.	*L'antenne **n'émet** en principe **de signal** que lorsque tu écoutes la radio.*
A műszer a kicsi teljesítmény ellenére használható nagyságú **jelet ad**.	*L'instrument **émet un signal utilisable** malgré la faible puissance.*
A körülöttünk lévő világ analóg **jeleket ad**.	*Le monde qui nous entoure, **émet des signaux analogues**.*
Az agy működésekor elektromos **jeleket ad**, amik az elektródákkal érzékelhetők.	*Lorsque le cerveau fonctionne, il **émet des signaux électriques** qui peuvent être détectés par les électrodes.*
Ez az opció valamiért nagyon kis **jelet ad**.	*Cette option **émet** juste **un signal très faible**, je ne sais pas pourquoi.*

Tableau 146 : Des exemples avec « jel/jelek » comme CODs et « ad » (conjugaison indéfinie).

Nous voyons ainsi trois groupes se détacher : (1) la communication liée au quotidien et au comportement humain ou animal, (2) le contexte religieux (Dieu donne des signes), (3) le langage technique. Dans le troisième cas, nous observons que le mot « jelet/jeleket » (signal/signaux) est systématiquement précédé d'un adjectif qui précise la nature du signal. Dans le contexte religieux, le nom n'est pas précédé d'adjectif (le contexte ne laissant aucune ambiguïté que ces signaux ne peuvent venir que de Dieu), et dans les énoncés liés à la communication quotidienne, nous trouvons quelques adjectifs décrivant la qualité du signal (par exemple, clair ou net).

Le corpus contient 745 énoncés avec la conjugaison définie. Dans ces phrases, le nom « jelét » (le signe de X) est systématiquement au singulier. Selon l'ordre des mots, nous trouvons 528 exemples avec COD + verbe (« jelét adja ») et 217 avec verbe + COD (« adja jelét ») (tableau 147).

4.2) Conjugaison définie (1) : « jelét + adja » et « adja + jelét » (donne le signe de X)

Nem tudom, hogy fáj-e a hasa, mert nem **adja jelét**, hogy fájna neki, nem sír.	*Je ne sais pas si son ventre lui fait mal parce qu'il* ***ne donne aucun signe*** *que ça lui fait mal, il ne pleure pas.*
A jógi még mindig nem **adja jelét**, hogy magához térne hosszú önkívületéből.	*Le yogi* ***ne donne*** *toujours* ***aucun signe*** *de ressortir de sa longue transe.*
Ha tudomást is vett jelenlétemről, ennek nem **adja** semmi **jelét.**	*S'il a remarqué ma présence, il n'en* ***donne aucun signe****.*
Én mindig azt hittem, hogy ha egy pasi nem adja **jelét** a vonzalmának, attól még tetszem neki.	*J'ai toujours pensé que même si un gars* ***ne donne pas de signe*** *de son attirance, je lui plais.*
X nem érzi jól magát a kapcsolatban, de nem **adja** semmi **jelét** változtatási szándékának.	*X ne se sent pas bien dans la relation, mais* ***il ne donne aucun signe*** *de son intention de changer.*
Azt hiszem ő érzékeli rajtam a változást, de nem **adja jelét.**	*Je pense qu'il sent le changement en moi, mais* ***il n'en donne aucun signe****.*
Mintha már rendben lenne a karja, legalábbis nem **adja jelét** az ellenkezőnek.	*Comme si son bras allait déjà bien, au moins* ***il ne donne pas de signe*** *pour (penser) le contraire.*
A kínai ingatlanpiaci boom nem **adja jelét** annak, hogy kifulladna.	*Le boom immobilier chinois* ***ne montre aucun signe*** *de s'essouffler.*
Szép tőle, hogy így **adja jelét** annak: megőrzött emlékezetében.	*C'est gentil de sa part de* ***signaler de cette façon*** *qu'il m'a gardée dans sa mémoire.*

Tableau 147 : Exemples avec « jelét » comme COD et « adja » (conjugaison définie).

Nous pouvons observer que les phrases négatives (« X ne donne aucun signe de + INF ») prédominent largement parmi les exemples (637 sur 745). C'est donc un schéma grammatical qui s'associe nettement plus fréquemment à la conjugaison définie qu'à la conjugaison indéfinie. Il y a des différences structurelles par rapport au français : le suffixe « -nAk » (littéralement : « il donne

le signal de quelque chose ») correspond à la structure « de + INF » en français. Si l'étudiant s'appuie sur ses connaissances de français, il risque de se tromper dans la conjugaison puisque dans la phrase française le mot négatif est placé devant le nom (aucun signe) alors qu'en hongrois, « nem » (ne pas, ne aucun) se trouve devant le verbe (« nem adja »). La structure française amènerait à la conclusion (fausse) qu'il faut utiliser la conjugaison indéfinie puisque le COD « signe » n'est pas défini. La continuation de la phrase française (« de + INF ») pourrait confirmer cette impression ; voici deux raisons qui expliquent les difficultés de décider – sans exemple authentique – quelle conjugaison est appropriée. Là encore, une approche contrastive informée par le corpus peut contribuer à souligner ces différences et fournir ce faisant des outils pertinents dans le cadre pédagogique.

Nous trouvons nettement moins d'exemples avec l'article défini (« adja a jelet », donne le signal, 107 occurrences dans le corpus) et avec l'article défini + pronom démonstratif (« azt a jelet adja », donne ce signal, 42 occurrences) : 147 énoncés au total. Le premier usage est clairement associé à un contexte technique (émission d'un signal), le deuxième peut faire référence à l'émission d'un signal mais aussi aux signaux que des personnes se donnent (tableau 148). La traduction française de ces phrases contient également l'article défini rendant le choix de la conjugaison plus clair.

4.3) Conjugaison définie (2) : « adja + a jelet » et « azt a jelet + adja »

(1) adja a jelet	**(1) X émet le signal (appareil technique)**
Egy infravörös emitter **adja a jelet** a szemüveg számára, hogy tudja, mikor melyik szemre kell küldeni az adott képkockát.	*Un émetteur infrarouge **donne le signal** pour que les lunettes sachent quand envoyer cette monture à quel œil.*
Elvileg mostantól nagyobb távolságra stabilabban **adja a jelet** a router.	*En principe, le routeur **transmet** désormais **le signal** de manière plus stable sur de plus longues distances.*
Ha a hdmi kábel nincs bedugva akkor nyilván a dvi **adja a jelet**, így a monitor működik rendesen.	*Si le câble HDMI n'est pas branché, il est évident que le DVI **émet le signal** pour que le moniteur fonctionne correctement.*
(2) azt a jelet adja	**(2) X émet le signal (appareil technique)**

A Trim úgy működik, hogy **azt a jelet adja** agyunknak, hogy tele vagyunk, még akkor is, ha semmit sem ettünk. Ha a férfi meglehetősen közel áll hozzánk nyilvánosan, **azt a jelet adja**, hogy ő már elkötelezte magát. Krisztus **azt a jelet adja** híveinek, hogy „Arról ismerjenek meg benneteket, hogy egymást szeressétek”.	**X donne le signe que (personne)** *Trim fonctionne en **donnant** à notre cerveau **le signal** que nous sommes rassasiés, même si nous n'avons rien mangé.* *Si l'homme est assez proche de nous en public, il **signale** ouvertement qu'il est déjà pris.* *Le Christ **donne** à ses disciples **le signe** que « A ceci tous connaîtront que vous êtes mes disciples, si vous avez de l'amour les uns pour les autres. »*

Tableau 148 : Exemples avec « a jelet » (le signal) et « azt a jelet » (ce signal) comme CODs et « adja » (conjugaison définie).

Pour conclure cette première partie de l'analyse, nous constatons que les exemples observés montrent, du moins dans le cas de « ad » et « adja », qu'il existe des différences remarquables entre l'usage des deux conjugaisons à la troisième personne du singulier, notamment :

- Il est possible d'établir des préférences lexicales pour les CODs car chacun s'associe avec une fréquence différente aux deux conjugaisons. Certaines unités multi-lexicales (X donne la possibilité, X donne une opportunité) favorisent clairement l'usage de la conjugaison indéfinie alors que d'autres ont une préférence pour le COD défini (donner sa bénédiction, donner son accord).
- Dans les rares cas où le nombre d'exemples avec le même COD est comparable (« jelet ad », « jelét adja »), nous observons des différences de sens entre les deux usages.
- Nous pouvons également observer des tendances lexico-grammaticales. Les CODs indéfinis n'ont souvent pas d'articles ni modificateur(s) alors que les CODs définis ont majoritairement une terminaison possessive.
- Les noms abstraits prédominent avec les deux conjugaisons, du moins dans le corpus étudié.

D) L'usage de la première personne du singulier

Dans les pages suivantes, nous examinerons des exemples avec les formes du verbe à la première personne du singulier afin d'analyser dans quelle mesure les schémas sont similaires ou différents à la troisième personne. Nous suivrons le même procédé que pour la troisième personne. Nous

comparerons tout d'abord les collocatifs des deux conjugaisons et nous prendrons ensuite trois exemples : (1) un nom toujours utilisé avec la conjugaison indéfinie dans le corpus, (2) un qui a le même nombre d'occurrences avec les deux conjugaisons et (3) un groupe d'exemples associés à la conjugaison définie.

1) Aperçu des collocatifs à la première personne : émergence de nouveaux groupes

Comme pour la troisième personne, nous avons utilisé l'outil « Word Sketch Difference » pour récupérer la liste des collocatifs fréquents à la première personne du singulier. Nous trouvons globalement moins d'exemples à la première personne qu'à la troisième, ce qui s'explique par la nature du corpus : la première personne est avant tout utilisée dans les commentaires, dans les blogs et dans les forums, partie relativement limitée du corpus. Le spectre des sujets est varié, l'élément fédérateur de leur multitude est qu'il s'agit de récits ou d'échanges autour des expériences et opinions personnelles. En revanche, les contenus d'autres sites Web ayant pour but de fournir des informations plus ou moins objectives, contiennent moins d'énoncés à la première personne.

La forme « adok » apparaît 56 630 fois dans le corpus et « adom » 34 563 fois ; l'écart entre les chiffres est donc moins important que dans le cas de « ad » et « adja » (1,5 : 1 versus 4 : 1). Dans les deux cas, la conjugaison indéfinie est utilisée plus souvent. Le tableau suivant présente les collocatifs des deux formes : la colonne de gauche indique le nombre d'occurrences de la conjugaison indéfinie et celle de droite le nombre d'exemples avec la conjugaison définie. Notons que nous avons dû manuellement inclure un nom que le software a identifié comme adjectif : « igazat » (raison). Il s'agit, en fait, d'un adjectif nominalisé (l'adjectif signifiant « vrai ») qui apparait 4763 fois dans le corpus, utilisé exclusivement dans l'expression « igazat adok XnAk » (« je donne raison à X »). Cette unité multi-lexicale prédomine largement dans le corpus (tableau 149).

igaz	4 763	0	
hála	3,230	0	•••
puszi	202	0	•••
parancsolat	92	0	•••
életjel	73	0	•••
ötös	80	0	•••
tápszer	75	0	•••
borravaló	69	0	•••
jóisten	62	0	•••
ízelítő	72	0	•••
felvilágosítás	74	0	•••
sors	171	0	•••
megnyugvás	62	0	•••
hússzív	49	0	•••
árajánlat	50	0	•••
esély	1,315	36	•••
tipp	327	8	•••
hang	497	24	•••
vélemény	662	38	•••
gáz	296	14	•••
isten	689	44	•••
ötlet	274	19	•••
interjú	168	11	•••
tanács	1,030	79	•••
engedély	146	10	•••
pont	428	37	•••
csillag	148	11	•••
csók	106	9	•••
link	229	32	•••
pénz	701	138	•••

úr	173	51	•••
kegyelem	72	16	•••
hír	192	65	•••
táp	65	14	•••
parancs	154	62	•••
elérhetőség	59	18	•••
utasítás	74	41	•••
cumi	46	17	•••
ajándék	218	171	•••
szeretet	120	119	•••
szív	173	189	•••
tájékoztatás	151	137	•••
szó	294	429	•••
kéz	398	685	•••
bér	33	56	•••
eledel	11	13	•••
írás	44	120	•••
remény	58	150	•••
szegény	21	46	•••
falat	9	19	•••
cumisüveg	8	14	•••
kincs	9	25	•••
békesség	31	82	•••

kulcs	32	116	•••
örökség	8	27	•••
áldás	26	175	•••
korona	11	81	•••
tudta	11	267	•••
állásfoglalás	0	16	•••
félár	0	10	•••
beleegyezés	0	13	•••
koszorú	0	14	•••
négyszeres	0	11	•••
bölcsi	0	16	•••
hozzájárulás	0	32	•••
hajnalcsillag	0	16	•••
leány	0	41	•••
utód	0	56	•••
értés	0	33	•••
voks	0	36	•••
juh	0	34	•••
becsületszó	0	58	•••
béke	0	139	•••

(raison (rajouté à la liste)
grâce (rendre grâce à Dieu)
bisou
commandement
signe de vie
cinq (note)
alimentation pour le bébé
pourboire
bon Dieu (non COD)
échantillon
renseignements, informations
destin (non COD)

Seigneur (non COD)
grâce
nouvelles
alimentation pour le bébé
commande
coordonnées
biberon
cadeau
amour
cœur
renseignements
mot

clé
héritage (non COD)
bénédiction
couronne
savoir (expr. id. : faire savoir qqch à qqn)
point de vue
moitié prix (plus COD)
consentement, accord
couronne de fleurs
quatre fois (non COD)
étoile du berger (non COD)

tranquillité	*main*	*fille*
cœur de chair	*location (non COD)*	*progéniture*
offre	*nourriture*	*comprendre (non COD)*
chance	*écriture*	*vote*
tuyau, suggestion	*espoir*	*mouton*
voix (exprimer qqch)	*pauvre (non COD)*	*parole d'honneur*
opinion (non COD)	*bouchée*	*paix)*
gas	*biberon*	
dieu	*trésor*	
idée	*paix (intérieur)*	
interview		
conseil		
autorisation		
point (dans une évaluation)		
étoile (dans une évaluation)		
baiser		
lien		
argent		

Tableau 149 : Les collocatifs les plus usités avec « adok » et « adom » dans « Word Sketch Difference ».

Nous retrouvons dans les deux listes (troisième et première personnes) un grand nombre de collocatifs, certains d'entre eux typiques à la troisième personne n'apparaissent cependant qu'avec une faible fréquence à la première personne. Il s'agit, logiquement, avant tout des phrases dont le sujet n'est pas une personne. Une similarité entre les deux personnes grammaticales est la forte présence de CODs définis avec une terminaison possessive.

Nous voyons onc émerger deux nouvelles catégories relatives à l'usage à la troisième personne : (1) les CODs désignant des actions et des objets concrets du quotidien (pourboire, bisou, alimentation) et (2) un nombre plus important de CODs exprimant une évaluation personnelle. Le tableau 150 en donne quelques exemples illustratifs.

(1) Le COD est une chose concrète	***(1) Le COD est une chose concrète***
Cumisüveget **adok** neki meleg tejjel, mosolyog.	*Je lui* ***donne une bouteille*** *de lait chaud, elle sourit.*

A macimnak egy nagy puszit **adok**, és mindketten elalszunk.	*Je **donne un gros bisou** à mon nounours et nous nous endormons tous les deux.*
Nem **adok** borravalót senkinek, aki csak úgy simán a munkáját elvégzi.	*Je **ne donne pas de pourboire** à ceux qui font juste leur travail.*
Több borravalót **adok**, mint amennyit megérdemel, de annyi baj legyen.	*Je **donne** plus **de pourboire** que ce qu'il mérite, tant pis.*
(2) Évaluation	***(2) Évaluation***
De teljesen igazat **adok** abban Xnek, hogy vita nélkül nem működhet semmi.	*Mais **je donne** tout à fait **raison** à X que rien ne peut fonctionner sans débat.*
És természetesen igazat **adok** abban, amit a rendrakásról írsz.	*Et bien sûr, **je** te **donne raison** par rapport à ce que tu dis sur le ...*
Én erre most kettő csillagot **adok**, lehet ezen még javítani.	***Je** lui **donne deux étoiles** maintenant, il y a place à amélioration.*
Izgalmasnak, üdítőnek találtam, négy csillagot **adok** rá.	*J'ai trouvé ça excitant, rafraîchissant, **je** lui **donne quatre étoiles**.*
Annyira hangulatos volt a film, hogy 9 pontot **adok** neki!	*Le film était si plaisant que **je** lui **donne 9 points** !*
Ez az a könyv, amelyre tízből tizenegy pontot **adok** : zseniális és iszonyat erős.	*C'est le livre auquel **je donne onze** sur dix : brillant et extrêmement puissant.*
Összességében 4 pontot **adok** rá tízből, biztos van, akinek bejön ez a humor, de nekem nem.	*Globalement, **je** lui donne **4 points** sur dix, je suis sûr qu'il y a des gens qui apprécient cet humour, mais pas moi.*
A koreográfusnak egy hatalmas csillagos ötöst **adok**.	***Je donne un énorme cinq sur cinq** au choréographe.*

Tableau 150 : Catégories divergeant de la troisième personne du singulier.

Ces énoncés nous donnent un aperçu des sujets discutés abondamment sur les réseaux sociaux. Ils concernent les événements du quotidien (éducation, services, actualités) et leur évaluation fait très

clairement partie du discours (cf. Cardon 2013 ; Carter et McCarthy 2006 ; Domonkosi 2018a ; Domonkosi 2018b ; Rühlemann 2007, 2018 ; McCarthy 2000, 2002, 2003)[101].

Cette analyse permet de constater la forte présence des énoncés relatifs à la religion chrétienne tels que « je vous donne ma paix » (le locuteur cite les mots de Jésus), « je rends grâce à Dieu/au bon Dieu pour ... » (le locuteur remercie Dieu), « Dieu donne un cœur de chair à l'homme » (citation biblique). L'usage de ces collocatifs se limite au contexte religieux, il ne s'agit pas d'une tendance générale. De tels usages ne devraient donc pas justifier une présentation spécifique dans le cadre pédagogique (aux premiers niveaux au moins).

2) Étude de cas (1) : le nom « esély » (chance) comme COD

De la même façon que pour l'étude des occurrences à la troisième personne, nous avons sélectionné un nom qui apparaît nettement plus souvent avec la conjugaison indéfinie qu'avec la conjugaison définie. Nous avons ainsi examiné des énoncés avec « esélyt adok » (je donne une chance), nom apparaissant 1315 fois avec la conjugaison indéfinie et 36 fois avec la conjugaison définie. Les instances avec la conjugaison définie peuvent se regrouper en deux grandes catégories que montre le tableau suivant (tableau 151).

(1) Donner une chance à qqn ou à qqch / être prêt à (ré)essayer quelque chose	***(1) Donner une chance à qqn ou à qqch / être prêt à (ré)essayer quelque chose***
Úgy döntöttem, **esélyt adok** egy tő petrezselyemnek és egy tő bazsalikomnak.	*J'ai décidé* ***de donner une chance à une plante*** *de persil et à une plante de basilic.*
Azután eltelt egy kis idő, és elhatároztam, hogy **esélyt adok** a sorozatnak.	*Puis un certain temps s'est écoulé, et j'ai décidé* ***de donner une chance à la série****.*
Át akarom írni a téves kiindulópontot, és akkor **esélyt adok** annak, hogy minden megváltozzon.	*Je veux réécrire le mauvais point de départ, c'est comme ça que* ***je donne une chance à ce que*** *tout change.*
Elhatároztam, hogy újabb **esélyt adok** a 3D-s műszempillának.	*J'ai décidé* ***de donner une autre chance aux faux cils 3D****.*
A következő telefonomnál **esélyt adok**	

[101] Il convient de noter que les CODs à sens concrets utilisés avec la première personne présentent une très grande variété dans le corpus. Dans la structure fréquemment utilisée « ad + COD + COI » (donner qqch à qqn), la position du COD peut être remplie par un grand nombre de mots comme « linket » (lien), « egy aszpirint » (une aspirine), « egy kis tál tejet » (un petit bol de lait), « házi feladatot » (des devoirs).

az Android OS-nek. Előítéletes lettem tisztára, de **adok** majd **esélyt** a könyvnek, ha azt mondod jó. Két epizód után ott tartok, hogy még egy utolsó **esélyt adok** a sorozatnak, és elkezdem a harmadik részt is.	***Je donnerai une chance à Android OS*** *pour mon prochain téléphone.* *Je manque totalement d'objectivité, mais* ***je donnerai une chance au livre*** *si tu dis qu'il est bien.* *Après deux épisodes, je pense* ***que je donne une dernière chance à la série*** *et je commence à regarder la troisième partie.*
(2) Quantifier la probabilité (sport, compétition) Még most is 50-50 százalékos **esélyt adok**. Az a csapat fog nyerni, amelyik nyugodtabban játszik. A visszavágón 20% **esélyt adok** nekik a továbbjutásra, a mi 15-ünkkel szemben. Én személy szerint kb. annyi **esélyt adok** a Róma 3. helyére, mint a Lazioéra. Nem sokat.	***(2) Quantifier la probabilité (sport, compétition)*** *Même maintenant,* ***je donne une chance de 50-50****. L'équipe qui jouera plus calmement, gagnera.* *Sur le match retour,* ***je leur donne 20% de chances*** *de se qualifier, par rapport à nos 15.* *Personnellement,* ***je donne autant de chances*** *à la 3e place de la Roma que du Lazio. Pas beaucoup.*

Tableau 151 : Catégories d'usage avec « esélyt adok » (je donne une chance) dans le corpus écrit.

Le premier groupe comprend des énoncés exprimant le fait que le locuteur est prêt à essayer ou à réessayer quelque chose. En général, il s'agit d'une première, d'une deuxième ou d'une dernière chance avant que le locuteur abandonne ses essais et passe à autre chose. De nombreux énoncés sont introduits par un ou plusieurs éléments lexicaux indiquant l'intention ou la volonté du locuteur : « elhatároztam » (j'ai pris la décision), « úgy döntöttem » (j'ai décidé), « most ott tartok, hogy » (je crains que). Notons également que l'expression française requiert un article défini alors que l'expression hongroise n'a pas d'article du tout.

Une comparaison entre les schémas grammaticaux de la troisième et la première personnes indique quelques différences impactant le sens. Dans le cas de la troisième personne, la colligation typique est « X esélyt ad vkinek vmire » (X donne une chance à une personne ou à un groupe de personnes pour faire quelque chose) et les phrases signalent le potentiel d'une évolution positive. Les phrases à la première personne montrent le schéma suivant : « X esélyt ad vminek/vkinek » (X donne une chance à qqch ou à qqn). Elles indiquent la volonté du locuteur de (ré)essayer quelque chose (premier groupe) ou expriment une estimation personnelle, non-scientifique, de la probabilité qu'un événement se déroulera d'une certaine façon (deuxième groupe). *Ces exemples illustrent ainsi la*

possibilité du changement de l'espace sémantique de la même unité multi-lexicale, en passant d'une personne grammaticale à une autre.

3) Étude de cas (2) : collocatifs fréquent avec la conjugaison définie

Le corpus renferme 35 320 exemples avec le verbe « adom », dont 3 243 exemples avec un préfixe, ce qui nous laisse avec 32 077 occurrences sans préfixe. Nous trouvons les collocatifs les plus usités à la fin de la liste dans le tableau 149, mais les compléments listés ne sont pas tous des CODs. Nous y trouvons ainsi des compléments de lieu comme « bölcsi » (crèche) et d'autres compléments comme « szeretettel » (avec amour) ou « féláron » (à moitié prix), parmi ces phrases nous trouvons de nombreuses expressions idiomatiques comme « értésére adom » (faire comprendre qqch à qqn), « tudtára adom » (faire savoir qqch à qqn). Nous avons donc supprimé ces occurrences avant de procéder à l'analyse.

3.1) CODs avec une terminaison possessive

Comme à la troisième personne du singulier, les collocatifs avec une terminaison possessive sont en majorité à la première personne (69% des occurrences) : quatre catégories sémantiques se dégagent, présentées dans le tableau 152.

(1) Consentement, accord

Az adataim közléséhez nem **adom** a hozzájárulásomat.

Levelem leközléséhez hozzájárulásomat **adom**, azzal a kikötéssel, hogy szó szerint és változtatás nélkül tegyék.

Ezúton is hozzájárulásomat **adom** ahhoz, hogy személyes adataim a Bizottsági döntésekkel összefüggésben nyilvánosságra hozhatók.

Addig nem **adom** a beleegyezésemet, míg azt a fiatalembert nem láttam.

Nos, konferenciákat nézve én a Sonynak **adom** a voksot.

Két párt van akire soha nem **adom** a voksomat, az pedig nem más mint X és Y.

(1) Consentement, accord

Je ne donne pas mon consentement *à la divulgation de mes données personnelles.*

Je donne mon consentement *à la publication de ma lettre, à condition qu'elle apparaisse littéralement et sans changement.*

Par la présente, ***je donne mon consentement*** *à la divulgation de mes données personnelles dans le cadre des décisions de la Commission.*

Je ne donnerai pas mon consentement *tant que je n'aurai pas vu ce jeune homme.*

Eh bien, en regardant les conférences, ***je suis pour Sony*** *(litt. : Je donne mon vote à Sony).*

Il y a deux partis pour lesquels ***je ne voterai jamais****, et ce n'est autre que X et Y.*

Azt nem tudom még kire **adom** a voksomat, de azt igen, hogy ilyen társadalomban nem szeretnék élni.
Marcus Pollard 2 éve visszavonult NFL-játékos is ott lesz, én rá **adom** a voksomat.
Áldásomat **adom** a viszonyotokra.
Azt mondtam, hogy bár nem örülök neki, de ha ennyire fontos, akkor áldásomat **adom** a dologra.

*Je ne sais pas encore pour qui **je vais voter**, mais je sais que je ne veux pas vivre dans une société comme celle-ci.*
*Marcus Pollard, un joueur de la NFL qui a pris sa retraite il y a 2 ans, sera également là, **je** lui **donnerai mon vote**.*
*Je **donne ma bénédiction** à votre relation.*
*Je lui ai dit que je n'étais pas content mais si c'est si important pour lui, je lui **donne ma bénédiction**.*

(2) Engagement (sur l'honneur)
Becsületszavamat **adom**, hogy a fentiek megfelelnek a valóságnak.
De arra a becsületszavamat **adom**, hogy ha rendszeresen látogatod a kurzust, akkor az eredmény meglesz.
Amire én általában a szavamat **adom**, az 100 %.
Ha valakinek szavamat **adom**, hogy ekkor és ekkor ott leszek, akkor már ott vagyok öt perccel azelőtt.
Én mindig a nevemet **adom** a véleményemhez.
Én még a régi iskola híve vagyok, és azt mondom, hogy a munkámhoz a nevemet **adom**.

(2) Engagement (sur l'honneur)
***Je donne ma parole d'honneur** que ce qui précède est vrai.*
*Mais **je donne ma parole d'honneur** que si vous assistez régulièrement au cours, le résultat suivra.*

*Quand **je donne ma parole**, c'est à 100%.*

*Si **je donne** à quelqu'un **ma parole** que je serai là à ce moment-là, j'y suis déjà cinq minutes avant.*

***J'assume complètement** ce que je pense. (litt. : Je donne toujours mon nom à mon avis.)*
*Je crois toujours à la vieille école et **j'assume** le travail que je fais (litt. : je donne mon nom à mon travail.*

(3) Contexte religieux (limité à ces phrases)
Az én békémet **adom** neked.
Az én békémet **adom** nektek.
Neked **adom** a mennyország kulcsait.

(3) Contexte religieux (limité à ces phrases)

Je te donne ma paix.
Je vous donne ma paix.
***Je** vous **donne les clés du Royaume des cieux**.*

(4) Inscrire un enfant à une institution
Egyévesen nem **adom** a gyerekem bölcsibe**.**

(4) Inscrire un enfant à une institution

	***Je n'inscris pas mon enfant** âgé d'un an **à la crèche**.*
Még lehet, hogy később én is bölcsibe **adom** a picúrt.	*Peut-être que **j'inscrirai le petit à la crèche** plus tard.*
Ha megoldja az állam, akkor bölcsibe **adom** a fiam, és akkor visszamegyek dolgozni.	*Si l'état trouve une solution, **j'inscrirai mon fils à la crèche** et je retournerai travailler.*

Tableau 152 : Des exemples avec un COD avec terminaison possessive et « adom » (conjugaison définie).

Les quatre groupes sémantiques sont listés dans l'ordre de leur fréquence. Néanmoins, nous constatons que la variance des collocatifs s'associant à la forme « adom » est limitée au sein de chaque groupe : chaque sous-ensemble forme ainsi un groupe cohérent, associé à un sens spécifique. Ces occurrences font donc apparaître des usages bien particuliers.

3.2) CODs avec l'article défini

Les exemples avec l'article défini (31% de toutes les occurrences) contiennent des CODs dont la majorité (79%) sont des noms concrets. En opposition avec le groupe précédent, leur variété au niveau lexical est grande, néanmoins nous pouvons établir quelques grandes catégories selon le sens du verbe (tableau 153).

(1) ad = árul Xt	**(1) Vendre quelque chose**
Tudomásomra adta, hogy nagyon drágán **adom a** festéket.	*Il m'a fait comprendre que **je vendais la peinture à un prix très élevé**.*
Ki volt akadva, hogy neki 5 centtel drágábban **adom a** sütiket, mint a piacon.	*Elle était furieuse que **je vende les biscuits 5 cents de plus** qu'au marché.*
Azt kérdezte, hogy mennyiért **adom a** rózsákat.	*Il m'a demandé à combien **je vendais les roses**.*
(2) ad = közöl Xt	**(2) Donner des informations**
Akit érdekel ez a lehetőség, keressen és **adom a** kontaktot.	*Si cette opportunité vous intéresse, contactez-moi, **je vous donnerai les coordonnées**.*
Szóval ha szeretnétek profi operatőrtől tanulni, akkor keressetek meg és **adom a** bővebb infókat.	*Alors si vous voulez apprendre d'un caméraman professionnel, contactez-moi et je vous **donnerai plus d'informations**.*
(3) ad = X kezébe ad Yt	**(3) Donner Y dans la main de X**
Kérik az úti okmányokat... **Adom a** jogsit, **a** forgalmit, **a** zöld kártyát...	*Ils me demandent les documents... **Je** leur **donne le permis, la carte grise, la carte verte**...*

Ha megbüntetnek, a drága polgármesternek **adom a** csekket. Fizesse be ő.	*Si je suis condamné à une amende,* ***je donnerai le chèque*** *au cher maire. Laissez-le payer.*
Neked **adom a** térképet, én anélkül is eltalálok a szigetre.	***Je*** *te* ***donne la carte****, je peux trouver le chemin de l'île sans elle.*
(4) ad = hozzáad XhEz Yt (gyakran receptben)	**(4) Ajouter Y à X (souvent dans des recettes)**
Csak akkor **adom a** zöldséget a húshoz, ha már puha a borjú.	***Je n'ajoute les légumes à la viande*** *que lorsque le veau est tendre.*
A tojásfehérjékhez **adom a** sót, és elkezdem felverni.	***J'ajoute le sel aux blancs d'oeufs*** *et je commence à fouetter.*
A tésztához **adom a** hajdinát, **a** koriandert.	***J'ajoute aux pâtes le sarrasin, la coriandre****.*

Tableau 153 : Exemples de CODs avec l'article défini et « adom » (conjugaison définie).

Nous pouvons observer différentes colligations et collocations selon le sens de « adom ». Le sens de « vendre » s'associe aux modificateurs indiquant le prix : « drágán/olcsón/Xért adom » (je donne/vends Y cher/pas cher/pour X), le sens d'« ajouter » est utilisé avec le suffixe -hEz dans les recettes : « XhEz adom Yt » (je donne Y à X, j'ajoute Y à X). Quand il s'agit du sens abstrait de « donner, fournir », les collocatifs de « adom » sont des mots abstraits, relatifs à un échange d'informations comme « coordonnées » ou « informations, renseignements ». Le sens concret est associé aux noms d'objets comme « permis », « carte grise » indiquant qu'il s'agit d'une action réelle. C'est donc une fois de plus l'environnement textuel qui rend le sens du verbe non ambigu, comme nous l'avons vu dans les chapitres 8 à 10, et détermine, au moins en partie, la conjugaison. Les deux premiers usages nécessitent la conjugaison définie, le troisième et le quatrième permettent plus de flexibilité mais la majorité des phrases exprimant ces sens prennent un COD défini avec la conjugaison définie, du moins dans le corpus étudié.

4) Étude de cas (3) : même collocatif, même nombre d'occurrences, conjugaison différente : le nom « tájékoztatás » (renseignement) comme COD

Dans le cas des unités multi-lexicales « tájékoztatást + adok » et « tájékoztatást + adom », il s'agit de vraies unités parallèles : « tájékoztatást » (renseignements) est en effet le COD des phrases avec la conjugaison indéfinie ainsi qu'avec la conjugaison définie.

4.1) Conjugaison indéfinie : « tájékoztatást + adok » (je donne des renseignements)

Cette unité multi-lexicale se trouve 151 fois dans le corpus, accompagnée des schémas grammaticaux suivants : « (ADJ +) tájékoztatást adok XrÓl » (je vous donne des renseignements (+ ADJ) sur X) et « XrÓl + MOD adok tájékoztatást » (je vous donne des renseignements + MOD sur X) (tableau 154).

(1) (ADJ +) tájékoztatást adok XrÓl	**(1) je vous donne des renseignements (+ ADJ) sur X**
Minden érdeklődőnek a felmerülő igények alapján részletes tájékoztatást **adok**.	***Je fournirai des informations détaillées*** *à tous les intéressés en fonction des besoins qui se présentent.*
A jövőbeli fejlődésről itt, és személyes oldalamon is tájékoztatást **adok** az érdeklődőknek.	***Je fournirai*** *aux intéressés* ***des informations*** *sur les développements futurs ici et sur ma page personnelle.*
A részletekről bővebb tájékoztatást **adok** január első napjaiban.	***Je donnerai plus de détails*** *dans les premiers jours de janvier.*
(2) XrÓl + MOD adok tájékoztatást	**(2) je vous donne des renseignements + MOD sur X**
A részletekről esetleges kérdés esetén magam is szívesen **adok** tájékoztatást.	***Je serai heureux de fournir des informations*** *sur les détails si vous avez des questions.*
Az addigi teendőkről e-mailen vagy telefonon **adok** tájékoztatást.	***Je*** *vous* ***fournirai*** *par e-mail ou par téléphone* ***des informations*** *sur ce qu'il faut faire.*
A felmérés menetéről majd a helyszínen **adok** tájékoztatást.	***Je donnerai des informations*** *sur place sur le déroulement de l'enquête.*

Tableau 154 : Exemples avec « tájékoztatást » comme COD indéfini et « adok » (conjugaison indéfinie).

Dans 45% des phrases du premier type, le mot « tájékoztatást » est précédé d'un des adjectifs suivants : « részletes » (détaillé), « pontos » (précis), « bővebb » (plus détaillé), « folyamatos » (continu), « rövid » (bref), további (autre), « egyedi » (individuel). Ces adjectifs sont par ailleurs les collocatifs typiques du nom « tájékoztatás » dans le corpus, indépendamment du verbe. Dans les phrases du deuxième type, le modificateur indique le moyen par lequel le locuteur donnera des renseignements : « telefonon » (par téléphone), « e-mailben » (par e-mail), « személyesen » (en

personne), a « helyszínen » (sur place). Dans quelques cas, le modificateur précise un autre aspect de l'action : « szívesen » ou « örömmel » (avec plaisir), « később » (plus tard).

4.2) Conjugaison définie : « tájékoztatást + adom »

Cette unité multi-lexicale est uniquement utilisée dans des communications officielles. 115 occurrences sur 137 (qui restent après avoir éliminé les doublons) sont des variantes de la phrase suivante : « XvAl kapcsolatosan az alábbi / a következő tájékoztatást adom : ... » (Concernant X, je vous fournis les renseignements suivants : ...). Il s'agit donc d'une phrase standard, formelle, qui sert à introduire des informations sur une question spécifique. Les occurrences dans le corpus illustrent donc un seul usage, contrairement à celui avec la conjugaison indéfinie. Là encore, le corpus révèle le lien parfois intime entre contexte, usage et grammaire.

E) Profil contrastif des deux conjugaisons du verbe « ad »

Les tableaux 155 et 156 résument les résultats les plus importants de notre recherche. Nous avons listé les usages/sens typiques avec les formes « ad », « adja », « adok » et « adom », l'information si l'usage donné est plutôt typique dans le corpus écrit ou dans le corpus oral ainsi que quelques collocations.

« ad » **(il/elle donne, conjugaison indéfinie)**	**« adja »** **(il/elle donne, conjugaison définie)**
1 Opportunité, chance, possibilité lehetőséget ad XrA (*possibilité de X*) esélyt ad XnAk (*chance à X*) alkalmat ad XrA (*occasion pour X*)	**1 Réaction, réponse, aide, soutien** támogatását adja XrA (*son soutien*) azt a választ adja (*la réponse que*)
2 Renseignements, informations, explication, réponse, description, aperçu, image ADJ képet ad XrÓl (*une image ADJ de X*)	**2 Actualité, propos, occasion d'un événement** X apropóját adja (*le propos de X*) X aktualitását adja (*l'actualité de X*)
3 Force, énergie, sécurité, espoir erőt ad XhEz (*de la force pour X*) energiát ad XhEz (*de l'énergie pour X*) biztonságot ad (*sécurité*) reményt ad (*espoir*)	**3 Accord, consentement, bénédiction** áldását adja XrA (*sa bénédiction à X*) beleegyezését adja XhEz (*son accord à X*) hozzájárulását adja XhEz (*son consentement à X*)
4 Nouvelles, signe, signal	**4 Particularité, singularité, valeur d'un événement/d'une chose**

jelt ad XrÓl (*signe de X*)	X különlegességét adja (*la particularité de X*)
életjelt ad XrÓl (magáról) (*signe de vie*)	X értékét adja (*la valeur de X*)
hírt ad XrÓl (*nouvelles*)	X egyediségét adja (*la singularité de X*)
5 Venue, cadre, espace, forum	**5 La base, le point de départ de X**
otthont ad XnAk (*héberger X*)	X kiindulópontját adja (*le point de départ de X*)
keretet ad XnAk (*du cadre à X*)	X alapját adja (*la base de X*)
teret ad XnAk (*espace à X*)	
fórumot ad XnAk (*forum à X*)	

Tableau 155 : Profil contrastif des deux conjugaisons à la troisième personne du singulier.

« adok » **(je donne, conjugaison indéfinie)**	**« adom »** **(je donne, conjugaison définie)**
1 Noms abstraits comme COD (usage similaire à la troisième personne)	**1 Noms abstraits comme COD : consentement, accord, bénédiction, vote**
esélyt adok XnAk (*chance à X*)	hozzájárulásomat adom XhOz (*mon consentement à X*)
lehetőséget adok XnAk YrA (*possibilité à X de Y*)	a voksomat adom XhOz (*mon vote pour X*)
teret adok XnAk (*de l'espace à X*)	áldásomat adom XhOz (*ma bénédiction à X*)
2 Évaluation du locuteur	**2 Inscrire un enfant à une institution**
ötöst adok XnAk (*cinq sur cinq à X*)	bölcsibe/iskolába adom Xt (*X à la crèche/à l'école)*
öt csillagot adok XnAk (*cinq étoiles à X*)	
tíz pontot adok XnAk (*dix sur dix à X*)	
3 Le COD est un objet concret	
cumit XnAk (*biberon à X*)	
puszit XnAk (*bisou à X*)	
borravalót XnAk (*pourboire à X*)	
pénzt XnAk (*argent à X*)	

Tableau 156 : Profil contrastif des deux conjugaisons à la première personne du singulier.

Nous pouvons observer qu'il est possible de catégoriser des usages typiques par type de conjugaison et par personne grammaticale. Les deux conjugaisons semblent donc être plus qu'un phénomène grammatical du hongrois. Il est possible d'identifier des CODs typiques, fréquents avec chacune des conjugaisons ainsi que d'observer des tendances générales qui vont au-delà de l'utilisation d'une seule unité multi-lexicale. *De nombreux exemples illustrent les façons dont le lexique et la*

grammaire s'influencent mutuellement : les informations grammaticales contribuent au sens et les éléments lexicaux émergent dans des environnements grammaticaux bien précis. Ainsi, les explorations au sein de ce chapitre nous amènent à la conclusion que *la présentation des deux conjugaisons pourrait, en effet, bénéficier d'une approche fondée sur le corpus.*

Les exemples semblent indiquer que les deux conjugaisons représentent bien plus qu'un phénomène grammatical particulier de la langue hongroise et elles ne devraient pas être étudiées ni décrites sans leurs environnements textuels plus larges. Ce constat est de première importance dans un contexte d'enseignement, en classe ou lors de la définition des ouvrages pédagogiques. L'étude plus approfondie des unités multi-lexicales peut en effet faire émerger des schémas susceptibles d'être utiles pour les apprenants. De fait, l'interrelation entre l'environnement textuel et le choix de conjugaison mérite d'être mise en valeur lors des présentations dans le cadre pédagogique.

Nous pouvons donc résumer les résultats les plus pertinents de notre recherche concernant les deux conjugaisons comme suit :

- Les noms remplissant la fonction du complément d'objet direct peuvent être catégorisés selon leurs propriétés sémantiques. Ces caractéristiques facilitent la prédiction du type de conjugaison.
- Le type de conjugaison dans la phrase semble restreindre les éléments lexicaux qui peuvent potentiellement devenir des CODs.
- Si les deux conjugaisons peuvent prendre, avec un pourcentage comparable, le même nom comme COD, l'environnement textuel ainsi que les propriétés sémantiques des énoncés seront différents dans les deux cas.
- Il semble également possible d'identifier des colligations typiques par conjugaison. Par exemple, le COD défini semble favoriser la terminaison possessive pour la conjugaison définie et le COD indéfini l'usage sans article.
- Les occurrences liées à des personnes grammaticales différentes semblent indiquer des différences qui peuvent s'exprimer soit par un écart de fréquence avec le même collocatif soit par des choix autres de collocatifs. Cela met l'accent sur la nécessité d'étudier séparément des personnes grammaticales.

Les études présentées dans ce chapitre ne constituent que le début des explorations d'un terrain largement vierge. L'étude d'un grand nombre d'exemples, incluant des verbes courants et plus

rares, serait nécessaire pour valider (ou réfuter) les résultats obtenus sur un petit échantillon. Ces résultats restent néanmoins des plus prometteurs autant par les pistes linguistiques qu'ils ouvrent (songeons à l'observation du lien souvent étroit entre contexte – lexical ou situationnel – et la forme grammaticale) mais aussi quant aux développements pédagogiques qu'ils permettent d'envisager. Ce dernier point sera l'objet du prochain chapitre.

Chapitre 12 : Présenter les résultats d'analyse de corpus dans le cadre pédagogique

Les études effectuées aux chapitres 8 à 11 démontrent que la mise en avant de l'interconnexion entre grammaire et lexique pourrait significativement enrichir la présentation des phénomènes linguistiques choisis. Dans ce chapitre, nous examinerons l'intégration possible de ces analyses dans la pratique pédagogique : notre attention portera donc sur *la/les manière(s) d'exposer de façon claire et accessible aux apprenants de niveaux de compétences linguistiques inférieurs des informations obtenues à partir du corpus.*

Nous nous concentrerons sur les trois composantes de la présentation, toutes trois utilisées dans les chapitres précédents – (1) lignes de concordance, (2) unités multi-lexicales fréquentes et (3) tableaux des schémas d'usage – et nous explorerons les questions suivantes :

- Comment choisir et simplifier les exemples issus du corpus ?
- Comment présenter les unités multi-lexicales contenant l'élément étudié ?
- Comment présenter les schémas d'usage ?
- Comment relier entre elles les trois composantes de la présentation ?
- Quelles activités proposer dans le cadre pédagogique ?

Pour répondre à ces questions, nous utiliserons les aspects étudiés aux chapitres 8 à 11 : « nehéz » (lourd, difficile) comme exemple des mots aux sens multiples, « tűnik » et « látszik » (sembler, paraître) et « megjön » et « eljön » (venir, arriver) comme exemples de synonymes et les deux conjugaisons comme exemples d'un phénomène considéré, avant tout, comme grammatical.

Nous commencerons notre démonstration par la description d'une technique que nous appellerons « zoom in, zoom out ». Cette technique implique l'observation de l'environnement immédiat et de l'environnement plus large de l'élément choisi. Les parties consécutives seront

dédiées aux trois composantes de la présentation mentionnées plus haut et la dernière section proposera des activités pour le cours de langues.

A) Technique de présentation : « zoom in, zoom out »

Comme le remarque Sinclair (2004b : 280-81), « [l]'élément langagier est mieux décrit au maximum, et non au minimum » (notre traduction) où une description minimale serait une description de la signification d'un mot individuel sans aucune information sur son utilisation. Dans notre cas, cette description s'effectue donc par le biais des lignes de concordance, des unités multi-lexicales et des schémas d'usage.

Dans le cadre de cette technique, les apprenants travaillent sur ces trois composantes et passent d'une composante à une autre dans l'ordre qu'ils choisissent. Par exemple, ils peuvent commencer leurs explorations en lisant des lignes de concordance, les continuer avec l'observation des unités multi-lexicales et finir par l'étude des tableaux des schémas, pour retourner par la suite à l'étude des lignes de concordance. Le procédé consiste donc en trois étapes et permet d'analyser l'élément choisi de plusieurs manières :

- Apprentissage avec environnement textuel plus large (lignes de concordance) : l'apprenant observe l'usage de l'élément étudié dans ses environnements textuels typiques.
- Apprentissage avec environnement textuel minimal (unités multi-lexicales à deux ou trois composantes) : l'attention porte sur l'environnement immédiat de l'élément étudié.
- Apprentissage avec le tableau des schémas d'usage : l'attention est délibérément concentrée sur une présentation plus abstraite des caractéristiques de l'élément choisi.

Nous avons intitulé cette technique *explorant l'élément choisi en trois étapes* « zoom in, zoom out ». Le nom de cette technique a été choisi parce que la technique rappelle le jeu de champs de caméra. Les lignes de concordance présentent l'environnement textuel plus large – on pourrait dire qu'ils élargissent le champ de notre caméra imaginaire –, alors que les unités multi-lexicales et les tableaux des schémas rétrécissent ce champ en révélant l'environnement immédiat de l'élément étudié. Les deux perspectives se complètent et attirent l'attention de l'apprenant sur des aspects différents de (ce que veut dire) « connaître le mot » comme défini par Nation (2013, voir le chapitre 8). *L'ordre des étapes du procédé est variable et une étape peut être répétée autant de fois que nécessaire.* En faisant des allers-retours entre unités multi-lexicales, lignes de concordance et tableaux des schémas, les apprenants acquièrent, d'une part, une expérience linguistique et, de l'autre, une méthode d'analyse qui ne

sépare pas les différents aspects du langage. Bien au contraire : elle montre à l'apprenant que ces aspects sont interdépendants et s'influencent mutuellement[102].

Les chapitres 8 à 11 ont traité de questions typiques que les étudiants sont susceptibles de poser en cours de langues. Comme nous l'avons vu, ces questions peuvent concerner des aspects linguistiques traditionnellement considérés avant tout comme grammaticaux ou avant tout comme lexicaux. L'utilisation des corpus peut préciser et compléter la description obtenue à partir d'un dictionnaire ou d'une grammaire. Ce chapitre se concentrera sur la présentation efficace des informations issues des corpus, dans le cadre pédagogique. Nous ciblerons les cours de langues dont la flexibilité inhérente permet l'adaptation des suggestions aux besoins des apprenants plus facilement que les ouvrages écrits.

B) Présenter des exemples : les lignes de concordance

1) L'intérêt des lignes de concordance : l'« exposition condensée »

Un des bénéfices les plus importants du travail sur un grand nombre d'énoncés est « l'exposition condensée » (Gabrielatos 2005), c'est-à-dire l'accumulation d'expériences linguistiques, grâce aux rencontres répétées avec l'élément donné. Porter au maximum les rencontres contribue à approfondir les connaissances sur cet élément car l'apprenant est encore loin d'y avoir été suffisamment exposé pour pouvoir *repérer* les schémas typiques. Contrairement aux natifs, il doit *comprendre et apprendre* comment l'élément choisi est utilisé. Par conséquent, une phase d'observation est nécessaire pendant laquelle l'apprenant peut étudier les utilisations du mot dans des environnements textuels compréhensibles à son niveau linguistique.

L'intérêt de cette approche est donc qu*'au lieu d'apprendre un schéma et d'étudier ensuite quelques exemples qui l'illustrent, l'apprenant fait l'expérience d'être d'abord exposé à des énoncés authentiques, accessibles et systématisés qu'il assimile avant d'en déduire des tendances d'usages.*

La théorie de Hoey (2005) sur le « Priming lexical » (section C.2 du chapitre 4) souligne également l'importance que les rencontres répétées avec un élément choisi dans des environnements textuels différents jouent dans l'acquisition de la langue. Il explique que chaque élément langagier devient « cumulativement chargé des contextes et des co-textes dans lesquels il est croisé » (p. 8, notre

[102] Pour faciliter le travail de l'apprenant, il faut lui donner la possibilité de se déplacer facilement entre les différents modes de présentation. Pour ne pas figer l'ordre de leur utilisation, on peut les proposer sur trois feuilles séparées ou sur trois pages d'un document.

traduction) et que cela s'applique probablement aussi aux unités multi-lexicales. D'après cette théorie, ce sont les rencontres répétées avec les éléments langagiers qui « permettent que leur usage s'enracine et ces enracinements forment eux-mêmes la base d'autres primings. » (Pace-Sigge and Patterson 2018 : ix-x, notre traduction). *Il est donc particulièrement bénéfique pour les apprenants d'avoir à leur disposition des collections d'exemples riches, bien organisés et accessibles sur les points problématiques de la langue.*

La première question qui se pose dans le contexte pédagogique est le nombre d'exemples à proposer aux apprenants. Dans les dictionnaires et dans les encyclopédies, chaque sens du mot est illustré par un ou deux exemples. Cette quantité peut être suffisante pour les natifs qui, ayant déjà rencontré assez souvent le mot dans des contextes variés, peuvent relier les informations à une expérience linguistique déjà existante (Hanks, 2013 : 286). D'après Sinclair (1993), « dix exemples sont un mauvais échantillon ; il en faut au moins cinquante pour décrire les significations d'un mot et cent cinquante pour les identifier de manière fiable » (p. 7, cité par Szirmai 2005 : 29, notre traduction). Dans le cadre pédagogique, l'étude d'un nombre plus limité d'exemples peut suffire pour ne pas submerger l'apprenant : dix à vingt exemples bien choisis, selon l'importance de l'usage à illustrer, peuvent ainsi en éclairer les aspects cruciaux.

2) Préparation : adapter les exemples au niveau de l'apprenant

Alors que la sélection des unités multi-lexicales est une tâche relativement simple, choisir des phrases appropriées dans le contexte pédagogique requiert une attention particulière. La première difficulté concerne le niveau de vocabulaire obtenu dans les exemples. La plupart des phrases authentiques tirées d'un grand corpus non pédagogique contiendra en effet des éléments qui dépassent les connaissances des apprenants aux niveaux de compétences linguistiques inférieurs. Dans l'intérêt de l'accessibilité, la simplification est donc souvent nécessaire. Mais que faut-il éliminer et que faut-il éventuellement rajouter pour rendre le contexte situationnel et le langage plus clairs tout en conservant le caractère authentique de l'énoncé ?

2.1) Utiliser un corpus large

L'outil « Good dictionary examples » permet de trouver de bons exemples d'anglais dans des corpus hébergés sur Sketch Engine. Les critères de sélection pour les phrases proposées sont les suivants : « Il s'agit d'un système d'évaluation des phrases en fonction de leur aptitude à servir d'exemples de dictionnaire ou de bons exemples à des fins pédagogiques. Les phrases sont évaluées en fonction de leur longueur, de l'utilisation du vocabulaire, de la présence de sujets controversés

(politique, religion, etc.), du contexte suffisant, des références pointant en dehors de la phrase (par exemple des pronoms), des noms de marque et d'autres critères. » (Sketch Engine 2021).

Cet outil pourtant très pratique ne peut malheureusement pas être utilisé pour le hongrois[103]. Les critères définis pour la recherche de bons exemples d'anglais peuvent cependant nous être utiles. Nous pouvons les appliquer pour identifier des exemples de hongrois, tâche qui est loin d'être simple comme nous le verrons par la suite. Pour illustrer les difficultés, nous avons choisi des exemples avec les verbes « látszik » (sembler, se voir, avoir l'air). Le tableau 157 montre les vingt premières lignes de concordance avec ce verbe :

1	tényleg nem tudtam mit mondani, hogy előttünk legyen a Tátra, Azért	látszik	, hogy menjek velük, Amíg én alig vártam a vizsgaidőszak végét, Sos
2	0--1037) és Avveroës (Ibn Rosd, 1126--1198) Arisztotelészt sikeresen	látszottak	az iszlám vallással kibékíteni. Az elutasítás tulajdonképpeni oka azon
3	úton elérünk az Operaházig , majd a palotákkal övezett út végén már	látszik	a Hősök tere Millenniumi emlékoszlopa . A Hősök terén vendégeink m
4	sen megszüntették volna. Aztán majdnem átkerült a Bartókra, ám úgy	látszik	, ennek a "kommunisták által kitalált" műsornak a temetőben, vagyis a
5	:: Az első három helyet a táblázatunk alapján határoztuk meg, amiben	látszik	a like gombra kattintások száma, az értéket növelte még a hozzászóla
6	ó és a kék kombinációja). A sárga foltosság a kéken "átütve" zöldnek	látszik	, tehát ezekben az esetekben zöld foltos, illetve zöldes színárnyalatú
7	is változói , ezek csak az alprogram (=függvény v. eljárás) futása alatt	látszanak	, illetve léteznek. A paraméterátadás kétféle lehet cím és érték szerint
8	ppról ezt nyilatkozza: "Egy hónappal a voksolás előtt még semmi sem	látszott	a későbbi súlyos vereségből, az akkori helyi közvélemény-kutatás sze
9	ttak, hogy ennek a tervnek a megvalósulása mostanra lehetségesnek	látszik	. A rehabilitációt irányító orvosok szakmai konferenciákon, publikációk
10	omként is magasabb sok templomnál... Domb tetején épült, messziről	látszik	... Az Intézet rövid története Miután a kolozsvári Ferencz József Tudo
11	adul az állattól. Kutyák minden mennyiségben. Némelyiken pontosan	látszik	, hogy tudatában van a lehetőségeinek, átérzi a hely végállomás jelleg
12	. még az első évadból, hogy egyszer mutattak valamilyen doksit, amin	látszott	a születési dátuma. Arra emlékszem, hogy fiatalították, mintha 1978 l
13	abban nő majd. Még ugyan csak picike bimbó volt rajta, de már így is	látszott	, hogy csodaszép rózsaszínű virágokat fog hozni. Lili különösen örült
14	gyasztást erősíti, hanem többlettel bír. Elsőre talán nem mindenkinek	látszik	mekkora erő van ebben az ügyben, hajrá! Mire jó az RSS? Az RSS
15	al helyettesítették, melyek pillanatnyi¬lag munka- és időtakarékosnak	látszottak	. Az azóta eltelt évek során teljesen mást kellett megtanulnunk. Ma m
16	között valószínűbb. Viszont attól függetlenül, hogy nem profi alvilági,	látszik	rajta, hogy sok tapasztalatot gyűjthetett már szerte a galaxisban és va
17	ak! Ez a férfi viszont biztosan nem veszítette el a szeme világát, ez jól	látszott	azon, hogy egész testével követte Marev minden mozdulatát. - Nem v
18	eglepetést okoztam az idegennek. Ekkor ő is eltette fegyverét, persze	látszott	, hogy teljes mértékben éber maradt és nem hagy előnyhöz jutni tová
19	; egy olyan fához, ami alatt rendszeresen elhaladt ez idáig a járőr. Bár	látszott	a rabszolga-tartókon, hogy titkán tették ki magukat közvetlen veszély
20	a bokron, ahová az ingjét is tette. A zsoldos megtorpant egy pillanatra.	Látszott	rajta, hogy mérlegel pár másodpercig, Marev pedig nem akarta megv

*(1 * Ça se voit pour que j'aille avec eux. (phrase incorrecte)*

2 Avveroès semblait pouvoir réconcilier Aristote avec la religion islamique.

3 En prenant la rue Andrássy, nous arrivons à l'Opéra, et à la fin de la rue bordée de palais, nous apercevons déjà la Place des héros avec le monument du Millénaire.

4 L'émission a été transférée à la chaine Bartók, mais il semble que cette émission « inventée par les communistes » a sa place dans le cimetière, c'est-à-dire sur une chaine régionale.

5 Les trois premières positions ont été déterminées par notre tableau qui montre le nombre de clics sur le bouton « like ».

6 On dirait que les taches jaunes qui « traversent » le bleu, sont vertes.

7 Ils ne deviennent visibles et n'existent qu'en démarrant le sous-programme (= fonction ou processus).

[103] L'outil fonctionne également pour quelques langues, par exemple pour l'allemand.

8 Rien ne laissait prévoir la défaite un mois avant l'élection.

9 Réaliser ce plan semble maintenant possible.

10 Il a été batî sur une colline, on le voit de loin...

11 Chiens en grande quantité. On voit que certains sont conscients de leurs possibilités.

12 Je crois me souvenir que dans la série 1, ils ont montré un document sur lequel on voyait sa date de naissance.

13 Il y avait juste un tout petit bourgeon mais on voyait déjà qu'il allait faire des fleurs magnifiques.

14 Ce n'est peut-être pas clair pour tout le monde à première vue la force qu'a cette affaire. Allez ! À quoi sert le RSS ?

15 Ils ont été remplacés par des nouveaux qui semblaient être plus économes au niveau du temps et du travail.

16 Indépendamment du fait qu'il n'est pas un criminel professionnel, on voit qu'il a déjà collecté pas mal d'expériences dans la galaxie.

17 Cet homme n'a pas perdu sa capacité de voir, ça se voyait de la manière dont il suivait tous les mouvements de Marev.

18 Il a rangé son arme mais ça se voyait qu'il restait vigilant.

19 Bien qu'on voyait que les propriétaires d'esclaves ne s'exposaient pas à un danger immédiat, ...

20 Le soldat s'est arrêté subitement. On voyait qu'il a hésité un instant.)

Tableau 157 : Lignes de concordance avec « látszik » (extrait).

En lisant ces phrases, nous pouvons relever de nombreux problèmes qui les rendent en pratique inutilisables et inappropriées sous leurs formes originales dans le cadre d'un cours de langues. Deux raisons excluent ainsi la première phrase comme un exemple pertinent permettant d'illustrer l'usage de « látszik » : d'une part, l'exemple est tiré d'une page présentant du contenu à caractère sexuel comme indiqué dans son titre et, de l'autre part, la phrase est grammaticalement incorrecte et, pour cette raison, incompréhensible. La deuxième phrase contient des faits historiques et évoque des personnages qui pourront être inconnus des utilisateurs. La troisième phrase est difficile à interpréter et contient une opinion politique, deux arguments que ne la rendent pas éligible dans notre cadre. La sixième phrase est également difficile à interpréter, les guillemets indiquent par ailleurs que le locuteur est conscient que sa phrase ne correspond pas à un usage habituel. La phrase 14 n'est pas claire (par exemple, on ne voit pas pourquoi le locuteur la termine par « Allez ») et la suite contient un acronyme (RSS) qui peut être difficile à interpréter. La phrase 16 est compréhensible mais le contexte (quelqu'un qui a collecté beaucoup d'expérience dans la « galaxie ») peut déstabiliser l'apprenant car, en plus de comprendre les mots dans la phrase, il a besoin de son imagination pour créer un contexte (ici un contexte de science-fiction). La phrase 19 contient l'expression « des propriétaires d'esclaves », il s'agit probablement d'un roman

historique, et le vocabulaire spécifique qui entoure le verbe « látszik » est déconcertant : l'apprenant doit tout d'abord saisir ces mots pour pouvoir interpréter la phrase et ceux-ci font dévier l'attention du verbe vers d'autres éléments. Ces exemples démontrent donc clairement la difficulté de trouver des énoncés susceptibles d'être considérés comme des bons « candidats » pour la simplification dans un contexte pédagogique.

Une fois les phrases appropriées en termes de leur contenu, trouvées, une autre étape doit être franchie : il faut les adapter pour que leur langage soit accessible aux étudiants. Frankenberg-Garcia (2014) suggère d'utiliser « une définition plus des exemples qui contiennent spécifiquement des indices contextuels pour faciliter la compréhension »[104] (p. 139). Nous présenterons ici le procédé d'adaptation en utilisant quelques phrases illustrant les deux usages du verbe « látszik » : (1) X donne une impression par la vue et (2) « úgy látszik » (il semble). La colonne de gauche montre les phrases originales, celle de droite contient les phrases adaptées. Les phrases adaptées sont précédées de quelques mots qui les mettent en contexte (par exemple, « blog de voyage », « forum sur les OVNI » ou « à propos d'un appartement »). Le nombre d'exemples par usage peut refléter la fréquence d'occurrences dans le corpus, la liste fournit ainsi de façon visuelle des informations sur la fréquence (tableau 158).

Phrases originales	**Phrases adaptées**
A városon pedig eleve **látszik**, hogy nem veszélytelen: túl nagy itt minden, óriási a forgalom, na és a trehányság egy kicsit Perura emlékeztet. *On voit tout de suite que la ville n'est pas sans danger : elle est trop grande, il y a beaucoup de circulation, et le laisser-aller me rappelle un peu le Pérou.*	Utazós blog: A városon azonnal **látszik**, hogy nem veszélytelen: túl nagy itt minden és óriási a forgalom. *Blog du voyageur : **On voit** tout de suite que la ville n'est pas sans danger : elle est trop grande et il y a beaucoup de circulation.*
Szerintem első pillantásra **látszik**, hogy ezt az építményt nem ember tervezte.	Fórum ufókról : Szerintem első pillantásra **látszik**, hogy ezt az építményt nem ember tervezte.

[104] A definition plus examples that specifically contain contextual clues to facilitate understanding.

Je pense qu'au premier coup d'œil, on peut voir que cette structure n'a pas été conçue par l'homme.	*Forum sur les OVNIs : Je pense qu'au premier coup d'œil,* ***on peut voir que*** *cette structure n'a pas été conçue par l'homme.*
Helyes kis lakás. **Látszik**, hogy még nincs teljesen belakva. *Joli petit appartement. On voit que les habitants ne se le sont pas encore complétement appropriés.*	Lakásról: Helyes kis lakás. **Látszik**, hogy még nincs teljesen belakva. *À propos d'un appartement : Joli petit appartement.* ***On voit que*** *les habitants ne se le sont pas encore complétement appropriés.*

Tableau 158 : Exemple d'adaptation des phrases avec le verbe « látszik » (se voit).

Nous voyons que les phrases adaptées sont essentiellement des versions raccourcies des phrases originales. Le vocabulaire dans les phrases choisies est suffisamment simple pour ne pas nécessiter de remplacements. Comme évoqué précédemment, l'expression au début de la citation, ne fait pas partie de l'exemple original ; elle a été rajoutée pour donner du contexte à l'énoncé et le rendre plus clair.

Considérons la première phrase afin d'illustrer le procédé d'adaptation. Cette phrase contient une conjonction (« pedig ») qui la connecte à la précédente, il convient donc de la supprimer. De plus, le mot « eleve » (à priori) ne semble pas être parfaitement à sa place (le lecteur a voulu sans doute dire « à première vue » ou « tout de suite ») et il faut l'éliminer. La phrase finit par une opinion personnelle, négative qui peut blesser le lecteur, il est donc également préférable de l'omettre. La phrase finale peut être considérée comme un bon exemple car elle donne une explication simple, facile à suivre qui justifie l'impression sur la ville.

Les autres phrases exemplifient le même usage dans des contextes légèrement différents : la deuxième phrase décrit une photo et la troisième un appartement. En rajoutant encore quelques phrases, les contextes seront suffisamment variés et accessibles pour que l'apprenant puisse se forger une image pertinente de cet usage de « látszik ».

Considérons à présent le cas de l'unité multi-lexicale « úgy látszik » (il semble). Comme dans le cas de « látszik », nous avons adapté des phrases du corpus pour les rendre accessibles aux apprenants. Le contexte est indiqué par quelques mots précédant l'exemple (tableau 159).

Phrases originales	Phrases adaptées
Úgy látszik, nehezen születnek itt bejegyzések, csak akkor törik meg a jég, ha örömködni kell egy-két húzósabb győzelmen.... De sebaj! ***Il semble que*** *les entrées se produisent lentement ici, la glace ne se brise que lorsqu'il faut se réjouir d'une ou deux victoires plus difficilement obtenues.... Mais ce n'est pas grave.*	Fórum a sportról: **Úgy látszik,** nehezen születnek itt bejegyzések, de sebaj. *Forum sur le sport:* ***Il semble que*** *les entrées se produisent lentement ici, mais ce n'est pas grave.*
Az emberi észnek, **úgy látszik,** az a természete, hogy mindent meg akar magyarázni. *La nature de la raison humaine* ***semble être de*** *vouloir tout expliquer.*	Vélemény: **Úgy látszik,** az emberi észnek az a természete, hogy mindent meg akar magyarázni. *Opinion : La nature de la raison humaine* ***semble*** *être de vouloir tout expliquer.*
De jó buli, be is telefonál egy-két emberke, **úgy látszik**, tetszik nekik a műsor. *C'est cool, une ou deux personnes appellent,* ***elles ont l'air d****'aimer l'émission.*	Vélemény rádióműsorról: Betelefonál egy-két ember, **úgy látszik**, tetszik nekik a műsor. *Avis sur une émission de radio : Une ou deux personnes appellent, elles* ***semblent*** *aimer l'émission.*
Ha mindenképpen szakállt szeretne viselni (amit, **úgy látszik,** a bőre rosszul tolerál), próbálja meg az Alksebor nevű, recept nélkül is kapható készítményt. *Si vous voulez absolument porter une barbe (que votre peau* ***semble*** *mal tolérer), essayez un produit en vente libre appelé Alksebor.*	Probléma leírása egy fórumon: Szakállt szeretne viselni, de **úgy látszik,** hogy a bőre ezt rosszul tolerálja. *Description d'un problème sur un forum : Vous voulez absolument porter une barbe que votre peau* ***semble*** *mal tolérer.*

Tableau 159 : Exemples d'adaptation des phrases avec l'unité multi-lexicale « úgy látszik » (il semble que).

Les phrases originales sont relatives à plusieurs contextes de la vie quotidienne : les locuteurs parlent de la participation aux forums, de la nature humaine, d'une émission de radio et de la possibilité de faire pousser une barbe. Comme dans le premier cas, les phrases adaptées gardent leur contexte mais elles sont plus courtes que les phrases originales. Les parties susceptibles de faire dévier l'attention de l'apprenant de l'élément à observer (les expressions argotiques, les

informations inutiles car trop spécifiques comme le nom du produit dans la dernière phrase) ont été éliminées mais, mis à part ce changement, le vocabulaire n'a pas été modifié.

Une fois ce travail de simplification réalisée, l'enseignant peut proposer les phrases adaptées de plusieurs manières :

- Il peut les pré-catégoriser comme nous l'avons fait dans les tableaux du haut, auquel cas l'apprenant n'a pas d'autre tâche que d'étudier les exemples dans les différents groupes, d'observer l'environnement textuel et de marquer les expressions qu'il souhaite mémoriser. L'analyse fournira le nom de la catégorie et, éventuellement, sa traduction[105]. Par exemple, le titre pour le premier groupe présenté peut être « quelque chose est visible, se voit au premier coup d'œil ».
- Il peut également regrouper les phrases, lister les catégories d'usage séparément et demander à l'apprenant d'attribuer les noms de catégories aux exemples[106].
- L'enseignant peut fournir les noms des catégories ainsi que la liste des phrases et demander aux apprenants de regrouper celles-ci. Cet exercice demande un travail analytique susceptible d'aider la mémorisation (Boers 2021 : 14, 136-138).

2.2) Prendre ses exemples dans un corpus pédagogique

Les exemples ci-dessus montrent que la sélection et l'adaptation des énoncés d'un corpus non pédagogique sont des tâches complexes et que créer une collection d'exemples avec l'élément choisi peut prendre un temps significatif. Avoir accès à un corpus pédagogique suffisamment important peut faciliter ce travail car il permettra de trouver des phrases au niveau de compétences de l'apprenant susceptibles d'être incluses dans la collection d'exemples sans modification. Le tableau suivant montre des occurrences avec « látszik » dans le corpus pédagogique du hongrois compilé par l'auteure de cette thèse (voir la Partie III pour une description détaillée) (tableau 160).

[105] Nous préconisons la traduction uniquement dans les cas où elle s'applique à toutes les phrases.
[106] Voir aussi les activités proposées dans la Partie III de cette thèse.

1	5. 1. Rossz kondícióban vagy. Sportolj többet! 2. Fáradtnak	látszol	. Pihenj egy kicsit! 3. Nagyon kevés vizet iszol. Igyál többet!
2	y jobban néz ki a szoba. - Aha. Nagyobbnak, tágasabbnak	látszik	. 9. Átlapoztam az újságot, de semmi érdekeset nem találta
3	nincs kapacitása. Dubar elméletét több tény is bizonyítani	látszik	. Egy-egy törzsi falu lakóinak száma általában 150 körül mo
4	am a tornacipőmet. - Már megint? 3. Olyan rosszkedvűnek	látszol	. - Mert az vagyok. Nem sikerült a vizsgám. 4. Talpraesett e
5	egészen. Nekem lenne egy kérdésem. 7. Olyan idegesnek	látszol	! - Ne is mondd, teljesen elegem van mindenből. 8. Úgy láto
6	ztáé viszont túl nagy. Az övét a lépcsőn kell felvinni. 7. Úgy	látszik	, a telefonom nem kapcsolódik az internethez. A tietek gono
7	olt. / Mi több: kabátja épen sárga volt, / És igy annál jobban	látszott	a folt. / "Eldobnám - szólt - de mással nem birok;" / Ez ok / I
8	em bántam. Elég volt rájuk nézni, különösen Mehmet Alira,	látszott	rajta, hogy nagyon boldog, hogy a vendégei lehetünk. És m
9	ak a meleg éghajlatot kedveli, de ez most erősen megdőlni	látszott	. A táj szépsége és a tudat, hogy Kínában biciklizünk, kiseb
10	izakodást - / elégedettséget fejezett ki. (= valamilyen érzés	látszik	rajta) 2. Kifejezi az együttérzését / a sajnálatát / a részvétét
11	nem akarsz változtatni. 6. Jól bírod az éjszakázást. Nem is	látszik	rajtad, hogy fáradt vagy. B2_MF9/6. (még ... is) 1. Még Mar

(1 Tu ***as l'air*** *fatigué. Repose-toi un peu.*

2 À mon avis, la chambre est mieux comme ça. – Oui, elle ***a l'air*** *plus spacieuse.*

3 Plusieurs faits ***semblent*** *soutenir la théorie de Dubar.*

4 Tu ***sembles*** *être de mauvaise humeur. – Parce que je suis de mauvaise humeur. J'ai raté mon examen.*

5 Tu ***sembles*** *être irrité. – Oh oui ! J'en ai marre de tout.*

6 Mon portable ne ***semble*** *pas pouvoir se connecter à l'Internet.*

7 (Extrait de poème) Puisque sa veste était jaune, la tache ***se voyait*** *d'autant plus.*

8 Il suffisait de les regarder, surtout Mehmet Ali pour ***voir*** *qu'il était très heureux de pouvoir nous accueillir chez lui.*

9 Nous pensions jusque-là que les chameaux n'aimaient que les climats chauds mais cette théorie ***semblait*** *s'effondrer maintenant.*

10 Explication du sens d'un mot : un sentiment se voit sur le visage de quelqu'un

11 Tu tiens bien le coup quand tu ne dors pas la nuit. On ne ***voit*** *même pas que tu es fatigué.)*

Tableau 160 : Exemples du corpus pédagogique dans le Concordancier.

Comme la taille du corpus pédagogique est plutôt limitée, nous ne trouvons pas un grand nombre d'exemples correspondant à l'ensemble des schémas d'usage. Nous n'y trouvons ainsi que onze occurrences avec le verbe « látszik ». Ce nombre ne suffit pas pour illustrer tous les environnements textuels dans lesquels le mot peut typiquement émerger. Il est donc nécessaire de compléter la collection en tirant d'autres exemples des grands corpus non pédagogiques et de les simplifier selon les méthodes que nous avons décrites dans la section précédente. Si le corpus pédagogique était de taille significative, cette étape ne serait bien sûr pas nécessaire, mais nous ne disposons pas actuellement d'un tel corpus pour le hongrois. Un corpus constitué d'exemples accessibles à l'apprenant lui permettrait par ailleurs d'effectuer des recherches en autonomie, sans l'aide de l'enseignant.

2.3) Créer des phrases parallèles

Les phrases tirées du corpus peuvent servir de modèles pour d'autres phrases créées par l'enseignant lorsqu'un usage particulièrement important mérite d'être approfondi. Par exemple, il peut produire quelques phrases analogues à la première phrase présentée dans le tableau 158 s'il souhaite étudier de plus près l'usage « valami látszik valamin/valakin » (qqch se voit à qqch/qqn) : « A városon azonnal **látszik**, hogy nem veszélytelen: túl nagy itt minden és óriási a forgalom. » (On voit tout de suite que la ville n'est pas sans danger : elle est trop grande et il y a beaucoup de circulation.) Les phrases analogues reprennent la même structure avec un vocabulaire que l'apprenant connaît : « A városon **látszik**, hogy sok az egyetemista: mindenhol kávézók, a teraszokon szinte csak fiatalok ülnek. » (Il y a beaucoup d'étudiants universitaires dans la ville : il y a des cafés partout, presque seulement des jeunes sont assis aux terrasses.) « A falun **látszik**, hogy a turizmusból él: szinte minden házban van kiadó szoba. » (Le village semble vivre du tourisme : il y a des chambres à louer dans presque toutes les maisons.) Ces phrases semi-authentiques contiennent du vocabulaire accessible et pertinent tout en exemplifiant un usage authentique et typique. Créer des phrases parallèles est particulièrement bénéfique aux niveaux A1 et A2 où les connaissances linguistiques des apprenants sont encore limitées et le vocabulaire important doit être répété de manière aussi variée que possible pour faciliter la mémorisation.

3) En cours : Analyser les exemples, observer les répétitions et les variations

L'apprenant observe un grand nombre d'exemples dans le but de se forger une image de l'utilisation de l'élément donné. L'importance de la phase d'observation a été soulignée par plusieurs chercheurs, par exemple par Schmidt (1990), fondateur de l'hypothèse d'observation (noticing hypothesis). Selon cette théorie, il faut que l'apprenant ait des opportunités de remarquer et d'observer consciemment des éléments langagiers qui, ensuite, peuvent être retenus. Sans la phase d'observation consciente, la mémorisation ne pourra pas avoir lieu. Frankenberg-García (2014 : 130) cite des études soutenant l'hypothèse de Schmidt (Leow 2000 ; Izumi 2002 ; Mackey 2006 et d'autres) en nous rappelant qu'en réalité, il ne s'agit pas que d'une seule opération cognitive mais de deux : l'apprenant doit (1) relever les phénomènes linguistiques mais aussi (2) les traduire en informations interprétables (*intake*) afin de les assimiler. Nation (2013 : 51) note par ailleurs que « [l]es connaissances contextuelles comme la collocation nécessitent une grande exposition à la langue. Il n'est pas indispensable pour un usage réceptif mais nécessaire à la production » (notre traduction). Prenons comme exemple du processus d'analyse le verbe « eljön » dans le sens où il indique l'arrivée d'un moment dans le temps. Le mode de présentation proposé dans le tableau ci-dessous attire l'attention de l'apprenant sur ces deux aspects et lui permet de les étudier à travers

un nombre suffisant d'exemples. Étudier des phrases systématisées pour y révéler des instances de répétition et de variation linguistique oriente le regard de l'apprenant vers des caractéristiques importantes de l'usage (Nation 2013 : 50-53) (tableau 161).

eljön (arriver + sujet = un moment dans le temps)	
1 « eljön X » (« X arrive », X = un moment dans le temps)	
2 « eljön X, hogy/amikor » (« X arrive où/quand », X = un moment dans le temps)	
1	*1*
Hamar **eljött** a december.	*Décembre* ***est venu*** *vite.*
Eljött a kapcsolati marketing kora.	***L'ère du marketing relationnel est arrivée.***
Eljött a videótechnika ideje.	***Le temps de la technologie vidéo est arrivé.***
Eljön az utolsó Rockmaraton!	***Le dernier Rock Marathon arrive!***
És **eljött** a várva várt délután.	*Et* ***vint l'après-midi tant attendu.***
Teljesen kimerült, mire **eljött** a nagy nap.	*Il était complètement épuisé quand* ***le grand jour est arrivé.***
Hamar **eljött** a túra vége.	*Bientôt,* ***la randonnée prit fin.***
2	***2***
Mindenki életében **eljön** az a pillanat, hogy költözne vagy költöznie kell.	***Vient un moment*** *dans la vie de chacun où il bougera ou devrait bouger.*
Egyszer csak **eljön** az a pillanat, hogy a természet ébredezni kezd.	*Tout à coup* ***vient le moment*** *où la nature commence à se réveiller.*
És **eljött** a nap, amikor már csak egy dologra lett volna szükség.	*Et* ***le jour est venu*** *où une seule chose aurait été nécessaire.*
Végre **eljött** az idő, hogy jobban megmutassuk magunkat.	***Le moment est enfin venu*** *pour nous de devenir plus visibles.*

Tableau 161 : Exemples à analyser par l'apprenant.

Comment s'expriment la variété et la répétition dans les exemples présentés ci-dessus ? Les sujets de ces phrases restent systématiquement au sein du même champ lexical, exprimant « un moment dans le temps » (élément de répétition) mais ils montrent une certaine variété lexicale (élément de variation). L'ordre des mots est le même dans toutes les phrases et le nombre des conjonctions possibles dans les phrases composées reste très limité (éléments de répétition) mais la deuxième partie des phrases composées varie (élément de variation).

En quoi cette démarche, nous est-elle utile ? Il suffit de rappeler que nous avons vu dans les chapitres 9 et 10 que les synonymes, sans environnement textuel, peuvent sembler très similaires à première vue. Or, l'environnement textuel révèle qu'il est possible d'attribuer à chaque synonyme des caractéristiques d'usage qui le rendent unique, non ambigu et non interchangeable.

L'étude des phrases présentées dans le tableau 161 permet ainsi à l'apprenant, comme évoqué plus haut, de prendre conscience que *l'exposition à l'élément donné par un grand nombre d'exemples peut contribuer à approfondir son usage. L'accent est donc mis sur l'observation et la construction d'une expérience linguistique permettant à l'apprenant de développer les compétences nécessaires non seulement pour reconnaître mais aussi pour utiliser correctement l'élément en question*.

C) Présenter les unités multi-lexicales et les schémas d'usage

La recherche en linguistique de corpus démontre que les unités multi-lexicales ainsi que les schémas d'usage semblent être au cœur de l'usage langagier, comme nous l'avons vu au chapitre 4. En effet, la connaissance d'un grand nombre d'unités multi-lexicales et des schémas permet aux apprenants de s'exprimer d'une façon qui les rapproche, au cours de l'apprentissage, de plus en plus des natifs. Présenter les unités multi-lexicales et les schémas à part des phrases-exemples est donc préconisé pour que l'apprenant puisse se concentrer sur leur apprentissage (cf. Nation 2013 : 479-513 ; Boers et al. 2006 ; Boers et al. 2017).

1) Réduire l'unité multi-lexicale à ses composantes essentielles

Dans les unités multi-lexicales présentées en cours de langues, *l'environnement du mot choisi doit être tout d'abord réduit à ses composantes essentielles afin que l'apprenant puisse se concentrer sur celles-ci*. La présentation par catégorie (éventuellement avec une traduction) peut faciliter la compréhension des unités choisies. Plusieurs chercheurs présentent des arguments convaincants en faveur d'un apprentissage conjoint des mots et de leurs unités multi-lexicales typiques afin de proposer une meilleure image de leurs sens authentiques[107] (Cheng et al. 2008 ; Lewis 1993 ; Sinclair 2004b ; Stubbs 2009 et autres).

Dans son ouvrage sur l'apprentissage du vocabulaire, Nation (2013 : 480) donne une description détaillée des propriétés des unités multi-lexicales dont nous reprenons ici les quatre points qui nous semblent particulièrement pertinents dans le cadre pédagogique :

[107] Sinclair (2004) démontre, par exemple, que le sens le plus courant du verbe anglais « see » est « comprendre » et non « voir (avec les yeux) ».

- *La plupart des unités multi-lexicales sont de nature variable*, bien qu'elles aient ce que Sinclair (2004b) appelle une forme « canonique ». Cela signifie que les membres de l'unité peuvent se retrouver plus ou moins proches l'un de l'autre. La morphologie des membres est également susceptible de changer.
- *Dans les unités multi-lexicales, les mots individuels ne sont pas combinés arbitrairement, mais se rassemblent de manière cohérente.* Même si certaines unités multi-lexicales sont plus que la somme de leurs composants, les significations des composants contribuent à l'ensemble. La compréhension du sens des parties facilite l'apprentissage et la rétention du sens de l'unité multi-lexicale (cf. Boers et al. 2017 ; Boers et Lindstromberg 2009 ; Liu 2010 ; Martinez et Schmitt 2012 ; Walker 2011).
- *De nombreuses unités multi-lexicales sont probablement stockées sous forme de choix uniques* (cf. Hoey 2005 ; Sinclair 2004 ; Siyanova-Chanturia et al. 2011), mais *ce stockage ne signifie pas nécessairement qu'elles ne sont pas analysées ou sont de forme fixe ou invariable*.
- Tout comme les mots simples sont utilisés à des fins de communication, *les unités multi-lexicales ont également des fins de communication*. Cette finalité communicative est une partie importante de ce qu'implique la connaissance de l'unité multi-lexicale.

Les outils tels que « Word Sketch » sur Sketch Engine permettent de produire des listes de collocations en incluant des informations sur la fréquence de chacune d'elles[108]. L'enseignant peut concevoir cette liste pour l'élément étudié et choisir les unités multi-lexicales les plus pertinentes pour ses apprenants. Ce travail est donc moins chronophage que la sélection et l'adaptation des phrases-exemples car il suffit de supprimer les unités multi-lexicales jugées peu judicieuses dans le cadre pédagogique.

La quantité et la complexité des unités multi-lexicales dépendent du niveau de l'apprenant et la tâche de sélection revient à l'enseignant qui peut catégoriser les unités d'après leurs propriétés sémantiques afin de présenter leurs usages habituels de manière claire et transparente. La fréquence d'usage joue un rôle important dans le processus de sélection car les niveaux de compétences linguistiques inférieurs doivent connaître les unités multi-lexicales courantes avant de passer à des unités moins usitées. Dans le tableau 162 ci-dessous, les couleurs et/ou les tailles différentes de lettres indiquant la fréquence de l'expression permettent une meilleure visualisation des

[108] Voir le chapitre 3 dans la Partie I pour la description plus détaillée de cet outil.

collocations : la taille des lettres indique à l'apprenant l'importance des collocations, les différentes couleurs (aide à la mémorisation) signalent quels mots appartiennent à la même catégorie. Pour illustrer ce mode de visualisation, nous avons choisi les collocations de l'adjectif « nehéz + N » (nom + « lourd, difficile ») dans le tableau suivant.

« nehéz » + N

Tâche, devoir difficile	
nehéz feladat	***tâche difficile***
nehéz dolog	***chose difficile***
nehéz kérdés	***question difficile***
nehéz munka	***travail difficile***
nehéz döntés	*décision difficile*
nehéz ügy	*affaire difficile*

Temps, période difficile	
nehéz időszak	***période difficile***
nehéz napok	***jours difficiles***
nehéz nap	***journée difficile***
nehéz idők	***temps difficiles***
nehéz pillanatok	*moments difficiles*

Circonstances difficiles	
nehéz helyzet	***situation difficile***
nehéz körülmények	***circonstances difficiles***

Autre	
nehéz téma	*sujet difficile*
nehéz ember	*homme difficile*
nehéz ételek	*plats lourds*

Tableau 162 : Visualisation possible des unités multi-lexicales avec l'adjectif « nehéz ».

Le tableau sert, avant tout, à visualiser le lexique en présentant une vingtaine de noms fréquemment associés à l'adjectif « nehéz », mais il contient aussi quelques informations grammaticales : il

indique notamment l'ordre des mots dans l'unité (l'adjectif précède le nom) et montre quels noms sont utilisés au singulier et lesquels sont utilisés au pluriel.

Les informations statistiques sur la fréquence des unités multi-lexicales permettent à l'enseignant de hiérarchiser les usages mais, contrairement aux listes produites dans les chapitres 8 à 11, le tableau ci-dessus n'inclut pas le nombre des occurrences par élément. À notre avis, une exposition visuelle de la fréquence telle que proposée peut être plus « digeste » pour l'apprenant qu'une liste de chiffres, et la visualisation peut également faciliter la mémorisation[109]. Comme le dit O'Keeffe et al. (2007) :

> « De toute évidence, pour l'apprenant d'une deuxième langue ou d'une langue étrangère, l'apprentissage des collocations de cette langue n'est pas un luxe si une maîtrise supérieure à un niveau de survie est souhaitée, car la collocation imprègne même les mots les plus élémentaires et les plus fréquents. » (p. 60, notre traduction)

Il est également possible de cartographier les verbes les plus fréquents accompagnant ces unités multi-lexicales. Il convient cependant d'examiner comment présenter à l'apprenant l'ensemble des informations issues de l'analyse effectuée dans le chapitre 8. Pour répondre à cette question, nous proposons de créer un tableau indiquant les unités les plus utiles aux niveaux A2-B1. Notre sélection a été motivée par les facteurs suivants : (1) la récurrence d'usage de ces unités dans le langage non spécifique, (2) par la simplicité de leurs constituants (la majorité des mots appartenant au vocabulaire de base) et (3) par le fait que ces unités multi-lexicales différentes s'inscrivent dans le même schéma (tableau 163).

Ez nehéz + N	*C'est un N diffcile*
Ez **nehéz kérdés**.	*C'est* ***une question difficile***.
Ez **nehéz ügy**.	*C'est* ***une affaire difficile***.
Ez **nehéz feladat**.	*C'est* ***une tâche difficile***.
Ez **nehéz döntés**.	*C'est* ***une décision difficile***.
Ez **nehéz téma**.	*C'est* ***un sujet difficile***.

[109] L'enseignant peut bien évidemment choisir d'inclure des chiffres exacts dans la présentation si cela lui semble approprié.

nehéz + N *ez(ek) (temps)*	*être un N difficile*
Nehéz idők ezek.	*Ce sont* ***des temps difficiles***.
Nehéz pillanatok ezek.	*Ce sont* ***des moments difficiles***.
Nehéz időszak ez.	*C'est* ***une période difficile.***
nehéz N+Poss *van/volt/lesz*	*avoir un N difficile*
Nehéz dolgom van.	*Je suis confronté à* ***une tâche difficile***.
Nehéz napom volt/lesz.	*J'ai eu/J'aurai* ***une journée difficile***.
Nehéz természete van.	*Il a* ***un caractère difficile***.
nehéz N *előtt állunk*	*se trouver devant un N difficile*
Nehéz nap előtt állunk.	*Nous nous trouvons devant* ***une journée difficile***.
Nehéz döntés előtt állunk.	*Nous nous trouvons devant* ***une décision difficile***.
Nehéz feladat előtt állunk.	*Nous nous trouvons devant* ***une tâche difficile***.
Nehéz helyzet előtt állunk.	*Nous nous trouvons devant* ***une situation difficile***.
nehéz N *jön/jönnek* (*temps*)	*s'annonce un N difficile*
Nehéz napok jönnek.	***Des jours difficiles*** *s'annoncent.*
Nehéz idők jönnek.	***Un temps difficile*** *s'annonce.*
Nehéz időszak jön.	***Une période difficile*** *s'annonce.*
nehéz N *él* (*temps*)	*vivre un N difficile*
Nehéz napokat élünk.	*Nous vivons* ***des jours difficiles***.
Nehéz időket élünk.	*Nous vivons* ***un temps difficile***.
Nehéz időszakot élünk.	*Nous vivons* ***une période difficile.***
nehéz N-*On vagyunk túl* (*temps*)	*traverser un N difficile*
Nehéz időszakon vagyunk túl.	*Nous avons traversé* ***une période difficile***.
Nehéz napokon vagyunk túl.	*Nous avons traversé* ***des jours difficiles***.
nehéz körülmények között + V	*dans les circonstances difficiles* + V
nehéz körülmények között él	*il vit* ***dans des circonstances difficiles***
nehéz körülmények között nőtt fel	*il a grandi / passé son enfance* ***dans des circonstances difficiles***
nehéz körülmények között kellett *dolgoznia, helyt állnia, újrakezdenie az életét*	*il a dû travailler/s'établir/recommencer sa vie* ***dans des circonstances difficiles***

nehéz körülmények között is + V	*malgré les circonstances difficiles* + V
nehéz körülmények között is sikeres lehet	*il peut réussir* ***même dans des / malgré les circonstances difficiles***
nehéz körülmények között is kitartóan dolgozik	*il travaille assidûment* ***même dans des / malgré les circonstances difficiles***
nehéz körülmények között is lehet eredményt elérni	*on peut obtenir des résultats* ***même dans des / malgré les circonstances difficiles***
X nehéz helyzetbe hozza Y-t	*X met Y dans une situation difficile*
A döntés **nehéz helyzetbe** hozta a céget.	*La décision a mis la companie* ***dans une situation difficile***.
A döntés **nehéz helyzetbe** hozhatja a cég vezetőjét, ha ...	*La décision peut mettre le directeur de la companie* ***dans une situation difficile*** *si ...*
nehéz helyzetbe kerül X	*Y se retrouve dans une situation difficile*
Nehéz helyzetbe került a cég.	*La companie s'est retrouvée* ***dans une situation difficile***.
Nehéz helyzetbe kerülhet a cég vezetője, ha ...	*Le directeur de la companie peut se retrouver* ***dans une situation difficile*** *si …*
nehéz helyzetben van/volt X	*Y est dans une situation difficile*
Nehéz helyzetben van a cég.	*La companie est* ***dans une situation difficile***.
Nehéz helyzetben volt a cég vezetője.	*Le directeur de la companie était* ***dans une situation difficile***.

Tableau 163 : Liste proposée dans le cadre pédagogique présentant des unités multi-lexicales fréquentes avec l'adjectif « nehéz ».

Cette liste peut sembler longue à première vue mais cette longueur s'explique par le fait qu'elle présente plusieurs exemples concernant la même unité multi-lexicale ; il ne s'agit donc pas d'usages différents dans chaque exemple. Les verbes accompagnant l'unité « ADJ + N » sensibilisent l'apprenant à la variation lexicale tout en montrant que leur nombre (du moins celui des variations fréquentes) est limité.

2) Faire découvrir les schémas d'usage

Dans les chapitres 4, 8 à 11, nous avons exposé les tableaux de schémas d'usage de plusieurs éléments langagiers à partir du travail de Sinclair (1991, 2003, 2004a et 2004b) et de Hoey (2005). Nous reprenons ci-dessous l'un de ces résultats pour rappeler les informations qu'un tel tableau peut fournir à l'apprenant (tableau 164).

« megjön » 2 (une personne est de retour)

Collocations typiques	**Sujets** : personne ou groupe de personnes **Modificateurs** : « végre » (enfin), « amióta » (depuis), « amikor » (quand) et d'autres mots indiquant un moment dans le temps
Colligations typiques, ordre des mots	(1) Verbe au passé : **(MOD « végre », « amikor » +) megjött + X** (2) Verbe au présent : **MOD (« hatra », « amikor » etc.) + megjön + X**
Composantes sémantiques	Indique le retour d'une personne, perçue par le locuteur comme un moment positif.
Composantes pragmatiques	Annonce le retour d'une personne (souvent par la personne elle-même), généralement suivi d'une réaction de l'interlocuteur

Tableau 164 : Schémas d'usage que l'apprenant peut établir.

Dans le cadre pédagogique, ces tableaux n'ont, en eux-mêmes, qu'une fonction relativement limitée : *ils sont utiles pour attirer l'attention de l'apprenant sur le fait qu'il existe des schémas et il y a un intérêt à rechercher les schémas. Cependant, ils ne sauraient remplacer la partie la plus enrichissante du travail, notamment l'expérience linguistique que l'apprenant construit lors de l'étude de l'élément choisi à travers les exemples*. Leur rôle le plus important est en effet de faire apparaître l'interconnexion des différents types de schémas.

3) Travail actif sur les trois modes de présentation

Les trois modes de présentation correspondent à trois étapes d'analyse ; il est cependant important de noter qu'elles ne doivent pas se suivre nécessairement dans le même ordre car elles ne sont pas organisées de manière hiérarchique. Comme nous l'avons évoqué lors de la description de la technique de « Zoom in, zoom out » dans la partie A, les phrases-exemples ne sont pas destinées à être les premières à être présentées à l'apprenant, et les tableaux des schémas les derniers à être mis en évidence. L'enseignant peut ainsi décider de suggérer un autre parcours en commençant par l'exposé des collocations, par exemple. En étudiant les modes de présentation en autonomie, l'apprenant a par ailleurs la possibilité de passer d'une modalité à l'autre à sa convenance. Il peut notamment commencer ses explorations par l'étude des unités multi-lexicales, porter par la suite son attention sur des exemples, puis étudier les schémas et relire les unités multi-lexicales et/ou

les phrases, et ainsi de suite. Il pourra aussi retourner plus tard à des modes de présentation déjà étudiés, son regard enrichi d'informations obtenues depuis la première analyse.

L'enseignant peut également proposer des tâches qui rendent plus actif le travail sur les trois modes de présentation. Voici quelques exercices qui engagent les apprenants dans l'identification des schémas et dans l'analyse des exemples :

- L'enseignant pré-catégorise les phrases et/ou les unités multi-lexicales. Les apprenants analysent les phrases ou les unités, seuls ou en petits groupes, et essaient de trouver la catégorie qui les relie. L'enseignant peut fournir les noms des catégories ou demander aux apprenants de les trouver.
- L'enseignant peut proposer des phrases dans lesquelles le collocatif (ou un des collocatifs) de l'élément étudié manque(nt). Les apprenants complètent les phrases en s'appuyant sur le co-texte. Ils peuvent effectuer la même tâche pour les unités multi-lexicales, basées sur leur traduction.
- L'enseignant fournit les catégories d'usage avec quelques exemples ; le reste des phrases ou des unités multi-lexicales doit être complété par les apprenants.
- Les apprenants doivent compléter certains éléments (par exemple, les collocatifs fréquents ou les colligations) dans les tableaux de schémas[110].

4) Les principes de base pour des modes de présentation efficaces

Les points ci-dessous résument les principes de base pour des modes de présentation efficaces :

- Pour les niveaux de langue inférieurs, nous proposons trois modes de présentation, correspondant à trois étapes analytiques dont l'ordre est à déterminer par l'enseignant et l'apprenant. Ces éléments sont (1) les phrases tirées du corpus qui montrent l'élément étudié dans des environnements textuels similaires mais présentant des variations linguistiques, (2) les unités multi-lexicales réduisent l'usage à observer à ses éléments essentiels et (3) les tableaux de schémas offrent une vue d'ensemble des résultats de l'analyse.
- Dans le cadre pédagogique, l'adaptation des énoncés authentiques se justifie par le niveau linguistique de l'apprenant. Les exemples doivent lui être accessibles et pertinents et

[110] D'autres activités ciblant la pratique avec l'élément étudié seront proposées au chapitre 15 de la Partie III.

doivent cependant garder leur caractère naturel, ce qui nécessite un travail significatif de la part de l'enseignant.

- Catégoriser les phrases servant d'exemples et les unités multi-lexicales par usage permet à l'apprenant d'interpréter plus facilement les résultats. Il peut observer non seulement des exemples avec le mot-clé mais aussi des schémas sous-jacents (des répétitions et des variations) si les occurrences sont ordonnées par catégories. Un avantage supplémentaire dans ce processus est que l'apprenant reçoit un nombre d'exemples nettement supérieur à ce que n'importe quel dictionnaire pourrait proposer.
- Les phrases-exemples ainsi que les unités multi-lexicales peuvent être hiérarchisées par la fréquence d'usage. La taille des lettres et/ou le nombre d'exemples fournis par usage peuvent être proportionnels à la fréquence. Ainsi, l'apprenant est à même d'observer au premier coup d'œil les expressions les plus courantes, donc les plus pertinentes, qu'il doit prioriser dans son apprentissage.
- La présentation doit inclure des informations lexicales, grammaticales, sémantiques et pragmatiques. Indiquer l'ordre des mots, les temps du verbe, la terminaison typique du complément, le message positif ou négatif associé avec un usage, sa fonction dans le discours, et ainsi de suite, tout ceci contribue à fournir une image plus complète de l'usage de l'élément en question. Ainsi, l'apprenant acquiert non seulement des informations propres à l'usage mais aussi une nouvelle manière de voir la langue : il prend conscience de l'interconnexion des différents aspects linguistiques, ce qui pourra l'aider à interpréter plus facilement de nouveaux énoncés.

Dans ce chapitre, nous avons avancé trois modes de présentation des éléments langagiers dont l'usage pose problème aux apprenants : phrases-exemples adaptées à partir des lignes de concordance, unités multi-lexicales catégorisées, tableaux de schémas d'usage. Ces outils permettent à l'apprenant d'obtenir des informations complémentaires sur l'usage de l'élément qu'il considère comme difficile à saisir. Nous avons également proposé une méthode pour adapter les phrases-exemples ainsi que quelques exercices pour le travail actif dans le cadre du cours de langues. L'apprenant qui acquiert ces techniques d'analyse à un niveau de compétences linguistiques inférieur pourra les utiliser tout au long de son parcours. En avançant de niveau à niveau, cela lui permettra de trouver des réponses à certaines de ses questions de manière autonome, sans l'aide de l'enseignant. L'enseignant, de l'autre côté, enrichit sa boîte-à-outils de nouveaux outils qui lui permettent de fournir des renseignements sur l'usage langagier à partir des

observations plutôt qu'à partir de son intuition. Ces outils lui offrent également la possibilité d'imaginer et de créer des types d'exercices novateurs.

Résumé de la Partie II

Explorer les corpus dans le cadre pédagogique : l'enseignant comme acteur

Dans cette partie, nous avons exploré les possibilités de l'utilisation des corpus dans le cadre pédagogique, nous concentrant sur les cas où l'enseignant peut fournir des réponses précises et complètes à l'aide des grandes bases de données linguistiques. Nous avons constaté que la description plus formelle de certains aspects de la langue hongroise pourrait également bénéficier des résultats d'analyse de corpus. Nous avons étudié des mots à usages multiples, les synonymes et les deux conjugaisons, particularité de la langue hongroise. Les résultats de ce travail ont révélé l'interconnexion forte entre les caractéristiques lexicales, grammaticales, sémantiques et pragmatiques des aspects langagiers étudiés.

Quels avantages ?

L'utilisation des corpus offre de toute évidence des avantages certains pour l'enseignement des langues : les corpus fournissent une matière linguistique authentique ainsi que des outils d'analyse statistique. Plutôt que d'inventer des exemples et d'improviser une explication basée sur l'introspection, l'enseignant peut élucider les points langagiers qui échappent à des règles bien précises et cohérentes, en consultant des corpus. Les observations collectées lors de ce processus d'analyse permettent de *valider, compléter et préciser les informations dans les dictionnaires et dans les grammaires pédagogiques ainsi que d'augmenter le nombre d'exemples relatif à l'élément linguistique choisi* basés sur des usages réels et authentiques.

Les principaux avantages du travail avec les corpus dans le cadre pédagogique peuvent se résumer comme suit :

- En étudiant un grand nombre d'exemples systématisés et en observant les contextes d'utilisation les plus fréquents, *l'apprenant peut accéder à une meilleure compréhension de l'usage du mot. Il a l'opportunité de retenir des informations concernant les propriétés grammaticales, sémantiques et pragmatiques des unités multi-lexicales fréquentes* avec le mot-clé, ce qui contribue à augmenter l'efficacité de l'apprentissage. Il peut également se rendre compte que les différentes

composantes de la langue sont interconnectées et que chaque usage a ses caractéristiques particulières qui le rendent unique, non ambigu et non interchangeable.

- Au cours de l'observation de l'élément choisi, *l'apprenant obtient des informations sur l'ensemble des éléments faisant partie des énoncés.* Par l'élargissement graduel de l'environnement textuel, par les répétitions et les variations dans les différentes phases, l'observation ne se limite pas au mot-clé. Même si celui-ci reste tout au long au centre des explorations, l'apprenant s'expose également à d'autres éléments du langage.
- Assimiler des usages fréquents identifiés dans le corpus permet à l'apprenant d'observer et d'acquérir des éléments linguistiques qui le rendront capable de développer des compétences linguistiques nécessaires qu'exigent les interactions avec les natifs. Les schémas observés lui permettent également de créer de nouveaux énoncés et d'utiliser en conséquence la langue de façon créative.
- Un avantage indirect de l'utilisation des corpus est qu'elle modifie favorablement la façon dont les étudiants perçoivent la langue : des phénomènes qui étaient auparavant « impossibles à apprendre » deviennent catégorisables et, par conséquent, abordables.

Les résultats de l'analyse vont bien au-delà d'une simple étude au cas par cas. En partant d'un seul élément, nous progresserons vers des unités multi-lexicales ainsi que vers des schémas d'usage qui donnent un caractère naturel à l'usage langagier.

Quelles limites ?

Nous ne pouvons pas prétendre que l'analyse de corpus et la présentation des résultats n'offrent que des avantages. S'il en était ainsi, tous les créateurs de matériels pédagogiques utiliseraient aujourd'hui des corpus, ce qui n'est pas le cas. Comme tout instrument dans la boîte-à-outils de l'enseignant, celui-ci a aussi ses limites.

- Il est clair que toute question linguistique ne se prête pas à être explorée à l'aide des grandes bases de données linguistiques. Par exemple, l'étude des éléments langagiers suivant des règles claires ne nécessite pas de consultation de corpus.
- Les grands corpus à fins linguistiques ne se prêtent pas automatiquement à l'utilisation dans le cadre pédagogique, car ils ne comprennent pas toujours de matériel adéquat pour les niveaux de compétences inférieurs. Cela nécessite une sélection et une adaptation soigneuses de la matière linguistique à présenter, tout en conservant l'équilibre délicat entre accessibilité et authenticité.

- Même si le nombre des unités multi-lexicales typiques avec l'élément étudié est limité, chacune d'elles peut être explorée plus en détail. Puisque ces approfondissements fournissent des résultats de plus en plus précis et que chaque analyse révèle de nouvelles informations, la tentation de continuer les explorations et de transmettre aux apprenants toutes les informations tirées du corpus, est grande. Néanmoins, il ne faut pas oublier qu'en fournissant une quantité d'informations trop abondante, nous n'aidons pas l'apprenant. Une sélection et une adaptation basées sur des considérations pédagogiques sont nécessaires pour rendre les données lisibles et utiles aux apprenants.
- L'analyse et la systématisation des résultats ne permettent la généralisation que dans une mesure limitée. Même le corpus le plus grand ne contient pas toutes les occurrences jamais produites. Nous pourrons donc identifier des usages typiques de l'élément choisi sans jamais prétendre à l'exhaustivité et l'enseignant doit être conscient de cette limite.
- Les réponses fournies par le corpus seront plus détaillées et plus précises que les réponses que nous pourrions obtenir en utilisant notre intuition et introspection d'expert mais elles seront aussi plus complexes. Les « traduire » dans le cadre pédagogique demande un effort non négligeable de la part de l'enseignant. En même temps, les réponses basées sur le corpus aident les étudiants non pas en leur fournissant nécessairement des « règles » mais en leur donnant accès à une grande collection de phrases-exemples systématisées ainsi qu'en leur permettant d'observer les caractéristiques d'usages typiques.

Nous tenons à évoquer un autre point (dont l'analyse approfondie dépasserait le cadre de ce travail de recherche) qui influence significativement le « destin » de tous les outils proposés dans le contexte pédagogique : l'attitude de l'enseignant. Comme pour d'autres ressources, c'est à l'enseignant de décider si et comment les corpus seront utilisés. Or, le choix des outils, l'étude des collocations et des lignes de concordance, l'analyse statistique des résultats, leur systématisation et interprétation demandent de l'expertise et, avant tout, du temps. Ce travail très significatif que demandent l'analyse et la didactisation des résultats est sans doute l'une des raisons pour lesquelles l'utilisation des corpus ne s'est pas davantage répandue dans l'enseignement (Cavalla et Loiseau 2013). Certaines études montrent également que les enseignants ne se sentent pas tous aptes à lire les lignes de concordance, d'y identifier les schémas et de proposer les résultats de façon claire et cohérente[111] (Boulton 2010a, 2017 ; Szita a) à paraître). Ceci est un autre facteur qui peut contribuer à décourager l'enseignant d'intégrer l'analyse outillée du langage-cible dans ses cours.

[111] Les raisons de cette attitude sont diverses, les facteurs les plus importants étant : les enseignants ne sont pas sûrs d'être capables d'analyser les énoncés correctement, ils ne se sentent pas à l'aise avec les outils

Pour remédier aux difficultés liées à la position des enseignants, nous aurions besoin d'une meilleure intégration des résultats de l'analyse de corpus dans les ouvrages pédagogiques, pour le hongrois et pour d'autres langues. Cela permettrait à l'enseignant de s'appuyer sur du matériel existant s'il ne se sent pas apte à effectuer des recherches par lui-même. Les exemples proposés dans les ouvrages pédagogiques peuvent également servir de modèles qui montrent des manières possibles d'utiliser les corpus ; ils peuvent donc guider l'enseignant dans ses explorations de corpus[112].

En outre, la création de corpus pédagogiques permettant d'accéder à des exemples au niveau de compétences linguistiques de l'apprenant, serait également souhaitable pour aider le travail de l'enseignant et pour permettre à l'apprenant d'explorer ce corpus de façon autonome. Grâce à ce travail, il pourrait trouver des réponses à ses questions sans l'aide de l'enseignant, s'exposer à la langue, gagner plus d'expérience linguistique et se forger une image plus précise de l'usage de l'élément étudié.

Ces corpus, nécessaires pour un travail en autonomie, sont encore trop peu nombreux et de faible taille ; par conséquent, les phrases tirées de ces collections ne reflètent pas tous les schémas typiques d'usage et ne fournissent pas suffisamment d'exemples pour ceux inclus dans le corpus. Il est possible de compléter ces phrases par d'autres sources provenant de grands corpus non pédagogiques, de préférence associées aux situations que les apprenants doivent maîtriser à leur niveau de compétences. Néanmoins, la solution idéale semblerait être la création de corpus pédagogiques de plus grande taille. La Partie III de cette thèse sera consacrée aux considérations relatives à la création de ce type de corpus pédagogiques.

numériques, il leur manque du temps pour faire des recherches qui amènent à des résultats fiables, ils ne voient pas la plus-value de l'utilisation des corpus (Szita à paraître 2022a).

[112] Voir les chapitres 5 et 6 pour l'analyse de quelques ouvrages existants.

PARTIE III : Les corpus au service des apprenants

Introduction à la Partie III

La Partie II de cette thèse a été réservée aux possibilités offertes aux enseignants par les grands corpus à fins linguistiques. À cette fin, nous avons étudié certaines questions fréquemment posées dans les cours de hongrois. Celles-ci portent, avant tout, sur des aspects linguistiques dont la particularité commune est qu'ils ne se laissent pas définir par des règles simples. Nous avons montré que l'utilisation des corpus peut aider l'enseignant à fournir des réponses plus fiables et à apporter davantage de précisions sur l'usage langagier concernant l'élément choisi. Nous avons également proposé quelques idées pour la présentation des résultats dans le cadre des cours de langues.

Dans la troisième et dernière partie de la thèse, nous nous intéresserons aux avantages que la consultation de corpus peut apporter aux apprenants. Pour cela, nous proposerons des corpus pédagogiques susceptibles d'être utilisés soit de façon autonome, soit dans le cadre d'un cours. Les chapitres 13 et 14 traiteront respectivement des corpus écrits et des corpus oraux construits pour compléter la série de manuels « MagyarOK », exposée brièvement au chapitre 6. La présentation de ces corpus se concentrera sur le processus de leur composition, sur leurs bienfaits ainsi que sur les problèmes rencontrés en cours de création. Dans ces chapitres, nous exposerons également l'intérêt de rendre les textes inclus dans le corpus accessibles dans leur intégralité.

La méthode de travail s'appuie sur les résultats de la linguistique de corpus, notamment sur le constat que les natifs utilisent leur langue de façon similaire quand il s'agit de participer à des interactions analogues[113]. Ainsi, l'usage langagier des natifs est loin d'être toujours original : ils construisent leurs produits langagiers en utilisant des éléments et des schémas récurrents. Travailler avec les textes dans nos corpus écrit et oral permet d'attirer l'attention de l'apprenant sur cette particularité de l'usage langagier, car la lecture attentive fera émerger le grand nombre de répétitions et de variations linguistiques dans le même texte et dans plusieurs textes. Nous pourrions dire que lors de l'étude de ces textes, l'apprenant utilise les méthodes de la linguistique de corpus — sans outils numériques. La validité de cette approche sera mise en évidence par l'exploration de différents sous-ensembles des corpus.

Le chapitre 15 sera consacré au travail avec les corpus présentés en utilisant des outils numériques et proposera des activités pour les explorer afin d'augmenter l'efficacité de l'apprentissage du

[113] Voir la section C) du chapitre 4.

hongrois. Les activités sélectionnées s'articuleront autour de plusieurs axes : elles viseront à sensibiliser l'apprenant à l'existence des unités multi-lexicales dans sa langue maternelle et dans la langue-cible et elles lui permettront d'observer et de pratiquer des phénomènes linguistiques choisis. De nombreuses activités reprennent l'idée de l'importance de la répétition et de la variation dans l'usage langagier ; d'autres exercices auront pour objectif à apprendre aux étudiants comment identifier des schémas et construire le profil de mots comme présenté dans les sections précédentes. Une première partie de ces activités a été conçue dans le but d'enrichir le cours de langues et implique la participation de l'enseignant. L'autre partie peut être effectuée par l'apprenant de manière autonome, sans l'aide de l'enseignant.

Les activités ainsi proposées remplissent plusieurs objectifs pédagogiques. Elles servent, d'une part, à développer un nouveau regard sur la langue-cible en soulignant l'interconnexion de ses différentes composantes. Elles cherchent par ailleurs à enrichir les activités « classiques » comme la révision du vocabulaire ou l'expression écrite et orale. Ces activités préparent également l'apprenant au travail avec des corpus à des niveaux plus avancés en les introduisant à la manipulation de grandes collections de données linguistiques et à l'usage des outils numériques pour leur exploitation.

Chapitre 13 : Corpus écrits au service des apprenants aux niveaux de compétences linguistiques inférieurs

Ce chapitre présente un corpus écrit qui comprend et élargit le contenu linguistique de « MagyarOK », série de quatre manuels (niveaux A1-B2) pour le hongrois, rédigée par l'auteure de cette thèse et Katalin Pelcz. Nous nous concentrerons uniquement sur les niveaux de compétences inférieurs (A1-B1) et ne discuterons pas des corpus pour le niveau B2. Nous décrirons en premier lieu le processus de collecte des données en mettant l'accent sur les points suivants :

- Comment générer du langage pour les corpus pédagogiques écrits ?
- Comment collecter du langage pour ces corpus ?
- Comment adapter les énoncés réels au niveau de l'apprenant ?
- Dans quelle mesures les textes sur le même thème contiennent-ils des répétitions et des variations linguistiques ? Comment peut-on les faire émerger ?

Comme nous le montrerons dans ce chapitre, la construction d'un corpus à fins pédagogiques pour les niveaux A1-B1 va bien au-delà de la sélection de textes réels. L'édition de ces textes et la création de nouveaux textes sont également nécessaires pour que le corpus puisse présenter *un langage à caractère naturel (authentique ou semi-authentique), accessible et pertinent pour l'apprenant*. Le résultat final de ce travail est un corpus qui comprend un ensemble de textes adaptés suivant certains critères que nous exposerons dans ce chapitre. Ce compromis semble nécessaire pour que la collection finale puisse remplir son objectif, comme nous le verrons dans les pages suivantes[114].

A) Les sous-ensembles du corpus

1) Aperçu général

Il est généralement admis dans la littérature que la lecture extensive est susceptible de faciliter l'apprentissage des langues (par exemple, Alan et Widdowson 1974 ; Borsos 2014 ; Ellis R. 1985 ; Krashen 1989 ; Nation 2013 ; Nation et Waring 1997 ; Robinson et Ellis 2008 ; Robb et Susser 1989 ; Webb 2005, 2007 ; Webb et Chang 2014). Un apport linguistique significatif sous la forme de textes écrits, riches et variés peut aider les apprenants non seulement à améliorer leur compréhension écrite en général mais aussi à prendre conscience des caractéristiques spécifiques

[114] Il convient de noter que de nombreux éléments récurrents dans le corpus pédagogique apparaissent également avec une fréquence significative dans les grands corpus généraux et peuvent ainsi représenter un usage langagier plutôt typique.

de la langue-cible. Ils peuvent observer des unités de plusieurs mots dans divers textes et développer une « intuition » (feeling) de leur utilisation (Gabrielatos 2005).

Le corpus pédagogique peut fournir du matériel au service ces explorations. Dans le cas de « MagyarOK », le corpus pédagogique écrit a été fondé sur un corpus authentique que nous avons compilé en amont de la production de la série de manuels. Ce corpus initial contient environ 15 millions de tokens de données linguistiques authentiques liées aux sujets et aux types de textes pour les niveaux A1-B1 du CECRL. La majorité des textes provient de l'Internet. Pour les interactions écrites formelles et semi-formelles, nous avons puisé dans notre correspondance électronique personnelle et dans celle d'autres locuteurs natifs se portant volontaires pour partager certains textes (anonymisés) avec nous.

Une fois les manuels terminés, *nous avons compilé un corpus plus restreint pour les compléter avec du matériel accessible au niveau de compétences linguistiques des apprenants*. Ce matériel s'appuie sur le corpus initial de données authentiques, bien que de nombreux textes aient été édités et raccourcis. *À cette collection ont été rajoutés des textes semi-authentiques, rédigés par des natifs sur les thèmes des manuels*. Le corpus final comprend donc des textes informés par le corpus, semi-authentiques et authentiques qui constituent les trois sous-ensembles suivants :

- Le sous-corpus 1 comprend tous les textes des manuels « MagyarOK A1-B1 » et contient environ 150 000 tokens.
- Le sous-corpus 2 regroupe des récits semi-authentiques écrits par des locuteurs natifs sur les sujets traités dans les manuels. Ce sous-ensemble compte actuellement 61 000 tokens.
- Le sous-corpus 3 se compose de textes authentiques de différents genres (articles de journaux, entrées de blogs et de forums) et de ressources personnelles telles que des courriels et des interactions sur les médias sociaux. Comme le sous-corpus 2, il comprend des textes dont les sujets sont pertinents pour les niveaux de compétences linguistiques inférieurs. Ce sous-ensemble contient environ 500 000 tokens.

Cette collection a pour but d'initier les apprenants à l'utilisation des corpus, leur permettant de construire progressivement leurs compétences analytiques et de reconnaître des avantages de la consultation d'un corpus pour trouver des réponses à leurs questions. En outre, le matériel linguistique dans le corpus peut servir aussi de *modèle* lors de la rédaction de leurs propres textes, comme le suggèrent Kennedy et Miceli (2010, 2017).

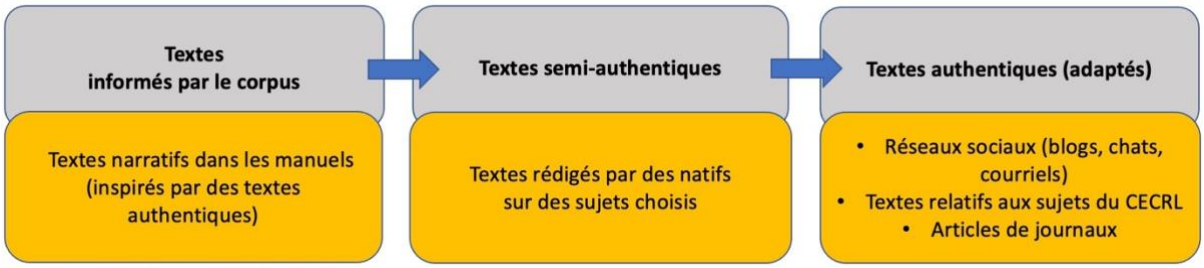


Tableau 165 : Sous-ensembles écrits pour les niveaux A1-B1.

Le corpus écrit contient actuellement environ 710 000 tokens et est continuellement enrichi. Les textes sont catégorisés et annotés en fonction du sujet et de leur degré de formalité. Les e-mails et les clavardages (*chats*) sont également annotés en fonction de l'âge et du genre des participants, du nombre de participants, de leurs relations (par exemple, mère et fille, amis intimes, collègues) et nous indiquons également si le texte est un fragment (juste un e-mail ou une partie d'un e-mail) ou s'il est complet (un échange d'e-mails incluant tous les messages). Outre la consultation des ensembles de données du corpus ouvert « MagyarOK » sur Sketch Engine, les apprenants peuvent lire de nombreux textes dans leur intégralité sur le site Web des manuels.

Il convient de remarquer qu'il est de même possible de compiler des corpus pédagogiques à partir d'autres types de textes. Allan (2009) recommande, par exemple, l'utilisation des ouvrages littéraires adaptés (*graded readers*) ; le principal atout de ces corpus étant leur accessibilité pour les niveaux inférieurs. Allan a créé un corpus à partir des livres adaptés (*graded readers*) pour les niveaux B1-B2, édités par « Penguin », totalisant environ 1,5 million de mots. Pour mesurer la représentativité de ce corpus, elle a mené une étude sur les unités multi-lexicales en comparant leur fréquence dans ces textes avec celle du British National Corpus (BNC). Tout en reconnaissant l'existence de certaines disparités, elle affirme que « [i]l semble que les corpus classés offrent un équilibre raisonnable entre accessibilité et authenticité des données qu'ils fournissent » (Allan 2009 : 30). Nous avons décidé de ne pas inclure des textes littéraires adaptés dans l'intérêt d'assurer le caractère systématique du contenu linguistique du corpus.

2) Sous-ensemble (1) : matériel linguistique des manuels informés par le corpus

2.1) Contenu

Le manuel peut être la première source à explorer avec les outils de corpus pour les niveaux de compétences linguistiques inférieurs. Si le livre est informé par le corpus, il peut préserver un degré

significatif d'authenticité, même aux niveaux inférieurs (Szita 2014). En reproduisant un langage à caractère naturel, il peut servir de source fiable pour les observations linguistiques.

Les manuels présentent des caractéristiques spécifiques qui doivent être prises en compte lors de la conception du corpus. Dans leur forme initiale, ils contiennent des exercices et des textes avec des blancs, des instructions, des numéros d'exercices. Dans l'intérêt d'une bonne lisibilité, ces éléments doivent être supprimés, complétés ou édités avant la construction effective du corpus. Pour le corpus « MagyarOK », les mesures suivantes ont été prises :

- Tout le contenu des manuels et des cahiers d'exercices *ainsi que les transcriptions des enregistrements audio et vidéo ont été inclus dans le corpus*.
- Les doublons ont été supprimés (par exemple, lorsque le même texte était d'abord proposé dans son intégralité, puis comme exercice à trous) afin d'éviter les biais de fréquence.
- Pour la même raison, les instructions ont également été supprimées. Ces instructions ont été systématiquement formulées de la même façon pour faciliter leur compréhension, ce qui a inévitablement entraîné un grand nombre de répétitions au sein du corpus.
- Les phrases des exercices à trous ont été complétées.
- Les numéros des exercices et des textes ont été conservés pour indiquer où ils se trouvent dans le manuel.
- Chaque fois qu'un exercice était composé de courts dialogues, ces dialogues ont été numérotés pour indiquer les phrases appartenant au même exercice.
- Dans le cas des exercices de grammaire, la grammaire abordée est indiquée pour créer une cohésion entre les phrases listées, facilitant ainsi le repérage.
- Les métadonnées suivantes ont été incluses : niveau de compétence, thème, numéro de chapitre et d'exercice, type d'exercice et, dans le cas des transcriptions, une note signalant que le texte est basé sur un enregistrement audio/vidéo.
- Les utilisateurs peuvent rechercher le corpus par niveau, nom de fichier, numéro de chapitre et type de données (textes de manuels, textes supplémentaires, transcriptions d'entretiens, etc.).

Le corpus du manuel contient environ 150 000 tokens au total et a été téléchargé sous forme de « corpus ouvert » sur le site web de Sketch Engine où il peut être consulté en libre accès[115].

[115] https://app.sketchengine.eu/#dashboard?corpname=preloaded%2Fmagyarok

2.2) Utilité

La suggestion de construire des corpus pédagogiques basés sur les textes utilisés en classe a été envisagée dans la littérature depuis un certain temps (par exemple, Aston 2001 ; Braun 2006 ; Charles 2014 ; Flowerdew J. 2009 ; Flowerdew L. 2009 ; Hunston 2002 ; Tyne 2012 ; Timmis 2015 ; Meunier et Reppen 2015 ; Weber 2001 ; Widdowson 2003 ; Willis 2003). Timmis (2015 : 3), citant Willis, suggère notamment qu'un corpus à fins pédagogiques peut être « constitué des *textes déjà utilisés par les apprenants en classe*, *qui sont ensuite exploités pour l'étude de caractéristiques linguistiques particulières*. L'avantage de tels corpus [...] est que les apprenants seront déjà familiarisés avec le co-texte, c'est-à-dire l'environnement textuel de la caractéristique-cible, puisqu'ils auront préalablement étudié le texte entier en classe » (notre traduction, nous soulignons). Ceci est d'autant plus important que les outils d'analyse de corpus ne présentent pas les textes dans leur intégralité. Hunston (2002 : 16) s'accorde également à dire qu'un corpus pédagogique pourrait contenir « toute matière linguistique à laquelle l'apprenant a été exposé ». Cet ensemble de données présente les avantages-clés suivants : il est « représentatif des besoins de l'apprenant » (Aston 2001) ; il est de taille gérable et comporte du matériel pertinent et linguistiquement accessible pour l'apprenant. Charles (2014), Tyne (2012) et Weber (2001), entre autres, soulignent aussi qu'« une approche « descendante » (*top down*), du texte au corpus, peut être extrêmement utile pour permettre aux apprenants de travailler avec les corpus » (Tyne 2012 : 144). Comme le formule Reppen (2016 : 411) :

> Un corpus pédagogique spécialisé est un outil précieux [qui], utilisé de manière appropriée, fournit une ressource puissante pour le développement de matériel et d'activités pour les enseignants et pour les activités pratiques pour les apprenants » (notre traduction).

Le principal avantage d'un tel corpus est qu'il peut être introduit dès le début du processus d'apprentissage. D'après notre expérience, les apprenants reconnaissent plus facilement la valeur ajoutée des corpus lorsqu'ils ont eu l'occasion de travailler avec eux dès le début, et ce pour une bonne raison. S'ils ont atteint le niveau B2 ou C1 de la langue-cible sans jamais consulter de corpus, pourquoi changeraient-ils leurs habitudes alors qu'ils peuvent déjà communiquer sans difficulté ? Ainsi, les corpus peuvent avoir un impact plus important sur l'apprentissage des langues lorsqu'ils sont intégrés dès le départ dans la boîte à outils de l'apprenant. Pour cela, des ensembles de données appropriés sont cependant nécessaires.

3) Sous-ensemble (2) : récits semi-authentiques

3.1) Contenu

L'objectif principal du sous-ensemble 2 est de fournir *un apport linguistique utile, semi-authentique, à caractère naturel.* Les récits des locuteurs natifs qui y sont inclus traitent des sujets des manuels. Les textes revoient, restructurent et complètent l'apport linguistique proposé dans le matériel de cours. Dix enseignants de langue maternelle ont été chargés de créer des textes pour les niveaux A1 et A2 et dix locuteurs natifs non enseignants ont rédigé ceux du niveau B1. Nous avons considéré qu'il était important que les rédacteurs des niveaux A1 et A2 connaissent le vocabulaire et les aspects grammaticaux traités dans les manuels ; c'est la raison pour laquelle nous avons demandé à des enseignants travaillant avec les livres de cours « MagyarOK » de rédiger ces textes. Comme ils connaissaient bien le matériel pédagogique, ils ont pu *intégrer un maximum d'éléments couverts par le livre dans leurs récits afin de montrer comment ils peuvent être utilisés de façon naturelle*. De nombreuses unités multi-lexicales importantes apparaissent plusieurs fois dans les récits et ces rencontres répétées contribuent à leur consolidation. Les textes ont été inclus dans le corpus sous leur forme originale, sans aucune modification.

La sélection des contributeurs affecte la qualité des produits finaux. Les enseignants et les utilisateurs experts, non enseignants, doivent bien comprendre ce que l'on attend d'eux, sinon les textes qu'ils produisent risquent de ne pas convenir au corpus (Frankenberg-García 2012). Les enseignants qui ont tendance à réduire leur langue à une version artificielle lorsqu'ils s'adressent à leurs élèves (comme ce peut être le cas pour les langues considérées comme « difficiles », telles que le hongrois), ne sont pas des contributeurs idéaux pour le corpus pédagogique. D'autre part, les contributeurs non-enseignants peuvent craindre de « ne pas écrire ce qu'il faut » et de « ne pas utiliser le bon langage », faite de maitriser exactement ce que les apprenants sont censés savoir à un niveau donné[116]. Notre sélection finale de contributeurs non-enseignants comprenait des professionnels de la langue (linguistes, écrivains, critiques littéraires, étudiants universitaires) à même de nous fournir les textes souhaités[117]. Ils ont rédigé des textes à caractère naturel (parfois même ludique et humoristique) pour le niveau B1, intéressants non seulement pour leur langue et leur style mais aussi pour leur contenu.

[116] Communications personnelles des enseignants.

[117] Les participants ont été instruits du sujet et des objectifs attendus avant l'écriture.

Ce sous-corpus contient environ 61 000 tokens au total. Le niveau A1 comprend 80 textes d'environ 150 mots (dix textes pour chacun des huit chapitres), l'ensemble du jeu de données totalisant 12 000 mots. Le corpus A2 compte 80 textes de 180 à 250 mots, soit environ 16 000 tokens. Enfin, le corpus B1 comporte 12 chapitres et les textes sont de 250 à 300 mots, ce qui représente 33 600 mots. Dans tous les cas, les textes sont suffisamment courts pour que les apprenants puissent les lire dans leur intégralité[118].

3.2) Utilité de ce corpus

Il est clair que *ces textes ne représentent pas tout ce qu'on peut dire du sujet en question ; ils montrent seulement la façon dont le langage du manuel peut être réutilisé de manière naturelle*. Néanmoins, ils remplissent plusieurs fonctions pédagogiques essentielles. Premièrement, *la lecture des textes permet aux apprenants d'élargir et de consolider le vocabulaire de base lié à un sujet donné, grâce à des rencontres répétées*. Deuxièmement, ces textes sont d'une grande utilité pour la production écrite, car *ils montrent tout le spectre de ce que les apprenants sont déjà capables de dire et servent de modèles pour leurs propres récits*. Ils présentent différentes manières dont les éléments lexicaux (mots et unités multi-lexicales) connus de l'apprenant peuvent être assemblés pour former des récits cohérents. La possibilité de consulter un tel corpus peut ainsi contribuer à améliorer la qualité linguistique des récits des apprenants (Boers 2021 : Kennedy et Miceli 2010, 2017 ; Szita 2020).

Considérons les textes suivants (tableaux 166-168), écrits par trois enseignantes natives pour compléter le chapitre 4 (« L'endroit où j'habite ») du manuel A1. Jusqu'à 95 % du vocabulaire est tiré du manuel ; les éléments-clés contribuant à la cohérence et au caractère naturel des textes sont indiqués en gras. Les expressions reprises du contenu du manuel sont soulignées.

Texte 1

Hosszúhetényben lakom. Ez egy kis falu Dél-Magyarországon, kb. 3000 ember él itt. Én egy szép új házban lakom a férjemmel és a kislányommal. A kislányom 13 éves, és iskolába jár. Hosszúhetényben van egy nagyon jó és **viszonylag nagy** iskola. Körülbelül 320 gyerek jár ide.

A falu csendes és tiszta. Gyönyörű a környék és nagyon jó a közbiztonság. Nekem **ez a két dolog nagyon fontos**. **Igaz, hogy** Hosszúhetényben nincsenek híres múzeumok, de van egy tájház **és persze** vannak gyönyörű erdők. **Nagyon**

[118] Une partie de ces textes est déjà accessible sur le site du manuel (magyar-ok.hu), d'autres y seront rendus disponibles en 2022. Le corpus ouvert sur Sketch Engine contient tous les textes.

szeretek itt túrázni. **Az is jó, hogy** közel van Pécs: autóval 20 perc, de busszal is lehet utazni. A buszok **elég gyakran** járnak. Pécsett minden van: vannak jó színházak, mozik, kocsmák, kávézók, sok jó étterem és múzeumok.

Hosszúhetény **elég gazdag falu, ahol** magas az életszínvonal. Vannak kulturális és sportolási lehetőségek, **lehet például** karatézni, kézilabdázni, focizni **és persze** kirándulni az erdőben. Van egy érdekes sportesemény is, a Talicskaolimpia. **Ez nem vicc**: a Talicskaolimpia **tényleg** létezik! Májusban van, és más országokból, például Horvátországból, Németországból és Olaszországból is jönnek emberek.

(J'habite à Hosszúhetény. C'est un petit village dans le sud de la Hongrie, environ 3000 personnes y vivent. Je vis dans une belle maison neuve avec mon mari et ma petite fille. Ma fille a 13 ans et va à l'école. Il y a une très bonne et ***relativement grande*** *école à Hosszúhetény. Environ 320 enfants y vont.*

Le village est calme et propre. La région est magnifique et la sécurité publique est très bonne. ***Ces deux choses sont très importantes pour*** *moi. Il est* ***vrai qu****'il n'y a pas de musées célèbres à Hosszúhetény, mais il y a un musée de traditions populaires* ***et bien sûr*** *de belles forêts.* ***J'aime vraiment*** *faire des randonnées ici.* ***Une autre chose qui est bien est que*** *Pécs soit proche : 20 minutes en voiture, mais vous pouvez aussi prendre le bus. Les bus circulent* ***assez souvent.*** *Pécs a tout : il y a de bons théâtres, des cinémas, des pubs, des cafés, de nombreux bons restaurants et des musées.*

Hosszúhetény est un ***village assez riche avec un*** *niveau de vie élevé. Il existe des possibilités culturelles et sportives,* ***comme le*** *karaté, le handball, le football* ***et, bien sûr, la*** *randonnée dans la forêt. Il y a aussi un événement sportif intéressant, les Olympiades de brouette.* ***Ce n'est pas une blague*** *: les Olympiades de brouette existent* ***vraiment*** *! C'est en mai, et des gens d'autres pays comme la Croatie, l'Allemagne et l'Italie viennent.)*

Texte 2

Én Ausztriában, Bécsben élek. Sok barátom van itt, és itt is dolgozom. Bécs nagyon szimpatikus város, szeretek itt élni. Nagyon sok történelmi emlék és szép épület van a belvárosban. Körülbelül kétmillió ember él Bécsben. **Előny, hogy** jó a közbiztonság, magas az életszínvonal, és vannak kulturális és sportolási lehetőségek is.

Egy szép kerületben, egy csendes utcában lakunk. A környék nyugodt és tiszta. Van a közelben bolt, óvoda, iskola, kórház, gyógyszertár, posta és játszótér **is**. Szeretünk színházba és koncertre menni. A közlekedés jó, praktikus és olcsó, a metró és a

villamos is tiszta. A városban nem járok autóval. Busszal, rollerrel és gyalog közlekedem.

Hétvégén sok család kirándul. Sok park, játszótér van a városban, **és** a bécsi erdő sincs messze.

(Je vis à Vienne, en Autriche. J'y ai beaucoup d'amis et c'est là que je travaille. Vienne est une ville très agréable, j'aime y vivre. Il y a beacoup d'histoire et de beaux bâtiments au centre-ville. Environ deux millions de personnes vivent à Vienne. ***Elle a l'avantage d****'une bonne sécurité publique, d'un niveau de vie élevé et de possibilités culturelles et sportives.*

Nous vivons dans un quartier agréable, dans une rue tranquille. Le quartier est calme et propre. Il y a un magasin, un jardin d'enfants, une école, un hôpital, une pharmacie, un bureau de poste et une aire de jeux à proximité. Nous aimons aller au théâtre et aux concerts. Les transports sont bons, pratiques et peu chers, le métro et le tramway sont propres. Je ne conduis pas en ville. Je me déplace en bus, en scooter et à pied.

De nombreuses familles partent en voyage le week-end. La ville compte de nombreux parcs et terrains de jeux, ***et*** *les bois de Vienne ne sont pas loin non plus.)*

Texte 3

Most a szülővárosomról, Kaposvárról mesélek, **ami** a Dél-Dunántúl központja. Nagyon szeretem a várost, **mert** most is itt élnek a szüleim és a nagymamáim. Egy családi házban laknak a város szélén, van egy nagy kertjük. Kaposvár nem nagy város, hatvanezer ember lakik itt. A város gyorsan fejlődik. Híres a színház, van egyetem, modern kórház, néhány múzeum, kávézók és új zeneház is. A szüleimmel minden nyáron elmegyünk a komolyzenei fesztiválra.

A sétálóutca és a főtér szép és barátságos, rengeteg a szökőkút. **Szerintem** a tömegközlekedés nem rossz, sok új busz jár a városban. Lehet sportolni is, az öcsém rendszeresen fut a sportpályán a barátaival. Van egy nagy uszoda is. Szép a környék, nagyon szeretek a dombokon és az erdőkben kirándulni.

(Je vais maintenant vous parler de ma ville natale, Kaposvár, ***qui*** *est le centre de la Transdanubie du Sud. J'aime beaucoup la ville* ***car*** *mes parents et mes grands-mères y vivent encore. Ils vivent dans une maison familiale à la périphérie de la ville et ont un grand jardin. Kaposvár n'est pas une grande ville, elle a 60 000 habitants. La ville se développe rapidement. Elle possède un théâtre célèbre, une université, un hôpital moderne, quelques musées, des cafés et un nouveau music-hall. Mes parents et moi allons au festival de musique classique chaque été.*

La rue piétonne et la place principale sont agréables et conviviales, avec de nombreuses fontaines. Je pense que les transports publics ne sont pas mauvais, il y a beaucoup de nouveaux bus en ville. On peut aussi faire du sport, mon frère court régulièrement sur le terrain de sport avec ses amis. Il y a même une grande piscine. C'est une belle région, j'aime beaucoup les randonnées dans les collines et les forêts.)

Tableaux 166-168 : Trois textes-modèles sur le thème « L'endroit où j'habite ».

Le fait que les noms des bâtiments, des événements et des éléments d'infrastructure, entre autres, se répètent dans plusieurs textes (puisque tous les locuteurs ont utilisé le vocabulaire du chapitre), est susceptible de faciliter la mémorisation. Malgré ces éléments récurrents, l'organisation des textes, la manière dont la cohésion textuelle est assurée et la façon dont les éléments d'information sont reliés, ainsi que la façon dont les locuteurs formulent leur point de vue personnel, rendent chaque contribution différente. Les apprenants peuvent également observer que la cohésion textuelle est assurée par l'organisation des arguments même s'il n'y a pas d'élément lexical mais la logique implicite est claire. Par exemple : « Bécs nagyon szimpatikus város, szeretek itt élni. » (Vienne est une ville très agréable, j'aime y vivre.) « Persze vannak gyönyörű erdők. Nagyon szeretek itt túrázni. » (Il y a bien sûr de belles forêts. J'aime vraiment faire des randonnées ici.) « Lehet sportolni is, az öcsém rendszeresen fut a sportpályán a barátaival. » (On peut aussi faire du sport, mon frère court régulièrement sur le terrain de sport avec ses amis.)

Que certaines expressions utiles, non thématiques reviennent plus d'une fois permet à l'apprenant de les reconnaître (par analyse manuelle ou outillée) et d'en intégrer quelques-unes dans ses propres textes. Dans le cas du sujet « L'endroit où j'habite », ces éléments expriment souvent le point de vue du locuteur comme le montre le tableau suivant. Les unités multi-lexicales listées apparaissent dans au moins deux textes différents (tableau 169).

Az a jó, hogy ...	*C'est bien/bon que...*
Igaz, hogy ...	*C'est vrai que...*
Csak az a baj, hogy …	*Le seul problème est que...*
Előny/Hátrány, hogy ...	*Un avantage/désavantage est que...*
Ez tényleg/elég/egészen + ADJ	*C'est vraiment/assez/entièrement + ADJ*
szerintem	*A mon avis, je pense que*
Lehet például + V*inf*	*Par exemple, vous pouvez + Vinf*
sajnos	*Malheureusement*

Tableau 169 : Éléments exprimant des points de vue personnels dans au moins deux textes sur le thème « L'endroit où je vis ».

Une recherche dans le corpus « huTenTen12 » confirme que ces expressions sont également fréquentes dans ce grand corpus général. L'intégration de ces éléments dans les énoncés des apprenants peut donc contribuer à augmenter le caractère naturel de leur usage langagier et rapproche leur mode d'expression de celui des natifs.

4) Sous-ensemble (3) : récits et interactions édités

4.1) Le contenu

Selon les descripteurs du CECRL[119], les apprenants doivent être capables d'écrire des messages simples, des courriels et des entrées de forum aux niveaux A1-B1. Ils doivent également être aptes à comprendre des informations de base dans une brochure de musée (thème de l'exposition, prix d'un ticket, tarifs réduits), sur la page Internet d'un service (départ et arrivée des trains, horaires d'ouverture d'un magasin ou d'un restaurant, météo, etc.). Ils sont censés pouvoir donner des informations sur eux-mêmes (loisirs, routine quotidienne, voyages, par exemple) et comprendre des textes simples, relatifs à ces sujets. En outre, ils doivent pouvoir participer à des interactions écrites (principalement des messages courts pour prendre un rendez-vous, s'excuser pour un retard, etc.).

Comme les sujets de la vie quotidienne sont discutés en permanence sur Internet, il peut sembler facile de collecter des exemples de ces textes à partir de cette source. Il existe une quantité significative d'échanges dans l'espace virtuel (médias sociaux, blogs, forums) à la portée du constructeur de corpus. Des articles de journaux, des brochures de musées, des critiques de restaurants, en bref : toutes sortes de textes susceptibles de faire partie du corpus peuvent être trouvés sur Internet. Les messages personnels du créateur de corpus peuvent compléter cette collection avec des exemples d'interactions écrites formelles et informelles. Il semblerait donc que nous puissions nous appuyer dans une large mesure sur les sources existantes. Cependant, la difficulté est que, même si de tels corpus peuvent contenir des textes « idéaux » quant à leur choix de sujets, le vocabulaire et les structures grammaticales risquent de dépasser les connaissances de l'apprenant.

Notre principe directeur le plus important lors de la construction de cette partie du corpus a été de *conserver autant que possible la formulation originale et de se contenter de raccourcir les textes et de corriger les fautes d'orthographe. Nous avons décidé d'investir un temps considérable dans la recherche de textes appropriés –*

[119] Pour la liste des descripteurs, voir le chapitre 2.

contenant un vocabulaire de base lié au sujet, avec un langage et un style adaptés – *plutôt que de réécrire des textes* (même si le contenu de certains d'entre eux était vraiment intéressant) *et de compromettre l'authenticité plus que nécessaire*. Les sources les plus utiles pour notre objectif se sont avérées être des entrées de forum et des recensions sur Internet. Ces textes représentent le langage standard (expressions du quotidien, langage semi-formel ou informel) et traitent des sujets du CECRL (description d'hôtels, de villes, d'événements, de services, de restaurants, de films et autres). Nous avons ainsi collecté une très grande quantité de matériel linguistique (plusieurs milliers de recensions sur TripAdvisor et sur d'autres sites d'appréciations) à partir duquel la sélection pouvait être faite relativement facilement. Outre la sélection de sources appropriées, les textes ont été également révisés au niveau de l'orthographe, du style et de la longueur.

Dans les pages suivantes, nous présenterons la méthode utilisée pour éditer les textes existants, en prenant les trois exemples suivants : (1) de courtes contributions sur un forum liées au confinement lors de la pandémie de Covid-19, (2) de courtes interactions (« chats » avant une visioconférence informelle) et (3) des récits plus longs (évaluations de restaurants).

4.2) Contributions sur un forum

Les interactions sur les réseaux sociaux contiennent de grandes quantités de langage authentique lié à la vie quotidienne. Les commentaires, les blogs, les forums, les e-mails et les « chats » sont généralement courts et relativement simples, ce qui en fait des candidats idéaux pour une collection présentant le langage écrit informel. Le corpus des textes de manuels de cours peut être complété par de tels textes car ils figurent parmi les types de textes que les apprenants sont censés lire et écrire aux niveaux A1-B1.

Il convient de noter que, même si leur contenu diffère, les interactions produites dans des contextes similaires tendent à présenter une dynamique comparable. Cela s'applique également aux courtes contributions ci-dessous. La question à explorer par l'apprenant est la suivante : « Comment vivez-vous le confinement ? ». Les exemples ci-après (tableau 170) présentent un sous-ensemble de vingt contributions (sur un total de 52) en réponse à cette question dans un forum en ligne ; les éléments décrivant l'état d'âme des participants sont indiqués en gras :

Hogy bírjátok a bezártságot?
Comment vivez-vous le confinement ?

1 **Nehezen**. / *Difficilement.*

2 Általában 2-3 napig **semmi bajom, ha** itthon vagyok... De **ez már kezd sok lenni**... *D'habitude, je n'ai aucun problème pendant 2-3 jours quand je suis à la maison ... Mais ça commence à être trop ...*

3 **Nehezen**. / *Difficilement.*

4 **Nagyon jól**, de ehhez hozzájárul az, hogy van kertünk.
Très bien, mais le fait que nous ayons un jardin y contribue.

5 **Jól**. / *Bien.*

6 Mindig családban éltem, éppen úgy, mint most, így **nincs alkalmam unatkozni**...
J*'ai toujours vécu en famille, comme maintenant, donc je n'ai pas le temps de m'ennuyer ...*

7 **Van tennivalóm**, családban élek. / *Je m'occupe, je vis dans une famille.*

8 **Nehezen**. Oké, haverokkal lógok telón, de a legjobb barátom szlovák, vele nem lóghatok.
Difficilement. D'accord, je passe mon temps avec des amis au téléphone, mais mon meilleur ami est Slovaque, je ne peux pas traîner avec lui.

9 **Nem valami jól.** Azt vettem észre egyébként, hogy olyanok is felhívnak, írnak, akik amúgy nem.
Pas très bien. Au fait, j'ai remarqué qu'il y a des gens qui m'appellent et m'écrivent, qui sinon ne le font jamais.

10 **Volt már jobb is,** főleg ilyen szép időben, mert nagyon szeretek gyalogolni. Na, és **az sem vidít fel, hogy** egyhamar nem fogok senkit sem látni a családból - **talán ez a legrosszabb**. *C'était déjà mieux, surtout quand il fait aussi beau qu'aujourd'hui, car j'aime beaucoup marcher. Et cela ne me rend pas heureux non plus de ne plus voir personne dans la famille - peut-être que c'est le pire.*

11 **Én váltakozva.** Ha az idő jó, sokkal jobban, de néha vágynék a városba, vagy egy-egy üzletbe.
Ça change. S'il fait beau, beaucoup mieux, mais parfois j'aurais envie d'aller en ville ou dans un magasin.

12 **Nehezen**.
Difficilement.

13 Legutóbb azért kaptam szobafogságot, mert kipróbáltam a cigarettát, kb. 10 éves koromban...
La dernière fois que j'ai été enfermé dans ma chambre, c'est parce que j'ai essayé une cigarette quand j'avais environ 10 ans ...

14 Hétfőtől péntekig dolgozom idehaza is. 8-10 órát ülök a gépnél, mert itthonról nem olyan gyors a rendszer. **Nem unatkozom, van mit csinálnom**. Hétköznap dolgozom, főzök. Hétvégén takarítok, főzök. Ma ablakokat pucoltam. **De hiányzik** az élőbeszéd, társaság, pörgés.
Je travaille à la maison du lundi au vendredi. Je suis devant l'ordi pendant 8 à 10 heures parce que le système n'est pas très rapide de chez moi. Je ne m'ennuie pas, j'ai toujours à faire. Je travaille en semaine, je cuisine. Je nettoie et cuisine le week-end. J'ai nettoyé les fenêtres aujourd'hui. Mais les discussions en direct, la compagnie, l'action me manquent.

15 **Én imádom,** hogy csak hármasban vagyunk egész nap. Mindennap együtt sétálunk egy nagyot.
J'adore être en trio toute la journée. Nous faisons un bon tour tous les jours.

16 **Egész jól. Főleg** amíg csak ketten vagyunk a gyerekkel.
Plutôt bien. Surtout quand nous ne sommes que deux avec mon enfant.

17 **Nagyon nehezen. Főleg** hogy nincsenek gyerekek, unokák.
C'est très difficile. Surtout que je n'ai pas d'enfants, ni de petits-enfants.

18 **Nehezen, de muszáj**. Tegnap sétálni, ma biciklizni mentünk, **ez így elment**. **De hiányoznak** a havi kétszeri-háromszori baráti összejövetelek. **A telefonos, internetes kapcsolattartás nem olyan**.
Difficilement, mais je suis obligé. Hier, nous sommes allés nous promener, aujourd'hui nous avons fait du vélo, ça allait. Mais les rencontres avec les amis deux ou trois fois par mois me manquent. Le téléphone, le contact par Internet, c'est pas pareil.

19 **Imádom**! Itthon vagyok a két gyerekemmel, a férjemet várjuk haza délután. **Semmi, de semmi problémám nincs ezzel**!
Je l'adore ! Je suis à la maison avec mes deux enfants, nous attendons mon mari à la maison l'après-midi. Je n'ai absolument aucun problème avec ça!

20 Én egy hete dolgozom itthonról, de előtte 3 hónapig úgyszólván teljes karanténban voltam műtét miatt. **Bírom. Mert bírni kell**.

Je travaille à domicile depuis une semaine, mais avant cela, j'étais, pour ainsi dire, complètement mis en quarantaine pendant 3 mois en raison d'une intervention chirurgicale. Je tiens le coup. Parce que j'y suis obligé.

Tableau 170 : Exemples de réponses sur un forum à la question « Hogy bírjátok a bezártságot ? » (« Comment vivez-vous le confinement ? »). Les fautes d'orthographe ont été corrigées. (Source : forum « hoxa.hu », avril 2020)

Si les activités mentionnées diffèrent d'une personne à l'autre (d'où une répétition lexicale limitée), chaque entrée est construite de manière identique : Les participants répondent d'abord avec un adverbe ou une phrase rapide à la question : « difficilement », « plutôt bien », « je l'adore » …), puis donnent quelques arguments. Les explications justifient l'adjectif choisi mais contiennent des informations plutôt générales. Il est frappant de constater à quel point certaines réponses sont vagues : « je m'occupe », « je ne m'ennuie pas », « j'ai toujours à faire ». Quelques participants énumèrent des activités mais même leurs récits restent courts et généraux. La raison en est peut être que la question initiale (Comment vivez-vous le confinement ?) n'incite pas à une réponse détaillée : *l'intérêt des contributions n'est pas de donner des récits précis sur sa vie mais de signaler aux autres participants que l'on fait partie du groupe et que l'on partage une expérience commune*. Comme nous le verrons, par ailleurs, au chapitre 14, ce genre de langage est imprécis mais pouvoir créer un lien entre les locuteurs, est une caractéristique typique de toutes sortes d'interactions, formelles ou informelles. Il est donc tout à fait utile et justifié d'attirer l'attention des apprenants sur ce phénomène.

4.3) Courtes interactions

Le sous-ensemble des interactions écrites contient non seulement des contributions dans les réseaux sociaux, mais aussi des « *chats* » et des courriels. Les conversations par « *chat* » ont tendance à évoluer autour d'un nombre limité de sujets tels que « prendre un rendez-vous », « informer une autre personne que l'on est arrivé ou que l'on sera en retard », « envoyer des photos d'un plat ou d'un événement », avec de courts commentaires, généralement suivis d'une réponse enthousiaste du partenaire à l'échange[120].

[120] Les photos n'ont pas pu être incluses dans le corpus en raison de problèmes de confidentialité et du manque d'options pour le stockage des données multimédia.

L'édition de ces textes, comme l'illustrent les échanges ci-dessous (tableau 171), n'a impliqué que des modifications mineures, comme la suppression des noms propres et autres informations personnelles.

Conversation 1	**Conversation 1**
A: Hahó! Mikor beszélünk? 3? 4? 5?	*A : Salut ! Quand est-ce qu'on se parle ? A 3 heures ? 4 ? 5 ?*
B: A három tökéletes.	*B : 3 est parfait.*
Conversation 2	**Conversation 2**
A: (*name*) írt vissza, neki péntek este 8 körül vagy szombat napközben lenne a legjobb. Esetleg a vasárnap délelőtt. Nekem mindegy. Mikor találkozzunk?	*A : (nom) a répondu qu'elle préférerait vers 20 heures vendredi ou samedi. Ou peut-être dimanche matin. Ça m'est égal. Quand voulez-vous qu'on se voit ?*
B: Szombaton napközben lenne talán a legjobb. 11 után bármikor.	*B : Samedi serait peut-être le mieux, pendant la journée. Après 11 heures, n'importe quand*
Conversation 3	**Conversation 3**
A: Akkor majd írj, hogy mikor jó Neked! :)	*A : Fais-moi savoir quand tu as le temps ! :)*
B: Szia, bocsánat, hogy csak most jelentkezem, nekem péntek előtt sajnos ezen a héten semmi nem jó. Az késő van? Ha igen, akkor megpróbálok keresni korábban valami időpontot.	*B : Bonjour, désolé(e) pour cette réponse tardive, malheureusement rien ne me va avant vendredi de cette semaine. Est-ce trop tard ? Si oui, je vais essayer de trouver une date plus tôt.*
A: Dehogyis, semmi baj! Nekem csak a szombat nem jó. Minden más igen, de szombaton is biztos lesz olyan idő, amikor jó.	*A : Ne t'inquiète pas, aucun souci ! Samedi ne me convient pas, mais ça va bien pour le reste. Même le samedi, je pourrais trouver un moment qui convient.*

Tableau 171 : Extraits du corpus de clavardage (« *chat* »).

De nombreux échanges d'e-mails autour des thèmes du CECRL (description d'un week-end ou d'un événement, souhaits d'anniversaire, etc.) ont également pu être intégrés dans le corpus. Comme indiqué plus haut, nous avons conservé la formulation originale de la plupart des messages, mais les textes ont été raccourcis pour une meilleure lisibilité. Parfois, des paragraphes entiers ont dû être supprimés car ils impliquaient des références à des lieux, des noms et autres informations nécessitant des connaissances préalables, ou contenaient des informations personnelles.

Il convient de noter que la collecte des interactions informelles (« *chats* » et courriels) s'est avérée une tâche plus difficile que prévu. Nombre de nos amis et collègues n'ont pas accepté de publier leurs textes, même anonymement. Certaines justifications étaient liées au style du texte (« Je n'ai pas vraiment fait attention quand j'ai écrit cela », « c'est mal écrit ») ou à des doutes sur l'utilité de la contribution (« Es-tu sûr(e) de vouloir utiliser cela ? » « Il n'y a vraiment rien de remarquable là-dedans ! ») ou au fait que l'on puisse reconnaître l'auteur par le contenu et l'opinion exprimée (« Même si tu enlèves mon nom, il y aura beaucoup de choses par lesquelles les gens pourront m'identifier »)[121].

De telles difficultés, inhérentes au processus de collecte des données ont un impact sur la distribution des textes dans le corpus. Tout d'abord, il est biaisé par le genre des auteurs, car il contient plus d'interactions entre femmes que d'interactions entre hommes ou hommes et femmes et il n'y a qu'occasionnellement des interactions entre hommes ; les sources féminines étant jusqu'à présent plus disposées à contribuer que leurs homologues masculins. La collecte d'interactions plus semi-formelles et d'interactions entre hommes est prévue pour pallier ce déséquilibre. D'autre part, le style de ces interactions est essentiellement informel, allant de conversations légèrement informelles à des conversations très informelles.

Des commentaires qui mettent l'usage de ces éléments en perspective, peuvent aider l'apprenant de niveaux de compétences linguistiques inférieurs à comprendre à quel degré de familiarité il peut utiliser telle ou telle expression avec son partenaire de conversation.

4.4) Récits écrits

Il existe de nombreuses façons de compiler des récits écrits qui permettent aux apprenants d'observer des modèles stylistiques, structurels et lexico-grammaticaux tout en pratiquant et en consolidant les éléments observés. On peut constituer un corpus de plusieurs *textes provenant de sources différentes et présentant la même* information (par exemple, le *même* produit, le *même* article de presse, la *même* ville ou la *même* biographie). Inclure des textes sur le même smartphone, le même musée ou la même ville provenant de sources différentes en serait un exemple.

L'autre possibilité est de rassembler des *textes sur des sujets similaires contenant des informations différentes* (par exemple, des recensions de produits différents mais apparentés, des présentations de villes ou

[121] Communications personnelles des enseignants.

de musées). L'inclusion de textes décrivant différents types de smartphones ou différentes capitales en serait un exemple. De telles collections peuvent être compilées sur presque tous les sujets pertinents pour l'apprenant. Comme pour les interactions écrites, une sélection minutieuse et une étape préalable d'édition sont généralement nécessaires, et les textes courts (100 à 250 mots) semblent être les plus appropriés.

La première approche – plusieurs textes, même contenu – donne l'occasion aux apprenants d'observer ce qui est répété (avec ou sans variation), donc typique et utile. D'après notre expérience, il faut au moins cinq textes sur le même contenu pour que les répétitions et les variations deviennent perceptibles. Les textes doivent être choisis par l'enseignant qui peut évaluer leur niveau de complexité ainsi que leur degré d'authenticité et de caractère naturel[122]. En outre, la lecture de plus d'un texte facilite la compréhension : ce qui n'est pas compris dans un texte l'est généralement dans un autre.

La deuxième approche – textes différents, sujets similaires – peut également être utilisée avec n'importe quel sujet. Elle présente l'avantage que les informations présentées dans les textes sont différentes et que les apprenants ne « décrochent » pas après le troisième ou le quatrième texte en raison du manque de nouveauté. De tels ensembles de données sont relativement simples à construire, par exemple on peut facilement compiler, et modifier si nécessaire, un corpus de présentations de smartphones ou de critiques de restaurants en quelques heures. Les apprenants peuvent lire quelques textes dans leur intégralité et analyser ensuite l'ensemble des textes avec des outils de corpus. Il est également possible d'inclure le texte original dans un corpus pour un niveau de compétences supérieur (dans notre cas, dans le corpus B1) et le texte édité (modifié et/ou raccourci) dans ceux pour les niveaux A1 et A2.

Nous illustrerons cette procédure en prenant l'exemple des rexcensions de restaurants (textes complémentaires pour le thème « Nourriture et boissons » du CECRL). Cette partie du corpus comprend environ 100 critiques (3 000 tokens) et se divise en deux sections dont la première contient les textes intégraux, recommandés pour les niveaux A2+ et B1. Dans cette section, les textes ont été conservés sous leur forme originale mais les fautes d'orthographe et les phrases incongrues ont été supprimées. La seconde section comprend les versions modifiées et abrégées

[122] Il est important de vérifier le caractère naturel de ces textes car les sites Web copient souvent des informations les uns sur les autres et les réécrivent pour éviter les problèmes de droits d'auteur. Les textes créés de cette manière sonnent souvent peu naturels.

des mêmes textes, recommandées pour les niveaux A1 et A2. Dix exemples de textes sont publiés sur le site Web du manuel « MagyarOK » (à la fois dans leur version originale et dans leur version modifiée), dont trois sont présentés ci-dessous (tableau 172). Les éléments lexicaux identifiés comme appartenant au vocabulaire de base en raison de leur fréquence élevée dans les textes sont marqués en gras – ces éléments ont tous été inclus dans les versions abrégées. Les expressions modifiées (une seule dans les textes ci-dessous) sont soulignées.

Textes originaux (version hongroise)	**Textes édités (version hongroise)**
1 Kedves, udvarias személyzet, **nagyon finom ételek, italok**. **Gyors volt a kiszolgálás, a tálalás pedig ízléses**. Kétszemélyes tálat ettünk, isteni finom volt. Volt a tálon minden, ami szem-szájnak ingere. Érdemes volt ide betérni, máskor is fogunk még ide jönni.	**1 Kedves, udvarias személyzet**, **nagyon finom ételek, italok**. **Gyors volt a kiszolgálás, a tálalás pedig ízléses**. Kétszemélyes tálat ettünk, isteni finom volt. Máskor is jövünk még ide.
2 2019 augusztusában voltunk X vendéglőben. **Nagyon tetszett a hely. A felszolgálás gyors volt.** Amint betértünk **azonnal jött a pincér**, és kettőnknek ajánlott is asztalt. Az italfogyasztástól az ételig **minden nagyon finom volt**. Érdemes volt ide betérni, nyugodt hely és romantikus. Nagyon jól éreztük magunkat a párommal. Jó hogy itt fogyasztottuk az ebédet.	**2** 2019 augusztusában voltunk X vendéglőben. **Nagyon tetszett a hely. A felszolgálás gyors volt. Azonnal jött a pincér,** és ajánlott asztalt. Az italoktól és az ételekig **minden nagyon finom volt**. Nyugodt hely és romantikus, nagyon jól éreztük magunkat a párommal.
3 Udvarias pincérek, finom ételek. **Nem kellett sokat várni,** bár hétköznap délben ez talán könnyebben összejön. Az egyszemélyes tál egy plusz körettel simán elmegy két személynek is, ha valaki nem akarja degeszre enni magát, így **elég gazdaságos is volt**. Természetesen ezért az árért nem számítottam arra, hogy gasztronómiai különlegességet kapunk. **Normál ételek**	**3 Udvarias pincérek, finom ételek**. Nem kellett sokat várni. Az egyszemélyes tál két személynek is elég, így elég gazdaságos is volt. **Normál ételek voltak, finoman elkészítve**, ízlésesen tálalva. Erre számítottunk, és ezt is kaptuk.

voltak, finoman elkészítve, ízlésesen tálalva. Erre számítottunk, és ezt is kaptuk.

4 Nagyon szép környéken van, a személyzet nagyon kedves, segítőkész. A vártnál sokkal gyorsabban megkaptuk az ebédet, ami **ízlésesen volt tálalva**. **Finom és nagy adag** volt mindkettőnké, a húsok szuper puhák! Megérte itt ebédelni, ja, és **nem is drága egyáltalán**. A pécsi barna sört pedig kóstolja meg mindenki, kötelező! Desszerteket sajnos nem tudtuk megkóstolni, már nem maradt nekik hely.

4 Nagyon szép környéken van, a személyzet nagyon kedves, segítőkész. Gyorsan megkaptuk az ebédet, ami **ízlésesen volt tálalva**. **Finom és nagy adagokat kaptunk**! És **nem is volt drága egyáltalán**. A pécsi barna sör pedig *nagyon finom*! A desszerteket sajnos nem tudtuk megkóstolni, már nem maradt nekik hely.

Textes originaux (traduction française)

1 ***Personnel agréable et poli, nourriture et boissons délicieuses. Le service était rapide. Tout était présenté avec goût.*** *Nous avons mangé un plat pour deux, c'était divinement délicieux. Il y avait dans l'assiette tout ce que l'on peut désirer. Cela valait la peine de venir ici, nous y reviendrons.*

2 *En août 2019, nous étions au restaurant Tettyei Garden.* ***J'ai beaucoup aimé cet endroit. Le service était rapide. Dès que nous sommes arrivés, le serveur est venu*** *et nous a donné une table pour nous deux.* ***Tout, des boissons à la nourriture, était délicieux.*** *Cela valait la peine de venir ici, c'était un endroit calme et romantique. J'ai passé un bon moment avec mon partenaire. C'était bien d'avoir déjeuné ici.*

Textes originaux (traduction française)

1 ***Personnel agréable et poli, nourriture et boissons délicieuses. Le service était rapide. Tout était présenté avec goût.*** *Nous avons mangé un plat pour deux, c'était divinement délicieux. Nous y reviendrons.*

2 *En août 2019, nous étions au restaurant Tettyei Garden.* ***J'ai beaucoup aimé cet endroit. Le service était rapide. Dès que nous sommes arrivés, le serveur est venu*** *et nous a donné une table.* ***Tout, des boissons à la nourriture, était délicieux.*** *C'est un endroit calme et romantique. J'ai passé un bon moment avec mon partenaire.*

3 ***Serveurs polis, nourriture délicieuse. Nous n'avons pas eu à attendre longtemps****, bien qu'à midi en semaine, cela puisse être plus facile. Le plat pour une personne avec un accompagnement supplémentaire est suffisant pour deux personnes si vous ne voulez pas trop manger, donc c****'était assez économique, aussi****. Bien sûr, pour ce prix, je ne m'attendais pas à recevoir une spécialité gastronomique.* ***Il s'agissait de plats normaux, délicieusement préparés, présentés avec goût.*** *C'est ce que nous attendions et c'est ce que nous avons eu.*	3 ***Serveurs polis, nourriture délicieuse. Nous n'avons pas eu à attendre longtemps.*** *Le plat pour une personne était suffisant pour deux personnes, donc c****'était assez économique, aussi. C'était des plats normaux, délicieusement préparés, présentés avec goût****. C'est ce que nous attendions et c'est ce que nous avons eu.*
4 ***Il est situé dans un quartier très agréable, le personnel est très gentil, serviable. Le déjeuner a été servi*** *beaucoup* ***plus rapidement*** *que prévu, et* ***était présenté avec goût****. De délicieuses grandes portions pour nous deux, la viande était super tendre ! Cela valait la peine de déjeuner ici, oh, et ce n'était pas cher du tout. Vous devriez goûter la bière brune de Pécs, c'est obligatoire ! Malheureusement, nous n'avons pas pu goûter les desserts, on n'avait plus de place pour ça.*	4 ***Il est situé dans un quartier très agréable, le personnel est très gentil, serviable. Le déjeuner a été servi rapidement*** *et* ***était présenté avec goût****. Nous avons eu de délicieuses grosses portions et ce n'était pas cher du tout. La bière brune de Pécs est délicieuse. Malheureusement, nous n'avons pas pu goûter les desserts, on n'avait plus de place pour ça.*

Tableau 172 : Critiques de restaurants originales et éditées.

4.5) L'avantage de lire les textes dans leur intégralité : observer des répétitions et des variations linguistiques « à l'œil nu »

Les textes présentés dans les sections précédentes sont accessibles dans leur intégralité sur le site Web de la série « MagyarOK ». Comme évoqué dans le chapitre 2, l'apprenant « entre » dans une langue par des textes car ceux-ci lui permettent d'observer les éléments langagiers en tant que composantes d'une entité cohérente. Il est donc logique de proposer le contenu du corpus sous forme de collection de textes à lire pour qu'il puisse se familiariser avec leur contenu. Cela lui permet d'interpréter plus facilement les extraits de textes affichés à l'interface du corpus et de les analyser avec les outils numériques.

Deux autres avantages, inhérents à l'étude de plusieurs textes authentiques et semi-authentiques autour du même thème sont les suivants : premièrement, des textes construits de manière similaire et au contenu comparable constituent un apport riche pour les niveaux inférieurs. Deuxièmement, ces textes comportent de nombreuses variations et répétitions linguistiques et illustrent ainsi la créativité limitée des natifs dans des situations identiques. Ils mettent également en exergue les schémas lexicaux et grammaticaux les plus usités, tout cela sans outils numériques mais en utilisant les méthodes et les résultats de l'analyse de corpus présentés au chapitre 4.

Ce sont les exemples de la section 4.4 qui illustrent le plus clairement le fait que le vocabulaire de base émerge naturellement lorsque les textes sont judicieusement sélectionnés (c'est-à-dire qu'ils sont en accord avec les sujets et les types de texte requis par le CECRL). Si le choix des textes de départ est pertinent, il suffit de les raccourcir pour les rendre accessibles tout en conservant la majeure partie de la formulation originale. *Proposer un certain nombre de textes au contenu similaire constitue ainsi une approche originale*, car les manuels (principalement en raison de leur longueur) présentent généralement le vocabulaire de base lié à un sujet dans un ou deux textes seulement. En outre, l'inclusion de plusieurs versions du même texte[123] offre l'avantage aux apprenants de différents niveaux de compétences linguistiques d'en trouver une qui leur soit adaptée[124].

Les textes présentés dans la section 4.3 servent d'exemples pour illustrer comment les collègues et les amis prennent rendez-vous par « *chat* ». Les trois exemples cités contiennent encore peu de répétitions textuelles mais si les apprenants lisent vingt ou vingt-cinq exemples sur ce sujet, ils peuvent observer un grand nombre d'éléments récurrents ainsi que des variations. Quelques exemples : « mikor találkozunk ? » (quand se voit-on ?), « ha neked/nektek is jó » (si ça te/vous convient), « ráérek" (j'ai le temps), « jó » (entendu), « nekem mindegy » (ça m'est égal, tout me va), « mit szólsz ? » (qu'en dis-tu ?).

Le même constat s'applique à la collection de réponses de forum dans la section 4.2. C'est grâce à la quantité de données linguistiques que *les apprenants pourront observer le phénomène de répétition et de variation sur les plans lexical et pragmatique.* Tout d'abord, les apprenants sont confrontés à au moins dix réponses différentes à la question initiale « Comment supportez-vous le confinement ? » : « nagyon nehezen » (très difficilement), « nehezen » (difficilement), « nem valami jól » (pas très

[123] Selon le CECRL, les thèmes de A1, A2 et B1 sont similaires (voir aussi le chapitre 2), ce qui garantit la répétition et l'élargissement du vocabulaire entre les différents niveaux. Certains textes figurent dans le corpus dans leur version originale ainsi que sous forme(s) adaptée(s) comme exposé plus haut.
[124] Ils peuvent même comparer le texte original et le(s) texte(s) modifié(s).

bien), « volt már jobb is » (j'ai déjà été mieux), « kezd sok lenni » (ça commence à faire), « jól » (bien), « egész jól » (plutôt bien), « nagyon jól » (très bien), « imádom » (je l'adore). Les activités s'articulent autour des axes suivants : être dehors, contact avec les amis par téléphone et Internet, être seul ou avec sa famille, travail à domicile. Certaines phrases reviennent plusieurs fois sous la même forme : « hiányzik/hiányoznak a ... » (il me manque/manquent), « muszáj » (on est obligé), « jól, főleg hogy/amíg » (bien, surtout parce que/tant que). D'autres expriment le même message avec des variations : « Nincs alkalmam unatkozni. » (Je n'ai pas d'occasion de m'ennuyer.), « Nem unatkozom. » (Je ne m'ennuie pas.), « Van tennivalóm. » (J'ai à faire.). Ou : « ha az idő jó » (s'il fait beau) et « ilyen szép időben » (avec ce beau temps). Ou encore : « Semmi bajom, ha … » (Je n'ai pas de problème si/quand …), « Semmi problémám nincs ezzel. » (Je n'ai aucun problème avec ça.).

Ces textes peuvent enrichir le vocabulaire des apprenants de façon efficace : la plupart des éléments lexicaux sont simples mais leur combinaison peut être nouvelle si l'apprenant n'y a pas encore été exposé en classe ou en dehors du cours. Le corpus étant de petite taille, le nombre des répétitions est plutôt limité ; l'enseignant peut donc compléter l'observation des éléments avec un enseignement explicite des éléments particulièrement utiles, fréquents et/ou typiques (Nation 2013). Ces textes peuvent également servir de modèles pour les textes des apprenants, comme nous le verrons au chapitre 15.

Ce chapitre a présenté les principes généraux de conception de corpus écrits aptes à fournir des textes utiles afin de compléter le contenu du manuel. *En prenant les descripteurs du CECRL comme lignes directrices, nous avons proposé que de tels corpus contiennent des interactions et des récits simples relatifs à la vie quotidienne et nous avons montré comment construire un corpus basé sur trois sous-ensembles en s'appuyant sur des données linguistiques semi-authentiques et authentiques.* D'une part, le sous-ensemble des manuels permet tout d'abord aux apprenants d'explorer des textes qui leur sont familiers. De l'autre part, les textes-modèles semi-authentiques qui traitent de sujets pertinents, rédigés par des locuteurs natifs aident à systématiser et à élargir les connaissances linguistiques. Enfin, des collections d'interactions et de récits authentiques (adaptés ou non), centrés sur des sujets et des types de textes pertinents, offrent la possibilité d'observer comment les natifs s'expriment dans diverses situations. Les textes peuvent être explorés par les outils numériques (voir le chapitre 15 sur les activités basées sur le corpus) ou par une lecture classique. Lire les textes dans leur intégralité

permet de à se familiariser avec le contenu du corpus, c'est-à-dire d'observer le vocabulaire et les thèmes inclus et d'interpréter plus facilement les résultats de l'analyse de corpus.

Nous avons également montré que la création des corpus à fins linguistiques obéit à d'autres critères que celle des corpus pédagogiques. La différence la plus importante repose sur le fait que, dans le cas des corpus pédagogiques, des compromis concernant l'authenticité du contenu semblent être nécessaires pour rendre les données du corpus accessibles aux apprenants de niveaux inférieurs. Modifier des textes authentiques et/ou guider les locuteurs, au moins en partie, lors de la production de leurs écrits pour le sous-ensemble 2, apparaissent ainsi comme des étapes incontournables de la construction de tels corpus au service de l'enseignement des langues (en particulier aux niveaux A1 et A2).

Le prochain chapitre abordera les questions liées à la création de corpus pédagogiques oraux, ainsi que leurs avantages et leurs limites. Il suit une structure similaire à celle de ce chapitre.

Chapitre 14 : Corpus oraux au service des apprenants aux niveaux de compétences linguistiques inférieurs

La présentation d'un corpus pédagogique oral pour les apprenants de hongrois constitue le cœur de ce chapitre. Comme le corpus écrit présenté au chapitre 13, ce corpus a été créé pour compléter la série de manuels « MagyarOK ». L'intérêt d'une telle collection est que l'usage langagier dans des interactions parlées a ses propres caractéristiques. Le fait d'en prendre conscience dès le début du processus d'apprentissage peut améliorer de manière significative la compréhension orale, la compétence communicative et l'expression orale des apprenants. Ce corpus a été construit en suivant une approche en trois étapes, identique à celle décrite dans le chapitre précédent. Nous exposerons le contenu des sous-ensembles ainsi que le processus de collection et d'adaptation de textes et les problèmes qui ont émergé lors de la création du corpus oral.

Les questions que nous nous poserons sont les mêmes que dans le cas du corpus écrit, complétées par la question abordant l'usage du langage interactionnel :

- Comment générer du langage pour les corpus pédagogiques oraux ?
- Comment collecter du langage pour ces corpus ?
- Comment adapter les énoncés réels au niveau de l'apprenant ?
- Dans quelle mesures les conversations sur le même thème contiennent-elles des répétitions et des variations linguistiques ? Comment peut-on les faire émerger ?
- Comment l'utilisation du langage interactionnel se manifeste-t-elle dans ces collections ?

A) Les sous-ensembles oraux

1) Aperçu général

O'Keeffe et al. (2007 : 30) décrivent l'intérêt des corpus oraux pour l'enseignement des langues de la façon suivante :

> « Lorsque nous observons ce que font les locuteurs, [...] nous entendons de vraies personnes qui interagissent les unes avec les autres, [...] qui sont créatives, affectives, interpersonnelles et, surtout, qui s'expriment en s'engageant dans les processus de communication qui se trouvent au centre de nos vies. Il est difficile d'imaginer que l'apprenant d'une deuxième langue ne souhaite pas être un bon communicateur dans cette nouvelle langue [...]. L'enseignement des langues ne

peut que bénéficier d'une étude encore plus approfondie de ces processus fondamentalement humains. » (notre traduction)

Un corpus de récits et d'interactions oraux peut éclairer les apprenants sur la façon dont les natifs réalisent ces tâches. Comme pour les corpus écrits, nous avons compilé plusieurs sous-ensembles avec un degré croissant d'authenticité. Dans notre corpus oral pour le hongrois, une approche en trois étapes conduit les apprenants des interactions à caractère naturel incluses dans le manuel vers le sous-ensemble de données linguistiques authentiques, en passant par les contributions semi-authentiques. Chaque sous-ensemble sert des objectifs spécifiques dans le processus de développement des compétences orales des apprenants.

- Le sous-ensemble 1 contient les transcriptions des dialogues des manuels (environ 62 100 tokens). Ces conversations inspirées par le corpus mais éditées offrent des modèles pour la réalisation de diverses interactions et présentent de nombreuses caractéristiques typiques des interactions orales.
- Le sous-ensemble 2 comprend les transcriptions de deux types d'interactions semi-authentiques : (1) des improvisations d'acteurs (20 700 tokens) et (2) de courts entretiens avec des locuteurs natifs (110 400 tokens). Ces interactions sont non scénarisées mais guidées.
- Le sous-ensemble 3 comprend des transcriptions de données linguistiques authentiques et est également divisé en deux parties : (1) les rencontres dans les lieux de service (environ 35 880 tokens) et (2) les conversations semi-formelles et informelles (environ 34 500 tokens). Il contient de nombreux enregistrements qui ont fourni la base des dialogues du manuel. Les transcriptions ont été réalisées avec le logiciel « Alrite » et corrigées manuellement. Certaines transcriptions ont également été éditées.

L'ensemble du corpus oral (niveau A1-B1) compte actuellement environ 262 200 tokens (plus de 38 heures d'enregistrement, soit environ 115 tokens/minute). Il s'agit d'une quantité de données supérieure à celle que pourrait fournir n'importe quel manuel. Les transcriptions ont été annotées en fonction de leur sujet, du nombre, du sexe et de l'âge des participants, du degré de formalité et de leur longueur. L'organisation de ces ensembles de données est un projet en cours, mais certaines des vidéos et enregistrements audio des sous-ensembles 1 et 2 sont d'ores et déjà disponibles sur le site Web du livre de cours et sur la chaîne YouTube de « MagyarOK ». Les transcriptions peuvent être analysées avec des outils numériques proposés par Sketch Engine. Le matériel vidéo

et audio du sous-corpus 3 est en cours de traitement pour améliorer la qualité des enregistrements et, lorsque nécessaire, anonymiser leur contenu afin de respecter la protection des données personnelles.

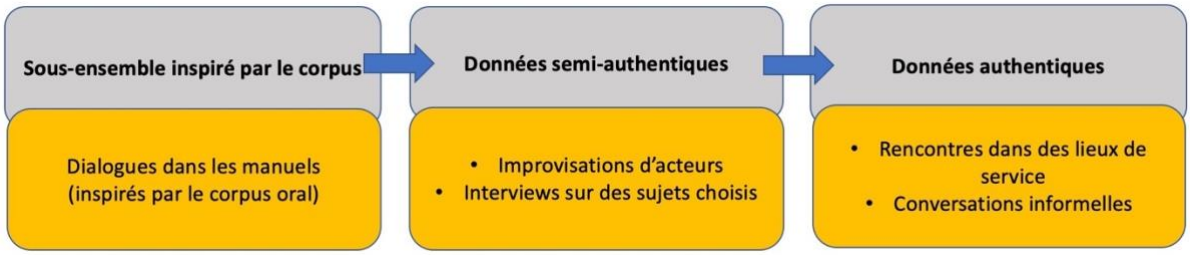


Tableau 173 : Sous-ensembles de données linguistiques pour les niveaux A1-B1.

2) Sous-ensemble (1) : dialogues dans les livres de cours

L'une des principales limites des dialogues rédigés des manuels aux niveaux inférieurs est qu'ils « reflètent rarement l'imprévisibilité des conversations et les caractéristiques […] des interactions réelles » (Burns (2001) cité dans O'Keeffe et al. 2007 : 21). Cela s'explique, du moins en partie, par leur fonction : ces dialogues sont censés montrer aux apprenants comment gérer des situations de la vie quotidienne et, par souci de clarté, ils sont généralement assez simples, avec une structure linéaire[125].

En dépit de cette contrainte, les interactions informées par le corpus peuvent conserver un degré d'authenticité significatif. Cette qualité en fait des « tremplins » idéaux pour les données semi-authentiques et authentiques. Nous illustrons cela avec un dialogue du chapitre 1 du manuel A1 (tableau 174). Il s'agit de l'une des premières conversations introduisant le vocabulaire dont les apprenants ont besoin pour se présenter. Les éléments du langage interactionnel sont écrits en gras.

Tímea: Szia! Tímea vagyok.	*T : Bonjour, je suis Tímea.*
Anna: Anna. **Nagyon örülök.**	*A : Anna.* ***Enchantée.***
Tímea: **Én is**... Magyar vagy, **ugye**?	*T :* ***Enchanté****. Tu es Hongroise,* ***n'est-ce pas*** *?*
Anna: Igen.	*A : Oui.*
Tímea: Budapesten élsz?	*T : Tu vis à Budapest ?*
Anna: Nem, Debrecenben.	*A : Non, à Debrecen.*

[125] Il convient de rappeler au lecteur les difficultés à interpréter les enregistrements authentiques ainsi que leurs transcriptions, que ce soit dans sa langue maternelle ou dans une autre langue.

Tímea: **És miért** tanulsz csehül?	*T : **Et pourquoi** apprends-tu le tchèque ?*
Anna: **Mert** a barátom cseh.	*A : **Parce que** mon petit ami est Tchèque.*
Tímea: **Tényleg**? **És** milyen nyelven beszéltek otthon?	*T : **Vraiment** ? **Et** quelle langue parlez-vous à la maison ?*
Anna: Magyarul és néha egy kicsit csehül. Patrik, a barátom nagyon jól tud magyarul.	*A : Hongrois et parfois un peu le tchèque. Patrik, mon ami, parle très bien le hongrois.*
Tímea: **Érdekes**! **Az én** barátom **is** cseh, **de mi** otthon csehül beszélünk.	*T : **Intéressant**. **Mon** petit ami **est aussi** Tchèque**, mais nous** parlons tchèque à la maison.*

Tableau 174 : Dialogue dans le manuel pour le niveau A1, inspiré par une conversation enregistrée.

Malgré l'adaptation du dialogue original, cette conversation a un caractère naturel parce que les éléments du langage transactionnel y ont été inclus ou, plus précisément, n'ont pas été retirés du dialogue. Acquérir ces expressions permet aux apprenants d'interagir de façon relativement naturelle dès les premières étapes de l'apprentissage.

Un exemple qui souligne que l'oralité possède ses propres règles vient de l'utilisation de la conjonction « És » (Et) dans le dialogue. La plupart des natifs s'accorderaient probablement à dire que placer « És » en début de phrase est une faute de style que l'on doit éviter (Schirm 2019). Cette recommandation est, en effet, justifiée pour les textes écrits. Cependant, notre corpus oral indique que commencer un énoncé par « És » est extrêmement courant dans le discours oral et non sans raison : ce mot relie les énoncés du locuteur 2 à ceux du locuteur 1 après le tour de parole et contribue ainsi à la co-construction du discours.

Il est instructif de comparer ce dialogue avec l'original enregistré entre deux Hongroises avant leur cours de tchèque aux Pays-Bas (tableau 175). Les éléments intégrés dans la conversation du manuel sont en caractères gras.

(*A hallja, hogy B magyarul telefonál*)	(*A entend B parler hongrois sur son téléphone portable*)
Anna: **Szia!** De jó! Te is **magyar vagy?**	*Anna : **Salut** ! Comme c'est bien ! **Tu es Hongroise**, toi aussi ?*
Tímea: Aha. Ezek szerint te is?	*Tímea : Ouais. Toi aussi alors ?*
Anna: Aha. Nem is tudtam, hogy rajtam	

kívül is vannak még itt magyarok.
Tímea: Ja, ja, vannak. De még nem olyan régóta vagyok itt. Amúgy **Tímea vagyok.**
Anna: **Anna. Nagyon örülök.**
Tímea: **Én is…** És amúgy, Magyarországon hol laksz?
Anna: **Debrecenben.**
Tímea: **Tényleg?** Egy nagyon jó barátom is ott lakik. Lehet, hogy ismered is. (*Mondja a nevét*).
Anna: Sajnos nem ismerem.
Tímea: Mondjuk, ahhoz nagy Debrecen, hogy mindenkit ismerj… Na, de hogyhogy itt, Hollandiában **tanulsz csehül?**
Anna: Hát, a barátom cseh, úgyhogy meg akarom tanulni a nyelvet.
Tímea: **Tényleg? És milyen nyelven beszéltek otthon?**
Anna: **Magyarul és néha egy kicsit csehül. Patrik, a barátom nagyon jól tud magyarul.**
Tímea: **Érdekes! Az én barátom is cseh, de mi otthon csehül beszélünk.** Próbálunk legalábbis. (*nevet*)
Anna: Nekünk is azt kéne, akkor biztos, hogy gyorsabban megtanulnám. De nem bírom rávenni magamat. Egyszerűbb a magyar. (*nevet*)
(…)

Anna : Oui. Je ne savais même pas qu'il y avait d'autres Hongrois ici.
Tímea : Si, si, il y en a. Mais je ne suis pas ici depuis très longtemps. Au fait, ***je suis Tímea****.*

Anna : ***Anna. Enchantée.***
Tímea : ***Enchantée.*** *Au fait, où habites-tu en Hongrie ?*
Anna : ***À Debrecen****.*
Tímea : ***Vraiment*** *? Un très bon ami à moi y vit aussi. Tu le connais peut-être aussi.*
(Elle donne le nom.)
Anna : Malheureusement, je ne le connais pas.
Tímea : C'est vrai que Debrecen est trop grand pour connaître tout le monde... Alors, comment ça se fait que ***tu apprennes le tchèque*** *ici aux Pays-Bas ?*
Anna : Eh bien, mon petit ami est Tchèque, donc je veux apprendre la langue.
Tímea : ***Vraiment ? Et quelle langue parlez-vous à la maison ?***
Anna : ***Le hongrois et parfois un peu le tchèque. Patrik, mon ami, parle très bien le hongrois.***
Tímea : ***Intéressant ! Mon ami est aussi Tchèque, mais nous parlons tchèque à la maison****. Du moins, nous essayons. (Elle rit.)*
Anna : Nous devrions aussi faire ça. Je l'apprendrais plus vite, c'est sûr. Mais j'ai du mal à me motiver. Parler en hongrois est plus simple.
(Elle rit.)
(...)

Tableau 175 : Interaction authentique comme base du dialogue présenté dans le tableau 174.

Ce dialogue illustre la structure non linéaire des conversations authentiques évoquée précédemment. Il est aussi plus long que sa version adaptée car les locutrices vont au-delà d'une simple « collecte d'informations », en faisant l'effort d'établir un lien l'une avec l'autre. Ce faisant, elles ont tendance à poser des questions plus générales afin de laisser à leur partenaire la liberté de décider les informations qu'elle souhaite partager (cf. Bencze 2020 ; McCarthy 2002, 2003 ; Rühlemann 2007, 2018). En outre, elles nomment des lieux et des personnes que la partenaire est susceptible de connaître et s'expriment plus longuement en répondant aux questions. En raison de ces caractéristiques, le dialogue sous sa forme originale est plus difficile à suivre que sa version adaptée.

Le texte contient quelques expressions du langage parlé dont l'introduction au chapitre 1 du manuel A1 serait prématurée même si les natifs les utilisent souvent. Nous pensons notamment à des expressions informelles comme « na, de hogyhogy … » (mais comment ça se fait que …), « mondjuk » (disons), « aha » et « ja » (les deux comparables au « ouais » français) qui, si utilisées dans un contexte inadéquat, peuvent être perçues comme des formulations impolies. Les natifs effectuent leurs choix stylistiques en évaluant le contexte, alors que les débutants ne peuvent s'appuyer que sur des moyens limités pour s'exprimer. Ces moyens doivent pouvoir « passer partout », c'est-à-dire qu'ils doivent appartenir au langage standard pour éviter les faux-pas. Les éléments propres à des registres particuliers devraient donc être introduits graduellement à des stades plus avancés de l'apprentissage. Pour la majorité, ils ne sont pas en lien avec un sujet mais plutôt avec des situations de communication, comme nous le verrons par la suite.

Il convient de noter que le thème « Rencontrer quelqu'un pour la première fois » est abordé aux trois niveaux de compétences linguistiques inférieurs, ce qui offre un certain degré de flexibilité à la présentation des éléments caractéristiques de cette situation. À mesure que le niveau augmente, les dialogues des manuels révisent le vocabulaire important du ou des niveaux précédents et l'enrichissent de nouveaux éléments. Les dialogues sont plus longs, les réponses deviennent de plus en plus longues et les questions moins directes, se rapprochant ainsi de la dynamique des dialogues authentiques. Cette structure cyclique permet aux apprenants d'étendre graduellement leurs connaissances linguistiques. Ils ont également la possibilité de consulter de nombreux textes originaux dans le sous-ensemble 3 (dont le dialogue précédent entre deux Hongroises). La conversation originale citée plus haut se trouve par ailleurs dans le corpus pédagogique (dans la sous-partie pour les niveaux B1 et B2), l'apprenant a donc, au cours de son apprentissage, l'opportunité de lire le texte original et d'observer ses particularités.

La manière dont nous avons simplifié le texte présenté plus haut a été appliquée pour l'adaptation d'autres textes. Bien que notre objectif dans les dialogues des manuels ait été de garder le langage proche des enregistrements, de nombreux textes ne convenaient pas pour les niveaux A1 et A2 sous leur forme originale. Dans de tels cas, nous avons procédé comme dans le cas du dialogue présenté : l'« ossature » de la conversation authentique a été conservée dans le manuel, mais les digressions spontanées ont été supprimées et certaines réponses raccourcies. Ces compromis sont apparus être nécessaires pour créer des textes que les apprenants peuvent facilement suivre. Lors de la modification de certaines parties des dialogues, nous avons suivi l'approche proposée par McCarten (2010) qui recommande que les dialogues des manuels devraient avoir un caractère naturel mais doivent être également clairs et transparents dans l'intérêt de l'apprentissage.

3) Sous-ensemble 2(1) : enregistrements vidéo avec des acteurs

Les informations visuelles font partie de la plupart des interactions réelles. Elles permettent de clarifier le contexte situationnel, de montrer les gestes et les expressions faciales et gestuelles accompagnant les énoncés ainsi que les objets auxquels les locuteurs font référence. Ces divers éléments justifient l'intérêt des enregistrements vidéo (Fortanet-Gómez et Querol-Julián 2010 ; Montenero Perez et Rodgers 2019).

Les enregistrements vidéo de « MagyarOK » se situent à mi-chemin entre les données semi-authentiques et authentiques. Ils ont été réalisés par des acteurs mais sans textes pré-rédigés. Les acteurs n'ont donc pas reçu de script, seulement un bref aperçu de la situation qu'ils étaient censés improviser. Ils ont joué des variations de la même interaction à plusieurs niveaux linguistiques (A1, A2 et/ou B1). Pour cela, ils ont seulement reçu l'instruction plutôt succincte de parler naturellement et d'éviter les longs monologues. 41 vidéos ont été enregistrées pour le niveau A1, 47 pour le niveau A2 et 34 pour le niveau B1 ; les sujets ont été déterminés par le CECR (voir la liste au chapitre 2). Les commentaires des locuteurs natifs (enseignants et non enseignants) ont confirmé que les scènes « respectent le niveau des apprenants sans compromettre l'authenticité », comme l'a dit une collègue[126].

Toutes les vidéos ont la même structure : une courte introduction écrite présente la scène avant qu'elle ne commence. La mise en scène est minimaliste, les acteurs portent des vêtements simples, l'arrière-plan est uniforme. Cette disposition, exempte d'éléments distrayants, favorise la concentration sur la langue. Dans l'extrait ci-dessous (tableau 176), les acteurs jouent deux

[126] Conversation enregistrée le 18 juin 2019 avec Mme Tímea Baumann, enseignante de hongrois à l'Université de Pécs.

étudiants qui se rencontrent dans le train et engagent une conversation. Ce dialogue fait partie de l'ensemble de textes pour le niveau A2 et explore le thème « Rencontrer quelqu'un pour la première fois », tout comme le dialogue du manuel présenté précédemment[127]. Les éléments du langage interactionnel sont écrits en gras.

Máté: Ne haragudj, szabad?	*M : Excuse-moi. Je peux (m'asseoir) ?*
Flóra: **Persze**.	*F : **Bien sûr**.*
Máté: Köszönöm... Bocsáss meg, hova utazol?	*M : Merci... Excuse-moi, où vas-tu ?*
Flóra: Szegedre.	*F : À Szeged.*
Máté: Én is.	*M : Moi aussi.*
Flóra: Egyetemista vagy?	*F : Tu es étudiant ?*
Máté: **Hát**, **igen**. Táncművészetire járok.	*M : **Eh bien, oui**. J'étudie à l'Académie de danse.*
Flóra: **Tényleg**?	*F : **Vraiment** ?*
Máté: Igen. Szegedre megyek próbálni... Te mi járatban?	*M : Oui. J'ai une répétition à Szeged... Et toi ?*
Flóra: **Hát**, egyetemista vagyok, **úgyhogy** az egyetemre megyek. Angol szakos vagyok.	*F : **Eh bien**, je suis étudiante **donc**, je vais à l'université. J'étudie l'anglais.*
Máté: **Angol szakos.**	*M : **L'anglais**.*
Flóra: **Ühüm**.	*F : **Ouais**.*
Máté: **Szuper**. És ott tanulsz Szegeden?	*M : **Super**. Et tu étudies à Szeged ?*
Flóra: Igen, **meg... meg igazából** most amiatt jöttem, hogy lesz edzésem. Szinkronúszom.	*F : Oui... **et maintenant,** j'y vais **en fait** parce que j'ai un entraînement. Je fais de la natation synchronisée.*
Máté: **Szinkronúszol**?	
Flóra: Igen.	*M : **De la natation synchronisée ?***
Máté: **Az tök szuper**.	*F : Oui.*
(...)	*M : **C'est trop cool**.*
	(…)

Tableau 176 : Dialogue improvisé par deux acteurs (niveau A2).

[127] Au total, 12 situations similaires sont incluses dans les corpus A1, A2 et B1 : deux anciens camarades de classe se croisent à l'aéroport, des apprenants de langue se présentent lors de leur première leçon d'italien, deux personnes âgées se rencontrent dans un musée, un professeur d'université se présente à son auditoire, etc.

Ce court extrait indique déjà que l'improvisation contient de nombreux éléments du langage interactionnel et les acteurs utilisent des stratégies conversationnelles courantes de manière naturelle. Par exemple, ils répètent les mots de leur partenaire (« J'apprends l'anglais. – L'anglais. »), posent des questions (« Je fais de la natation synchronisée. - De la natation synchronisée ? ») ou donnent un feedback positif sur ce que dit leur partenaire (« Super. » « Génial. »). *Ces stratégies simples, fréquemment utilisées sont faciles à employer et même les apprenants des niveaux inférieurs peuvent les intégrer dans leurs conversations pour créer des liens avec leurs partenaires.*

Travailler avec des acteurs présente plusieurs avantages. Tout d'abord, ils ne sont pas familiarisés avec le contenu des manuels de cours et se fient donc à leur propre jugement pour décider quel niveau de langage était « suffisamment simple ». Ils ne sont pas tentés de créer des énoncés non naturels, faciles à comprendre mais non typiques (un danger quand nous travaillons avec des professeurs de langues qui produisent occasionnellement des phrases peu typiques dans leurs cours[128]). Deuxièmement, ils sont familiers de l'improvisation et peuvent immédiatement compris ce qu'on attendait d'eux et ont pu jouer les scènes avec aisance, devant la caméra. Comme les acteurs ont réalisé la plupart des dialogues à plusieurs niveaux (puisque de nombreux sujets sont couverts à chaque niveau du CECR), nous avons pu observer que la complexité croissante impliquait un langage transactionnel qui fournit de plus en plus de détails (des réponses plus longues aux questions), mais que *le nombre d'éléments interactionnels est resté le même. Une observation intéressante et utile sur laquelle l'enseignant peut attirer l'attention de l'apprenant.*

4) Sous-ensemble 2(2) : entretiens scénarisés

Un autre sous-ensemble de textes semi-authentiques comprend un ensemble d'entretiens dans lesquels les locuteurs natifs racontent leurs expériences liées à des sujets du quotidien, définis par le CECRL. Bien que l'entretien en tant que genre ne représente pas une situation de communication naturelle et spontanée, nous considérons qu'il a toute sa place dans le processus d'apprentissage : lorsque plusieurs personnes interrogées répondent aux mêmes questions, les apprenants peuvent en effet remarquer les répétitions et les variations lexicales et grammaticales dans leurs réponses[129]. Ces réponses peuvent également servir de modèles pour leurs propres énoncés.

[128] Pour un aperçu de la recherche sur le langage de l'enseignant dans le cours de langues voir, par exemple, Cullen 1998 ; Walsh 2010 ; Basra et Toyyibah 2017 ; Moser et al. 2012 ; Thornbury 1996.
[129] Pour une approche similaire voir Braun (2006 et 2010) ainsi que le chapitre 2 de cette thèse.

Huit à dix entretiens ont été réalisés pour dix sujets par niveau (soit un total de 80 à 100 interviews par niveau), d'une durée d'une à six minutes chacun. Cet ensemble constitue un corpus d'environ 16 heures de matériel au total, principalement formé d'enregistrements audio mais également accompagné de quelques vidéos. Les listes des questions posées lors des entretiens sont téléchargeables sur le site Web et peuvent être utilisées par les apprenants et les enseignants pour mener des entretiens similaires avec les personnes de leur choix. Les conversations enregistrées n'ont été modifiées en aucune façon et leurs transcriptions sont incluses dans le corpus ouvert « MagyarOK ».

Compte tenu des exigences spécifiques de chaque niveau de compétences, les entretiens ont été menés de manière légèrement différente aux différents niveaux. Le corpus pour le niveau A1 consiste en de courtes questions et réponses autour de divers sujets du CECRL. Les personnes interrogées ont été limitées à des enseignants, car *il était crucial de trouver un équilibre entre l'authenticité et l'accessibilité afin que les débutants puissent réellement bénéficier de la collection*. Pour la même raison, les mêmes questions ont été posées dans chaque entretien sur un sujet donné. Par exemple, les questions sur les « Activités quotidiennes » étaient systématiquement les suivantes : « Que fais-tu le lundi ? » « Où passes-tu la matinée en général ? » « À quelle heure déjeunes-tu ? » « Où déjeunes-tu ? » « Que fais-tu après le déjeuner ? » « Que fais-tu le soir ? » « Que fais-tu vendredi soir ? » « Que fais-tu pendant le weekend ? » Pendant les conversations, l'intervieweuse est restée en retrait : elle a réagi aux réponses (elle a fourni quelques exemples du langage interactionnel) mais n'a pas fait de commentaire sur le sujet lui-même (elle n'a pas fourni d'exemples pour le langage transactionnel).

Un problème majeur auquel nous avons été confrontée *lors de l'enregistrement de ces conversations était la difficulté d'obtenir un niveau de langage à caractère naturel que les apprenants de niveau A1 pourraient comprendre*. Certaines entrevues ont dû être écartées parce qu'elles semblaient artificielles, les participants n'étant que trop disposés à s'adapter au public débutant en simplifiant à l'extrême leurs énoncés. La simplification dans les textes que nous avons décidé d'écarter, impliquait l'omission des dispositifs de connexion, celle des modificateurs (comme « assez », « surtout », « plutôt ») et des réactions appropriées si elles utilisaient un langage idiomatique – en bref, les caractéristiques essentielles de l'utilisation d'un langage à caractère naturel. Ainsi, la compilation d'un corpus oral de petite taille pour le niveau A1 s'est avérée être un processus long et fastidieux, car les enregistrements supprimés ont dû être remplacés par de nouveaux. Cet effort souligne là encore l'importance d'un travail de sélection du contenu du corpus pédagogique critique et raisonné, suivant une logique claire.

Les trois entretiens présentés ci-dessous (tableau 177) ont été enregistrés pour le niveau A1. Les personnes interviewées parlent de leur semaine. Les éléments que l'on retrouve dans plusieurs conversations sont écrits en caractères gras. Les éléments modifiés sont indiqués par une parenthèse. Les éléments appartenant au langage interactionnel et aux stratégies conversationnelles sont soulignés.

R: Zsófi, milyen programod van hétfőn?
Zs: Hétfőn először a munkahelyemre megyek, nyolctól négyig dolgozom, utána hazamegyek. Utána elmegyek futni, **ha** jó idő van, **akkor** a szabadban, **ha** rossz idő van, **akkor** futópadon.
R: Hol vagy általában délelőtt?
Zs: [Délelőtt mindig] a munkahelyemen vagyok.
R: Hánykor ebédelsz?
Zs: 12-kor, délben.
R: Hol ebédelsz?
Zs: A munkahelyemen a konyhában, vagy pedig étterembe megyünk.
R: Hová mész ebéd után?
Zs: Ebéd után visszamegyek dolgozni.
R: Milyen programod van pénteken este?
Zs: Péntek este a barátaimmal találkozom, és kávézóba megyünk vagy moziba.
R: Milyen programod van szombat délelőtt?
Zs: Szombat délelőtt **először mindig** takarítok, aztán főzök, és persze a családommal vagyok.
R: Mikor sportolsz?
Zs: Sportolni hetente kétszer-háromszor szoktam, esténként.
R: Mit csinálsz hétvégén?
Zs: Hétvégén próbálok pihenni, olvasni, filmet nézni, és ha jó idő van, kirándulni.
R: Mikor pihensz?
Zs: Este. Este szoktam pihenni.

(R: Zsófi, quel programme as-tu le lundi?
Zs: Le lundi, je vais d'abord sur mon lieu de travail, je travaille de 8h à 4h, puis je rentre chez moi, puis je vais courir dehors, [s'il] fait beau, ou [encore] sur un tapis roulant [s'il] fait mauvais,.

R: Où es-tu d'habitude le matin?

Zs: Je suis [toujours] au travail [le matin].

R: A quelle heure déjeunes-tu ?

Zs: À midi.

R: Où déjeunes-tu ?

Zs: Sur mon lieu de travail, à la cuisine ou on va au restaurant.

R: Où vas-tu après le déjeuner?

Zs: Je retourne au travail après le déjeuner.

R: Quel programme as-tu le vendredi soir?

Zs: Je rencontre mes amis le vendredi soir et nous allons dans un café ou au cinéma.

R: Quel programme as-tu le samedi matin?

Zs: Je fais ***toujours*** *le ménage* ***en premier*** *le samedi matin, puis je cuisine et, bien sûr, je suis avec ma famille.*

R: Quand fais-tu du sport?

Zs: Je fais de l'exercice deux ou trois fois par semaine, le soir.

R: Que fais-tu le week-end?

Zs: Le week-end, j'essaie de me détendre, de lire, de regarder un film et, s'il faut beau, de faire une randonnée.

R: Quand te reposes-tu?

Zs: Le soir, je me détends généralement le soir.))

R: Timi, mi a heti programod? Például mit csinálsz hétfőn?

T: Hétfőn az egyetemen dolgozom.

R: Egész nap?

T: Igen, igen. Nyolctól ott vagyok.

R: Mikor ebédelsz?

T: 11 óra 30 perckor kezdődik az ebédszünet.

R: És hol ebédelsz?

T: Az egyetemen.

R: Mit csinálsz délután?

T: Délután is dolgozom, háromig, és három órakor hazamegyek... bocsánat, először az iskolába megyek.

R: Iskolába? iskolába jársz? (*nevet*)

T: Nem, nem. (*nevet*) A fiam és a lányom jár iskolába. Ezért az iskolába megyek, és a fiammal és a lányommal együtt hazamegyünk.

R: Mi a délutáni program, a gyerekekkel?

T: Aha. **Délután általában** otthon vagyunk. Játszunk, főzök valamit, kenyeret sütök, vacsorázunk. Néha kicsit sportolunk is.

R: A gyerekekkel együtt sportolsz?

T: Igen, a gyerekek bicikliznek, és én is biciklizem, vagy futok a gyerekekkel együtt.

R: Akkor van időd sportolni, az jó.

T: Ühüm. Igen, egy kis időm van, igen.

R: Mit csinálsz a hétvégén?

T: **Hétvégén általában** itthon vagyunk, és a kertben dolgozunk, vagy biciklizünk valahol, vagy csak játszunk, néha takarítok.

R: Az kell. Mikor pihensz?

T: Mikor pihenek? Éjszaka. (*nevet*)

(R: Timi, quel est ton programme hebdomadaire? Par exemple, que fais-tu le lundi ?

T: Je travaille à l'université le lundi.

R: Toute la journée?

T: Oui, oui. J'y suis à partir de huit heures.

R: Quand déjeunes-tu?

T: La pause déjeuner commence à 11h30.

R: Et où déjeunes-tu?

T: À l'université.

R: Que fais-tu l'après-midi?

T: Je travaille aussi l'après-midi, jusqu'à trois heures, et je rentre à la maison à trois heures ... Pardon, je vais d'abord à l'école.

R: À l'école? Tu vas à l'école ? (Elle rit.)

T: Non, non. (Elle rit.) Mon fils et ma fille vont à l'école. C'est pourquoi je vais à l'école et je rentre à la maison avec mon fils et ma fille.

R: Et quel est le programme de l'après-midi avec les enfants?

T: Nous sommes ***généralement*** *à la maison* ***l'après-midi****. Nous jouons, je cuisine quelque chose, je fais du pain, nous dînons. Parfois, nous faisons aussi un peu de sport.*

R: Tu fais du sport avec les enfants?

T: Oui, les enfants font du vélo et je fais du vélo ou je cours avec eux.

R: Tu as le temps de faire du sport, c'est bien.

T: Ouais, oui, j'ai un peu de temps, oui.

R: Que fais-tu le week-end?

T: ***Le week-end****, nous sommes* ***généralement*** *à la maison et nous travaillons dans le jardin, nous faisons du vélo quelque part ou nous jouons simplement, parfois je fais le ménage.*

R: Il le faut. Quand te reposes-tu ?

T: Quand je me repose ? La nuit. (Elle rit.))

R: Milyen programod van hétfőn?

O: **Hétfőn általában** reggeltől délutánig dolgozom, **este általában** otthon vagyok.

R: Hol vagy délelőtt?

O: Délelőtt az egyetemen vagyok, az egyetemen dolgozom.

R: Hánykor ebédelsz?

O: Fél kettőkor.

R: És hol ebédelsz?

O: **Ha** dolgozom, és az egyetemen vagyok, **akkor** az egyetemen, a munkahelyemen ebédelek.

R: Hová mész ebéd után?

O: Ebéd után néha még dolgozom, **ha** nem dolgozom, **akkor** általában hazamegyek.

R: Mit csinálsz munka után?

O: **Munka után általában egy kicsit** [pihenek], [olvasok] **egy kicsit**, zenét hallgatok, néha sétálok, és már este is van.

R: Mikor sportolsz?

O: Általában hetente kétszer-háromszor sportolok. **Hétvégén mindig** konditerembe megyek, és a héten egyszer vagy kétszer jógázom, vagy otthon tornázom.

R: Hű, akkor te nagyon aktív vagy. És mit csinálsz este?

O: **Este általában** vacsorázom, filmet nézek, beszélgetek és borozok a férjemmel.

R: Milyen a hétvégéd?

O: Hétvégén [próbálok dolgozni] egy kicsit és [pihenni] **is egy kicsit**. Általában takarítok és főzök is, mindig bevásárolok, és sétálok vagy [kirándulok] **egy kicsit**, és **ha tudok**, találkozom a barátaimmal is.

(R: Quel est ton programme le lundi?

O: Je travaille ***généralement*** *du matin à l'après-midi le* ***lundi****, et je suis* ***généralement*** *à la maison* ***le soir****.*

R: Où es-tu le matin?

O: Je suis à l'université le matin, je travaille à l'université.

R: À quelle heure déjeunes-tu?

O: À deux heures et demie.

R: Et où déjeunes-tu ?

O: Si je travaille et que je suis à l'université, je déjeune là-bas, au travail.

R: Où vas-tu après le déjeuner?

O: Parfois, je travaille encore après le déjeuner, ***si*** *je ne travaille pas,* ***alors*** *je rentre chez moi* ***généralement****.*

R: Que fais-tu après le travail?

O: ***Après le travail,*** *je me détends* ***un peu****, je lis* ***un peu****, j'écoute de la musique, parfois je marche, et c'est déjà le soir.*

R: Quand fais-tu du sport?

O: Je fais habituellement du sport deux ou trois fois par semaine. Je vais toujours à la gym le week-end et je fais du yoga une ou deux fois par semaine ou je m'entraîne à la maison.

R: Wow, tu es très active alors. Et que fais-tu le soir?

O: Je dîne ***généralement le soir****, je regarde un film, je discute et je bois du vin avec mon mari.*

R: Comment se passe ton week-end?

O: Le week-end, j'essaye de travailler ***un peu*** *et de me détendre* ***un peu****. Habituellement, je fais le ménage et je cuisine aussi, je fais toujours du shopping et je fais des promenades ou des randonnées* ***un peu****, et* ***si je peux****, je rencontre aussi mes amis.)*

Tableau 177 : Trois exemples d'interviews au niveau A1 sur le sujet « Vie quotidienne ».

Que peuvent observer les apprenants dans ces interviews ? Bien que l'on ne puisse considérer ces dialogues comme véritablement naturels et authentiques, le travail avec ces entretiens apparaît cependant pertinent car même si les conversations sont fortement guidées et les questions de l'interviewer artificielles, elles font émerger un grand nombre de stratégies conversationnelles et d'expressions utiles pour le quotidien. Cela montre à quel point ces éléments font parties de manière organique de n'importe quel échange à l'oral : ils sont répétés et variés dans les différents textes, par différents locuteurs, soulignant ainsi qu'il ne s'agit pas de réalisations langagières singulières et rendant par la même occasion leur observation relativement aisée.

Notre attention se portera maintenant sur les répétitions et les variations concrètes, émergeant de ces trois entretiens. Le schéma grammatical « ha X, akkor Y » (si X, alors Y) exprimant que l'action

est liée à une condition revient plusieurs fois dans les entretiens. On remarque cependant que l'une des phrases cette structure sans la conjonction « akkor » (alors) : « Ha tudok, találkozom a barátaimmal is. » (Si je peux, je rencontre aussi mes amis.), offrant la possibilité d'en observer une variation. Un autre schéma grammatical qui se dégage de ces phrases est lié à l'ordre des mots avec un complément de temps suivi de l'adverbe « általában » (en général, habituellement) : « délelőtt általában, este általában, hétfőn általában, hétvégén általában, munka után általában » (le matin en général, le soir en général, lundi en général, après le travail en général). Dans deux phrases « általában » est remplacé par un autre élément lexical « mindig » (toujours) sans que le schéma observé ne change. Il ne s'agit donc que d'une variation : « hétvégén mindig, először mindig », (le week-end toujours, d'abord toujours). Puisque l'ordre des mots est très différent du français, cet aspect de la langue mérite l'attention de l'apprenant qui peut en observer de nombreux exemples dans les entretiens.

Les textes contiennent également des répétitions et des variations de composantes lexicales, par exemple avec le verbe « pihenek » (je me détends) : « este pihenek, próbálok pihenni, egy kicsit pihenek » (je me détends le soir, j'essaie de me détendre, je me détends un peu). L'apprenant observe donc le même élément lexical utilisé dans trois environnements textuels légèrement différents. Le verbe le plus usité dans ces dialogues est « dolgozom » (je travaille), l'apprenant a ainsi de nombreuses opportunités d'observer son usage. Par exemple : « Ebéd után néha még dolgozom » (Je travaille parfois après le déjeuner), « Ha nem dolgozom, akkor általában hazamegyek. » (Si je ne travaille pas, je rentre en général à la maison), « Nyolctól négyig dolgozom, utána hazamegyek » (Je travaille de 8 à 16 heures, puis je rentre à la maison), « Hétfőn általában reggeltől délutánig dolgozom. » (Lundi, je travaille en général du matin à l'après-midi), « Délután is dolgozom » (Je travaille aussi l'après-midi), « Hétfőn az egyetemen dolgozom » (Lundi je travaille à l'université) et « Ebéd után visszamegyek dolgozni. » (Je retourne au travail après le déjeuner).

On notera également de noter que les textes contiennent quelques stratégies conversationnelles que l'apprenant a déjà pu observer dans les interactions semi-authentiques des acteurs (section 3, ce chapitre). Des exemples de telles stratégies se produisent lorsque le locuteur répète la question de son partnenaire : « Mikor pihensz? – Mikor pihenek? » (Quand te reposes-tu ? – Quand je me repose ?) ou exprime une réaction (ici son étonnement) par la répétition : « Bocsánat, először az iskolába megyek. – Iskolába? Iskolába jársz? » (Pardon, je vais d'abord à l'école. – À l'école ? Tu vas à l'école ?) Les locuteurs peuvent aussi répondre en évaluant l'énoncé du partenaire : « Hű, akkor te nagyon aktív vagy » (Wow, tu es vraiment active alors.), « Az jó. » (C'est bien.) ou « Az

kell » (Il le faut). Il arrive enfin que les participants répètent plusieurs fois un élément de leur énoncé, soit pour lui donner plus d'emphase, soit pour gagner du temps : « Nem, nem. » (Non, non.) « Igen, igen. » (Oui, oui.) « Ühüm, igen egy kis időm van, igen. » (Ouais, oui, j'ai un peu de temps, oui.) « Este. Este szoktam pihenni. » (Le soir. Je me détends le soir.)

Que ces répétitions et variations apparaissent à l'apprenant de façon évidente en raison du vocabulaire limité, constitue un net avantage des entretiens au niveau A1, car *le but principal de cette comparaison approfondie de plusieurs textes sur le même sujet est la consolidation du vocabulaire-clé et des schémas typiques.* Le travail avec les textes dans leur intégralité peut par ailleurs être complété avec l'analyse des textes au moyen d'outils numériques, comme nous le verrons dans le chapitre 15.

L'avantage de ces entretiens est cependant aussi source de contraintes. Rappelons que les participants sont des professeurs de hongrois qui savent exactement quel vocabulaire les apprenants peuvent comprendre. Pour que leurs réponses soient pertinentes, accessibles et faciles à suivre, les enseignants ont utilisé un maximum de langage transactionnel connu des élèves (l'accent dans le chapitre étant sur la présentation de ces éléments) et ont limité l'usage du langage interactionnel. Par conséquent, ces dialogues contiennent davantage de langage transactionnel que les enregistrements pour les niveaux A2 et B1 et, ne reflètent pas les interactions typiques. Néanmoins, certains éléments du langage interactionnel sont observables, comme nous l'avons vu précédemment et il est probable que leur nombre limité aide leur interprétation correcte ainsi que leur mémorisation.

La tâche d'attirer l'attention sur ces éléments revient initialement à l'enseignant qui peut, ensuite, habituer les apprenants à remarquer eux-mêmes ces éléments. Une fois appris comment identifier les répétitions et les variations, les apprenants peuvent travailler de façon autonome et demander éventuellement à l'enseignant de vérifier si leurs observations sont correctes. L'enseignant peut également proposer des activités en utilisant ces textes pour des jeux de rôle et pour la personnalisation des informations. Les apprenants peuvent, par exemple, marquer ce qui est vrai pour eux, répondre au nom de la personne interviewée ou en leur propre nom[130].

Les enregistrements pour le niveau A2 ont permis une plus grande flexibilité dans le choix des participants et dans la manière dont les discussions ont été réalisées que les enregistrements pour

[130] Voir aussi les activités proposées au chapitre 15.

le niveau A1. Les personnes interrogées étaient des locuteurs natifs « ordinaires » – cinq hommes et cinq femmes – de différents groupes d'âge. *Ils ont tous reçu les mêmes questions ; la seule demande de l'interviewer était d'éviter les réponses longues et complexes, mais ils n'étaient limités dans leur choix linguistique en aucune façon.* Par exemple, les questions initiales relatives au sujet « La vie à la campagne et en ville » étaient les suivantes : « Préfères-tu la vie en ville ou à la campagne ? Pour quelles raisons ? » « As-tu une ville ou un village préféré ? » Dans ces dialogues, l'interviewer a également contribué à la discussion, en donnant son opinion argumentée ; les questions et les réactions de l'interviewer étaient ainsi plus variées en fonction des réponses. Par exemple, elle a confirmé qu'elle connaissait la ville mentionnée, qu'elle l'aimait également et, occasionnellement, elle a proposé d'autres arguments. Cela a rendu les interactions plus proches des conversations naturelles[131]. Comme les personnes interrogées ne connaissaient pas le contenu des manuels, leur usage langagier n'était pas déterminé (voire biaisé) par des connaissances préalables. De même, l'intervieweur ne pouvait pas anticiper leurs réponses et réagissait souvent spontanément aux énoncés de la personne interrogée. Dans les exemples ci-dessous (tableaux 178) les dispositifs communicatifs sont soulignés et le langage d'argumentation indiqués en caractère gras.

R: Melyik életformát szereted jobban? A vidékit vagy pedig a nagyvárosi életet?
A: Igazából **én a városi életet szeretem**. **Szeretem a nyüzsgést**… **Szeretem a nyüzsgést, a nagyváros lüktetését, szeretek színházba, moziba járni.** Persze nagyon szeretem a természetet is, de... de **jobban kedvelem a nagyvárost**.
R: Akkor **te nem is költöznél vidékre**.
A: Hát, lehet, hogy egyszer szeretnék, de **most jobban érzem magam a városban**.
R: Van kedvenc falud vagy városod? Mindegy, hogy hol, akár Brazíliában, akár Magyarországon.
A: Természetesen Magyarországon, és **nekem a kedvenc városom Budapest, ahol születtem**.
R: Akkor te budapesti vagy?
A: Igen. **Budapesten minden van**, és miután Sao Paolóban élek, számomra Budapest egy kisvárosnak számít, és bejárható, belátható, a vendéglátás, a gasztronómia, a művészetek, koncertek, parkok, **minden megtalálható ott, amire az embernek szüksége van**, azt gondolom.

[131] Une manière similaire de collecter des données est discutée par Tyne (2012) qui présente une base de données pour l'espagnol contenant des entretiens que les étudiants ont enregistrés avec leur assistant de langue.

(R: Quel style de vie préférez-vous? Campagne ou grande ville?

A: En fait, j'aime la vie en ville. J'aime cette effervescence… J'aime cette effervescence, la pulsation de la grande ville, j'aime aller au théâtre et au cinéma. Bien sûr, j'aime aussi beaucoup la nature, mais ... mais je préfère la grande ville.

R: Alors tu ne déménagerais pas à la campagne.

A: Eh bien, j'aurais peut-être envie plus tard, mais maintenant, ***je me sens mieux en ville****.*

R: As-tu un village ou une ville préféré? Peu importe où, que ce soit au Brésil ou en Hongrie.

A: Bien sûr en Hongrie, et ma ville préférée est Budapest, où je suis née.

R: Alors tu es de Budapest?

A: Oui. Il y a tout à Budapest, et depuis que je vis à Sao Paolo, pour moi Budapest est une petite ville et on peut se promener, elle est de taille humaine avec une hospitalité, la gastronomie, les arts, les concerts, les parcs, tout ce dont on a besoin, on le trouve là-bas, je pense.)

R: Melyik életformát szeretitek jobban, a vidékit vagy a nagyvárosi életet?

Á: A vidékit... a vidékit. Egyértelműen.

R: És miért?

P: **Van tér, van udvar, van kapcsolat az emberekkel, mindenki szóba áll egymással, tehát közvetlenebbek az emberi kapcsolatok egy faluban, mint a városban. Mindenki mindenkit ismer**, legalábbis sok embert.

R: Ágota, te azt mondtad, hogy **szeretnél belvárosba költözni**. Miért?

Á: Csak a gyerekek miatt. Arra az időszakra, amikor kamaszok, amikor még otthon laknak, hogy ne legyen az akadály a mozgásukban, **hogy ne kelljen busszal bemenni mindenhova**.

R: Akkor **a közlekedés miatt**?

Á: Igen. Abszolút. Meg idő, idő… Rengeteg időt visz el az ingázás a falu és a város között.

Igen.

R: Van esetleg kedvenc falvatok vagy városotok?

P: Hosszúhetény. Igen. **Én jól érzem itt magam**.

Á: Én is.

R: És hol van ez a hely?

P: Hát, Hosszúhetény Pécs mellett van, Pécstől 10 kilométerre, Délnyugat-Magyarországon, a Mecsek lábánál. Egy ilyen dimbes-dombos, szép környezetben lévő település.

R *(parle cette fois-ci à deux personnes) : Quelle forme de vie préférez-vous, la vie rurale ou métropolitaine?*

Á: *La campagne ... la campagne. De toute évidence.*

R*: Et pourquoi?*

(P: Il y a de l'espace, on a une cour, des contacts avec les gens, tout le monde se parle, donc les relations humaines sont plus simples dans un village que dans une ville. Tout le monde se connaît, du moins beaucoup de gens se connaissent.

R*: Ágota, tu as dit que tu voudrais déménager au centre-ville. Pourquoi?*

Á*: Juste pour les enfants. Pour la période où ils sont adolescents et vivent encore à la maison, pour qu'ils puissent se déplacer facilement sans prendre tout le temps le bus.*

R*: Alors à cause du trafic?*

Á: Oui. Absolument. Il faut du temps... du temps... Il faut beaucoup de temps pour se déplacer entre le village et la ville.

Á*: Oui.*

R*: Avez-vous des villages ou villes préférés?*

P: Hosszúhetény. Oui. Je me sens bien ici.

Á: Moi aussi.

R*: Et où est cet endroit?*

P: Eh bien, Hosszúhetény est près de Pécs, à 10 kilomètres de Pécs, dans le sud-ouest de la Hongrie, au pied du (montagne) Mecsek. Un sorte de village vallonné dans un endroit magnifique.)

R: Melyik életformát szereted jobban, a vidéki, vagy a nagyvárosi életformát?

A: Most már a vidékit. Fiatalon a nagyvárosit szerettem, most már a vidékit jobban szeretem.

R: És miért volt ez a váltás?

A: Mert még fiatalon **fontos volt a társaság, a szórakozóhelyek, minden elérhető volt, és a nyüzsgés.** Most így, **családdal kényelmesebb a nagy udvar, kevés szomszéd, és a hétvégi nyugalom**.

R: Egyszer nekem azt mondtad, hogy szeretnél a városhoz közelebb lakni, közelebb költözni.

A: Igen, egy hasonló kertvárosias övezetbe, **hogy ne kelljen autózni** a munkahelyemre.

R: Van kedvenc településed?

A: Vác, **ahol születtem, azt szerettem**. Az egy város.

R: Hol van Vác?

A: Budapesttől északra 30 kilométerre.

R: De ez egy elég kicsi hely, nem?

A: **Pont méretes város**. A közelben **van nagyváros**, Budapest, de Vácon **van elég munkahely**, és ott **van iskola, munkahely, lehet élni.** Város, de nem nagyváros.

(R: Quel style de vie préfères-tu, rural ou métropolitain?

A: Maintenant, c'est la campagne. J'aimais la grande ville quand j'étais jeune, maintenant je préfère la campagne.

R: Et pourquoi ce changement?

A: Parce que quand j'étais jeune, les copains, les boîtes de nuit et l'effervescence était importants, tout était accessible. Maintenant, comme ça, avec la famille on est plus à l'aise avec une grande cour, peu de voisins et un week-end de tranquillité.

R: Tu m'as dit une fois que tu voulais vivre plus près de la ville.

A: Oui, dans une banlieue comme ici pour ne pas avoir à conduire pour me rendre au travail.

R: As-tu une ville ou un village préféré?

A: Vác, ***là où je suis né, je l'ai aimé****. C'est une ville.*

R: Où est Vác?

A: À 30 kilomètres au nord de Budapest.

R: Mais c'est un assez petit endroit, n'est-ce pas ?

A: Une ville de juste la bonne taille. A proximité se trouve une grande ville, Budapest, mais il y a suffisamment de possibilités de travail à Vác, et il y a des écoles, du travail, on peut y vivre. C'est une ville, mais pas une grande ville.)

Tableau 178 : Trois exemples d'interviews au niveau A2 sur le sujet « Vivre dans une ville et à la campagne ».

L'analyse de ces entretiens a un avantage évident. Ils contiennent un grand nombre d'éléments lexicaux thématiques, par exemple : « van iskola, munkahely » (il y a des écoles, du travail), « mindenki ismer mindenkit » (tout le monde se connaît), « nagy udvar » (grande cour), « kevés szomszéd » (peu de voisins), « hétvégi nyugalom » (calme le weekend), « Budapesten minden van » (il y a tout à Budapest) et ainsi de suite. Mais cela n'est pas leur seule utilité. Comme dans le cas des entretiens pour le niveau A1, ces conversations permettent également aux apprenants d'explorer des questions telles que : « Y a-t-il des cas de répétitions et de variations dans les textes et entre les textes ? » « Des questions similaires génèrent-elles des réponses similaires ? » À première vue, les réponses semblent toutes différentes, par exemple, chaque locuteur formule une

réponse plus ou moins différente aux questions « Que fais-tu le lundi ? », « As-tu une ville préférée ou un village préféré ? » ou « Qu'est-ce qui t'aide à te concentrer ? » Néanmoins, une étude plus approfondie révèle également des similitudes.

Au-delà de ces éléments, les conversations comprennent également du vocabulaire qui révèle des manières différentes de donner son opinion, de nommer ses préférences ou d'argumenter. Quelques phrases qui expriment les préférences : « **szeretem a** nyüzsgést », « **szeretek** moziba, színházba járni » (j'aime aller au cinéma et au théâtre), « **most már jobban szeretem a** vidéket » (je préfère maintenant la vie à la campagne), « én **jól érzem itt magam** (je me sens bien ici), « **a kedvenc városom** Budapest, **ahol születtem** » (ma ville préférée est Budapest où je suis née), « **kedvelem a** falut is » (j'aime aussi le village), « **jobban kedvelem a** nagyvárost » (je préfère la grande ville). Dans ces phrases, l'apprenant peut observer plusieurs variantes de la structure « szeretem a ...-t » (j'aime ...) : « jobban szeretem a ...t » (je préfère ..., j'aime plus ...) qui exprime une comparaison en rajoutant le mot « jobban » (plus) à l'expression, « kedvelem a ...t » (j'aime ...), remplacement de « szeret » par un synonyme moins fort, « jobban kedvelem a ...t » (je préfère ..., j'aime plus ...), comparaison en rajoutant le mot « jobban » (plus) à l'expression.

Un des schémas grammaticaux qui ressort concerne l'utilisation du comparatif ; observation peu surprenante dans la mesure où les participants comparent deux modes de vie. En revanche, les différentes manières d'intégrer le comparatif dans les textes (pour les adjectifs et les adverbes) mérite l'attention de l'apprenant : « **jobban** kedvelem a nagyvárost » (je préfère la grande ville), « **jobban** érzem magam a városban » (je me sens mieux dans une grande ville), « az emberek **közvetlenebbek** » (les gens sont plus sympathiques), « **kényelmesebb** a nagy udvar » (une grande cour est plus confortable). L'utilisation dominante de « jobban » (mieux) et son environnement textuel sont ainsi observables de même que le fait que l'adverbe se trouve habituellement devant le verbe conjugué (« jobban » + V). Cette observation, guidée par les exemples tirés du corpus permet donc une prise de conscience inductive de l'importance d'ordre des mots dans la phrase comparative.

Ces dialogues offrent également de multiples opportunités d'observation des caractéristiques des interactions orales. Ils contiennent des hésitations et des répétitions, par exemple : « szeretem a természetet is, de... de jobban kedvelem a nagyvárost » (j'aime aussi la nature mais... mais je préfère les grandes villes), « Meg idő, idő... » (Et puis le temps, le temps), « A vidékit... a vidékit. Egyértelműen. » (La campagne... la campagne. C'est évident.), « van elég munkahely, és ott van iskola, munkahely, lehet élni. » (Il y a assez de travail, il y a une école, du travail, on peut y vivre.),

des phrases redondantes : « most így, a családdal » (maintenant, comme ça, avec la famille), des expressions non grammaticales et des cas d'autocorrection : « Meg idő, idő... rengeteg időt visz el az ingázás. » (Et puis le temps, le temps... le trajet domicile-travail prend énormément de temps). Le fait que ces caractéristiques sont extrêmement courantes dans les interactions orales, mérite être souligné car les apprenants ont tendance à penser qu'ils doivent produire des phrases parfaitement formulées dans la langue-cible pour être compris. Les données réelles montrent que les énoncés sonnent, en fait, plus naturels lorsqu'ils contiennent quelques imperfections.

Le dernier groupe des données semi-authentiques se compose des conversations enregistrées pour le niveau B1. Ces entretiens suivent un cours assez naturel et ressemblent à des échanges naturels. Tout comme pour les entretiens au niveau A2, les participants étaient des non-enseignants qui ne connaissaient pas le contenu linguistique du manuel du niveau B1. Lors de la préparation, ils ont reçu une liste de 15 sujets avant l'entretien et en ont choisi cinq dont ils voulaient parler. Par exemple les questions sur le thème « Se concentrer » étaient les suivantes : « Ton niveau de concentration est-il stable ? Sinon, de quoi cela dépend-il ? », « Qu'est-ce qui t'aide à te concentrer ? Qu'est-ce qui te dérange ? », « Que fais-tu lorsque tu remarques que tu ne peux plus te concentrer ? » Les deux parties ont posé des questions et donné des réponses, ce qui a donné lieu à un large éventail de langage transactionnel et interactionnel. De temps en temps, ils se sont également écartés du sujet pour suivre des idées spontanément émergeantes.

Ces entretiens se rapprochent le plus des conversations naturelles informelles et mettent davantage en relief leurs caractéristiques. Nous observons également plus de différences dans le vocabulaire lié à la singularité des expériences des locuteurs qu'ils partagent cette fois-ci dans des réponses plus longues. Néanmoins, de nombreux éléments interactionnels émergent systématiquement, certains d'entre eux associés aux actions langagières plus complexes (par exemple, donner une réponse négative en respectant les règles de la politesse quand l'opinion de l'interviewer et de la personne interviewée sur un film qu'ils sont vu tous les deux, est différente).

Il est également intéressant d'observer que ces conversations sont plus légères, moins denses en informations. Si nous comparons leur longueur et leur contenu informationnel avec les entretiens des niveaux A1 et A2, nous trouvons que la longueur des interactions augmente alors que leur contenu informationnel reste essentiellement le même. En outre, l'information factuelle est moins précise, les locuteurs hésitent à donner des réponses claires et tranchées, l'information est « noyée » dans les éléments interactionnels dont ces conversations regorgent.

Le tableau 179 montre, à titre d'exemple, un extrait d'entretien à ce niveau sur le sujet de la concentration. Les informations répétées sont marquées en gras, les éléments interactionnels sont soulignés :

B : **Vannak olyan napok, amikor gyakorlatilag nem tudok koncentrálni. Nem tudok odafigyelni arra, amit csinálni kéne, széjjelfolyik minden feladat.**
R : Ez ismerős. És van, amikor pedig nagyon gyorsan helyére kerülnek a dolgok és tudok haladni, sikeresen el tudok látni ilyen-olyan feladatokat. **Veled is így van ez?**
B : Hát persze. **Van, amikor teljesen mindegy, mit csinálsz, egyszerűen nem jönnek össze a dolgok.** Vagy egyszerűen kell egy olyan pszichés, vagy nem is tudom milyen állapot, hogy tényleg így jól tudj működni.
R : Neked vannak valami trükkjeid, hogyan tudsz koncentrálni?
B : Hát azt tudom, hogy... na, szóval **vannak olyan napok, amikor tényleg semmi nem sikerül.**
R : Nálam is van ilyen, hogyha nem alszom.
B : Ja, ja, nálam is. Ha nem alszom ki magam, **akkor egyszerűen nem megy és szétszórt vagyok, és nem tudok sokáig egy dologra figyelni**, meg én alapból is kicsit ilyen vagyok. Nekem az a nagyon egyhelyben ülni, nyugodtan ülni, egy dolgot csinálni sokáig, az egyébként sem megy.

*(B: **Il y a des jours où je ne peux pratiquement pas me concentrer. Je ne peux pas faire attention à ce que je devrais faire, rien n'aboutit.***
R: Je connais ça, moi aussi. Et puis, il y a des moments où les choses se mettent en place très rapidement et je peux avancer, je peux faire des choses. Ça t'arrive, à toi aussi ?
*B: Bien sûr. **Parfois, quoi que je fasse, les choses ne marchent pas**. Ou tu as juste besoin d'un état psychologique ou je ne sais pas quel état pour pouvoir vraiment bien fonctionner.*
R: Tu as des astuces pour t'aider à te concentrer ?
*B: Eh bien, je sais que... eh bien, **il y a des jours où je ne peux vraiment rien faire.***
R: Oui, oui, moi aussi. Quand je ne dors pas assez.
*B: Oui, moi aussi. Si je ne dors pas assez, **je n'y arrive pas, je suis distraite et je ne peux pas me concentrer sur une chose pendant longtemps,** et je suis un peu comme ça par défaut. Pour moi, rester assise, faire une seule chose pendant un long moment, je ne suis pas faite comme ça de toute façon.)*

Tableau 179 : Extrait d'entretien, niveau B1.

La valeur informationnelle de la conversation ci-dessus est très faible : nous apprenons seulement que la personne interviewée est quelqu'un d'actif et qu'elle a du mal à se concentrer quand elle n'a pas assez dormi. Sinon, les deux locutrices constatent qu'il y a des jours où les choses marchent et d'autres où les choses ne marchent pas. Néanmoins, l'importance de l'échange se noue sur un autre plan : outre le fait de partager des informations, elles se rapprochent l'une de l'autre, à travers leurs paroles. Nous pourrions également dire que le but de cette interaction – identique à tant d'autres interactions du quotidien – est la construction d'une relation à travers des expériences partagées. En observant ces entretiens, les apprenants peuvent donc se rendre compte que personne n'attend pas nécessairement d'eux des réponses précises dans des conversations informelles. Ils peuvent rester délibérément vagues et ils peuvent aussi reprendre les mots de leur(s) interlocuteur(s), car c'est ce que font les natifs (et probablement eux-mêmes aussi dans leur première langue) à condition qu'ils alimentent la conversation autrement : par des répétitions et par des éléments interactionnels (cf. André 2010 ; Bencze 2020 ; Evison et al. 2007 ; Rühlemann 2007, 2018). Les apprenants observent également quelques éléments langagiers qui les aident à formuler ce type de phrases.

En résumant les observations des données semi-authentiques, nous pouvons donc constater que ce sous-ensemble peut aussi bien illustrer le vocabulaire thématique, des échanges d'informations, des différentes fonctions du langage (argumenter, poser des questions polies, etc.) que le langage transactionnel. Par le biais des répétitions et des variations dans et entre les textes, lus et écoutés dans leur intégralité, les apprenants acquièrent une grande quantité d'informations sur l'usage oral du langage ainsi que du vocabulaire sur les sujets abordés dans les interactions. Ils contribuent également à l'observation de la dynamique et des éléments langagiers typiques des interactions orales. Ces analyses pourront être approfondies par l'étude du sous-ensemble des données authentiques, présenté dans la section suivante.

B) Collecte de données authentiques

1) Aperçu général

La collecte des contributions orales authentiques est généralement moins aisée que celle des textes écrits. Plusieurs questions doivent être clarifiées concernant le processus d'enregistrement et la protection des données ainsi que la manière de transcrire les textes.

Lorsque nous avons enregistré notre première collecte de données orales (2012–2014), les lois concernant la protection des données personnelles étaient moins strictes qu'aujourd'hui[132]. Notre approche était donc conforme aux lois en vigueur mais elle ne correspond plus à celles d'aujourd'hui. À l'époque, nous procédions de la manière suivante : au moment de l'enregistrement, les interlocuteurs ne savaient pas que nous étions en train de les enregistrer, mais ils en étaient informés par la suite. Ils pouvaient alors décider s'ils autorisaient ou non l'utilisation des données. Lors d'un refus, l'enregistrement était immédiatement effacé. La raison d'être de ce procédé était de s'assurer que les conversations se déroulaient de manière naturelle. Certaines interactions n'ont pas été enregistrées mais notées à la main, soit parce que l'équipement technique n'était pas disponible au moment de l'interaction, soit parce que l'interaction était si brève et si rapide qu'il était plus facile de la noter que d'essayer de l'enregistrer.

Si ces enregistrements ont été d'une grande utilité lors de la rédaction des manuels, la plupart d'entre eux n'ont pu être utilisés pour le corpus pédagogique en raison de leur qualité moyenne d'enregistrement. Ils contenaient des bruits de fond, des voix faibles (les locuteurs étant situés à une distance inégale de l'appareil) et de l'écho. Pour cette raison, nous avons décidé de ne publier que les transcriptions des dialogues et nous prévoyons de les réenregistrer de nouveau, avec des acteurs[133].

Une autre question importante qu'il est nécessaire d'aborder en relation avec ces dialogues est celle de leur transcription. La question du traitement des erreurs, des hésitations, des cas d'autocorrection, se pose inévitablement. Les « Directives pour l'encodage de textes » (Thompson 2004) recommandent ainsi d'inclure les caractéristiques suivantes dans une bonne transcription :

- Les énoncés et les pauses,
- Des phénomènes vocaux comme la toux ou le rire,
- Les phénomènes kinésiques tels que les gestes et la mimique,
- Des événements entièrement non linguistiques (par exemple, le bruit d'un camion faisant marche arrière sur la route à côté de la salle de conférence),
- Des changements de la qualité de la voix (la voix devient rauque ou aigue, par exemple).

[132] Le Règlement Général sur la Protection des Données (RGPD) (UE 2016/679) a été adopté le 27 avril 2016 par le Parlement européen. Ses dispositions ont été directement applicables à l'ensemble des 27 États membres de l'Union européenne à partir du 25 mai 2018. Aujourd'hui, une autorisation préalable est requise avant toute collecte de données personnelles.

[133] Nous comptons proposer en libre accès la reproduction littérale de dialogues choisis ainsi que des improvisations complètement libres, relatives à une situation définie par avance.

Nos transcriptions n'incluent pas les deux dernières caractéristiques car elles ne semblent pas pertinentes du point de vue de l'apprentissage. Les événements non linguistiques n'ont été inclus que lorsqu'ils étaient essentiels pour l'interprétation correcte de l'énoncé, lorsqu'il s'agissait par exemple d'un rire, d'un sourire ou d'un geste clarifiant le message. De plus, nous avons décidé de créer deux transcriptions de nos corpus parlés : une transcription « propre » et une transcription littérale. Les transcriptions « propres » sont constituées de textes cohérents, dans lesquels certaines répétitions, hésitations, ont été supprimées dans l'intérêt d'une meilleure lisibilité. Les transcriptions littérales servent quant à elle de base à une nouvelle série de vidéos dont nous présenterons le concept dans le paragraphe suivant.

Au cours de 2020, nous avons commencé à collecter les énoncés naturels d'une manière différente[134]. Dans un premier temps, nous enregistrons ou transcrivons l'interaction aussi précisément que possible (la plupart des rencontres de service sont assez courtes) et nous ajoutons les transcriptions au corpus. Nous les modifions, lorsque nécessaire, en supprimant les parties requérant des connaissances implicites ou trop d'explications en raison de l'absence de données visuelles. Nous fournissons également des informations supplémentaires concernant les phénomènes non-lexicaux (gestes, mimiques, objets). À un stade ultérieur, nous prévoyons d'enregistrer certaines de ces interactions avec les acteurs afin de générer du matériel multimédia proche des dialogues originaux. Cette procédure présente plusieurs avantages. Premièrement, il n'y a pas de problèmes liés à la protection des données ; deuxièmement, nous pouvons éviter de longues explications aux participants sur la raison pour laquelle nous voulons enregistrer leurs conversations. Troisièmement, la qualité des enregistrements est supérieure à celle des enregistrements authentiques et la situation peut être recréée dans la vidéo. En bref, nous pouvons obtenir de bien meilleurs résultats avec une procédure plus simple.

Toutes les transcriptions du sous-ensemble 3 (rencontres dans les lieux de service, conversations entre locuteurs natifs) peuvent être analysées manuellement ou avec des outils de corpus. L'analyse manuelle des textes complets peut fournir des informations sur les répétitions et variations relatives aux situations de communication, sur leur dynamique intrinsèque, sur leurs séquences typiques ainsi que sur les éléments du langage interactionnel. Elle peut également faire ressortir les mots, les unités multi-lexicales fréquents et quelques schémas grammaticaux.

[134] Malheureusement la pandémie a mis provisoirement fin pour l'instant à ce travail mais nous comptons le reprendre dès que possible.

2) Sous-ensemble 3(1) : rencontres dans les lieux de service

Être capable de mener à bien des rencontres dans les lieux de service est un objectif essentiel pour la majorité des apprenants en langues. On attend des apprenants qu'ils soient capables de participer à ces situations dès le niveau A1. Les enregistrements d'interactions de service constituent donc un corpus particulièrement utile pour les niveaux de compétences linguistiques inférieurs. Le sous-ensemble 3(1) comprend les transcriptions de dix à quinze interactions de la vie quotidienne (salon de coiffure, boulangerie, glacier) pour un total de huit à dix situations par niveau. Elles comprennent des rencontres dans des magasins, des cafés, des restaurants, des musées, des cinémas, par exemple, et durent entre 30 secondes et deux minutes, la durée moyenne d'enregistrement étant inférieure à une minute. Notre travail de construction de cette base de données nous a permis de réunir un total d'environ 320 minutes et de plus de 200 dialogues.

Ces conversations contiennent le vocabulaire thématique de base ainsi qu'un grand nombre d'unités multi-lexicales essentielles pour une communication réussie (Scott et Thompson 2000 ; Wray 2007). Travailler à partir de nombreuses interactions similaires présente l'avantage que les apprenants peuvent observer et apprendre les éléments linguistiques pertinents ainsi que leurs usages typiques beaucoup plus rapidement que s'ils devaient les déduire des conversations de la vie quotidienne. Par exemple, un sous-corpus de vingt-et-un courts dialogues entre des commerçants et des clients, enregistrés dans trois boulangeries différentes, aide les apprenants à identifier le langage transactionnel idoine, mais aussi à « enrichir les répertoires linguistiques [des apprenants] de manière à les préparer à des interactions imprévisibles en dehors du cours » (O'Keeffe et al. 2007 : 21).

Les trois exemples suivants (tableaux 180) donnent une idée de l'importance des répétitions dans ces interactions. Ils illustrent le fait que de nombreuses interactions de la vie quotidienne ne sont pas particulièrement complexes en termes de vocabulaire, mais que leur utilisation linguistique est hautement idiomatique et souvent ritualisée (Wray 2008). Ils montrent également comment les phénomènes non verbaux (gestes) ont été intégrés dans les transcriptions.

Dialogue 1	***Dialogue 1***
– **Tessék!**	- ***Que puis-je faire pour vous ?***
– A rozskenyérből kérek szépen egy kilót.	*- Je peux avoir un kilo de pain de seigle ?*
– **Más valamit adhatok?**	- ***Bien sûr. Avec ceci ?***
– Nem, köszönöm.	*- Non, merci.*
– **250 forint lesz**.	- ***Ça vous fera 250 forints.***

Dialogue 2	***Dialogue 2***
– Jó napot kívánok!	*- Bonjour !*
– Jó napot! Tessék!	***- Bonjour ! Que puis-je faire pour vous ?***
– Milyen pogácsáik vannak?	*- Quel genre de pogácsa (pâtisserie hongroise) avez-vous ?*
– Ez tepertős (*a jobb oldali pogácsákra mutat*), ez pedig túrós (*a bal oldali pogácsákra mutat*).	*- Ceux-ci (désigne les pogácsas de droite) sont avec des craquelins, et ceux-ci avec du fromage blanc (désigne les pogácsas de gauche).*
– A túrósból kérek szépen 20 dekát.	*- Je peux en avoir 200 grammes avec du fromage blanc, s'il vous plaît ?*
– Máris adom. (*leméri*) **Más valamit adhatok még?**	*- Tout de suite. (pèse les articles)* ***Avec ceci ?***
– Ennyi lesz, köszönöm.	*- Ce sera tout, merci.*
– **420 forint lesz**. Kártyás fizetés lesz?	***- Ça vous fera 420 forints.*** *Allez-vous payer par carte ?*
– Igen.	*- Oui.*

Dialogue 3	***Dialogue 3***
– Jó napot kívánok!	*- Bonjour !*
– Jó napot! Tessék!	***- Bonjour ! Que puis-je faire pour vous ?***
– Egy kakaós csigát kérek szépen.	*- Puis-je avoir un rouleau de cacao (pâtisserie hongroise) s'il vous plaît ?*
– Nincsen már kakaós csiga, csak diós.	*- Nous n'avons plus de rouleau au cacao, seulement des rouleaux aux noix.*
– Nem baj, az is jó lesz, köszönöm.	*- Ce n'est pas grave, alors je vais en prendre, merci.*
– Egyet kér?	*- Vous en voulez un ?*
– Igen.	*- Oui.*
– 140 forint lesz.	***- Ça vous fera 140 forints.***

Tableau 180 : Exemples de conversations à la boulangerie.

Ces interactions renferment plusieurs phrases utiles pour des interactions similaires : elles illustrent comment demander quelque chose, comment répondre aux questions du commerçant et comment payer les marchandises. L'apprenant peut donc observer ces phrases dans plusieurs dialogues et être sensibilisé ainsi à leur usage.

3) Sous-ensemble 3(2) : conversations entre locuteurs natifs

Ce sous-ensemble comprend des conversations de locuteurs natifs sur différents sujets. Il contient les transcriptions de nombreux dialogues enregistrés avant la rédaction des manuels, par exemple, la conversation entre Anna et Tímea avant leur leçon de tchèque (voir la section A.2). Les sujets ont été sélectionnés à partir de la liste thématique du CECRL.

Les enregistrements comprennent des conversations avec des amis, des collègues et des membres de la famille, les interactions étant d'une longueur entre 30 secondes et quatre minutes. L'enregistrement commence généralement par une question liée à un sujet figurant dans le CECRL, par exemple : « As-tu passé un bon week-end ? » « As-tu des projets pour ce soir ? », etc. L'enregistrement a été interrompu lorsque les participants ont commencé à s'écarter sensiblement du sujet. Au moins six enregistrements ont été réalisés sur dix sujets par niveau (environ 180 enregistrements), soit un total de 300 minutes, avec des chevauchements occasionnels entre les niveaux. La plupart des interactions ont été largement conservées sous leur forme originale – bien que parfois raccourcies – ; les modifications ont consisté à supprimer les allusions à des connaissances implicites (personnes ou lieux connus de tous les participants mais non de l'apprenant), ainsi que les expressions idiomatiques dépassant le vocabulaire de base.

Collecter des données linguistiques orales a réservé quelques surprises qui ont attiré notre attention sur la nature de l'oralité fondamentalement différente de l'usage langagier écrit. Nous nous servirons d'une conversation enregistrée lors d'une pause déjeuner au bureau pour mettre en évidence certaines caractéristiques-clés de ces interactions informelles. La conversation était dans cet exemple destinée à compléter un exercice du chapitre 4 du manuel A1 invitant les apprenants à parler d'une ville où ils pourraient s'imaginer vivre. Le livre énumère dix arguments pour et dix arguments contre liés à plusieurs aspects de la question (infrastructures, sécurité publique, événements culturels et sportifs, éducation). Lorsque le sujet a été abordé pendant la pause déjeuner (grâce à l'intervenant A qui a accepté d'en discuter à un moment opportun), nous nous attendions à ce que les participants mentionnent au moins un ou deux arguments figurant dans le manuel[135]. Voici ce qu'ils ont dit à la place (tableau 181) :

A: Te hol szeretnél élni?	*Où aimerais-tu vivre ?*
B: Úgy érted, melyik városban?	*Tu veux dire, dans quelle ville ?*

[135] Les arguments proposés dans le manuel ont été collectés sur un forum où les participants échangeaient à propos de leurs villes préférées.

A: Aha, úgy.	*Ouais. C'est ça.*
B: Szerintem Madridban.	*Peut-être à Madrid.*
C: Mit mondtál, Madridban?	*Tu as dit à Madrid ?*
B: Aha.	*Oui.*
C: Ja, jó... De miért? Mert nagyváros?	*Ah, d'accord... Mais pourquoi ? Parce que c'est une grande ville ?*
B: Igen.	*Oui.*
C: Értem.	*Je vois.*
B: Miért? Szerinted nem jó hely?	*Pourquoi ? C'est pas bien à ton avis ?*
C: De, biztos az… Nem ismerem.	*Oh, mais si, certainement... Je ne la connais pas.*

Tableau 181 : Conversation authentique sur le thème « Où aimerais-tu vivre ? »

Dans ce dialogue, la proportion de langage véhiculant des informations est étonnamment faible : la seule chose que nous apprenons est que B aimerait vivre à Madrid, probablement parce que c'est une grande ville et que C ne connaît pas Madrid. Il est identique à la conversation semi-authentique au niveau B1 sur la capacité de concentration, citée plus haut. Comme elle, cette discussion (comme de nombreuses autres dans ce sous-ensemble) remplit une fonction essentielle : tout en échangeant peu d'informations, *les locuteurs se renvoient des signaux affirmant qu'ils s'écoutent et souhaitent poursuivre la conversation*. Cet échange, comme beaucoup d'autres dans ce sous-ensemble, illustre que les partenaires de conversation attendent généralement un certain degré de « légèreté » et de lien relationnel, en plus (quelquefois même plutôt que) des informations factuelles. Comme l'observe McCarthy (2003 : 10), il semble que les individus font volontairement un effort quand il s'agit d'être sociables. Apprendre à être sociable dans la langue-cible est donc essentiel, et les interactions orales enregistrées peuvent être d'une aide précieuse dans ce processus.

Le travail avec des conversations authentiques et semi-authentiques, informelles révèle des différences significatives entre l'usage oral et l'usage écrit de la langue. Les manuels ne pouvant présenter que certaines caractéristiques du langage parlé, il est difficile pour l'apprenant de construire ses connaissances de cette variété langagière à partir des textes proposés dans ces livres, d'où le grand intérêt du corpus oral pédagogique ainsi constitué.

Ce chapitre a présenté le corpus oral pédagogique qui complète la série de manuel « MagyarOK » et souligné les avantages qu'il offre pour l'étude du langage parlé. Nous avons montré comment

les dialogues de manuels, les entretiens avec des locuteurs natifs sélectionnés et les improvisations d'acteurs sur un thème peuvent constituer la base de sous-ensembles de données semi-authentiques à caractère naturel. Ces sous-ensembles constituent une riche source d'interactions à caractère naturel pour les niveaux inférieurs. L'apprenant peut y observer la dynamique typique des interactions, le vocabulaire-clé lié au sujet choisi ainsi que des éléments du langage interactionnel et enrichir, de façon graduelle, ses compétences linguistiques. Dans notre exposé, nous avons également souligné les principales difficultés liées à la collecte de données authentiques. Ce processus est plus complexe que la collecte de textes écrits, en raison de plusieurs facteurs. Une fois résolue les questions relatives à la protection des données, il est nécessaire de s'assurer une haute qualité des enregistrements et de trouver un mode de transcription des données qui produise des textes utiles aux apprenants. Comme solution alternative, nous avons suggéré d'enregistrer les interactions authentiques (par écrit ou avec un outil d'enregistrement) et de laisser les acteurs rejouer les scènes.

Au chapitre suivant, nous passerons de la construction du corpus à des activités basées sur des corpus pour l'enseignement des langues. Ces activités impliquent l'usage des outils numériques d'analyse et peuvent compléter l'analyse manuelle des textes dans leur intégralité, présentée dans ce chapitre et au chapitre 13.

Chapitre 15 : De l'observation à la pratique : analyse linguistique et textes-modèles

Les chapitres 13 et 14 ont présenté le processus de construction du corpus pédagogique écrit et oral ainsi que les bénéfices qu'apporte la lecture des textes complets. Ce chapitre présentera des activités qui, d'une part, exploreront les corpus pédagogiques avec des outils numériques et, de l'autre, montreront comment l'apprenant peut produire ses propres récits à partir des textes-modèles inclus dans le corpus pédagogique. Les activités de l'analyse outillée reposent majoritairement sur les méthodes de l'« Apprentissage sur corpus » (data-driven learning), en y ajoutant quelques nouvelles composantes. La présentation du cadre méthodologique de cette approche est suivie par des activités dont la première série repose sur l'exploration des schémas alors que la deuxième reprend le travail autour de la répétition et de la variation dans le corpus, travail outillé cette fois-ci. Dans la troisième et dernière série, nous proposerons des activités axées sur la production langagière : ces tâches guident l'apprenant dans l'utilisation des textes du corpus pour améliorer la qualité de ses propres textes. Toutes les activités présentées nécessitent bien évidemment un corpus pédagogique, créé explicitement pour l'usage par l'apprenant lui-même.

A) L'Apprentissage sur corpus « revisité »

1) L'Apprentissage sur corpus – de quoi s'agit-il ?

Il est clair que chaque exercice mené en classe devrait rapprocher les apprenants de leur objectif ultime qui est la maîtrise de la langue-cible. Les activités utilisant le corpus pédagogique peuvent apporter une contribution précieuse à la réalisation de cet objectif. Elles stimulent les réflexions sur la langue-cible, consolident le vocabulaire-clé par des rencontres répétées et offrent aux apprenants la possibilité d'intégrer les résultats de leurs observations dans leurs propres textes. Citant Spöttl et McCarthy (2003) et O'Keeffe et al. (2007), Salazar (2014 : 158) affirme qu'« *au moins un certain degré d'analyse linguistique consciente est nécessaire pendant le processus d'apprentissage* » et que « le cours de langue est exactement le lieu où ce type de réflexion peut et doit être encouragé » (nous soulignons, notre traduction).

« L'Apprentissage sur corpus » (*data-driven learning, DDL*) offre un cadre méthodologique permettant d'intégrer de nombreuses facettes de l'analyse linguistique dans l'apprentissage des langues. Le terme a été inventé par Johns (1991) qui résume les éléments essentiels de cette approche comme suit :

« Premièrement, nous devons fournir des occasions adéquates [...] aux étudiants de soulever des questions [...]. Deuxièmement, nous devrions essayer de rendre notre enseignement transposable de sorte que les stratégies développées en classe pour comprendre le fonctionnement de la langue soient également applicables en dehors de la classe. » (p. 295, notre traduction)

D'après Johns (1991 : 2), « l'apprenant en langues est aussi, essentiellement, un chercheur dont l'apprentissage doit être guidé par l'accès aux données linguistiques – d'où le terme "Apprentissage sur corpus" pour décrire cette approche ». L'hypothèse sous-tendant cette approche est que l'apprentissage efficace des langues est lui-même une forme de recherche linguistique, et que l'analyse des lignes de concordance (l'outil favorisé par Johns) offre une ressource unique pour la stimulation des stratégies d'apprentissage inductif – en particulier les stratégies de perception des similitudes et des différences (Johns 1994 : 297).

Dans un objectif plus limité mais peut-être plus réaliste, Frankenberg-García (2014 : 130) définit le DDL comme « la capacité à utiliser des données de corpus pour comprendre – au lieu de se faire dire – ce que signifient les mots ou comment ils sont utilisés » (notre traduction). C'est dans ce sens que nous nous y référerons par la suite.

Les questions qui conviennent comme point de départ pour l'exploration du corpus pédagogique sont les mêmes que celles que nous avons proposées pour l'enseignant dans la Partie II de cette thèse. En voici quelques exemples : « Comment utilise-t-on le mot X ? » « Que signifie exactement X ? » « Quelle est la différence entre X et Y ? » « Peut-on dire ... ? » Il s'agit donc des questions auxquelles il n'est pas possible de répondre de manière adéquate à l'aide d'autres moyens tels que les livres de grammaire, les dictionnaires ou l'intuition du locuteur expert. La raison de l'efficacité limitée de ces outils est qu'ils ne peuvent fournir que quelques exemples, des règles simples et/ou des définitions alors que les réponses judicieuses à ces questions spécifiques tendent à émerger de l'analyse d'un grand nombre d'énoncés, d'où l'intérêt de l'analyse de corpus.

L'étude d'un nombre relativement important d'occurrences d'un même élément lexical offre des moments d'« exposition condensée » (*condensed exposure*) (Gabrielatos 2005 : 10), comme déjà évoqué à plusieurs reprises dans cette thèse. Le fait d'observer le même élément dans des environnements textuels différents présente plusieurs avantages : ces rencontres systématiques

peuvent clarifier son ou ses sens. Le potentiel de l'approche sur corpus pour l'apprentissage lexical est également souligné par Allan (2009 : 24) : « [cette méthode] est particulièrement utile en ce qu'*elle offre aux apprenants de multiples expositions aux mots en contexte, leur permettant d'approfondir leurs connaissances des mots grâce aux informations sur ses collocations, son environnement textuel et les registres dans lesquels ils ont tendance à émerger* » (notre traduction, nous soulignons). Nous nous proposons de compléter cette liste par des informations grammaticales, car nous avons vu dans les chapitres 8 à 11 que considérer grammaire et lexique en tant que deux facettes interconnectées des énoncés peut être très bénéfique pour se forger une image plus précise du comportement langagier des natifs.

Johns (1991 : 5) suggère une méthodologie en trois étapes pour la consultation de corpus ; la première étape étant l'observation, la deuxième la classification et la troisième la généralisation. Lors de la phase d'observation, les apprenants identifient un ou plusieurs schémas (lexicaux, grammaticaux, sémantiques, pragmatiques), puis ils regroupent leurs résultats et ils évaluent si les schémas observés peuvent être généralisés à d'autres exemples.

Le processus d'analyse peut être aidé par l'enseignant qui identifie et rédige les échantillons de corpus afin qu'ils soient utiles à l'apprenant. Cela change le rôle de l'enseignant (celui qui « sait ») en celui du guide qui aide l'élève à réaliser sa recherche linguistique. Une fois que les apprenants sont familiarisés avec les outils et qu'ils savent à quels types de questions il est possible de répondre en utilisant le corpus, ils peuvent l'explorer par eux-mêmes. D'après Boulton (2009 : 37), « il se peut même, dans certains cas, que l'apprentissage soit plus efficace sans enseignant, c'est-à-dire lorsque les apprenants découvrent des choses par eux-mêmes » (notre traduction). Les avantages de cette approche à plus long terme incluent également une conscience métalinguistique et métacognitive plus aiguisée de l'apprenant (Aston 2001), l'augmentation de ses compétences linguistiques ou encore l'amélioration de sa capacité à gérer le langage authentique (Boulton 2008, 2010b).

Les réserves concernant l'Apprentissage sur corpus s'articulent autour de quatre domaines (Gilquin et Granger 2010) : (1) les préoccupations concernant la logistique ; (2) le point de vue de l'enseignant ; (3) le point de vue de l'apprenant ; (4) la nature des données à explorer. Nous examinerons ces points dans les paragraphes suivants.

Concernant le premier point, Guilquin et Granger constatent que l'Apprentissage sur corpus nécessite l'accès aux corpus et aux logiciels appropriés ainsi que des connaissances techniques de

la part de l'usager[136]. Dans ce contexte, une interface transparente, facile à manipuler est indispensable. Or, le problème est que la majorité des logiciels gratuits ont été conçus pour les explorations de la langue par les linguistes, pour cette clientèle ; le travail sur l'esthétique et sur la transparence est secondaire par rapport aux fonctionnalités accessibles. En revanche, les apprenants et les enseignants se servent de ces outils comme ils utilisent d'autres applications. Par conséquent, ils ont les mêmes attentes au niveau du design et de la pratique : apprendre à travailler avec les nouveaux outils doit être un processus simple et rapide car le corpus est un outil qui entre en compétition avec des applications plus accessibles à première vue. De plus, dans le cadre de l'enseignement des langues, le temps requis pour atteindre le niveau choisi est plutôt limité : si l'enseignant et/ou l'apprenant ne sont pas immédiatement attirés, il est peu vraisemblable qu'ils se donneront le temps de regarder les logiciels de plus près.

Le deuxième point évoque les implications pour le rôle de l'enseignant. Comme mentionné précédemment, dans le cadre de l'Apprentissage sur corpus, les enseignants sont considérés comme des guides qui aident les apprenants à formuler leurs questions et leurs hypothèses et apportent leur soutien lorsque les apprenants analysent les données dans le corpus. Les apprenants sont considérés quant à eux comme des constructeurs actifs plutôt que comme des destinataires passifs de connaissances. Timmis (2015 : 138) remarque ainsi que « dans certains contextes éducatifs, ce changement de rôle peut exiger une adaptation considérable de la part des apprenants et des enseignants ». Il est vrai que cette approche ne peut fonctionner si le cours et le manuel reposent sur une approche « classique » (dans laquelle l'enseignant explique et l'apprenant se cantonne à exécuter les exercices).

Le troisième point aborde la perspective de l'apprenant. Il est important de noter que, tout comme les corpus pédagogiques ne remplissent pas les mêmes objectifs que les corpus à fins linguistiques, *l'apprenant n'est pas un expert de l'analyse de corpus*. Ainsi, pour les niveaux A1-B1, la recommandation de Chambers (2019 : 472–3) selon laquelle nous devrions « donner aux apprenants des compétences utiles en matière de consultation et d'analyse de corpus sans essayer de faire d'eux des linguistes de corpus » semble des plus judicieuses. Pour illustrer certaines implémentations possibles de cette approche, nous présenterons des exemples concrets d'activités dans les pages suivantes.

[136] Boulton (2010a) propose une analyse sur papier comme alternative mais cette méthode utilise également les corpus traités par des logiciels comme point de départ.

Le quatrième point concerne la nature des données linguistiques à explorer. Certains défenseurs de l'Apprentissage sur corpus tels que Johns (1991), Chambers et al. (2011) ou Tyne (2012) recommandent l'utilisation des corpus à fins linguistiques pour ces explorations en soulignant la nécessité de travailler avec des exemples réels de langage. Pour les étudiants de langues et pour l'entraînement à l'écriture académique, cette approche peut convenir, cependant nous avons démontré dans les chapitres 1 et 2 que ces corpus ne sont pas vraiment appropriés pour les niveaux de compétences linguistiques inférieurs. En effet, ils contiennent, d'une part, du vocabulaire complexe, ce qui rend difficile la recontextualisation des énoncés ; et, de l'autre part, les textes ne sont pas systématisés, ni disponibles dans leur intégralité. Les corpus à fins pédagogiques construits selon les recommandations exposées au chapitre 2 (Partie I) permettent aux apprenants de consulter des exemples qui leur sont accessibles et pertinents.

Les principaux types d'activités dans le cadre de l'Apprentissage sur corpus tels que proposé dans cette thèse sont les suivants :

- Recherche à partir d'une question lexicale
- Recherche à partir d'une question de grammaire
- Pratique des éléments observés
- Analyse de plusieurs textes autour du même thème dans leur intégralité
- Observation des répétitions et des variations linguistiques dans ces textes

La majorité des activités présentées peuvent être mises en œuvre dès les premiers stades de l'apprentissage des langues.

B) Explorer les schémas

Dans les pages suivantes, nous présenterons quelques exercices classiques de l'« Apprentissage sur corpus » en démontrant comment leur efficacité peut être accrue par l'utilisation d'un corpus pédagogique au lieu d'un corpus à fins linguistiques. Les activités 1 et 2 impliquent que les étudiants fassent des observations linguistiques systématiques et détectent des schémas. Ces tâches ne constituent toutefois que la première étape de l'exploration du corpus. Elles doivent être suivies d'une pratique active de l'élément linguistique en question et de son ou de ses environnements textuels. C'est ce que visent les activités 3 à 6. Les activités 3 et 4 utilisent le corpus comme support pour des tâches que les apprenants connaissent bien (par exemple, des exercices où il faut compléter des phrases) et leur donnent une nouvelle tournure. Les activités 5 et 6 illustrent

comment le corpus peut aider à réviser et à consolider le vocabulaire de base. Les activités montrent également comment explorer *les répétitions et les variations dans les textes avec les outils numériques*.

1) Commencer l'exploration du corpus par une question liée au lexique

La question de l'apprenant : *Comment utilise-t-on l'adjectif « nagy » ? Que signifie-t-il exactement ?*
Outil : *Word Sketch*

Comme nous l'avons vu tout au long de cette thèse, un des résultats importants en linguistique de corpus est la prise de conscience que les mots apparaissant avec une fréquence particulièrement élevée sont susceptibles d'avoir de nombreux collocatifs et des significations différents (Nation 2013 ; Szudarski 2017). L'adjectif hongrois « nagy » ne fait pas exception à cette observation. Selon le Dictionnaire hongrois-français (szotarnet.hu, 2021), il peut ainsi être traduit par « grand », « large », « grave », « significatif » ou « gros ». C'est l'environnement textuel plus étendu qui permet de désambiguïser son sens dans un énoncé donné.

La méthode d'exploration du corpus par l'apprenant suit les principes proposés par Sinclair et adaptés par Hoey[137] (comme nous les avons appliqués lors de l'exploration du corpus par l'enseignant pour établir les profils du mot, voir les chapitres 8 à 12). L'intérêt d'utiliser le même procédé est que l'enseignant le connaît déjà et peut facilement initier l'apprenant à son usage. Nous souhaitons également démontrer qu'il fonctionne non seulement pour l'enseignant mais aussi pour l'apprenant, à condition que le corpus soit approprié.

La première étape consiste donc dans notre exemple à observer les unités multi-lexicales avec « nagy ». Cet adjectif apparaît 361 fois dans le corpus des manuels (soit 885 par million), ce qui indique son importance : il occupe la place 5 dans la liste des adjectifs les plus fréquents de notre corpus et la place 1 dans le corpus « huTenTen12 ». L'outil « Word Sketch » permet d'identifier ses collocatifs fréquents (voir le chapitre 3 pour la description plus détaillée de l'outil). À des fins d'illustration, nous présentons dans le tableau 182 une seule colonne : les noms fréquemment modifiés par « nagy ».

[137] Pour une description plus détaillée, voir la section C) du chapitre 4 sur la construction du profil de l'élément linguistique choisi.

succès

a été un grand succès

partie

la majeure partie des exercices écrits

maison

dans une belle grande maison

ville

sont également de grandes villes

appartement

un appartement plus grand

calme

ennuyeux. C'est trop calme, rien d'intéressant (litt. : Le calme est trop grand.)

agitation, émoi

Ces articles ont suscité un tel émoi que

jardin

qui a un grand jardin bureau

Nora : Un beau et grand bureau.

bureau

Nora : Un beau bureau grand.

Gabor

trafic

parce qu'il n'y a pas beaucoup (lit. grand) de trafic.

musée

la ville entière est un grand musée. Dans le centre

Tableau 182 : Noms modifiés par « nagy » dans le corpus A1-B1 de « MagyarOK ».

Un des avantages – et l'une des originalité – de la construction d'un corpus pédagogique à partir des chapitres (voire des manuels entiers) réside dans le fait que les apprenants sont déjà familiarisés avec son contenu durant le cours, les listes issues de corpus ne contenant ainsi que des collocations

connues. En cliquant sur les trois points à côté de la collocation choisie, des lignes de concordance apparaissent. Comme le corpus pédagogique est plutôt petit, le nombre de phrases par collocation reste raisonnable. Les apprenants peuvent lire et sélectionner les phrases qu'ils trouvent particulièrement utiles et, ils peuvent également consulter le texte complet dans le manuel si nécessaire[138].

À l'étape suivante, les apprenants classent leurs résultats. Une liste de questions peut les guider dans ce processus. Adapté des travaux de Flowerdew J. (2009), Hoey (2005), Sinclair (2004b) et Stubbs (2009), les apprenants peuvent utiliser les instructions suivantes pour établir les profils de mots. Les explications que les apprenants reçoivent sur la pertinence d'une caractéristique donnée sont incluses entre parenthèses :

- Observez les mots accompagnant l'élément sélectionné avec une certaine régularité. De quels mots s'agit-il ? (Chaque mot a tendance à apparaître avec certains autres mots. Ce sont des composantes lexicales avec lesquelles il forme des « collocations »).
- Observez les schémas grammaticaux. Y a-t-il un temps ou un mode de verbe, ou d'autres caractéristiques grammaticales qui apparaissent avec une certaine régularité ? (Il est souvent possible d'associer certains schémas grammaticaux aux mots et aux unités multi-lexicales. Ce sont leurs composantes syntaxiques ou « colligations »).
- Observez le sens du mot. Le mot et son environnement textuel font-ils référence à quelque chose de positif ou de négatif, d'abstrait ou de concret ? Y a-t-il d'autres caractéristiques notables liées à leur signification ? (Les mots et les multi-lexicales ont des significations typiques. Ce sont leurs « composantes sémantiques »).
- Observez le contexte de communication. Le mot est-il utilisé dans des situations de communication formelles ou informelles, écrites ou orales ? (Il existe des situations typiques dans lesquelles un mot et une unité multi-lexicale émergent. Ce sont ses « composantes pragmatiques »).

Sur la base de ces critères, le profil du mot « nagy » pourrait ainsi ressembler aux tableaux présentés dans la Partie II de cette thèse (tableau 183) :

[138] Les numéros de pages sont inclus dans le corpus ; ils précédent chaque exercice.

Collocations fréquentes : noms à sens concret	*város (ville), ház (maison), kert (jardin), múzeum (musée)*
Collocations fréquentes : noms à sens abstrait	*rész (partie), siker (succès), probléma (problème), gond (souci), öröm (plaisir), szerep (rôle), szükség (besoin), dolog (chose), meglepetés (surprise), segítség (aide), előny (avantage), nap (jour), változás (changement), mérték (mesure), ember (homme)*
Colligations	**Adj + N.** Typiquement superlatif : *a legnagyobb rész/probléma* (*la majeure partie, le plus gros problème);* Typiquement comparatif: *nagyobb mértékben* (*dans une plus grande mesure);* Typiquement forme de base : *nagy meglepetés* (*grande surprise), nagy segítség (aide significative), nagy siker (grand succès), nagy gond (gros souci), nagy nap (grand jour), nagy ember (grand homme)* etc.
Composantes sémantiques : noms à un sens concret	Fait référence à la taille ou à la proportion de quelque chose
Composantes sémantiques : noms à un sens abstrait	(1) Choses; événements ou personnes importantes et/ou agréables; (2) Exprime la gravité d'un problème
Composantes pragmatiques	Pas de composante pragmatique particulière.

Tableau 183 : Le profil de l'adjectif « nagy ».

L'étape suivante est la généralisation. Les apprenants vérifient leurs observations en étudiant d'autres exemples et en les ajustant, si nécessaire.

L'enseignant (et plus tard aussi l'apprenant) peut comparer les collocations fréquentes dans le corpus pédagogique avec celles d'un grand corpus général et montrer aux apprenants les résultats obtenus. Dans le corpus « huTenTen2012 », une recherche donne les collocations suivantes pour « nagy » : partie, succès, problème, trouble, plaisir, rôle, besoin, chose, surprise, aide, avantage, surface, jour, changement, étendue. Bon nombre de ces collocations sont présentes dans le corpus du manuel, avec moins d'occurrences cependant[139].

L'exploration ne doit pas s'arrêter là ; c'est là l'un des plus grands avantages de la consultation de corpus. Comme souligné par Charles (2007 : 297), une fois les étudiants familiarisés avec les outils et les procédures, ils peuvent être encouragés à explorer le corpus par eux-mêmes. De nombreuses requêtes à la base de données auront tendance à susciter de nouvelles questions, qui, à leur tour, pourront conduire à de nouvelles observations non envisagées à l'origine par l'apprenant (ni éventuellement par l'enseignant). Dans le cas de « nagy », les apprenants peuvent rassembler des unités multi-lexicales plus longues, par exemple en enrichissant les collocations sélectionnées à

[139] Cette confrontation à un corpus général peut ainsi renforcer la confiance de l'apprenant.

deux composantes (adjectif et nom) avec des verbes typiques. Les unités multi-lexicales dominantes dans le corpus avec l'expression « nagy siker » (grand succès) sont : « nagy sikert arat » (atteindre un grand succès), « nagy sikernek számít » (être considéré comme un grand succès), « nagy sikere van » (a un grand succès). Une analyse rapide du corpus « huTenTen2012 » révèle que ce sont également les unités multi-lexicales les plus fréquentes dans ce corpus[140].

2) Commencer l'exploration du corpus par une question grammaticale

La question de l'apprenant : *« Pourquoi la conjonction « pedig » est-elle parfois placée au début de la proposition et parfois plus à droite ? Ces deux positions sont-elles interchangeables ? »*
Outil : *Concordancier*

L'exploration du corpus peut aussi commencer par une question concernant la grammaire (même si la réponse, comme dans le cas de la première activité débutant par une question lexicale, comprendra à la fois des éléments lexicaux et grammaticaux). Nous utiliserons la conjonction « pedig » pour illustrer ce processus. « Pedig » peut être traduit par « mais », « et » ou « alors que ». Cependant, comme d'autres conjonctions sont les équivalents exacts de « mais » (de), « et » (és) ou « alors que » (viszont), les apprenants sont évidemment perplexes. L'analyse approfondie d'un certain nombre d'exemples permet de mettre en lumière les particularités de ce mot.

Le corpus A1-B1 de « MagyarOK » contient 236 occurrences de « pedig », le corpus A1-A2 en contient 161 et le corpus A1 84. Même le nombre des exemples dans le corpus d'un seul niveau est donc significativement plus élevé que ce que peut offrir un simple dictionnaire ou une grammaire. Le tableau suivant montre dix lignes de concordance du corpus A1. Le mot-clé en contexte (KWIC) est marqué en rouge (ici : le mot « pedig ») (tableau 184) :

Spanyolországban. Gyönyörű volt az óceán, a régi város, a spanyolok	**pedig**	nagyon kedvesek. Szeretnék még egyszer odautazni. Amikor külf
n, kávézom, elolvasom a híreket, felkészülök az óráimra, a tanítás után	**pedig**	sokszor sétálok egy nagyot az iskola melletti szép parkban. Néha
án jól megtanulni, de szeretném folytatni a tanulást. A következő célom	**pedig**	az, hogy olaszul is megtanuljak. Szerintem nagyon igaz az a mon
véből és az internetről is meg lehet tanulni. Sajnos nekem nem sikerült,	**pedig**	néztem sok Spongya Bobot. Ezért most magánórákra járok, és pr
munkalehetőségek motiválják, Spanyolországban és Franciaországban	**pedig**	sok britnek van második otthona, és ez is segíti a nyelvtanulást. N
franciául, de az általános iskolában oroszul tanultam, a gimnáziumban	**pedig**	olaszul. Még sok mindent értek olaszul, ha valaki lassan beszél. N
grosszabb nyelvtanulója! Soha nem tudtam leülni és szavakat magolni,	**pedig**	szeretem a nyelveket. Az iskolában németül és franciául tanultam,
az ő nyelvét. Ezért tudok néhány szót, mondatot kínaiul és arabul, most	**pedig**	szerbül és spanyolul tanulok. Spanyolul már régóta szeretnék meg
volt: 6 órától este 8 óráig dolgoztam. 5 óráig voltam az iskolában, 5-től	**pedig**	otthon tanítottam angolt és magyart. Ma már nincs átlagos munka
ak? Adél: Én egy eszpresszót kérek, és két deci narancslevet. Nóra: Én	**pedig**	egy hosszúkávét és egy szénsavas ásványvizet. Pincér: Máris hoz

[140] Il convient de noter que les résultats fondés sur le corpus de manuels ne coïncideront pas systématiquement avec ceux des grands corpus généraux, car la portée du corpus de manuels est limitée : il offre un aperçu de l'utilisation d'éléments lexicaux liés à des situations de la vie quotidienne.

(1 L'océan et la vieille ville étaient magnifiques ***et*** *les Espagnols étaient très sympathiques.*

2 Je me lève tôt, puis je prends un café et je lis les nouvelles, je prépare mes leçons ***et****, après mes leçons, je me promène dans le parc à côté de l'école.*

3 Je n'ai pas appris beaucoup d'espagnol, mais j'aimerais continuer à le faire. ***Mais*** *mon prochain objectif est d'apprendre l'italien.*

4 Mes parents pensaient que l'on pouvait apprendre l'anglais en regardant la télévision. Je n'ai pas réussi ***bien que*** *j'aie regardé beaucoup de « Bob l'éponge ».*

5 Les apprenants de langues sont motivés par de bonnes opportunités d'emploi, ***et*** *de nombreux Britanniques ont une résidence secondaire en Espagne ou en France, ce qui est également motivant.*

6 J'ai appris le russe à l'école primaire ***et*** *l'italien au lycée.*

*7 Je ne pourrais jamais m'asseoir et apprendre des mots****, même si*** *j'aime les langues.*

8 Je connais quelques mots en arabe et en chinois ***et*** *maintenant j'apprends le serbe et l'espagnol.*

9 J'étais jusqu'à 17 heures à l'école ***et*** *après 17 heures, je donnais des cours à la maison.*

10 Je vais prendre un Expresso et un jus d'orange. - ***Et*** *je prendrai un Americano et une eau minérale.)*

Tableau 184 : Lignes de concordance avec la conjonction « pedig ».

Comme dans la première activité, les apprenants identifient, dans un premier temps, les principales caractéristiques du mot. Pour cela, ils sont invités à lire au moins vingt lignes de concordance avec « pedig ». Si le co-texte n'est pas clair, l'apprenant peut cliquer sur le mot-clé ; cela ouvre un nouveau tableau affichant l'environnement textuel plus large (à savoir les phrases avant et après la phrase contenant le KWIC). L'enseignant peut faciliter l'analyse des occurrences en posant des questions ciblées, par exemple : « Où le mot est-il placé dans la phrase ? Qu'est-ce qui le précède et qu'est-ce qui le suit ? » « Y a-t-il des similitudes dans la structure de la première et de la deuxième partie de la phrase ? » « Y a-t-il une différence dans la signification de « pedig » selon l'ordre des mots ? »[141]

L'étape suivante est la classification. Dans le cas de « pedig », deux grands groupes émergent : la conjonction est placée en position 1, au début d'une phrase ou d'une proposition (phrases 3, 4 et 7) ou en position 2, après un mot ou une phrase en position 1 (toutes les autres phrases). Sur la base des phrases 3, 4 et 7, les apprenants peuvent arriver à la conclusion (correcte) que lorsque « pedig » est placé en position 1, il est traduit par « alors que », « bien que ». Dans toutes les autres

[141] L'enseignant peut également télécharger les lignes de concordance, les mettre en groupes et demander aux apprenants d'identifier les critères selon lesquels les groupes sont organisés.

phrases, le sens du mot est plus proche de « et » (« és »), à la différence que « pedig » indique non seulement une addition mais aussi un contraste ; deux éléments de la proposition s'opposent à deux éléments de la première partie de la phrase. Par exemple : russe/école primaire ↔ italien/lycée ; jusqu'à 17 heures/à l'école ↔ après 17 heures/à la maison. L'un des mots/expressions contrastés précède immédiatement la conjonction, l'autre suit.

Une fois qu'ils ont répondu aux questions listées ci-dessus, les apprenants généralisent leurs observations. Pour cela, ils peuvent étudier d'autres exemples dans le corpus et vérifier si ceux-ci confirment leur classification[142]. Le résumé des résultats pourrait ressembler aux tableaux 185 et 186 :

Pedig (1)

Collocations	Pas de collocations particulières.
Colligations	**MOT/EXPRESSION + *pedig* + proposition** *Pedig* est le deuxième mot dans la phrase. Un des deux éléments opposés/confrontés le précède.
Composantes sémantiques	Deux items dans la première phrase sont opposés/comparés à deux autres dans la deuxième : *A régi város gyönyörű volt, a spanyolok **pedig** nagyon barátságosak. (L'ancienne ville était magnifique **et** les Espagnols étaient très sympathiques.) Az általános iskolában oroszul tanultam, a gimnáziumban **pedig** olaszul. (J'ai appris le russe au collège **et** l'italien au lycée.)*
Composantes pragmatiques	Apparait souvent dans des récits, quand le locuteur raconte une histoire.

Tableaux 185 : Profils de la conjonction « pedig » (1)

Pedig (2)

[142] Si un mot ou une unité multi-lexicale important n'apparaît pas avec une fréquence suffisante dans le petit corpus pour faire émerger des schémas, l'enseignant peut enrichir le corpus avec des exemples (simplifiés) provenant d'un grand corpus général.

Collocations	Pas de collocations particulières. Des mots à sens négatif dans une des deux phrases (*nem (ne pas), ritkán (rarement) ...*)
Colligations	***pedig* + proposition** *Pedig* est le premier mot de la deuxième phrase.
Componsantes sémantiques	Opposition avec la première partie de la phrase, qui exprime souvent une sorte d'échec : *Soha nem tudtam szavakat magolni, **pedig** szeretem a nyelveket. (Je n'a jamais su apprendre des mots par coeur **bien que** j'aime les langues.) Nem sikerült megtanulni angolul, **pedig** néztem sok Spongya Bobot. (Je n'ai pas réussi à apprendre l'anglais **bien que** j'aie regardé plein de Sponge Bob.)*
Composantes pragmatiques	Discours exprimant l' attitude du locuteur envers une personne, une chose ou un événement. Décrit majoritairement une conséquence surprenante, plutôt frustrante.

Tableaux 186 : Profils de la conjonction « pedig » (2).

Les activités 1 et 2 démontrent que le fait de disposer d'un corpus pédagogique peut être bénéfique pour explorer l'usage et le sens d'un terme. Cette approche est particulièrement adaptée pour les mots fréquents pour lesquels même un petit corpus fournit un grand nombre d'occurrences. En enregistrant leurs résultats comme proposé ci-dessus, les apprenants peuvent produire un résumé détaillé du comportement linguistique de l'élément lexical ou grammatical choisi[143]. Que le même procédé puisse être appliqué dans les deux cas permet à l'apprenant d'acquérir au fil du temps une certaine expertise pour l'usage des outils et de perfectionner ses compétences analytiques, renforçant par là-même le rôle actif de l'apprenant-chercheur.

Les avantages vont cependant bien au-delà du développement des techniques d'analyse : ils améliorent les compétences linguistiques de l'apprenant, but ultime de toutes les activités proposées en cours de langues. Lors des activités 1 et 2, les apprenants font ainsi bien plus que « collecter » des informations relatives à un seul mot. Ils étudient un certain nombre d'exemples qui mettent tous en évidence la pertinence de l'environnement textuel, et ils apprennent à poser les « bonnes questions » lorsqu'ils explorent l'utilisation d'un élément choisi. En outre, ils revoient des expressions, des phrases et des textes de leurs manuels et peuvent ainsi mémoriser des unités multi-lexicales utiles.

[143] Différents points grammaticaux mériteront une attention particulière en fonction de la nature de la langue-cible ainsi que de la langue maternelle de l'apprenant. Le corpus du manuel peut être utilisé pour élucider ces points en présentant de nombreux exemples. Par exemple, l'observation de l'utilisation des articles en anglais peut être extrêmement bénéfique pour les locuteurs de langues sans articles (par exemple ceux des langues slaves), l'exploration de l'utilisation des temps des verbes en anglais est utile pour les locuteurs de langues ayant moins de temps (par exemple le hongrois, le chinois mandarin), et l'étude de l'ordre des mots en hongrois aide les locuteurs de langues ayant moins de flexibilité dans la structure des phrases (par exemple les langues romanes et germaniques).

3) Même mot, phrase différente

Outil : *Concordancier*

Cette activité se concentre sur l'ordre des mots et peut être utile aussi bien pour les langues dont l'ordre des mots est flexible que pour celles dont l'ordre des mots est fixe. C'est le seul exercice qui doit être préparé par l'enseignant. Les apprenants placent un mot ou une unité de plusieurs mots dans un certain nombre de phrases du corpus du manuel. Une fiche de travail peut être créée en téléchargeant les lignes de concordance et en y retirant le ou les mots sélectionnés (tableau 187).

Helyezze a 'pedig' szót a mondatba! Ügyeljen a szórend és a jelentés kapcsolatára!

1. Koncerten voltunk este, napközben csak pihentünk.
2. Amikor én fiatal voltam, még nem lehetett utazgatni. De szerettem volna!
3. Mindig meglepődöm, mennyi ember elmegy Albert előadásaira. Unalmasan ad elő, ráadásul nem artikulál.
4. Mindjárt elalszom, még tanulnom kell.
5. Erdély és a Balkán irányába innen indultak a vonatok. A Nyugati pályaudvarról Bécs és Párizs felé.

(Placez « pedig » dans les phrases. Faites attention au lien entre l'ordre des mots et le sens.

1. Nous sommes allés à un concert le soir, nous nous sommes reposés pendant la journée.

2. Quand j'étais jeune, je ne pouvais pas voyager aussi facilement. Je l'aurais aimé !

3. Je suis toujours surpris de voir combien de personnes assistent aux conférences d'Albert. Il est ennuyeux et ne s'exprime pas clairement.

4. Je m'endors, je dois étudier.

5. D'ici, les trains partaient pour la Transylvanie et les Balkans. De la gare de l'Ouest, ils partent pour Vienne et Paris.)

Tableau 187 : Exercice sur l'ordre des mots à partir d'exemples de corpus.

Une fois les phrases complétées et vérifiées, les apprenants peuvent les utiliser de manière créative. Ils peuvent, par exemple, inventer de nouvelles fins aux phrases, par ex : « Nous sommes allés à un concert le soir et *nous avons fait une randonnée* dans la journée ». Ils peuvent consulter le corpus et construire leurs phrases en cherchant des exemples avec les mots qu'ils souhaitent utiliser. De cette façon, ils apprennent aussi à utiliser les corpus pour les aider à formuler leurs énoncés. Enfin, en s'éloignant du seul travail sur le corpus, ils peuvent relire et écouter des textes sélectionnés contenant la conjonction « pedig », pour renouveler le vocabulaire qui y est abordé.

4) Trouver le mot manquant

Outil : *Concordancier*

Cet exercice est une variation de l'activité précédente, à la différence que ce sont cette fois les apprenants qui créent les fiches de travail pour leurs camarades. Pour cela, ils copient quelques phrases sélectionnées du corpus et en enlèvent le mot-clé. Pendant la leçon, chaque apprenant complète au moins une fiche créée par un autre élève et vérifie ses réponses avec le Concordancier. Cette activité permet d'approfondir les connaissances des apprenants concernant l'utilisation des éléments linguistiques étudiés[144] et peut être complétée avec ou sans le corpus. L'exercice a été proposé par Johns (1991) pour illustrer comment les corpus peuvent offrir un regard neuf sur des tâches traditionnelles et accroître leur efficacité (tableau 188).

Melyik szó hiányzik minden mondatból? Használja a « MagyarOK A1 » korpuszt!

1. A székem egy nagyon _____ fitneszlabda.
2. A szobám nem túl _____: 10–12 négyzetméter.
3. A lányok mániákus ember, a szobájában is _____ a rend.
4. Kaposvár nem _____ város: hatvanezer ember él itt.
5. Gyalog járok az irodába. Ez Budapesten _____ luxus!

(Quel mot manque dans toutes les phrases ? Consultez le corpus « MagyarOK A1 ».

1. Mon fauteuil est un ballon de _____fitness.

2. Ma chambre n'est pas trop grande _____: 10-12 mètres carrés.

3. Ma fille est une personne ordonnée : sa chambre est toujours très bien rangée. (litt. : dans sa chambre il y a toujours_____ de l'ordre).

4. Kaposvár n'est pas une _____ville : 60 000 personnes y vivent.

5. Je vais à pied à mon bureau. C'est un véritable luxe à Budapest. (litt. : qui est un _____ luxe))

Tableau 188 : Exploration de l'utilisation d'un adjectif dans divers environnements textuels.

[144] En comparant cette liste avec les 26 adjectifs les plus fréquents dans le corpus « huTenTen12 », nous trouvons des similitudes importantes. Ces adjectifs dans le grand corpus sont les suivants : « grand, tel, bon, nouveau, hongrois, petit, propre, autre, important, complet, agréable, vrai, long, entier, donné, européen, haut/grand, tardif, dernier, difficile/lourd, suivant, ancien, approprié, sûr, similaire ». Les deux listes partagent 11 adjectifs, et les 15 autres du corpus à fins pédagogiques figurent également parmi les 100 adjectifs les plus fréquents du grand corpus. Si l'on ajoute les corpus A2-B1 au A1, 67 des 100 premiers adjectifs se correspondent. Cela indique une bonne fiabilité du corpus de manuels pour les adjectifs récurrents.

Une fois de plus, l'exercice de préparation (création de la fiche) ainsi que la tâche effectuée lors du cours de langue permettent aux apprenants de prendre conscience du lien étroit qui existe entre l'environnement textuel et l'élément lexical étudié. Ils peuvent également noter que certaines des phrases ci-dessus ne semblent pas naturelles lorsqu'elles sont traduites en français ; cela devrait leur rappeler que chaque langue a sa propre idiomaticité et que les énoncés peuvent rarement être construits en les traduisant mot à mot d'une langue à une autre.

5) Lexique autour des mots-clés

Les textes portant sur un sujet similaire sont utiles aux apprenants afin d'observer les répétitions et les variations linguistiques. Pour illustrer ce point, nous utiliserons ici comme exemple les textes sur « L'endroit où j'habite » dans le sous-ensemble 2 du corpus écrit (récits rédigés par des professeurs de langues avec le vocabulaire-clé du chapitre) pour les niveaux A1.

Tout d'abord, les apprenants sont enjoints à lire attentivement les textes-modèles. Lors de cette étape, ils rencontrent des éléments-clés à plusieurs reprises, dans des environnements textuels légèrement différents[145]. Dans l'étape suivante, les apprenants identifient les éléments lexicaux les plus courants dans les textes, à l'aide de l'outil « Wordlist ». Dans le corpus des dix textes reliés au chapitre 4, le verbe « lakom » (j'habite), les noms « ház » (maison), « város » (ville), « falu » (village) sont des exemples de mots-clés pertinents. Les apprenants peuvent créer le profil de ces mots comme indiqué au chapitre 4. Le profil de mot de « lakom » basé sur les textes-modèles se présente de la façon suivante (tableau 189) :

[145] Voir les exemples au chapitre 13, section 3.2.

Collocations	Compléments indiquant un lieu : *Egy kisvárosban / faluban / nagy házban / Monoron / Hollandiában / a városközpontban lakom. (J'habite dans une petite ville / dans un village / dans une grande maison / à Monor / aux Pays-Bas / au centre-ville.)*
Colligations: suffixes	(1) N-*ban/-ben* + *lakom*; (2) N-*on/-en/-ön* + *lakom*.
Colligations: l'ordre des mots	***CompLieu + lakom*** Le verbe suit le complément de lieu. En cas de plusieurs compléments de lieu, le verbe est placé entre les compléments.
Composantes sémantiques	*Lakom* s'utilise avec des noms propres de pays, ville et village ainsi qu'avec des noms indiquant une partie d'une ville ou village ou d'une maison.
Composantes pragmatiques	Pas de composante pragmatique spéciale.

Table 189 : Profil du verbe « lakom ».

6) Observer la grammaire : même suffixe, différente signification

Le point de départ de cette activité est une question grammaticale : l'utilisation du suffixe « -ban/-ben »[146]. Dans les collocations avec « lakom », la plupart des adverbes de lieu se terminent par « -ban/-ben », suffixe dont la signification est comparable aux prépositions françaises « à » et « dans ». Pour examiner si ce suffixe a d'autres fonctions et si c'est sa fonction la plus répandue, les apprenants peuvent utiliser l'option « Recherche avancée » dans le Concordancier, choisir « Word » comme type de requête et écrire « .*ban|.*ben » sur la ligne.

Dans les textes-modèles du chapitre 4, les mots suivants ont été trouvés avec ce suffixe : « a hegyekben » (dans les montagnes), « a városban » (dans la ville), « a városközpontban » (au centre-ville), « iskolában » (à l'école), « egy faluban » (dans un village), « a lakásomban » (dans mon appartement), « a házban » (dans la maison), etc. D'autres occurrences sont : « általában » (en général), « májusban » (en mai), « gyerekkoromban » (dans mon enfance), « napközben » (dans la journée), « mostanában » (ces jours-ci) ».

Le mot le plus fréquent avec le suffixe « -ban/-ben » est « általában » (en général). L'analyse montre que « -ban/-ben » est fréquemment utilisé avec d'autres compléments de temps. En élargissant leur recherche aux textes d'autres chapitres, les apprenants peuvent constater qu'en effet, de

[146] La voyelle du suffixe dépend des voyelles du radical, d'où les deux variantes.

nombreux compléments de temps se terminent par « -ban/-ben ». Ces résultats sont intéressants car « -ban/-ben » est présenté dans les grammaires hongroises principalement comme un suffixe typique des compléments de lieu, et ses autres fonctions sont considérées comme secondaires (voir par exemple en ce sens Budai 2016 ; Keresztes 1995).

D'autres aspects grammaticaux (l'utilisation d'autres suffixes, le temps des verbes, l'ordre des mots) peuvent également faire l'objet d'une exploration à partir du corpus. L'intérêt d'une analyse à l'aide d'outils de corpus est là encore clair : ce serait une tâche fastidieuse que de parcourir manuellement un grand nombre de textes pour identifier les schémas d'utilisation de l'aspect grammatical choisi.

D) De l'observation à la production langagière

Les activités dans cette section représentent des prolongements de l'Apprentissage sur corpus en ce qu'*elles utilisent les textes du corpus non seulement comme sources d'informations linguistiques mais aussi comme modèles pour les textes des apprenants.* Ils peuvent y identifier les unités multi-lexicales fréquentes ainsi que les répétitions et les variations langagières, étudier leur usage et faire une sélection de ces éléments afin de les réutiliser dans leurs propres textes. Ce procédé sert à améliorer la qualité des produits linguistiques de l'apprenant, comme le suggèrent Kennedy et Miceli (2017) et Szita et Pelcz (2022 à paraître).

1) Réviser le vocabulaire (1)

Outil : *Wordlist*

Les deux activités suivantes explorent le potentiel du corpus pédagogique pour améliorer l'acquisition et la consolidation du vocabulaire, deux éléments centraux de l'apprentissage des langues (Allan 2009). La première activité utilise l'outil « Wordlist » pour réviser le vocabulaire dans le(s) sous-ensemble(s) sélectionné(s) ou dans tout le corpus. Par exemple, les apprenants peuvent générer la liste ci-dessous (tableau 190) pour réviser l'utilisation des adjectifs les plus fréquents dans le manuel A1.

1. magyar (hongrois) : 134
2. amilyen (comme) (erreur d'annotation) : 114
3. jó (bon) : 85
4. kicsi (petit 1) : 82
5. nagy (grand) : 71
14. érdekes (intéressant) : 33
15. nehéz (lourd/difficile) : 39
16. finom (délicieux) : 28
17. gyönyörű (magnifique) : 26
18. kényelmes (confortable) : 25

6. szép (beau) : 60	19. német (allemand) : 25
7. új (nouveau) : 70	20. friss (frais) : 24
8. kis (petit 2) : 46	21. könnyű (facile/léger) : 23
9. éves (ans) : 41	22. magas (haut/grand) : 22
10. kedvenc (favori) : 39	23. türelmes (patient) : 22
11. kedves (gentil, sympathique) : 39	24. angol (anglais) : 22
12. olyan (tel) : 39	25. fáradt (fatigué) : 20
13. régi (vieux) : 33	26. drága (cher) : 19

Tableau 190 : Les 26 adjectifs les plus fréquents dans le corpus « MagyarOK A1 ».

Même dans notre corpus de taille limitée, ces adjectifs apparaissent avec une fréquence notable, ce qui offre de nombreuses occasions d'observer leur comportement linguistique. Avant de réviser le vocabulaire en cours, les apprenants peuvent ainsi revoir l'utilisation des adjectifs en étudiant dix à quinze occurrences avec chacun d'eux. En cliquant sur les trois points après un mot, les apprenants peuvent sélectionner le Concordancier pour étudier des exemples ou l'outil « Word Sketch » pour explorer les collocations fréquentes. Ce procédé est en accord avec Nation (2013 : 119, 127) qui souligne que l'apprentissage de chaque mot est un « processus cumulatif ». L'apprenant construit ses connaissances au fur et à mesure en étant exposé à des instances d'utilisation ou à un enseignement précis exposant certaines caractéristiques du mot ou en effectuant des analyses linguistiques. Dans notre contexte, cela signifie que l'étudiant voit d'abord ces adjectifs dans différents textes, avec des rencontres espacées dans le temps. L'outil « Wordlist » lui propose une vue d'ensemble qui se prête à une exploration avec « Word Sketch » afin d'identifier leurs collocations ou avec le « Concordancier » pour observer et analyser les différents environnements textuels dans lesquels le mot s'est produit.

2) Pratiquer les mots et les unités multi-lexicales

Cette activité offre aux apprenants la possibilité d'approfondir les caractéristiques typiques du discours oral en les intégrant dans leurs propres énoncés.

Certains mots récurrents ont des fonctions différentes à l'oral et à l'écrit. Le mot hongrois « jó » (bon) en est un bon exemple : il est couramment utilisé dans notre corpus oral et écrit mais son rôle typique n'est pas le même dans les deux modalités. Sa principale fonction à l'écrit est de qualifier un nom (« un bon film », « un bon livre », « une bonne professeure ») alors que dans les interactions orales, « jó » est essentiellement un marqueur de discours à multiples significations et

peut être traduit par plusieurs mots selon le contexte comme « bon », « bien », « ça marche », « super ». Dans nos collections, les exemples avec « jó » incluent des suggestions, des réponses positives à des invitations, des compliments, ainsi que des exemples pour indiquer la compréhension, demander une opinion, accepter une suggestion et souligner des avantages.

Les apprenants peuvent observer et identifier ces fonctions dans des sous-corpus sélectionnés ou dans l'ensemble du corpus, en utilisant le Concordancier, comme l'illustre l'activité suivante. Si elle est menée avec des apprenants du niveau A1 qui n'ont que peu d'expérience de la langue-cible et des outils, l'enseignant peut fournir les catégories et demander aux apprenants d'y attribuer les occurrences.

Le tableau 191 présente des exemples d'occurrences de 'Réactions positives' extraites de tous les dialogues dans le corpus. Seul l'environnement textuel immédiat nécessaire à la compréhension de la fonction de « jó » a été inclus.

Majd szólok, ha legközelebb koncertezünk. – **Jó**, mindenképp!	*Je te ferai savoir quand nous aurons un concert la prochaine fois. -* ***Super****. Fais-le, s'il te plaît.*
Meglátogatlak, jó? – Jaj, **de jó**!	*Je viendrai te voir, d'accord ? -* ***Super****.*
De jó, hogy itt vagy, Adél! Már nagyon vártalak. – Miért? Valami baj van?	***Je suis si heureuse que*** *tu sois venu, Adél ! J'avais tellement envie de te voir. - Pourquoi ? Il y a un problème ?*
Milyen volt a nyarad? – Hát, voltunk tíz napot Horvátországban. – Fú, az **jó**.	*Comment s'est passé ton été ? - Eh bien, nous avons été 10 jours en Croatie. - Wow, c'est* ***génial****.*

Tableau 191 : « Jó » dans des phrases exprimant un retour positif et de l'enthousiasme.

Après avoir observé les fonctions de « jó » dans les courts dialogues, les apprenants peuvent être invités à changer un ou deux mots dans ces dialogues. Par exemple, le locuteur 1 peut dire dans le premier dialogue « Je te ferai savoir quand nous aurons une réunion » au lieu de dire « Je te ferai savoir quand nous aurons un concert la prochaine fois ». Les apprenants peuvent produire une ou plusieurs variantes du même dialogue. De cette façon, les apprenants créent de nouvelles situations dans lesquelles cette réaction est appropriée tout en s'appuyant sur les exemples. Dans un deuxième temps, ils peuvent conserver les réactions et changer complètement la phrase initiale.

3) Même sujet, différents textes

Cette tâche sert à observer et à consolider le vocabulaire lié à un type de texte et à un sujet donnés. À titre d'exemple, nous utiliserons un petit ensemble de données authentiques créé pour le chapitre 8 (habitudes alimentaires) de « MagyarOK » A2. Le corpus contient 100 critiques de restaurants (environ 3 000 tokens) ; quelques textes ont été légèrement modifiés ou raccourcis. Dix exemples de textes sont publiés dans leur intégralité sur le site Web de « MagyarOK », d'autres textes peuvent être lus sur le site Web de « TripAdvisor ». L'enseignant peut préparer la série de questions suivantes pour que les apprenants puissent explorer le corpus :

- Compréhension générale : Lisez dix critiques et décidez si elles sont positives ou négatives. Marquez les expressions qui vous ont aidé à trouver la réponse.
- Vocabulaire de base (1) : Qu'est-ce qui a été examiné (nourriture, personnel ou autre) ? Utilisez la fonction Liste de mots pour le découvrir.
- Structure : Les examens ont-ils une structure typique ?
- Vocabulaire de base (2) : Cherchez des adjectifs qui qualifient les mots-clés suivants : « adag » (portion), « kiszolgálás » (service), « ár » (prix), « hangulat » (ambiance), « étel » (nourriture, plat). Utilisez le Concordancier.
- Vocabulaire de base (3) : Trouvez les cinq adjectifs les plus fréquents (utilisez Wordlist). Rassemblez les noms qui sont leurs collocatifs typiques. Utilisez le Concordancier ou l'outil Word Sketch.
- Vocabulaire de base (4) : Quelles phrases pouvez-vous utiliser pour recommander un restaurant ? Cherchez les occurrences avec « ajánl » (recommander) dans le Concordancier.

Les apprenants peuvent rassembler les noms appartenant au même adjectif ou les adjectifs appartenant au même nom et observer les similitudes (répétitions) et les variations dans l'usage langagier dans les différentes critiques. Les lignes de concordance avec le verbe « ajánl » (recommander) constituent un bon exemple de répétitions et de variations lexicales (tableau 192) :

Tableau 192 : Phrases avec « ajánl » (recommander) dans les critiques de restaurants.

En étudiant ces occurrences, les apprenants peuvent reconnaître qu'il n'y a qu'un nombre limité de variations que les locuteurs natifs utilisent pour recommander un restaurant. Celles-ci sont présentées dans le tableau 193.

Mindenkinek ajánlom, aki …	*Je le recommande à tout le monde (qui ...)*
Mindenkinek **szívből** ajánlom(, aki …)	*Je le recommande de tout cœur à tout le monde (qui ...)*
Mindenkinek **jó szívvel** ajánlom(, aki …)	*Je le recommande de tout cœur à tous ceux (qui ...) (litt. :*
/ **Szívesen** ajánlom mindenkinek(, aki …)	*de bon cœur, cordialement)*
Csak ajánlani **tudom** (mindenkinek, aki …). / Mindenkinek **csak** ajánlani **tudom**.	*Je ne peux que le recommander à tout le monde (qui ...)*
Bátran ajánlom mindenkinek.	*Je le recommande sans hésitation. (litt. : courageusement)*

Tableau 193 : Répétitions et variations dans les critiques de restaurants.

Enfin, les apprenants peuvent utiliser les critiques de restaurants comme modèles pour leurs propres textes. Par exemple, ils peuvent présenter leur restaurant préféré en incluant un certain nombre (ou un maximum) d'unités multi-lexicales typiques dans le corpus[147].

[147] Dans sa publication sur les activités d'Apprentissage basées sur les données, Tyne (2012) suggère d'autres activités qui utilisent des corpus de textes courts dans la langue maternelle des apprenants et dans la langue-cible.

4) Reconstruire un texte

Outil : *Wordlist ou Wordle*

Cette activité permet de réviser le vocabulaire ainsi que les éléments de la cohésion textuelle. L'enseignant choisit un texte traité pendant le cours d'une longueur de 50 à 100 mots et en retire une phrase ou une proposition sur deux. Au cours de la leçon, les apprenants lisent et écoutent d'abord le texte pour se rafraîchir la mémoire. Ensuite, sur la base d'une liste de tous les mots apparaissant dans le texte, générée par l'outil « Wordlist », ils essaient de reconstituer les parties manquantes. Leurs textes ne doivent pas nécessairement être des copies exactes de l'original, mais ils doivent être cohérents. Les éléments contribuant à la cohésion textuelle peuvent être marqués sur la liste pour attirer sur eux l'attention des apprenants. En cas de doute, les apprenants peuvent observer l'utilisation de ces éléments dans le Concordancier avant de les intégrer dans leur texte. Enfin, les apprenants comparent leur version avec le texte original (v. les tableaux 194-196).

1	a	22	14	idejét	1
2	és	7	15	ezek	1
3	az	5	16	ez	1
4	majd	4	17	jobban	1
5	nem	3	18	fiatal	1
6	munkaidő	2	19	felerősödik	1
7	nyolctól	2	20	döntheti	1
8	között	2	21	kommunikációs	1
9	egyre	2	22	fog	1
10	négyig	2	23	fiatalabb	1
11	is	2	24	fogják	1
12	dolgozni	2	25	fognak	1
13	valószínűleg	2	26	el	1

27	bébiszitter	1	40	könnyen	1
28	be	1	41	teljesen	1
29	múlva	1	42	lehet	1
30	idő	1	43	lesz	1
31	ilyenkor	1	44	ma	1
32	irodákban	1	45	utána	1
33	otthonról	1	46	magánélet	1
34	reggelig	1	47	maga	1
35	jövő	1	48	választják	1
36	jövőben	1	49	megbeszéléseken	1
37	szabadon	1	50	gondoltunk	1
38	kommunikálni	1			
39	kontinensek	1			

Tableau 194 : Les mots du texte « Le lieu de travail du futur » (extrait).

A jövő munkahelye

Régen a munkaidő nyolctól négyig tartott. Utána hazamentünk, és másnap reggelig nem gondoltunk a munkára. A munka és a magánélet között húsz-harminc év múlva valószínűleg teljesen eltűnik majd a határ. Az emberek nem irodákban fognak dolgozni, és a munkaidő nem nyolctól négyig fog tartani. Mindenki szabadon oszthatja majd be az idejét, és maga döntheti el, mikor és hol akar dolgozni. Valószínűleg egyre többen fogják otthonról végezni a munkájukat, és a cégnél csak megbeszéléseken találkoznak majd. Ilyenkor a munkahelyi bébiszitter vigyáz a gyerekekre. Egyre több lesz a nemzetközi projekt is, hiszen az új kommunikációs eszközökkel a kontinensek között is könnyen lehet kommunikálni. A fiatalabb generációk számára ezek a tendenciák természetesek. Ők már ma sem választják szét a tanulást, a munkát és a szabadidőt. Sok fiatal az idő nagy részét a számítógép, az okostelefon vagy a táblagép előtt tölti. A jövőben ez a tendencia még jobban felerősödik majd.

*(**Le lieu de travail du futur***

Dans le passé, les heures de travail duraient de huit à quatre heures. Ensuite, nous sommes rentrés chez nous et n'avons pensé au travail que le lendemain matin. Dans vingt à trente ans, la frontière entre le travail et la vie privée est susceptible de disparaître complètement. Les gens ne travailleront pas dans les bureaux et les heures de travail ne dureront pas de huit à quatre heures. Chacun sera libre de

partager son temps et de décider lui-même quand et où il veut travailler. Il est probable que de plus en plus de personnes effectueront leur travail à domicile et ne se rencontreront que lors de discussions au sein de l'entreprise. Dans ce cas, la baby-sitter au travail s'occupe des enfants. Il y aura également de plus en plus de projets internationaux, car de nouveaux outils de communication facilitent la communication entre les continents. Pour les jeunes générations, ces tendances sont naturelles. Ils ne séparent plus les études, le travail et les loisirs. De nombreux jeunes passent la plupart de leur temps devant un ordinateur, un smartphone ou une tablette. Cette tendance s'intensifiera à l'avenir.)

Tableau 195 : Texte-modèle sur « Le lieu de travail du futur » au chapitre 6 du manuel A2.

Une variation de l'activité est de donner les débuts de phrases et laisser l'apprenant remplir les trous à l'aide de la liste de mots. Le tableau 196 illustre cette possibilité.

A jövő munkahelye

Régen a munkaidő nyolctól .... Utána hazamentünk, és .... A munka és a magánélet között .... Az emberek nem irodákban fognak dolgozni, és a munkaidő .... Mindenki szabadon oszthatja majd be az idejét, és maga .... Valószínűleg egyre többen fogják otthonról ..., és a cégnél csak .... Ilyenkor a munkahelyi bébiszitter .... Egyre több lesz a nemzetközi projekt is, hiszen .... A fiatalabb generációk számára ezek a tendenciák természetesek. Ők már ma sem .... Sok fiatal az idő nagy részét ..... A jövőben ez a tendencia ....

Le lieu de travail du futur

Dans le passé, les heures de travail ... de huit .... Ensuite, nous sommes rentrés chez nous et .... Dans vingt à trente ans, la frontière .... Les gens ne travailleront pas dans les bureaux et les heures .... Chacun sera libre de partager son temps et de .... Il est probable que de plus en plus de personnes .... Dans ce cas, la baby-sitter au travail .... Il y aura également de plus en plus de projets internationaux, car .... Pour les jeunes générations, ces tendances .... Ils ne .... De nombreux jeunes passent la plupart de leur temps .... Cette tendance ....

Tableau 196 : Exemple de texte à trous (à combiner avec la liste de mots présentée dans le tableau 194).

Si l'enseignant attend une reproduction exacte du texte original, il peut indiquer le nombre de mots manquants entre parenthèses. L'effort conscient de reconstruction du texte aide l'apprenant à mémoriser le vocabulaire-clé thématique ainsi que les éléments contribuant à la cohésion textuelle (Szita 2022b à paraître).

5) Écrire son propre texte

La production d'un récit est un moyen efficace de systématiser ses connaissances. Par conséquent, certaines des activités présentées ci-dessus se terminent par la recommandation faite aux apprenants d'écrire leur propre texte. En suivant les étapes conseillées, les apprenants peuvent améliorer la qualité de leurs textes grâce à une utilisation plus consciente des éléments fréquents, souvent idiomatiques, de la langue. Tout d'abord, les apprenants observent les répétitions lexicales (mots et phrases récurrents), les répétitions structurelles (textes suivant la même logique) et les répétitions grammaticales (la même terminaison ou le même ordre des mots utilisés dans divers énoncés). Ils sont ensuite invités à « recycler » autant de mots et de phrases des textes-modèles qu'ils le souhaitent dans leur récit (Kennedy et Micelli 2017 ; Szita et Pelcz 2022 à paraître). Le défi n'est donc pas de « réinventer la langue », c'est-à-dire d'essayer de dire les choses différemment des textes-modèles, mais de créer des *textes cohérents, lexicalement variés, naturels et précis à l'aide d'un apport linguistique authentique*. Les apprenants peuvent également être encouragés à publier leurs textes sur les médias sociaux et, ce faisant, à participer à des interactions réelles significatives. Les activités impliquant des relations positives avec des locuteurs de la langue-cible semblent en effet conduire à une meilleure mémorisation du vocabulaire en raison d'une motivation accrue (Dörnyei et Csizér 1998).

Les tableaux suivants montrent deux exemples de textes écrits sur le sujet « L'endroit où j'habite » reposant sur des textes-modèles. Les unités multi-lexicales que les apprenants ont intégrées des textes-modèles sont imprimées en gras (tableau 197).

> **Strasbourgban élek, de nem Strasbourgi vagyok**. Strasbourg **egy éleg nagy város**, körülbelül 280000 ember él itt. Szeretek itt élni, az egyetemen tanulok kínai nyelvet. **Szerintem jó a közbiztonság és magas az életszínvonal.** Rengeteg a közlekedést vannak, nagyon praktikus: villamosok, buszok, vonatok vannak. Tiszta és kenyelmes. **A közlekedés egészen megbízható és gyors**. **Természetesen, mint mindenhol, itt is vannak dugók**, ézer biciklivel gyakran járok. **Strasbourgban minden itt van**: szép parkok, érdékes múzeumok, tipikus étteremek, híres fő tér és **a központ gyönyörű**. A területem nagyon nyugodt. **A Orangerie park mellet lakom, igy** gyakran sétálok a parkban. **Strasbourgban nagyon sok park és zöld terület van**, ézert szeretek a várost. Az utcámban két buszmegálló is van, **szerintem nagyon praktikus**. Strasbourgban sok megnéznivaló van. A városban például **különleges kiállitások, kiváló színházi előadások és jó sportolási lehetőségek**. **Korábban azt hittem,**

hogy Strasbourg csúnya és unalmas. Most, ugy tudom, hogy Strasbourg nagyon szimpatikus és hangulatos város.

(J'habite à Strasbourg, mais je ne suis pas de Strasbourg. Strasbourg est une grande ville animée, avec environ 280 000 personnes qui y vivent. J'aime vivre ici, j'étudie le chinois à l'université. Je pense que la sécurité publique est bonne et que le niveau de vie est élevé. Il y a beaucoup de transports, très pratiques : il y a des tramways, des bus, des trains. Ils sont propres et confortables. Le trafic est assez fiable et rapide. Bien sûr, comme partout, il y a des embouteillages ici, je fais souvent du vélo. À Strasbourg, on a tout : de beaux parcs, des musées intéressants, des restaurants typiques, la célèbre place principale et le centre sont magnifiques. Mon quartier est très calme. J'habite à côté du parc de l'Orangerie, donc je marche souvent dans le parc. Il y a beaucoup de parcs et d'espaces verts à Strasbourg, j'adore la ville. Il y a aussi deux arrêts de bus dans ma rue, je pense que c'est très pratique. Il y a beaucoup à voir à Strasbourg. Dans la ville, par exemple, il y a des expositions spéciales, d'excellentes représentations théâtrales et de bonnes installations sportives. Avant cela, je trouvais Strasbourg moche et ennuyeux. Maintenant, je sais que Strasbourg est une ville très sympathique et chaleureuse.)

A városom fantasztikus. Franciaországban, **Strasbourgban élek**. Az egyetemen kínaiul és magyarul tanulok. Strasbourgban **körülbelül négyszáz ezer ember él**. **Természetesen** nem a nagyváros.

Nagyon szép a város. Jó a közbiztonság, magas az életszínvonal és vannak kulturális és sportolási lehetőségek is.

Központban vannak **szép és régi épületek, modern könyvtárok, híres múzeumok, nagy mozik és színházak** is. **A közlekedés jó, praktikus de egy kicsit drága**. A városban nem járok autóval. **Általában villamossal és gyalog közlekedem.**

A kollégiumban lakom, a kollégium elég modern és világos. **Egy csendes utcában lakom**, de sok diák lakik is itt. Szerintem Strasbourg nagyon érdekes és **hangulatos, mert a városban sok olcsó és népszerű bár, kocsma és étterem van**. Nagyon tetszik Strasbourgi kávézók. Itt lehet könyvet olvasni, kávét inni, pogácsát enni és e-mailt írni. A városban **kitűnő cukrászdák** vannak is. Strasbourg **nagyon biztonságos város, az emberek kedvesek, barátságosak és nyugodtak**.

Strasbourgban **nincsenek magas hegyek**, tengerek, tók vagy nagy erdők. Tél elég hideg is, de mégis **nagyon szeretem itt élni**.

(Ma ville est fantastique. J'habite à Strasbourg, en France. J'étudie le chinois et le hongrois à l'université, il y a environ quatre cent mille personnes vivant à Strasbourg. Ce n'est certainement pas une grande ville.

La ville est très belle, la sécurité publique est bonne, le niveau de vie est élevé et il existe des opportunités culturelles et sportives.
Le centre possède également de beaux et anciens bâtiments, des bibliothèques modernes, des musées célèbres, de grands cinémas et des théâtres. Le transport est bon, pratique mais un peu cher. Je ne conduis pas en ville. Je prends généralement les tramways et à pied.
Je vis dans un dortoir, le dortoir est assez moderne et lumineux. Je vis dans une rue calme, mais de nombreux étudiants vivent également ici. Je pense que Strasbourg est très intéressante et confortable car il y a de nombreux bars, pubs et restaurants bon marché et populaires dans la ville. J'aime beaucoup les cafés strasbourgeois. Ici, vous pouvez lire un livre, boire du café, manger un gâteau et écrire un e-mail. Il y a aussi d'excellentes pâtisseries dans la ville. Strasbourg est une ville très sûre, les gens sont gentils, sympathiques et détendus.
Strasbourg n'a pas de hautes montagnes, de mers, de lacs ou de grandes forêts. L'hiver est assez froid aussi, mais j'aime toujours beaucoup vivre ici.)

Tableau 197 : Deux exemples de textes écrits sur le sujet « L'endroit où j'habite » reposant sur des textes-modèles (textes hongrois non corrigés).

Au vu de la quantité d'unités multi-lexicales reprises à partir des textes-modèles, on peut s'interroger sur l'effet d'apprentissage de ce qui n'est apparemment qu'un simple copier-coller. Pour rendre cet exercice efficace, il convient, en effet, d'expliquer en cours le processus et ses bénéfices. La lecture des modèles et la rédaction du texte de l'apprenant sont donc précédées d'un enseignement explicite de la technique qui consiste en plusieurs étapes :

(1) Lire tous les textes-modèles
(2) Identifier des éléments récurrents dans plusieurs textes et les énoncés qui s'appliquent à la vie de l'apprenant.
(3) Rédiger son propre texte en réutilisant des unités multi-lexicales.
(4) Correction des textes par l'enseignant
(5) Recopier le texte corrigé à la main (même si la première version a été rédigée à l'ordinateur)
(6) Pratique active avec le texte au cours suivant.

Lors des deux premières étapes, l'apprenant observe les modes d'expression des natifs et identifie les expressions qu'il pourra utiliser dans son récit. On ne l'encourage pas à rédiger un texte avec ses « propres mots » car, aux niveaux débutants, l'apprenant n'a pas encore « ses propres mots ». Sans modèle, il produira très probablement un texte assez pauvre linguistiquement (par peur de

faire des fautes ou par manque d'inspiration) ou, en essayant d'être créatif, il prendra des mots isolés et en composera des phrases qui ne seront probablement pas typiques, ni parfaitement compréhensibles dans la langue-cible. Rappelons ici que la majorité de ces unités multi-lexicales ont été présentées dans le manuel – les textes-modèles ne font donc que reprendre ces éléments, les réorganisent pour les présenter dans des récits cohérents et les complètent avec quelques nouveaux éléments si besoin. L'étude des textes et l'intégration consciente de certains éléments dans un texte personnel permettent la révision et la pratique active du vocabulaire-clé. À l'aide des modèles proposés, l'apprenant sera capable de rédiger un texte sur ses expériences qui ait un caractère naturel. Le fait que le texte soit corrigé et commenté par l'enseignant, puis recopié par l'apprenant, renforce la consolidation des unités multi-lexicales. Le travail avec le texte lors du cours suivant contribuera également à la consolidation du vocabulaire et offrira la possibilité d'une révision espacée. Ce travail peut consister en la transformation du texte écrit dans une présentation orale ou dans une conversation sur le sujet. Ainsi, la pratique de la langue qui reposait seulement jusque-là sur la lecture et sur l'écriture, s'enrichit par l'expression et la compréhension orales (voir aussi 3.4.3 et 3.4.4). Ainsi, cette séquence d'exercices forme à toutes les compétences, assure la possibilité d'un entrainement espacé dans le temps et facilite l'intégration des textes corrigés dans des activités appropriées pendant le cours.

6) Transformer ses propres récits écrits en interactions orales

Dans cette activité, les apprenants sont invités à transformer leurs propres récits en interactions à consonance naturelle. Pour cela, ils peuvent ajouter à leurs textes des marqueurs de discours, du langage approximatif, bref, toutes les caractéristiques des interactions naturelles – et supprimer les parties qui, selon eux, ne conviendraient pas à une interaction orale.

Une interaction fondée sur le premier texte présenté dans la section 3.2 du chapitre 13 pourrait ressembler à ceci (tableau 198) Le langage interactionnel indiqué écrit en caractère gras :

(…)	*(…)*
– Budapesten laksz?	– *Tu habites à Budapest ?*
– Nem, Hosszúhetényben.	– *Non, à Hosszúhetény.*
– **Hol?**	– ***Où ça ?***
– Hosszúhetényben.	– *À Hosszúhetény.*
– **Igen?** És ez egy falu? Vagy egy kis város?	– ***Oh.*** *C'est un village ? Ou une petite ville ?*
– Egy falu Dél-Magyarországon.	– *Un village du sud de la Hongrie.*
– **Sajnos** nem ismerem.	– ***Malheureusement****, je ne le connais pas.*

– **Pedig** szép hely. Gyönyörű a környék, **tényleg szép helyeken** lehet kirándulni. **Eljöhetnél** májusban a Talicskaolimpiára.	– ***Pourtant****, c'est un endroit agréable. La région est magnifique et on peut faire des randonnées* ***dans des endroits vraiment jolis****.* ***Tu pourrais venir*** *aux Olympiades de la Brouette en mai.*
– **A Talicskaolimpiára?**	– ***Aux Jeux Olympiques de la Brouette ?***
– **Aha.**	– ***Oui.***
– **Az micsoda?**	– ***C'est quoi ça ?***
– **Egy ilyen verseny.** Több országból is jönnek rá az emberek.	– ***Un genre de compétition.*** *Des gens de plusieurs pays y participent.*
– **Tényleg?**	– ***Vraiment ?***
– **Aha.**	– ***Oui.***
– **Hát, érdekes. Lehet, hogy** egyszer megnézem. (...)	– ***Hm, intéressant. Peut-être que*** *j'irai le voir un jour. (...)*

Tableau 198 : Interaction possible à la base d'un récit écrit.

Grâce à ce processus, (1) les *apprenants peuvent réutiliser le vocabulaire contenu dans leurs textes et le consolider à l'oral, et* (2) *ils peuvent s'entraîner à intégrer les informations d'un texte narratif dans le flux d'une conversation*.

Ce chapitre a présenté diverses activités qui suivent et enrichissent le concept de l'« Apprentissage sur corpus ». Ces activités mettent en pratique toutes les compétences : l'étudiant lit, écrit, écoute, parle – les quatre compétences-clés – mais il acquiert également des connaissances métalinguistiques et métacognitives ainsi que des compétences interculturelles, toutes importantes pour un apprentissage réussi. Ce faisant, *l'apprenant prend conscience de l'interdépendance de plusieurs aspects langagiers* : il réalise, entre autres, la forte présence des unités multi-lexicales dans tous les types de textes. Les observations révèlent également que la plupart des unités multi-lexicales existent en plusieurs variations mais que le nombre de ces variations est souvent limité. Les apprenants cherchent dans le corpus des réponses à des questions, lisent et analysent les textes-modèles dans leur intégralité ou avec les outils d'analyse et intègrent les résultats de leurs observations dans leurs propres textes. Cette démarche améliore la qualité linguistique de leurs productions en rapprochant l'usage langagier de l'apprenant de celui des natifs.

Ainsi, toutes les activités présentées contribuent à la consolidation des connaissances linguistiques, au développement des compétences linguistiques et, à travers l'ensemble de ces dimensions participant à la maîtrise de la langue-cible.

Résumé de la Partie III

Pour pouvoir comprendre et produire un langage à caractère naturel, les phases d'observation des énoncés authentiques et semi-authentiques semblent nécessaires ; là se situe l'un des intérêts des corpus pédagogiques. La partie III de cette thèse a présenté un corpus pédagogique pour l'enseignement du hongrois. L'approche sous-jacente à la création de tels corpus est que, contrairement aux corpus à fins linguistiques, ils sont créés en tenant compte des besoins de l'apprenant. Dans ce but, ils remplissent trois critères essentiels : (1) ils contiennent des données linguistiques à caractère naturel ; (2) ces données concernent des sujets pertinents pour l'apprenant et (3) ils proposent enfin un langage accessible aux apprenants.

Le corpus pédagogique présenté dans cette partie est fondé sur la série de manuels « MagyarOK » pour l'apprentissage du hongrois. Il consiste en différentes parties consacrées aux usages langagiers écrit et oral. Au sein de ces deux grandes catégories, la sélection des textes appropriés s'articule autour des interactions du quotidien et des sujets pour les niveaux A1-B1, définis par le CECRL. À travers trois étapes, ce corpus amène progressivement l'apprenant de l'étude des textes adaptés et semi-authentiques vers les textes authentiques. La volonté d'adaptation des textes authentiques dans la première étape repose sur un compromis dans l'intérêt de l'accessibilité ; l'apprenant doit être capable de comprendre et de recontextualiser les énoncés du corpus, tâche qui dépasserait largement ses compétences linguistiques si nous lui proposions, *ab initio*, des textes complétement authentiques. Ainsi, en accord avec les suggestions évoquées dans la littérature, la première sous-partie de notre corpus pédagogique comprend un ensemble de textes adaptés à partir d'un corpus de textes authentiques : il s'agit des textes de manuels et de textes supplémentaires tels que des transcriptions de vidéos étroitement liés au contenu des manuels. Dans le processus d'apprentissage, l'élève se familiarise d'abord avec les textes entiers et ne les explore avec les outils numériques que dans un deuxième temps. La deuxième sous-partie du corpus contient des textes semi-authentiques tels que des improvisations d'acteurs, des entretiens et des textes thématiques à caractère naturel. Cette partie ouvre la voie vers le langage authentique que l'apprenant pourra explorer dans la troisième sous-partie du corpus.

Les corpus pédagogiques peuvent également être exploités par l'apprenant de façon autonome, sans l'aide de l'enseignant, en dehors du cadre des cours de langues. Leur utilisation permet de fournir des réponses à certaines questions posées par les apprenants et d'améliorer la qualité linguistique de leurs textes écrits et oraux. Organisés par niveau et par thème, l'exploration de ces corpus fait émerger des répétitions et des variations. Cette observation est permise par l'accès à la grande quantité de textes liés aux mêmes thèmes et/ou à des situations de communications semblables. En étudiant ces manifestations langagières, les apprenants peuvent se rendre compte des formes d'expression typiques (le « langage de tout le monde ») ainsi que de leurs variantes possibles. Les textes inclus dans le corpus peuvent ainsi servir de modèles pour les textes des apprenants.

Les activités proposées dans cette partie reposent sur les principes de l'« Apprentissage sur corpus » (Data-driven learning) et peuvent servir, entre autres, à consolider le vocabulaire thématique, l'usage de certains éléments linguistiques (mots à significations multiples, synonymes et d'autres éléments) ainsi qu'à établir le profil de l'élément langagier choisi. Aux niveaux A1, A2 et B1, ces exercices peuvent être effectués à l'aide des différentes parties du corpus pédagogique. Cet entrainement à l'usage des corpus et des outils d'analyse peut préparer les apprenants à manipuler de manière autonome des corpus linguistiques plus complexes aux niveaux de compétences supérieurs.

Conclusions

Ce travail de recherche s'inscrit dans une approche interdisciplinaire, au croisement de la linguistique et de la didactique des langues. Il avait pour but d'étudier comment les résultats récents dans le domaine des sciences du langage peuvent contribuer à augmenter l'efficacité de l'approche pédagogique de l'enseignement des langues, avant tout pour les niveaux de compétences linguistiques inférieurs (A1-B1 du Cadre européen commun de référence pour les langues). La principale caractéristique de ces trois niveaux (A1, A2 et B1) est que les apprenants ne sont pas encore des utilisateurs autonomes de la langue-cible. Il est utile de rappeler ici les deux hypothèses par lesquelles nos explorations ont été guidées :

> (1) Il est possible de démontrer une interconnexion étroite entre le lexique et la grammaire dans le cas du hongrois, langue morphologiquement complexe. Si une telle interconnexion existe, elle peut être exploitée au service non seulement de la création de nouveaux contenus pédagogiques mais aussi du développement de nouvelles approches méthodologiques pour l'enseignement des langues.
> (2) Introduire l'utilisation des corpus dès le début de l'apprentissage peut enrichir l'expérience de cet apprentissage et fournir des opportunités d'observation de l'usage langagier que d'autres ressources (manuels, dictionnaires...) ne peuvent offrir. Pour répondre aux besoins des apprenants aux niveaux de compétences linguistiques inférieurs (A1–B1), il est nécessaire de construire des corpus qui leur soient spécifiquement adaptés. Les principes de construction seront différents de ceux définis pour les corpus à des fins linguistiques.

Afin de valider (ou réfuter) ces hypothèses, nous avons examiné des questions liées, d'un côté, au contenu de l'enseignement (*enseigner quoi ?*) et à ses méthodes de l'autre (*enseigner comment ?*). Nous avons opté pour une démarche s'appuyant sur les questions suivantes :

> (1) Que révèlent précisément les résultats linguistiques récents à propos de l'usage langagier susceptible d'être pertinent dans le cadre de l'enseignement des langues ? Dans quelles mesures ces résultats, permettent-ils une meilleure description de l'usage langagier des natifs dans un objectif pédagogique ?
> (2) Comment appliquer les méthodes empiriques à une langue à morphologie complexe telle que le hongrois dans l'objectif d'améliorer son enseignement ?

(3) Comment peut-on intégrer les résultats de la recherche empirique dans les ouvrages pédagogiques ?
(4) Est-il utile de créer des corpus spéciaux à fins pédagogiques ? Si oui, quels sont les principaux critères de leur construction ?
(5) Comment explorer les corpus pédagogiques en tant qu'enseignant et en tant qu'apprenant ?

Pour pouvoir répondre à ces questions, les trois parties de cette thèse se sont concentrées sur trois axes d'utilisation de corpus par trois publics différents :

- Au cœur de la Partie I a été placée *la recherche en linguistique de corpus* ainsi que l'intégration des résultats dans quelques ouvrages pédagogiques existants.
- La Partie II a été consacrée à l'utilisation de corpus par *les acteurs de la scène pédagogique* : non seulement les linguistes souhaitant explorer les corpus existants pour une meilleure description du hongrois dans le cadre pédagogique mais aussi les auteurs de manuels et les enseignants).
- La Partie III a enfin présenté diverses approches par lesquelles *les apprenants* peuvent explorer les corpus créés à fins pédagogiques.

Comme nous l'avons démontré dans cette thèse, ces divers publics sont susceptibles de s'intéresser à l'utilisation de corpus pour des raisons différentes : le linguiste dans le but de fournir une meilleure description de la langue, les auteurs de manuels pour rendre les textes de leurs ouvrages plus proches de la langue utilisée par les natifs, les enseignants afin d'enrichir leur boîte-à-outils et les apprenants pour améliorer leurs connaissances linguistiques. Les paragraphes suivants résumeront les éléments les plus importants de ces différents modes d'exploitation.

Dans la **Partie I** de ce travail, nous avons présenté les résultats de la recherche linguistique empirique concernant l'usage langagier des natifs et des utilisateurs experts de la langue. Les grandes collections d'énoncés authentiques (les corpus) nous renseignent sur des expériences linguistiques probables des locuteurs (Hoey 2005 : 185) ; leur étude peut révéler des caractéristiques échappant à l'analyse manuelle et à l'introspection ou à l'intuition du linguiste.

Ces résultats indiquent, en premier lieu, que l'usage langagier des natifs est loin d'être original. Différents locuteurs se trouvant loin l'un de l'autre dans l'espace et dans le temps ont tendance à

utiliser leur langue de façon similaire dans les interactions analogues. Une grande partie de nos énoncés semble consister en unités de plusieurs mots (unités multi-lexicales) que la communauté des locuteurs natifs utilise de manière répétée. L'explication de l'omniprésence des unités mufti-lexicales proposée dans cette thèse est fondée sur les travaux de Firth et de l'école néo-firthienne : elle implique que les éléments de ces unités ne sont pas sélectionnés séparément, l'un après l'autre mais ensemble, car c'est ainsi qu'ils peuvent former des unités de sens non-ambiguës, simples à décoder par les locuteurs. L'analyse outillée nous permet de cartographier les environnements textuels typiques de ces unités ainsi que leurs composantes lexicales, grammaticales, sémantiques et pragmatiques. Ces composantes ne sont pas séparables mais représentent différentes facettes de l'usage langagier étroitement interconnectées. D'autres résultats pertinents présentés dans cette partie incluent l'inséparabilité du lexique et de la grammaire, l'importance du registre et celle du langage interactionnel.

Les derniers chapitres de cette partie ont été consacrés à la présentation de quelques ouvrages pédagogiques. Deux grammaires et deux manuels illustrent les autres manières permettant d'intégrer les résultats de la linguistique de corpus dans les matériels de cours. Ces ouvrages démontrent que les exemples tirés des corpus ainsi que les informations reposant sur l'analyse outillée peuvent largement contribuer à la présentation d'un langage à caractère naturel. Il est également possible d'enrichir le matériel pédagogique avec des exercices originaux (activités d'observation, études des lignes de concordance, intégration des mots fréquents dans des textes, etc.) pour rendre l'apprentissage plus efficace.

Dans la **Partie II**, nous avons étudié diverses façons d'utiliser les grands corpus du hongrois pour augmenter l'efficacité de son enseignement. Il s'agit d'une langue à morphologie complexe qui n'a été jusqu'à présent que peu explorée au moyen d'outils numériques. Les corpus à fins linguistiques et pédagogiques sont également peu nombreux pour cette langue. En nous basant sur les résultats présentés dans la Partie I de cette thèse, nous avons, avant tout, examiné si la théorie d'interconnexion forte entre les composantes lexicales et grammaticales s'appliquait aussi à la langue hongroise.

Notre étude s'est concentrée sur une série de questions « problématiques », fréquemment posées par les apprenants dont une partie est considérée traditionnellement principalement de nature grammaticale et l'autre de nature lexicale. Nous avons analysé des mots à usages multiples dont le sens est largement défini par l'environnement textuel comme « nehéz » (difficile, lourd), les

synonymes « megjön/eljön » (qui se traduisent par « venir » ou « arriver ») et « tűnik/látszik » (comparables aux verbes français « sembler » et « paraître ») ainsi que les deux conjugaisons, particularité de la langue hongroise. Les étapes de notre recherche empirique présentées dans les chapitres de la Partie II ont révélé l'importance de l'environnement textuel pour l'usage des éléments étudiés. Notre analyse démontre, en particulier, l'importance de l'usage de mots concrets en tant que COD pour les deux conjugaisons. Notre travail éclaire également le mécanisme de la co-sélection des noms avec les verbes à préfixe et les mots à usages multiples. Les résultats de ces explorations ont validé notre première hypothèse : *nous avons pu constater qu'il était, en effet, possible de révéler une interconnexion étroite entre lexique et grammaire pour la langue hongroise, au moins pour les phénomènes étudiés*. La mise en avant de cette interconnexion pourrait significativement enrichir la présentation des phénomènes linguistiques choisis dans le cadre pédagogique.

La **Partie III** de notre thèse a porté sur les questions concernant la création et l'exploitation des corpus à fins pédagogiques. Nous avons analysé la construction et les possibilités d'utilisation du corpus de « MagyarOK », série de manuels pour la langue hongroise. Ce corpus a été construit en parallèle à l'écriture de cette thèse.

Que les grands corpus à fins linguistiques ne se prêtent pas automatiquement à l'utilisation dans le cadre pédagogique, car leurs contenus ne sont pas toujours adéquats pour les apprenants aux niveaux de compétences linguistiques inférieurs, a déjà été mentionné dans la Partie I. Cette partie a également résumé les caractéristiques communes aux corpus pédagogiques identifiés pour des langues plus largement enseignées comme l'anglais, le français et l'italien. Selon ces critères, les corpus pédagogiques doivent tenir compte des besoins des apprenants et proposer des contenus linguistiques qui leur sont adaptés et accessibles.

Comment remplir ces conditions aux niveaux A1–B1 ? Pour ces niveaux, une exigence (non applicable aux corpus linguistiques) s'impose : la cohérence intertextuelle. Cela signifie que les corpus doivent contenir *des textes concernant des sujets pertinents* pour les apprenants, c'est-à-dire des énoncés relatifs à des situations diverses que les apprenants doivent maitriser à leur niveau de compétences. Cette approche garantit que les apprenants seront capables d'interpréter les textes proposés et d'en intégrer les éléments choisis dans leur propre usage langagier. Dans l'intérêt de l'accessibilité et de la pertinence, les textes initiaux doivent être présélectionnés et soumis à un contrôle de qualité, une étape de « filtrage » qui impacte nécessairement l'authenticité du contenu

du corpus. Le résultat d'une telle adaptation n'est plus un langage authentique mais *un langage produit dans des situations authentiques, modifié* ou *un langage à caractère naturel*.

Comment construire des corpus de taille significative pour les niveaux qui n'ont encore que peu de connaissances linguistiques ? Comme nous l'avons vu, il est possible de proposer des variations de textes (variations écrites et orales, monologues et dialogues, interactions authentiques, improvisations et textes rédigés de façon plus ou moins contrôlée) autour du même sujet et/ou de situations de communication analogues.

Il n'existe à l'heure actuelle qu'un nombre très limité de corpus construits pour les niveaux de compétences linguistiques inférieurs (A1–B1), le corpus de « MagyarOK » étant le premier pour le hongrois. Nous avons tâché de démontrer que ce corpus, ainsi que les corpus construits d'une façon similaire, peuvent offrir de nombreux avantages :

> (1) Les corpus pédagogiques bien construits seront d'une taille bien plus large que n'importe quel ouvrage pédagogique. En étudiant un grand nombre d'exemples et en observant les contextes d'utilisation les plus courants, l'apprenant peut accéder à une meilleure compréhension de l'usage des éléments lexicaux inclus dans le corpus.
> (2) L'apprenant ainsi que l'enseignant ont l'opportunité de retenir des informations concernant les propriétés grammaticales, sémantiques et pragmatiques des unités multi-lexicales fréquentes avec un mot-clé choisi. Ils peuvent également se rendre compte que les différentes composantes de la langue sont interconnectées. Au cours de l'observation de l'élément choisi, l'apprenant obtient des informations sur tous les éléments contenus dans les énoncés.
> (3) L'élargissement graduel de l'environnement textuel et l'étude de répétitions et de variations permettent l'observation au-delà du mot-clé. Même si celui-ci reste tout au long au centre des explorations, l'apprenant s'expose également à d'autres éléments langagiers.
> (4) Assimiler des usages récurrents identifiés dans le corpus permet à l'apprenant d'observer et d'acquérir des éléments linguistiques qui lui permettent de participer à des interactions avec les natifs, de comprendre des textes et de développer des compétences linguistiques. Les schémas observés lui permettent également de créer de nouveaux énoncés et d'utiliser en conséquence la langue de façon créative.

(5) L'enseignant enrichit son répertoire pédagogique de nouveaux outils qui lui permettent de fournir des renseignements sur l'usage langagier à partir des observations plutôt qu'à partir de son intuition.
(6) Il peut utiliser le contenu du corpus pédagogique à illustrer un phénomène considéré comme problématique par l'apprenant ou composer des exercices (du format classique ou novateur) à partir des exemples issus du corpus.

La Partie III de ce travail de recherche a ainsi validé notre deuxième hypothèse : il est possible et utile d'introduire le travail avec les corpus pédagogiques dès le début de l'apprentissage. Ces corpus doivent être construits selon des principes adaptés pour pouvoir être explorés de façon efficace par les apprenants aux niveaux de compétences linguistiques inférieurs. Les opportunités d'apprentissage offertes par ces ressources (observation d'usage langagier dans une variété de textes, révision des aspects sélectionnés du vocabulaire en contexte, intégration des éléments langagiers dans les textes de l'apprenant) sont uniques ; ces conditions ne pourraient être créées que dans une mesure bien plus limitée en utilisant seulement les ressources « classiques » (manuels, dictionnaires...).

Limites

Malgré les avantages évidents de l'utilisation des corpus dans le cadre pédagogique, nous ne pouvons cependant pas prétendre que leur utilisation soit une panacée à toutes les questions émergeant au cours de l'apprentissage et à l'enseignement des langues. Même s'il s'agit d'outils puissants qui permettent d'augmenter l'effectivité de notre démarche pédagogique, certaines limites s'imposent. Les paragraphes suivants les présenteront brièvement.

La première limite concerne le choix du corpus approprié. Dans le cas des langues les plus largement enseignées, au moins en Europe (allemand, anglais, français), les utilisateurs ont de nombreux corpus à leur disposition et ils peuvent sélectionner celui qui convient le mieux à leurs objectifs. Pour les langues moins fréquemment enseignées (hongrois, suédois, roumain ...), ce choix est bien plus limité. Ainsi, le corpus pédagogique de hongrois présenté dans la Partie III de cette thèse est approprié à un certain public (apprenants aux niveaux A1 à B1 dans les cours généraux), mais la taille et le contenu de ce corpus sont limités.

Construire son propre corpus est, en raison de la facilité de trouver et de collectionner des textes existants, une option envisageable pour créer d'autres ressources plus adaptées selon les différents

contextes pédagogiques. Ainsi, les étudiants de niveaux de compétences linguistiques supérieurs peuvent eux-mêmes compiler leurs collections de textes, en fonction de leurs besoins (sur un sujet qui les intéresse ou des textes d'un certain genre, par exemple des dossiers semestriels dans une certaine matière). En revanche, cette option n'est pas accessible aux apprenants aux niveaux linguistiques inférieurs qui ne sont pas encore des utilisateurs autonomes de la langue-cible. Pour ces niveaux, d'autres intervenants (professeurs, auteurs/éditeurs d'ouvrages pédagogiques) doivent créer des corpus idoines. Or, la création d'un corpus est un travail chronophage et fastidieux même s'il s'agit de produire un corpus de taille modeste, comme nous l'avons vu dans les derniers chapitres de cette thèse (il faut trouver les textes qui conviennent, les adapter, solliciter des natifs à écrire des textes, travailler avec des acteurs ...). Ce travail ne peut être effectué que si les acteurs de la scène pédagogiques se sentent tous concernés et partagent les tâches qu'impose une telle entreprise.

Le dernier point concerne l'interprétation des résultats. Les réponses fournies par le corpus seront plus détaillées et plus précises que les réponses que l'on peut obtenir en s'appuyant sur le dictionnaire ou encore sur son intuition/introspection, mais elles seront aussi plus complexes. Analyser ces résultats exige non seulement de la pratique mais aussi de la confiance de la part de l'apprenant et de l'enseignant d'être capables d'interpréter leurs observations. Or, les études évoquées dans la deuxième partie de cette thèse montrent que les enseignants ne se sentent pas tous aptes à lire les lignes de concordance, à en identifier les schémas et à en proposer des résultats de façon claire et cohérente. Il est ainsi peu probable qu'ils fassent découvrir ces outils à leurs apprenants.

Certaines de ces limites peuvent être écartées, au moins en partie, en prenant en considération quelques suggestions pour l'avenir que nous allons esquisser dans cette dernière section.

Considérations futures

Quel avenir pour les corpus et pour les outils numériques dans le cadre de l'enseignement des langues ? Les résultats de ce travail de recherche suggèrent que l'« approche corpus » peut contribuer à l'amélioration de la pratique pédagogique de plusieurs manières. Ils indiquent également que le travail très significatif que demandent l'analyse outillée, la didactisation des résultats ainsi que la création des corpus ne peuvent pas être réalisées par des individus séparés, mais nécessitent une collaboration étroite entre les acteurs de la scène pédagogique, c'est-à-dire entre linguistes, auteurs de manuels, éditeurs, formateurs d'enseignants, enseignants et apprenants.

Seule une telle démarche permettra (1) d'analyser les langues peu explorées à présent (comme le hongrois) avec des outils numériques et (2) d'intégrer les résultats de ces études dans les ouvrages pédagogiques. Un tel effort collectif pourrait également rapprocher la recherche dans le domaine des sciences du langage et dans celui de la didactique des langues (deux domaines qui prennent souvent des chemins bifurquants) et faciliter ainsi la création de nouveaux corpus pédagogiques ainsi que le développement d'une méthodologie pour leur application efficace.

Ce travail seul ne garantira cependant pas le succès d'une « approche corpus » sans l'engagement actif des acteurs principaux de la dissémination : les enseignants. Ce sont eux qui définissent le cadre de l'apprentissage et dirigent l'apprenant vers les outils et méthodes qu'ils considèrent les plus efficaces. Ainsi, leur attitude envers les nouvelles ressources est décisive : s'ils n'ont que peu de connaissances les concernant et ne se sentent pas aptes à travailler avec les outils proposés (ou avec les ouvrages fondés sur l'analyse de corpus), il est peu probable qu'ils introduiront le travail avec les corpus dans leurs cours.

Comment sensibiliser les enseignants à l'utilisation de corpus ? Lors de nos explorations se sont dessinés cinq types de connaissances qui mériteraient d'être inclus dans le curriculum de formation des futurs enseignants de langues pour leur permettre de développer les compétences nécessaires et mieux les préparer à travailler avec ces nouveaux outils :

> (1) Connaître les résultats majeurs issus du domaine de la linguistique empirique. Ces résultats dirigent l'attention sur l'usage langagier observable et les habitudes langagières des natifs telles qu'elles se révèlent à travers des énoncés réels. Ils accentuent également l'interconnexion de différents aspects de la langue, avant tout celle du lexique et de la grammaire.
> (2) Connaître les bases de l'analyse outillée de corpus. Acquérir ces connaissances permet aux enseignants de mieux comprendre le processus d'analyse et de savoir interpréter ses résultats par eux-mêmes.
> (3) Connaître les ouvrages pédagogiques intégrant des résultats d'analyses empiriques. Ces ouvrages peuvent significativement améliorer la qualité de l'enseignement en proposant un usage langagier à caractère naturel, proche de celui des natifs.
> (4) Connaître les corpus pédagogiques existants pour la langue enseignée. Ces corpus offrent la possibilité d'explorations linguistiques au niveau des apprenants et permettent la création de nouveaux exercices et de nouvelles activités.

(5) Connaître les principes de base pour de création de corpus pédagogiques. Cela permettrait aux enseignants de construire leurs propres corpus selon les besoins des apprenants.

L'intégration de ces connaissances dans le parcours des futurs enseignants constituerait un pas significatif vers la diffusion de l'utilisation compétente des corpus dans le cadre pédagogique, alors que la coopération entre les acteurs décrite ci-dessus permettrait de développer de nouveaux contenus pédagogiques ainsi que de nouvelles approches méthodologiques cohérentes et efficaces à l'enseignement des langues.

Littérature

Ouvrages, chapitres d'ouvrage, articles

Ädel, A. (2010). Using corpora to teach academic writing : Challenges for the direct approach. *Corpus-based approaches to English language teaching,* 6(7), 39–55.

Ädel, A. & Reppen, R. (2008). *Corpora and discourse. The challenges of different settings*. John Benjamins.

Alan, D. & Widdowson, H. G. (1974). Reading and writing. In Allen, J. P. B. & Pit Corder, S. (dir.), *The Edinburgh course in applied linguistics 3*. Cambridge University Press, 155–201.

Allan, R. (2009). Can a graded reader corpus provide 'authentic' input? *ELT Journal*, 63(1), 23–32.

André, V. (2020a). Faire de la linguistique de corpus avec des apprenants de français langue étrangère. In Larrivée, P. & Lefeuvre, F. (dir.), *La didactisation du français vernaculaire.* Presses Universitaires de Caen, 37–66.

André V. (2020b). Corpus d'interactions et apprentissage du français langue étrangère. In Benzitoun, C. & Rebuschi, M. (dir.), *Les corpus en sciences humaines et sociales*. Presses Universitaires de Nancy, 101–121.

André, V. (2019). Des corpus oraux et multimodaux authentiques pour acquérir des compétences sociolangagières. In Gajo, L., Luscher, J.-M., Racine, I. & Zay, F. (dir.), *Variation, plurilinguisme et évaluation en français langue étrangère*. Peter Lang, 209–223.

André, V. (2017). Un corpus multimédia pour apprendre à interagir en situations universitaires en France. *Troisième colloque international de l'ATPF Enseigner le français: s'engager et innover,* Bangkok.

André, V. (2010). Éléments de construction collaborative du discours au sein de réunions de travail : la reprise et le couple oui non. *Pratiques. Linguistique, littérature, didactique,* 147–148, 199–222.

André V. & Ciekanski M. (2018). Apprendre à interagir à l'oral à partir d'un concordancier multimodal : effets sur le développement de la conscience langagière et sur l'autonomie de l'apprenant dans le dispositif FLEURON. In Dejean-Thircuir C., Mangenot F., Nissen E., Soubrié T. (dir.), *EPAL – Échanger pour apprendre en ligne. Actes de la 6ème édition du colloque EPAL*, Grenoble.

Arnon, I., and Snider, N. (2010). More than words: frequency effects for multi-word phrases. *Memory and Language.* 62, 67–82.

Asención-Delaney, Y. (2014). A Multi-Dimensional analysis of advanced written L2 Spanish. In Berber-Sardinha, T. & Veirano Pinto, M. (dir.), *Multi-Dimensional Analysis, 25 years on. A tribute to Douglas Biber.* John Benjamins, 239–270.

Aston, G. (2001). *Learning with corpora.* Athelstan.

Aston, G. (1997). Small and large corpora in language learning. *Practical applications in language corpora*, 51–62.

Atta-Allah, F., Agnaou, F., Ansar, K., Bouhjar, A., Boulaknadel, S. et al. (2017). *Actes de l'atelier « Diversité Linguistique et TAL » (DiLiTal).* France.

Balogh, J. et al. (2000). *Magyar grammatika.* [*Grammaire hongroise.*] Nemzeti Tankönyvkiadó.

Baraldi, C. & Gavioli, L. (dir.) (2012). *Coordinating participation in dialogue-interpreting*. John Benjamins.

Bárczi, G. & Országh, L. (dir.) (2003). *A magyar nyelv értelmező szótára I–VII. kötet*. http://mek.oszk.hu/adatbazis/magyar-nyelv-ertelmezo-szotara/elolap.php

Bárdos, J. (2005). *Élő nyelvtanítás-történet*. [*Histoire de l'apprentissage des langues vivantes.*] Nemzeti Tankönyvkiadó.

Barlow, M. (2004). Software for corpus access and analysis. Sinclair, J. McH. (dir.), *How to use corpora in language teaching*. John Benjamins, 205–221.

Barlow, M. & Kemmer, S. (dir.) (2000). *Usage-based models of language*. CSLI Publications, University of California.

Basra, S. et Thoyyibah, L. (2017). A speech act analysis of teacher talk in an EFL classroom. *International Journal of Education,* 10(1), 73–81.

Bencze, N. (2020). Önkonstruálás és benyomáskeltés az első találkozás során [La construction du soi et la creation des premières impressions lors d'une première rencontre]. In Ludányi, Zs. & Gráczi, T. E. (dir.), *Actes de la XIV. Conférence de la lingustique appliquée*. Nyelvtudományi Intézet, 6–20.

Berber-Sardinha, T. (2019). Lexicogrammar. In : Chapelle, C. (dir.), *The Encyclopedia of applied linguistics*. Wiley-Blackwell, 1–5.

Berber-Sardinha, T. ; Kaufmann, C. & Mayer Acunzo, C. (2014). Dimensions of register variation in Brazilian Portuguese. In Berber-Sardinha, T. & Veirano Pinto, M. (dir.), *Multi-Dimensional Analysis, 25 years on. A tribute to Douglas Biber*. John Benjamins, 35–80.

Berger, C. R. & Roloff, M. E. (2019). *Interpersonal communication*. Routledge.

Bernardini, S. (2004). Corpora in the classroom: An overview and some reflections on future developments. In Sincair, J. (dir.), *How to use corpora in language teaching*. John Benjamins, 15–26.

Bernardini, S. (2000). Systematising serendipity: Proposals for concordancing large corpora with language learners. In Burnard, L. & McEnery, T. (dir.), *Rethinking language pedagogy from a corpus perspective*. Peter Lang, 225–234.

Biber, D. (2012). Register as a predictor of linguistic variation. *Corpus Linguistics and Linguistic Theory,* 8, 9–37.

Biber, D. (2009). A corpus-driven approach to formulaic language: multi-word patterns in speech and writing. *International Journal of Corpus Linguistics,* 14, 381–417.

Biber, D. (1993). The multi-dimensional approach to linguistic analyses of genre variation: An overview of methodology and findings. *Computers and the Humanities,* 26, 331–345.

Biber, D. (1988). *Variation across speech and writing*. Cambridge University Press.

Biber, D. ; Connor, U. & Upton, T. (2007). *Discourse on the move. Using corpus analysis to describe discourse structure*. John Benjamins.

Biber, D. & Conrad, S. (2009). *Register, genre and style*. Cambridge University Press.

Biber, D. ; Conrad, S. & Leech, G. (2002). *The Longman student grammar of spoken and written English*. Longman.

Biber, D. ; Conrad, S. ; Reppen. R. ; Byrd, P. & Helt, M. (2002). Speaking and writing in the university: A multi-dimensional comparison. *TESOL Quarterly,* 36, 9–48.

Biber, D. ; Conrad, S. & Reppen, R. (1998). *Corpus linguistics: Investigating language structure and use*. Cambridge University Press.

Biber, D. & Gray, B. (2010). Challenging stereotypes about academic writing: Complexity, elaboration, explicitness. *Journal of English for Academic Purposes,* 9, 2–20.

Biber, D. & Egbert, J. (2018). *Register variation online.* Cambridge University Press.

Biber, D. & Reppen, R. (2002). What does frequency have to do with grammar teaching? *Studies in Second Language Acquisition,* 24(2), 199–208.

Bisson, M. ; Van Heuven, W. J. B.; Conklin, K. & Tunneya, R. Y. (2014). The role of repeated exposure to multimodal input in incidental acquisition of foreign language vocabulary. *Language Learning,* 64(4), 855–977.

Blanche, P. ; Bertrand, R. ; Ferré, G. ; Pallaud, B. ; Prévot, L., & Rauzy, S. (2017). The corpus of interactional data: A large multimodal annotated resource. In Ide, N. & Pustejovsky, J. (dir.), *Handbook of linguistic annotation.* Springer, 1323–1356.

Boers, F. (2021). *Evaluating second language vocabulary and grammar instruction*. Routledge.

Boers, F. ; Demecheleer, M. ; He, L. ; Deconinck, J. ; Stengers, H. & Eyckmans, J. (2017). Typographic enhancement of multiword units in second language text. *International Journal of Applied Linguistics,* 27, 448–469.

Boers, F. & Lindstromberg, S (2009). *Optimizing a Lexical Approach to instructed second language acquisition*. Palgrave-MacMillan.

Boers, F. ; Eyckmans, J. & Stengers, H. (2006). Motivating multiword units: Rationale, mnemonic benefits, and cognitive style variables. In Foster-Cohen, S. H. ; Krajnovic, M. M. & J. M. Djigunovic (dir.), *EUROSLA yearbook 6*. John Benjamins, 169–190.

Bogaards, P. (2001). Lexical units and the learning of foreign language vocabulary. *Studies in Second Language Acquisition,* 23(3), 321–343.

Boronkai, D. (2011). Az interakciós szerkezet és a szociokulturális tényezők összefüggései a spontán társalgásokban. [La relation entre la structure interactionnelle et les facteurs socioculturels dans les conversations spontanées.] *Alkalmazott Nyelvtudomány,* 11(1–2), 151–168.

Borsos, L. (2014). Az extenzív olvasás a magyar mint idegen nyelv tanításában és tanulásában. [La lecture extensive dans l'enseignement et dans l'apprentissage du hongrois langue étrangère.] *Teaching Hungarian Language*, 2(1), 61–80.

Boulton, A. (2017). Research timeline: Corpora in language teaching and learning. *Language Teaching,* 50(4), 483–506.

Boulton, A. (2010a). Data-driven learning: Taking the computer out of the equation. *Language Learning,* 60(3), 534–572.

Boulton, A. (2010b). Consultation de corpus et styles d'apprentissage. *Cahiers de l'APLIUT,* 29(1), 98–115.

Boulton A. (2008). Esprit de corpus : promouvoir l'exploitation de corpus en apprentissage des langues, *Texte et Corpus*, 3, 37–46.

Boulton, A. & Cobb, T. (2017). Corpus use in language learning: A meta-analysis. *Language Learning* 67(2), 348–393.

Boulton, A. & Thomas, J. (2012). Hands-on/hands-off: Alternative approaches to data-driven learning. In J. Thomas and A. Boulton (dir.), *Input, process and product: Developments in teaching and language corpora*. Masaryk University Press, 153–169.

Boulton, A. & Tyne, H. (2014). *Des documents authentiques aux corpus. Collection langues et didactique*. Les Édition Didier.

Braun, S. (2010). Getting past Groundhog day. Spoken multimedia corpora for student-centred corpus exploration. In Harris T. & Moleno Jaén, M. (dir.), *Corpus linguistics in language teaching.* Peter Lang, 75–98.

Braun, S. (2007). Integrating corpus work into secondary education: From data-driven learning to needs-driven corpora. *ReCALL* 19(3), 307–328.

Braun, S. (2006). ELISA – a pedagogically enriched corpus for language learning purposes. In Braun, S. ; Kohn, K. & Mukherjee, J. (dir.), *Corpus technology and language pedagogy: New resources, new tools, new methods*. Peter Lang, 25–47.

Braun, S. (2005). From pedagogically relevant corpora to authentic language learning contents. *ReCALL,* 17, 47–64.

Brazil, D. (1995). *A grammar of speech.* Oxford University Press.

Breyer, Y. A. (2011). *Corpora in language teaching and learning: Potential, evaluation, challenges.* Peter Lang.

Brezina, V. (2018). *Statistics in corpus linguistics.* Cambridge University Press.

Brown, P. & Levinson, S. C. (1978). Universals in language usage : Politeness phenomena. In Goody, E. N. (dir.), *Questions and politeness: strategies in social interaction*. Cambridge University Press, 56–310.

Brown, G. & Yule, G. (1983). *Teaching the spoken language.* Cambridge University Press.

Budai, L. (2016). *A magyar mint idegen nyelv nyelvtana. Elmélet és gyakorlat. [Grammaire du hongrois comme langue étrangère. Théorie et pratique*.] Tinta Kiadó.

Burnhard, L. & McEnery, T. (dir.) (2000). *Rethinking language pedagogy from a corpus perspective*. Peter Lang.

Burns, A. (2001). Analysing spoken discourse: implications for TESOL. In Burns, A. and Coffin, C. (dir.), *Analysing English in a Global Context: A Reader*. Routledge, 123–148.

Buttery, P. ; McCarthy, M. J. & Carter, R. (2015). Chatting in the academy: informality in spoken academic discourse. In Charles, M. ; Groom, N. & John, S. P. (dir.), *Corpora, Grammar and Discourse*. John Benjamins.

Cardon, D. (2013). Les formes de la conversation sur Facebook et les réseaux sociaux. In *Le français parlé dans les médias. Discours, médias, technologies: que change le numérique? Colloque international*.

Carter, R. (2004). *Language and creativity. The art of common talk.* Routledge.

Carter, R. & McCarthy, M. (2006). *Cambridge grammar of English: A comprehensive guide. Spoken and written English grammar and usage.* Cambridge University Press.

Carter, R. ; McCarthy, M. ; Mark, G. & O'Keeffe, A. (2016). *English grammar today: An A–Z of spoken and written grammar*. Cambridge University Press.

Cavalla, C. (2019a). Comment former les étudiants de Master FLE à l'utilisation pédagogique des corpus numériques ? In Goes J., Meneses-Lerin L., Mangiante J.M., Olmo F. & Pineira-Tresmontant C. (dir.), *Apports et limites des corpus numériques en analyse de discours et didactique des langues de spécialité*, Editura Universitaria, 79–92.

Cavalla, C. (2019b). Corpus numériques : critères pour l'enseignement des langues. *PERL : Entre présence et distance. Enseigner et apprendre les langues à l'université à l'ère numérique*. Equipe PERL, Paris.

Cavalla, C. & Loiseau M. (2013). Scientext comme corpus pour l'enseignement. In Tutin A. & Grossman F. (dir.), *L'écrit scientifique : du lexique au discours. Autour de Scientext*. Rennes, PUR, 163–182.

Chambers, A. (2019). Towards the corpus revolution? Bridging the research-practice gap. *Language Teaching,* 52(4), 460–475.

Chambers, A. (2010). L'apprentissage de l'écriture en langue seconde à l'aide d'un corpus spécialisé. *Revue française de linguistique appliquée*, 15(2), 9–20.

Chambers, A. (2005). Integrating corpus consultation in language studies. *Language Learning and Technology,* 9(2), 111–25.

Chambers, A.; Farr, F. & O'Riordan, S. (2011). Language teachers with corpora in mind: From starting steps to walking tall. *Language Learning Journal,* 39(1), 85–104.

Chang, J-Y. (2014). The use of general and specialized corpora as reference sources for academic English writing: A case study. *ReCALL*, 26(2), 243–259.

Charaudeau, P. (2009). Dis-moi quel est ton corpus, je te dirai quelle est ta problématique. *Corpus*, 8, 37–66.

Charles, M. (2015). Same task, different corpus. In Boulton, A. & Leńko-Szymańska, A. (dir.), *Multiple Affordances of Language Corpora for Data-driven Learning*. John Benjamins.

Charles, M. (2014). Getting the corpus habit: EAP students' long-term use of personal corpora. *English for Specific Purposes,* 35(1), 30–40.

Charles, M. (2007). Reconciling top-down and bottom-up approaches to graduate writing: Using a corpus to teach rhetorical functions. *Journal of English for Academic Purposes*, 6(4), 289–302.

Chen, M., & Flowerdew, J. (2018). A critical review of research and practice in data-driven learning (DDL) in the academic writing classroom. *International Journal of Corpus Linguistics*, 23(3), 335–369.

Cheng, W. (2013). Semantic prosody. In Chapelle, C. A. (dir.), *The encyclopedia of applied linguistics*. Wiley-Blackwell, 1–7.

Cheng, W. ; Greaves, C. & Warren, M. (2008). *A corpus-driven study of discourse intonation*. John Benjamins.

Cheepen, C. (2014). Small talk in service dialogues: The conversational aspects of transactional telephone talk. In Coupland, J. (dir.), *Small Talk* (reprint). Longman, 288–311.

Chomsky, N. (1968). *Language and mind*. Harper and Row.

Clancy, B. & McCarthy, M. (2015). Co-constructed turn-taking. In Aijmer, K. & Rühlemann, C. (dir.). *Corpus pragmatics: A handbook*. Cambridge University Press. 430–453.

Cobb, T. (2014). A resource wish-list for data-driven learning in French. In Tyne, H. ; Andre, V. ; Boulton, A. ; Benzitoun, C. & Greub, J. (dir.), *French through corpora: Ecological and data-driven perspectives in French language studies*. Cambridge Scholars.

Cobb, T. & Boulton, A. (2015). Classroom applications of corpus analysis. In Biber, D. & Reppen, R. (dir.), *The Cambridge handbook of English corpus linguistics*. Cambridge University Press.

Conseil d'Europe (2021). *Common European Framework of Reference for Languages*. https://www.coe.int/en/web/common-european-framework-reference-languages

Conrad, S. (2009). *Real grammar: A corpus-based approach to English grammar*. Pearson Education.

Conrad, S. (2000). Will corpus linguistics revolutionize grammar teaching in the 21th century ? *TESOL Quarterly*, 34(3), 548–560.

Cook, G. (1998). The uses of reality: A reply to Ronald Carter. *ELT Journal* 52(1), 57–63.

Coupland, J. (dir.) (2014). *Small talk* (reprint). Routledge.

Cowie, A. P. (2002). Examples and collocations in the French Dictionnaire de langue. *Lexicography and Natural Language Processing. A Festschrift in honour of B. T. S. Atkins*. Göteborg University, 73–90.

Cowie, A. P. (1992). Multiword lexical units and communicative language teaching. *Vocabulary and Applied Linguistics*, 1–12.

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly,* 34, 213–238.

Crosthwaite, P. (2019). *Data-Driven Learning for the next generation: Corpora and DDL for pre-tertiary learners*. Routledge.

Cullen, R. (2002). Supportive teacher talk: The importance of the F- move. *ELT Journal*, 56(2), 117-127.

Cullen, R. (1998). Teacher talk and the classroom context. *ELT Journal*, 52(3), 179-187.

Davies, M. (2015). Corpora: An introduction. In Biber, D. & Reppen, R. (dir.): *The Cambridge handbook of English corpus linguistics*. Cambridge University Press, 11–32.

Davies, M. (2009). The 385+ million word Corpus of Contemporary American English (1990–2008+): Design, architecture, and linguistic insights. *International Journal of Corpus Linguistics*, 14(2), 159–190.

Davis, P. & Kryszewska, H. (2012). *The company words keep*. Delta Publishing.

De Fornel, M. (1990). Sémantique du prototype et analyse de conversation. *Cahiers de linguistique française,* 11, 159–179.

De Fornel, M. & Verdier, M. (2018). Corpus, classes d'exemples et collections en analyse de conversation. *Corpus,* 18, 1–15.

Di Vito, S. (2013). L'utilisation des corpus dans l'analyse linguistique et dans l'apprentissage du FLE. *Revue des linguistes de l'université Paris X Nanterre 68–69*, 159–176.

Dóla, M. (2014). Lexikon és grammatika kapcsolatáról – különös tekintettel az idegennyelv-tanulásra. [La relation entre lexique et grammaire – en particulier dans l'enseignement des langues]. *Hungarológiai Évkönyv,* 15, 8–25.

Domonkosi, Á. (2018a). Megszólítások a közösségi oldalak társalgásaiban [Formes de politesse dans les conversations sur les réseaux sociaux]. In Bank, B. (dir.), *Utak és útkereszteződések*. EKE Líceum Kiadó, 147–157.

Domonkosi, Á. (2018b). Nyilvánosság és attitűddeixis a közösségi oldalak diskurzusaiban. [Publicité et déixis attitudinal dans les discours sur les réseaux sociaux.] *Alkalmazott Nyelvészeti Közlemények*, 13(2), 63–75.

Dörnyei, Z. & Csizér, K. (1998). Ten commandments for motivating language learners: results of an empirical study. *Language Teaching Research,* 2(3), 203–229.

Egbert, J., Larsson, T. & Biber, D. (2020). *Doing linguistics with a corpus*. Cambridge University Press.

Ellis, N. C. (2008). Usage-based and form-focused SLA: The implicit and explicit learning of constructions. In Tyler, A. ; Kim, Y. & Takada, M. (dir.), *Language in the context of use: Cognitive and discourse approaches to language and language learning*. Mouton de Gruyter, 93–120.

Ellis, N. C. (2008). Phraseology: The periphery and the heart of language. In Meunier, F. & Granger, S. (dir.), *Phraseology in foreign language learning and teaching*. John Benjamins. 1–13.

Ellis, N. C. (2006). Selective attention and transfer phenomena in L2 acquisition: Contingency, cue competition, salience, interference, overshadowing, blocking, and perceptual learning. *Applied Linguistics,* 27, 164–194.

Ellis, N. C. (2002). Frequency effects in language acquisition: A review with implications for theories of implicit and explicit language acquisition. *Studies in Second Language Acquisition*, 24, 143–188.

Ellis, N. C. & Ogden, D. C. (2017). Thinking about multiword constructions: Usage-based approaches to acquisition and processing. *Topics in Cognitive Science,* 9(3), 604–620.

Ellis, N. C. ; Simpson-Vlach, R. ; Römer, U. ; Brook O'Donnell, M. & Wulff, S. (2015). Learner corpora and formulaic language in second language acquisition. In S. Granger, G. Gilquin, & F. Meunier (dir.), *The Cambridge handbook of learner corpus research,* Cambridge University Press, 357–378.

Ellis, R. (1985). *Understanding second language acquisition*. Oxford University Press.

Erjavec, T. (2004). MULTEXT - East Version 3: Multilingual Morphosyntactic Speci cations, Lexicons and Corpora. *Proceedings of the Fourth International Conference on Language Resources and Evaluation*. European Language Resources Association, 1535–1538.

Evert, S. (2009). Corpora and collocations. In Lüdeling, A. & Kytö, M. (dir.), *Corpus linguistics. An international handbook*, *vol. 2*. De Gruyter Mouton, 1212–1248.

Evison, J. ; McCarthy, M. & O'Keeffe, A. (2007). "Looking out for love and all the rest of it": Vague category markers as shared social space. In J. Cutting (dir.), *Vague language explored,* Palgrave Macmillan, 138–160.

É. Kiss, K. ; Kiefer, F. & Siptár, P. (2003). *Új magyar nyelvtan*. [*Nouvelle grammaire du hongrois*]. Osiris.

Farr, F. (2008). Evaluating the use of corpus-based instruction in a language teacher education context: Perspectives from the users. *Language Awareness*, 17(1), 25–43.

Fellbaum, C. (dir.) (2007). *Idioms and collocations*. Continuum.

Firth, J. R. (1957). *Papers in Linguistics 1934–1951*. Oxford University Press.

Fligelstone, S. (1993). Some reflections on the question of teaching, from a corpus linguistics perspective, *ICAME journal*, 17, 87–109.

Flowerdew, J. (2009). Corpora in language teaching. In Long, H. & Doughty, C. J. (dir.), *The handbook of language teaching*. Blackwell, 327–350.

Flowerdew, L. (2015). Data-driven learning and language learning theories. Whither the twain shall meet. In Boulton, A. & Leńko-Szymańska, A. (dir.), *Multiple affordances of language corpora for Data-driven learning,* 15–37.

Flowerdew, L. (2009). Applying corpus linguistics to pedagogy. A critical evaluation. *International Journal of Corpus Linguistics* 14(3), 393–417.

Forgács, T. (2007). *Ungarische Grammatik*. Praesens.

Fortanet-Gómez, I. & Querol-Julián, M. (2010). The video corpus as a multimodal tool for teaching. In Campoy-Cubillo, M., Belles-Fortuno, B. & Gea-Valor, L. (dir.), *Corpus-based approaches to English language teaching. Corpus and discourse*. Continuum, 261–270.

Forti, L. & Spina, S. (2019). Corpora for linguists vs. corpora for learners: Bridging the gap in Italian L2 learning and teaching. *EL.LE*, 8(2), 349–362.

Francis, G., Hunston, S. & Manning, E. (1998). *Collins COBUILD grammar patterns 2: Nouns and adjectives*. HarperCollins.

Francis, G., Hunston, S. & Manning, E. (1996). *Collins COBUILD grammar patterns 1: Verbs*. HarperCollins.

Frankenberg-García, A. (2014). The use of corpus examples for language comprehension and production. *ReCALL*, 26(2), 128–146.

Frankenberg-García, A. (2012). Raising teachers' awareness of corpora. *Language Teaching* 45(4), 475–489.

Frankenberg-García, A., Flowerdew, L. & Aston, G. (dir.) (2011). *New trends in corpora and language learning*. Bloomsbury.

Frérot, C. & Pecman, M. (2021) (dir.). *Des corpus numériques à l'analyse linguistique en langues de spécialité*. UGA Éditions.

Friginal, E. (2018). *Corpus linguistics for English teachers: Tools, online resources, and classroom activities*. Routledge.

Friginal, E. (2017). Developing research report writing skills using corpora. *English for Specific Purposes*, 32, 208–220.

Gabrielatos, C. (2005). Corpora and language teaching: Just a fling or wedding bells? *Teaching English as a Second Language – Electronic Journal,* 8(4), 1–35.

Gablasova, D., Brezina, V. & McEnery, T. (2017). Collocations in corpus-based language learning research: Identifying, comparing, and interpreting the evidence. *Language Learning,* 67(1), 155–179.

Gavioli, L. (1997). Bookshop service encounters in English and Italian: Notes on the achievement of information and advice. In Bargiela-Chiappini, F. & Harris, S. *Languages of business. An international perspective.* Edinburgh University Press, 136–158.

Geluso, J. & Hirch, R. (2019). The reference corpus matters: Comparing the effect of different reference corpora on keyword analysis. *Register Studies,* 1(2), 209–242.

German, C. & Seguin, H. (1995). *Le point sur la grammaire en didactique des langues*. Montréal CEC.

Girard, M. & Sionis, C. (2004). The functions of formulaic speech in the L2 class, *Pragmatics,* 14(1), 31–53.

Gilquin, G. & Granger, S. (2015). Learner language. In Biber, D. & Reppen, R. (dir.), *The Cambridge handbook of English corpus linguistics*. Cambridge University Press, 418–437.

Gilquin, G. & Granger, S. (2010). How can data-driven learning be used in language teaching? In O'Keeffe, A. & McCarthy, M. (dir.), *The Routledge handbook of corpus linguistics*. Routledge, 359–371.

Givón, T. (2001). *Syntax. Volume 1*. John Benjamins.

Godwin-Jones, R. (2017). Data-informed language learning. *Language Learning and Technology*, 21(3), 9–27.

Goldberg, A. (2005). *Constructions at work. The nature of generalization in language*. Oxford University Press.

Goldberg, A. (1995). *Constructions : A construction grammar approach to argument structure*. University of Chicago Press.

Granger, S. (2001). Didactique des langues étrangères, linguistique de corpus et traitement automatique des langues. *Questions d'épistémologie en didactique du français (langue maternelle, langue seconde, langue étrangère)*, 105–109.

Granger, S. (1998). Prefabricated patterns in advanced EFL writing: Collocations and formulae. In Cowie, A. P. (dir.), *Phraseology: Theory, analysis, and applications*. Oxford University Press, 145–160.

Gray, B. & Biber, D. (2011). Corpus approaches to the study of discourse. *Continuum companion to discourse analysis,* 3, 138–154.

Gréa, P. (2017). *Probabilités et statistiques en psychologie et en linguistique : Petit tour d'horizon.* Textes et Cultures, Institut Ferdinand de Saussure.

Gries, S. (2015). Quantitative designs and statistical techniques. In Biber, D. & Reppen, R. (dir.), *The Cambridge handbook of English corpus linguistics*. Cambridge University Press, 50–72.

Grieves, C. & Warren, M. (2010). What can a corpus tell us about multiword units? In O'Keeffe, A. & McCarthy, M. (dir.), *The Routledge handbook of corpus linguistics*. Routledge, 212–227.

Habert, B. (2001). Des corpus représentatifs : de quoi, pour quoi, comment ? In Bilger, M. (dir.), *Linguistique sur corpus. Études et réflexions*. Presses Universitaires de Perpignan, 11–58.

Habert, B., Nazarenko, A. & Salem, A. (1997). *Les linguistiques de corpus*. Armand Colin.

Halliday, M. (1985). An introduction to functional grammar. London, Edward Arnold.

Hanks, P. (2013). *Lexical analysis. Norms and exploitations.* MIT Press.

Hanks, P. (2012). The corpus revolution in lexicography. *International Journal of Lexicography,* 25(4), 398–436.

Harrington, M. & Dennis, S. (2002). Input-driven language learning. *Studies in second language acquisition*, 261–268.

Hegedűs, R. (2004). *Magyar nyelvtan. Formák, funkciók, összefüggések. [Grammaire hongroise. Formes, fonctions, connexions.]* Tinta Könyvkiadó.

Henry, A. & Roseberry, R. L. (2001). A narrow-angled corpus analysis of moves and strategies of the genre: 'Letter of Application'. *English for Specific Purposes,* 1(1), 153–167.

Hilpert, M. (2014). *Construction grammar and its application to English*. Edinburgh University Press.

Hilpert, M. & Mair, C. (2015). Grammatical change. In Biber, D. & Reppen, R. (dir.), *The Cambridge handbook of English corpus linguistics*. Cambridge University Press, 180–201.

Hoey, M. (2014). Words and their neighbours. In Taylor, J. R. (dir.), *The Oxford handbook of the word.* Oxford University Press, 141–156.

Hoey, M. (2009). Corpus-driven approaches to grammar. In Römer, U. & Schulze, R. (dir.), *Exploring the grammar-lexis interface.* John Benjamins, 33–49.

Hoey, M. (2005). *Lexical priming: A new theory of words and language*. Routledge.

Hoey, M. (1991). *Patterns of lexis in text*. Oxford University Press.

Holmes, J. (2014). Institutional identity work: a better lens. In Coupland, J. (dir.), *Small talk*. Longman, 84–109.

Hong Zang, V. (2020). What do you know about semantic prosody? Teaching and evaluating implicit knowledge of English with corpus-assisted methods. *English in Education,* https://doi.org/10.1080/04250494.2020.1838896

Howarth, P. (1998a). Phraseology and second language proficiency. *Applied Linguistics* 19(1), 24–44.

Howarth, P. (1998b). The phraseology of learners' academic writing. In Cowie, A. P. (dir.), *Phraseology: Theory, analysis, and applications*. Clarendon Press, 161–86.

Hughes, R. (2010). What a corpus tells us about grammar teaching materials. In O'Keeffe, A. & McCarthy, M. (dir.), *The Routledge handbook of corpus linguistics*. Routledge, 401–413.

Hunston, S. (2015). Lexical grammar. In Biber, D. & Reppen, R. (dir.), *The Cambridge handbook of English corpus linguistics*. Cambridge University Press, 201–216.

Hunston, S. (2010). How can a corpus be used to explore patterns? In O'Keeffe, A. & McCarthy, M. (dir.), *The Routledge handbook of corpus linguistics*. Routledge, 152–167.

Hunston, S. (2009). The usefulness of corpus-based descriptions of English for learners: The case of relative frequency. In Aijmer, K. (dir.), *Corpora and language teaching*, 141–157. John Benjamins.

Hunston, S. (2008). Starting with the small words. *International Journal of Corpus Linguistics* 13(3), 271–295.

Hunston, S. (2007). Semantic prosody revisited. *International Journal of Corpus Linguistics,* 12, 249–268.

Hunston, S. (2002). *Corpora in applied linguistics*. Cambridge University Press.

Hunston S. & Francis, G. (2000). *Pattern grammar. A corpus-driven approach to the lexical grammar of English.* John Benjamins.

Izumi, S. (2002). Output, input enhancement, and the noticing hypothesis: An experimental study on ESL relativization. *Studies in second language acquisition,* 24(4), 541–577.

Jantunen, J. H. & Brunni, S. (2013). Morphology, lexical priming and second language acquisition. A corpus study on learner Finnish. In Granger, S. ; Gilquin, G. & Meunier, F. (dir.), *Twenty years of learner corpus research. Looking back, moving ahead*. Presse universitaire de Louvain, 235–245.

Jones, M. & Durrant, P. (2010). What can a corpus tell us about vocabulary teaching materials? In O'Keeffe, A. & McCarthy, M. (dir.), *The Routledge handbook of corpus linguistics*. Routledge, 387–401.

Jones, C. & Waller, D. (2015). *Corpus linguistics for grammar. A guide for research*. Routledge.

Johns, T. (1994). From printout to handout: Grammar and vocabulary teaching in the context of data-driven learning. In Odlin, T. (dir.). *Perspectives on pedagogical grammar.* Cambridge University Press, 293–313.

Johns, T. (1991). Should you be persuaded: Two samples of data-driven learning materials. *English Language Research Journal*, 4, 1–16.

Kálmán, L. (dir.) (2001). *A magyar nyelv leíró nyelvtana [Grammaire descriptive de la langue hongroise]*. Budapest: Tinta Könyvkiadó.

Kaltenböck, G. & Mehlmauer-Larcher, B. (2005). Computer corpora and the language classroom: on the potential and limitations of computer corpora in language teaching. *ReCALL,* 65–84.

Kamber, A. (2011). Contexte et sens : utilisation d'un corpus écrit dans l'enseignement/apprentissage du FLE. *Travaux neuchâtelois de linguistique (Tranel)*, 55, 199–218.

Kamber, A. & Dubois, M. (2016). Corpus, grammaire et français langue étrangère : une concordance nécessaire. *Linguistik Online,* 78(4), 3–9.

Katinskaia, A. & Sharoff, S. (2015). Applying Multi-dimensional Analysis to a Russian webcorpus: Searching for evidence of genres. *Conference on Balto-Slavic Natural Language Processing*. Hissar.

Kennedy, C. & Miceli, T. (2017). Cultivating effective corpus use by language learners. *Computer Assisted Language Learning,* 30(1–2), 91–114.

Kennedy, C. & Miceli, T. (2010). Corpus-assisted creative writing: Introducing intermediate Italian learners to a corpus as a reference resource. *Language Learning and Technology,* 14(1), 28–44.

Keresztes, L. (1995). *Praktische ungarische Grammatik*. [*Grammaire pratique du hongrois*.] Debreceni Nyári Egyetem.

Keszler, B. (2017). *Magyar nyelvtan*. [*Grammaire du hongrois*.] Műszaki Könyvkiadó.

Kiefer, F. (dir.) (2006). *Magyar nyelv*. [*Langue hongroise*.] Akadémiai Könyvkiadó.

Kilgarriff, A. (2014). The Sketch Engine: Ten years on. *Lexicography,* 1(1), 7–36.

Kilgarriff, A. (2009). Simple maths for keywords. In Mahlberg, M., González-Díaz, V. & Smith, C. (dir.), *Proceedings of Corpus Linguistics Conference CL2009*. University of Liverpool.

Kilgarriff, A. (2006). Collocationality and how to measure it. In E. Corino, Marello C. & Onesti, C. (dir.), *Proceedings XII Euralex International Congress*. Edizioni dell'Orso.

Kiss, J. (1999). Nyelvi intuíció és elfogadhatósági ítéletek. [Intuition linguistique et des critères d'acceptance]. *Magyar Nyelv,* 95(2), 129–137.

Kjellmer, G. (1984). Why great : greatly but not big : bigly ? On the formation of English adverbs in -ly. *Studia Linguistica*, 38(1), 1–19.

Konopka, M. ; Wöllstein, A. & Felder, E. (2020). *Bausteine einer Korpusgrammatik des Deutschen,* vol. 1. Heidelberg University Publishing.

Kramer, A. (2011). Contexte et sens: utilisation d'un corpus écrit dans l'enseignement/apprentissage du FLE. *Revue TRANEL (Travaux neuchâtelois de linguistique),* 55, 199–218.

Krashen, S. (1989). We acquire vocabulary and spelling by reading: Additional evidence for the input hypothesis. *The Modern Language Journal,* 73, 440–464.

Krishnamurty, R. (2006). Collocations. In Brown, K. & Anderson, A. (dir.), *Encyclopedia of language and linguistics*. Elsevier, 596–600.

Kübler, N. (2014a). Mettre en œuvre la linguistique de corpus à l'université. Vers une compétence utile pour l'enseignement/apprentissage des langues ? *Recherches en didactique des langues et des cultures. Les cahiers de l'Acedle*, 11(1), 2–35.

Kübler, N. (2014b). Corpora and LSP translation. In Zanettin, F. ; Bernardini, S. & Stewart, D. (dir.), *Corpora in translator education.* Routledge, 29–46.

Kübler, N. ; Mestivier-Volanschi, M & Pecman, M. (2018). Teaching specialised translation through corpus linguistics: quality assessment and methodology evaluation by experimental approach. *META : Journal des traducteurs,* 63(3), 806–824.

Kytö, M. & Smitterberg, E. (2015). Diachronic registers. In Biber, D. & Reppen, R. (dir.), *The Cambridge handbook of English corpus linguistics*. Cambridge University Press, 330–346.

Landure, C. & Boulton, A. (2010). Using corpora in language learning. Language and use. *Recherche et pratiques pédagogiques en langues de spécialité – Cahiers de l'APLIUT*, 35(2), 52–67.

Langacker, R. W. (2008). Cognitive grammar as a basis for language instruction. In Robinson, P. & Ellis, N. C. (dir.), *Handbook of cognitive linguistics and second language acquisition*. Routledge, 66–88.

Langacker, R. (1991a). *Foundations of cognitive grammar*, v*ol 2*. Stanford University Press.

Langacker, R. (1991b). *Cognitive grammar: A basic introduction*. Oxford University Press.

Langacker, R. (1987). *Foundations of cognitive grammar, vol. 1*. Stanford University Press.

Larsson, T. (2019). Grammatical stance marking across registers. Revisiting the formal-informal dichotomy. *Register Studies,* 1(2), 243–268.

Leblanc, J-M. (2016). Phraséologie et formules rituelles dans le discours politique, l'expérimentation en lexicométrie, *Lidil*, 53, 43–69.

Lee, H. ; Warschauer, M. & Lee, J.H. (2019). The effects of corpus use on second language vocabulary learning: A multilevel analysis. *Applied Linguistics*, 40(5), 721–753.

Leech, G. (1997). Teaching and language corpora: A convergence. In Wichmann, A. ; Fligelstone, S. ; McEnery, T. & Knowles, G. (dir.), *Teaching and language corpora*. Longman, 1–23.

Leech, G. (2015). Descriptive grammar. In Biber, D. & Reppen, R. (dir.), *The Cambridge handbook of English corpus linguistics*. Cambridge University Press, 146–160.

Leech, G. (2007). New resources, or just better old ones? The Holy Grail of representativeness. In Hundt, M. ; Nesselhauf, N. & Biewer, C. (dir.), *Corpus linguistics and the web*. Brill Rodopi, 133–149.

Legallois, D. (2012). La colligation : autre nom de la collocation grammaticale ou autre logique de la relation mutuelle entre syntaxe et sémantique ? *Corpus,* 11, 31–54.

François, J. & Legallois, D. (2006). Autour des grammaires de construction et de patterns. *Cahiers du CRISCO*, 1–73.

Leńko-Szymańska, A. (2017). Training teachers in data-driven learning: Tackling the challenge. *Language Learning & Technology*, 21(3), 217–241.

Leńko-Szymańska, A. (2014). Is this enough? A qualitative evaluation of the effectiveness of a teacher-training course on the use of corpora in language education. *ReCALL*, 26(02), 260–278.

Leow, R. P. (2000). A study of the role of awareness in foreign language behavior: Aware versus unaware learners. *Studies in second language acquisition,* 22(4), 557–584.

Lewis, M. (1997). *Implementing the Lexical Approach: Putting theory into practice*. Hove.

Lewis, M. (1993): *The lexical approach*. Hove.

Liu, D. (2010). Is it a *chief*, *main*, *major*, *primary*, or *principal* concern? A corpus-based behavioral profile study of the near-synonyms. *International Journal of Corpus Linguistics,* 15(1), 56–87.

Mackey, A. (2006). Feedback, noticing and instructed second language learning. *Applied Linguistics,* 27(3), 405–430.

Mahlberg, M. (2017). Corpus stylistics. In Burke, M. (dir.), *The Routledge handbook of stylistics*. Routledge, 396–410.

Mahlberg, M. & Stockwell, P. (2015). Mind-modelling with corpus stylistics in David Copperfield. *Language and Literature,* 24(2), 129–147.

Malinowski, B. (1923). The problem of meaning in primitive languages. In Ogden, C. K. & Richards, I. A. *The meaning of meaning*. Kegan Paul, Supplement 1.

Manca, E. (2012). *Context and language*. Coordinamento, Università del Salento.

Mauranen, A. (2004). Spoken corpus for an ordinary learner. In Sincair, J. (dir.), *How to use corpora in language teaching*. John Benjamins, 89–105.

Markova, V. (2012). *Synonyme unter dem Mikroskop. Eine korpuslinguistische Studie*. Gunter Narr Verlag.

Martinez, R. & Schmitt, N. (2012). A phrasal expression list. *Applied Linguistics,* 33(3), 299–320.

McCarten, J. (2010). Corpus-informed course book design. In O'Keeffe, A. & McCarthy, M. (dir.), *The Routledge handbook of corpus linguistics*. Routledge, 413–428.

McCarten, J. & McCarthy, M. (2010). Bridging the gap between corpus and course book: the case of conversation strategies. In Chambers, A. & Mishan, F. (dir.), *Perspectives on language learning materials development*. Peter Lang, 11–32.

McCarthy, M. (2008). Assessing and interpreting corpus information in the teacher education context. *Language Teaching*, 41(4), 563–574.

McCarthy, M. (2003). Talking back: 'small' interactional response tokens in everyday conversation. *Research on Language in Social Interaction*, 36(1), 33–63.

McCarthy, M. (2002). Good listenership made plain: British and American non-minimal response tokens in everyday conversation. In Reppen, R. ; Fitzmaurice, S. & Biber, D. (dir.): *Using corpora to explore linguistic variation*. John Benjamins, 49–71.

McCarthy, M. (2000). Captive audiences: the discourse of close contact service encounters. In Coupland, J. (dir.), *Small talk*. Longman, 84–109.

McCarthy, M. (1999). What constitutes a basic vocabulary for spoken communication. *Studies in English Language and Literature,* 1, 233–249.

McCarthy, M. & Carter, R. (2017). Spoken grammar: Where are we and where are we going? *Applied Linguistics*, 38, 1–20.

McCarthy, M. & McCarten, J. (2019). Interaction management in academic speaking. *Revue des linguistes de l'université Paris X Nanterre,* 79, 1–19.

McCarthy, M. ; McCarten, J. & Sandiford, H. (2005–2011). *Touchstone*. Cambridge University Press.

McCarthy, M. ; McCarten, J. & Sandiford, H. (2012–2013). *Viewpoint*. Cambridge University Press.

McEnery, T. (2013). History of corpus linguistics. In Allen, K. (dir.), *Oxford handbook of the history of linguistics*. Oxford University Press, 749–768.

McEnery, T. & Hardie, A. (2012). *Corpus linguistics: Method, theory and practice*. Cambridge University Press.

McEnery, T. ; Xiao, R. & Tono, Y. (2006). *Corpus-based language studies: An advanced resource book*. Routledge.

McEnery, T. & Wilson, A. (1997). Teaching and language corpora. *ReCALL,* 9(1), 5–21.

McGarrell, H. M. & Lee, D. (2011). Corpus-based/Corpus-informed English language learner grammar textbooks: An example of how research informs pedagogy. *Research Symposium, TESL Ontario Conference, Toronto. Refereed Proceedings* 37(2), 78–100.

Mel'čuk, I. (2018). Theory and practice of lexicographic definition. *Journal of Cognitive Science,* 19(4), 417–-470.

Mel'čuk, I. (2003). Les collocations : définition, rôle et utilité. *Travaux et recherches en linguistique appliquée*, 1, 23–32.

Mel'čuk, I. (1998). Collocations and lexical functions. In A.P. Cowie (dir.), *Phraseology. Theory, Analysis, and Applications*, Clarendon Press, 23–53.

Meunier, F. (2012). Formulaic language and language teaching. *Annual review of applied linguistics*, 32(1), 111–129.

Meunier, F. & Granger, S. (dir.) (2008). *Phraseology in foreign language learning and teaching*. John Benjamins.

Meunier, F. & Reppen, R. (2015). Corpus versus non-corpus-informed pedagogical materials: Grammar as the focus. In Biber, D. & Reppen, R. (dir.), *The Cambridge handbook of English corpus linguistics*. Cambridge University Press, 498–514.

Mondada, L. (2002). Pour une approche interactionnelle de la catégorisation des ressources linguistiques pour les locuteurs. *Cahiers de l'Institut de linguistique de Louvain,* 28(3), 23–35.

Montenero Perez, M. & Rodgers, M. (2019). Video and language learning. *The Language Learning Journal,* 47(4), 403–406.

Moser, J. ; Harris, J. & Carle, J. (2012). Improving teacher talk through a task-based approach. *ELT Journal,* 66(1), 81–88.

Mukherjee, J. (2004). Bridging the gap between applied corpus linguistics and the reality of English language teaching in Germany. In Connor, U. & Upton, T. (dir.). *Applied corpus linguistics: A multi-dimensional perspective*. Rodopi, 239–250.

Nagyházi, B. (2017). *Szórendtanítás a magyar mint idegen nyelvben I-II.* [*Enseigner l'ordre des mots, hongrois langue étrangère*.] GlobeEdit.

Naismith, B. (2017). Integrating corpus tools on intensive CELTA courses. *ELT Journal*, 71(3), 273–283.

Nation, I. S. P. (2013). *Learning Vocabulary in Another Language*. Cambridge University Press.

Nation, I. S. P. & Waring, R. (1997). Vocabulary size, text coverage and word lists. In Schmitt, N. & McCarthy, M. J. (dir.), *Vocabulary: Description, acquisition and pedagogy*. Cambridge University Press, 6–19.

Née, É. ; Sitri F., & Veniard, M. (2016). Les routines, une catégorie pour l'analyse de discours : le cas des rapports éducatifs, *Lidil*, 53, 71–93.

Németh, Zs. ; Nagy C., K. & Németh T. E. (2018). Az adatforrások elmosódott határai a konverzációelemzésben. [Les frontières floues des sources d'informations dans l'analyse de conversations.] *Argumentum,* 14, 301–312.

Nesselhauf, N. (2005). *Collocations in a learner corpus*. John Benjamins.

Nini, A. (2019). The Multi-Dimensional Analysis Tagger. In Berber Sardinha, T. & Veirano Pinto M. (dir.), *Multi-dimensional analysis: Research methods and current issues*. Bloomsbury Academic, 67–94.

Oakes, M. P. (1998). *Statistics for corpus linguistics*. Edinburgh University Press.

O'Keeffe, A. (2000). Varieties of spoken English: same difference? Colloquium Paper with McCarthy, M., Koester, A. & Prodromou, L. *34thIATEFL Conference*, Dublin.

O'Keeffe, A. ; McCarthy, M. & Carter, R. (2007). *From corpus to classroom.* Cambridge University Press.

Omidian, T. & Siyanova-Chanturia, A. (2020). Semantic prosody revisited. Implications for language learning. *TESOL Quarterly,* 54(2), 1–13.

Oravecz, Cs. ; Váradi, T. & Sass, B. (2014). The Hungarian Gigaword Corpus. In Calzonari, Nicoletta (dir.). *Proceedings of LREC. European Language Resources Association*. Reykjavik.

O'Sullivan, I. & Chambers, A. (2006). Learners' writing skills in French: Corpus consultation and learner evaluation. *Journal of Second Language Writing,* 15, 49–68.

Pace-Sigge, M. T-L. (2013a). *Lexical priming in spoken English usage*. Palgrave McMillan.

Pace-Sigge, M. T-L. (2013b). The concept of Lexical Priming in the context of language use. *ICAME Journal,* 37, 149–173.

Pace-Sigge, M. T-L. (2018). *Spreading activation, lexical priming and the semantic web*. Palgrave Macmillan.

Pace-Sigge, M. T-L. & Patterson, K. (2017). *Lexical priming. Applications and advances*. John Benjamins.

Page, R. ; Barton, D. ; Unger, J. W. & Zappavigna, M. (2014). *Researching language and social media: A student guide*. Routledge.

Partington, A. (2004). Utterly content in each other's company: Semantic prosody and semantic preference. *International Journal of Corpus Linguistic*s, 9, 131–156.

Partington, A. (2001). Corpus-based description in teaching and learning. In Allison, G. (dir.), *Learning with corpora*. Athelstan, 46–63.

Partington, A. (1998). *Patterns and meanings: Using corpora for English language research and teaching.* John Benjamins.

Pawley, A. & Syder, F. (1983). Two puzzles for linguistic theory: native-like selection and native-like fluency. In Richards, J. & Schmidt, R. (dir.), *Language and communication*. Longman, 191–266.

Pérez-Paredes P. (2021). *Corpus linguistics for education. A guide to research*. Routledge.

Pérez-Paredes, P. & Mark, G. (2021). What can corpora tell us about language learning? In McCarthy, M. & O´Keeffe, A. (dir.) *The Routledge Handbook of Corpus linguistics*. Routledge.

Pérez-Paredes P. ; Mark G. & O'Keeffe, A. (2020). *The impact of usage-based approaches on second language learning and teaching. Cambridge educators research report.* University of Cambridge.

Pérez-Paredes, P. & Sánchez-Tornel, M. (2019). The linguistic dimension of L2 interviews: A multidimensional analysis of native speaker language. *Focus on ELT Journal*, 1(1), 4–26.

Pérez-Paredes, P. & Bedmar, B. D. (2009). Language corpora and the language classroom. *Materiales de formación del profesorado de lengua extranjera*, CARM, 1–48.

Poole, R. (2020). "Corpus can be tricky": revisiting teacher attitudes towards corpus-aided language learning and teaching. *Computer Assisted Language Learning*, 1–22.

Poole, R. (2018). *A guide to using corpora for English language learners*. Edinburgh University Press.

Prodromou, L. (1997). From corpus to octopus. *IATEFL Newsletter*, 137, 18–21.

Pustet, R. (2004). Zipf and his heirs. *Language Sciences,* 26(1), 1–25.

Recski, G. (2014). Hungarian noun phrase extraction using rule-based and hybrid methods. *Acta Cybernetica*, 1(1), 461–479.

Reppen, R. (2016). Designing and building corpora for language learning. In Farr, F. & Murray, L. (dir.). *The Routledge handbook of language learning and technology*. Routledge.

Robb, T. & Susser, B. (1989). Extensive reading versus skill building in an EFL context. *Reading in a Foreign Language* 5(2), 239–251.

Robinson, P. (1995). Attention, memory and the noticing hypothesis. *Language Learning*, 75(2), 283-331.

Robinson, P. & Ellis, N. C. (dir.) (2008). *Handbook of cognitive linguistics and second language acquisition*. Routledge.

Robinson, P., Mackey, A., Gass, S. & Schmidt, R. (2012). Attention and awareness in second language acquisition. In Gass, S. & Mackey, A. (dir.), *The Routledge handbook of second language acquisition*. Routledge.

Rounds, C. (2008). *Hungarian : an essential grammar.* Routledge.

Römer, U. (2011). Corpus research applications in second language teaching. *Annual Review of Applied Linguistics*, 37, 205–225.

Römer, U. (2006). Pedagogical applications of corpora: some reflections on the current scope and a wish list for future developments. *Zeitschrift für Anglistik und Amerikanistik,* 54(2), 121–134.

Römer, U. & Schulze, R. (dir.) (2009). *Exploring the lexis-grammar interface.* John Benjamins.

Rühlemann, C. (2018). *Corpus linguistics for pragmatics*. Routledge.

Rühlemann, C. (2007). *Conversation in context. A corpus-driven approach*. Continuum.

Salazar, D. (2014). *Lexical bundles in native and non-native scientific writing*. John Benjamins.

Sánchez-Cárdenas, B. (2010). Les restrictions sémantiques des arguments verbaux : une question de fréquence d'usage. *Synergies France,* 6, 41–50.

Schaeffer-Lacroix, E. (2012). Qu'est-ce qui rend les corpus 'pédagogiques' ? *Procedia - Social and Behavioral Sciences,* 34, 198–201.

Schirm, A. (2014). A diskurzusjelölők stilisztikai és pragmatikai megközelítése. [Les marqueurs du discours : Approches stylistiques et pragramtiques] In Dobi, E ; Domonkosi Á. & Pethő, J. (dir.), *Nyelvről, stílusról – sokszínűen.* Université de Debrecen, 294–308.

Schmid, H-J. (2016). *Language and the human lifespan*. De Gruyter Mouton.

Schmidt, R. (1990). The role of consciousness in second language learning. *Applied Linguistics*, 11, 129–158.

Schmidt, R. (2010). Attention, awareness, and individual differences in language learning. In Chan, W. M. ; Chi, S. ; Cin, K. N. ; Istanto, J. ; Nagami, M. ; Sew, J. W. ; Suthiwan, T. & Walker, I. (dir.) *Proceedings of CLaSIC 2010*. National University of Singapore, Centre for Language Studies, 721–737.

Schmitt, N. (2004). *Formulaic sequences: Acquisition, processing and use*. John Benjamins.

Scott, M. & Thompson, G. (2000). *Patterns of text.* John Benjamins.

Scott, M. & Tribble, C. (2006). *Textual patterns – Key words and corpus analysis in language education*. John Benjamins.

Siepmann, D. (2015). L'élaboration d'une grammaire pédagogique à partir de corpus : l'exemple du subjonctif. In Tinnefeld, T. (dir.), *Grammatographie und didaktische Grammatik - gestern, heute, morgen. Gedenkschrift für Hartmut Kleineidam anlässlich seines 75. Geburtstages.* Htw saar.

Sinclair, J. (2004a). Meaning in the framework of corpus linguistics. *Lexicographica* 20, 20–32.

Sinclair, J. (2004b). *Trust the text*. Routledge.

Sinclair, J. (2003). *Reading concordances. An introduction*. Pearson/Longman.

Sinclair, J. (2001). *Collins COBUILD English Language Dictionary*. Collins.

Sinclair, J. (2000). Lexical grammar. *Naujoji Metodologija*, 24, 191–203.

Sinclair, J. (1997). Corpus evidence in language description. In Wichmann, A., Fligelstone, S., McEnery, T. & Knowles, G. (dir.), *Teaching and language corpora*. Longman, 27–39.

Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford University Press.

Sinclair, J. (1984). Naturalness in language. *Ilha do desterro. A Journal of English language, literature in English and cultural studies*, 5(11), 45–55.

Siyanova-Chanturia, A. & Martinez, R. (2014). The Idiom Principle revisited. *Applied Linguistics,* 36(5), 549–569.

Siyanova-Chanturia, A. & Spina, S. (2015). Investigation of native speaker and second language learner intuition of collocation frequency. *Language Learning*, 65, 533–562.

Spina, S. & E. Tanganelli (2012). Les collocations comme indice pour distinguer les genres textuels. *Corpus,* 11, 73–89.

Siyanova-Chanturia, A. ; Conklin, K. & van Heuven, W. J. B. (2011). Seeing a phrase "time and again" matters: The role of phrasal frequency in the processing of multiword sequences. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37*(3), 776–784.

Spöttl, C. & McCarthy, M. J. (2003). Formulaic utterances in the multilingual context. In Cenoz, J., Jessner, U. & Hufeisen, B. (dir.), *The multilingual lexicon*. Kluwer, 133–151.

Staples, S. (2015). Spoken discourse. In Biber, D. & Reppen, R. (dir.), *The Cambridge handbook of English corpus linguistics*. Cambridge University Press, 271–292.

Stefanowitsch, A. (2020). *Corpus linguistics. A guide to the methodology*. Language Science Press.

Stubbs, M. (2009). The search for units of meaning: Sinclair on empirical semantics. *Applied Linguistics,* 30(1), 115–137.

Stubbs, M. (2001). *Words and phrases: Corpus studies of lexical semantics*. Blackwell.

Stubbs, M. (1996). *Text and corpus analysis: Computer-assisted studies of language and culture*. Wiley.

Szende, T. & Kassai, G. (2001). *Grammaire fondamentale du hongrois*. Langues et mondes, L'Asiathèque.

Szili, K. (2001). A perfektivitás mibenlétéről a magyar nyelvben a *meg-* igekötő funkciói kapcsán. [Les fonctions du préfixe *meg- :* la perfectivité.] *Magyar Nyelvőr,* 97, 263–282.

Szili, K. (1999). *Hogyan tanítsuk? I. Az igekötők. [Comment enseigner les préfixes ?]* Dolgozatok a magyar mint idegen nyelv és a hungarológia köréből, 38. Eötvös Lóránd University.

Szirmai, M. (2005). *Bevezetés a korpusznyelvészetbe. A korpusznyelvészet alkalmazása az anyanyelv és az idegen nyelv tanulásában és tanításában*. [*Introduction à la linguistique de corpus. Appliquer la linguistique de corpus pour l'apprentissage et pour l'enseignement de la langue maternelle et des langues étrangères.*] Tinta Könyvkiadó.

Szita, S. (2022a, à paraître). Overwhelming, time-consuming, user-unfriendly… and now what? Teacher training for the successful implementation of a corpus-informed methodology. In Curry, N. ; Tyne, H. ; Bilger, M. ; Buscail, L ; Leray, M. & Pérez-Sabater, C. (dir.), *À la découverte de la langue: apprentissages et affordances/Discovering language: Learning and affordance/Descubrir la lengua: aprendizaje y oportunidades*. Peter Lang.

Szita, S. (2022b, à paraître). A MagyarOK nyílt korpusz használatáról. [L'utilisation du corpus ouvert de « MagyarOK »]. *Hungarológiai Évkönyv,* 21(1–2).

Szita, S. (2021). Au-delà du glossaire. Les mots difficiles en contexte. In Berk, S. (dir.), *Dictionnaire et apprentissage des langues*. Éditions des archives contemporaines.

Szita, S. (2020). Korpuszépítés és korpuszhasználat alacsonyabb nyelvtudási szinteken. [Construction et utilisation des corpus aux niveaux de compétences inférieurs]. *Hungarológiai Évkönyv,* 19(1–2), 173–179.

Szita, S. (2014). Invent content, not language. Meaningful interaction and natural language use in the classroom. In Hegedűs, R. & Görbe, T. (dir.), *Kleine Sprachen, was nun? Studies on language and culture in Central and Eastern Europe.* Kubon und Sagner Verlag, 112–127.

Szita, S. & Pelcz, K. (2017). Modellalapú nyelvoktatás, természetes nyelvhasználat [Enseignement fondé sur des modèles et usage langagier naturel]. *Journal of Teaching Hungarian as a second language and culture,* 1(2), 262–269.

Szita, S. & Pelcz, K. (2013–2019). *MagyarOK A1–B2.* Université de Pécs.

Szita, S. & Pelcz, K. (2023 à paraître). *A modellalapú nyelvtanítás kézikönyve.* [*Manuel de l'Apprentissage de langues fondé sur des modèles*]. Insitute of Model-based Language Learning.

Szita, S. & Görbe, T. (2009). *Gyakorló magyar nyelvtan. [Grammaire pratique du hongrois]*. Akadémiai Kiadó.

Szudarski, P. (2017). *Corpus linguistics for vocabulary*. Routledge.

Taylor, D. (2012). *The mental corpus*. Oxford University Press.

Teubert, W. (2010). *Meaning, discourse and society*. Cambridge University Press.

Thompson, P. (2004). Spoken language corpora. In Wynne, M. (dir.), *Developing linguistic corpora. A guide to good practice*. Oxbow Books for the Arts and Humanities Data Service.

Thornbury, S. (1996). Teachers research teacher talk. *ELT Journal,* 50(4), 279–289.

Timmis, I. (2015). *Corpus linguistics for ELT. Research and practice.* Routledge.

Tognini-Bonelli, E. (2010). Theoretical overview of the evolution of corpus linguistics.
In O'Keeffe, A. & McCarthy, M. (dir.). *The Routledge handbook of corpus linguistics*. Routledge, 14–29.

Tognini-Bonelli, E. (2004). Working with corpora. In Coffin, C. ; Hewings A. & O'Halloran K. (dir.), *Applying English Grammar*. London Arnold, 11–24.

Tomasello, M. (2003). *Constructing a language. A usage-based theory of language acquisition.* Harvard University Press.

Tribble, C. (2015). Teaching and language corpora. Perspectives from a personal journey. In Boulton, A. & Leńko-Szymańska, A. (dir.), *Multiple Affordances of Language Corpora.* John Benjamins, 37–62.

Tribble, C. (2001). *Small corpora and teaching writing*. John Benjamins.

Tribble, C. & Jones, G. (1997). *Concordances in the classroom: A resource guide for teachers.* Athelstan.

Trofimovich, P. & McDonough, K. (dir.) (2011). *Applying priming methods to L2 learning, teaching and research Insights from Psycholinguistics*. John Benjamins.

Tyler, A. (2010). Usage-based approaches to language and their applications to second language learning. *Annual Review of Applied Linguistics,* 30, 270–291.

Tyne, H. (2012). Corpus work with ordinary teachers: Data-driven learning activities. In Thomas, J. & Boulton, A. (dir.), *Input, process and product: Developments in teaching and language corpora.* Masaryk University Press, 114–129.

Váradi, T. (2002). The Hungarian National Corpus. *Proceedings of the Third International Conference on Language Resources and Evaluation*, 5., 385–389.

Walsh, S. (2010). What features of spoken and written corpora can be exploited in creating language teaching materials and syllabuses? In O'Keeffe, A. & McCarthy, M. (dir.), *The Routledge handbook of corpus linguistics*. Routledge, 333–344.

Walsh, S. (2002). Construction or obstruction: Teacher talk and learner involvement in the EFL classroom. *Language Teaching Research,* 6(1), 3–23.

Warren, M. (2006). *Features of naturalness in conversation*. John Benjamins.

Waseda, M. (2017). Mi a különbség a „Megjött a tavasz" és „Eljött a tavasz" között? [Quelle est la différence entre „Megjött a tavasz" et „Eljött a tavasz" (Le printemps est arrivé.) ?] *Hungarológiai Évkönyv* 2017(1), 94–99.

Watzlawik, P ; Beavin Baàvelas, J. & Jackson, D. (2011). *Pragmatics of human communication: A study of interactional patterns, pathologies and paradoxes*. W. W. Norton & Company.

Webb, S. (2005). Receptive and productive vocabulary learning: The effects of reading and writing on word knowledge. *Studies in Second Language Acquisition*, 27(1), 33–52.

Webb, S. (2007). The effects of repetition on vocabulary knowledge. *Applied Linguistics,* 28(1), 46–65.

Webb, S. & Chang, A. (2014). Second language vocabulary learning through extensive reading with audio support: How do frequency and distribution of occurrence affect learning? *Language Teaching Research,* 18(1), 1–20.

Weber, J. J. (2001). A concordance- and genre-informed approach to ESP essay writing. *ELT Journal,* 55(1), 14–20.

Webster, J. & Kit, C. (1992). Tokenisation as the initial phase in NLP. *Computer Science.* Actes de Coling, Nantes, 1106–1110.

Widdowson, H. (2003). *Defining issues in English language teaching.* Oxford University Press.

Widdowson, H. (1998). Context, community and authentic language. *TESOL Quarterly*, 32(4), 705–716.

Widdowson, H. (1978). *Teaching language as communication*. Oxford University Press.

Willis, D. (2003). *Rules, patterns and words: grammar and lexis in ELT*. Oxford University Press.

Wray, A. (2008). *Formulaic language. Pushing the boundaries.* Oxford University Press.

Wray, A. (2007). Set phrases in second language acquisition. In Burger et al. (dir.), *Phraseologie*: *Ein internationales Handbuch zeitgenössischer Forschung*. Mouton de Gruyter (1), 870–881.

Wray, A. (2002). *Formulaic language and the lexicon*. Cambridge University Press.

Wray, A. (2000). Formulaic sequences in second language teaching: principle and practice. *Applied Linguistics*, 21(4), 463–489.

Wynne, M. (dir.). (2004). *Developing linguistic corpora. A guide to good practice*. Oxbow Books for the Arts and Humanities Data Service.

Xiao, R. (2015). Collocation. In Biber, D. & Reppen, R. (dir.), *The Cambridge handbook of English corpus linguistics*. Cambridge University Press, 106–125.

Xiao, R. & McEnery, T. (2006). Collocation, semantic prosody, and near synonymy: A cross-linguistic perspective. *Applied Linguistics*, 27(1), 103–129.

Yan, R. ; Tutin, A. & Tran, T. T. H. (2018). Routines verbales pour les français langue étrangère : des corpus d'experts aux corpus d'apprenants. *Lidil,* 58, 1–19.

Zipf, G. K. (dir.) (1965). *Human behavior and the principle of least effort*. MIT Press.

Zyzik, E. (2009). The role of input revisited: Nativist versus usage-based models. *L2 Journal* (1)1, 42–61.

Corpus et logiciels

Corpus pédagogiques

Les corpus Backbone et Sacodeyl : http://projects.ael.uni-tuebingen.de/backbone/moodle/

Le corpus ouvert de MagyarOK sur Sketch Engine :
https://app.sketchengine.eu/#dashboard?corpname=preloaded%2Fmagyarok

ELISA : http://universal.elra.info/product_info.php?cPath=25&products_id=1835

FLEURON : https://fleuron.atilf.fr/contact.php?lg=fr

Corpus linguistiques

British National Corpus : http://www.natcorp.ox.ac.uk/corpus/index.xml

COBUILD Corpus : https://collins.co.uk/pages/elt-cobuild-reference-the-collins-corpus

Collins Teaching grammar patterns : https://grammar.collinsdictionary.com/grammar-pattern/teaching-resources

Corpora Collection, Université de Leipzig : http://corpora.uni-leipzig.de/

Corpus of Contemporary American English : https://corpus.byu.edu/coca/

Corpus national du hongrois (Magyar Nemzeti Szövegtár) : http://corpus.nytud.hu/mnsz/

COSMAS : https://www2.ids-mannheim.de/cosmas2/

Dictionnaire Merriam-Webster : https://www.merriam-webster.com

Digitales Wörterbuch der deutschen Sprache: https://www.dwds.de

Ortolang : repository.ortolang.fr

Pattern Grammar : http://arts-ccr-002.bham.ac.uk/ccr/patgram/

SketchEngine : http://sketchengine.eu

Touchstone :
https://www.cambridge.org/gb/cambridgeenglish/catalog/adult-courses/touchstone

Index

A

B

C

D

E

F

G

H

I

L

M

N

O

P

R

S

T

U

V

W

Z

Cette thèse s'inscrit dans un cadre interdisciplinaire, au croisement de la linguistique appliquée – plus précisément, de la linguistique de corpus – et de la didactique de l'enseignement des langues étrangères. Au cœur de ce travail se situent les considérations concernant la création et l'analyse de corpus dans le cadre pédagogique. Notre premier but est de démontrer une interconnexion étroite entre lexique et grammaire dans le cas des langues morphologiquement complexes et le potentiel d'une approche lexico-grammaticale pour l'enseignement de ces langues. Les démonstrations seront effectuées en utilisant le hongrois comme exemple. Nous argumenterons également que pour parvenir aux besoins des apprenants, il est nécessaire de créer de nouveaux corpus pédagogiques selon des principes particuliers.

La Partie I est dédiée au recensement de la littérature sur les avancées de la linguistique de corpus. Elle présente les méthodes de ce domaine et ses résultats pertinents ainsi que les principes de base de création de corpus et d'ouvrages pédagogiques se fondant sur une « approche corpus ». La Partie II traite de l'exploration des corpus non pédagogiques au service de l'enseignement du hongrois. À cette fin, deux aspects lexicaux et deux aspects grammaticaux sont analysés en détail. La Partie III concerne la création et l'utilisation des corpus pédagogiques pour l'observation et la pratique langagières au cours de l'apprentissage. Nous explorerons leur potentiel et leurs limites à travers l'exemple concret des corpus compilés pour la série de manuels « MagyarOK ». Les considérations autour de l'équilibre entre authenticité et accessibilité seront également abordées.

Mots-clés : linguistique appliquée ; linguistique de corpus ; nouvelles technologies ; corpus pédagogiques ; didactique des langues ; data-driven learning ; apprentissage sur corpus ; langue hongroise

The nature of this doctoral thesis is interdisciplinary: its research topic is situated at the crossroads of applied corpus linguistics and the didactics of foreign language teaching. The thesis explores the possibilities and potentials of the creation and analysis of corpora in the pedagogical context. We aim to demonstrate a close interconnection between lexicon and grammar in the case of morphologically complex languages on the one hand, and the potential of a lexico-grammatical approach for the teaching of these languages on the other. Demonstrations will be made using Hungarian as an example. We will also argue that in order to meet the needs of learners (especially at lower proficiency levels), it is necessary to create new pedagogical corpora based on specific principles.

The thesis is divided into three parts. Part I is dedicated to a review of the literature on advances in corpus linguistics. It presents the methods of this field and its relevant results, as well as the basic principles of creating corpora and corpus-informed teaching materials. Part II deals with the exploration of non-pedagogical corpora in the service of Hungarian teaching. To this end, two lexical and two grammatical aspects are analysed in detail. Part III concerns the creation and use of pedagogic corpora for linguistic observation and practice in the classroom. We explore their potential and limitations through the concrete example of the corpora compiled for the *MagyarOK* textbook series. Issues related to the balance between authenticity and accessibility will also be addressed.

Keywords: applied linguistics; corpus linguistics; new technologies; educational corpora; language didactics; data-driven learning; Hungarian language