Utilisation de méthodes linguistiques pour la détection et la correction automatisées d'erreurs produites par des francophones écrivant en anglais

Marie Garnier

THÈSE

En vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par :
Université Toulouse 2 Le Mirail (UT2 Le Mirail)

Cotutelle internationale avec :

Présentée et soutenue par :
Marie GARNIER

Le 19 septembre 2014

Titre :

Utilisation de méthodes linguistiques pour la détection et la correction automatisées d'erreurs produites par des francophones écrivant en anglais

---

ED ALLPH@ : Anglais

Unité de recherche :
Laboratoire Cultures Anglo-Saxonnes, EA 801

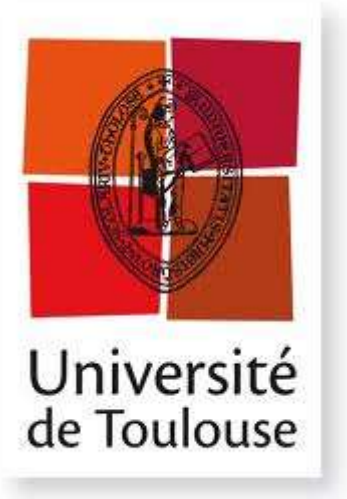
Directeur(s) de Thèse :

M. Dennis PHILPS, Professeur, CAS, Université Toulouse 2
M. Patrick SAINT-DIZIER, Directeur de recherches, IRIT, Université Toulouse 3

Rapporteurs :
Mme Alda MARI, Directrice de recherches, IJN, École Nationale Supérieure, Paris
M. Jean-Marie MERLE, Professeur, BCL, Université de Nice-Sophia Antipolis

Autre(s) membre(s) du jury :

Mme Blandine PENNEC, Maître de Conférences, CAS, Université Toulouse 2
Mme Cornelia TSCHICHOLD, Senior Lecturer, Swansea University



THÈSE en vue de l'obtention du
DOCTORAT DE L'UNIVERSITE DE TOULOUSE
Délivré par l'Université Toulouse 2
École doctorale : ALLPH@
Spécialité : Anglais

Présentée et soutenue par **Marie GARNIER**
Le 19 septembre 2014

Utilisation de méthodes linguistiques pour la détection et la correction automatisées d'erreurs produites par des francophones écrivant en anglais

MEMBRES DU JURY

Mme Alda MARI	Directrice de recherches (Linguistique), École Nationale Supérieure, Paris (Rapporteur)
M. Jean-Marie MERLE	Professeur (Linguistique), Université de Nice-Sophia Antipolis (Rapporteur)
Mme Blandine PENNEC	Maître de conférences (Linguistique), Université Toulouse 2
Mme Cornelia TSCHICHOLD	Senior Lecturer (Linguistique appliquée), Swansea University
M. Dennis PHILPS	Professeur (Linguistique), Université Toulouse 2 (Directeur de thèse)
M. Patrick SAINT-DIZIER	Directeur de recherches (Informatique), Université Toulouse 3 (Directeur de thèse)

À mes grands-parents

Résumé

Le point de départ de cette recherche est le constat des difficultés persistantes rencontrées par les francophones de niveau intermédiaire à avancé lors de la production de textes en anglais, dans des contextes personnels ou professionnels. Les premiers outils utilisés pour remédier à ces erreurs, les correcteurs grammaticaux automatiques, ne prennent pas en compte de nombreuses erreurs produites par les francophones utilisant l'anglais, notamment car ces correcteurs sont rarement adaptés à un public ayant l'anglais comme L2. Nous proposons d'identifier précisément les difficultés rencontrées par ce public cible à partir du relevé des erreurs dans un corpus adapté, et d'élaborer une modélisation linguistique des erreurs et des corrections à apporter. Cette modélisation est fondée sur une analyse linguistique approfondie des phénomènes concernés, à partir d'indications grammaticales, d'études de corpus, et de l'analyse des segments erronés. La validité de l'utilisation de méthodes linguistiques est établie par l'implémentation informatique des règles de détection et de correction, suivie de l'évaluation des résultats de l'application de ces règles sur des corpus d'anglais L1 et L2.

Les deux types d'erreur concernés sont le placement des adverbes, en particulier les adverbes de manière et l'adverbe *also*, et l'utilisation de structures N+N. Ces deux phénomènes ont la particularité d'interroger les normes de grammaticalité et d'acceptabilité. Ces types sont sélectionnés par le biais de l'application de la méthodologie de l'analyse des erreurs à notre recherche. Un corpus d'anglais L2 est constitué à partir de productions représentatives de l'utilisation de l'anglais par les francophones. Les segments erronés y sont relevés manuellement, puis sont classés selon un système reposant sur des catégories linguistiques. Les erreurs de placement des adverbes et d'utilisation des structures N+N sont deux des six types les plus fréquemment relevés dans ce corpus.

La détection et la correction automatisées reposent sur la modélisation linguistique des schémas d'erreurs et de correction. Cette modélisation aboutit à la création de 11 schémas au total pour les adverbes, et de cinq schémas pour les erreurs N+N, certains schémas incluant plusieurs propositions de correction.

Les règles de détection et de correction utilisent des patrons de détection associés à des instructions de réécriture, et sont implémentées en Dislog dans la plateforme <TextCoop>, un analyseur de discours programmé en Prolog. Les règles sont évaluées sur des corpus d'anglais L1 et L2. Pour le corpus d'anglais L2, les taux de précision et de rappel atteignent 100 % et 67 % pour *also*, et 95 % et 91 % pour les adverbes de manière. L'évaluation des règles de correction des erreurs N+N sur le corpus d'anglais L1 génère un nombre de faux positifs important. De plus, pour être détectées et corrigées de manière appropriée, ces erreurs requièrent des recherches approfondies en sémantique lexicale. Ces deux facteurs repoussent leur évaluation à une période ultérieure.

Le traitement des erreurs dans cette recherche inclut également l'élaboration de messages correctifs, dans l'objectif de permettre la prise d'autonomie des personnes utilisatrices du système. Le canevas proposé repose sur cinq étapes modulables en fonction du public cible et du contexte d'utilisation. Ces étapes incluent le marquage de l'erreur, le diagnostic d'erreur, la rétroaction métalinguistique, les instructions de remédiation et l'illustration.

Mots-clés : linguistique de l'anglais, correction grammaticale automatisée, adverbes anglais, structures N+N, acquisition des langues secondes, enseignement des langues assisté par ordinateur, traitement automatisé des langues

Abstract

The starting point of this research is the observation that French speakers writing in English in personal or professional contexts still encounter grammatical difficulties, even at intermediate to advanced levels. The first tools they can reach for to correct those errors, automatic grammar checkers, do not offer corrections for a large number of the errors produced by French-speaking users of English, especially because those tools are rarely designed for L2 users. We propose to identify the difficulties encountered by these speakers through the detection of errors in a representative corpus, and to create a linguistic model of errors and corrections. The model is the result of the thorough linguistic analysis of the phenomena at stake, based on grammatical information available in reference grammars, corpus studies, and the analysis of erroneous segments. The validity of the use of linguistic methods is established through the implementation of detection and correction rules in a functional platform, followed by the evaluation of the results of the application of those rules on L1 and L2 English corpora.

The two error types this research focuses on are adverb placement, especially for manner adverbs and the adverb *also*, and the use of N+N structures. These phenomena both question the factors leading to grammaticality and acceptability judgments. Error analysis is used to select those error types. A corpus of L2 English is compiled using productions that are representative of the use of English by French speakers. Erroneous segments are detected manually and classified using a system based on linguistic categories. Adverb placement errors and errors in the use of N+N structures are two of the six most frequent types in our corpus.

The automatic detection and correction of errors are based on the linguistic modeling of error and correction schemas. The use of this method results in the creation of 11 schemas in total for adverbs, and five schemas for N+N structures. Some schemas include several correction propositions.

Detection and correction rules rely on detection patterns and rewriting instructions, and are implemented in Dislog in the <TextCoop> platform, a Prolog-based discourse analyzer. Rules are evaluated on L1 English and L2 English corpora. For the L2 English corpus, precision and recall rates reach 100 % and 67 % for *also*, and 95 % and 91 % for manner adverbs. The evaluation shows that the rules for N+N errors generate a high number of false positives. Additionally, these errors require thorough research in lexical semantics to be adequately detected and corrected. These two factors result in their evaluation being postponed to a future stage in this research.

The processing of errors in this research also includes the creation of corrective feedback messages, with the objective of allowing system users to improve their autonomy. We propose a five-step plan, which can be adapted to the needs of the user and the context of use. The plan includes error marking, error diagnosis, metalinguistic feedback, directions for remediation, and illustrations.

Keywords : English linguistics, automatic grammar checking, English adverbs, N+N structures, second language acquisition, computer-assisted language learning, natural language processing

Table des matières

Chapitre 3

Liste des tableaux

Liste des abréviations

ACG	*A Comprehensive Grammar of the English Language*, Quirk et al.
Adj	Adjectif
Adv	Adverbe
AdvM	Adverbe de manière
AE	Analyse des erreurs
ALF	Anglais comme *lingua franca*
ALI	Anglais comme langue internationale
ALM	Anglais comme langue maternelle
ALS	Acquisition des langues secondes
Aux	Auxiliaire
CECL	*Center for English Corpus Linguistics*
CGE	*The Cambridge Grammar of the English Language*, Huddleston et Pullum et al.
CGA	Correction grammaticale automatisée
Det	Déterminant
DP	Détachement prosodique
EFL	*English as a foreign language*
ELAO	Enseignement des langues assisté par ordinateur
ELF	*English as a lingua franca*
ELFA	*English as a Lingua Franca in Academic Settings*
eM	Placement Médian-final
GAdj	Groupe adjectival
GAdv	Groupe adverbial
GDét	Groupe déterminant
GN	Groupe nominal
GP	Groupe prépositionnel
GV	Groupe verbal
ICICLE	*Interactive Computer Identification and Correction of Language Errors*
ICLE	*International Corpus of Learner English*
iE	Placement Final-initial
iM	Placement Médian-initial
L1	Langue maternelle
L2	Langue seconde/étrangère
LGSWE	*Longman Grammar of Spoken and Written English*, Biber et al.
LN	Personne locutrice native
LNN	Personne locutrice non-native
LONGDALE	*Longitudinal Database of Learner English*
LOCNESS	*Louvain Corpus of Native English Essays*
mM	Placement Médian-médian
NICT JLE	*National Institute of Information and Communications Technology Japanese Learner of English Corpus*
NLP	*Natural Language Processing*
Npr	Nom propre
Ns	Nom pluriel/avec morphème *-s*
Prep	Préposition
TAL	Traitement automatique des langues
VESPA	*The Varieties of English for Specific Purposes Database*
Ving	Forme verbale non-finie en *-ing*
Vlex	Verbe lexical
VOICE	*Vienna-Oxford International Corpus of English*
XML	*Extensible mark-up language*

Remerciements

Je tiens tout d'abord à remercier chaleureusement mes directeurs de thèse, Dennis Philps et Patrick Saint-Dizier, pour leurs conseils avisés, leur regard critique et bienveillant sur mes travaux, et leur soutien scientifique et moral pendant toute la durée de ma thèse et même bien avant son commencement.

Je remercie également Alda Mari, Jean-Marie Merle, Blandine Pennec et Cornelia Tschichold de m'avoir fait l'honneur d'accepter d'être membres du jury de soutenance de cette thèse.

Mes remerciements vont aussi à l'école doctorale ALLPH@, et plus particulièrement à ses directrices et directrices adjointes successives, Monique Martinez, Nathalie Dessens et Françoise Knopper, qui m'ont permis de commencer et de poursuivre mes recherches dans les meilleures conditions possibles.

J'aimerais remercier tous les membres du laboratoire Cultures Anglo-Saxonnes, et en particulier les personnes qui se sont succédé à sa tête pendant la durée de ma thèse, Catherine Lanone, Wendy Harding, Philippe Birgy et Anne Stéfani, pour leur accueil, leur disponibilité et leur soutien. Merci aux membres de l'axe linguistique du laboratoire, pour leurs encouragements et leurs conseils : Henri Le Prieult, Andrew McMichael, Blandine Pennec, et en particulier Linda Terrier pour le coup de pouce nécessaire des derniers mois. Un grand merci également aux doctorants et doctorantes du CAS, et en particulier à Sarah Bourse, Adèle Cassigneul et Candice Lemaire. Merci à Hanane Serjaouan, responsable administrative de l'équipe, pour son aide toujours patiente et efficace.

Mes remerciements vont également au personnel enseignant et administratif du Département Etudes du Monde Anglophone de l'Université Toulouse 2, et tout particulièrement aux membres de l'équipe LANSAD pour leur accueil chaleureux, leur professionnalisme, et leur compréhension lors des derniers mois de rédaction.

Un grand merci aux personnes ayant généreusement accepté d'effectuer des tâches dans le cadre de ma thèse : Camille Albert, Anna Cumbie, Erika Hennings, Mathilde Janier, Caitlin Smith.

Merci à mes parents, Josiane Villa et Jacques Garnier, pour leur soutien inconditionnel.

Merci à Nora Moujane, pour son amitié précieuse et sa confiance totale en mes capacités.

Merci à Bruno Bastiani, pour tout.

Introduction générale

Exposé de la problématique

Qu'on s'en réjouisse ou qu'on le déplore, l'anglais est à ce jour la langue véhiculaire la plus utilisée dans le monde, ainsi que la langue seconde la plus souvent enseignée (Crystal, 2003 : 106-109). Un certain niveau de maîtrise de cette langue est devenu une nécessité pour de nombreuses activités professionnelles et non professionnelles. L'accession rapide de l'anglais à ce statut de langue de communication internationale, la diversité des situations personnelles, ainsi qu'un certain contexte culturel français ont pour conséquence l'existence de nombreux groupes d'adultes ayant une connaissance imparfaite de l'anglais. Si l'anglais oral reste une difficulté importante pour les francophones, la communication orale permet au moins une certaine négociation du sens, moins évidente lorsqu'il s'agit de communication écrite. La production de documents écrits est par ailleurs une tâche redevenue incontournable au quotidien avec l'utilisation d'internet. Les personnes ayant une maîtrise imparfaite de l'anglais peuvent donc être confrontées à l'obligation de communiquer en anglais écrit (ex. : courriels privés, professionnels et semi-professionnels, communication internationale de travaux, communications personnelles) sans pour autant pouvoir bénéficier facilement de corrections de leurs travaux. Ceci a pour résultat la production de textes comprenant de nombreuses erreurs qui peuvent, d'une part, gêner leur compréhension par des anglophones ou des non-anglophones ne partageant pas la même langue maternelle, et, d'autre part, poser des problèmes de crédibilité, de professionnalisme et d'image.

Il existe bien sûr déjà des outils utilisables en autonomie, allant de sites internet gratuits proposant des aides ponctuelles (ex. : *Anglais Facile*) à des programmes complets d'apprentissage de l'anglais sur internet (ex. : *Cyberteachers*) ou sous forme de logiciel (ex. : *Rosetta Stone*). Cependant, l'aide linguistique à laquelle toute personne utilisatrice d'ordinateurs est confrontée, parfois même involontairement, est la correction automatisée. L'immense majorité des traitements de texte et des systèmes de courriel intègrent désormais des correcteurs d'orthographe et de grammaire, et les proposent par défaut au public utilisateur. Malgré le fait qu'ils soient utilisées simultanément et qu'ils présentent des similitudes de surface, les correcteur d'orthographe et les correcteurs grammaticaux sont en réalité des outils très différents. Nous n'abordons ici que les correcteurs grammaticaux, étant entendu que les erreurs morphosyntaxiques, par exemple les erreurs d'accord, sont des erreurs de grammaire et non d'orthographe.

Si ces outils permettent de corriger de nombreuses erreurs à moindre coût, c'est-à-dire sans passer par des correctrices et correcteurs humains, ils n'atteignent pas la perfection et laissent

encore passer un grand nombre d'erreurs. Notons en passant que la correction humaine est également loin d'être infaillible, comme le montrent les résultats de l'étude comparative de Nadasdi et Sinclair (2007). Les erreurs qui constituent des angles morts pour les correcteurs grammaticaux automatiques sont souvent des erreurs syntaxiques complexes ou des erreurs de sélection lexicale.

L'objectif général de notre recherche est donc d'élaborer un système pour la détection et la correction automatisées d'erreurs qui ne sont pas encore traitées par des correcteurs automatiques, mais qui représentent une difficulté pour notre public cible, c'est-à-dire les francophones utilisant l'anglais à un niveau intermédiaire à avancé. De cet objectif de départ découlent trois principales problématiques de recherche pour lesquelles nous proposons des réponses au fil de ce document.

La première d'entre elles concerne la sélection des erreurs à traiter, et implique d'apporter des réponses à au moins trois questions. Tout d'abord, quelles sont les erreurs qui sont produites par le public auquel les corrections s'adressent ? Parmi ces erreurs, lesquelles ne sont pas encore prises en charge par les correcteurs grammaticaux existants, et plus largement, quels sont les phénomènes linguistiques et les problématiques liées à la correction qui sont peu abordées dans ce domaine ? Par ailleurs, quels sont les types d'erreur les plus pertinents à traiter dans le cadre de notre recherche, dans le respect de nos exigences scientifiques et pratiques ?

L'élaboration technique de règles pour le traitement automatique des ces erreurs constitue une deuxième problématique importante de la présente recherche. Nous postulons que les problèmes linguistiques doivent être abordés d'un point de vue linguistique ; l'élaboration des règles de correction automatique s'attache donc explicitement à utiliser des méthodes linguistiques. Cette approche conditionne l'ensemble de notre recherche, et elle a donc des conséquences sur le choix des techniques utilisées pour la création des règles. Quelles sont les techniques de correction automatisée qui reposent sur l'observation linguistique des phénomènes ? Quelles sont les limites et les points forts de ces méthodes ? Dans quelle mesure ces méthodes sont-elles adéquates pour le traitement des erreurs sélectionnées pour notre recherche, et quels sont les ajustements qui devraient leur être apportés ? Les résultats obtenus par l'utilisation de ces méthodes sont-ils à la hauteur des défis engendrés par le choix d'un public cible de personnes apprenantes ? La réponse à ces questions s'accompagne d'un questionnement sur les ressources à rassembler ou à créer, en lien avec la méthode sélectionnée.

Les deux problématiques mentionnées ci-dessus ont été identifiées en amont de l'amorce de la recherche. La troisième problématique à laquelle nous tâcherons de répondre a émergé au cours de la recherche, en particulier à la suite de la sélection des erreurs à traiter. Ces erreurs concernent globalement deux phénomènes, le placement des adverbes dans le groupe verbal et l'utilisation des séquences N+N, qui ont en commun le fait d'être régis par des règles complexes, soumises à des facteurs très variés, d'ordre syntaxique, sémantique et lexical. Elles s'opposent ainsi à d'autres types d'erreur aux contours grammaticaux beaucoup mieux définis, comme peuvent l'être les erreurs morphosyntaxiques par exemple. Les descriptions existantes du fonctionnement et de l'utilisation de ces structures reflètent cette complexité et offrent peu de jugements tranchés. Face à ces incertitudes et à la pluralité des variables à prendre en compte, comment faire pour offrir des solutions de correction automatisées, sachant que ce domaine s'accommode mieux de certitudes ? Des difficultés émergent dès l'identification des erreurs, juger de la grammaticalité ou de l'acceptabilité des segments s'apparentant à un exercice d'équilibriste. On se demande d'ailleurs légitimement jusqu'à quel point l'axe grammatical/agrammatical est pertinent pour ces phénomènes. La question des références grammaticales s'impose : les sources habituelles sont-elles suffisantes ? Par quoi peut-on avantageusement les compléter ? Proposer des corrections est une tâche tout aussi épineuse : comment corriger ces segments sans tomber dans la prescription ?

Ces questions de recherche soulignent le caractère résolument pluridisciplinaire des travaux présentés ici. Nous tâchons d'aborder cette recherche en nous appuyant sur tous les domaines pertinents de la linguistique et de la linguistique appliquée susceptibles de l'enrichir. Le traitement automatisé des erreurs y est abordé en priorité d'un point de vue linguistique, en s'efforçant de concilier précision linguistique et pertinence scientifique des méthodes de traitement automatique des langues utilisées. Nous souhaitons également que nos choix méthodologiques soient informés par les résultats de recherche du domaine de l'acquisition des langues secondes, qui a des contributions précieuses à apporter au domaine de la correction grammaticale automatique, bien que ces contributions soient, selon nous, trop peu souvent utilisées. Cette ambition pluridisciplinaire implique de consacrer des parties importantes de ce document à la présentation des travaux existants dans les domaines concernés.

Délimitation du cadre pratique

Nous abordons ici la question des choix et exigences pratiques et scientifiques qui composent le cadre de notre recherche. La première exigence à laquelle nous devons répondre

est celle de l'utilité et de la pertinence. Nos règles de correction doivent pallier un manque existant du point de vue des correcteurs grammaticaux déjà disponibles, ainsi que répondre à une véritable difficulté de communication en anglais du point de vue du public utilisateur. Nous devons aussi prendre en compte le statut particulier de l'anglais dans le monde, et identifier les conséquences que son statut de langue internationale peut avoir dans nos travaux. Le mode de sélection des erreurs est présenté dans le Chapitre 1. Les règles de détection/correction doivent pouvoir être modifiées, améliorées et maintenues dans le temps, ainsi qu'être adaptées à des erreurs proches, par exemple à d'autres types sémantiques d'adverbe. Leur création et leurs caractéristiques sont par conséquent documentées avec précision dans le Chapitre 3.

La deuxième exigence à intégrer dans le cadre de notre recherche concerne l'approche linguistique des questions de détection et correction automatique. L'élaboration des règles de correction doit permettre d'explorer les phénomènes linguistiques et grammaticaux qui correspondent aux erreurs sélectionnées, dans le but de mieux comprendre et modéliser ces phénomènes.

La troisième exigence porte sur le point d'aboutissement de la recherche. Celle-ci doit donner naissance à un système fonctionnel implémenté dans une plateforme informatique et pouvant faire l'objet de démonstrations. Nous ne posons pas comme exigence le fait d'aboutir à un "logiciel" complet avec une interface ergonomique, car cela s'apparente à une étape de développement et non de recherche. Cependant, le projet doit aller au-delà de la modélisation linguistique et de la preuve de concept, en proposant un système de correction pouvant être appliqué à des textes authentiques, avec une étendue et une précision de correction relativement satisfaisantes.

Le cadre posé par ces exigences est complété par un certain nombre de choix scientifiques et pratiques. Le premier de ces choix est d'ordre pragmatique et concerne l'étendue de nos travaux, qui ont pour objectif la création de règles de détection et correction dans un cadre de recherche. Ces règles pourraient être intégrées à des systèmes existants, tels des tuteurs de langue ou des correcteurs grammaticaux à visée pédagogique. Comme nous venons de le souligner, ce projet n'a donc pas pour objectif la création d'un système complet et autonome. Ce choix trouve sa justification évidente d'une part dans le temps et les ressources qui peuvent être consacrés à ces travaux, et d'autre part dans sa cohérence avec les objectifs d'un projet de recherche et non d'un projet de développement industriel. De la même façon, nos travaux n'entendent pas apporter de solution à l'ensemble des erreurs non traitées à ce jour, mais se

concentrent sur deux types d'erreur qui font l'objet d'un traitement détaillé. Ce nombre réduit est suffisant pour faire la présentation de notre méthode tout en répondant à l'exigence d'exploration des structures et systèmes linguistiques concernés.

Un autre choix important concerne celui d'associer détection et correction des erreurs dans un même projet. Cette approche n'est pas systématiquement adoptée dans les projets de recherche et/ou développement en correction automatisée, et certains projets privilégient détection ou correction, qu'il s'agisse d'un objectif final ou d'une phase intermédiaire. Par exemple, dans un article de 2009, Hermet et Désilets présentent une approche de la correction des erreurs de choix des prépositions pour les francophones écrivant en anglais. Leur méthode utilise une boucle de traduction automatique combinée à des modèles linguistiques, et fonctionne à partir d'erreurs déjà détectées, ce qui leur permet d'affiner les règles de correction à partir d'erreurs définies. À l'inverse, Nadasdi et Sinclair ont développé à partir de 2001 le correcteur grammatical *Spell Check Plus* (disponible en ligne) dont l'objectif est de détecter les erreurs et d'apporter un diagnostic correctif, mais sans proposer de correction. Cette approche originale est justifiée par l'orientation pédagogique du système, qui est conçu pour être utilisé dans un cadre d'apprentissage, avec l'aide d'enseignants et enseignantes. Nous avons choisi d'aborder détection et correction simultanément, d'une part parce que cette approche reste la plus fréquente pour ce type d'étude, et d'autre part par intérêt scientifique : nous souhaitons pouvoir étudier tous les aspects d'un phénomène linguistique, depuis son fonctionnement précis en anglais jusqu'à la question de son acquisition et des difficultés qu'elle peut poser aux personnes n'ayant pas l'anglais comme langue maternelle.

Le troisième choix que nous évoquons ici concerne le public auquel s'adressent les règles de correction. Même si nos travaux n'ont pas pour vocation d'aboutir à un système complet, il est important de déterminer le public cible du système, car ce paramètre a une influence sur certains volets importants comme la constitution du corpus ou le style de réponse corrective à apporter. Il y a trois principales caractéristiques à prendre en compte lors du choix d'un public cible pour un système de correction automatisée.

La première de ces caractéristiques est la langue maternelle des utilisatrices et utilisateurs, qui peut être soit la même que la langue sur laquelle les corrections s'appliquent (ex. : francophones utilisant un correcteur grammatical pour leurs textes en français), soit une langue différente (ex. : francophones utilisant un correcteur grammatical pour leurs textes en anglais ou dans une autre langue étrangère/seconde). Les correcteurs grammaticaux ne sont pas nécessairement créés à destination de personnes utilisant une langue étrangère/seconde, et

sont souvent conçus comme une aide à la rédaction pour les locuteurs et locutrices de cette langue. Ceci n'empêche pas leur utilisation par un public de langue étrangère/seconde, mais ils n'y sont pas spécifiquement adaptés. Si on fait le choix de s'intéresser à la correction à destination d'un public utilisateur de l'anglais en tant que langue étrangère/seconde (L2), il faut également déterminer s'il l'on souhaite que ce système s'adresse à des personnes ayant des langues maternelles (L1) variées, ou bien une seule L1. Ces deux approches ont des avantages et des inconvénients. La première permet de disposer de plus de ressources, par exemple sous la forme de corpus de textes produits par des publics apprenants variés, et d'être utile pour un plus grand nombre de personnes. La seconde permet d'affiner l'étape de la détection en utilisant les connaissances disponibles sur la L1, ainsi que de fournir une aide à la correction adaptée à la L1 du public cible. Nous avons opté pour cette seconde approche en créant des règles de détection/correction à l'usage des francophones.

La seconde caractéristique à prendre en compte dans le choix d'un public cible est le niveau de maîtrise de la langue cible par ce public. Puisque notre recherche n'a pas l'ambition de traiter l'ensemble des erreurs produites par les francophones écrivant en anglais, cette caractéristique est particulièrement importante : les erreurs produites par des utilisateurs et utilisatrices de l'anglais de niveau débutant sont différentes des erreurs produites par des personnes de niveau intermédiaire ou avancé, il faut donc choisir les erreurs à traiter en fonction d'un niveau de maîtrise précis. On peut également avancer le fait qu'un correcteur grammatical automatique n'est pas l'outil le mieux adapté pour un public de niveau débutant, puisqu'un tel outil s'adresse par nature à des personnes ayant les compétences nécessaires pour rédiger des phrases globalement bien construites. Il est probable qu'une personne utilisatrice de niveau intermédiaire à avancé, qui bénéficie d'une plus grande autonomie et de connaissances plus étendues concernant la L2, mais qui rencontre encore des difficultés de rédaction, soit le type de public le mieux indiqué pour cet outil.

Les objectifs des utilisateurs et utilisatrices, c'est-à-dire le type de production que ces personnes seront amenées à rédiger en anglais, font également partie des caractéristiques à prendre en compte. L'utilisation de l'anglais écrit se fait actuellement majoritairement dans un cadre professionnel ou semi-professionnel, et pour des types de document précis, comme les courriels ou les rapports. Une analyse détaillée de cette question est présentée dans le Chapitre 1, car elle influence le choix du corpus de documents utilisé comme base empirique de cette recherche.

Pour résumer les paragraphes précédents, notre cadre est donc le suivant : nous proposons d'élaborer des règles de détection et de correction pour un nombre d'erreurs restreint, à destination de francophones maîtrisant l'anglais écrit à un niveau intermédiaire à avancé, et l'utilisant principalement dans un cadre professionnel ou semi-professionnel. La recherche répond à des exigences de pertinence, d'innovation et de faisabilité concernant les erreurs sélectionnées. Enfin, les méthodes utilisées pour automatiser ces processus reposent dans la mesure du possible sur une modélisation des phénomènes linguistiques étudiés.

Les deux types d'erreur traités concernent le placement des adverbes en fonction de modifieur au niveau du groupe verbal ou de la proposition, et l'utilisation des noms en fonction de modifieur ou complément dans le groupe nominal. La catégorie des adverbes étant une des plus nébuleuses, nous nous penchons plus longuement sur la délimitation de ses contours dans le Chapitre 2. Nous pouvons pour l'instant nous contenter de définir les adverbes comme les termes dont la caractéristique principale est de modifier les catégories autres que les noms, en particulier les verbes, les adjectifs et les autres adverbes (Pullum et Huddleston, 2002 : 563). Les termes "séquence N+N" ou "structure N+N" sont utilisés de manière interchangeable pour faire référence à l'utilisation de noms en fonction de modifieur ou de complément dans le groupe nominal. Le statut linguistique de ces séquences particulières est interrogé en détail dans le Chapitre 2. Tous les termes syntaxiques utilisés, comme "modifieur" et "complément", sont également définis dans ce chapitre.

L'expression "règles de correction" est utilisée pour faire référence aux règles implémentées dans la plateforme informatique utilisée. Nous utilisons à certains endroits la précision "règles de détection et de correction" afin d'attirer l'attention sur les deux étapes du traitement des erreurs. Toutefois, même lorsque le terme "détection" n'est pas employé, il faut entendre l'expression comme faisant référence à ces deux étapes, détection et correction, dans le traitement des erreurs. Cette généralisation est cependant levée lorsque nous évoquons d'autres travaux de recherche. L'expression "schémas d'erreur et de correction" est employée, avec des variantes, pour renvoyer à la modélisation linguistique des segments erronés et de leurs corrections, en amont de leur automatisation.

Présentation des étapes de la recherche

La suite de cette introduction est consacrée à la présentation des étapes amenant à la réalisation de nos objectifs. Nous identifions cinq étapes principales : la sélection des erreurs à traiter, la modélisation linguistique des erreurs et des corrections à apporter, l'implémentation

des règles de détection et de correction, l'évaluation du système, et enfin la réalisation de l'accompagnement correctif. Chaque étape principale est décomposée en tâches plus courtes, que nous présentons ci-dessous.

Pour la première de ces étapes, c'est-à-dire la sélection des erreurs à traiter, il faut d'une part étudier des productions émanant du public cible choisi afin de déterminer quelles erreurs subsistent dans ces productions, et d'autre part connaître précisément les capacités des correcteurs grammaticaux déjà existants. Une des tâches consiste en la constitution d'un corpus de productions représentatives, suivie de sa validation et de son analyse. Nous proposons de procéder selon la méthode de l'analyse des erreurs, méthodologie classique dans le domaine de l'acquisition des langues secondes, grâce à laquelle nous relevons, catégorisons et annotons les erreurs trouvées dans le corpus. Dans un deuxième temps, nous passons en revue une sélection de correcteurs grammaticaux existants, et testons leur couverture grammaticale grâce à un échantillon d'erreurs présentes dans le corpus. À l'issue de cette première étape, il est possible d'identifier un certain nombre d'erreurs candidates à la correction. Deux types d'erreur sont sélectionnés à partir de critères scientifiques et pratiques.

La seconde étape concerne la modélisation linguistique des erreurs existantes et des corrections à apporter. La méthode utilisée ici est une méthode mixte, ascendante et descendante, qui s'appuie sur les observations empiriques des erreurs dans le corpus, sur des prévisions d'erreurs probables, et sur une modélisation du phénomène linguistique en question. Cette modélisation s'appuie sur les informations grammaticales fournies notamment par des grammaires de référence. Elle nécessite de confronter et de synthétiser les différentes approches et analyses des deux phénomènes étudiés qui sont faites dans les grammaires, afin de dégager les points de consensus et de faire des choix informés lorsqu'un point de désaccord est rencontré. Ces informations sont complétées par l'observation des phénomènes dans des corpus de productions écrites en anglais natif et non-natif. Cette méthode nous amène à la création d'une modélisation des schémas d'erreur et de leurs corrections associées pouvant mener à une implémentation informatique.

L'implémentation des règles de détection et de correction automatisées implique la sélection préalable d'une plateforme au sein de laquelle l'implémentation sera effectuée, ainsi que sa présentation et la justification de son adéquation à notre recherche. Il convient ensuite d'évaluer la nature et la quantité des ressources nécessaires au bon fonctionnement des règles automatisées, puis de les rassembler à partir de sources variées et bien documentées. Il peut également s'avérer nécessaire de créer de nouvelles ressources spécifiques aux présents

travaux. L'étape suivante consiste en la gestion de la programmation des règles et à l'évaluation technique du fonctionnement adéquat et coordonné des différents éléments du système.

L'évaluation technique, effectuée sur des textes contrôlés, doit être complétée par une évaluation de la qualité des règles sur des textes authentiques similaires aux textes utilisés dans le corpus initial. Il faut donc constituer des corpus adaptés à l'évaluation. De plus, il faut mettre sur pied une méthode d'interprétation des résultats de l'évaluation qui permette d'en extraire des pistes d'amélioration des règles de correction. Une fois l'évaluation conduite, les résultats doivent être analysés en détail afin d'identifier, si nécessaire, les insuffisances des différents modules du système, et planifier son développement futur.

La dernière étape de la recherche concerne l'accompagement des corrections. Il est prévu que la présentation des propositions de corrections au public utilisateur soit accompagnée de messages correctifs dont l'objectif est d'encourager l'apprentissage. La première tâche dans la création de cet outil est l'estimation de sa pertinence. Celle-ci est évaluée à l'aune des recherches effectuées sur ce thème dans les domaines de l'acquisition des langues secondes (ALS) et de l'enseignement des langues assisté par ordinateur (ELAO). Ces recherches nous permettent également d'obtenir des indications quant aux règles correctives les plus efficaces et les plus appropriées à ce genre de système. Cette tâche est complétée par l'observation des messages correctifs fournis par les correcteurs grammaticaux existants, et l'évaluation de leurs points forts et faibles. Ces recherches mènent à la création d'éléments de méthode pour la rédaction de tels messages correctifs, qui sont ensuite générés automatiquement dans le système choisi pour les règles de détection et correction.

Organisation de la thèse

Le présent document est organisé en trois chapitres. Si certaines problématiques sont abordées principalement dans un seul chapitre, d'autres sont présentes en filigrane tout au long du texte. Par ailleurs, la réalisation des cinq étapes principales présentées ci-dessus n'est pas systématiquement effectuée de manière linéaire, afin de respecter une certaine cohérence thématique dans les chapitres.

Le premier chapitre se concentre sur la première phase de notre recherche, et expose les différentes étapes menant à la sélection des erreurs à traiter. La première partie de ce chapitre introduit la méthode de l'analyse des erreurs, héritée du domaine de l'acquisition des langues secondes. Cette partie contient également une réflexion autour de ce qui constitue une

"erreur" lors de l'apprentissage ou de l'utilisation d'une L2, ainsi que sur les causes de ces erreurs. Une courte introduction à la perspective de l'anglais comme *lingua franca* ouvre un questionnement sur les conséquences du statut international de l'anglais sur la correction automatisée des erreurs. Dans la seconde partie de ce chapitre, nous exposons en détail les étapes de l'analyse des erreurs appliquées à notre problématique. La sélection des erreurs à traiter passe par leur relevé dans un corpus constitué selon les exigences de notre recherche, puis par la catégorisation de ces erreurs selon un système original. Nous proposons également un schéma adapté pour l'annotation des erreurs. À l'issue de ce chapitre sont présentés les deux types d'erreur sélectionnés, c'est-à-dire le placement des adverbes de manière en fonction de modifieur dans le groupe verbal et la proposition, et l'utilisation de noms en fonction de modifieur ou complément dans le groupe nominal.

L'objectif du deuxième chapitre est de fournir une modélisaton des erreurs produites et de leur correction, cette modélisation devenant la base des règles de correction dont l'implémentation est présentée dans le Chapitre 3. La première partie du Chapitre 2 est consacrée à l'exposé de la méthode et des sources utilisées pour la modélisation des deux phénomènes étudiés. Il s'agit d'une méthode mixte fondée sur la synthèse d'informations grammaticales et l'observation du fonctionnement des éléments concernés dans des corpus d'anglais natif et non-natif. Les sources de départ sont les trois grammaires de référence les plus récentes pour l'anglais, *A Comprehensive Grammar of the English Language* (Quirk et al., 1985), *Longman Grammar of Spoken and Written English* (Biber et al., 1999) et *The Cambridge Grammar of the English Language* (Huddleston et Pullum et al., 2002). Le reste du chapitre est organisé en deux parties, respectivement consacrées au placement des adverbes dans la proposition et à l'utilisation des noms en fonction de dépendant dans le groupe nominal. Chacune de ces parties intègre une présentation de la syntaxe générale des phénomènes étudiés fondée sur la synthétisation des informations données dans les sources de référence. Les erreurs relevées dans le corpus pour ces deux types sont ensuite analysées, ces analyses étant informées et complétées par des indications issues des domaines de la linguistique et de l'acquisition des langues secondes. Nous procédons de manière progressive afin de clarifier au maximum les incertitudes subsistant quant à l'usage des adverbes et des noms en fonction de dépendants, afin de présenter une modélisation des erreurs et des corrections aussi précise que possible pour ces deux types d'erreur. Les difficultés rencontrées ne manquent pas d'être soulignées.

Le troisième et dernier chapitre de ce document présente l'implémentation technique des règles de correction, leur évaluation, et nos propositions pour la création de messages correctifs accompagnant les corrections. La première partie de ce chapitre présente l'état des recherches dans le domaine de la correction grammaticale automatisée à destination des personnes apprenantes, présentation qui est complétée par une évaluation des capacités d'une sélection de correcteurs automatiques existants. La deuxième partie du chapitre présente en détail les aspects techniques de l'implémentation des règles de correction, ainsi que les résultats des évaluations et les pistes d'amélioration découlant de l'interprétation des résultats. Pour finir, nous nous penchons sur la rétroaction corrective en acquisition des langues, et plus particulièrement pour l'enseignement des langues assisté par ordinateur. La présentation de l'état des recherches dans ce domaine et l'observation des aides correctives utilisées dans les correcteurs grammaticaux existants mènent à l'élaboration d'un canevas en cinq étapes pour la création de messages correctifs dans notre système.

La thèse présente notre recherche de manière progressive, ce qui correspond également à une approche par domaines de recherche. Le premier chapitre regroupe les aspects de notre recherche ayant trait à l'acquisition des langues, le deuxième est ancré dans la linguistique et la grammaire de l'anglais, avec un détour par les études de corpus, alors que le troisième chapitre s'inscrit dans le domaine du traitement automatisé des langues et de l'enseignement des langues assisté par ordinateur. La longueur plus importante du deuxième chapitre est le signe de la place primordiale de la linguistique de l'anglais dans notre projet.

De manière générale, le terme "partie" fait référence au premier niveau de division des chapitres (ex. : 1.2), celui de "sous-partie" renvoie au deuxième niveau de division (ex. : 1.2.2), et enfin celui de "section" est utilisé pour les divisions des sous-parties (ex. : 1.2.2a). Certaines sections plus longues sont organisées en plusieurs parties recevant chacune un titre non-numéroté ; nous utilisons le terme de "sous-section" pour faire référence à ces parties, et le titre de celles-ci est donné en entier lorsque nous incluons un renvoi à l'une d'entre elles (ex. : 1.2.2b, "Proposition de schéma d'annotation").

Ce document adopte des pratiques de rédaction épicènes.

Chapitre 1
L'analyse des erreurs appliquée à la correction automatisée : Éléments théoriques et méthodologiques

Introduction

Ce premier chapitre introduit les processus et réflexions menant à la sélection des erreurs qui feront l'objet d'une correction automatisée. Ces types d'erreur doivent satisfaire trois exigences, que nous posons *a priori* : l'exigence d'utilité, l'exigence d'innovation, et l'exigence de faisabilité. Chaque exigence implique la prise en compte de facteurs spécifiques et la mise en place de processus de sélection croisés.

Evaluer la faisabilité de la détection et de la correction d'un certain type d'erreur implique de prendre en compte le cadre que nous avons défini, en particulier en ce qui concerne l'utilisation de méthodes basées sur des modèles linguistiques, toutes les erreurs ne se prêtant pas avec succès à ce mode de traitement. Cette question sera abordée plus en détail à la fin du présent chapitre, mais elle restera présente tout au long de ce document.

Satisfaire l'exigence d'innovation implique de choisir des types d'erreur n'ayant pas fait l'objet d'études et de traitements approfondis dans le domaine de la correction grammaticale automatisée. Ceci nécessite d'une part de passer en revue les différents travaux de recherche à ce sujet, et d'autre part de connaître les capacités des correcteurs grammaticaux actuels. Ces deux études, sous la forme d'un état de l'art et d'une étude comparative, sont présentées dans le troisième chapitre de ce document, pour des raisons de cohérence thématique, ce dernier étant consacré plus spécifiquement à la correction grammaticale automatisée. Nous ferons néanmoins référence aux résultats de ces études dès le premier chapitre lorsque cela s'avèrera nécessaire.

Instaurer une exigence d'utilité signifie simplement que la détection et la correction de ces erreurs doit être utile au public cible choisi : ces erreurs doivent figurer parmi les plus répandues dans leurs productions, et présenter des difficultés liées à leur correction et à l'apprentissage des structures et phénomènes dont elles dépendent. Il paraît nécessaire d'observer les textes produits en anglais par notre public cible afin de rechercher et catégoriser les erreurs qui s'y trouvent, pour déterminer lesquelles mériteraient le développement de règles de correction. Recueillir des informations sur les erreurs produites par les apprenants de langue seconde est l'objectif principal de la méthode de l'"analyse des erreurs", codifiée par Corder au cours des années 1960 et 1970, et appartenant au domaine de l'acquisition des langues secondes. Nous proposons d'appliquer cette méthode et d'en adapter les étapes afin de la rendre cohérente dans le cadre de notre recherche. C'est donc plus particulièrement à la

présentation de cette méthode, de son adaptation et des résultats obtenus que nous consacrons notre premier chapitre.

La première partie de ce chapitre consiste en une présentation générale de la méthode de l'analyse des erreurs. Nous évoquons l'émergence de cette méthode, puis nous nous penchons plus spécifiquement sur deux aspects importants pour nos travaux : une réflexion autour de la définition et la délimitation des erreurs, ainsi qu'une présentation de deux de leurs causes principales en lien avec les processus d'apprentissage, c'est-à-dire le transfert et la surgénéralisation. Une sous-partie est également consacrée à la présentation du concept de l'anglais comme *lingua franca*, qui propose un changement de paradigme dans la façon dont sont jugées les productions de personnes n'ayant pas l'anglais comme langue maternelle. Nous verrons comment prendre en compte les conclusions de cette re-conceptualisation dans notre recherche. Dans un second temps, nous présentons l'application des étapes de l'analyse des erreurs à un projet de correction automatisée. Ceci nous amènera à donner les détails de la constitution de notre corpus de productions d'utilisateurs et utilisatrices de l'anglais et de la catégorisation des erreurs repérées dans les productions. Nous évoquerons également le schéma d'annotation des erreurs qui a été élaboré, malgré le fait que toutes les erreurs n'aient pas été annotées dans le corpus numérisé. À la fin de cette seconde partie, nous reviendrons sur les trois exigences citées dans cette introduction et justifierons notre choix de traiter les erreurs liées au placement des adverbes et à l'utilisation de noms en fonction de modifieur ou complément du nom noyau dans un groupe nominal.

1.1 L'analyse des erreurs, un héritage de l'étude de l'acquisition des langues secondes

Le questionnement autour des erreurs langagières, qu'elles concernent les productions de personnes locutrices natives ou apprenantes d'une L2, est une considération tout aussi ancienne que courante. À défaut de la création d'une institution soi-disant gardienne du bon usage de l'anglais, comme l'Académie Française peut l'être pour le français, de nombreux ouvrages plus ou moins prescriptifs ont été publiés dans l'objectif de tenter d'endiguer des usages "erronés" de l'anglais. L'un des premiers représentants modernes de ces ouvrages fut *The King's English* de Fowler (1906), plusieurs fois réédité et mis à jour ; l'un des plus récents et des moins prescriptifs est le *Merriam-Webster's Dictionary of Contemporary English Usage* (1994), fondé sur des bases empiriques rigoureuses, du propre aveu des auteurs de *The Cambridge Grammar of the English Language* (2002). Il existe également des ouvrages plus

spécifiquement destinés aux publics apprenants et enseignants de l'anglais, et rédigés dans le but de recenser les erreurs les plus fréquentes. *Common Mistakes in English* de Fitikides, publié pour la première fois en 1936, en est un exemple connu. L'ouvrage de Swan et Smith, *Learner English: A Teachers's Guide to Interference and Other Problems* (2001) attire l'attention des enseignantes et enseignants sur l'influence que peut avoir la langue maternelle sur les erreurs qui sont produites (la question du transfert est abordée dans la section 1.1.3 de ce chapitre).

L'intérêt que suscitent les erreurs est également partagé par la communauté scientifique du domaine de l'acquisition des langues secondes. La méthode de l'"analyse des erreurs" (AE) a été élaborée à partir de l'idée selon laquelle observer les erreurs permettrait d'étudier les processus liés à l'acquisition des langues secondes/étrangères (Ellis, 2008 : 45). Ellis précise que si cette approche scientifique de l'analyse des erreurs est relativement récente, l'observation des erreurs est depuis longtemps utilisée dans un cadre pédagogique plus informel (op. cit. : 960).

La méthode de l'AE est composée de cinq étapes de traitement des erreurs, de la constitution d'un échantillon d'erreurs à l'interprétation de leurs causes. Nous présentons les conditions d'émergence, les principaux objectifs et les critiques formulées à son égard dans la section suivante, mais il nous paraît important d'avertir dès à présent notre lectorat que notre utilisation de la méthode de l'AE ne constitue pas un parti pris en faveur de sa validité en tant que théorie de l'acquisition, et au détriment d'autres approches qui peuvent sembler entrer en contradiction avec elle. N'étant pas en position de juger du bien-fondé de chaque théorie d'acquisition des langues secondes, nous nous contentons ici d'utiliser le cadre méthodologique de l'AE comme un outil de recueil de données reconnu comme valide en tant que tel (Cook, 1993 : 22). Nos travaux étant centrés sur les erreurs produites par des personnes utilisatrices d'une langue seconde/étrangère, il nous semble nécessaire d'emprunter les méthodes classiques des spécialistes de l'acquisition des langues secondes.

Les sous-parties et sections qui suivent constituent une présentation des différents aspects de la méthode de l'AE, que nous appliquons ensuite à la sélection d'erreurs pertinentes. Cette présentation générale nous permet également d'aborder des questions centrales comme la définition et l'identification des erreurs, ainsi que l'analyse de leurs causes.

1.1.1 Une méthode de recueil de données

Dans cette section, nous abordons tout d'abord le contexte d'émergence de la méthode de l'AE, son évolution et ses applications récentes, ainsi que ses limites. Dans un second temps, nous présentons les cinq étapes qui la constituent, et à partir desquelles nous menons notre propre analyse (voir partie 1.2). Deux de ces étapes font l'objet d'une présentation plus détaillée, en 1.1.2 et 1.1.3.

a. Émergence, objectifs et limites de l'analyse des erreurs

Les fondements théoriques de l'analyse des erreurs ainsi que les procédures empiriques qui la composent ont été énoncés principalement par Corder dans une série d'articles au cours des années 1960 et 1970, et dont les principales conclusions sont rassemblées dans son ouvrage *Error Analysis and Interlanguage* (1981). Dans sa monographie consacrée à l'AE, James en donne la définition suivante : "Error Analysis is the process of determining the incidence, nature, causes and consequences of unsuccessful language" (1998 : 1). Nous reviendrons dans la section 1.1.2 sur ce qui constitue un "segment langagier infructueux", notre tentative de traduction de l'expression tout aussi vague qu'évocatrice *unsuccessful bit of language*, et que James donne comme définition provisoire du terme "erreur".

L'émergence de l'AE a constitué un changement de paradigme théorique dans le domaine de l'ALS, qui était dominé dans les années 1950 et 1960 par le cadre de la linguistique contrastive, et par son application, l'analyse contrastive. Cette méthode d'analyse vise à comparer les systèmes linguistiques de deux langues afin d'en identifier les ressemblances et les différences, en faisant l'hypothèse, dans le cadre de l'enseignement et de l'acquisition des langues étrangères, que ces observations permettront de prédire les facilités et les difficultés d'apprentissage d'une langue donnée pour une locutrice ou un locuteur d'une autre langue, elle aussi déterminée. D'après cette hypothèse, il serait possible de découvrir la source du tout ou partie des erreurs produites (Ellis, 2008 : 958).

Les résultats obtenus par cette approche furent cependant jugés comme décevants car ils se révélaient peu informatifs et peu prédictifs (James, 1998 : 4). Par ailleurs, cette approche était associée au béhaviorisme, théorie de l'apprentissage de plus en plus remise en question pendant cette période (op. cit. : 4). Ces deux limites ont ouvert la voie à une méthode proposant de comparer non pas langue maternelle (L1) et langue cible (L2), mais les productions de personnes apprenantes dans la L2 et des productions en L2 "bien formées", c'est-à-dire émanant de personnes ayant l'anglais comme langue maternelle. L'AE propose

ainsi de substituer à la tentative de prédiction des erreurs l'étude d'erreurs réelles, afin d'en découvrir les différentes caractéristiques (ex. : incidence, nature, causes) (James, 1998 : 5). James indique cependant que l'influence de la L1 n'est pas totalement mise de côté : comme nous le verrons dans la section 1.1.3, le transfert, ou influence translinguistique, est un des facteurs possibles d'erreur, et ce phénomène est pris en compte lors de la phase d'explication des erreurs.

Le fait de comparer les productions de personnes apprenantes à des productions "natives" a rapidement donné lieu à des critiques de la part de groupes dans la communauté scientifique de l'ALS qui estimaient que les productions linguistiques des personnes apprenantes devaient être considérées comme cohérentes en elles-mêmes et donc exemptes d'"erreurs", puisqu'elles étaient simplement la manifestation du développement de leur système linguistique au moment de la production. Corder a lui-même développé le concept de "dialecte idiosyncratique" (*idiosyncratic dialect*) (1981 [1971]), un concept proche de celui d'"interlangue" (*interlanguage*) introduit par Selinker en 1972. Le terme d'"interlangue" est aujourd'hui largement accepté et utilisé pour décrire le système linguistique intermédiaire utilisé par les apprenantes et apprenants d'une L2. L'adoption de cette nouvelle façon d'envisager les productions des apprenants a mené à un nouveau changement de paradigme qui a rapidement éclipsé l'AE en tant que méthode privilégiée pour la découverte des processus d'acquisition d'une L2. Nous reparlerons des questions posées par la comparaison des productions de personnes utilisatrices de l'anglais à des productions natives dans la section consacrée à la définition des erreurs (1.1.2), ainsi que dans celle consacrée à la présentation du concept de l'anglais comme *lingua franca*.

Les objectifs de l'AE sont principalement de deux sortes : faire avancer les connaissances concernant les processus d'acquisition d'une L2, et obtenir des résultats permettant de développer ou d'enrichir des approches pédagogiques. D'un point de vue pédagogique, l'observation et la description des erreurs peut permettre d'apprécier l'écart entre ce qui est enseigné (*input*) et ce qui a été intégré par les personnes apprenantes (*intake*) (Corder, 1981 [1967]), et de mettre en lumière les phénomènes linguistiques qui présentent plus de difficultés pour un groupe de personnes donné. Répondant aux hypothèses chomskyennes de l'acquisition des langues, selon lesquelles les apprenants et apprenantes tirent un bénéfice de l'obtention d'informations positives comme négatives sur ce qui constitue un segment de langue bien formé, James indique que l'AE permet de fournir un échantillon de segments de

langue mal formés mais représentatifs de ce que les apprenants sont en capacité de produire (1998 : 54), qui peut ensuite être utilisé comme information négative.

Si la capacité de l'AE à fournir les résultats évoqués ci-dessus n'est pas remise en question, elle est néanmoins critiquée pour son intérêt unique pour les erreurs, arbres qui cachent la forêt constituée de tous les segments corrects produits par les personnes apprenantes. Puisqu'elle quantifie les erreurs visibles, elle ne permet pas non plus d'attirer l'attention sur les stratégies d'évitement souvent mises en place consciemment ou inconsciemment pour éviter de révéler un manque de connaissances (Schachter, 1974, cité par James). Du point de vue de l'avancée des connaissances sur les processus d'acquisition, Ellis souligne le fait que le manque d'études longitudinales dans le cadre de l'AE ne permet par d'isoler et d'étudier les différentes étapes de développement des capacités des apprenants (2008 : 61). À l'exception de celle portant sur les stratégies d'évitement, ces limites ne sont cependant pas inhérentes à la méthodologie de l'AE, mais concernent plutôt des problèmes de réalisation dans les recherches effectuées dans le cadre de cette méthodologie ; par exemple, rien dans cette méthodologie n'empêche la mise en place d'études longitudinales, mais ces études sont notoirement difficiles à mener et coûteuses, notamment sans l'aide des technologies disponibles actuellement, ce qui explique leur absence ici.

Ces limites ont mené à une baisse de popularité de cette méthode, mais elle a continué à être utilisée comme moyen codifié de recueillir des données sur les productions des personnes apprenantes, au service de projets précis. L'AE a connu un certain renouveau avec l'avènement des techniques de traitement informatisées (Ellis, 2008 : 64), notamment dans le cadre des projets liés à la création du *International Corpus of Learner English*, dont les erreurs sont annotées et ont donné lieu à une analyse (Granger et al., 2009). Dagneaux et al. présentent d'ailleurs une défense convaincante de l'intérêt scientifique de l'analyse des erreurs assistée par ordinateur et une présentation du rôle que cette méthode pourrait jouer dans la création d'outils pédagogiques innovants (1998 : 163).

b. L'analyse des erreurs en cinq étapes

Les paragraphes suivants présentent les cinq étapes de l'AE, ainsi que les questionnements et problèmes méthodologiques qu'elles soulèvent. Nous reprenons les termes donnés par Ellis (2008), qui sont également ceux employés par Corder (1981 [1974]). Nous utilisons également l'ouvrage de James consacré à l'AE dans notre synthèse, et tâchons de souligner les différences dans les approches. Deux des étapes de l'AE, l'identification des erreurs et

l'explication des erreurs, donnent lieu à un traitement plus détaillé dans les sous-parties 1.1.2 et 1.1.3 du fait de leur importance pour notre recherche.

Constitution d'un corpus de productions d'apprenants

La première étape de l'AE concerne la constitution d'un corpus de productions de personnes apprenantes. Cette étape est décisive pour les résultats qui seront obtenus, les différents paramètres du corpus de productions choisis influençant grandement ces résultats. Les productions spontanées permettent par exemple d'observer des erreurs très différentes de celles que l'on trouve dans des productions "réfléchies" ou qui vont être soumises à une évaluation extérieure, comme les productions issues de copies d'examens. Le tableau suivant, que nous empruntons à Ellis (2008 : 47), présente de manière synthétique les différents facteurs à prendre en compte lors de la constitution d'un corpus pour l'AE. Nous reprenons ces différents facteurs dans la section 1.2.1, consacrée à la constitution du corpus ayant servi de base aux présents travaux.

Facteurs	Variables	Description
Personne apprenante	Niveau de compétence	Ex : Élémentaire, intermédiaire, avancé
	Connaissance d'autres langues	L1 ; autres L2
	Expérience d'apprentissage des langues	Contexte dans lequel la personne a appris la L2
Type de document	Medium	Production orale ou écrite
	Genre	Conversation, rédaction, lettre, présentation, etc.
	Contenu	Thème sur lequel la personne communique
Type de production	Spontanée	Le discours (écrit ou oral) est produit spontanément
	Planifiée	Le discours (écrit ou oral) est produit dans un contexte qui permet à la personne de prévoir sa production

Tableau 1. Facteurs à prendre en compte lors de la constitution d'un corpus pour l'AE

Ellis déplore que les corpus soient généralement limités à des productions écrites (rédactions, traductions) faites lors d'examens ou de contrôles (op. cit. : 47). Ces corpus peuvent cependant prendre de nombreuses formes, et peuvent être dans certains cas plus proches d'"échantillons" de langue que de véritables corpus. James évoque notamment la technique d'observation de classe menant à une liste d'erreurs produites à l'oral (op. cit. : 20).

Identification des erreurs

La tâche d'identification des erreurs qui suit la constitution d'un corpus peut à première vue sembler relativement simple. Il n'en est rien : il s'agit d'un processus soumis à de nombreuses variables, et qui implique d'avoir mené une réflexion préalable sur ce qu'est en réalité une erreur de langue. Comme nous l'avons indiqué précédemment, James définit l'erreur comme "un segment langagier infructueux" (1998: 1, notre traduction ; l'humour et la concision de la formule anglaise, *unsuccessful bit of language*, sont malheureusement perdus en français), définition qui entraîne des questions en cascade. De quoi l'erreur est-elle l'échec ? Par rapport à quelle(s) norme(s) l'échec est-il jugé ? Comment délimiter ces segments ? Toutes les erreurs sont-elles visibles à l'œil nu ? Nous tentons d'apporter, sinon des réponses, au moins des pistes de réflexion plus approfondies dans la sous-partie 1.1.2.

Description des erreurs

Si Ellis assimile la description des erreurs à leur catégorisation, James a une conception plus étendue de cette étape (op. cit. : 94). Il indique que la catégorisation des erreurs est précédée par une étape de description grammaticale à partir d'un système théoriquement neutre, comme peuvent l'être les grammaires descriptives de Quirk et al. (1985) ou Huddleston et Pullum et al. (2002), et qui offrent d'après James le meilleur cadre pour l'aspect pratique de la description des erreurs. Malheureusement, James reste très vague sur ce qu'implique réellement cette phase de description menant à la catégorisation. Il indique cependant que cette description "neutre" des erreurs sert trois objectifs. Le premier, sur lequel il s'étend peu, est de rendre explicite ce qui serait resté implicite si l'erreur n'était pas clairement décrite, et ainsi rendre possible la confrontation des intuitions sur l'erreur. Le deuxième objectif de la description est la quantification des erreurs. Le troisième en est la catégorisation.

Il existe deux principaux types de système de catégorisation des erreurs, ou taxonomies, selon le phénomène que l'on souhaite mettre en avant : les systèmes construits à partir de catégories linguistiques, et ceux fondés sur l'observation des phénomènes de surface. Notons que les erreurs peuvent également être organisées en "dictionnaires d'erreurs", comme les ouvrages que nous citons dans l'introduction de la présente section.

Les systèmes fondés sur l'observation des phénomènes de surface proposent de catégoriser les erreurs selon les différences observables entre la production de l'apprenant et celle qui

serait attendue dans la L1. Dulay, Burt et Krashen (1982 : 150) identifient quatre types principaux de modification de surface (exemples empruntés à Dulay, Burt et Krashen) :

- l'omission :

  [1] **She sleeping.* (omission de l'auxiliaire)

- l'ajout :

  [2] **We didn't went there.* (le passé est considéré comme doublement marqué)

- la malformation :

  [3] **The dog ated the chicken.* (forme correcte *ate* déformée en *ated*)

- les erreurs d'ordre des mots :

  [4] **What daddy is doing?* (inversion de l'ordre de l'auxiliaire et du verbe)

James propose d'y ajouter un cinquième type, les erreurs hybrides (*blends*), qui peuvent se produire lorsque deux structures proches sont en compétition, ce qui amène l'apprenant à les combiner (exemple emprunté à James : 111) :

[5] **according to Erica's opinion*

Les systèmes construits à partir de catégories linguistiques utilisent quant à eux les différents aspects linguistiques affectés par les erreurs comme critères de classification (Dulay, Burt et Krashen, 1982 : 146). Ils peuvent inclure plusieurs niveaux de critères, à commencer par le domaine général de l'erreur, qu'il s'agisse de la phonologie, de l'orthographe, de la grammaire, du lexique, ou du discours. Dans le cas des erreurs concernant le domaine grammatical, James suggère d'indiquer ensuite la partie du discours concernée (nom, verbe, adjectif, etc.), puis le "rang" de l'erreur, c'est-à-dire la position du segment erroné dans la hiérarchie des unités liées à cet élément ; dans le cas du nom, les différents rangs possibles seraient ceux du morphème, du mot, du groupe nominal et de la proposition nominale. Le dernier critère mentionné concerne le système grammatical affecté par l'erreur, comme par exemple le système temporel, la transitivité, le nombre, etc. Ce système est particulièrement utile dans le cadre des erreurs grammaticales, et pour la description des erreurs d'apprenants et apprenantes de niveau avancé (op cit. : 5).

Le principal intérêt de ces deux modes de classification est leur application pédagogique : elles permettent d'obtenir des informations sur les erreurs qui peuvent être directement utilisées pour améliorer l'enseignement d'une L2 à l'attention d'un groupe déterminé de

personnes apprenantes. James propose d'ailleurs de les combiner afin d'obtenir un système plus complet, à deux, voire trois dimensions. Ces systèmes sont moins intéressants du point de vue de la recherche sur les processus d'acquisition d'une L2, car ils n'offrent pas d'éclairage sur les règles d'apprentissage (Ellis, 2008 : 51). Corder (1981 [1974]) a proposé un système basé sur l'évaluation de la systématicité des erreurs, mais il implique d'avoir accès aux réflexions des apprenants, ce qui le rend difficile à mettre en place.

Nous terminerons en identifiant deux problèmes qui se posent lors de cette étape dans le traitement des erreurs. Tout d'abord, il peut être difficile de décider dans quelle catégorie une erreur doit être classée, même si le système utilisé inclut des distinctions fines. La description d'une erreur, et donc sa classification, implique d'une part la compréhension du message d'un point de vue sémantique, et d'autre part la reconstruction du segment dans la langue cible d'un point de vue formel, deux aspects qui peuvent poser problème pour un certain nombre d'erreurs.

L'erreur suivante, extraite de notre corpus, en est un exemple :

[6] **People are decided to keep their identity.*

Il est impossible ici de savoir si la correction du segment doit être *People have decided to keep their identity*, ou *People are determined to keep their identity* ; ces deux reconstructions ont des sens différents et seraient également classées dans des catégories différentes dans un système fondé sur des critères grammaticaux. Le second problème lié à la catégorisation des erreurs concerne l'un des buts identifiés par James pour cette étape, c'est-à-dire la quantification des erreurs et l'évaluation de leur fréquence. Les fréquences d'erreur obtenues grâce à leur catégorisation sont le plus souvent absolues, et les corpus utilisés pour les diverses études sont de natures si différentes que ces fréquences n'ont que peu de valeur comparative. Rechercher des fréquences relatives, c'est-à-dire le nombre d'erreurs produites comparé au nombre de fois où les erreurs auraient pu être produites, serait plus efficace de ce point de vue, mais s'avère extrêmement difficile voire impossible si on utilise des productions spontanées et/ou libres (conversations, rédactions, etc.).

Explication des erreurs

L'objectif de l'étape d'explication des erreurs est de découvrir leurs causes, c'est-à-dire les phénomènes liés à l'apprentissage d'une L2 qui ont mené à la production des segments erronés. D'après Ellis, il s'agit de l'étape la plus importante pour le domaine de l'ALS (op. cit. : 53). Deux causes sont généralement identifiées : l'influence translinguistique, ou transfert, et

les phénomènes menant à des erreurs intralinguales (*intralingual errors*), aussi appelées erreurs liées au développement (*developmental errors*).

Dans le cas du transfert, l'apprenante ou apprenant applique à la L2 ses connaissances concernant la L1 (ou une autre langue également connue), ce qui peut donner lieu à des erreurs. Voici un segment erroné, tiré de notre corpus, et dans lequel l'erreur peut être attribuée à un transfert :

[7] *?[It] won't change completely the life of its citizens.*

Dans cet exemple, l'adverbe *completely* est placé entre le verbe et le complément d'objet, ce qui est le placement par défaut en français, mais qui crée une erreur en anglais, langue dans laquelle ce placement n'est pas accepté. Nous revenons plus en détail sur cet aspect dans le Chapitre 2, section 2.2.2a.

Les erreurs intralinguales sont quant à elles dues à des problèmes dans l'application des règles de la L1, comme par exemple la surgénéralisation. Celle-ci est visible dans l'exemple suivant, où la construction de la proposition interrogative indirecte est calquée sur la construction des interrogatives directes en anglais, dans lesquelles l'ordre canonique est, pour simplifier, [Mot Interrogatif + Verbe + Sujet] :

[8] **Let's first have a look at what is Europe.*

Cet exemple nous permet d'aborder un point sur lequel Ellis et d'autres chercheurs tels que Flick (1979 : 60) et Schachter et Celce-Murcia (1977), cités par Ellis, ont particulièrement insisté : il est en réalité très difficile d'évaluer la source véritable des erreurs, et toute conclusion en ce sens doit être prudente et solidement étayée. Ainsi, l'exemple cité plus haut pourrait tout aussi bien être interprété comme un exemple de transfert, puisque la structure [Mot Interrogatif + Verbe + Sujet] est possible dans les subordonnées interrogatives (cf. en français, "… ce qu'est l'Europe"). Nous présentons plus longuement les phénomènes du transfert et de la surgénéralisation dans la sous-partie 1.1.3.

Évaluation des erreurs

D'après James, l'objectif principal de l'évaluation des erreurs est d'établir des priorités pédagogiques (op. cit. : 205). Évaluer les erreurs implique d'apprécier leur "gravité" selon un certain nombre de critères. Cette démarche peut être vue de manière négative, comme un jugement porté sur les personnes apprenantes et leurs productions "imparfaites". Cependant, James répond à cette critique en soulignant la chose suivante : "Passing judgment on error is

not a matter of devaluation of learners and their language, but rather one of assigning relative values to errors". Il insiste également sur l'intérêt de l'évaluation : "Evaluation is indeed a matter of ethics, since society rewards those who get things right" (op. cit. : 205). Nous verrons en section 1.1.4 que cette approche de l'évaluation et des erreurs en général n'est pas uniformément acceptée dans la communauté scientifique de l'acquisition des langues secondes.

Les critères retenus par James pour évaluer la gravité d'une erreur sont au nombre de cinq : les critères d'ordre linguistique, la fréquence des erreurs, la compréhension, la visibilité, et l'irritation causée par les erreurs. Notons que ces critères se recoupent largement, comme nous allons le voir, et qu'ils représentent chacun une façon d'aborder la question de la gravité plutôt que l'ajout d'un degré dans celle-ci.

D'un point de vue linguistique, les erreurs peuvent être jugées comme étant plus ou moins graves selon la nature de la règle qui est concernée. Ainsi, une erreur liée à une règle de grammaire hautement généralisable serait jugée par un relecteur ou une relectrice comme étant plus grave qu'une erreur touchant un domaine peu généralisable, comme le lexique. Par ailleurs, James indique qu'une erreur peut être jugée plus grave si elle concerne un segment de langue long plutôt qu'un seul mot, mais ce phénomène s'explique aussi du point de vue de la compréhension, qui est rendue plus difficile si une partie plus longue du discours doit être reconstruite.

Concernant la longueur du segment de langue concerné, Burt (1975, citée par Ellis) distingue les erreurs globales des erreurs locales. Les erreurs globales affectent l'organisation générale de la phrase, comme l'omission de connecteurs ou les erreurs d'ordre des mots, tandis que les erreurs locales n'affectent qu'un seul élément de la phrase, comme les erreurs de morphologie. Les erreurs globales auraient tendance à gêner la compréhension des phrases, et seraient donc considérées comme plus graves.

La fréquence d'une erreur peut également être utilisée pour évaluer la gravité d'une erreur de manière indirecte : une erreur peut être considérée comme "grave", c'est-à-dire nécessitant une remédiation, soit parce qu'elle est produite fréquemment par un apprenant, soit parce que la structure sur laquelle porte l'erreur est fréquemment utilisée en L1, ces deux raisons pouvant bien sûr être corrélées. Nous avons cependant vu que la mesure de la fréquence des erreurs relativement aux nombres de formes de même nature utilisées dans un corpus donné est problématique, ce qui rend ce critère difficile à exploiter.

Les deux critères suivants concernent l'effet que les erreurs ont sur les destinataires du discours des apprenants et apprenantes. Le critère de compréhension est jugé comme étant celui à considérer en priorité lors de l'évaluation de la gravité d'une erreur (Johansson, 1973 cité par Ellis). La "compréhension" d'un message est liée à deux facteurs : l'intelligibilité de celui-ci, c'est-à-dire l'accessibilité de son contenu propositionnel, et ce que James nomme la valeur communicative du message (*communicativity* ; James, 1998 : 216). Cette dernière est liée aux aspects pragmatiques et sociaux de la langue, et fait partie des aspects les plus difficiles à maîtriser pour les personnes apprenantes. Concernant l'intelligibilité d'un message, soulignons qu'elle ne dépend pas obligatoirement de sa grammaticalité mais est également largement influencée par d'autres facteurs, notamment par les erreurs lexicales.

Le second critère représentatif de l'effet des productions d'interlangue sur leurs destinataires est la réponse affective négative engendrée par les erreurs, qualifiée d'"irritation". Cette réaction dépend cependant d'un grand nombre de facteurs ayant parfois peu à voir avec la linguistique, comme la relation qu'entretiennent les co-énonciatrices ou co-énonciateurs, les préjugés sociaux, ou simplement l'état d'esprit des destinataires. Pour cette raison, le critère d'irritation a été jugé comme difficile à exploiter efficacement.

Nous avons abordé les différents critères qui peuvent être utilisés pour évaluer les erreurs. Il nous reste à évoquer un autre aspect important de cette étape, c'est-à-dire la question de *qui* évalue les erreurs. Ces personnes peuvent être locutrices natives ou non-natives de la L2, être des enseignants et enseignantes de la L2, ou bien des personnes "non-expertes". Les personnes locutrices natives et non-natives réagissent différemment aux erreurs. Les non-natives ont tendance à être globalement plus sévères que les natives, notamment sur les erreurs portant sur des mots grammaticaux, alors que les personnes locutrices natives jugent plus sévèrement les erreurs liées à la compréhension et au lexique (Ellis, 2008 : 56-57).

Dans l'ensemble, la question de l'évaluation de la gravité des erreurs laisse de nombreuses interrogations en suspens, car il est difficile de généraliser les résultats obtenus par différentes études. Le manque d'homogénéité et de reproductibilité est d'ailleurs une des critiques générales adressées aux recherches effectuées dans le cadre de la méthodologie de l'AE (op. cit. : 47). Nous reviendrons rapidement sur ces critères d'évaluation de la gravité lors de la sélection des erreurs à traiter, en cohérence avec notre exigence d'utilité pour les personnes apprenantes et utilisatrices de l'anglais.

1.1.2 Qu'est-ce qu'une erreur ?

Dans la section précédente, nous avons formulé un ensemble de questions découlant de la définition préliminaire de l'"erreur", donnée par James. La synthèse qui suit présente des pistes de réflexion pour y répondre. Les réponses à ces questions renvoyant le plus souvent aux mêmes problématiques, nous n'y répondons pas de manière linéaire, mais nous proposons de revenir sur chaque question en conclusion.

Nous avons indiqué précédemment que l'erreur constituait un "échec", à partir de la traduction d'un des éléments de la définition de James (*unsuccessful*). Tout comme James affine sa définition au fil des pages, ce choix de terme doit à présent être révisé, car il attache à la notion d'erreur une connotation négative qui n'a pas sa place lorsque l'on fait référence à des processus d'apprentissage. "Erreur" n'est pas non plus synonyme de "faute", terme qui implique la présence d'un jugement moral (cf. Dictionnaire Larousse en ligne). James fait appel à la notion d'"écart" (*discrepancy*, 63 ; *deviance*, 64), définissant l'erreur comme la façon dont l'interlangue s'écarte des productions natives. Nous allons à présent nous intéresser aux différentes façons dont cet écart peut se manifester.

a. Grammaticalité et acceptabilité

La première intuition est de prendre en compte la grammaticalité d'un segment pour déclarer s'il est bien formé ou s'il est erroné. D'après James, ce critère devrait en théorie permettre de juger les segments de manière objective : "Appeal to grammaticality is an attempt to be objective, to take decisions such as whether some bit of language is erroneous or not out of the orbit of human whim" (65). L'intérêt d'utiliser ce critère est que la grammaticalité d'un segment peut être évaluée en dehors de tout contexte, puisqu'elle n'en dépend pas : si un segment est agrammatical, c'est-à-dire qu'il ne correspond pas aux codes de la grammaire, alors il n'existe aucun contexte dans lequel ce segment pourrait être jugé correct.

L'utilisation pratique de ce critère est pourtant problématique, pour au moins deux raisons. Tout d'abord, la grammaire d'une langue n'est pas monolithique, et il existe des variations en fonction du médium (oral vs. écrit), et des différentes variétés de la langue. C'est ce que souligne James dans la citation suivante : "Nobody would seriously think of designating the differences between, say, British English and any of the "colonial" varieties of English as "error"" (1990 ; 209). D'autre part, le critère de grammaticalité ne peut être utilisé que dans les cas ou le défaut de grammaticalité est net et non controversé, or ce genre de cas est

relativement réduit du fait des possibles variations que nous venons d'évoquer. La grammaticalité semble donc être un critère idéal à première vue, mais il ne permet d'identifier qu'un nombre restreint d'erreurs.

Par ailleurs, certains segments peuvent être parfaitement grammaticaux, mais être erronés d'un point de vue pragmatique, ou du point de vue de l'organisation de l'information dans la phrase. Ellis cite l'exemple suivant :

[9] *I want to read your newspaper.*

Cette phrase peut être interprétée comme une erreur si elle est énoncée à destination d'une personne inconnue dans une salle d'attente, par exemple (op. cit. : 48). Ce type d'erreur relève de l'acceptabilité plutôt que de la grammaticalité. James indique que, contrairement à la grammaticalité, l'acceptabilité d'un segment est évaluée en fonction de son contexte d'utilisation : "To decide on the acceptability of a piece of language we refer not to rules, but to contexts, trying to contextualize the utterance in question. […] [It] is seldom clear-cut and takes some thought, even imagination" (1998 : 67). Si la grammaticalité dépend du co-texte d'un mot ou segment, l'acceptabilité dépendrait donc de son contexte.

Il reste à définir ce qui constitue un énoncé acceptable ; Lyons en donne la définition suivante, fondée sur une acceptation réciproque en production et en réception par un public natif : "An acceptable utterance is one that has been, or might be, produced by a native speaker in some appropriate context and is, or would be, accepted by other native speakers as belonging to the language in question" (1968 : 137). Cette définition ne fournit cependant pas de critère plus précis qui permettrait d'expliquer le jugement du public natif. James en propose quant à lui huit ; nous en retenons trois pour leur intérêt dans notre recherche :

- le non-respect de collocations courantes et d'expressions, telles que :

  [10] *black and white*

  [11] *smiling like a Cheshire cat*

- la production de configurations grammaticales difficiles à traiter d'un point de vue cognitif, telles que dans le segment suivant extrait de notre corpus :

  [12] *?different access administration concept*

- la production de phrases ne respectant pas les normes d'équilibre et de présentation de l'information, comme dans l'énoncé suivant extrait de notre corpus :

  [13] *?A general definition of metaphors can be the following: […].*

Le critère de grammaticalité fait également partie de cette liste, puisqu'on peut supposer qu'un public natif jugera un énoncé comme étant inacceptable s'il ne correspond pas aux règles de grammaire de la variété d'anglais qu'il ou elle maîtrise. Il faut cependant se demander si un défaut de grammaticalité entraîne systématiquement un jugement d'inacceptabilité. James identifie quatre combinaisons possibles :

- [+Grammatical +Acceptable]
- [-Grammatical -Acceptable]
- [+Grammatical -Acceptable]
- [-Grammatical +Acceptable]

Parmi celles-ci, les deux premières ne sont pas problématiques, tandis que la troisième correspond au cas de figure que nous avons évoqué dans le paragraphe précédent. La quatrième configuration fait référence à des énoncés qui seraient jugés acceptables tout en étant reconnus comme agrammaticaux. D'après Milroy et Milroy (1985 : 74), cette configuration est possible, notamment pour des formules orales. Voici l'exemple qui est cité :

[14] *?This is the house that its roof fell in.*

L'interprétation de James concernant ce phénomène est la suivante : "The general rule that ungrammaticality precipitates unacceptability is relaxed by saying that what is deemed ungrammatical in written English may nevertheless be acceptable in spoken English" (op. cit. : 71). Nous y opposons deux interrogations. Pour commencer, l'acceptation d'un tel segment à l'oral n'est-elle pas le signe que la grammaire de la langue orale diffère de celle de la langue parlée sur ce point, et que ce segment est en réalité jugé comme grammatical dans ce médium ? D'autre part, comme nous l'avons souligné, le contexte d'énonciation d'un segment influence le jugement de son acceptabilité, et on peut se demander si de tels segments seraient jugés acceptables par un public natif s'ils étaient énoncés par une personne apprenante.

b. Compétence et performance

La première question que nous avions posée concernant la définition des erreurs était "De quoi l'erreur est-elle l'échec ?" (section 1.1.1b, "Identification des erreurs"). Si nous substituons la notion d'"écart" à celle d'"échec", cette question peut être reformulée ainsi : "De quelle nature est l'écart que constitue l'erreur ?". Dans les paragraphes précédents, nous avons repris, à la suite de James et d'autres chercheuses et chercheurs, les notions de grammaticalité

et d'acceptabilité afin de définir cet écart. Une autre problématique fréquemment évoquée concerne la distinction entre compétence et performance : l'écart symbolisé par une erreur se situe-t-il au niveau de la compétence d'une personne apprenante, ou bien de sa performance dans un contexte d'énonciation donné ? Dès son article de 1967, Corder a attiré l'attention sur l'importance de distinguer les erreurs dues à un manque de connaissances (*errors*), de celles dues à des facteurs extérieurs tels que la fatigue ou l'énervement, et qui relèvent de la performance (*mistakes*, que l'on peut traduire par "méprises") :

> We must therefore make a distinction between those errors which are the product of such chance circumstances and those which reveal his (*sic*) underlying knowledge of the language to date, or, as we may call it his transitional competence. The errors of performance will characteristically be unsystematic and the errors of competence, systematic. (1981 [1967] : 10)

Puisque chaque locuteur ou locutrice produit des méprises dans sa langue maternelle, Corder estime que ces dernières n'ont aucune signification pour l'apprentissage des langues, et que l'AE devrait se limiter à identifier les erreurs (au sens strict du terme, expliqué ci-dessus). Il est cependant très problématique de différencier ces deux types d'écart, et dans son article de 1971, Corder propose d'utiliser la faculté d'auto-correction comme critère de distinction, arguant du fait que les erreurs de performance sont en général facilement corrigées par la personne les ayant produites (1981 [1971] : 18). Ce critère est cependant difficile à appliquer en pratique, car d'une part il nécessite de pouvoir interagir avec les auteurs des productions, et d'autre part il n'est pas jugé comme fiable d'après James :

> It is well known that people who compose in a [foreign language], in an exam for instance, can be given ample time to review and monitor their text […] but still overlook ungrammaticalities that they would be immediately and effortlessly able to self-correct if only the fact that such and such a form is wrong were pointed out to them. In that case the question of whether the learners were able to auto-correct is 'yes' and 'no'. (op. cit. : 79)

Cette réflexion l'amène à proposer une classification en quatre catégories :

- les *lapsus* (*slips*), qui sont facilement corrigés de manière autonome par l'auteure ou auteur de la production,
- les *méprises*, qui peuvent être corrigées après que la présence d'une anomalie a été indiquée (méprise de premier ordre), ou bien avec plus d'indications, comme sa localisation précise (méprise de second ordre),

- les *erreurs*, qui ne peuvent pas donner lieu à une auto-correction sans apprentissage supplémentaire,
- les *solécismes*, ou "erreurs" liées à l'infraction d'une règle de grammaire prescriptive, dont l'application n'est pas forcément justifiée.

Le principe fondamental de la distinction compétence/performance, c'est-à-dire le fait que l'on puisse en théorie distinguer erreurs et méprises puisque les erreurs de compétence sont systématiques, est également critiqué. Ellis souligne que ce principe implique que la compétence soit homogène, alors qu'il est possible qu'elle soit variable et dépende des contextes, notamment en ce qui concerne les combinaisons lexicales, dans lesquels les structures sont produites (2008 : 48).

La question de la distinction entre erreurs et méprises a récemment été étudiée dans le cadre de l'enseignement des langues assisté par ordinateur. Thouësny est l'auteure d'une thèse intitulée "Modeling second language learners' interlanguage and its variablity: A computer-based dynamic assessment approach to distinguishing between errors and mistakes" (2011). Son étude repose sur la mise en place d'un système informatisé permettant l'évaluation dynamique d'un groupe de personnes apprenantes du français, dans le but de distinguer erreurs et méprises. L'évaluation dynamique a pour objectif de mener l'apprenant ou apprenante à améliorer sa performance lors de l'évaluation, par exemple en lui fournissant une aide extérieure. La distinction entre erreurs et méprises est effectuée à partir de la capacité de la personne apprenante à proposer une solution alternative et à corriger sa propre production avec ou sans l'assistance proposée. Une des ses principales conclusions concernant la stabilité de la compétence du public apprenant est que celle-ci est aussi variable que la performance, et que l'évolution des connaissances dans le temps est systématique comme non-systématique. Les erreurs de compétence peuvent donc être systématiques mais aussi imprévisibles, contrairement à la déclaration de Corder à ce sujet. Thouësny conclut également qu'il est possible de distinguer erreurs et méprises, mais que l'interaction de la personne apprenante avec le système lors de l'évaluation est un processus nécessaire pour établir cette distinction.

Pour clore notre présentation de la distinction entre erreurs et méprises, nous souhaitons revenir sur l'affirmation de Corder selon laquelle les méprises ne devraient pas être prises en compte dans le cadre de l'AE, et ne sont pas pertinentes pour l'étude de l'acquisition des langues en général. Nous ne souhaitons pas nous prononcer sur la dernière partie de cette affirmation, mais nous partageons l'avis de James concernant l'intérêt de prendre en compte

les méprises d'un point de vue pédagogique. Puisque les méprises peuvent être assez facilement corrigées par les apprenants et apprenantes, leur correction représente le moyen le plus rapide pour améliorer leurs productions : "Mistakes, by contrast, can be attended to: feedback can be given, the learners can learn how to monitor, and opportunities for further practice can be provided" (James, 1998 : 86). Thouësny, comme James, souligne le fait que les apprenantes et apprenants ont une compétence plus étendue que ce que reflètent leurs performances ; leur permettre d'exprimer leurs compétences réelles dans leurs productions en corrigeant les méprises, par exemple de manière automatisée, ne nous semble pas être un objectif dénué d'intérêt.

c. La visibilité des erreurs

Un autre aspect à prendre en compte lors de la définition et de l'identification des erreurs (au sens général) est leur visibilité. Ce paramètre a été relevé par Corder dans son article de 1971. Les erreurs visibles (*overt errors*) sont facilement identifiables car elles constituent un écart net en termes de grammaticalité ou d'acceptabilité, comme par exemple le segment suivant, relevé dans notre corpus :

[15] **Why should we loss our identity?*

Les erreurs invisibles (*covert errors*) sont des segments grammaticaux mais inacceptables, non pas à cause d'une rupture de collocation ou d'organisation de la phrase, mais à cause du fait qu'ils ne signifient pas ce que la personne apprenante avait l'intention de dire. Le segment suivant est un exemple d'erreur invisible, car le sujet *It* fait référence au vent (exemple emprunté à Corder, 1981 [1971] : 21) :

[16] *It was stopped.*

Cette distinction rejoint la remarque de James concernant l'importance de la notion d'intention dans la définition des erreurs : "an error arises only when there was no intention to commit one" (op. cit. : 77). Dans le cas des erreurs invisibles, la déviance du segment est mesurée en fonction du message que la personne apprenante avait l'intention de transmettre. Ces erreurs sont problématiques en particulier car elles créent une difficulté de communication qui, contrairement au cas des erreurs visibles, peut passer inaperçue :

> These [miscommunications] are serious, not only because they unintentionally misinform, but also because they tend to go unnoticed, by speaker and listener alike. They can become the basis for cumulative [miscommunication] and alienation in the 'communicative' process (op. cit. : 218).

d. Éléments de réponse et synthèse

Nous sommes à présent en mesure de fournir des réponses plus détaillées aux questions que nous avons soulevées dans la section 1.1.1b. La première question concernait la nature des critères permettant de juger qu'un segment est erroné. Les deux critères à retenir sont la grammaticalité, liée aux règles de la grammaire et non-tributaire du contexte, et l'acceptabilité, entièrement dépendante du contexte et liée à un ensemble d'usages reconnus par le public natif. Ces deux critères entrent le plus souvent en combinaison, l'agrammaticalité d'un segment déclenchant le plus souvent son inacceptabilité. L'intention de la personne apprenante lors de la production du segment est également un facteur à prendre en compte lors de la détection des erreurs, puisqu'on ne peut juger qu'un segment est erroné que si son auteure ou auteur n'avait pas l'intention de produire une erreur. Il faut en effet laisser la liberté aux personnes apprenantes d'user de licence poétique ou d'humour dans leurs productions de temps en temps, au moins en théorie.

La seconde question portait sur le choix des normes utilisées pour juger qu'un segment est erroné. Cette question est particulièrement problématique car il ne semble pas exister de norme qui puisse s'appliquer uniformément à toute production. Comme nous l'avons vu, le critère de grammaticalité, qui est le plus objectif, est lui aussi soumis à des variations en fonction du médium, du niveau (formel vs. informel) et des variétés de la langue utilisés. L'acceptabilité est jugée du point de vue des normes natives, qui sont soumises à ces mêmes variations.

Dans un article consacré à une réflexion autour de la définition, de l'identification et de la distinction des erreurs, Lennon incorpore la notion de variabilité de la norme ainsi que l'importance du contexte de production dans sa définition de l'erreur : "A linguistic form or combination of forms which, in the same context and under similar conditions of production, would, in all likelihood, not be produced by the speaker's native speaker counterparts" (1991 : 182). La formulation de cette définition, qui propose en fin de compte de "comparer ce qui est comparable", permet de limiter l'influence des variations de norme sur le jugement d'erreur, mais son application pratique est compliquée car les conditions exactes de production ainsi que le groupe exact de locuteurs et locutrices sont difficiles à identifier et à reproduire.

L'utilisation de la norme native pour détecter les erreurs est également considérée comme problématique pour d'autres raisons que sa variabilité. Nous avons déjà mentionné le concept d'interlangue et l'hypothèse qui lui est associée (Selinker, 1972), et qui met l'accent sur l'étude descriptive et non-comparative des productions des personnes apprenantes, considérées

comme représentatives de leur propre version de la langue cible. Corder (1981 [1971]) a identifié le même concept sous le nom de "dialecte idiosyncratique" (*idiosyncratic dialect*). Les approches de l'acquisition des langues secondes fondées sur ce concept ont rejeté la comparaison systématique des productions des personnes apprenantes à celles de personnes natives, ainsi que le concept même d'"erreur" comme manifestation d'un écart de compétence entre ces deux publics (Corder, 1981 [1971] : 18-19), proposant de réserver ce terme à l'identification des erreurs de performance.

Il existe d'autres objections à l'utilisation de normes natives pour l'évaluation des productions de personnes apprenantes dans le cas de l'anglais. La branche de l'acquisition des langues secondes fondée sur la conception de l'anglais comme *lingua franca* (*English as a Lingua Franca*), et que nous présentons plus longuement dans la sous-partie 1.1.4, rejette également le recours à ces normes pour des raisons scientifiques tout autant qu'éthiques. Dans un monde qui comporte environ quatre fois plus de personnes utilisatrices de l'anglais que de personnes dont c'est la langue maternelle (Kachru, 1996 : 241), l'utilisation systématique des normes natives est envisagée comme le signe d'une hiérarchie entre ces différents groupes, et d'un manque de prise en compte des règles communicatives développées par les personnes utilisatrices de l'anglais (Seidlhofer, 2011). Nous reviendrons sur cette question et sur les conséquences de sa prise en compte dans notre recherche dans la sous-partie 1.1.4.

La troisième question posée en 1.1.1b portait sur la délimitation des segments erronés. Lennon propose de les délimiter selon deux dimensions : l'étendue (*extent*) et le domaine (*domain*). Le domaine est défini ainsi : "the rank of the linguistic unit which must be taken as context in order for the error to become apparent," et constitue un moyen d'identifier la quantité de contexte qu'il est nécessaire de prendre en compte pour attester de la présence d'une erreur. L'étendue est définie ainsi : "the rank of the linguistic unit, from minimally the morpheme to maximally the sentence, which would have to be deleted, replaced, reordered, or supplied in order to repair production", et fait référence à la profondeur de l'erreur en termes de hiérarchie linguistique (1991 : 191). Lennon ajoute l'explication suivante, qui donne un éclairage différent sur la distinction entre domaine et étendue :

> Another way of looking at 'domain' and 'extent' is to regard 'domain' as reflecting the hearer's perspective, and extent the speaker's. Domain refers to the amount of (linguistic or non-linguistic) context the hearer needs to recognize the error. Extent refers to the amount of linguistic context which the speaker needs to refashion in order to repair the error. (op. cit : 191)

Lennon précise que le domaine d'une erreur relèvera toujours d'un rang supérieur ou égal à celui de son étendue. Par exemple, pour l'erreur **an advice*, tirée de notre corpus, l'étendue est l'article *an*, et le domaine est le groupe nominal en entier. Nous reprenons cette distinction lors de la détection des erreurs ainsi que dans la mise en place du schéma d'annotation présenté dans la sous-partie 1.2.3.

Concernant la visibilité des erreurs, sur laquelle porte la quatrième et dernière question, nous avons vu que certaines erreurs sont visibles car clairement identifiables comme des segments agrammaticaux et/ou inacceptables, alors que d'autres erreurs peuvent ne pas être détectées car les segments sont grammaticaux. L'erreur réside alors dans l'écart entre ce que l'apprenant ou apprenante souhaitait communiquer, et ce qui est en réalité communiqué par leur énoncé.

1.1.3 Les causes des erreurs

Dans le paragraphe intitulé "Evaluation des erreurs" de la section 1.1.1b, nous avons rappelé que les causes des erreurs sont généralement attribuées à deux types de phénomène : les phénomènes d'interférence et les phénomènes liés au développement de la compétence linguistique. Il est souvent difficile d'attribuer une erreur strictement à un de ces phénomènes (cf. Ellis, 2008 ; Jarvis et Pavlenko, 2007), c'est une des raisons pour lesquelles nous ne tenterons pas de donner une évaluation complète des erreurs relevées dans notre corpus. Avoir des informations concernant les causes probables des erreurs que nous choisissons de traiter en correction automatisée sera néanmoins très utile lors de la création des messages correctifs. Nous présentons plus en détail dans les paragraphes suivants le phénomène du transfert et celui de la surgénéralisation, qui semblent être des facteurs dans la production des erreurs traitées ici.

a. Le rôle du transfert dans la production d'erreurs

Ellis donne la définition suivante des erreurs de transfert : "These are errors in learner language that can be accounted for in terms of differences between the structures of the L1 and the L2" (2008 : 981). L'analyse contrastive, que nous avons déjà évoquée, était fondée sur l'hypothèse selon laquelle tout ou partie des erreurs serait prévisible grâce à l'observation des différences et ressemblances entre L1 et L2. Cette approche a été largement critiquée, mais il semble à présent que s'il est difficile, voire impossible, de prédire qu'une erreur de transfert sera produite à partir de l'observation des langues en présence, il est tout aussi impossible de

faire la prédiction que ce type d'erreur ne sera pas produit (Odlin, 1989 : 157). L'intérêt porté au phénomène du transfert concerne donc désormais l'observation des ses manifestations en tant que sujet d'étude à part entière plutôt que comme facteur d'explication des erreurs.

Il est admis que les manifestations de ce phénomène sont en réalité beaucoup plus diversifiées, ce que reflète la définition volontairement vague qui est donnée du phénomène dans la monographie la plus récente consacrée à ce sujet : "The influence of a person's knowledge of one language on that person's knowledge or use of another language" (Jarvis et Pavlenko, 2007 : 1). D'un point de vue terminologique, les termes de "transfert" et "influence translinguistique" sont utilisés indifféremment par Ellis et Jarvis et Pavlenko, même si le second terme est considéré comme plus adapté, le terme de "transfert" étant hérité des théories béhavioristes de l'acquisition des langues.

Afin de s'orienter dans la littérature abondante sur le phénomène du transfert et également de rendre visible la diversité de ses manifestations, Jarvis et Pavlenko ont développé une schématisation en dix dimensions de ce phénomène, qui permet d'identifier de manière très précise la nature de la manifestation étudiée. Ces dix dimensions sont présentées dans le tableau 2 ; la colonne de droite donne les différentes valeurs ainsi que des précisions sur l'intitulé de la dimension en question s'il n'est pas jugé transparent.

Dimension	**Valeurs**
Aire d'utilisation ou de connaissance de la langue	**Phonologie, orthographe, lexique, sémantique, morphologie, syntaxe, discours, pragmatique, sociolinguistique**
Direction	Le transfert peut s'effectuer de la L1 à une L2 (*forward*, **vers l'avant**), d'une L2 à la L1 (*reverse*, **à rebours**), entre deux L2 (**latéral**), ou aller dans plusieurs directions en même temps (**bi- ou multi-directionnel**)
Niveau cognitif	Le transfert peut concerner uniquement le niveau **linguistique**, ou bien découler d'un transfert **conceptuel** entre les deux langues
Type de savoir	La connaissance d'une L2 peut être intuitive et tacite (**savoir implicite**), ou bien il peut s'agir de connaissances des règles et structures précises (**savoir explicite**)
Intentionnalité	**Intentionnel**, **non-intentionnel**
Mode	**Productif, réceptif**
Canal	**Audio/oral, visuel/écrit**
Forme	**Verbal, non-verbal**
Manifestation	**Visible, invisible**
Résultat	Le transfert peut avoir des conséquences **positives** sur la maîtrise de la langue cible (ex. : facilitation de la compréhension), ou des conséquences **négatives** (ex. : erreurs)

Tableau 2. Caractérisation de l'influence translinguistique en dix dimensions

D'après cette schématisation, le type de transfert auquel nous nous intéressons et que nous sommes susceptibles d'observer dans notre corpus d'erreurs a les caractéristiques suivantes : il touche le domaine de la syntaxe, constitue un transfert vers l'avant (de la L1, le français, à la L2, l'anglais) non-intentionnel et visible à l'écrit, donc sous forme verbale, et son résultat est jugé négatif puisqu'il a pour résultat la production d'erreurs. Nous ne nous prononcerons pas sur le niveau cognitif et le type de savoir concernés car l'étude de ces phénomènes ne fait pas partie de notre cadre de recherche.

Les recherches sur le phénomène du transfert se sont également penchées sur les facteurs pouvant le déclencher ou avoir une influence sur ses manifestations. Avant de donner plus de détails sur ces facteurs, précisons que la nature du transfert qui est étudiée ici est celle qui a trait aux règles d'apprentissage, c'est-à-dire au type de transfert qui se produit lorsque la personne apprenante utilise ses connaissances concernant une langue afin de formuler des hypothèses quant aux structures, aux significations, aux règles, etc. d'une autre langue. On identifie un autre type de transfert, utilisé comme stratégie de communication, qui consiste à utiliser sciemment une forme connue pour pallier un manque de connaissance ponctuel. Le phénomène du transfert a d'ailleurs été longtemps réduit à cette seule manifestation (Jarvis et Pavlenko, 2007 : 9).

Jarvis et Pavlenko font une présentation détaillée de l'ensemble des facteurs interagissant avec le transfert. Nous n'en mentionnons qu'une sélection, choisie pour sa pertinence dans les phénomènes que nous analysons. L'un des facteurs linguistiques et psycholinguistiques les plus reconnus est le degré de congruence existant entre la langue source et la langue de réception du transfert. Ce facteur est également appelé "similarité translinguistique" (*crosslinguistic similarity*, op. cit. : 176), terme qui englobe les ressemblances et les différences entre les deux langues.

La similarité existe de manière objective, c'est-à-dire qu'elle relève de différences et ressemblances réelles, ou de manière subjective, lorsque le degré de congruence est celui qui est apprécié par la personne apprenante (op. cit. : 177). Cette dernière diffère de la similarité objective car elle est susceptible d'évoluer dans le temps, selon l'avancement de la maîtrise de la langue. C'est la similarité subjective qui est en réalité source de transfert ; la similarité objective n'intervient pas dans le déclenchement du phénomène mais détermine si ses conséquences sont positives (si la similarité existe réellement et constitue une facilitation) ou négatives (si elle n'existe pas et cause donc une erreur) (op. cit. : 179). De plus, ce sont généralement les ressemblances subjectives, plutôt que les différences, qui mènent au

transfert, car les apprenantes et apprenants tendent à rechercher les analogies et à éviter de transférer des structures qu'ils jugent comme étant très différentes.

La similarité subjective entre deux langues est souvent éloignée de la similarité objective, et cet écart peut être le résultat de trois situations (op. cit. : 178) :

- la personne apprenante ne reconnaît pas les ressemblances et différences qui existent réellement,
- elle interprète de manière inadéquate la nature de certaines ressemblances ou différences,
- elle postule l'existence de ressemblances ou différences qui n'existent pas en réalité.

On distingue deux types de similarité subjective : la similarité perçue, qui résulte d'une observation consciente ou inconsciente quant à une ressemblance entre les deux langues, et la similarité présumée, qui relève d'une hypothèse consciente ou inconsciente ne tenant pas compte d'une observation réelle. Ces deux types de similarité subjective ont des conséquences différentes en termes de transfert : certains types de transfert, comme dans les transferts pragmatiques et sémantiques, peuvent prendre leur source dans une similarité présumée, alors que d'autres types plus formels, comme les transferts syntaxiques, résultent de similarités perçues (op. cit. : 179). Ces deux types de similarité ne sont pas mutuellement exclusifs et travaillent souvent de concert : lorsque les ressemblances perçues entre deux langues atteignent un certain seuil, elles amènent la personne apprenante à présumer de l'existence d'autres ressemblances. Les transferts négatifs liés à l'utilisation de formes ou de structures (par exemple dans le domaine de la syntaxe) semblent ainsi se produire plus fréquemment lorsque les deux langues ont été perçues comme globalement similaires (op. cit. : 181).

Si la similarité translinguistique et les différentes perceptions de son existence sont des facteurs importants dans l'apparition du phénomène du transfert, ils entrent en interaction avec un grand nombre d'autres facteurs. Nous ne les citerons pas tous, mais retenons les facteurs suivants : la fréquence des formes dans la L1 et la L2, le niveau de contrôle exercé lors de la production et le savoir explicite concernant la langue, et le niveau de maîtrise de la L2.

Concernant la fréquence des formes dans la L1 et la L2, nous attirons particulièrement l'attention sur les résultats de l'étude de Selinker sur le placement des "adverbiaux" (*adverbials* ; nous reprenons la terminologie de Selinker même si le terme de *adjunct* semblerait plus indiqué), qui montre que des personnes locutrices de l'hébreu apprenantes de l’anglais manifestaient une préférence pour un placement en particulier, et qui était le résultat

d'un transfert de l'hébreu à l'anglais (Selinker, 1969 ; cité par Jarvis et Pavlenko). La fréquence de certaines structures dans la L1 peut donc avoir un effet sur leur utilisation dans la L2.

En ce qui concerne le savoir explicite sur la langue, son influence sur le transfert est visible dans le fait que les productions des personnes apprenantes ne présentent pas les mêmes manifestations de ce phénomène lorsque ces dernières utilisent ce savoir lors de productions sur lesquelles elles exercent un contrôle conscient. Odlin suggère que les contextes dans lesquels l'attention à la langue est importante (*focused environments*) donnent lieu à des productions comportant moins d'erreurs provenant d'un transfert négatif (1989 : 146).

L'influence du niveau de maîtrise de la L2 sur le transfert a été étudiée dans le but de découvrir si celle-ci tendait à disparaître ou à se modifier avec l'avancement de la compétence. Les résultats des recherches sont fortement partagés à ce sujet, notamment car le facteur de la similarité translinguistique interagit avec celui de la compétence dans la L2, puisque l'observation des ressemblances et différences évolue avec l'augmentation des connaissances sur la L2. Même s'il est reconnu que le niveau de maîtrise est un facteur influençant les manifestations du transfert, peu de conclusions fiables sont disponibles à ce sujet (Jarvis et Pavlenko, 2007 : 203).

En raison des types d'erreur sélectionnés pour notre recherche, les manifestations du transfert qui retiennent notre attention sont principalement celles qui concernent le domaine de la syntaxe. L'importance, et même l'existence du transfert dans ce domaine ont fréquemment été remises en question, en particulier car ses manifestations fonctionnent souvent en interaction avec d'autres variables, comme la surgénéralisation. Jarvis et Pavlenko indiquent cependant que l'observation de ce type de transfert a en réalité été largement documentée. Parmi les phénomènes syntaxiques étudiés, celui qui semble le plus direct et le plus facilement observable est le placement des adverbes et adverbiaux. Alonso (2002, citée par Jarvis et Pavlenko, op. cit. : 99) montre que le transfert est la principale source de placements erronés d'adverbiaux en anglais oral chez des personnes ayant l'espagnol pour L1, mais observe que ce phénomène est moins fréquent dans les productions contrôlées, ainsi que dans les productions de locuteurs et locutrices avancés. Osborne (2008) montre cependant que ce phénomène touche aussi les productions écrites contrôlées de francophones apprenantes et apprenants à un niveau avancé. Nous reviendrons plus longuement sur cette étude dans la section 2.2.2a.

b. Le rôle de la surgénéralisation dans la production d'erreurs

À notre connaissance, le rôle de la surgénéralisation dans la production d'erreurs dans une L2 a été beaucoup moins étudié que celui du transfert. Ce phénomène en lui-même semble avoir suscité moins d'intérêt que l'influence translinguistique, et n'est pas un objet d'étude à part entière. Comme son nom l'indique, la surgénéralisation fait référence à l'application générique d'une règle à des éléments qui ne sont en réalité pas régis par cette dernière (Ellis, 2008 : 974). Richards en donne la définition suivante : "Overgeneralization covers instances where the learner creates a deviant structure on the basis of his *[sic]* experience of other structures in the target language" (1974 [1971] : 174). Les erreurs découlant de ce phénomène appartiennent à la catégorie des erreurs liées au développement de la compétence linguistique (et non à l'influence de la L1), et la surgénéralisation est étudiée comme une facette de ces processus de développement. Ce phénomène est également observable dans l'acquisition de la langue maternelle.

L'exemple le plus fréquemment cité de surgénéralisation observable, que ce soit dans le domaine de l'acquisition d'une L2 ou de la L1, est celui de la régularisation des verbes irréguliers en anglais (cf. Ellis, 2008 : 358 ; Tarone, 2006 : 749 ; Selinker, 1974 [1972] : 38), lors de laquelle des verbes comme *drink* et *eat* sont surgénéralisés par analogie en *drinked* et *eated*. Elle est souvent combinée à d'autres processus liés au développement de la compétence, comme la simplification, l'utilisation trop fréquente (*overuse*), et l'évitement de certaines structures. L'utilisation erronée d'un nom en fonction de modifieur d'un autre nom peut par exemple relever de la surgénéralisation d'une structure courante en anglais mais non applicable à tous les groupes nominaux, comme dans les segments suivants issus de notre corpus :

[17] **money inequality*

[18] **his love obsession*

On peut également supposer que la stratégie de surgénéralisation est ici combinée avec l'évitement de l'utilisation d'une préposition. Nous reviendrons sur les causes possibles de ce type d'erreur, car il s'agit d'une des erreurs que nous proposons de traiter.

Richards avance l'hypothèse que la surgénéralisation serait un moyen pour la personne apprenante de limiter sa charge cognitive lors de l'apprentissage (op. cit. : 174). Tarone souligne pour sa part l'aspect positif de la présence d'erreurs de surgénéralisation dans une production : "The overgeneralization error shows clear evidence of progress, in that it shows

that the learner has mastered a target language rule" (op. cit.). Étant le signe de l'existence d'une compétence, l'observation d'erreurs liées à la surgénéralisation donne donc des indications sur une éventuelle opportunité d'apprentissage.

1.1.4 L'anglais comme *lingua franca* : un paradigme alternatif

L'objectif principal de cette thèse est de proposer des règles de détection et de correction pour des erreurs produites à l'écrit par des personnes utilisatrices de l'anglais. Les sections précédentes ont présenté le cadre théorique de l'analyse des erreurs, qui s'inscrit dans le domaine de l'acquisition des langues secondes. Ce cadre, qui peut être appliqué à l'étude de n'importe quelle L2 ou paire L1-L2, s'appuie sur une vision de l'apprentissage des langues secondes comme étant un cheminement d'un stade initial de maîtrise inexistante ou limitée de la L2 à un stade final de maîtrise proche de celle des personnes locutrices natives. La notion d'interlangue est représentative de ce cadre, car elle désigne le stade intermédiaire atteint par une personne apprenante à un moment donné. Même si certaines approches sont réticentes à considérer les productions de ces publics comme pouvant être erronées, la compétence native reste la norme en fonction de laquelle les productions sont évaluées.

Cependant, l'anglais a indéniablement un statut particulier dans le monde, en ceci qu'il est largement reconnu comme une langue internationale. Comme nous l'avons mentionné plus haut, le nombre de personnes utilisatrices de l'anglais est à ce jour très supérieur au nombre de personnes l'ayant comme langue maternelle. Dans le but de répondre à notre exigence de pertinence et d'utilité du point de vue des publics visés, il est nécessaire de prendre en compte le statut particulier de la langue que nous étudions, ainsi que les approches développées autour de la reconnaissance de ce statut et des conséquences qu'il a sur son utilisation, son apprentissage et son enseignement, et sur les normes qui sont utilisées dans ces différents contextes. Cela est d'autant plus important que les erreurs que nous choisissons de traiter interrogent les normes grammaticales, puisqu'elles correspondent à des phénomènes qui ne sont pas clairement codifiés. Dans ce cadre, la question du bien-fondé de leur correction doit être posée et résolue.

Nous présentons dans cette section l'approche de l'anglais comme *lingua franca* (ALF, *English as a Lingua Franca*), et nous nous appuyons sur l'ouvrage de Barbara Seidlhofer consacré à la présentation de cette approche, *Understanding English as a Lingua Franca* (2011). Nous abordons la façon dont ces constatations sont prises en compte dans le cadre de notre recherche dans la section 1.1.4d.

a. L'anglais, langue internationale

Le modèle le plus répandu de l'utilisation de l'anglais dans le monde est celui des "cercles concentriques" de B. Kachru (1985). À partir de données géographiques et historiques, ce modèle identifie trois cercles :

- le Cercle central (*Inner Circle*),
- le Cercle extérieur (*Outer Circle*),
- le Cercle en expansion (*Expanding Circle*).

Le Cercle central regroupe les populations ayant l'anglais pour langue maternelle et étant issues des premières vagues de dispersion de l'anglais dans le monde, comme celles du Royaume-Uni, de l'Irlande, des États-Unis, du Canada anglophone, de l'Australie, de la Nouvelle-Zélande, de l'Afrique du Sud et d'autres zones comme les Caraïbes anglophones.

Le Cercle extérieur concerne les populations dont la langue maternelle n'est pas l'anglais, mais qui l'utilisent néanmoins pour la communication intra-nationale, notamment suite à leur colonisation par le Royaume-Uni (Inde, Nigéria, Singapour, etc.).

Le Cercle en expansion fait référence à l'ensemble des populations dont la langue maternelle n'est pas l'anglais et qui ne l'utilisent pas pour communiquer à l'intérieur du pays, mais pour qui il s'agit du moyen privilégié de communication internationale. La France, ainsi que la plupart des autres pays (Chine, Russie, Japon, Allemagne, etc.) fait partie du Cercle en expansion.

D'après Crystal, le Cercle central comprendrait de 320 à 380 millions de personnes et le Cercle extérieur en comprendrait de 300 à 500 millions. Le nombre de personnes pouvant être incluses dans le Cercle en expansion est difficile à évaluer, car il est impossible d'identifier exactement ce qui constitue une personne utilisatrice de l'anglais tant les utilisations et les moyens d'acquisition sont variés, mais Crystal estime qu'il peut concerner jusqu'à un milliard de personnes (2003 : 69).

L'état de fait décrit par ce modèle a valu à l'anglais le titre de "langue internationale", ce qui signifie qu'il est utilisé pour communiquer à l'intérieur des cercles décrits par Kachru mais également entre ces cercles. Cette expression décrit une fonction de la langue plutôt qu'une variété : l'anglais comme langue internationale est à distinguer des "anglais du monde" (*World Englishes*), variétés de l'anglais utilisées par les populations du Cercle extérieur. Les anglais du monde ne sont pas des langues internationales, mais le résultat d'une évolution locale de la

langue après son exportation (Seidlhofer, 2011 : 3). L'expression "anglais comme langue internationale" fait référence à une utilisation mondialisée, non localisée, de l'anglais.

L'expression "anglais comme *lingua franca*" est parfois considérée comme un synonyme de "anglais comme langue internationale" (cf. Seidlhofer, 2003), mais elle fait référence à l'utilisation de l'anglais internationalement avec une perspective différente. Seidlhofer donne la définition suivante de l'ALF :

> [A]ny use of English among speakers of different first languages for whom English is the communicative medium of choice, and often the only option. Due to the numbers of speakers involved worldwide, this means that ENL [English as a Native Language] speakers will generally be in a minority. (op. cit., 2011 : 7)

Le terme *lingua franca*, ou "langue franche", renvoie à l'utilisation d'une langue véhiculaire composite dans des situations de contact entre personnes n'ayant pas la même langue maternelle. Pour comparaison, la définition donnée par Ellis de l'anglais comme langue internationale est moins précise, et elle laisse une moins grande ouverture à la question des normes natives dans la communication entre personnes non-natives : "[English as an International Language] is used to refer to the use of English across a wide range of contexts throughout the world" (Ellis, 2008 : 960).

Il apparaît clairement que l'anglais n'est pas une langue étrangère "comme les autres", ce qui a des conséquences sur son évolution :

> The ownership (by which I mean the power to adapt and change) of any language in effect rests with the people who use it, whoever they are, however multilingual they are, however monolingual they are. […] Statistically, native speakers [of English] are in a minority for language use, and thus in practice for language change, for language maintenance, and for the ideologies and beliefs associated with the language. (Brumfit, 2001 : 116)

Cette proposition peut paraître radicale, mais Seidlhofer insiste sur le fait que si l'ALF ne se résume pas à une utilisation mondialisée de l'anglais comme langue maternelle (ALM, cf. *English as a Native Language*, op. cit. : 5), comme nous l'expliquons dans les sections suivantes, l'ALM n'a pas non plus vocation à être influencée par l'ALF, chacune des deux versions de la langue pouvant conserver et développer leurs propres caractéristiques. L'avantage de cette approche est qu'elle permet de dénouer les problématiques qui entourent l'avancée de l'utilisation de l'anglais dans le monde, la question de la diversité linguistique et la domination des cultures anglophones :

> [O]nce one denies this right of exclusive ownership and dissociates the language from its native speakers and recognizes it as a partial and expedient resource that anyone can make use of – in other words, once one thinks of English as [English as a Lingua Franca], then the language obviously no longer poses the same threat of domination. […] [L]inguistic diversity is more likely to be protected by a new conceptualization of English as ELF. (op. cit. : 68).

Cette reconceptualiation de l'anglais dans ce contexte amène cependant à se poser la question des normes de grammaticalité et d'acceptabilité qui sont utilisées pour évaluer les productions de personnes utilisatrices de l'anglais dans un cadre pédagogique ou professionnel. Ces questions sont d'une grande importance pour une recherche centrée sur la correction grammaticale automatisée à destination de personnes utilisatrices de l'anglais, et particulièrement si l'on prend en compte leur utilisation réelle de la langue (voir 1.2.1b) qui correspond à celle de l'anglais comme *lingua franca*.

b. Quelles normes pour l'anglais comme lingua franca *?*

La compétence linguistique des personnes locutrices natives de l'anglais est le point de repère par excellence pour identifier les règles et normes de la langue. Nous avons vu que dans le domaine de l'acquisition des langues secondes, cette compétence sert de critère pour fixer l'objectif idéal à atteindre par les apprenants et apprenantes. L'idiomaticité et l'authenticité du point de vue de l'ALM sont vues comme des caractéristiques positives que ces personnes doivent chercher à atteindre dans leurs productions.

Seidlhofer attribue notamment l'adoption de la compétence native comme point de repère à l'influence de la linguistique générative, approche qui repose en grande partie sur la notion de compétence absolue des personnes locutrices natives d'une langue (op. cit. : 89). Précisons cependant, comme le fait Seidlhofer, que Chomsky faisait reposer ce concept sur un type de personne locutrice et des circonstances qu'il jugeait lui-même comme "idéales" et donc peu représentatives de situations réelles (cf. Chomsky, 1957). Les normes utilisées par le public locuteur natif de l'anglais sont en réalité difficiles à identifier : les jugements de grammaticalité et d'acceptabilité peuvent ne pas être homogènes même à l'intérieur d'un groupe linguistique identifié (Seidlhofer, 2011 : 90). La constitution et l'analyse de corpus linguistiques informatisés permettent néanmoins d'apporter une solution à ce problème.

À ce stade, il est nécessaire de souligner que l'enseignement, l'apprentissage et l'utilisation de l'anglais ne rentrent pas automatiquement dans le cadre de l'ALF ou de l'anglais comme

langue internationale (ALI). Comme toute langue, l'anglais peut être étudié comme une langue étrangère, dans le cas où l'élève porte un intérêt à la culture d'un ou plusieurs pays anglophones, pour des raisons diverses. C'est d'ailleurs le cadre proposé par les filières universitaires françaises en Langues, Littératures et Civilisations du Monde Anglophone, dont l'auteure de cette thèse est issue. La mise en avant des normes de l'ALM pour l'apprentissage et l'évaluation des productions écrites comme orales est dans ce cas entièrement justifiée. Cependant, au vu des chiffres donnés dans la section précédente, il est logique d'envisager qu'une part importante des personnes utilisatrices de l'anglais ont pour principal objectif de communiquer efficacement avec un public international. Pour ce genre de public, qui est celui que nous visons majoritairement ici (voir 1.2.1b), il est utile de réévaluer la pertinence de l'utilisation de normes natives comme points de repère, pour plusieurs raisons.

Commençons par la raison la plus évidente : pour atteindre les standards d'idiomaticité, d'acceptabilité et de grammaticalité de l'ALM en production comme en réception, il est nécessaire de maîtriser un grand nombre d'exceptions et particularités (par exemple, les verbes irréguliers ou les proverbes), ce qui rend la langue plus difficile à apprendre et à utiliser (op. cit. 55). Cette exigence pourrait cependant être légitime si ces particularités étaient requises pour une communication efficace, ce qui n'est apparemment pas le cas : "[C]ommunicative efficiency and grammatical correctness are not taken as being in a straightforward relationship whereby the former can be guaranteed through the latter" (op. cit. : 99). Les erreurs grammaticales sont souvent considérées comme problématiques car elles risquent de gêner la compréhension des productions (cf. "Evaluation des erreurs", 1.1.1b), argument que Seidlhofer nuance, arguant du fait que l'intelligibilité d'un message repose sur un ensemble de facteurs bien plus large que la simple correction linguistique (op. cit. : 35). De plus, certaines normes de l'ALM, comme la recherche d'idiomaticité, peuvent même avoir un effet négatif sur l'intelligibilité d'un message et donc aller à l'encontre de l'objectif de communication efficace, comme par exemple lorsque les locutrices ou locuteurs choisissent d'utiliser des verbes à particules ou des expressions dont le sens est non-compositionnel (*to chill out*, *in the nick of time*, etc.) (op. cit. : 134). Ces observations amènent Seidlhofer à faire la constatation suivante : "There is now empirical research that shows that there is no generally valid, direct relationship between communicative effectiveness and correctness in terms of the norms of native speakers" (op. cit. 53).

Cette constatation permet d'identifier les normes qui ne sont pas pertinentes, mais ne fournit pas de solution concernant le choix ou l'établissement de nouvelles normes. Cela

touche à une des problématiques les plus importantes dans la reconnaissance de la pertinence linguistique des utilisations de l'anglais comme *lingua franca* et de la légitimité de leur étude : "The central issue, then, is whether ELF users should be accorded the right to be 'norm-developing' rather than simply 'norm-dependent'" (op. cit. : 60). En raison du nombre de personnes utilisant l'ALF mondialement, Seidlhofer conclut que les caractéristiques qui émergent des productions d'ALF doivent en effet être considérées comme productrices de normes. Ces normes peuvent être étudiées grâce à deux corpus d'un million de mots chacun centrés sur les productions orales d'ALF (plusieurs corpus de plus petite taille ont été créés pour des projets précis ; voir [Prodromou, 2008 : 94] pour une liste de ceux-ci) :

- le corpus *VOICE* (*Vienna-Oxford International Corpus of English*, 2013), développé à l'Université de Vienne, est un corpus d'enregistrements audio d'interactions en ALF accompagnés de leurs transcriptions, dans des contextes professionnels, pédagogiques ou privés ;
- le corpus *ELFA* (*English as a Lingua Franca in Academic Settings*, 2008), développé à l'Université de Tampere en Finlande, concerne uniquement des productions orales dans un cadre universitaire et ayant eu lieu dans des universités finlandaises.

À notre connaissance, il n'existe pas à ce jour de corpus de productions écrites dans un cadre d'ALF, d'une part car l'oral est le lieu privilégié de la variation linguistique, et d'autre part car l'écrit ne permet pas d'observer des interactions en temps réel donnant lieu aux négociations de sens caractéristiques des communications en ALF (op. cit. : 23). Les résultats de l'analyse de ces corpus amènent Seidlhofer à considérer l'ALF non pas comme une variété clairement démarquée d'un anglais standard dont les contours sont flous, mais comme le résultat d'une exploitation différente des ressources linguistiques fournies par la langue (op. cit. : 110).

Du point de vue théorique, Seidlhofer reprend le concept d'"anglais virtuel", emprunté à Widdowson (2003). L'ALM et l'ALF seraient tous les deux des réalisations de ce modèle, et il servirait de repère à l'évaluation de productions en ALF et à l'acceptation de nouvelles formes :

> [W]e need to be able to refer to a construct that can accommodate the dynamic and fluid character of ELF while also accounting for what its realizations across the globe, despite all their diversity, have in common: the underlying encoding possibilities that speakers make use of. It is these possibilities that we can (speculatively) call the virtual language. […] It is with reference to this virtual English that new words […] are coined in both

> ENL and ELF, and more generally, that speakers can be said to be 'using English', whether the forms they produce conform to attested ones or not. (op. cit. : 111-112).

Cette perspective permet d'envisager les modalités de la variation que constitue l'ALF d'un point de vue théorique. Dans la section suivante, nous abordons ces variations du point de vue de leur pratique réelle (en quoi l'ALF est-il différent de l'ALM ?) et d'un point de vue pédagogique (comment enseigner l'utilisation de l'ALF ?).

c. Pratique et enseignement de l'anglais comme lingua franca

L'intérêt de l'étude de la pratique de l'ALF réside moins dans le catalogue de ses caractéristiques que dans l'observation de la façon dont ses utilisateurs et utilisatrices font usage des ressources à leur disposition afin de communiquer de manière efficace, et des raisons expliquant l'utilisation des formes choisies (op. cit. : 97). Ces dernières sont le plus souvent liées à l'optimisation de l'intelligibilité et à une innovation linguistique qui peut servir plusieurs objectifs, comme la minimisation de la charge cognitive ou l'exploitation de connaissances communes. Comme dans toute communication, le sens est "négocié" et "co-construit", mais la particularité des utilisateurs et utilisatrices de l'ALF est leur propension à réinterpréter les normes de la langue lorsque cela sert un intérêt communicationnel (op. cit. : 99). Ce qui est interprété comme une erreur dans un cadre où l'ALM est pris comme point de repère peut en réalité être le résultat d'une recherche de clarté ou d'économie.

Un certain nombre de caractéristiques formelles fréquentes dans les productions d'ALF correspondent à cette stratégie. L'abandon du morphème flexionnel *-s* à la troisième personne du singulier au présent en est un exemple : constituant une irrégularité apparente dans la conjugaison des verbes, cette caractéristique est largement abandonnée dans les productions d'ALF (Cogo et Dewey, 2006 : 77). Le domaine de la morphologie dérivationnelle est également intéressant de ce point de vue. La dérivation des noms en verbes et vice-versa n'est pas régulière : on trouve *communication* et *communicate*, mais *pronunciation* et *pronounce*. S'il existe des raisons diachroniques pour expliquer ces phénomènes, elles ne sont pas connues des personnes utilisatrices de l'ALF, qui produisent des formes "régularisées", par exemple *pronunciate* ou *examinate*, comme l'indiquent les données du corpus *VOICE* (Seidlhofer, 2011 : 102). On observe également une recherche de régularité lexicale, par exemple dans le cas de l'utilisation de la préposition *about* avec le verbe *discuss*, fréquemment observée dans le corpus *VOICE* (op. cit. : 145), cette préposition étant utilisée avec des verbes de sens proche, comme *talk* ou *tell*.

Notre objectif dans cette étude n'étant pas de présenter un compte-rendu exhaustif de ces phénomènes, nous nous limiterons donc aux cas présentés ci-dessus. Ils permettent à Seidlhofer de souligner le fait que les innovations introduites par les personnes utilisatrices de l'ALF ne sont pas soumises au hasard mais suivent en réalité les règles morphologiques et lexicales de l'anglais :

> [T]he bulk of innovations we observe ELF speakers introducing […] is not a matter of arbitrarily replacing a [Standard English] pattern with 'just anything'. Rather, what we observe is the unfolding of familiar processes of language variation in language use, but extended in non-canonical, creative ways. (op. cit. : 108)

Même si l'on dispose de plus en plus d'informations concernant la pratique de l'ALF, la question de son enseignement demeure épineuse. Cette question a un intérêt particulier pour nous, puisque notre recherche inclut une composante didactique. Nous avons vu que si l'objectif des personnes apprenantes/utilisatrices est de faire usage de l'anglais dans des contextes internationaux sans rechercher de liens particuliers avec les cultures anglophones, l'adoption des normes de l'anglais comme langue maternelle n'est pas pertinente. Tout enseignement implique pourtant un choix prescriptif concernant ce qu'il convient d'enseigner. D'après Seidlhofer, la conceptualisation de l'enseignement de l'anglais dans un cadre d'ALF est centrée sur les deux points suivants :

- l'enseignement doit prendre en compte les objectifs communicationnels des élèves,
- l'objectif de l'enseignement est d'encourager l'utilisation de la langue afin d'activer les processus d'apprentissage.

Ce second point repose sur l'hypothèse selon laquelle l'apprentissage et l'utilisation d'une langue ne sont pas des processus consécutifs mais cumulatifs, et que la mobilisation des compétences linguistiques, même limitées, favorise l'extension de ces compétences (op. cit. : 198).

La question du choix des éléments à enseigner ne peut être résolue globalement mais est jugée au cas par cas : "What decisions teachers will make for particular learners with their particular needs will always be a local matter that a general book about ELF cannot (or rather should not) address" (op. cit. : 175). L'accent est mis sur une utilisation stratégique de la langue, orientée vers la recherche de l'intelligibilité et de la coopération, plutôt que sur la maîtrise de formes précises, qu'elles proviennent de descriptions de l'anglais comme *lingua franca* ou de l'anglais comme langue maternelle. (op. cit. : 198).

d. La prise en compte de l'approche de l'anglais comme lingua franca *dans notre recherche*

L'objet de cette thèse n'est bien sûr pas l'étude de l'ALF, et si nous utilisons les méthodes et les résultats de recherche des domaines de l'acquisition et de la didactique des langues secondes, notre degré d'expertise ne nous permet pas de nous prononcer sur les débats importants au sein de ces domaines. Les arguments en faveur de la prise en compte des nouvelles utilisations de l'anglais sont cependant suffisamment convaincants pour justifier de s'interroger sur leurs conséquences pour nos travaux, même si nous n'avons pas abordé l'ensemble des points de vue à ce sujet. Il semble cependant que le choix de ce qui constitue une erreur, la correction de ces erreurs, et l'angle sous lequel les productions de non-anglophones sont considérées ne soient pas neutres idéologiquement et scientifiquement.

Les recherches sur la pratique de l'ALF s'intéressent majoritairement aux productions orales, notamment aux productions en interaction. Les deux principaux corpus d'ALF, *VOICE* et *ELFA*, n'intègrent des documents écrits que dans la mesure où ils sont les transcriptions d'interventions orales. Comme nous l'avons mentionné, ce choix s'explique par le fait que les productions orales, notamment dans le cadre d'interactions spontanées, sont plus sensibles aux variations linguistiques. De plus, les variations observées dans ces productions, et dont Seidlhofer propose une synthèse, concernent en particulier le lexique (ex. : sélection de prépositions, créations de mots), la morphologie (ex. : modification de préfixes et suffixes) et la morphosyntaxe (ex. : conjugaisons). Nous ne disposons donc pas de résultats fiables concernant les variations qui peuvent être observées à l'écrit et dans le domaine de la syntaxe. En ce qui concerne les normes à utiliser pour la correction, il semble alors prudent pour l'instant de considérer que l'ALF est suffisamment semblable à l'ALM pour que les mêmes normes puissent être utilisées.

La question du bien-fondé des corrections se pose également à un autre niveau. Si l'objectif de l'utilisation de l'anglais pour une personne francophone est d'abord de pouvoir communiquer internationalement, et donc en ALF, est-il réellement intéressant de proposer des corrections grammaticales ? Ces corrections sont-elles utiles du point de vue de l'optimisation de la communication, ou bien participent-elles simplement de la perpétuation d'une "tradition linguistique" fondée sur le standard de l'anglais comme langue maternelle ? On ne peut donner de réponses à ces questions que pour le contexte actuel, car comme le montre Seidlhofer, la façon dont l'anglais est conceptualisé est en train d'être modifiée. En parallèle, elle insiste sur le fait que l'idée selon laquelle les normes de l'ALM doivent être

adoptées par toutes les personnes utilisant l'anglais est solidement ancrée, même parmi ces personnes. Ceci est corroboré par les résultats de recherche concernant l'évaluation des erreurs, et que nous avons présentés dans la section 1.1.1b. Pour rappel, un des résultats les plus stables de ces recherches est que les personnes utilisatrices non-natives de l'anglais sont plus sévères que les personnes locutrices natives lorsqu'on leur demande de juger de la gravité des erreurs dans des productions (cf. Ellis, 2008 : 56-57). Puisque les productions de personnes utilisatrices de l'ALF sont destinées à un public majoritairement non-natif, il nous semble que ce facteur est à prendre en compte, même s'il nous pousse au paradoxe. Par ailleurs, l'un des principaux éléments de la pratique de l'ALF, c'est-à-dire la négociation du sens, n'est pas possible dans des écrits formels, ce qui rend d'autant plus importante la production de textes clairs et sans ambiguïtés découlant d'éventuelles erreurs. Pour ces raisons, il apparaît que la correction des erreurs grammaticales conserve un intérêt important pour le public utilisateur de l'ALF.

La prise en compte du cadre conceptuel de l'anglais comme *lingua franca* a cependant un ensemble de conséquences directes sur des éléments importants de notre recherche. Tout d'abord, nous adoptons en partie la terminologie de ce domaine, privilégiant le terme d'"utilisatrice/utilisateur" au lieu d'"apprenante/apprenant", afin de représenter de manière plus réaliste l'expérience de l'anglais de notre public cible. Vivian Cook donne la définition suivante, très large, du terme "utilisateur/utilisatrice d'une L2" : "An L2 user is any person who uses another language than his or her first language (L1), that is to say, the one learnt first as a child" (2002 : 1). Nous n'adoptons cependant pas systématiquement la distinction entre anglais comme *lingua franca* et anglais comme langue maternelle, cette distinction n'étant pas centrale pour nos objectifs et les productions que nous étudions. Par ailleurs, nous conservons également le terme de "erreur" pour identifier les segments non-conventionnels produits par des personnes utilisatrices de l'ALF ; comme nous l'avons vu dans la section 1.1.2, nous n'associons pas de connotation négative à ce terme et nous l'utilisons afin d'alléger nos expressions et de conserver une formulation compréhensible par le plus grand nombre.

Le choix des documents à inclure dans notre corpus, que nous présentons en détail dans la section 1.2.1b, découle de la conception de notre public cible comme des personnes utilisatrices et non strictement apprenantes de l'anglais ; les productions choisies correspondent à un ensemble d'utilisations différentes de l'anglais dans des contextes de communication internationale. Pour finir, le choix des erreurs et leur traitement reflètent également les résultats de l'observation de la pratique de l'ALF : nous sélectionnons des

erreurs pouvant causer des problèmes de compréhension ou générer une ambiguïté, et proposons des corrections ayant pour objectif de maximiser l'intelligibilité des segments.

Ainsi s'achève la présentation des principaux aspects théoriques qui sous-tendent notre méthode de sélection des erreurs à traiter. Nous avons présenté les différentes étapes de la méthode de l'analyse des erreurs, ainsi que réaffirmé la pertinence de son utilisation en tant que méthode de recueil de données, malgré les critiques qu'elle a pu recevoir concernant ses fondements théoriques. Rappelons que cette méthode comprend cinq étapes : constitution d'un corpus, identification et description des erreurs, explication et évaluation. Un projet reposant sur les erreurs linguistiques ne pouvant faire l'économie d'une réflexion sur ce qui constitue une erreur, nous nous sommes penchés plus précisément sur l'identification et la délimitation de celles-ci. Nous avons ainsi pu poser les bases théoriques du repérage que nous présentons dans la partie 1.2. La définition de l'erreur comme un "échec" a amené une réflexion autour des phénomènes en jeu dans la production d'erreurs (compétence vs. performance, visibilité des erreurs, causes des erreurs). Nous nous sommes focalisés sur les notions de grammaticalité et d'acceptabilité dans le jugement d'erreur ; la question du choix des normes de grammaire et d'usage à prendre en compte lors de la détection et de la correction des erreurs est alors apparue comme particulièrement épineuse. En effet, le statut particulier de l'anglais, langue internationale, pousse à repenser l'établissement des normes et à voir au-delà de la simple norme native, encore largement dominante. Notre recherche porte sur l'anglais écrit, qui reste le médium le plus homogène du point de vue des normes de grammaticalité ; nous prendrons cependant en compte la perspective de l'anglais comme *lingua franca* dans la sélection des erreurs à traiter, présentée dans la partie suivante.

1.2 De l'analyse des erreurs à la sélection des erreurs à traiter

Dans les sections suivantes, nous présentons l'application des étapes de l'analyse des erreurs à un projet de détection et correction automatique de celles-ci. L'utilisation de cette approche empirique sert deux objectifs principaux. Comme nous l'avons déjà dit, nous souhaitons traiter des erreurs qui représentent une difficulté significative pour les personnes utilisatrices de l'anglais L2. Cependant, certains types d'erreur très fréquents en anglais L2, comme par exemple le choix des prépositions et des articles, sont déjà largement traités dans des projets de correction grammaticale automatisée à visée de recherche ou de développement commercial (ex. : De Felice, 2008 ; Hermet et Désilets, 2009 ; Tetreault et Chodorow, 2008).

Il est nécessaire de se pencher de nouveau sur des productions authentiques si l'on souhaite débusquer d'autres types d'erreur en attente de traitement, et qui peuvent être détectées et corrigées en utilisant des méthodes linguistiques.

Le deuxième objectif de cette approche empirique est de recueillir des informations sur les erreurs : une fois les types d'erreur sélectionnés, l'analyse de l'échantillon des segments erronés repérés dans le corpus nous permet de déterminer plus précisément les caractéristiques de ces segments. Notre utilisation de la méthode de l'analyse des erreurs est donc à ranger du côté de ses usages pédagogiques plutôt que théoriques. Le traitement des erreurs assisté par ordinateur (*computer-aided error analysis*) est un domaine de recherche qui s'est développé à la suite de la constitution et de l'étude de corpus d'interlangue informatisés, dans les années 1990 (cf. Dagneaux et al., 1998). Nous présentons dans cette section certains des aspects importants de ce domaine, tel l'annotation des erreurs, mais le traitement que nous présentons ne peut être qualifié d'"assisté par ordinateur" car les erreurs n'ont pas été annotées à l'aide d'un outil informatique dans le corpus.

Les sous-parties suivantes présentent l'application à notre étude de trois des étapes de l'analyse des erreurs. La sous-partie 1.2.1 est consacrée à la constitution d'un corpus de productions écrites de personnes utilisatrices de l'anglais. Ce choix de sujets, ou du moins de perspective sur leur relation à l'anglais, marque la première distinction entre l'AE "classique" et son application dans nos travaux, qui ciblent les personnes *utilisatrices* et non strictement *apprenantes* de l'anglais. La sous-partie 1.2.2 traite de l'identification des erreurs, c'est-à-dire leur relevé dans le corpus, ainsi que de leur description, effectuée sous la forme d'une catégorisation. Une description grammaticale "neutre" des erreurs sélectionnées (cf. James, 1998, voir sous-section 1.1.1b "Description des erreurs") sera donnée dans le Chapitre 2. Nous abordons également dans la sous-partie 1.2.2 les critères choisis pour l'identification des erreurs et les limites de notre étude dans la gestion de cette étape. La dernière partie de cette section est consacrée à la présentation d'un schéma d'annotation des erreurs fournissant des informations pouvant mener à la création de messages correctifs. La quatrième étape de l'AE, l'explication des erreurs, sera abordée dans le Chapitre 2, lorsque nous évoquons les difficultés d'apprentissage posées par ces deux types d'erreur en anglais, et dans le Chapitre 3 dans le cadre de la mise en place de messages correctifs.

La phase d'évaluation de la gravité des erreurs n'est pas traitée dans une partie spécifique, mais nous prenons en compte certains des facteurs de gravité identifiés par la recherche en AE (ex. : portion de phrase concernée par l'erreur, intelligibilité) afin de sélectionner les erreurs à

traiter. La sous-partie 1.2.3 passe en revue les erreurs fréquentes relevées dans notre corpus afin d'évaluer leur adéquation avec nos exigences d'innovation et de faisabilité. Nous développons ensuite l'intérêt linguistique et pédagogique du traitement des deux types d'erreur sélectionnés. Notons ici que nous serons amenée à faire brièvement référence aux résultats d'études comparatives des différents correcteurs grammaticaux existants et à présenter un bref état de l'art du domaine de la correction grammaticale automatisée du point de vue des types d'erreur fréquents. Un état de l'art des méthodes de traitement automatique d'erreurs linguistiques est présenté dans le Chapitre 3, qui est consacré à l'implémentation de nos propres règles de détection et de correction.

1.2.1 Constitution d'un corpus de productions écrites d'utilisatrices et utilisateurs de l'anglais

Les sections suivantes sont consacrées à la constitution du corpus qui fournit la base de notre analyse des erreurs. Nous commençons par présenter une sélection des corpus d'interlangue existants, et expliquons pourquoi nous avons choisi de constituer un corpus original. Nous exposons ensuite les critères de sélection des documents intégrés au corpus, ainsi que leurs sources et leurs caractéristiques. Nous évoquons également les problèmes qui sont posés par l'utilisation de certains documents, comme les courriels personnels et les publications scientifiques. En dernier lieu, nous donnons quelques détails concernant la mise en place du corpus par le biais de la plateforme *Sketch Engine*.

a. Les corpus d'interlangue, ou pourquoi constituer un corpus original

Un corpus d'interlangue (*learner corpus*) est un corpus de documents authentiques écrits ou oraux produits par des personnes apprenantes d'une L2. Les L1 des sujets peuvent être variées à l'intérieur d'un même corpus, mais la L2, ou langue cible, est déterminée. Le terme d'"interlangue" (cf. 1.1.1a) fait référence au système linguistique intermédiaire des personnes apprenantes d'une L2. Nous choisissons d'utiliser ce terme plutôt que son alternative plus courante de "corpus d'apprenants", notamment car celle-ci rend difficile le respect des pratiques de rédaction épicène.

Les corpus d'interlangue sont utilisés dans plusieurs types de recherche, comme l'analyse des erreurs bien sûr, ou des études dans le domaine de l'acquisition des langues secondes qui concernent divers aspects de l'acquisition, comme les règles d'évitement ou l'influence de la L1 sur les productions en L2. Ils sont aussi utilisés dans des projets de détection et correction

automatisées fondés sur l'analyse automatique de données ; dans ce cadre, ils servent de données d'entraînement pour des classifieurs (Leacock et al., 2010 : 27).

D'après le recensement effectué par le *Centre for English Corpus Linguistics* de l'Université Catholique de Louvain et disponible en ligne, il existe à ce jour près de 130 corpus d'interlangue. Le tableau 3 en présente une sélection en fonction de nos paramètres. Nous retenons uniquement les corpus de documents écrits, avec l'anglais comme L2, dans lesquels les sujets sont strictement non-natifs (certains corpus intègrent des productions de personnes locutrices natives de la langue étudiée en tant que L2 par les personnes apprenantes, dans un but comparatif).

Intitulé	**Nature des documents**	**Niveau**	**Taille (nombre de mots)**	**Direction**
The Cambridge Learner Corpus (CLC)	Copies d'examen, série ESOL L1 variées	Divers	25 millions (total)	Cambridge University Press, Cambridge ESOL
The Longman's Learners Corpus	Dissertations et copies d'examen L1 variées	Divers	10 millions (total)	Pearson Longman
The Scientext English Learner Corpus	Dissertations, dossiers L1 français	Intermédiaire à avancé	1,1 million (L1 français)	LIDILEM Université de Grenoble 3
The International Corpus of Learner English (ICLE)	Dissertations L1 variées	Intermédiaire à avancé	3,7 millions (total)	Sylviane Granger CECL Université Catholique de Louvain
Longitudinal Database of Learner English (LONGDALE)	Documents varies L1 variées	Intermédiaire à avancé	n/c	Fanny Meunier CECL Université Catholique de Louvain
The Varieties of English for Specific Purposes Database (VESPA)	Productions en anglais de spécialité L1 variées	Divers	n/c	Magali Pacquot CECL Université Catholique de Louvain

Tableau 3. Sélection de corpus d'interlangue

De plus, et il s'agit là du critère le plus restrictif, nous nous concentrons sur les corpus qui regroupent des productions de francophones : comme nous l'avons indiqué dans l'introduction générale de ce document, notre public cible est celui des adultes francophones. En effet, nous souhaitons exploiter les éventuelles conséquences de l'influence translinguistique français-anglais, et ainsi limiter le nombre de configurations erronées possibles dans la phase de détection. Nous souhaitons également fournir un retour correctif dans la langue maternelle du public cible dans la phase de correction. Cela n'est possible que si l'ensemble du public cible partage la même L1. Le fait que cette L1 soit également celle de l'auteure des travaux est un facteur de facilitation additionnel. Nous incluons également des corpus qui ne sont pas

spécifiquement consacrés au français L1 mais qui incluent ce module parmi d'autres langues. Les corpus sont présentés par ordre de taille, du plus important au plus réduit, mais la taille donnée est celle du corpus en entier : il est le plus souvent impossible de connaître le détail chiffré des sous-corpus.

Il convient à présent d'expliquer notre choix de constituer un corpus original pour le projet au lieu d'utiliser un corpus existant. En observant la synthèse présentée dans ce tableau, on obtient déjà plusieurs éléments de réponse. Pour commencer, les corpus d'interlangue de productions écrites français L1 – anglais L2 sont en nombre limité. Ce petit nombre est encore réduit par le fait que certains de ces corpus ne sont pas mis à disposition de l'ensemble de la communauté scientifique, pour des raisons commerciales. C'est le cas du *Cambridge Learner Corpus* et du *Longman Learners' Corpus*, dont l'accès est limité aux personnes collaborant avec ces éditeurs. Deux des corpus créés au CECL de l'Université Catholique de Louvain, *LONGDALE* et *VESPA*, ne sont pas encore utilisables par des personnes extérieures à l'équipe ou ne collaborant pas avec celle-ci sur le projet. De plus, la constitution de ces corpus ayant été initiée en 2008, leur développement est contemporain de celui du corpus utilisé pour notre recherche, et ils n'auraient donc pas pu s'y substituer.

Par ailleurs, ces corpus sont étiquetés "interlangue", et même la partie français L1 – anglais L2 du corpus *Scientext* (ce corpus incluant également des productions scientifiques en français natif et anglais natif), qui regroupe des productions académiques, utilise les textes produits dans le cadre de leurs études par des personnes étudiant l'anglais en spécialité à l'université. Même si, d'après la définition de Cook (voir 1.1.4d), ces personnes sont utilisatrices de l'anglais, ce contexte est clairement axé sur l'apprentissage plutôt que l'utilisation.

Afin de pouvoir satisfaire aux critères présentés dans la section suivante, nous avons donc constitué un corpus de productions authentiques pour nos travaux. Cela n'exclut cependant pas d'utiliser les ressources existantes, puisqu'un des sous-ensembles provient du module français L1 du corpus *ICLE*. Dans l'éventualité d'une poursuite du développement de ce corpus, nous n'excluons pas non plus de le modifier ou de l'enrichir avec des apports d'autres corpus.

b. Critères de sélection des documents et composition du corpus

Afin de constituer un corpus représentatif de l'utilisation de l'anglais par des adultes francophones, nous tentons d'identifier l'ensemble des contextes et des documents pour

lesquels l'anglais est le plus souvent utilisé. Comme nous l'avons vu dans la sous-partie 1.1.4, l'anglais comme langue internationale est fréquemment utilisé à l'oral, mais notre étude étant axée sur les productions écrites, nous nous limiterons aux contextes nécessitant l'utilisation de l'anglais écrit. Le tableau 4 présente une liste de ces types de contexte. Cette liste n'est pas exhaustive, les utilisations de l'anglais étant infiniment variées, mais elle regroupe un ensemble de contextes dont nous estimons qu'ils sont relativement fréquents.

Nous prenons comme point de départ le profil des locuteurs et locutrices (personne travaillant ou étudiant dans le domaine de la recherche, personne travaillant dans d'autres domaines privés ou publics, population étudiante, particuliers). À partir de ces profils, nous faisons la liste des utilisations possibles de l'anglais pour ces personnes, et nous y adossons les types de public destinataire des productions. Ce paramètre a un impact sur le niveau de contrôle qui est exercé lors de la production ainsi que sur le niveau de langue utilisé et le degré de précision de l'expression qui est requis. Nous identifions les destinataires les plus probables, étant entendu qu'un document peut être lu par un ensemble de personnes différentes (ex. : rapports de projets scientifiques, qui peuvent être lus par des partenaires hors recherche comme par les organismes financeurs).

Prenant en compte les observations faites par les spécialistes de l'anglais comme *lingua franca*, nous partons du principe que les destinataires ne sont pas nécessairement des locutrices et locuteurs de l'anglais L1. Il faut également garder à l'esprit que ce tableau présente une schématisation hautement simplifiée, une seule et même personne pouvant être concernée par tous les profils selon les situations, simultanément ou successivement. Le tableau présentant ces différents paramètres est visible en page suivante.

Profil	Productions	Destinataires
Personnes travaillant ou étudiant dans le domaine de la recherche	Publications scientifiques Mémoires, thèses	Communauté de recherche Comités de lecture Jurys Futurs organismes employeurs
	Rapports de projets scientifiques	Partenaires hors-recherche Organismes financeurs
	Supports de présentations orales ("diapositives")	Communauté de recherche Partenaires hors recherche Jurys
	Courriels professionnels	Partenaires recherche Partenaires hors recherche Population étudiante étrangère Administrations et organismes étrangers (comités d'organisation de colloques, universités étrangères, etc.)
	Sites internet professionnels Blogs professionnels	Communauté de recherche Futurs organismes employeurs Public large
Personnes travaillant dans d'autres domaines (publics et privés)	Rapports de projet	Partenaires professionnels Organismes financeurs Responsables Clientèle
	Documentation technique Plaquettes de présentation	Clientèle Public utilisateur
	Courriels professionnels	Partenaires professionnels Responsables Clientèle Fournisseurs
	Sites internet professionnels	Partenaires professionnels Clientèle
	Curriculum vitae Lettres de candidature/motivation	Organismes employeurs
Population étudiante (hors recherche)	Dissertations (*essays*)	Personnel enseignant
	Curriculum vitae Lettres de candidature/motivation	Organismes employeurs (ex. : stages) Comités de sélection (ex. : programmes d'échange)
	Courriels professionnels	Universités étrangères
Particuliers	Courriels semi-professionnels (communication avec des professionnels dans le cadre de la vie privée)	Commerces, entreprises Partenaires semi-professionnels (ex. : associations)
	Courriels personnels	Famille, amis
	Sites internet et blogs personnels	Famille, amis Public large
	Sites internet et blogs semi-professionnels (ex. : sites de présentation d'associations, de projets artistiques, etc.)	Public large
	Commentaires sur sites internet, forums et blogs	Public large

Tableau 4. Liste des contextes de l'utilisation de l'anglais par des adultes francophones

Dans l'idéal, un corpus d'utilisation de l'anglais devrait inclure des échantillons de toutes ces productions. Dans le cadre de nos travaux, dont l'ampleur est limitée et dans lesquels la constitution d'un corpus est un objectif secondaire, nous avons sélectionné certains types de production pour des raisons pratiques. Ces productions sont les suivantes : publications scientifiques, productions d'interlangue, courriels professionnels, semi-professionnels et personnels, extraits d'un rapport technique. Le tableau 5 en présente la synthèse chiffrée. Nous détaillons ensuite les caractéristiques de chaque sous-ensemble.

Types de production	Nbr. d'auteur.es	Nbr. de textes	Taille (mots)
Publications scientifiques	26	12	40020
Productions d'interlangue	49	49	32430
Courriels (toutes catégories)	8	125	17267
Extraits d'un rapport technique	2	1	11177
TOTAL	**85**	**187**	**100894**

Tableau 5. Synthèse de la composition du corpus

Le principal sous-ensemble du corpus est celui qui concerne les publications scientifiques. Avec le domaine de l'entreprenariat privé, le domaine académique est sans doute celui dans lequel on rencontre le plus de communication écrite en anglais au niveau international. Un des deux principaux corpus d'anglais comme *lingua franca* est d'ailleurs consacré uniquement aux productions orales dans le milieu académique (cf. corpus *ELFA*). Les documents sélectionnés pour notre corpus sont des articles scientifiques qui sont soit disponibles librement sur internet, soit mis à disposition du projet par leurs auteures et auteurs. Il peut s'agir d'articles non publiés ou de versions antérieures à la publication. Ils ont été rendus entièrement anonymes, comme tous les documents du corpus (les prénoms sont conservés à l'intérieur des courriels, mais les signatures ont été retirées afin de ne pas identifier les auteurs et auteures).

Nous avons cherché à diversifier les domaines scientifiques représentés, mais avons rencontré deux difficultés. Dans certains domaines des sciences humaines n'étant pas liés directement au monde anglophone, la publication d'articles de recherche en anglais par des francophones est moins fréquente que dans les domaines des "sciences dures" ou liés aux technologies. De plus, afin de garantir un relevé fiable des erreurs, le contenu des articles doit être compréhensible pour la personne relevant les erreurs, ce qui rend difficile l'inclusion d'articles dans des domaines tels que la biologie, la médecine, les mathématiques ou la

physique. Les articles sélectionnés proviennent donc en majorité de domaines tels que le traitement automatisé des langues, la linguistique, la psychologie et l'histoire.

L'utilisation de ce genre de document pose cependant le problème de la traçabilité de l'anglais utilisé : même s'ils sont en général rédigés par une seule personne, ces articles ont plusieurs auteurs et auteures, qui peuvent avoir effectué des corrections ou modifications linguistiques. Ils peuvent également avoir été relus par des anglophones. Ce problème peut être contourné si on sélectionne uniquement des articles rédigés par une seule personne, et si on s'assure par le biais d'un questionnaire que l'article n'a pas fait l'objet de corrections par un ou une anglophone. Une telle rigueur, même si elle est souhaitable, n'est pas absolument nécessaire ici, puisque notre objectif principal est de repérer les erreurs qui subsistent à un niveau de maîtrise intermédiaire à avancé, et non d'obtenir des données sur l'utilisation de l'anglais international à proprement parler. Il conviendrait cependant de modifier le corpus afin d'intégrer les dispositions mentionnées plus haut si l'on souhaite qu'il soit également utilisable pour des recherches de cette nature.

Le deuxième sous-ensemble est celui des productions d'interlangue. Le choix d'inclure ces productions dans un corpus d'utilisation de l'anglais peut paraître paradoxal. Ce choix est avant tout pragmatique, car comme nous l'avons vu, le recours au corpus *ICLE* permet d'avoir accès à des productions en anglais écrit du même niveau que celui de notre public cible, entièrement annotées et disponibles en format numérique. Les textes sont de courtes dissertations autour d'une question de société à débattre ou autour d'une œuvre littéraire, et sont donc proches de l'utilisation scientifique de l'anglais par certains aspects. Notre corpus ne contient qu'une sélection du module de français L1 du corpus. Nous avons inclus des productions par des sujets ayant étudié l'anglais pendant une durée totale pouvant aller de six à huit ans, ayant uniquement le français comme langue maternelle, n'utilisant pas d'autre langue dans leur milieu familial, et n'utilisant pas d'autre langue étrangère en général, c'est-à-dire n'ayant pas déclaré la connaissance d'une L3. Par ailleurs, une partie de ces productions n'a pas été incluse dans le corpus principal mais conservée dans le but de servir de données pour la phase d'évaluation.

Le troisième sous-ensemble regroupe les courriels. Nous n'indiquons pas ici les distinctions de catégories (personnels, semi-professionnels, professionnels), la taille très réduite de ce sous-ensemble rendant les distinctions entre les différents types peu pertinentes. Les documents ont été mis à la disposition du projet par des volontaires. Comme le remarque Seidlhofer (2011 : 4), et comme nous l'indiquons dans notre synthèse des utilisations de

l'anglais international, une part importante de la communication en anglais écrit se fait par le biais d'internet. Il serait donc particulièrement intéressant d'avoir accès à ce genre de document, mais il est très difficile d'en obtenir de grandes quantités car ces communications relèvent le plus souvent soit de la discrétion professionnelle, soit de la vie privée.

Le sous-ensemble le plus réduit du corpus est celui qui regroupe des extraits courts d'un rapport technique rédigé dans le cadre d'un projet en entreprise. Ici également, il serait intéressant d'avoir accès à plus de données de cette nature, mais leur utilisation, même à but de recherche, peut poser des problèmes de confidentialité.

Parmi les productions qui ne sont pas incluses dans notre corpus, certaines communications sur internet seraient particulièrement intéressantes à prendre en compte dans le cadre de la création d'un véritable corpus d'utilisation de l'anglais. Les commentaires laissés sur des sites ou blogs, grâce à l'instantanéité et aux interactions rendues possibles par ce mode de communication, permettraient d'observer à l'écrit la négociation du sens et la construction collaborative qu'évoque Seidlhofer. Le problème de la traçabilité de la production se pose également ici, car il est difficile de s'assurer que l'auteur ou auteure n'a pas l'anglais comme L1. Le même problème existe en ce qui concerne les sites internet et blogs professionnels, semi-professionnels et personnels. Cette considération dépassant largement notre cadre et nos objectifs pour l'instant, nous n'avons pas inclus ces documents dans notre corpus.

En plus du type d'utilisation de l'anglais, le corpus doit refléter le niveau de maîtrise de l'anglais choisi ici, c'est-à-dire le niveau intermédiaire à avancé. Les productions incluses dans le projet *ICLE* sont clairement identifiées comme relevant de ce niveau. La rédaction d'articles de recherche et de rapports nécessite de fait une maîtrise relativement avancée de la langue de rédaction, et comme nous le verrons dans la partie suivante (voir 1.2.2), la fréquence des erreurs dans ces productions est peu élevée. Le *Cadre Européen Commun de Référence pour les Langues* donne les descripteurs suivants pour les productions écrites du niveau C1, qui correspond à la fourchette haute du niveau que nous avons décrit comme "intermédiaire à avancé" (2001 : 27) :

> Je peux m'exprimer dans un texte clair et bien structuré et développer mon point de vue. Je peux écrire sur des sujets complexes dans une lettre, un essai ou un rapport, en soulignant les points que je juge importants. Je peux adopter un style adapté au destinataire.

Si un auteur ou une auteure est capable de produire un article ou un rapport clair et avec une proportion d'erreurs relativement faible, cette personne peut être considérée comme

maîtrisant l'anglais à un niveau au moins intermédiaire. Les productions dont les niveaux sont les plus hétérogènes sont les courriels. Même si certains d'entre eux comportent plus d'erreurs que les autres et seraient donc à ranger dans la fourchette basse des niveaux de maîtrise, ils correspondent dans l'ensemble au descripteur cité ci-dessus.

Le corpus a été rassemblé sur la plateforme *Sketch Engine*, disponible en ligne. Il est annoté pour les parties du discours et lemmatisé avec l'analyseur incorporé dans *Sketch Engine*, Tree Tagger 2.2. Cette plateforme donne accès à un ensemble d'outils facilitant l'analyse du corpus, dont les deux principaux sont un concordancier et l'outil *Word Sketches*, qui permet la génération de synthèses de l'environnement grammatical et des collocations d'un terme. Cet outil ne nous est malheureusement pas accessible en raison de la taille réduite du corpus ; les analyses des structures et contextes grammaticaux des segments erronés sont donc effectuées à la main. *Sketch Engine* ne permet pas non plus d'annoter les erreurs directement dans le corpus. Un des intérêts de cette plateforme est la possibilité de partager ses données avec les autres personnes abonnées à *Sketch Engine*. Le corpus n'est cependant pas accessible à des personnes extérieures à la date de conclusion de nos travaux, en juin 2014.

1.2.2 Identification et description des erreurs détectées dans le corpus

Cette section présente l'application des étapes 2 et 3 du schéma traditionnel de l'analyse des erreurs, c'est-à-dire l'identification et la description de celles-ci. Notre approche du traitement des erreurs peut être qualifiée d'ascendante (*bottom-up*), dans le sens où leur relevé précède leur catégorisation, qui est elle-même utilisée dans la création du schéma d'annotation. Cette approche est justifiée par l'objectif premier de cette analyse, qui est de repérer des erreurs "orphelines" et pouvant être corrigées avantageusement grâce à des méthodes linguistiques. Ces étapes faisant partie des phases préliminaires de nos travaux, les relevés et analyses ont été effectués manuellement ; la numérisation des résultats serait une étape à inclure obligatoirement dans un développement éventuel du corpus.

a. Relevé des erreurs

Dans notre présentation de l'identification et de la définition des erreurs (voir 1.1.2), nous avons vu qu'il existe au moins trois problématiques à considérer lors du relevé de ces dernières dans un corpus d'AE : les critères utilisés pour juger qu'un segment est erroné, la distinction entre "erreurs" de compétence et "méprises" liées à la performance, et la visibilité

des erreurs. Les erreurs invisibles sont par définition le plus souvent impossibles à détecter et à relever, si nous n'avons pas accès au sens que le locuteur ou la locutrice avait l'intention de donner au segment. Cette distinction n'est donc pas prise en compte ici. Comme nous l'avons vu, la distinction entre "erreurs" et "méprises" est d'une part impossible à effectuer à partir d'un corpus de productions écrites uniquement (cf. Thouësny, 2011), et d'autre part n'est pas pertinente dans notre cadre, puisque corriger les unes comme les autres permettrait d'améliorer la justesse grammaticale des productions (cf. James, 1998). Les critères utilisés ici pour relever et classer les erreurs sont donc ceux de la grammaticalité et de l'acceptabilité des segments, étant entendu que la frontière entre ces deux concepts est mobile, et qu'ils sont le plus souvent corrélés (cf. 1.1.2a). Observons les trois segments suivants, issus de notre corpus :

[19] **three list of files*

[20] **on the bottom and the right of the picture*

[21] **To commemorate the abolition of slavery in France, it is to speak about humanity and Republican values*

Le premier de ces segments est facilement jugé comme grammaticalement erroné, l'accord obligatoire du déterminant pluriel avec le nom n'étant pas respecté. Le second n'est pas une erreur d'un point de vue grammatical, mais ne respecte pas les normes d'usage de ce type d'expression. Nous verrons d'ailleurs qu'une proportion importante des erreurs est due à des questions de lexique (sélection de termes, réalisation des contraintes lexicales, etc.). Le troisième segment est à la frontière de ces deux critères, avec un redoublement du sujet qui n'est pas accepté en anglais, mais dont on peut se demander s'il constitue une erreur de grammaire ou non.

La question de "qui juge les erreurs ?" a été abordée lors de notre présentation de l'étape de l'évaluation des erreurs, et il est nécessaire d'y répondre ici pour compléter notre présentation des critères utilisés. La personne ayant effectué les jugements d'erreurs est l'auteure principale des travaux, dont le profil est celui d'une utilisatrice francophone non-native de l'anglais et enseignante de l'anglais L2. Les productions ne sont pas jugées en fonction d'une variété précise de l'anglais, mais, en ce qui concerne les jugements de grammaticalité et acceptabilité, le cadre cible est celui de l'anglais comme langue maternelle, dans la mesure du possible. Les segments erronés sont relevés manuellement, cette étape constituant le développement d'une étude préliminaire effectuée avant la numérisation du corpus sur la plateforme *Sketch Engine*.

Nous identifions au moins deux limites à cette méthode. Le traitement manuel des segments, que nous venons d'évoquer, constitue une première limite. Comme indiqué dans le paragraphe d'introduction à cette section, une des modifications à apporter concerne la numérisation des données rassemblées. La seconde limite concerne la fiabilité des jugements de grammaticalité et d'acceptabilité. Ces derniers sont variables en fonction des personnes effectuant ces jugements, si bien qu'il est préférable de pouvoir comparer les jugements de plusieurs personnes, en particulier lorsque sont concernées des erreurs pour lesquelles les règles de grammaire ou d'acceptabilité sont difficiles à identifier ou relèvent de facteurs multiples (Leacock et al., 2010 : 81). Ceci n'a pas été possible dans notre étude pour des raisons pratiques, et cette limite influence nécessairement les résultats obtenus.

Ce problème pourrait être résolu à moindre coût par le recours au *Mechanical Turk* lancé par la compagnie Amazon en 2005. Ce système, disponible sur internet, permet de mettre en ligne des tâches simples requérant néanmoins une réponse humaine, comme des jugements de grammaticalité. Les tâches sont ensuite effectuées par des *Turkers* volontaires pour un coût très bas.

Cette méthode de recherche de données, appelée *crowdsourcing*, gagne en popularité dans de nombreux domaines. Dans une étude pilote sur la sélection des prépositions et la détection de l'utilisation erronée de prépositions, Tetreault et al. ont montré qu'il était possible d'obtenir un niveau de fiabilité comparable entre les résultats des *Turkers* et ceux de personnes qualifiées, mais le nombre de *Turkers* doit être bien supérieur, avec un ratio allant de trois à treize *Turkers* pour deux personnes qualifiées. Les tâches de détection d'erreurs sont jugées par les auteurs de l'article comme étant plus lourdes d'un point de vue cognitif que les tâches de sélection de prépositions, ce qui explique la nécessité de faire appel à un nombre de *Turkers* élevé (Tetreault et al., 2010).

L'utilisation de ce système peut constituer une aide d'appoint ou servir à nuancer des résultats *a posteriori*, mais la détection d'erreurs dans des textes longs semble être une tâche trop complexe pour reposer uniquement sur cette méthode. Notons qu'elle n'est pas sans poser de problème éthique, le salaire des *Turkers* pouvant avoisiner les 1,5$/heure (Cushing, 2012).

b. Catégorisation et annotation des erreurs

Dans la sous-section 1.1.1b "Description des erreurs", nous avons vu que cette description passe notamment par leur catégorisation. Avec l'avènement de l'informatisation du traitement des erreurs, leur description est effectuée par le biais de l'annotation, dont le schéma est le

plus souvent basé sur le système de catégorisation choisi. Rappelons ici que si une réflexion est menée concernant le schéma d'annotation pouvant être utilisé pour décrire les erreurs, ces dernières n'ont pas fait l'objet d'annotations informatisées systématiques. Cette approche étant néanmoins la plus souhaitable et la plus répandue à ce jour, notamment dans le cadre de la détection et de la correction automatisées, la synthèse présentée dans les paragraphes suivants est centrée sur ces systèmes. Après avoir abordé les probématiques et objectifs courants de l'annotation des erreurs par le biais d'un bref état du domaine, nous identifions les distinctions propres à notre recherche, avant de proposer un système de catégorisation pour les erreurs relevées, et un schéma d'annotation consacré à la correction automatisée des erreurs.

Annoter les erreurs d'interlangue : état du domaine

Dans leur article de 2006, Díaz-Negrillo et Fernàndez-Domínguez passent en revue douze systèmes d'annotation des erreurs et présentent les caractéristiques des quatre systèmes d'annotation les mieux documentés. Ces quatre systèmes, adossés à des corpus d'interlangue, sont présentés dans le tableau suivant (lorsque les systèmes n'ont pas d'intitulé précis, on leur donne généralement le nom de l'équipe ou du corpus auquel ils sont associés – nous reprenons les titres utilisés par Díaz-Negrillo et Fernàndez-Domínguez dans leur article). Nous détaillons leurs caractéristiques importantes dans les paragraphes qui suivent.

Nom du corpus	L1	L2	Objectif	Description
Cambridge Learner Corpus	Variées	Anglais	Commercial – Création de supports d'enseignement de l'anglais	Nicholls, 2003
Projet FreeText	Variées	Français	Académique – Analyse des erreurs, création d'un système d'ELAO	Granger, 2003
International Corpus of Learner English	Variées	Anglais	Académique – Recherches diverses (ELAO, ALS)	Dagneaux et al., 1998
NICT Japanese Learner of English Corpus	Japonais	Anglais	Académique – Détection automatique d'erreurs, création d'un modèle d'interlangue	Izumi et al., 2005

Tableau 6. Caractéristiques principales de quatre systèmes d'annotation

Leacock et al. (2010) incluent également un état de l'art et une réflexion sur les problématiques relatives à l'annotation des erreurs dans leur ouvrage sur la détection automatisée. Nous nous servons principalement de ces deux sources comme base de la synthèse suivante.

Comme il a déjà été mentionné, la possibilité de numériser les corpus d'interlangue a constitué une relance importante de la méthode de l'analyse des erreurs, notamment grâce aux possibilités offertes par l'utilisation des fonctions de recherche. C'est principalement dans cet objectif que l'annotation des erreurs a été développée : la description des segments erronés grâce à des étiquettes précises permet ensuite d'effectuer des recherches ciblées et de rassembler les erreurs relevant de la même caractéristique dans un corpus de grande taille (Nicholls, 2003 : 572). À la suite de Granger (2003 : 467), on identifie généralement quatre exigences à satisfaire afin qu'un schéma d'annotation soit utile et pertinent : le schéma doit permettre un équilibre entre quantité d'information et maniabilité, ainsi qu'être réutilisable, flexible et cohérent.

Les principales informations contenues dans un schéma d'annotation concernent généralement la catégorisation des erreurs, et la première exigence renvoie à la granularité du système taxonomique choisi, c'est-à-dire à la quantité de détail ou à la finesse des sous-catégories qu'il inclut. Si un degré de granularité fin ouvre des possibilités de recherche plus vastes, il peut se révéler problématique au stade de l'exécution, puisque l'annotation nécessitera plus de temps et l'intervention de personnes sans doute plus qualifiées.

L'exigence de réutilisabilité concerne la possibilité pour un schéma d'annotation d'être appliqué à des textes dans d'autres langues que celle pour laquelle il a été élaboré. Elle rejoint l'exigence précédente dans le sens où les catégories devraient en théorie être suffisamment larges pour correspondre à tous les systèmes linguistiques. Elle est néanmoins peu mise à l'épreuve, les schémas d'annotation étant souvent créés pour un projet précis, à l'exception notable du système créé par Granger à l'Université de Louvain, qui est disponible commercialement. Il est possible que certaines des catégories utilisées soient transposables, comme celles concernant les grands domaines linguistiques des erreurs (ex. : lexique, grammaire, pragmatique, etc.), et que d'autres soient spécifiques à une langue ou une paire de langue.

Si le système d'annotation et l'outil d'exploitation qui lui est adossé font preuve de suffisamment de flexibilité, comme le stipule la troisième exigence énoncée par Granger, les catégories et étiquettes peuvent être adaptées en fonction des caractéristiques d'une langue différente, ou bien selon les besoins de recherche précis d'un projet (Díaz-Negrillo et Fernàndez-Domínguez, 2006 : 90).

Enfin, la quatrième exigence concerne la cohérence qui doit exister entre les annotations effectuées par différentes personnes afin de garantir la fiabilité de l'annotation. Granger

(2003 : 467) préconise à cet effet la création d'un manuel détaillé donnant la définition des différentes étiquettes et catégories. La question de la subjectivité est un des points les plus épineux de la création de corpus d'interlangue annotés, notamment en raison des problèmes de standards et de normes linguistiques que nous avons déjà soulevés.

Avant d'évoquer les autres problématiques liées à l'annotation des erreurs, telles que les éléments d'information supplémentaires qui peuvent faire partie du schéma ou bien le choix entre annotation exhaustive et annotation ciblée, nous souhaitons revenir plus en détail sur les taxonomies, ou systèmes de catégorisation, qui sont utilisés pour l'annotation. Du point de vue du domaine de l'acquisition des langues secondes, on distingue traditionnellement les systèmes de catégorisation fondés sur l'observation des phénomènes de surface et ceux reposant sur les niveaux et catégories linguistiques en lien avec les erreurs. Le choix du type de catégorisation ainsi que de son organisation structurelle en étiquettes constituent la base d'un schéma d'annotation.

D'après l'étude de Díaz-Negrillo et Fernàndez-Domínguez, la plupart des systèmes de catégorisation pour l'annotation sont fondés principalement sur des catégories linguistiques, avec l'inclusion de certaines informations liées aux phénomènes de surface suivant les systèmes. L'organisation structurelle des annotations varie également largement parmi les systèmes étudiés. Dans les paragraphes suivants, nous présentons le choix de taxonomie et d'organisation des quatre schémas les plus développés (voir Tableau 6 ci-dessus).

Le schéma développé pour le *Cambridge Learner Corpus* est basé sur une combinaison de catégories linguistiques et de descriptions des phénomènes de surface (Nicholls, 2003 : 573-575). Les étiquettes comprennent en général deux lettres, la première représentant un phénomène de surface (ex. : choix erroné de forme, élément manquant, élément en trop, etc.), et la seconde identifiant la partie du discours correspondant au mot à substituer, et non au mot erroné. À ces combinaisons viennent se rajouter des étiquettes spécifiques aux erreurs de ponctuation, de dénombrabilité, de "faux-amis", et d'accord. Le schéma comprend également 13 étiquettes pour d'autres types d'erreur, comme les erreurs de registre de langue, de collocation, ou d'ordre des arguments du verbe. Enfin, l'étiquette *Compound Error* est attribuée aux erreurs complexes impossibles à catégoriser. Le schéma utilise les normes XML (*Extensible Mark-up Language*) et intègre une correction du segment. Voici un exemple de segment comprenant une erreur d'organisation des arguments à la suite du verbe (AS), et que nous empruntons à Nicholls :

> [<#AS>it caused trouble to me|it caused me trouble</#AS>].

L'erreur concernée est le segment [it caused trouble to me], pour lequel la correction proposée est [it caused me trouble], qui consiste en un réagencement des arguments à la suite du verbe, représenté par l'étiquette AS. Cette étiquette est présentée dans des balises typiques des normes XML : une balise ouvrante indiquant le début d'une annotation (<#AS>) et une balise fermante en indiquant la fin (</#AS>). Erreur et correction sont séparées par une barre verticale.

Dans le projet *FreeText*, la particularité de l'annotation est sa catégorisation en trois parties, qui se traduit par une organisation en trois niveaux. Les catégories employées donnent des informations concernant le domaine linguistique général de l'erreur (ex. : morphologie, syntaxe, lexique, etc.), la "catégorie" de l'erreur, c'est-à-dire ce que nous avons appelé le "phénomène linguistique" en question (ex. : voix, ordre des mots, accord en nombre, etc.), et enfin la partie du discours concernée. Certaines étiquettes d'ordre linguistique peuvent être combinées avec des descriptions de surface. Ainsi, les erreurs liées à la syntaxe et à la ponctuation peuvent également porter mention de la nature du phénomène de surface observé, comme par exemple une erreur dans l'ordre des mots, ou une confusion entre deux mots (Díaz-Negrillo et Fernàndez-Domínguez, 2003 : 93-94). Le schéma utilise également les normes XML, et intègre la correction du segment. Voici un exemple de segment de français L2 comprenant une erreur de grammaire (G) liée à l'accord en genre (GEN) touchant un adjectif (ADJ). L'exemple est emprunté à Granger (2003 : 470) :

> [L'héritage du passé est très <G><GEN><ADJ> #fort$ forte </ADJ></GEN></G> et le sexisme est toujours présent].

On voit également ici l'intégration des étiquettes dans des balises ouvrantes et fermantes, mais présentées à la suite plutôt que combinées.

Le système développé à l'Université Catholique de Louvain pour le corpus *ICLE* a servi de base à la création du schéma d'annotation de *FreeText*, et partage donc des caractéristiques avec ce dernier. Il comporte entre deux et trois niveaux hiérarchiques selon les catégories, et les étiquettes ne sont pas en format XML. Le premier niveau, comme dans le projet *FreeText*, est composé d'étiquettes donnant des informations sur le domaine linguistique général de l'erreur (ex. : lexique, grammaire, style, etc.) ainsi que d'une étiquette concernant les phénomènes de surface, comme des mots manquants ou des erreurs d'ordre des mots (Dagneaux et al., 1998 : 166). Le second niveau est subdivisé en plusieurs catégories selon la catégorie principale de l'erreur. Dans le cas des erreurs de grammaire, l'étiquette inclut le groupe syntaxique de l'erreur, puis le phénomène linguistique concerné, comme le temps, la

voix ou l'utilisation des auxiliaires. Une correction est fournie dans le cas de ce schéma également. Voici un exemple cité par Dagneaux et al., pour deux segments erronés ; le premier concerne une erreur de forme (F) liée à l'utilisation d'un "faux-ami" (S), et la seconde est un exemple d'erreur lexico-grammaticale liée au groupe verbal (V) et relevant de l'utilisation des prépositions (PR) :

> [a herd of tiny (FS) braun $brown$ cows was grazing quietly, (XVPR) watching at $watching$ the toy train going past].

Dans ce schéma, les annotations ne sont pas signalées par des balises XML, mais les corrections sont indiquées à l'aide de symboles spécifiques ($). Les étiquettes sont quant à elles données entre parenthèses, et combinées entre elles.

Le dernier schéma que nous présentons ici est celui qui a été élaboré pour l'annotation du *Corpus of Japanese Learner English*. Parmi les quatre systèmes présentés, il est celui qui fait la plus grande utilisation des catégories linguistiques, et se concentre uniquement sur les erreurs morphologiques, grammaticales et lexicales (Izumi et al., 2005 : 75). Les étiquettes, également créées en format XML, reposent sur deux niveaux de catégories : la partie du discours du mot erroné, et le système grammatical concerné. Ces dernières sont complétées par une série de cinq étiquettes pour les erreurs qui ne peuvent pas être catégorisées uniquement en termes de catégories linguistiques, comme par exemple les mots donnés en japonais (L1) au lieu de l'anglais (L2), les erreurs dans l'ordre des mots, ou les segments incompréhensibles. L'annotation inclut une correction du segment. L'exemple suivant, emprunté à Izumi et al., montre l'annotation d'une erreur portant sur un nom (n) dans le domaine de l'accord en nombre (num) :

> [I belong to two baseball <n_num crr="teams">team</n_num>].

On remarque la présence de balises XML ouvrantes et fermantes, la première contenant également la correction, qui consiste au passage d'un nom du singulier au pluriel, en accord avec le nombre donné ([two teams] et non [*two team]).

Comme nous venons de le voir, un schéma d'annotation peut également contenir d'autres éléments d'information, le plus fréquent étant une proposition de correction du segment. Certains projets d'annotation, comme le projet *MELD* (*Montclair Electronic Language Database*), incluent d'ailleurs uniquement la correction du segment, dans le but de rendre l'annotation plus rapide et de réduire les différences dans les classifications proposées par des annotatrices et annotateurs différents (Fitzpatrick et Seegmiller, 2004 : 226, cité par Díaz-Negrillo et Fernàndez-Domínguez, 2006 : 97). Les créateurs et créatrices du projet *FALKO*

(*Fehlerannotiertes Lernerkorpus*), partant du principe qu'il est impossible de ne pas interpréter les erreurs lors de leur annotation, ont inclus la possibilité de donner plusieurs interprétations des erreurs grâce à une organisation des annotations effectuée non pas de manière linéaire mais en profondeur, à l'aide d'un tableau à plusieurs niveaux (Lüdeling et al., 2005). Tetreault et Chodorow, dans leur projet d'annotation ciblée sur les erreurs liées aux prépositions, incluent dans le schéma d'annotation une auto-évaluation de la confiance de l'annotateur ou annotatrice en la validité de son propre jugement sur une échelle de 2 points (1 = confiance faible, 2 = confiance élevée) (2008 : 28).

Ce dernier point nous amène à évoquer la question de l'exhaustivité de l'annotation, c'est-à-dire le choix d'annoter toutes les erreurs ou bien uniquement certaines d'entre elles. Parmi les projets que nous avons évoqués, seul le dernier fait le choix de ne cibler qu'un seul type d'erreur. D'après Izumi et al, le projet du *NICT JLE* inclut également des restrictions sur les erreurs qui sont relevées, dans le sens où seules les erreurs de morphologie, de grammaire et de lexique font l'objet d'une annotation (op. cit. : 75). La majorité des schémas d'annotation propose donc une annotation exhaustive de tous les types d'erreur. Leacock et al. (2010 : 85) soulignent les avantages et les inconvénients de chaque approche, tout en indiquant que, si les schémas exhaustifs sont hautement informatifs, le nombre d'étiquettes et de valeurs à retenir pour un schéma exhaustif risque de mener à des incohérences entre les différentes personnes ayant fourni des annotations. Le choix d'un schéma ciblé simplifie la phase d'annotation, mais empêche d'utiliser le corpus pour des recherches liées à d'autres erreurs que celles que l'on a choisi d'annoter. Les objectifs de la création d'un schéma et de l'annotation des erreurs en elle-même (ex. : recherche en acquisition des langues secondes vs. création d'un outil de détection automatisée pour un type d'erreur) vont orienter le choix du degré d'exhaustivité de l'annotation.

En dehors des choix pratiques et scientifiques évoqués ci-dessus, le domaine de l'annotation des erreurs d'interlangue est également confronté à des problématiques générales en lien avec les technologies du traitement de la langue. La subjectivité de l'annotation est un de ces problèmes. La création de consignes précises et la recherche d'un équilibre entre la quantité d'information fournie et la maniabilité du schéma sont les principales solutions proposées en réponse à ce problème. La seconde problématique importante liée à l'annotation concerne sa standardisation. Díaz-Negrillo et Fernàndez-Domínguez (op. cit. : 86) relèvent le manque de standardisation entre les différents systèmes existants, notant que chaque projet mène le plus souvent au développement d'un système *ad hoc*. Tout en étant le signe d'un

manque de références claires pour l'annotation des erreurs, ceci marque également le dynamisme du domaine ainsi que la volonté d'explorer différentes approches émergentes. Meurers (2010) souligne la nécessité d'avoir accès à des standards d'annotation afin de maximiser la cohérence et de pouvoir évaluer plus efficacement les différents systèmes de traitement des langues reposant sur de telles annotations. Leacock et al. (op. cit. : 90) reconnaissent également la nécessité de la création de tels standards pour faire évoluer un domaine qu'il jugent n'en être qu'à ses balbutiements.

La catégorisation des erreurs dans notre recherche

Lors de la phase de catégorisation des erreurs, notre objectif est d'obtenir une vue d'ensemble des phénomènes linguistiques posant des difficultés au public cible que nous avons sélectionné. Après avoir passé en revue les projets de recherche en correction grammaticale automatisée existants, et pris en compte les capacités des correcteurs grammaticaux disponibles, nous sélectionnons les erreurs à traiter en fonction de nos exigences et du cadre théorique et pratique de nos travaux.

Prenant en compte cet objectif ainsi que les recherches effectuées sur la catégorisation des erreurs, nous avons identifié trois exigences pour le choix d'un système de catégorisation pour notre recherche :

- le système doit être adapté à la description de n'importe quel segment erroné : notre relevé ne cible pas un type d'erreur précis mais toutes les erreurs, et correspond donc aux schémas d'annotation exhaustifs,
- le système doit permettre de repérer les erreurs d'un point de vue linguistique, afin de préparer la modélisation linguistique des phénomènes concernés,
- les catégories doivent avoir une cohérence interne, avec des catégories de second niveau constituant un sous-ensemble des catégories de premier niveau, plutôt que la superposition d'un autre type de critère.

Nous souhaitions également rendre le système réutilisable pour toutes les autres langues, mais cette exigence est en conflit avec notre exigence de traitement linguistique détaillé : la référence à des phénomènes et systèmes linguistiques caractéristiques de l'anglais rend difficile la transposition de l'ensemble des catégories à d'autres systèmes linguistiques. Le premier niveau de catégorie est cependant transposable à des langues présentant des systèmes

similaires à l'anglais, comme les langues germaniques et latines. Les niveaux supplémentaires peuvent être adaptés en fonction des phénomènes spécifiques à la langue étudiée.

Les systèmes de catégorisation présentés ci-dessus utilisent une combinaison des différents critères suivants :

- le domaine général de l'erreur (ex. : morphologie, syntaxe, orthographe, etc.),
- la partie du discours concernée par l'erreur ou par sa correction,
- les systèmes linguistiques en lien avec l'erreur (ex. : accord, structure des arguments du verbe, etc.),
- la description des phénomènes de surface (ex. : omission d'un mot, mot en trop, ordre des mots, etc.).

Pour qu'un système de catégorisation soit le plus informatif possible, il devrait en théorie comporter toutes ces informations ; dans le cadre de notre recherche, où l'annotation des erreurs ne constitue pas l'objectif principal, nous nous limitons à une sélection de ces critères. Nous souhaitons cependant leur ajouter une dimension qui n'a pas été prise en compte dans les systèmes présentés par Díaz-Negrillo et Fernàndez-Domínguez, c'est-à-dire la question du domaine des erreurs. Cette notion, que nous avons déjà évoquée, a été introduite par Lennon (1991 : 191), qui en donne la définition suivante (déjà citée précédemment mais que nous reproduisons ici pour plus de clarté) : "the rank of the linguistic unit which must be taken as context in order for the error to become apparent". Cet élément d'information est donc nécessaire pour la détection de l'erreur, et peut se révéler important dans le cadre de travaux de recherche dans le domaine de la détection et de la correction automatisées.

Notre système de catégorisation se présente en deux niveaux hiérarchiques. Le premier niveau de catégorisation est celui du domaine de l'erreur selon la définition de Lennon, représenté par le groupe syntaxique (ou syntagme) concerné (groupe nominal, groupe adjectival, groupe prépositionnel, groupe verbal). Nous ajoutons à ces quatre catégories celle de "Proposition et phrase". Les groupes syntaxiques étant enchâssés, chaque erreur est classée dans la catégorie correspondant à la plus petite structure supérieure nécessaire pour détecter l'erreur. Par exemple, une erreur dans la morphologie d'une structure comparative est classée dans la catégorie du groupe adjectival, car la présence d'un nom n'est pas nécessaire pour la repérer, comme dans l'exemple suivant :

[22] **more stronger*

Une erreur dans le placement d'un adjectif par rapport à un nom, par exemple s'il est placé après ce dernier alors qu'il devrait le précéder, est par contre classée dans le domaine du groupe nominal. Les erreurs liées aux accords entre sujet et verbe sont classées dans le domaine de la proposition, puisqu'elles se produisent à la jonction entre deux groupes syntaxiques (GN et GV), formant ensemble la base d'une proposition.

Le classement par domaine permet d'éviter le plus souvent les possibilités d'interprétations multiples, au moins au premier niveau de catégorisation. Prenons l'exemple d'une erreur portant sur le choix d'une préposition suivant un verbe (nous soulignons la préposition pour plus de clarté, mais l'astérisque indiquant l'agrammaticalité du segment est placé devant le domaine concerné) :

[23] *the agreement needs to be *translated in French.*

Celle-ci pourrait être interprétée comme relevant du choix des prépositions, donc du groupe prépositionnel, ou des contraintes de sélection liées à un verbe, donc du groupe verbal, mais elle est interprétée ici comme faisant partie des erreurs du domaine verbal puisque la présence du verbe est nécessaire pour détecter l'erreur dans le choix de la préposition. L'ambiguïté peut néanmoins être reportée sur le classement en catégories de second niveau.

Le second niveau hiérarchique concerne le système ou phénomène linguistique concerné par l'erreur. Comme dans les systèmes de catégorisation que nous avons présentés, les catégories incluses dans ce niveau se révèlent assez hétérogènes, et certaines d'entre elles sont des descriptions de phénomènes de surface, notamment en ce qui concerne l'ordre et la sélection des mots. Ce second niveau inclut également une catégorie "Lexique" pour les domaines du groupe nominal et du groupe verbal, car ce type d'erreur ne peut généralement pas être classé différemment. Nous avons créé une catégorie "Autres erreurs" pour les erreurs les plus difficiles à classer dans les domaines pour lesquels cela s'est avéré nécessaire, mais nous y avons inclus aussi peu de segments que possible. Cette catégorie est aussi utilisée pour les segments dont l'agrammaticalité ou l'inacceptabilité provient de la combinaison de plusieurs caractéristiques qui ne peuvent être séparées. Ce type d'erreur correspond aux "erreurs composées/complexes" (*Compound Errors*) de la classification effectuée dans le cadre du *Cambridge Learner Corpus* (Nicholls, 2003 : 573-575).

Les catégories de second niveau sont parfois présentées en deux catégories plus fines selon les problématiques que nous avons souhaité observer ; par exemple, dans la catégorie

"Modification" du domaine "Groupe Nominal", nous avons distingué les erreurs liées à la modification par des adjectifs de celles liées à la modification par des noms.

Le tableau 7 présente l'ensemble des catégories de premier et second niveau, accompagnées d'exemples tirés du corpus. L'astérisque est placé au début du domaine identifié pour l'erreur, selon la définition de Lennon, et le mot ou les mots qui portent l'erreur sont soulignés pour plus de clarté.

1er niveau : Domaine	**2ème niveau : Système**	**Exemple**
Groupe Nominal	Détermination	*the intricate and tricky problem of *the feminism*
	Modification (Adjectif)	**a way decent*
	Modification (Nom)	**information extraction technology results*
	Choix et utilisations des prépositions	*a *rise of nationalism*
	Choix des pronoms	*but also *this of a new political configuration*
	Accord	*a folder containing *three list of files*
	Lexique	*to ease *the fulfill of some enriched aircraft documentation structures*
	Autres erreurs	**the Pagnol's dialogue*
Groupe Adjectival	Construction du comparatif et du superlatif	*The competition coming from Japan is a lot *more stronger*
	Choix des prépositions	*This is *typical for the 20th century*
	Autres	*One trait of her personality is *very much striking*
Groupe Prépositionnel	Choix des prépositions	**Since a few years*
Groupe Verbal	Placement des modifieurs après le verbe (Adverbe)	*To *index efficiently the soundtrack of multimedia documents*
	Placement des modifieurs après le verbe (Autres)	*we might *change a bit the statement proposed*
	Placement des compléments après le verbe	*I will *put on my website some mp3 files*
	Morphologie du GV	*I *will sent all the information required*
	Utilisation et construction de la négation	**Have you never opened a newspaper?*
	Choix et présence des prépositions	*this system *results of the study*
	Lexique	*to *remind a document*
	Autres erreurs	*he should also *feel inhabitant of the European nation*

Proposition et phrase	Construction des phrases interrogatives	*We are going to become Europeans very soon. *What implies this?*
	Construction des propositions subordonnées à verbe fini	*This introduction roughly defines what a procedure is, *what is its structure in linguistic and conceptual terms.*
	Construction des propositions subordonnées à verbe non-fini	*who has *imagined to divide the world into three categories of men*
	Accord sujet-verbe	*the *people who enjoys the long summer*
	Choix du temps	*Can you tell me quickly if I *receive a parcel this Thursday?*
	Choix de l'aspect	*They say that man has no opportunity to dream or imagine and that he *becomes a robot*
	Présence et choix de l'auxiliaire modal	*the idea according to which there would be a specific problem*
	Cohérence des pronoms	*Belgium is the capital of Europe and she is federal*
	Planification de l'information	*for your culture *august 15 in France, it is Assumption day*
	Lexique	**Least, but not last*
	Autres erreurs	**it is already of habit*

Tableau 7. Catégorisation des erreurs

Deux autres dimensions pourraient venir compléter ces catégories dans le cadre de la mise en place de l'annotation numérisée des erreurs. Des indications concernant l'étendue des erreurs, c'est-à-dire le mot ou groupe de mots qui doit être ajouté, supprimé, modifié ou déplacé afin de corriger l'erreur (Lennon, 1991 : 191), permettraient d'obtenir des informations croisées sur les mots qui sont sujets à erreur, et dans quelles configurations syntaxiques. Il serait par exemple possible d'évaluer la fréquence des erreurs de prépositions en lien avec les contraintes de sélection lexicale des noms, des adjectifs et des verbes, par rapport aux erreurs liées au choix des prépositions locatives. Par ailleurs, même si les catégories de second niveau font dans certains cas référence directement ou indirectement au domaine linguistique général des erreurs (ex. : lexique, accord sujet-verbe, construction du comparatif, placement des modifieurs), nous souhaiterions enrichir la catégorisation d'indications précises à ce sujet, en indiquant si l'erreur relève du lexique, de la syntaxe, de la morphologie, de la morphosyntaxe, ou de la sémantique.

Le tableau 8 présente la distribution des erreurs selon les quatre sous-corpus, ainsi que leur densité pour 1000 mots et leur fréquence, indiquée par le nombre de mots "entre" chaque erreur. Les taux ont été arrondis à l'entier inférieur ou supérieur en fonction de la première décimale :

Type de corpus	Nombre d'erreurs	Taille du corpus (nombre de mots)	Erreurs/1000 mots	1 erreur/x mots
Publications	210	40 020	5	191
Interlangue	137	32 430	4	237
Courriels	270	17 267	16	64
Rapport	44	11 177	4	254
TOTAL	**661**	**100 894**	**7**	**152**

Tableau 8. Distribution et fréquence des erreurs selon les sous-corpus

Nous avons mentionné précédemment que le sous-corpus des courriels contenait des productions d'un niveau inférieur à celui des autres sous-corpus. Ceci est confirmé par l'observation de la distribution des erreurs : la fréquence des erreurs est trois fois plus importante dans les courriels que dans les autres documents, pour l'ensemble desquels on retrouve en moyenne cinq erreurs pour 1000 mots et une erreur tous les 214 mots. Cette distinction peut également s'expliquer par le fait que la rédaction de courriels est plus spontanée et soumise à moins de contrôle que les autres productions incluses dans le corpus (ex. : degré d'attention lors de la rédaction, relectures, vérifications grammaticales). Les productions d'interlangue, qui proviennent de copies issues d'examens pour lesquels la maîtrise de l'anglais est évaluée en priorité, présentent le plus faible taux d'erreurs, avec une erreur tous les 246 mots.

Le tableau 9 montre la répartition des erreurs dans les cinq domaines identifiés et selon les quatre sous-corpus :

Corpus / Catégorie	Publications	Interlangue	Courriels	Rapport	TOTAL
G. Nominal	97	57	65	19	**238**
G. Adjectival	4	9	3	-	**16**
G. Prép.	3	4	4	-	**11**
G. Verbal	57	39	105	17	**218**
Prop. et Phrase	49	28	93	8	**178**
TOTAL	**210**	**137**	**270**	**44**	**661**

Tableau 9. Distribution des erreurs selon les catégories

On remarque un fort déséquilibre entre les domaines du Groupe Nominal, du Groupe Verbal, de la Proposition et de la Phrase, qui regroupent 95,9 % des erreurs, et ceux du Groupe Adjectival et du Groupe Prépositionnel, qui forment les 4,1 % restants. Ce résultat est

néanmoins loin d'être surprenant, puisque la présence d'un groupe verbal et d'un groupe nominal est essentielle pour la construction des propositions et phrases, les adjectifs et les groupes prépositionnels représentant des éléments largement optionnels.

Par ailleurs, le critère de catégorisation choisi pour le premier niveau, c'est-à-dire le niveau hiérarchique du segment qui doit être pris en compte pour détecter une erreur, implique souvent de classer l'erreur dans un domaine plus large que celui du mot qu'elle concerne. Ceci explique le faible nombre d'erreurs classées dans le domaine du Groupe Prépositionnel, alors que l'utilisation des prépositions est une source importante d'erreurs (Leacock et al., 2010 : 20) : une erreur dans le choix d'une préposition est le plus souvent visible uniquement si le groupe verbal ou nominal qui l'englobe est pris en compte. L'absence d'un domaine Groupe Adverbial découle également de ce choix de critère, puisque la plupart des erreurs liées aux adverbes concernent leur placement après le verbe, ce qui implique de les classer dans le domaine du Groupe Verbal.

Les six types d'erreur les plus fréquents dans l'ensemble du corpus sont les suivants :

1. Groupe Nominal/Détermination, 14,2 % des erreurs,
2. Groupe Verbal/Choix et présence des prépositions, 10 % des erreurs,
3. Groupe Nominal/Modification (Nom), 6,8 % des erreurs,
4. Groupe Verbal/Lexique, 5,4 % des erreurs,
4. (*ex aequo*) Groupe Verbal/Modification (Adverbe), 5,4 % des erreurs,
6. Proposition et Phrase/Accord sujet-verbe, 5,3 % des erreurs.

Nous avons choisi de présenter six types d'erreur plutôt que cinq, les trois derniers types ayant des fréquences très proches, et deux des catégories ayant la même fréquence. Le tableau 10 montre la présence de ces types d'erreur selon les quatre sous-corpus :

Corpus / Catégorie	Publications	Interlangue	Courriels	Rapport	TOTAL
Détermination	**33**	**33**	19	**9**	94
Prep. G.V.	13	7	**40**	6	66
Modification (Nom)	**33**	4	4	4	45
Lexique G.V.	9	5	22	-	36
Modif. GV (Adverbe)	13	16	1	6	36
Accord sujet-verbe	7	4	22	2	35
Total erreurs	/210	/137	/270	/44	/661

Tableau 10. Distribution des cinq types d'erreur les plus fréquents selon les sous-corpus

Ce tableau nous permet de constater que les erreurs fréquentes ne sont pas distribuées de manière uniforme dans les quatre sous-corpus. Nous avons indiqué en caractères gras la catégorie regroupant le plus d'erreurs pour chaque ensemble.

Les erreurs liées à la détermination représentent 24,1 % des erreurs dans le corpus interlangue, mais seulement 7 % des erreurs du corpus de courriels. Les erreurs liées à l'utilisation des prépositions dans le groupe verbal sont plus fréquentes dans le corpus de courriels, pour lequel elles représentent 14,8 % des erreurs, contre 5,1 % des erreurs dans le corpus d'interlangue.

Les erreurs dans l'utilisation des noms en tant que modifieurs sont plus fréquentes dans les documents techniques et scientifiques, avec 15,7 % des erreurs dans les publications et 9,1 % dans les extraits de rapport technique, mais seulement 2,9 % et 1,5 % dans les productions d'interlangue et les courriels. Ce déséquilibre s'explique par le fait que les structures N+N sont principalement utilisées dans les textes scientifiques (cf. Pastor-Gómez, 2011), comme nous le verrons dans la section 2.3.1 du deuxième Chapitre ; il est donc normal de ne pas retrouver ces erreurs en aussi grande quantité dans les courriels et productions d'interlangue, puisque la structure qui est source de difficulté y est peu utilisée.

Les erreurs liées au placement des adverbes après le verbe se retrouvent dans des proportions importantes dans les sous-corpus de publications, de productions d'interlangue, et d'extraits de rapports techniques, pour lesquels elles représentent respectivement 6,2 %, 11,7 % et 13,6 % des erreurs. Elles sont par contre quasiment inexistantes dans le corpus de courriels, avec 0,4 % des erreurs. Ceci peut s'expliquer par le caractère bref et concis des courriels, les adverbes étant des éléments le plus souvent optionnels dans les phrases.

Par ailleurs, les erreurs de lexique dans le groupe verbal et d'accord sujet-verbe sont plus fréquentes dans le corpus de courriels que dans les autres corpus, avec 8,1 % des erreurs pour chacune de ces catégories, contre entre 0 % et 4,3 % maximum dans les autres sous-corpus. Ces résultats sont en cohérence avec le niveau de contrôle moins élevé appliqué à la rédaction de courriels, ainsi qu'avec le niveau de maîtrise généralement plus faible dans ce sous-corpus.

Nous avons évoqué précédemment le fait que le système de catégorisation choisi ne permettait pas de représenter les erreurs liées aux prépositions dans une seule catégorie. Étant donné que ces erreurs, avec les erreurs d'articles, sont parmi les plus étudiées en raison de leur fréquence dans les productions non-natives (cf. Leacock et al., 2010 ; De Felice et Pulman, 2009 ; Hermet et Désilets, 2009), nous avons recherché la fréquence des erreurs de

prépositions de manière transversale. On retrouve ainsi 111 erreurs de ce type dans le corpus, soit 16,8 % du total des erreurs, ce qui les placerait en première place, juste avant les erreurs liées à la détermination. Nous revenons sur ces erreurs dans la section 1.2.3a consacrée à la sélection des erreurs à traiter.

Les résultats de la classification des erreurs énoncés ci-dessus sont en cohérence avec les résultats habituels de projets d'analyse des erreurs comparables, et dans lesquels les erreurs en lien avec la détermination, les prépositions et le lexique occupent une place importante. Les différents corpus font apparaître des distinctions liées au genre des documents pris en compte. Parmi les cinq types d'erreur les plus fréquents, deux d'entre eux font l'objet de nombreuses études (détermination, choix des prépositions), un d'entre eux correspond à des erreurs traitées par les correcteurs grammaticaux existants (accord sujet-verbe, cf. section 3.2.1), et un autre correspond à des erreurs notoirement difficiles à traiter automatiquement (lexique). Nous aborderons les conséquences de ces résultats sur la sélection des erreurs dans la section 1.2.3.

Proposition de schéma d'annotation

L'annotation systématique des erreurs du corpus n'a pas été effectuée dans notre étude, mais une réflexion a été menée concernant le schéma qui devrait être utilisé. Nous avons également annoté manuellement un ensemble évaluatif d'erreurs, correspondant aux erreurs données en illustration des catégories dans le tableau 7. Elles sont disponibles en annexe (Annexe 1). L'objectif de l'annotation est ici en partie de fournir des informations sur les erreurs pouvant servir à la recherche sur l'acquisition des langues secondes, comme dans tout projet d'analyse des erreurs, mais prioritairement de donner des indications pouvant être utilisées lors de la création de règles de détection et correction automatisées.

Le schéma que nous proposons utilise un format XML standard enrichi d'attributs accompagnés de leur valeur. XML, *Extensible Markup Language* ou "langage de balisage extensible" en français, est un langage informatique utilisant des balises encadrées par des chevrons ouvrants et fermants ([< >]. Les balises contiennent un identifiant, ainsi que des informations qui sont ici des attributs du segment annoté, et qui doivent être accompagnées d'une indication de valeur, numérique ou sous forme de texte. Comme nous l'avons vu précédemment, le format XML est souvent utilisé pour l'annotation en traitement automatisé des langues en raison de sa facilité d'utilisation, de sa grande adaptabilité, et de sa portabilité (les annotations faites en XML pouvant être facilement intégrées à de nombreux systèmes). Nous l'utilisons ici pour ces raisons précises.

Nous utilisons certains des éléments présents dans d'autres projets, comme l'inclusion de propositions de correction et l'évaluation du degré de certitude de la personne effectuant l'annotation. Le schéma comporte deux parties enchâssées : la caractérisation de l'erreur (<error>), et la caractérisation de sa ou ses corrections (<correction>). La zone d'erreur correspond au domaine de l'erreur, notion que nous avons définie précédemment. La zone de correction est intégrée à la zone d'erreur. Il est possible d'ajouter autant de zones de correction que nécessaire, suivant le nombre de possibilités de correction existantes.

La structure des attributs a été définie afin de permettre aux annotations d'être utilisées dans un cadre d'argumentation, par exemple pour la sélection d'une correction lorsque plusieurs sont possibles. De plus, ils permettent de faire apparaître les incertitudes dans les jugements de grammaticalité ou dans le choix des corrections à proposer, car, comme nous l'avons vu dans la section 1.1.2, il est parfois difficile de juger de la présence d'une erreur sans aucune ambiguïté. Les tableaux 11 et 12 suivants présentent ces attributs et les valeurs qui leur sont associées pour chaque zone, ainsi que l'explication de leur signification. Les termes utilisés pour définir les valeurs et attributs sont en anglais afin de les rendre plus facilement transférables et communicables à un large public :

Attributs	**Valeurs**	**Explications**
comprehension	de 0 à 2	Évalue l'intelligibilité du segment malgré la présence de l'erreur. Il peut être jugé comme : - incompréhensible ou compréhensible avec un haut degré d'interprétation (0), - compréhensible avec quelque interprétation (1), - complètement clair (2).
grammaticality	de 0 à 2	Évalue la grammaticalité et l'acceptabilité du segment. Il peut être jugé comme : - clairement agrammatical et inacceptable (0), - probablement agrammatical et inacceptable (1), - grammatical et inacceptable (2).
category	gn-det, gn-moda, gn-modn, gn-prep, gn-pro, gn-agr, gn-lex, gn-o ga-comp, ga-prep, ga-lex gprep gv-modadv, gv-modo, gv-comp, gv-morph, gv-neg, gv-prep, gv-lex, gv-o p-int, p-subf, p-subn, p-agr, p-ten, p-asp, p-aux, p-pro, p-inf, p-lex, p-o	Indique la catégorie de l'erreur. Chaque étiquette de valeur inclut le domaine de l'erreur (gn, ga, gprep, gv, p) et la catégorie de 2ème niveau (det, prep, lex, etc.)

Tableau 11. Caractérisation de l'erreur

Attributs	Valeurs	Explications
meaning	yes/no	Indique si le sens du segment risque d'être modifié par la correction proposée.
length	yes/no	Indique si la correction implique de modifier le nombre de mots utilisés dans le segment.
qualif	de 0 à 2	Évalue le degré de certitude de la personne effectuant l'annotation quant à la correction proposée. Cette personne évalue la pertinence de la correction, dont elle peut s'estimer être : - peu sûre (0), - assez sûre (1) - complètement certaine (2)
correct	(segment corrigé)	Propose une correction.

Tableau 12. Caractérisation de la correction

Le paragraphe suivant présente un exemple d'erreur annotée, avec deux propositions de correction :

> They are often based on
>
> <error comprehension="1" grammaticality="1" categ="gn-modn">
> the objects properties
>
> <correction meaning="no" length="no" qualif="2" correct="the object's properties"> </correction>
>
> <correction meaning="no" length="yes" qualif="2" correct="the properties of the object"> </correction>
>
> </error>
>
> and on their potential similarities.

L'annotation nous apprend qu'il s'agit d'une erreur pouvant gêner modérément la compréhension (comprehension="1"), qui semble relever d'un défaut de grammaticalité (grammaticality="1"), et qui appartient à la catégorie des erreurs liées à l'utilisation des noms en fonction de modifieurs dans le groupe nominal (categ="gn-modn"). Deux corrections sont proposées : aucune des deux ne risque de modifier le sens du segment (meaning="no") et elles bénéficient du même degré de certitude de la part de la personne conduisant l'annotation (qualif="2"). La seconde implique une modification de la longueur du segment (length="yes"), ce qui pourrait faire préférer la première correction à une utilisatrice ou un utilisateur du système.

L'utilisation de ce schéma n'est pas sans présenter un certain nombre de difficultés. La plus évidente est sa complexité, qui oblige à confier l'annotation à des personnes ayant des connaissances en grammaire anglaise et en linguistique, en plus d'avoir été formées à utiliser ce schéma. Par ailleurs, cette complexité a un impact sur le temps nécessaire à l'annotation, ce

qui fait augmenter son coût. Si ce schéma venait à être utilisé pour l'annotation systématique des erreurs, nous proposerions la mise en place d'un projet pilote afin d'évaluer la pertinence des différents attributs, dans le but d'envisager des allègements si nécessaires. Trouver un outil existant permettant d'annoter les erreurs en utilisant ce schéma peut également se révéler être une difficulté, mais l'utilisation du format XML étant courante dans les systèmes d'annotation, il ne devrait pas être nécessaire de créer un outil *ad hoc*, comme cela est le cas pour certains projets (cf. L'Haire et Vandeventer Faltin, 2003).

1.2.3 Sélection des erreurs à traiter

Pour commencer, rappelons que les erreurs traitées ici doivent satisfaire trois exigences : leur correction doit être utile au public cible, elles ne doivent pas avoir fait l'objet d'un traitement exhaustif dans d'autres projets de recherche en correction grammaticale automatisée ou par des correcteurs grammaticaux déjà disponibles, et doivent représenter un objectif atteignable dans le cadre d'un projet de recherche fondé sur l'utilisation de méthodes linguistiques. Bien que ces trois exigences soient interdépendantes, nous avons choisi d'accorder la priorité à l'exigence d'utilité, et y avons donc consacré la deuxième partie de ce chapitre. Les sections suivantes reprennent les résultats de l'analyse des erreurs et les confrontent à nos exigences.

a. Traitement des erreurs les plus fréquentes : aperçu de la recherche et perspectives de faisabilité

Le relevé et la catégorisation des erreurs a permis de faire apparaître les difficultés que rencontre notre public cible ; il paraît logique de prendre comme point de départ la liste des erreurs les plus fréquentes afin de sélectionner les erreurs à traiter. Nous passons en revue cette liste afin de déterminer l'adéquation de chaque type d'erreur avec les deux autres exigences que nous avons énoncées, certaines de ces erreurs étant déjà au centre de plusieurs travaux similaires, ou nécessitant une approche ou des ressources qui dépassent notre cadre pratique.

Nous reproduisons ici la liste des six types d'erreur les plus fréquents dans l'ensemble du corpus :

1. Groupe Nominal/Détermination,
2. Groupe Verbal/Choix et présence des prépositions,
3. Groupe Nominal/Modification (Nom),

4. Groupe Verbal/Lexique (*ex aequo*),
4. Groupe Verbal/Modification (Adverbe) (*ex aequo*),
6. Proposition et Phrase/Accord sujet-verbe.

Toutes les erreurs de la catégorie Groupe Nominal/Détermination concernent l'utilisation globale des articles, qu'il s'agisse du choix entre *a* et *the*, de l'absence ou de la présence superflue d'un article, et du choix entre *a* et *an* pour des raisons phonologiques. Dans la section précédente, nous avons indiqué que les erreurs liées à l'utilisation des prépositions dans l'ensemble des domaines représentaient globalement 16,8 % des erreurs du corpus, ce qui les placeraient en tête de cette liste.

La fréquence des erreurs liées aux articles et aux prépositions dans les productions d'interlangue en anglais a été largement remarquée dans les projets d'analyse des erreurs comme dans les projets de détection et/ou correction automatisée. Dans une étude sur les erreurs produites par un groupe d'étudiants et étudiantes de la *City University of New York* ayant des L1 variées, Dalgish (1991 : 46-49) indique des quantités allant jusqu'à 28 % du total des erreurs pour les erreurs liées aux articles, et 18 % pour les erreurs de préposition. Diab (1997 : 74-76) rapporte avoir relevé 28 % d'erreurs liées aux articles et 44 % d'erreurs de préposition dans les productions de personnes arabophones du Liban. Notre propre étude fait apparaître un total de 28,4 % d'erreurs pour ces types combinés, ce qui correspond à la fourchette, certes large, de 20 %-50 % qu'indiquent Leacock et al. pour ces erreurs (op. cit. : 20).

Une part importante des travaux de détection/correction automatisée pour l'anglais L2 est consacrée au traitement de ces deux types d'erreur. Parmi ces travaux figurent notamment ceux de Eeg-Olofsson et Knutsson (2003), Lee et Seneff (2006), Chodorow et al. (2007), De Felice (2008), Tetreault et Chodorow (2008), Yi et al. (2008), Hermet et Désilets (2009), et Gamon (2010). Ces deux types d'erreur ont également fait l'objet d'une tâche commune (*shared task*) en 2012, dont les résultats furent présentés lors du *7th Workshop on Innovative Use of NLP for Building Educational Applications*. Pour résumer, si la maîtrise de l'utilisation des articles et des prépositions constitue une des plus grandes difficultés pour les personnes utilisant et apprenant l'anglais, ainsi que pour le domaine de la détection et la correction automatisée, ces aspects sont également de loin les plus étudiés à l'heure actuelle. Cette constatation amène d'ailleurs Leacock et al. à encourager les recherches sur d'autres types d'erreur moins populaires (op. cit. : 102).

Les erreurs morphosyntaxiques, ou erreurs d'accord et de choix de formes, représentent globalement 11,8 % des erreurs relevées dans notre corpus, si l'on combine les domaines du Groupe Nominal, du Groupe Verbal et de la Proposition/Phrase. Elles sont également fréquentes dans d'autres corpus d'interlangue, d'après le classement des erreurs relevées dans le *Cambridge Learner Corpus* et cité par Leacock et al. (op. cit. : 16). Plusieurs projets de recherche en détection/correction automatisée se sont intéressés à ces erreurs, en particulier les projets dont l'objectif est la création d'un correcteur grammatical complet et autonome, comme *Language Tool* (Naber, 2003), *SpellCheckPlus* (Nadasdi et Sinclair, 2007), ou *ESL Assistant* de Microsoft Research (Gamon et al., 2009).

Elles suscitent cependant un faible intérêt en comparaison des deux autres types d'erreur que nous avons évoqués, d'une part parce que leur correction n'est pas problématique dans le sens où les règles qui régissent les accords et la morphologie flexionnelle en anglais sont la plupart du temps bien identifiées et qu'elles ne représentent donc pas un défi linguistique, et d'autre part parce qu'elles font partie des erreurs les plus efficacement prises en charge par les correcteurs grammaticaux courants, comme le correcteur présent dans le traitement de texte *Word* de Microsoft. Ces deux raisons sont bien évidemment corrélées. Nous revenons dans le Chapitre 3 sur les performances des correcteurs grammaticaux disponibles commercialement.

D'après le classement présenté ci-dessus, les erreurs lexicales dans le domaine du Groupe Verbal constituent la 4ème cause d'erreur (*ex aequo*). Globalement, 11,5 % des erreurs relevées dans le corpus concernent le domaine lexical. Nous regroupons sous cette étiquette les erreurs liées au non respect des caractéristiques lexicales de certains mots (ex. : noms dénombrables/indénombrables, noms invariables, voir [24]), à des confusions sur la nature d'un mot ([25]), à des problèmes de morphologie dérivationnelle ([26]), à des problèmes de collocation ([27]), des "faux-amis" ([28]) et des questions de choix lexical ([29]) :

[24] *for further *informations*

[25] **my apologize*

[26] **to determinate*

[27] **we meet a problem*

[28] *I hope you have *passed a good week-end*

[29] **I am in courses*

On peut se demander si les erreurs lexicales ont vocation à être prises en charge par un correcteur grammatical. Ce type d'outil peut bien sûr être envisagé d'une manière qui englobe cette catégorie d'erreur, avec une vision large de ce qui constitue la "grammaire". Certaines erreurs de ce type, en particulier les erreurs de collocation, ont fait l'objet de recherches approfondies dans le domaine de la correction automatisée, parmi lesquelles on peut citer les travaux de Wible et al. (2003), Chang et al. (2008), Yi et al. (2008), et Liu et al. (2009). Brockett et al. (2006) se sont intéressés à la correction des confusions entre noms dénombrables et indénombrables, et Gamon et al. (2009) aux confusions entre noms et adjectifs.

Les erreurs que nous avons regroupées sous l'appellation "Lexique" sont donc en réalité de natures différentes, et nécessitent un traitement précis adapté à chaque type. Les erreurs de collocation et de choix lexical totalisent ensemble près des deux tiers des erreurs lexicales (respectivement 19 et 23 erreurs pour 68 erreurs lexicales dans le corpus). Les erreurs des quatre autres types sont donc beaucoup moins nombreuses. Gamon et al. (op. cit. : 493) indiquent que les confusions entre noms et adjectifs représentent 3,19 % des erreurs relevées dans le corpus *Japanese Learner English*. D'après l'analyse faite par Leacock et al (op. cit. : 18) du classement des erreurs relevées dans le *Cambridge Learner Corpus*, les erreurs de choix lexical (*content word choice*) sont les erreurs les plus fréquentes. Si leur traitement automatisé apparaît donc comme utile, comme l'indiquent Leacock et al. (op. cit. : 102), on risque de se heurter à une importante difficulté dans la réalisation de cette tâche en raison de l'importance de l'aspect sémantique. De plus, comme les erreurs de collocations, ce type d'erreur appelle un traitement statistique, approche qui ne correspond pas au cadre que nous avons choisi.

De manière *a priori* surprenante, le troisième type d'erreur le plus fréquent dans notre corpus est lié à l'utilisation des noms en tant que modifieurs dans le groupe nominal. À notre connaissance, cette erreur n'a pas été remarquée dans d'autres projets d'analyse des erreurs. Comme nous l'avons vu, la majorité de ces erreurs provient du sous-corpus de publications scientifiques, ce qui est en cohérence avec le contexte d'utilisation des structures N+N (Pastor-Gómez, 2011). Étant donné que les corpus utilisés pour l'analyse des erreurs sont généralement constitués de productions d'interlangue faites dans un cadre pédagogique (copies d'examen, *essays*), il est logique que ces structures, et donc les erreurs qui leur seraient associées, ne soient pas présentes en des proportions suffisamment importantes pour être remarquées. Il est également possible que ces erreurs aient été attribuées à l'utilisation de

noms composés, la frontière entre compositions lexicales nominales et constructions syntaxiques étant ambiguë et donc controversée (voir Chapitre 3).

Le seul projet en lien avec cette problématique dont nous ayons connaissance est le projet *Compounds* (Boucher et al., 1993), dont l'objectif est la création d'un tuteur de langue intelligent pour l'apprentissage de la compréhension et la production des noms, adjectifs et verbes composés anglais à l'usage de francophones. Il ne cible pas uniquement les structures N+N, et, comme son nom l'indique, les structures prises en compte sont considérées comme des créations lexicales. La recherche en traitement automatique des langues s'est en revanche penchée sur la désambiguïsation des composés N+N, mais en prenant comme point de départ les structures bien formées relevées dans des productions en anglais L1. Ce genre de recherche est notamment développé dans les travaux de Fabre (1996) et Buckeridge et Sutcliffe (2002). Les erreurs dans les structures N+N pourraient être traitées dans le cadre de projets de correction des erreurs de collocations, que nous avons déjà évoqués, mais d'après la synthèse présentée par Leacock et al. (2010 : 68-71), la plupart des projets de ce type se concentrent sur les collocations [Verbe + Nom] (ex. : Wible et al., 2003 ; Chang et al., 2008 ; Liu et al. , 2009). Les erreurs liées à l'utilisation des noms en tant que modifieurs dans le groupe nominal constituent ainsi une opportunité d'innovation, et leur fréquence dans notre corpus indique que son traitement automatisé présente un intérêt pour les utilisatrices et utilisateurs de l'anglais.

Les erreurs liées au placement des adverbes dans le domaine du Groupe Verbal arrivent à la quatrième place (*ex aequo*), totalisant 5,4 % des erreurs. Ce pourcentage peut paraître peu élevé, mais il faut garder à l'esprit que les adverbes sont des éléments le plus souvent optionnels (Biber et al., 1999 : 66). Le nombre d'erreurs qui leur est associé court ainsi le risque de ne pas être représentatif des difficultés réelles posées par leur utilisation. Citant l'étude de Duskova (1969), Ellis insiste sur cet aspect : "the frequency of the errors did not necessarily reflect the level of difficulty the learners experienced with different linguistic features, as some features (such as articles) were attempted more often than others (for example, adverbs)" (2008 : 51). Par ailleurs, même le type d'erreur le plus fréquent ne regroupe que moins de 15 % des erreurs.

Si les erreurs de placement des modifieurs dans le groupe verbal ont été étudiées du point de vue du transfert syntaxique (Jarvis et Pavlenko, 2007 : 99), elles n'ont pas suscité d'intérêt important dans le domaine de la correction grammaticale automatisée. À notre connaissance, les seuls travaux portant explicitement sur la correction des erreurs de placement d'adverbes

sont ceux de Meurers et Metcalf (2006), présentés lors de deux interventions orales dans lesquelles les auteurs mentionnent l'existence d'un prototype pour la correction de ces erreurs. Les résultats détaillés de ces travaux n'ont cependant pas été publiés. Le traitement des erreurs liées au placement des adverbes est également mentionné dans Michaud et al. (2000) dans le cadre du projet *ICICLE,* mais dans ce cas également, il n'existe pas d'autre publication dans laquelle les résultats des travaux sont présentés.

La question de la prise en compte de la quantité de ressources et connaissances nécessaires à la détection et à la correction des erreurs est évoquée plus longuement dans le Chapitre 3, qui est consacré à l'implémentation des règles, mais nous pouvons d'ores et déjà amorcer cette réflexion au sujet des deux types d'erreur mentionnés ci-dessus. La première concerne le groupe nominal, ce qui implique d'une part de reconnaître les noms, et d'autre part de détecter la présence des groupes nominaux. La reconnaissance des noms peut être effectuée par le biais d'un lexique, existant ou bien à constituer, en gardant à l'esprit le caractère exploratoire de notre recherche.

La détection des groupes nominaux est plus problématique ; nous y reviendrons dans le Chapitre 3. L'aspect des structures N+N qui est le plus susceptible de remettre en question la faisabilité du traitement de telles erreurs concerne cependant les connaissances nécessaires à la détermination des relations sémantiques qui existent entre les noms de la structure. Ce problème a été soulevé dans les travaux portant sur la désambiguïsation des composés N+N. Cette tâche s'avère difficile lorsqu'elle concerne des composés relevés dans des productions en anglais L1, la rareté de ces composés limitant fortement les données utilisables ; on peut logiquement envisager que leur traitement sera encore plus difficile si l'on s'intéresse aux relations erronées entre les noms. Nous verrons qu'une partie de ces erreurs n'est pas liée aux relations sémantiques existant entre les noms, mais à des empilements abusifs et à des erreurs morphologiques, pour ne citer que ces deux types. Le traitement de ces types d'erreur précis se prête plus facilement à une approche linguistique utilisant des patrons de détection et des règles de réécriture.

Les difficultés en termes de faisabilité potentiellement posées par les erreurs liées au placement des adverbes sont différentes de celles que nous venons d'exposer. Ces erreurs appartenant à la catégorie des erreurs d'ordre des mots (*word order errors*), leur traitement par le biais de patrons et de règles semble *a priori* adapté. On identifie toutefois trois principaux défis en lien avec la question des ressources et connaissances nécessaires. Tout d'abord, ces erreurs se produisant dans le domaine du groupe verbal, il est impératif pour les détecter de

pouvoir identifier un grand nombre de parties du discours différentes, comme les verbes et les auxiliaires, ainsi que les noms et les éléments qui peuvent se trouver dans un groupe nominal. Comme pour les structures N+N, leur reconnaissance peut être effectuée à l'aide de ressources existantes, éventuellement adaptées à nos besoins. Ensuite, la polysémie de certains adverbes représente une difficulté supplémentaire en lien avec les ressources, qui peut néanmoins être surmontée grâce à la création d'un lexique d'adverbes donnant des informations sur leurs différents sens lorsque cela s'avère pertinent.

Afin de passer outre cette difficulté, les règles de détection et correction peuvent être limitées dans un premier temps à un type sémantique fréquent, comme celui des adverbes de manière, ou bien à un adverbe particulièrement usité, comme l'adverbe *also*. La pluralité des facteurs à prendre en compte lors de la détection et de la correction des erreurs de placement, et que nous évoquons dans la section suivante, représente également un obstacle au traitement automatisé des erreurs de placement. Ce dernier peut néanmoins être surmonté par le biais d'une modélisation linguistique précise du placement des adverbes ainsi que par l'observation des schémas d'erreurs relevées dans les productions authentiques.

Ce bref aperçu des travaux de recherches concernant la correction automatisée des erreurs fréquentes révèle que les erreurs liées aux articles, aux prépositions, et à la morphosyntaxe (accords) font ou ont fait l'objet d'un traitement détaillé dans d'autres projets. Le traitement des erreurs lexicales pose quant à lui un problème en termes de faisabilité et d'adéquation à l'orientation grammaticale de la présente recherche. Il semble cependant qu'il y ait un manque de recherches concernant les erreurs dans l'utilisation des noms en tant que modifieurs dans le groupe nominal, particulièrement fréquentes dans les productions scientifiques. Les erreurs dans le placement des adverbes n'ont pas non plus été étudiées en détail du point de vue de la correction automatisée. Lors d'un test comparatif (voir Chapitre 3), un seul des onze correcteurs grammaticaux que nous avons testés a détecté les erreurs dans le placement des adverbes et l'utilisation de noms en fonction de modifieurs dans le groupe nominal parmi l'échantillon d'erreurs tirées de notre corpus que nous leur avons soumis. Ce correcteur a néanmoins uniquement détecté certains des schémas d'erreur, et de manière non-systématique. Ces deux types d'erreur répondent donc en grande partie à notre exigence d'innovation ; ils partagent également un ensemble de caractéristiques qui rend leur traitement intéressant d'un point de vue linguistique. La section suivante s'étend plus longuement sur l'utilité de la correction de ces erreurs et sur les caractéristiques que nous venons de mentionner.

b. La correction des erreurs dans le placement des adverbes et l'utilisation de séquences N+N : intérêt linguistique et didactique

Dans la section précédente, nous avons insisté sur l'utilité de traiter les erreurs liées aux structures N+N en raison de leur fréquence importante dans nos deux sous-corpus de productions scientifiques (publications et rapport). Ces erreurs sont également source d'ambiguïtés dans les groupes nominaux, car il peut être difficile pour les lecteurs ou lectrices de déterminer les relations noyau-modifieur dans un groupe nominal incluant plusieurs modifieurs nominaux, ou une combinaison de noms et d'adjectifs. C'est le cas dans le segment suivant, extrait de notre corpus :

[30] **information system security strategies heterogeneity,*

Celui-ci inclut quatre modifieurs nominaux dans une structure qui peut être considérée comme empilée :

[infomation [system [security [strategies [heterogeneity]]]]],

ou bien enchâssée :

[[[information system] [security strategies]] [heterogeneity]].

Cette représentation n'est qu'un exemple des différentes configurations possibles d'un point de vue syntaxique, selon l'intention de l'auteure ou auteur. Lorsqu'elles sont construites de manière grammaticalement correcte, ces structures représentent déjà une difficulté cognitive pour le ou la destinataire, qui doit reconstruire la relation sémantique entre les deux noms (Pastor-Gómez, 2011). La difficulté est donc augmentée par d'éventuelles erreurs, comme le cas d'"empilement" que nous venons de citer, ou bien d'erreurs dans la sélection des relations sémantiques qui sont possibles dans ces structures :

[31] **the goal failure*

[32] **the discrimination experience.*

Comme nous l'avons indiqué dans la sous-section 1.1.1b "Évaluation des erreurs", l'intelligibilité du segment est un facteur important dans l'évaluation de la gravité d'une erreur. De plus, puisque nous avons souhaité prendre en compte le changement de perspective proposé par la reconceptualisation de l'anglais comme *lingua franca* (cf. 1.1.4d), il est nécessaire d'accorder une place importante au critère d'intelligibilité, ce dernier ayant la priorité dans les situations de communication en ALF. Nous revenons en détail sur l'aspect linguistique des structures N+N dans le Chapitre 3, dans lequel nous passons également en revue les raisons de leur utilisation de plus en plus fréquente, notamment dans les textes

scientifiques et académiques (Pastor-Gómez, 2011). Du point de vue de l'acquisition de l'anglais L2, il est possible que le recours à ces structures compactes soit indirectement lié aux difficultés rencontrées dans l'utilisation des prépositions : la juxtaposition de deux noms permet d'éviter d'avoir à introduire une préposition pour expliciter leur lien sémantique, et ainsi éviter une éventuelle erreur dans le choix de celle-ci.

Les erreurs liées au placement des adverbes dans le groupe verbal peuvent, quant à elles, être problématiques en termes d'intelligibilité lorsque l'interprétation du sens d'un adverbe dépend de son emplacement dans la phrase. Ainsi, certains adverbes peuvent recevoir l'interprétation "manière" ([33]) lorsqu'ils se trouvent en position centrale ou finale dans le groupe verbal, ou l'interprétation "évaluation" ([34]) s'ils sont en position initiale dans la phrase et détachés de cette dernière par une pause prosodique. Ceci est illustré par les deux exemples suivants :

[33] *They were packing their bags sadly.*

[34] *Sadly, they were packing their bags.*

Nous présentons une synthèse détaillée des différents placements possibles en fonction des types sémantiques d'adverbe dans le Chapitre 2. Dans d'autres cas, le placement erroné d'un adverbe ne pose pas de problème d'intelligibilité important pour les destinataires, mais peut néanmoins interrompre la fluidité du texte et compromettre la compréhension du message de manière indirecte.

Il semble que le placement erroné le plus fréquent pour les francophones rédigeant en anglais est en position postverbale à l'intérieur du groupe verbal, c'est-à-dire avant les compléments du verbe ; ce placement est susceptible de créer des confusions mineures quant à la composition du groupe verbal. De plus, le traitement des erreurs dans le placement des adverbes nous semble particulièrement pertinent car il a été montré que ce type d'erreur persiste à un niveau avancé dans les productions de locutrices et locuteurs de langues romanes, dont le français fait partie, ce qui indique l'existence d'un phénomène de transfert entre L1 et L2 (Osborne, 2008 : 127-128) :

> Learners whose L1 has obligatory verb-raising – Spanish, Italian and French – show the strongest tendency to use V-Adv-O order. […] [N]on target-like placement of adverbs continues to appear at a post-intermediate stage of learning, even in highly monitored production.

Par ailleurs, White a avancé l'hypothèse selon laquelle ce type d'erreur ne pourrait être éliminé simplement par la confrontation à des exemples bien formés, car la locutrice ou le locuteur ne peut être assuré de l'agrammaticalité d'autres configurations. White en tire la conclusion suivante : "It appears that this type of error is likely to be persistent, and a candidate for fossilization" (op. cit., 1989 : 154). Ellis la complète en indiquant que la correction de ces erreurs nécessite probablement le recours à un retour correctif (*corrective feedback*) (op. cit. : 847).

Comme il a déjà été mentionné, notre recherche est également soumise à une exigence linguistique, dans le sens où les problématiques choisies ainsi que leur traitement doivent se prêter à une approche linguistique. Nous venons de montrer que les deux types d'erreur sélectionnés répondent à nos exigences d'innovation et d'utilité ; ils présentent également un intérêt du point de vue de la recherche en linguistique, car elles ont trait à des phénomènes de modification répondant à des normes bien réelles pour les locutrices et locuteurs natifs mais néanmoins difficiles à identifier *a posteriori*. Huddleston et Pullum écrivent la chose suivante au sujet du placement des *adjuncts*, qui incluent les adverbes en fonction de modifieurs dans le groupe verbal et la proposition : "Only rather broad and approximate flexible generalisations about adjunct placement and sequence can be made. There is a great deal of variation in use, and features of context, style, prosody, and euphony play a role in some decisions" (2002 : 576). D'après Pastor-Gómez, qui a consacré une thèse et une monographie à l'utilisation des structures N+N, l'augmentation de leur fréquence est un phénomène relativement récent, qui a peu été étudié en tant que phénomène de construction syntaxique et non en tant que composition lexicale (2010 : 1).

Il est également difficile de trouver des indications claires concernant d'une part la forme que peuvent prendre de telles constructions (ex. : le nom modifieur doit-il être au singulier ou au pluriel ? Quel est le nombre de noms modifieurs acceptable ?), et d'autre part sur les relations sémantiques qui peuvent exister entre les deux termes (cf. Downing, 1977 ; Levi, 1978 ; Benczes, 2006). Cette problématique a également été soulevée dans le domaine du traitement automatisé des langues (cf. Costello et al., 2006 ; Nakov, 2007). Le manque de connaissances concernant les normes qui régissent l'utilisation de ces deux outils optionnels représente un défi pour la détection et la correction automatisée, et oblige à résoudre un ensemble de problèmes théoriques et pratiques. Une exploration de ces phénomènes employant différentes approches linguistiques est nécessaire à la mise en place d'une modélisation linguistique permettant de traiter ces erreurs.

Conclusion

L'objectif principal de ce premier chapitre a été de présenter la méthode raisonnée utilisée pour la sélection des erreurs à traiter. Ces erreurs devaient correspondre à un besoin des personnes utilisatrices de l'anglais L2, constituer une innovation, et être en accord avec notre cadre pratique en termes de faisabilité. Ce chapitre a été plus particulièrement consacré au premier de ces critères, et aborde donc la sélection des erreurs avec une approche ascendante, prenant appui avant tout sur les relevés d'erreurs dans un corpus de productions authentiques.

Dans la première partie, nous avons posé les bases méthodologiques de l'analyse dont les résultats sont présentés en deuxième partie. Nous avons exploré les différentes étapes canoniques de la méthode de l'analyse des erreurs, largement reconnue comme valable en tant que méthode de recueil de données. Ceci nous a amenée à introduire des concepts empruntés au domaine de l'acquisition des langues secondes, comme celui d'interlangue, d'utilisateur ou utilisatrice de l'anglais, d'influence translinguistique et de surgénéralisation. La réflexion menée autour de la définition et de la délimitation des erreurs nous a permis de déterminer le type de segment erroné à relever dans le corpus, même si les normes de grammaticalité et d'acceptabilité sont flexibles, comme nous avons pu le voir dans notre exposé de l'étude de l'anglais comme *lingua franca*. Dans la perspective de traiter des erreurs représentant une difficulté pour les utilisatrices et utilisateurs de l'anglais, nous ne pouvions faire l'économie de la prise en compte du statut international de l'anglais, et des conséquences que ce statut peut avoir dans le cadre d'une étude de cette nature.

L'objectif de la deuxième partie de ce premier chapitre a été d'appliquer les étapes pertinentes de la méthode de l'analyse des erreurs à une recherche en détection et correction automatisée. Le corpus constitué pour nos travaux est composé de publications scientifiques, de productions d'interlangue, de courriels personnels et professionnels, et d'extraits d'un rapport technique, pour un total d'environ 100 000 mots. Les documents inclus dans le corpus ont été sélectionnés à partir d'un ensemble de productions représentatives des différentes utilisations de l'anglais L2 à l'heure actuelle. Les erreurs sont relevées manuellement et catégorisées grâce à un système fondé sur deux niveaux de classification, le premier étant le domaine des erreurs (cf. Lennon, 1991) et le second le système linguistique touché par l'erreur.

Le schéma d'annotation proposé pour une éventuelle annotation systématique des erreurs utilise le format XML. Il donne notamment la possibilité d'inclure plusieurs propositions de correction, et intègre une mesure de la certitude de l'annotateur ou annotatrice quant à la

pertinence de la correction proposée. On relève 661 erreurs dans le corpus entier, et leur fréquence moyenne s'élève à 7 erreurs pour 1000 mots, ou une erreur tous les 152 mots environ. La fréquence et les types d'erreur relevés varient en fonction des sous-corpus, mais les résultats obtenus sont en cohérence avec les caractéristiques de ces derniers. Les types d'erreur les plus fréquents correspondent à ceux qui sont mis en avant dans les projets d'analyse des erreurs et de correction automatisée existants : les erreurs liées aux articles et aux prépositions sont parmi les types les plus fréquents, comme les erreurs de choix lexical et celles ayant trait aux accords. Il semble qu'il ait été pertinent de choisir de constituer un corpus original d'utilisation de l'anglais, au lieu d'utiliser un corpus d'interlangue existant, puisque nous avons pu relever des types d'erreur spécifiques à certaines productions, et n'ayant pas été relevés dans ces proportions auparavant.

Après avoir passé en revue les travaux existants dans le domaine de la correction automatisée ainsi que les perspectives de faisabilité pour chaque type d'erreur fréquent, nous avons présenté les deux types d'erreur traités dans notre projet. Ces erreurs touchent toutes les deux des éléments ayant un caractère optionnel : les adverbes sont le plus souvent des modifieurs dans le groupe verbal ou la proposition, et les séquences N+N sont une façon alternative de structurer les groupes nominaux. Outre le fait qu'elles sont la manifestation d'une difficulté pour les utilisateurs et utilisatrices de l'anglais, elles ont également en commun le fait de relever de normes grammaticales et sémantiques floues, ce qui constitue un défi pour la détection et la correction automatisée. La méthode de traitement linguistique envisagée pour surmonter ces difficultés est décrite et mise en application dans le Chapitre 2.

Ces difficultés ont également des conséquences sur la façon dont les règles doivent être implémentées. En ce qui concerne la détection, on prendra soin de limiter les faux positifs ; ces derniers sont de toute façon à éviter lorsque l'on crée des systèmes à destination d'un public apprenant, mais dans le cas de phénomènes aux contours grammaticaux flous, il est d'autant plus important d'y prêter attention. Du côté de la correction, il sera nécessaire d'offrir plusieurs possibilités de correction, afin de permettre à la personne utilisatrice du système d'intervenir pour choisir sa correction.

Nous finissons ce chapitre par un bref résumé des contributions concrètes apportées à notre recherche dans cette première partie, et qui sont les suivantes : un corpus de productions écrites représentatives des utilisations courantes de l'anglais de personnes francophones de niveau intermédiaire à avancé ; il n'existe à notre connaissance pas d'autre corpus de ce type, un système de catégorisation utilisant comme critère de classification le domaine des erreurs,

et tendant ainsi à minimiser les possibilités multiples de classement, et enfin un schéma d'annotation en format XML enrichi de sept attributs permettant d'inclure plusieurs possibilités de correction.

Le Chapitre 2 est consacré à la modélisation des deux phénomènes syntaxiques liés aux erreurs que nous avons sélectionnées. Nous y abordons également la question de la modélisation des erreurs produites, afin de préparer leur implémentation dans le dernier chapitre.

Chapitre 2
Méthodes linguistiques pour la modélisation des erreurs et de leurs corrections

Introduction

Les erreurs relatives au placement des adverbes dans la proposition et le groupe verbal et à l'utilisation des structures N+N se situent toutes deux à l'intersection de plusieurs problématiques, qui justifient d'ailleurs leur sélection pour notre recherche. Celles-ci ont déjà été mentionnées, mais nous proposons d'en explorer les corrélations pour débuter ce chapitre.

La première de ces problématiques concerne les difficultés que ces structures posent aux personnes utilisatrices de l'anglais, situation que nous avons exposée dans le chapitre précédent. Les deux types d'erreur sélectionnés font partie des six erreurs les plus fréquemment rencontrées dans le corpus. Le présent chapitre s'étend d'ailleurs plus longuement sur les modalités de la production de ces erreurs.

Cependant, ces erreurs portent sur des éléments dont l'utilisation peut être contournée, contrairement aux articles ou aux prépositions par exemple. La fonction caractéristique des adverbes est la modification, et même si nous verrons que ceux-ci peuvent être rencontrés en fonction de complément du verbe, ils sont le plus souvent optionnels. Les structures N+N sont optionnelles d'un autre point de vue, dans le sens où il est souvent possible de leur substituer une autre configuration syntaxique du type [Nom noyau + Prep + GN].

De plus, les phénomènes concernés sont régis par des règles grammaticales complexes, et l'utilisation correcte des adverbes et des noms en fonction de dépendant repose sur de nombreuses variables linguistiques. Si ces erreurs peuvent le plus souvent être corrigées aisément par des personnes spécialistes de l'anglais, qui s'appuient sur un ensemble de compétences et d'intuitions linguistiques, elles représentent une difficulté pour la correction automatisée, du fait du grand nombre de domaines linguistiques à prendre en compte dans leur traitement, et de l'impossibilité relative de fournir des jugements de grammaticalité généralisables.

Ces trois problématiques, loin d'être juxtaposées, dépendent en réalité les unes des autres. Le fait que ces deux types d'erreur concernent des structures et éléments optionnels, et soient donc utilisés moins fréquemment, peut expliquer le fait qu'ils n'aient pas été le sujet d'un grand nombre de travaux en correction grammaticale automatisée, d'autant plus que la complexité de leur utilisation rend cette tâche particulièrement délicate. Le manque de normes grammaticales claires pour la détection et la correction des erreurs liées à ces phénomènes peut également être source de lacunes dans les ressources pédagogiques mises à la disposition du personnel enseignant et des personnes autodidactes, mettant la maîtrise de ces structures

hors de la portée de nombreuses personnes utilisatrices ou apprenantes. En l'état, cette situation peut aisément être considérée comme une impasse.

Nous postulons toutefois qu'une approche prudente de ces phénomènes, dans le respect de leur complexité linguistique, peut amorcer la résolution de ce problème. Le présent chapitre propose ainsi avant de décrire les structures et systèmes linguistiques concernés de manière précise afin de tenter de surmonter l'obstacle de la complexité linguistique.

La première partie du chapitre est consacrée à la description de la méthode et des sources utilisées, qui forment notre cadre théorique. Nous adoptons une approche mixte, combinant les approches descendante et montante : nous nous appuyons d'une part sur les informations données par des grammaires de référence, et d'autre part sur les informations que nous avons pu recueillir sur les erreurs, leur typologie et les configurations syntaxiques concernées. Les grammaires de référence utilisées comme sources principales y sont également décrites, avec une attention particulière à leur portée scientifique, aux approches théoriques mobilisées par leurs auteurs et à leur façon de gérer l'ambiguïté et les normes grammaticales floues.

Les deuxième et troisième parties de ce chapitre sont une mise en application de cette méthode mixte aux deux phénomènes sélectionnés. Nous abordons dans la deuxième partie la question du placement des adverbes, et nous nous concentrons sur la modélisation des erreurs et des corrections liées aux adverbes de manière et à l'adverbe *also*. La troisième partie est consacrée aux structures N+N. Nous y présentons une modélisation des erreurs et de leurs corrections pour deux des cinq sous-types identifiés pour cette erreur, ainsi que des pistes pour le traitement des trois autres types, qui requièrent des règles différentes.

2.1 Présentation de la méthode

L'objectif de cette partie est de présenter clairement la façon dont nous avons choisi de procéder pour la modélisation linguistique devant mener à la détection et la correction automatisée des deux types d'erreur sélectionnés. La première sous-partie détaille cette méthode et donne des indications sur nos choix pratiques et théoriques, tandis que la seconde introduit les grammaires de référence sur lesquelles nous nous appuyons.

2.1.1 Une méthode mixte : synthèses grammaticales et études de corpus

Le placement des adverbes comme l'utilisation des structures N+N sont régis par un ensemble de contraintes grammaticales, sémantiques et lexicales plus ou moins strictes et parfois difficiles à identifier. Ceci complique considérablement le traitement automatisé des erreurs qui leur sont associées. En comparaison, le traitement des erreurs morphosyntaxiques, comme par exemple les erreurs d'accord et de conjugaison de verbes irréguliers, est facilité par l'existence de règles morphosyntaxiques connues et stables, et ne représente plus réellement un défi pour la recherche en correction grammaticale automatisée pour l'anglais. Comme nous le verrons dans le Chapitre 3, ce type d'erreur est maîtrisé par la majorité des correcteurs grammaticaux existants.

Si le domaine de la correction automatisée ne s'est pas intéressé de près aux types d'erreur choisis pour notre recherche, il serait pourtant faux d'affirmer que les erreurs les plus étudiées le sont en raison de leur facilité de traitement. Les erreurs dans le choix des articles et des prépositions, qui sont à l'heure actuelle la cible de la majorité des projets de détection et correction automatisée, sont parmi les erreurs les plus difficiles à aborder. Comme le soulignent Leacock et al., les prépositions sont utilisées dans un grands nombre de contextes syntaxiques, par exemple dans les adjoints ou dans la complémentation verbale et nominale, ainsi que dans des constructions lexicales, comme les idiomes et les verbes à particule (2010 : 20-22).

En ce qui concerne le choix des articles, Leacock et al. estiment que celui-ci pose au moins autant de difficultés aux personnes apprenantes que la sélection des prépositions : "As with prepositions, English language learners are faced with a bewildering array of inter-related lexical, syntactic, and discourse rules to master, and each with its own set of exceptions, as well as world knowledge" (op. cit. : 23). Ce dernier paramètre est particulièrement épineux dans le cadre d'une approche d'automatisation, les connaissances générales étant difficiles à modéliser et à mobiliser de manière pertinente.

Puisque ces erreurs présentent une difficulté de traitement comparable à celle des types que nous avons sélectionnés, il semble logique de nous pencher brièvement sur les règles mises en place pour la détection et la correction automatisées de ces erreurs dans les recherches à leur sujet. Rappelons qu'un exposé plus détaillé des différentes techniques et stratégies utilisées dans ce domaine est présenté dans le Chapitre 3.

Leacock et al. passent en revue les projets consacrés à la détection et à la correction de ces erreurs (op. cit. : 47-57). La détection des erreurs dans le choix des articles et des prépositions ne donne pas lieu exactement aux mêmes méthodes, mais puisque celles-ci sont relativement similaires et que notre étude n'est pas axée sur ces types d'erreurs, nous prenons la liberté de les présenter ensemble. La plupart des projets récents utilisent des méthodes statistiques, reposant sur un apprentissage automatique à partir de corpus annotés de très grande taille. Nous présentons ces méthodes plus longuement dans la sous-partie 3.1.1. Ces données sont complétées par des informations syntaxiques grâce à l'utilisation des étiquettes de partie du discours et d'analyses syntaxiques. Certaines études ont cependant utilisé des heuristiques sous la forme de règles de correction créées manuellement.

Si Leacock et al. insistent sur l'intérêt d'utiliser des méthodes statistiques pour le traitement de ces erreurs, les auteurs reconnaissent qu'un traitement à base de règles créées manuellement est envisageable pour d'autres types (op. cit. : 74-75) :

> In the case of preposition and article errors, machine learning approaches are especially advantageous because of the complexity of these errors and their interactions […]. However, for other error types, especially those where the disambiguating context is very local, […] manually constructed rules can be easier to develop than statistical ones without loss of performance.

Nous verrons cependant dans ce chapitre que les deux types d'erreur que nous avons sélectionnés ne se prêtent malheureusement pas aussi bien l'un que l'autre à l'utilisation de règles de correction, du moins en partie. Toutefois, ce choix s'explique également par la volonté d'enrichir les objectifs de nos travaux : en utilisant des méthodes linguistiques pour la création de règles de détection et de correction pour des erreurs "orphelines", nous espérons apporter un éclairage grammatical sur les phénomènes étudiés et sur les erreurs qui sont produites. Cet objectif secondaire a également un volet pédagogique, puisque ces informations peuvent être utilisées d'une part pour la création de messages correctifs à destination des personnes utilisant le système de correction, et d'autre part pour l'enseignement de ces phénomènes à des personnes apprenantes.

La sélection d'un type d'approche ne constitue que l'amorce du travail de modélisation linguistique. Comme nous l'avons déjà dit, les informations grammaticales disponibles au sujet des deux phénomènes étudiés ne font pas consensus, contrairement à d'autres aspects de la grammaire de l'anglais ; la mise en place de règles automatisées nécessitant une base

linguistique claire, nous passons en revue les informations existantes afin de tenter de construire cette base. Cette approche fait émerger trois questionnements :

- quelles ressources grammaticales doivent être utilisées ?
- quelle quantité d'informations doit-on, et peut-on recueillir dans le cadre restreint de ces travaux ?
- comment procéder lorsque la synthèse des informations grammaticales ne permet tout de même pas d'obtenir des réponses claires ?

Nous tentons de répondre à ces questions dans les paragraphes suivants.

La méthode que nous mettons en place pour aboutir à une modélisation linguistique utilisable dans les patrons de détection et les règles de correction adopte une approche progressive en trois étapes. S'il est vrai que nous souhaitons apporter un éclairage grammatical sur ces phénomènes, les présents travaux ne sont pas pour autant consacrés entièrement à leur étude. Il est donc nécessaire de restreindre le champ de nos analyses, et de laisser à d'autres le soin de les explorer dans une démarche exhaustive, assistés peut-être des synthèses que nous aurons pu fournir. Notre approche est donc restreinte sur deux points : d'une part par le choix des sources utilisées, et d'autre part par l'observation des erreurs et des configurations qui sont sources de difficultés.

Nous prenons comme point de départ et fondations de nos synthèses les conclusions présentées dans trois grammaires de référence de l'anglais parues depuis le milieu des années 1980. Puisqu'elles constituent nos principales ressources, ces grammaires sont présentées en détail dans la section suivante. Nous sommes aussi amenée, lors de cette étape et dans la troisième étape présentée ci-dessous, à faire appel à des sources supplémentaires, comme par exemple la monographie de Pastor-Gómez sur les structures N+N (2011). Cependant, afin de conserver une certaine homogénéité des approches théoriques, nous n'intégrons pas dans notre synthèse les travaux effectués dans le cadre de théories linguistiques précises, comme la grammaire transformationnelle générative ou la psychomécanique du langage. Les ouvrages sont néanmoins cités à titre indicatif lorsque cela paraît opportun.

Nous commençons par aborder le phénomène général auquel les erreurs sont liées, c'est-à-dire la syntaxe externe des adverbes et la syntaxe interne du groupe nominal. Ceci nous amène également à évoquer des problématiques connexes, comme la définition de la catégorie syntaxique de l'adverbe, le concept de la modification, et l'identification des noms composés. À l'issue de cette première étape générale, nous présentons une synthèse des aspects majeurs

des phénomènes étudiés, sous la forme de tableaux présentant les possibilités de placement des adverbes selon leur type sémantique, et l'organisation des modifieurs et compléments dans le GN.

La deuxième étape consiste en l'observation des segments erronés relevés dans le corpus, afin de repérer et modéliser les configurations dans lesquelles ces erreurs se retrouvent. Notre corpus étant de taille limitée, et les segments erronés en nombre réduit, nous utilisons également les résultats d'études sur le transfert ou l'emploi de ces structures par des personnes utilisatrices de l'anglais pour affiner notre "grammaire des erreurs". C'est l'utilisation de cette stratégie qui vaut à notre approche le qualificatif de "mixte". La partie consacrée aux adverbes est plus riche en ce sens en raison de l'existence de plusieurs études consacrées au placement des adverbes dans le domaine de l'acquisition des langues.

La troisième étape de notre approche a pour objectif d'affiner la synthèse effectuée précédemment, en se focalisant sur les pierres d'achoppement des personnes apprenantes dans le placement des adverbes et la construction des structures N+N, et que nous avons identifiées dans la seconde étape. Nous faisons varier pour cela les sources utilisées, tout en tentant de conserver l'homogénéité théorique que nous avons évoquée précédemment. Nous sommes également amenée à restreindre le champ des erreurs traitées pour des raisons de faisabilité, en nous concentrant sur les erreurs dans le placement des adverbes de manière ainsi que sur les erreurs liées à l'empilement et à l'emploi du génitif dans les séquences N+N. C'est dans cette partie que nous tentons de répondre à la troisième des questions que nous avons énoncées plus haut, concernant la façon d'aborder les aires qui restent incertaines, même après la synthèse des informations grammaticales disponibles sur le sujet. Dans le cas du placement des adverbes, nous faisons appel à des stratégies *ad hoc*, comme par exemple le recueil de jugements de grammaticalité auprès de personnes anglophones.

En ce qui concerne l'organisation des parties qui suivent, nous avons choisi de traiter le cas du placement des adverbes et des séquences N+N successivement, appliquant les trois étapes de notre approche à chacun de ces deux types, de manière plus ou moins stricte selon le type d'erreur concerné. Une organisation fonctionnelle par étapes aurait également été possible ; nous avons cependant préféré privilégier la cohérence linguistique de chaque partie.

2.1.2 Les grammaires de référence de l'anglais

La deuxième section de cette première partie présente les trois grammaires de référence utilisées comme base de nos synthèses des phénomènes étudiés. L'objectif de cet exposé

détaillé est d'une part de justifier leur utilisation conjointe, et d'autre part de contextualiser les jugements grammaticaux que nous en extrayons, pour éventuellement les nuancer.

a. Les différents types de grammaire

Le terme de "grammaire", lorsqu'il renvoie à un ouvrage écrit, est utilisé pour faire référence à trois types de documents : les grammaires pédagogiques, les grammaires théoriques, et les grammaires de référence (F. Aarts, 1988 : 173). Les premières sont utilisées pour l'enseignement et l'apprentissage des règles de grammaire à destination d'un public natif ou non-natif. L'objectif des secondes est de présenter une théorie scientifique précise appliquée au langage, dans le but d'expliquer l'organisation et les formes de celui-ci.

Les grammaires de référence ont quant à elles pour ambition de décrire les principes syntaxiques et morphologiques qui régissent une langue, de manière aussi exhaustive que possible, afin de pouvoir servir de référence à toute personne souhaitant obtenir des informations sur un point de grammaire. Les descriptions qu'elles contiennent proviennent de l'observation des utilisations de la langue par ses locutrices et locuteurs, d'un point de vue neutre et non prescriptif, ce qui leur vaut également le titre de grammaires "descriptives". Même si elles reposent généralement sur un socle théorique plus ou moins explicite, les explications qui y sont présentées ne sont pas utilisées comme arguments en faveur d'une théorie linguistique plutôt qu'une autre. Biber et al., dans *LGSWE* (voir ci-dessous), expriment cette distinction ainsi (1999 : 6) :

> Studies with a theoretical orientation focus on discovering abstract underlying principles in relation to a model of linguistic competence, typically analyzing relatively few grammatical constructions in depth. In contrast, descriptive studies (such as ours) attempt to provide a more comprehensive characterization of grammatical phenomena in an individual language like English.

Par souci d'efficacité, et dans le but de donner à nos analyses et synthèses un socle solide et reconnu, nous avons choisi de faire appel à des grammaires de référence de l'anglais comme base de nos recherches sur les phénomènes étudiés. Les trois grammaires utilisées sont les suivantes (par ordre chronologique) :

- *A Comprehensive Grammar of the English Language* (*ACG*), de Quirk et al., parue en 1985,
- *The Longman Grammar of Spoken and Written English* (*LGSWE*), de Biber et al., parue en 1999,

- *The Cambridge Grammar of the English Language* (*CGE*), de Huddleston et Pullum et al., parue en 2002.

Nous présentons ces trois grammaires dans les paragraphes suivants. Nous ne cherchons pas à en faire une critique exhaustive ; l'objectif de cette présentation est de souligner les caractéristiques de chaque ouvrage ainsi que leur orientation, en gardant à l'esprit les raisons de leur utilisation dans notre étude. Les caractéristiques auxquelles nous accordons notre attention concernent principalement leur contexte de rédaction, les aspects originaux de chaque grammaire, l'approche théorique adoptée, les variétés et standards de l'anglais représentés, et enfin la façon dont les cas d'indécision sont traités par leurs auteurs. Notre présentation des grammaires est effectuée à partir de ces thématiques, plutôt qu'en détaillant les caractéristiques de chaque ouvrage successivement. Les sources utilisées pour effectuer cette présentation sont les introductions de ces ouvrages ainsi que des critiques rédigées par des linguistes lors de leur parution. Le premier et le troisième de ces ouvrages ayant les mêmes initiales, nous utilisons les abréviations *ACG* et *CGE* respectivement pour y faire référence dans la suite de ce document ; comme on peut le voir dans les différents articles critiques consultés, il n'existe plus à l'heure actuelle de consensus concernant les abréviations à utiliser pour ces deux ouvrages.

b. Contexte de rédaction et réception critique des trois grammaires

Comme l'indique F. Aarts dans sa critique de l'ouvrage de Quirk et al., celui-ci est la première grammaire de grande ampleur à avoir été écrite par des grammairiens anglophones (1988 : 163). Elle est le point culminant d'une série de grammaires de l'anglais écrites par les mêmes auteurs, et qui inclut *A Grammar of Contemporary English* (1972), considérée précédemment comme ouvrage de référence. *ACG* fut considérée comme la principale ressource de référence pour toute recherche en grammaire anglaise jusqu'au début des années 2000, comme le reconnaissent les auteurs de *LGSWE* : "[*ACG*] is probably the most detailed grammar of present-day English yet written, and its grammatical system has gained a broad currency through its use in other grammars, textbooks, and academic publications" (op. cit. : 7).

D'après Hirst (2001 : 132), la grammaire de Biber et al., *LGSWE*, publiée en 1999, constituerait un complément plutôt qu'un ouvrage concurrent à *ACG* en raison de son approche empirique. Ceci est d'ailleurs souligné par les auteurs même de *LGSWE* : "In many ways, the two grammars complement rather than compete with each other" (op. cit. : viii).

L'approche adoptée par Biber et al. et les informations complémentaires contenues dans leur ouvrage justifient donc pleinement sa prise en compte dans nos travaux.

Dans un article critique publié en 2004, Leech, un des auteurs de *ACG*, n'hésite pas à donner à la grammaire de Huddleston et Pullum et al., publiée en 2002, le titre de "nouvelle *Gray's Anatomy* de la grammaire anglaise" (2004 : 121 ; notre traduction), soulignant ainsi son ampleur et le caractère innovant d'une grande partie des analyses qui y sont présentées. Dans un article similaire, B. Aarts évalue l'impact de cet ouvrage dans la tradition linguistique : "There is no doubt that the publication of this important and outstanding grammar by some of the most distinguished experts in the field is a landmark event in English linguistics" (2004 : 380). Pour sa part, Culicover juge *CGE* comme supérieure à *ACG* : "[*ACG*] feels almost superficial compared with [*CGE*]. The level of detail of [*CGE*] is such that the reader may begin to feel that s/he is being told everything that one could possibly know about the topics that it covers" (2004 : 128). B. Aarts et Leech adoptent une posture plus mesurée et s'accordent pour maintenir que, si *CGE* est une grammaire plus moderne que *ACG* du fait qu'elle bénéficie de deux décennies supplémentaires de recherche en linguistique, ces deux ouvrages ont des qualités complémentaires qui justifient leur étude respective (B. Aarts, 2004 : 379 ; Leech, 2004 : 146).

c. Originalités des trois grammaires

Au-delà de sa place particulière en tant que première grammaire de référence issue de linguistes anglophones, l'originalité de *ACG* réside dans le fait qu'elle dépasse la simple constatation de faits linguistiques pour tendre vers leur explication en des termes aussi simples que possible (F. Aarts, 1988 : 173). F. Aarts loue également la posture scientifique et pédagogique empruntée par les auteurs de cette grammaire (op. cit : 173) :

> [*ACG*] encourages readers to think critically by offering them alternative analyses, by pointing out that there are many cases of indeterminacy and gradience and by showing that the best analysis is always supported by the most convincing arguments.

L'originalité de *LGSWE* dans la tradition grammairienne est double. Tout d'abord, cette grammaire a la particularité d'adopter une approche empirique, se fondant sur des analyses de corpus pour décrire les phénomènes. Le corpus utilisé est le *Longman Spoken and Written English Corpus*, créé pour les éditions Pearson Longman et comprenant 40 millions de mots. Biber et al. divisent le corpus en quatre sections (*registers*) afin de pouvoir représenter les variations entre les différentes utilisations de l'anglais. Une de ces sections concerne la langue

orale (transcriptions de conversations), et trois sections sont consacrées à la langue écrite (fiction, articles de journaux, prose universitaire). Nous donnons plus de détails sur la composition du corpus dans la section consacrée aux variétés de l'anglais prises en compte. L'organisation des sections et le titre choisi pour la grammaire de Biber et al. permettent d'entrevoir sa seconde originalité, qui est d'inclure expressément l'étude de la langue orale dans les analyses. Ceci n'est pas le cas des deux autres grammaires que nous utilisons, ou du moins pas à la même échelle ; les auteurs de *CGE* déclarent par exemple que leurs analyses s'appliquent de la même manière aux utilisations orales de la langue qu'à leurs utilisations écrites, mais les corpus consultés pour sa conception ne contiennent que des documents écrits (op. cit. : 11). Cette caractéristique a cependant peu d'impact sur notre étude, puisque cette dernière ne concerne par nature que l'anglais écrit.

Une des premières caractéristiques à avoir attiré l'attention de Leech et B. Aarts dans la grammaire de Huddleston et Pullum et al. concerne l'étendue de l'équipe de linguistes ayant participé à sa rédaction. Cette équipe comprend quinze personnes, et même si Huddleston et Pullum sont les principaux auteurs de l'ouvrage, B. Aarts recommande l'utilisation du format de référence "Huddleston et Pullum et al.", ainsi que l'utilisation d'entrées bibliographiques différentes pour chaque chapitre de l'ouvrage, comme on le ferait dans le cas d'un ouvrage composé de plusieurs articles d'auteurs différents. C'est le format que nous avons adopté ici ; nous utilisons donc l'abréviation *CGE* ou la référence "Huddleston et Pullum et al." lorsque nous évoquons l'ouvrage en entier (par exemple dans la présente section), et les noms des auteurs dans l'ordre dans lequel le sommaire les présente lors de références à des chapitres précis (par exemple "Pullum et Huddleston" pour le chapitre "Adjectives and adverbs", et "Payne et Huddleston" pour le chapitre "Nouns and noun phrases").

Cependant, l'originalité majeure de *CGE* réside dans le regard neuf que ses auteurs portent sur un ensemble de problématiques : "[I]t contains a great deal that is new, if not daringly provocative, in its reworking of the well-tilled territory of English grammar" (Leech, 2004 : 121). Les innovations de *CGE* portent notamment sur la définition de la catégorie des prépositions, qui a une influence sur celle des adverbes, et le concept de *fused-head*, c'est-à-dire la capacité pour un seul mot d'avoir plus d'une fonction syntaxique dans un groupe nominal.

d. Approches théoriques adoptées dans les grammaires

En tant que grammaires de référence, ces trois ouvrages entendent présenter les phénomènes de la grammaire anglaise de manière avant tout descriptive, et ne sont pas explicitement sous-tendus par une approche théorique précise. Dans les faits, il est impossible de décrire une langue sans parti pris théorique, celui-ci ayant une influence sur la terminologie employée, les représentations utilisées, et les interprétations des phénomènes, comme l'indiquent les auteurs de *CGE* : "The problem with attempting to describe English without having a theory of grammar is that language is too big to be described without bringing things under generalisations, and without a theory there are no generalisations" (op. cit. : 18).

F. Aarts note que la terminologie utilisée dans *ACG* révèle l'influence de la grammaire transformationnelle, de la grammaire des cas et de la théorie des actes de langage. L'influence de la première est particulièrement visible dans la terminologie utilisée ainsi que leur approche de la description et de l'explication (1988 : 167-168, 171). Dans *LGSWE*, Biber et al. abordent les analyses avec une approche fonctionnelle clairement identifiée comme telle (op. cit : 41), mais ce type d'approche ne peut être assimilé à une posture théorique à proprement parler. Le cadre descriptif et terminologique qui y est adopté est largement emprunté à *ACG* (Schmid, 2003 : 1265).

L'ouvrage qui semble le plus ancré dans une approche théorique est *GCE* : "There is no doubt that [*CGE*] is more theory-laden than [*ACG*]. Many of their ideas and analyses show adherence to monostratal phrase structure models of theoretical syntax" (Leech, 2004 : 124). Ceci ne fait pas pour autant de *CGE* une grammaire théorique, mais il s'agit par contre bien d'une "grammaire descriptive à orientation théorique" (*descriptive theory-oriented grammar*, op. cit. : 125). D'après Leech, ce parti pris théorique est particulièrement visible dans les longues argumentations que les auteurs développent en faveur de leurs analyses, et contre les analyses concurrentes identifiées. Huddleston et Pullum et al. se défendent pourtant de sélectionner leurs arguments du point de vue d'un cadre théorique restreint, présentant leurs analyses comme étant mises en avant simplement parce qu'elles sont les plus convaincantes :

> [I]f ever the facts at hand can be presented in a way that is neutral between competing theoretical frameworks, we try to present them that way. However, a significant amount of space is devoted here to arguing carefully that the particular analysis we have decided to adopt, within the framework of theory we assume, is the right **analysis**. What we mean by that is that even someone with a different idea about how to design a theory of syntax

> would have to come to a conclusion tantamount to ours if they considered all the facts. (caractères gras dans l'original ; op. cit. : 19)

Nous verrons dans notre présentation du traitement des cas indécis que cette stratégie amène les auteurs de *CGE* à adopter des positions plus tranchées que celles présentées dans *ACG*.

e. Standards et variétés de l'anglais représentées

Les auteurs de *ACG*, *LGSWE* et *CGE* adoptent des postures très similaires en ce qui concerne l'attention apportée aux différentes variétés de l'anglais, ainsi qu'au concept d'"anglais standard". Quirk et al. déclarent qu'il existe d'une part une seule langue anglaise, et d'autres part un ensemble de variétés et variations (op. cit. : 11), soulignant l'existence d'un noyau commun (*common core*) liant ces différentes variétés, noyau qui est le sujet des analyses données dans cette grammaire (op. cit. : 16). Même si les auteurs consacrent une quinzaine de pages au passage en revue des différents types de variation et variété de l'anglais, les deux seules variétés prises en compte dans les analyses grammaticales sont l'anglais britannique et l'anglais américain standards (op. cit. : 33).

Biber et al. s'intéressent à la variation de l'anglais en fonction de ses différentes utilisations, représentées par les quatre sous-ensembles qui composent leur corpus, plutôt qu'aux variations régionales et sociales. Les auteurs avancent que celles-ci diffèrent peu du standard du point de vue lexico-grammatical (op. cit. : 17), et que les variations les plus marquées se situent au niveau des différentes utilisations de la langue plutôt qu'au niveau des dialectes (op. cit. : 21). Les analyses présentées concernent donc en priorité l'anglais standard, qui est défini de la façon suivante : "the dialectal variety that has been codified in dictionaries, grammars, and usage handbooks". Les auteurs indiquent néanmoins qu'il existe une grande marge de variabilité à l'intérieur même du standard (op. cit. : 18). Comme dans *ACG*, les deux principales variétés prises en compte sont les variétés britannique et américaine.

Huddleston et Pullum et al. insistent également sur l'existence d'un anglais standard relativement homogène, qui est l'objet de leurs descriptions et analyses, et que les auteurs délimitent de la façon suivante :

> [W]e are describing the kind of English that is widely accepted in the countries of the world where English is the language of government, education, broadcasting, news publishing, entertainment, and other public discourse. […] [T]he controversy about particular points stands out against a backdrop of remarkably widespread agreement

> about how sentences should be constructed […]. This widespread agreement defines what we are calling Standard English. (op. cit. : 4)

Les auteurs de *CGE* soulignent également que les variations entre les différentes variétés et dialectes de l'anglais concernent principalement la prononciation et le lexique, et sont très limitées en ce qui concerne la grammaire. Les deux variétés auxquelles une attention plus particulière est accordée sont l'anglais britannique et l'anglais américain, mais les auteurs citent également parmi leurs données un corpus d'anglais australien (voir ci-dessous).

Avant de passer au dernier point de notre passage en revue des caractéristiques de ces trois grammaires de référence, nous souhaitons profiter de cette section sur le traitement des différentes variétés de l'anglais pour donner plus de détails sur les données utilisées par les auteurs. Ces éléments sont présentés dans le tableau suivant :

Grammaires	Type de données	Type de document	Variété
ACG	Corpus *Survey of English Usage*	Textes écrits et transcriptions de conversations, sources variées	Brit.
	Corpus *Brown*	Textes écrits, sources variées	Amér.
	Corpus *Lancaster-Oslo/Bergen*	Textes écrits, sources variées	Brit.
	Données expérimentales sur des jugements de fréquence par des anglophones	n/d	n/d
LGSWE	*Longman Spoken and Written English Corpus*	Transcriptions de conversations	Brit.
		Fiction	Brit. et amér.
		Articles de presse	Brit.
		Prose universitaire	Brit. et amér.
CGE	Corpus *Brown*	Textes écrits, sources variées	Amér.
	Corpus *Lancaster-Oslo/Bergen*	Textes écrits, sources variées	Brit.
	Australian Corpus of English	Textes écrits, sources variées	Austr.
	Corpus *Wall Street Journal*	Articles de presse	Amér.
	Consultations informelles d'anglophones, dictionnaires, autres grammaires	n/d	n/d

Tableau 13. Données linguistiques utilisées dans les grammaires de référence

f. Traitement des cas d'indécision quant aux normes d'usage ou de grammaticalité

Le dernier point que nous abordons ici concerne le traitement de la grammaticalité et l'indécision par les auteurs de ces trois ouvrages. Ce point est particulièrement important dans

notre recherche puisque les cas que nous ciblons sont des exemples de la difficulté d'identifier les facteurs qui amènent une structure à être jugée comme grammaticale et acceptable, ou agrammaticale et inacceptable. F. Aarts (1988 : 170) et Hirst (2004 : 136) attirent l'attention sur les attentes contradictoires qui reposent sur les grammaires de référence : ces auteurs notent que si les auteurs de ces grammaires revendiquent la neutralité de leurs descriptions, une partie des personnes consultent ces ouvrages à la recherche de consignes prescriptives.

Quirk et al. se font le relais prudent de certaines opinions prescriptives à titre indicatif, tout en évitant de porter des jugements de grammaticalité ou d'agrammaticalité, préférant à cette notion celle d'acceptabilité (F. Aarts, 1988 : 170 ; Quirk et al., 1985 : 33). Ils insistent en effet sur le caractère par nature indéterminé du système grammatical :

> Grammar is to some extent an indeterminate system. Categories and structures, for example, often do not have neat boundaries. Grammarians are tempted to overlook such uncertainties, or to pretend that they do not exist. Our guiding principle in this grammar, however, will be to acknowledge them, and where appropriate to explore them through the study of GRADIENCE. (majuscules dans l'original ; op. cit. : 90)

L'analyse en gradients constitue donc un des moyens de représenter l'ambiguïté dans *ACG*. Les auteurs en donnent la définition suivante :

> A gradient is a scale which relates two categories of description […] in terms of degrees of similarity and contrast. At the ends of the scale are items which belong clearly to one category or to another; intermediate positions on the scale are taken by 'in-between' cases - items which fail, in different degrees, to satisfy the criteria for one or the other category. (op. cit. : 90)

Quirk et al. ont également recours à des "analyses multiples" (*multiple analysis*, op. cit. : 90), donnant lieu à plusieurs analyses en constituants différentes. Dans ce cadre, plusieurs analyses d'une même structure peuvent coexister, chacune d'entre elles mettant en avant un type de généralisation possible.

Les auteurs de *CGE* utilisent quant à eux les notions d'usages prototypique et périphérique pour évoquer les cas d'indécision, par exemple dans la délimitation des catégories syntaxiques (Leech, 2004 : 125). Même si le recours à ces notions indique une certaine reconnaissance de la nécessité d'accepter les analyses n'aboutissant pas à des résultats tranchés, Leech juge que les auteurs de *CGE* sont trop soucieux du statut des segments étudiés par rapport aux normes de grammaticalité et d'acceptabilité, dont il met en doute la pertinence :

> [T]hey show too strong a tendency to judge acceptability in clear-cut ways which suit the analytic point being illustrated. The desire to seek a decisive answer to all research questions is too strong, in particular, when examples of borderline acceptability are judged to be either fully grammatical or fully ungrammatical. […] [Huddleston and Pullum et al.] are somewhat too ready to brand an example as ungrammatical or accept it as grammatical, without allowing for the in-between cases. (op. cit. : 127-128)

La notion de gradience ne fait pas partie de l'appareil théorique de *CGE*, ce qui est également représentatif d'un parti pris en faveur de catégorisations strictes (B. Aarts, 2004 : 368). Cette caractéristique est une des distinctions importantes de *CGE* et *ACG*.

La question du traitement de l'indécision est moins pertinente en ce qui concerne *LGSWE*. En effet, cette grammaire étant fondée sur l'utilisation d'un corpus, l'accent est mis sur la description de l'utilisation des structures de la grammaire dans différents "registres" ou sous-corpus (op. cit. : 4). La notion de "fréquence" remplace celle de "grammaticalité" ou d'"acceptabilité", comme l'illustre le passage suivant tiré de l'introduction :

> Teachers, students, materials writers, and those with a purely academic interest will all find it useful to know which grammatical patterns are common and which are rare. Hitherto this information has been based on native-speaker intuition. (op. cit. : 8).

Comme nous l'avons déjà mentionné, ce choix d'approche constitue la principale originalité de *LGSWE*, et en fait un complément judicieux aux autres grammaires : "This book complements previous grammatical descriptions by investigating the linguistic patterns actually used by speakers and writers in the late twentieth century. […] Our focus on use constitutes an entire extra dimension for grammatical description" (op. cit. : 8).

Nous avons consacré la première partie de ce chapitre à l'exposé de la méthode mixte utilisée pour les synthèses grammaticales, puis à la présentation des trois ouvrages de référence utilisés comme points de départ de ces synthèses. La méthode utilisée est qualifiée de "mixte" car elle repose sur des bases théoriques et empiriques, et procède par étapes successives dans le but de parvenir à une modélisation fine et solidement étayée facilitant l'implémentation quasi-directe des règles de correction. Dans un premier temps, nous faisons appel aux conclusions des auteurs des grammaires de référence concernant le placement des adverbes et les structures N+N. Dans un second temps, nous recourons à une modélisation des erreurs relevées ainsi qu'à d'autres informations concernant la maîtrise de ces structures par des francophones, dans le but d'aboutir à une "grammaire des erreurs" permettant de restreindre le nombre de patrons de détection à utiliser.

Les ouvrages de référence sélectionnés sont tous les trois considérés comme essentiels pour toute étude de la grammaire de l'anglais, et offrent des caractéristiques complémentaires. Quirk et al., référence "historique", présentent des analyses très mesurées et laissant une grande place à l'indétermination, dans un cadre théorique jugé traditionnel bien qu'inspiré par la syntaxe générative. Dans ce domaine, les analyses de Huddleston et Pullum et al. offrent des jugements plus tranchés et une présentation innovante de certains points qui ont une grande importance pour notre étude, comme la délimitation de la catégorie des adverbes et la distinction entre noms composés et groupes nominaux complexes. Les informations grammaticales présentées dans ces deux ouvrages sont complétées par les données de corpus dont Biber et al. présentent l'analyse dans *LGSWE*.

Le seul aspect de ces grammaires qui peut être considéré comme un angle mort concerne la prise en compte d'un ensemble représentatif des différentes variétés de l'anglais, les auteurs de ces ouvrages se limitant en général aux variétés britannique et américaine. Ceux-ci déclarent néanmoins constater l'existence d'une grande similitude concernant les phénomènes grammaticaux dans les différentes variétés.

Nous terminons cette partie sur la question de la terminologie et des cadres syntaxiques utilisés lors de nos descriptions. L'emploi de trois grammaires de référence comme sources principales, s'il permet une couverture large des problématiques envisagées, n'est pas sans poser quelques difficultés dans la gestion des terminologies et des descriptions syntaxiques différentes. Nous tâchons de limiter les longues descriptions aux deux problématiques choisies, mais celles-ci entretiennent nécessairement des liens avec des structures et phénomènes connexes. Tout du moins, elles nous amènent à aborder le phénomène de la modification, qui ne peut être décrit sans faire référence à celui de la complémentation, et plus globalement des relations que les noyaux entretiennent avec leurs dépendants. Le cadre que nous adoptons en général est celui qui est décrit dans *CGE*. Nous effectuons les clarifications nécessaires concernant ce cadre au fil des descriptions.

Afin de rendre notre propos le plus clair et précis possible, nous tâchons dans les parties suivantes de donner la définition des termes choisis par les auteurs des trois ouvrages. Les définitions que nous utilisons pour les termes "adverbe" et "séquence/structure N+N" ont déjà été données dans l'Introduction générale de ce document ; elles sont rappelées et contextualisées dans ce chapitre. Nous choisissons de donner certaines définitions en introduction de section, surtout lorsqu'il existe une ambiguïté due aux utilisations qui en sont faites dans les trois ouvrages. Comme nous avons pu le voir dans la description des approches

théoriques des trois grammaires, les auteurs de *LGSWE* ont repris en grande partie les termes utilisés dans *ACG*, et les définitions qui leur sont associées. Les distinctions apparaissent donc principalement lorsque nous confrontons les approches de *ACG* et *LGSWE* à celles de *CGE*. Les définitions qui ne seraient pas données en introduction de section sont données soit lors de la première utilisation du terme, soit ultérieurement, avec une indication de renvoi, si celle-ci n'est pas primordiale pour l'explication en cours et risque d'en perturber la fluidité. Lorsque cela s'avère nécessaire, nous indiquons quelle définition est associée au terme en question pour la suite du document, et nous ne manquons pas de faire des rappels réguliers concernant l'origine de la définition choisie.

2.2 Le placement des adverbes dans la proposition

"Si je ne sais pas ce qu'est ce mot, alors ce doit être un adverbe" : voilà une pensée qui a dû traverser l'esprit de nombre de linguistes en herbe devant une tâche d'étiquetage. Si en plus de cela, on ne sait où placer cet élément dans la phrase, alors l'intuition première est sans doute la bonne.

Contours flous, mobilité et optionalité sont trois des principales caractéristiques de la catégorie des adverbes. Ces caractéristiques se combinent pour la condamner à recueillir peu d'attention en enseignement des langues secondes comme en traitement automatisé des langues. Fort heureusement, la catégorie des adverbes n'a pas été oubliée par les linguistes, qui s'échinent à les classer, les définir, et à percer les secrets de leur placement et de leurs significations depuis des décennies.

Notre objectif est ici d'utiliser les informations grammaticales à notre disposition, ainsi que nos propres analyses de corpus et les résultats de recherche concernant l'acquisition du placement des adverbes, afin d'aboutir à une modélisation des erreurs relatives au placement des adverbes dans la proposition, et des règles de correction que nous sommes en mesure d'apporter. La première sous-partie est consacrée à la syntaxe générale des adverbes, et plus particulièrement à leur syntaxe externe. La seconde se concentre sur le traitement de l'adverbe *also* et des adverbes de manière, ces deux types étant les plus fréquents dans les segments erronés relevés dans notre corpus.

2.2.1 Syntaxe générale du placement des adverbes

Les erreurs que nous avons relevées en relation avec l'utilisation des adverbes concernent surtout le placement de ceux-ci lorsqu'ils fonctionnent comme adjoints dans la proposition (voir définitions du terme "adjoint" ci-dessous) ; pour cette raison, nous nous intéressons principalement à la syntaxe externe du groupe adverbial (GAdv), c'est-à-dire à son comportement syntaxique en relation avec les autres éléments de la proposition. Cette question est abordée dans les sections 2.2.1d et 2.2.1e. Cependant, il ne nous paraît pas souhaitable de faire l'impasse sur certains aspects de l'étude des adverbes qui sont abordés dans toute présentation de la catégorie, et ont une influence sur notre traitement du phénomène. Les sections suivantes sont donc consacrées non seulement à la syntaxe externe du groupe adverbial, mais également à la délimitation de la catégorie des adverbes, à leur morphologie ainsi qu'à la construction des groupes adverbiaux. Nous incluons également en première partie une courte section d'introduction aux termes employés, les terminologies utilisées dans les trois grammaires étant loin d'être superposables.

a. Description de la syntaxe de l'adverbe : cadre syntaxique et questions de terminologie

Plusieurs termes faisant référence à des fonctions syntaxiques de l'adverbe sont utilisés dans les trois ouvrages avec des définitions différentes, mais il existe également des recoupements. Pour cette raison, nous avons choisi de présenter les définitions des termes dans un tableau synthétique.

Les approches des auteurs de *ACG* et *LGSWE* sont regroupées en raison de leur similarité, mais les différences sont indiquées lorsque cela est pertinent. Les définitions données sont des paraphrases succinctes plutôt que des citations, les auteurs ne fournissant pas systématiquement une définition précise en un seul endroit du texte. Certains termes ne sont utilisés que dans l'un des ouvrages (ex. : "adverbial"), la case correspondante est donc grisée dans la colonne de l'autre ouvrage. Les informations présentées dans le tableau 14 sont volontairement réduites au strict minimum des ressources nécessaires pour la description du fonctionnement syntaxique de l'adverbe ; nous donnons des explications plus détaillées dans les paragraphes suivant le tableau et dans les prochaines sections (ex. : définition du terme "dépendant"). La section 2.2.1b étant consacrée entièrement à la délimitation et à la définition de la catégorie des adverbes, nous ne nous étendons pas sur ce point ici.

	ACG* et *LGSWE	***CGE***
Modifieur	Élément **optionnel** (vs. complément) Terme limité aux éléments optionnels du **GP, GN, GAdj et GAdv**	Dépendant **optionnel** (vs. complément) Terme utilisé pour les dépendants optionnels à **tous les niveaux syntaxiques** (du GN à la proposition)
Adverbial	Élément généralement **optionnel** et parfois **obligatoire**, mais distinct des compléments Terme utilisé pour les éléments optionnels ou obligatoires dans la **proposition** (en dehors des compléments)	*Pas d'utilisation de ce terme dans* CGE
Adjoint	***ACG* :** Type d'**adverbial** le plus central (vs. subjoint, disjoint et conjoint)	Dépendant **optionnel dans le GV et la proposition** Terme regroupant les **modifieurs** (intégrés à la proposition) et les **suppléments** (détachés de la proposition syntaxiquement, et par des moyens intonatifs et typographiques)
	Pas d'utilisation de ce terme dans LGSWE	
Supplément	*Pas d'utilisation de ce terme dans* LGSWE *et* ACG	Élément **optionnel** Terme utilisé pour les éléments présents linéairement dans la proposition mais qui n'y sont **pas intégrés d'un point de vue syntaxique**

Tableau 14. Comparaison des définitions liées à la syntaxe de l'adverbe

Les choix terminologiques des auteurs sont liés à des visions différentes de la syntaxe en général, comme on peut le remarquer dans le choix de Quirk et al. de regrouper sous l'appellation "adverbial" des éléments optionnels et des éléments obligatoires. Il est en conséquence très difficile de proposer des équivalents pour chacun des termes.

Nous pouvons cependant compléter cette présentation par l'illustration des différentes définitions présentées dans le tableau 14. Dans le groupe adjectival suivant, l'adverbe *stunningly* est considéré comme fonctionnant comme un modifieur dans les trois grammaires (l'adverbe est souligné) :

[35] *stunningly beautiful*

Par contre, dans la phrase suivante, seuls les auteurs de *CGE* considère l'adverbe *slowly* comme un modifieur du verbe ; il y est considéré comme un adverbial dans *ACG* et *LGSWE*, puisque le terme de modifieur ne s'applique pas aux éléments optionnels dans la proposition :

[36] *His condition was slowly improving.*

Dans la phrase suivante, les auteurs de *CGE* identifient les deux élements soulignés comme des adjoints, catégorie qui regroupe les modifieurs et les suppléments ; le premier élément est un supplément, alors que le second est un modifieur :

[37] *Fortunately, his condition was slowly improving.*

Le premier élément est considéré comme un supplément et non un modifieur parce qu'il n'est pas intégré à la structure syntaxique de la proposition : dans une représentation sous forme d'arbre syntaxique, l'adverbe *fortunately* serait présenté seul dans une branche séparée. Nous avons vu que *slowly* est interprété comme un adverbial dans le cadre de *ACG* et *LGSWE*, tout comme *fortunately* ; pour les auteurs *d'ACG*, l'adjoint est une sous-classe des adverbiaux, et *slowly* est identifié comme un adjoint. *Fortunately* correspond à un disjoint, élément relativement extérieur à la proposition et exprimant une évaluation du propos qui y est tenu.

L'expression "groupe adverbial", que nous avons déjà utilisée, dénote les structures ayant un adverbe pour noyau (*adverb phrase*), et concerne donc la catégorie, ou "nature", de ces structures. Cependant, comme on peut le voir ici, le terme nominalisé d'"adverbial" (*adverbial*) est utilisé par les auteurs de *ACG* et *LGSWE* pour faire référence à une fonction syntaxique qui peut être instanciée par différentes catégories de groupes, comme les GP et les GAdv. Dans la section 2.2.1d, nous détaillons l'approche des auteurs de *ACG* et *LGSWE* concernant le rôle syntaxique des adverbes dans la proposition, et sommes donc amenée à utiliser le terme d'"adverbial". Cependant, il nous semble que ce terme est source de confusion, et nous adoptons donc dans les autres parties la terminologie de *CGE*, qui fait appel au terme d'"adjoint".

Le cadre syntaxique que nous utilisons est généralement celui qui est décrit par les auteurs de *CGE*, ces descriptions étant les plus modernes et, à nos yeux, les plus précises et les mieux étayées. Puisque nous traitons des adverbes en fonction d'adjoint dans la proposition, il est impossible de ne pas faire référence à l'organisation générale des dépendants du verbe et des éléments de la proposition. Le tableau ci-dessus contient déjà la présentation de certains concepts, comme ceux de modifieur et supplément.

Dans le cadre descriptif de *CGE*, le verbe est considéré comme le noyau de la proposition, et admet des dépendants. Le terme de "dépendants" (*dependents*) est utilisé pour faire référence à tous les éléments en relation avec un noyau, cette fonction regroupant les compléments aussi bien que les modifieurs, dans le GV comme dans les autres types de groupes. Les compléments sont plus étroitement liés au verbe que les modifieurs, et leur présence, leur réalisation syntaxique et leur rôle sémantique dépendent du type du verbe noyau. Le sujet est considéré comme un complément externe. Les compléments internes incluent les compléments d'objet, les compléments prédicatifs (*predicative complements*, ou attributs du sujet), et certains types moins prototypiques tels les compléments locatifs (*locative complements*, voir exemple [38]) et les compléments de manière (voir exemple [39],

ainsi que ci-dessous, section 2.2.1d) (Huddleston, 2002 : 52-54). Comme nous l'avons vu, les suppléments sont des éléments qui sont liés sémantiquement à la proposition mais qui ne sont pas des dépendants du verbe et ne sont pas intégrés à sa structure syntaxique (voir exemple [40]) (op. cit. : 66).

[38] *The cat is on the chair.*

[39] *You'll have to word your reply very carefully.*

[40] *Actually, I'm leaving tomorrow.*

L'approche la plus répandue du rôle des auxiliaires les présente comme des éléments attachés au verbe lexical, aussi nommé "verbe principal" (*main verb*) si l'on adopte cette approche de l'auxiliaire. *CGE* propose une approche différente, considérant ces verbes comme des verbes caténatifs, c'est-à-dire prenant pour complément une proposition non-finie, si bien que les deux exemples ci-dessous ont la même structure syntaxique (op. cit. : 104) :

[41] *She is writing a novel.*

[42] *She began writing a novel.*

Pour des raisons pratiques, nous privilégions dans nos descriptions l'analyse traditionnelle de l'auxiliaire comme dépendant du verbe lexical ; notre étude n'appelant pas une analyse poussée du groupe verbal en lui-même, il ne nous paraît pas utile d'introduire cette nouvelle analyse. Huddleston reconnaît par ailleurs l'intérêt pratique de l'identification d'un "groupe de verbes" (*verb group*), à ne pas confondre avec le groupe verbal (*verb phrase*, GV) (op. cit. : 1213, note 26) :

> The term 'verb group' is an ad hoc phrase […]. The ad hoc term reflects our view that the category is not theoretically justified but may have some practical descriptive value. We prefer, therefore, to reserve the term 'verb phrase' for the unit which includes the complements and modifiers of the verb, in accordance with widespread usage in modern grammars.

b. Délimitation de la catégorie des adverbes

Traditionnellement, la catégorie des adverbes a été considérée comme une catégorie "résiduelle" (*residual category*, Pullum et Huddleston, 2002 : 563), constituée de tous les mots qui ne correspondaient pas aux critères d'appartenance à d'autres catégories aux contours plus nets, tels les noms ou les verbes. Cette catégorie a été donc été définie principalement en creux, comme l'indique le point de départ de la définition qu'en font Pullum et Huddleston ;

ces derniers définissent en effet les adverbes comme les termes fonctionnant comme modifieurs des noyaux autres que les noms (op. cit. : 562).

Les auteurs des trois grammaires que nous utilisons insistent également sur le caractère hétérogène, voire "nébuleux" (Quirk et al., 1985 : 438) de cette catégorie. Comme nous le verrons dans la section 2.2.1c, cette hétérogénéité est marquée aux niveaux morphologique, sémantique, et syntaxique. D'après Pullum et Huddleston, c'est l'hétérogénéité des comportements syntaxiques des adverbes qui est la plus caractéristique (op. cit. : 563), la liste des types de noyaux qu'ils peuvent modifier comprenant non seulement les verbes, les adjectifs et les autres adverbes, mais aussi les propositions, les déterminants, les GP et les GN (op. cit. : 562). Les auteurs notent cependant que si tous les adverbes n'ont pas une distribution syntaxique aussi étendue, tous les éléments qui acceptent un adverbe en fonction de modifieur acceptent des adverbes prototypiques d'un point de vue morphologique, c'est-à-dire formés par l'ajout du suffixe *-ly* à un adjectif. Ainsi, même les fonctions les plus marginales des adverbes, tel que la modification du GN, peuvent être instanciées par au moins un adverbe prototypique, comme dans l'exemple suivant, emprunté à Pullum et Huddleston (563) (adverbe souligné pour plus de clarté) :

[43] *Absolutely the best way of handling the situation.*

Ceci constitue un critère d'unification syntaxique de la catégorie. On peut en voir un autre dans le fait qu'absolument aucun adverbe n'est observé en fonction de modifieur du nom dans un GN (op. cit. 563). Nous revenons sur la syntaxe externe de l'adverbe dans la section 2.2.1d.

Comme nous l'avons déjà évoqué dans notre présentation des ouvrages de référence, un des aspects novateurs de *CGE* réside dans la proposition qui y est faite de remettre sur le métier la délimitation de certaines catégories syntaxiques, en particulier celle des prépositions. L'existence d'une frontière floue entre la catégorie des adverbes et celle des prépositions est signalée dans *ACG* et *LGSWE*. Quirk et al. utilisent par exemple la notion d'"adverbe prépositionnel" (*prepositional adverb*), qu'ils définissent ainsi (op. cit. : 713) :

> A prepositional adverb is a particle which is formally identical to or related to a preposition, and which often behaves like a preposition with ellipted complement. […] Thus a prepositional adverb shares the form, but not the syntactic status, of a preposition. It is capable of standing alone as an adjunct, conjunct, postmodifier, etc without the addition of a prepositional complement.

Dans ces exemples, empruntés aux auteurs (op. cit. : 713), *past* est considéré comme une préposition dans la première phrase et un adverbe prépositionnel dans la seconde :

[44] *A car drove past the door.*

[45] *A car drove past.*

Les auteurs de *LGSWE* évoquent quant à eux l'existence de particules adverbiales (*adverbial particles*, op. cit. : 78) : "While prepositions have a special relationship to nouns, adverbial particles are closely linked to verbs : most typically, prepositions precede noun phrases, and adverbial particles are added to verbs". Ils notent également que si les particules adverbiales diffèrent des adverbes de nombreuses façons, il peut être plus difficile de les distinguer des prépositions (op. cit. : 78).

Quirk et al. relativisent l'importance des étiquettes de catégorie, utilisant l'argument selon lequel le questionnement autour de la composition d'une catégorie n'est pas assimilable à l'explication de la syntaxe de ses éléments : "Such arguments, however, have more to do with the labelling of categories than with the question of how we can best explain the grammatical behaviour of items on the basis of their various degrees of similarity and contrast" (op. cit. : 73). Pullum et Huddleston ne semblent pas partager entièrement cet avis, puisqu'ils proposent de rendre la catégorie des adverbes plus cohérente, et ainsi de simplifier les descriptions syntaxiques, en réassignant à d'autres catégories un certain nombre d'éléments qui y figurent traditionnellement, comme les mots *outside* ou *yesterday* (op. cit. : 564).

Cette refonte de la catégorie des adverbes est en partie une des conséquences de l'extension de la catégorie des prépositions proposée par Huddleston et Pullum et al. Dans *CGE*, contrairement à d'autres approches de la fonction des prépositions, celles-ci sont considérées comme le noyau du groupe syntaxique correspondant, ainsi nommé "groupe prépositionnel" (GP) (op. cit. : 598). Les auteurs abandonnent également la restriction traditionnelle de l'étiquette de "préposition" aux termes prenant pour complément des GN, pour l'étendre non seulement aux termes prenant d'autres types de compléments, mais également aux termes fonctionnant seuls (op. cit. : 599).

Pullum et Huddleston avancent l'argument de la simplicité et de la cohérence avec les autres groupes syntaxiques. Ainsi, si dans les GN ou les GV, les noyaux peuvent accepter des compléments comme ne pas en accepter, il semble logique que ce soit également le cas des prépositions. Dans les exemples donnés ci-dessus, *past the door* ([44]) est interprété comme un GP dans lequel la préposition noyau est complétée par un GN, et *past* ([45]) est interprété comme un GP constitué uniquement d'une préposition noyau.

Cette extension fait donc disparaître de la catégorie des adverbes l'immense majorité des termes dénotant les localisations spatiales, commes par exemple les termes *outside*, *away*, et bien sûr *past*, si bien que cette catégorie sémantique n'apparaît pas dans la liste des significations fréquemment portées par les adverbes (op. cit. : 576). L'exclusion de ces termes constitue la première modification apportée à la catégorie des adverbes.

La deuxième modification est d'ampleur plus réduite puisqu'elle concerne l'exclusion totale de termes qui étaient auparavant assignés soit à la catégorie des adverbes, soit à une autre catégorie suivant leur fonction. Plus précisément, ces modifications concernent tout d'abord un petit nombre de pronoms dénotant la localisation temporelle et qui sont réassignés à cette catégorie uniquement (*yesterday*, *today*, *tomorrow*, *tonight*, *now*, *then*) ; ces derniers sont d'ailleurs identifiés par Jespersen comme étant des "adverbes pronominaux" (2006 [1933] : 39). Le second groupe concerné est celui des déterminants fonctionnant généralement comme dépendants de noms indénombrables singuliers, mais pouvant également fonctionner comme modifieurs du verbe ou de l'adjectif (*the*, *this*, *that*, *all*, *any*, *a little*, *much*, *little*, *enough*). Les phrases suivantes, citées par Pullum et Huddleston, illustrent les utilisations de ces deux types de termes dans les configurations syntaxiques qui ont pu leur valoir d'être inclus dans la catégorie des adverbes (op. cit. : 564-565):

[46] *They arrived <u>yesterday</u>.*

[47] *She is <u>this</u> tall.*

Nous reconnaissons dans ce travail de délimitation rigoureuse des catégories, et dans les arguments donnés, le penchant des auteurs de *CGE* pour des jugements tranchés, et qui a été souligné par Leech dans sa critique de l'ouvrage (2004 : 127). Leech y soulève certains problèmes liés à cette refonte de la catégorie des prépositions, observant que si les contours de la catégorie des adverbes sont désormais bien visibles, ceux de la catégorie des prépositions ont perdu beaucoup de leur netteté (op. cit. : 135). Nous ne souhaitons pas nous prononcer sur l'ensemble des modifications proposées par Pullum et Huddleston concernant la catégorie des prépositions, qui touchent également les chevauchements avec la catégorie des conjonctions. Les arguments présentés en faveur de coupes dans la catégorie des adverbes nous paraissent cependant convaincants, et c'est cette délimitation que nous adoptons dans notre étude. Nous verrons dans les sections suivantes que cette nouvelle délimitation influence l'interprétation de la syntaxe interne et externe des GAdv.

c. Formation des adverbes et construction du groupe adverbial

Dans cette section, nous abordons brièvement les aspects morphologiques liés à la catégorie de l'adverbe, ainsi que la syntaxe interne des GAdv. Nous avons regroupé ces deux problématiques car elles sont toutes deux d'importance limitée en comparaison à la syntaxe externe des GAdv dans la question de leur placement.

Aspects morphologiques de la catégorie des adverbes

En plus de l'hétérogénéité des fonctions que ceux-ci peuvent assurer, les adverbes se distinguent des noms et des verbes par le fait qu'une majorité d'entre eux est complexe d'un point de vue morphologique (Pullum et Huddleston, 2002 ; 565). Les auteurs de *ACG* et *LGSWE* s'accordent sur une division des adverbes en trois classes principales selon leur type morphologique (Quirk et al, 1985 : 438 ; Biber et al., 1999 : 539) :

- les adverbes de forme simple (ex. : *well*, *too*, *quite*, *only*, *no*, etc.),
- les adverbes composés (ex. : *therefore*, *nevertheless*, *somehow*, *anyway*, etc.),
- les adverbes dérivés par suffixation (ex. : *clearly*, *finally*, *really*, *clockwise*, etc.).

Biber et al. mentionnent brièvement une quatrième catégorie comprenant les expressions figées, par nature non-compositionnelles, comme *of course* et *kind of*. Notons que cette catégorisation repose sur la délimitation "traditionnelle" de la catégorie des adverbes, les auteurs de *ACG* et *LGSWE* donnant comme exemple des termes qui sont réassignés à la catégorie des prépositions dans *CGE*.

Les auteurs de *CGE* synthétisent les différentes caractéristiques morphologiques des adverbes en les regroupant en quatre classes (Pullum et Huddleston, 2002 : 565-570) :

- les adverbes dérivés d'adjectifs par suffixation en *-ly* (ex. : *finally*, *individually*, *totally*, etc.),
- les autres adverbes composés (ex. : *already*, *sometimes*, *moreover*, *nowadays*, etc.),
- les adverbes ayant la même forme qu'un adjectif, avec un sens proche ou bien éloigné (ex. : *fast*, *hard*, *daily* / *just*, *even*, *still*, etc.),
- les autres adverbes de forme simple (ex. : *soon*, *though*, *often*, *rather*, etc.).

Pullum et Huddleston font une présentation exhaustive de la distribution morphologique des adverbes, mentionnant notamment la liste des contraintes reposant sur la dérivation d'adjectifs en adverbes par le biais de suffixes. Ils indiquent également qu'une large proportion

des adverbes formés par dérivation d'un adjectif avec le suffixe *-ly* héritent d'un sens que l'on peut paraphraser en "d'une manière ADJ" (ex. : *clearly*, "in a clear manner") (op. cit. : 565).

Biber et al. analysent la distribution des types morphologiques d'adverbes (selon leur délimitation) dans les quatre "registres" ou sous-corpus qui sont utilisés comme base des descriptions présentées dans leur ouvrage. Les types les plus couramment utilisés sont les adverbes de forme simple et les adverbes avec suffixation en *-ly*. Les adverbes de forme simple se retrouvent le plus fréquemment en conversation, alors que le registre de la prose universitaire présente le plus d'adverbes avec suffixation en *-ly*, totalisant près de 55 % des adverbes utilisés.

Syntaxe interne du groupe adverbial

Le seul des trois ouvrages à aborder la construction des groupes adverbiaux en elle-même est *CGE*. Celle-ci est abordée indirectement et de manière partielle dans les deux autres grammaires, lorsque les auteurs évoquent les fonctions syntaxiques des adverbes, qui, pouvant modifier d'autres adverbes, rentrent dans la composition du groupe adverbial.

Si les GAdv, comme les autres groupes syntaxiques, peuvent contenir des modifieurs et des compléments, la présence de ces dépendants (voir définition donnée en 2.2.1a) est beaucoup moins fréquente pour ce groupe syntaxique (Pullum et Huddleston, 2002 : 570). Les adverbes acceptant des compléments se limitent à un sous-ensemble des adverbes en *-ly*. Ces compléments prennent le plus souvent la forme de GP, la préposition sélectionnée étant soit celle qui est associée à l'adjectif dont l'adverbe est dérivé, soit la préposition *for*. Voici un exemple de chacun de ces cas, empruntés à *CGE* (le GAdv est présenté entre crochets, l'adverbe noyau est souligné et la préposition apparaît en gras) (op. cit. : 571) :

[48] *Purchase of State vehicles is handled [<u>similarly</u> **to** all State purchases]*.

[49] *[<u>Fortunately</u> **for** me], my mother was unusually liberal-minded.*

Les types de modification associés au GAdv sont similaires à ceux que l'on retrouve dans les GAdj. Les adverbes sont le plus souvent modifiés par d'autres adverbes, indiquant le degré, comme *very*, ou permettant la focalisation (*focusing modifiers*), comme *only*. Si l'on trouve plusieurs adverbes en fonction de modifieur dans le GAdv, il peut s'agir soit de modification empilée (*stacked modification*), soit de modification enchâssée (*submodification*), selon la structure hiérarchique du groupe concerné. Les deux phrases suivantes, citées par Pullum et Huddleston, sont des exemples de modification empilée, puis

enchâssée (la hiérarchie des modifieurs est représentée linéairement à l'aide de crochets, et l'adverbe noyau est souligné) (op. cit. : 572) :

[50] *She loses her temper [only [very rarely]].*

[51] *They had sung [[quite remarkably] well].*

Dans la première phrase, l'adverbe *only* s'applique au groupe *very rarely*, tandis que dans la seconde, *quite* modifie *remarkably*, et forme avec ce dernier un groupe qui modifie *well*.

Les groupes adverbiaux acceptent également comme modifieurs des GN, des GP, et des termes analysés par Pullum et Huddleston comme des déterminants (ex. : *a little*, *enough*, etc.). Ces trois possibilités sont illustrées dans les exemples suivants empruntés aux auteurs (op. cit. : 573-574) :

[52] *We arrived [three hours late].*

[53] *[Later in the day], the situation had improved slightly.*

[54] *He had answered [a little indiscreetly].*

Les règles de correction automatisée que nous mettons en place dans cette recherche reposent sur l'identification de "patrons", ou structures récurrentes, pour la détection, associés à des instructions de réécriture pour la correction, ces deux éléments formant une règle de correction. Dans cette optique, connaître la structure des groupes adverbiaux permet en théorie de prévoir les différentes configurations possibles. Cependant, comme nous l'avons vu, il est plus fréquent pour un adverbe d'être utilisé seul que d'être accompagné de dépendants. *LGSWE* apportent quelques informations concernant les dépendants les plus fréquemment utilisés ; en dehors des modifieurs liés à la comparaison (ex. : *too*, *much*, etc.) et de ceux utilisés dans un registre informel (ex. : *pretty*), les adverbes les plus utilisés sont *very*, *really*, *quite*, et *almost* pour le sous-corpus de prose académique (op. cit. : 546). Nous utilisons ces informations de fréquence pour la création du modèle présenté à la fin de cette partie, et des règles de correction qui en découlent dans le Chapitre 3.

d. Syntaxe externe de l'adverbe

Les rôles syntaxiques de l'adverbe

D'après Pullum et Huddleston, la principale caractéristique syntaxique de l'adverbe anglais est d'être utilisé comme modifieur d'éléments autres que les noms (op. cit. : 562). Derrière cette généralisation en creux se cache cependant une distribution syntaxique complexe.

Pour commencer, les adverbes peuvent fonctionner comme compléments d'un petit nombre de verbes et de prépositions. Le rôle de l'adverbe comme complément d'une préposition est mentionné dans les trois ouvrages ; notons cependant qu'en raison des différences dans la délimitation de la catégorie, certains des adverbes présentés dans cette fonction par les auteurs de *ACG* et *LGSWE* seraient analysés dans *CGE* comme des prépositions ou des pronoms (ex. : *until after*, *before today*, etc. ; Quirk et al., 1985 : 455). Le rôle de l'adverbe comme complément d'un verbe, d'après l'approche de *CGE*, est analysé comme un "adverbial obligatoire" (*obligatory adverbial*) dans *ACG* (op. cit. : 56) et *LGSWE* (op. cit. : 143, 381).

Les exemples suivants, extraits de *CGE*, montrent des groupes adverbiaux dans ces fonctions (op. cit. : 574):

[55] *You'll have [to word your reply very carefully]*.

[56] *I didn't hear about it [until recently]*.

Dans *CGE*, le GAdv *very carefully* est considéré comme un complément car il est obligatoire pour conserver le sens original de l'énoncé et la grammaticalité de la proposition. Il correspond aux compléments de manière, que nous avons évoqués plus haut. On retrouve cette configuration avec des verbes tels que *treat*, *phrase* et *behave*, qui, pour un de leurs sens au moins, doivent être complétés par un élément exprimant la manière.

La principale fonction des adverbes reste cependant celle de la modification. Comme il a été montré dans la section 2.2.1a, les auteurs de *ACG* et *LGSWE* n'utilisent pas le terme de "modifieur" avec une application aussi étendue que les auteurs de *CGE*. Par conséquent, ils distinguent les utilisations de l'adverbe en tant que modifieur des adjectifs, des adverbes et d'autres éléments (GN, GP, déterminants), de ses utilisations en tant qu'adverbial. *CGE* adopte une vision globale du rôle de modifieur des adverbes, indiquant qu'ils modifient un grand nombre d'éléments, allant des adjectifs et adverbes au verbe et à la proposition. Étant donné que les adverbes peuvent également fonctionner comme supplément au niveau de la proposition, *CGE* regroupe ces deux fonctions, modifieur et supplément, sous l'appellation d'"adjoint" (*adjunct*), avec une définition qui diffère de celle que *ACG* attribue au terme (cf. 2.2.1a).

D'après l'aperçu des erreurs produites dans notre corpus, l'utilisation des adverbes en tant que modifieur dans le GAdj et le GAdv n'est pas une source d'erreur importante. Nous ne nous attardons donc pas sur ce type de fonction. De la même façon, nous laissons de côté pour le moment les utilisations des adverbes dans des structures comparatives et superlatives, qu'ils

fassent partie de leur construction ou qu'ils en soient l'objet. Le reste de cette sous-partie est donc consacré à la fonction de modification au niveau du GV et de la proposition.

Les adverbes en fonction d'adjoint/adverbial dans la proposition

Avant de commencer, nous attirons l'attention sur le fait que la description des fonctions exercées par les adverbes dans la proposition implique d'aborder des phénomènes qui dépassent le cadre de la syntaxe de l'adverbe : pour preuve, la question des adverbiaux, pour *ACG* et *LGSWE*, ou des adjoints, pour *CGE*, occupe dans ces ouvrages un chapitre entier, distinct de celui consacré aux adjectifs et aux adverbes, et bien plus long. Notre objectif n'étant pas de produire une synthèse exhaustive des descriptions du fonctionnement des adverbiaux/adjoints, nous tâchons de limiter autant que possible la présentation des descriptions à celles qui concernent précisément les adverbes.

Les auteurs des trois ouvrages adoptent trois descriptions différentes de la fonction des adverbes en tant qu'adverbial (dans le sens de *ACG* et *LGSWE*) ou adjoint (dans le sens de *CGE*).

Dans la description donnée par Quirk et al., les adverbiaux constituent une catégorie de fonctions hétérogène, certains adverbiaux étant considérés comme périphériques à la proposition et optionnels, d'autres comme centraux et obligatoires (op. cit. : 51). Les auteurs identifient quatre types d'adverbiaux :

- les adjoints (*adjuncts*),
- les subjoints (*subjuncts*),
- les disjoints (*disjuncts*),
- les conjoints (*conjuncts*).

Les phrases suivantes, extraites de *ACG*, sont des exemples de ces quatre types, dans le même ordre (op. cit. : 440-441) :

[57] *He spoke to me about it <u>briefly</u>.* (adjoint)

[58] *We haven't <u>yet</u> finished.* (subjoint)

[59] *<u>Fortunately</u>, no one complained.* (disjoint)

[60] *All our friends are going to Paris this summer. We, <u>however</u>, are going to London.* (conjoint)

Les adjoints et les subjoints démontrent une plus grande intégration dans la proposition que les disjoints et les conjoints, qui sont à sa périphérie. Ceci est visible pour ces derniers dans les exemples ci-dessus en raison de leur détachement typographique, représentant un détachement prosodique. D'un point de vue sémantique, les disjoints véhiculent une évaluation de l'énoncé, liée à la façon dont ce dernier est communiqué, ou bien à son contenu propositionnel. Les conjoints expriment quant à eux la relation existant entre deux unités.

Dans l'approche de Quirk et al., les adjoints et subjoints ne sont pas adossés respectivement à un seul type sémantique d'adverbes, et peuvent véhiculer un ensemble de sens variés. Ils se distinguent des deux autres types, et l'un de l'autre, par des critères syntaxiques. Par exemple, Quirk et al. indiquent que certains sous-types d'adjoints sont obligatoires pour le sens et la grammaticalité de la proposition (*obligatory predication adjuncts*, op. cit. : 505), et qu'ils peuvent en général être le focus de structures clivées. Les subjoints ne présentent pas ces particularités syntaxiques, mais ne sont pas non plus détachés de la proposition d'un point de vue prosodique, comme le sont généralement les disjoints et les conjoints. Les adverbes *also* et *only*, identifiés par Pullum et Huddleston comme des "modifieurs focalisants additifs ou restrictifs" (*additive/restrictive focusing modifiers*, op. cit. : 586), sont considérés dans *ACG* comme des subjoints lorsqu'ils fonctionnent au niveau de la proposition. La syntaxe de ces adverbes particuliers est présentée dans la sous-section suivante.

Chaque type d'adverbial admet de trois à huit sous-types, pour un total de vingt-deux sur l'ensemble des adverbiaux. Ce degré de précision rend cette description dans le même temps exhaustive et difficile à résumer. Les distinctions entre les différents types sont parfois ténues, du propre aveu des auteurs (op. cit. : 567), et ces types reposent en réalité largement sur le rôle sémantique des éléments concernés. Ce paramètre mène par ailleurs à des recoupements avec les cadres des autres ouvrages. Par ailleurs, cette classification concerne les adverbiaux dans leur ensemble et non seulement les adverbes.

Les descriptions données par les auteurs de *LGSWE* sont généralement semblables à celles de *ACG*, et ceux-ci conservent d'ailleurs le terme de "adverbial", avec la même définition. Cependant, *LGSWE* abandonne l'analyse des adverbiaux en quatre catégories aux appellations relativement opaques pour adopter une division en trois catégories tirant leur nom du rôle sémantique de l'adverbial concerné (op. cit. : 549) :

- les adverbiaux circonstanciels (*circumstance adverbials*),
- les adverbiaux de positionnement (*stance adverbials*),

- les adverbiaux connecteurs (*linking adverbials*).

Les phrases suivantes, tirées de *LGSWE*, sont des exemples de l'utilisation d'adverbes dans ces fonctions (op. cit. : 549) :

[61] *He took it in slowly but uncomprehendingly.*

[62] *His book undoubtedly fills a need.*

[63] *The weight of bureaucracy still hangs a trifle heavy. Nevertheless, the review represents substantial progress.*

Les adverbiaux circonstanciels décrivent les circonstances ou conditions dans lesquelles l'action ou l'état présenté dans la proposition a lieu. Ces informations sont généralement d'ordre temporel ou spatial, ou concernent la manière dont l'action est effectuée.

Les adverbiaux de positionnement véhiculent l'attitude que la personne locutrice adopte au sujet du contenu du message ; cette attitude peut prendre la forme d'une évaluation épistémique, ou bien concerner un point de vue général sur le message, ou encore clarifier l'état d'esprit dans lequel le message est énoncé. L'exemple [62], cité plus haut, illustre l'utilisation épistémique des adverbiaux de positionnement ; les phrases ci-dessous, tirées de *LGSWE*, sont des exemples des deux autres types (op. cit. : 549) :

[64] *Then, amazingly, he would turn his microphone over to his daughter Maureen.*

[65] *And he sounded a bit low, quite frankly, to me yesterday on the phone.*

Le dernier type d'adverbial, les adverbiaux connecteurs, sert à expliciter le lien entre différentes portions de textes (propositions, phrases ou paragraphes), comme on peut le voir dans l'exemple [63].

Comme dans la classification de Quirk et al., ces quatre types d'adverbiaux diffèrent dans leur degré d'intégration à la proposition. Les adverbiaux circonstanciels sont décrits comme plus centraux à la proposition d'un point de vue syntaxique et sémantique, les deux autres types étant décrits comme périphériques. Bien que les auteurs de *LGSWE* aient opéré une simplification, tant théorique que terminologique, du cadre de *ACG*, il existe une grande correspondance entre ces deux cadres. Les adverbiaux circonstanciels correspondent ainsi largement aux adjoints de Quirk et al., les adverbiaux de positionnement aux disjoints, et les adverbiaux connecteurs au conjoints. Les subjoints se partagent entre les adverbiaux circonstanciels et les adverbiaux de positionnement, selon les sous-types sémantiques.

Au-delà d'une description remodelée des types d'adverbiaux, *LGSWE* présentent une étude de corpus de la distribution, de la réalisation, et du placement des adverbiaux. Nous revenons sur la question du placement des adverbiaux (dans le sens de *LGSWE* et *ACG*) ou adjoints (dans le sens de *CGE*) dans la section 2.2.1e, mais souhaitons d'ores et déjà donner un aperçu des résultats de Biber et al. concernant la distribution et la réalisation de cette fonction des adverbes. Rappelons cependant que ces auteurs adoptent une délimitation plus large que Pullum et Huddleston de la catégorie des adverbes, ce qui affecte nécessairement leurs résultats. Nous souhaitons par ailleurs mentionner le fait que le manque de résultats chiffrés dans le compte-rendu suivant est imputable au manque de données précises données dans *LGSWE*, certainement dû au choix des auteurs d'éviter de submerger leur lectorat par des résultats chiffrés d'analyse de corpus au mot près.

Les adverbiaux peuvent être réalisés par plusieurs types de structures (GP, GN, GAdv, propositions), et si les GP sont les plus fréquents, les adverbes, fonctionnant seuls ou comme noyaux d'un GAdv, tiennent la seconde place. Ces derniers sont le plus souvent utilisés dans les adverbiaux connecteurs, et le moins souvent dans les adverbiaux circonstanciels (op. cit. : 768-769).

Cependant, parmi les trois types d'adverbiaux identifiés, les adverbiaux circonstanciels sont de loin les plus fréquents dans l'ensemble des quatre sous-corpus qu'utilisent Biber et al. (op. cit. : 765). Parmi ceux-ci, les adverbiaux circonstanciels de processus (*process circumstance adverbials*) sont les plus fréquents, et à l'intérieur même de ce sous-type, ce sont les adverbiaux exprimant la manière qu'on retrouve en plus grand nombre (op. cit. : 783). En toute logique, les adverbiaux de manière qui ne sont pas réalisés par des GP prennent le plus souvent la forme d'adverbes, fonctionnant généralement seuls (op. cit. : 787). Il est également très fréquent pour des adverbes d'être utilisés dans l'expression du degré et de la focalisation additive/restrictive (op. cit. : 788). Biber et al. ont également étudié la diversité des adverbes utilisés en fonction d'adverbial circonstanciel. Les adverbes de manière sont ceux qui présentent la plus grande diversité, avec le plus faible taux de répétition d'un même adverbe et le plus grand nombre de termes différents utilisés (op. cit. : 793).

Comparé aux approches de Quirk et al. et Biber et al., Pullum et Huddleston adoptent un cadre descriptif des plus simples. Ceci est visible en partie dans le fait que ces derniers abandonnent toute velléité de classification des adjoints en plusieurs types (terme utilisé dans le sens de *CGE*). Ils sont également les seuls à donner des informations spécifiques sur le fonctionnement des adverbes comme adjoints dans la proposition, ainsi que sur leur

placement, sans renvoyer uniquement au placement des adjoints de toutes réalisations syntaxiques. Ce choix d'organisation est justifié par le fait que la position des adverbes dans la proposition diffère généralement de celle des GP, qui sont la réalisation syntaxique la plus fréquente pour les adjoints (op. cit. : 574).

Comme nous l'avons déjà mentionné, Pullum et Huddleston indiquent que la fonction principale des adverbes est celle de la modification, et que celle-ci peut s'appliquer à un grand nombre de structures différentes, sauf les noms. Ils offrent ainsi une vision unifiée du rôle des adverbes. Cependant, au niveau de la proposition, les adverbes peuvent fonctionner soit comme modifieur du GV, soit comme supplément de la proposition. Les suppléments sont des éléments qui occupent une position dans la séquence linéaire de la proposition, mais qui ne font pas partie de sa structure syntaxique, ce qui les distinguent des modifieurs qui sont eux pleinement intégrés à la structure qu'ils modifient. Dans un arbre syntaxique, les suppléments sont représentés par un arbre indépendant, existant en parallèle de celui de la structure à laquelle ils s'intègrent linéairement (Mittwoch et al., 2002 : 1354). Ce détachement est souvent signalé à l'écrit par l'utilisation de virgules. Le terme d'adjoint est donc utilisé dans *CGE* pour faire référence à ces deux fonctions des adverbes en rapport avec la proposition. Les deux exemples suivants, extraits de *CGE*, illustrent ces fonctions (op. cit. : 576) :

[66] *She walked <u>unsteadily</u> to the door.* (modifieur)

[67] *<u>Moreover</u>, he didn't even apologise.* (supplément)

La présentation des différents types d'adjoints dans *CGE* n'est pas fondée sur une catégorisation en types et sous-types, mais repose sur l'identification des types sémantiques les plus fréquents, puis sur le passage en revue de leur description en fonction de sept caractéristiques. Ces dernières incluent les questions de portée, de réalisation syntaxique, de capacité à être le focus de structures clivées et de questions, et de placement dans la proposition. La description du fonctionnement syntaxique des adverbes au niveau de la proposition est fondée sur le même ensemble de critères. Celle-ci étant largement dépendante des types sémantiques, nous incluons cette problématique dans la section suivante, consacrée au placement des adverbes, ou adverbiaux pour *ACG* et *LGSWE*. Avant d'aborder cette thématique, nous terminons cette section par une présentation de la fonction de modification spécifique attachée à un petit nombre d'adverbes.

Les adverbes en fonction de modifieur focalisant

Les adverbes *also* et *only* font partie, avec d'autres, des adverbes utilisés comme "modifieur focalisant" (*focusing modifier*, Pullum et Huddleston, 2002 : 586 ; notre traduction). La fonction est également identifiée comme celle de "subjoint focalisant" dans *ACG* (*focusing subjunct*, op. cit. : 604 ; notre traduction), ou encore assimilée à celle des adverbiaux circonstanciels, pour le type sémantique addition/restriction dans *LGSWE* (op. cit. : 780). Nous adoptons le choix de termes de *CGE* pour les descriptions, dans cette section ainsi que dans le reste du présent document.

Ces adverbes se distinguent des adverbes prototypiques par un ensemble de caractéristiques. Ils ne peuvent pas recevoir de modifieur, et ne peuvent donc pas faire l'objet de structures comparatives, cette caractéristique découlant également du fait qu'ils ne peuvent être gradués. Ils ne peuvent servir de focus pour la construction de structures clivées, et ne peuvent faire partie de structures coordonnées avec un autre adverbe (Quirk et al., 1985 : 610-611). Ces trois impossibilités sont présentées dans les exemples ci-dessous, associés à des exemples utilisant d'autres adverbes (le GAdv est souligné pour plus de clarté) :

[68] *Do it more carefully next time.*

[69] **She was more also surprised to see them here.*

[70] *It was carefully that she explained the situation to him.*

[71] **It is only that you can look at them.*

[72] *He was painting quickly and sloppily.*

[73] **He had also and carefully selected a new color for the kitchen.*

Ils partagent cependant la caractéristique syntaxique centrale des adverbes, qui est de modifier un ensemble large de structures, en dehors des noms ; ces structures peuvent être divisées en deux catégories : les groupes et les propositions (Pullum et Huddleston, 2002 : 586-587). Puisque la présente étude s'intéresse uniquement au placement des adverbes dans la proposition et dans le GV, nous nous concentrons en théorie sur l'utilisation des adverbes en fonction de modifieur focalisant dans ces structures. Nous verrons cependant en pratique que, ces modifieurs privilégiant majoritairement un type de placement, les distinctions entre les différents supports syntaxiques de la modification (ex. : GN vs. GV) s'estompent.

Bien qu'elles n'utilisent pas exactement les mêmes termes pour décrire cette fonction, les trois grammaires s'accordent pour classer ces modifieurs en deux catégories principales : les

modifieurs focalisants additifs, et les modifieurs focalisants restrictifs. Voici les définitions des adverbes additifs et restrictifs données par Biber et al (op. cit. : 780) :

> Even when they occur as adverbials at clause level, additive adverbs typically single out one particular part of the clause's meaning as being 'additional' to something else (often implied). […] Restrictive adverbs […] are similar to additive adverbs in that they focus attention on a certain element of the clause. They serve to emphasize the importance of one part of the proposition, by restricting the truth value of the proposition either primarily or exclusively to that part.

Les deux phrases suivantes, empruntées à Pullum et Huddleston (op. cit. : 586), sont des exemples de ces deux types sucessivement :

[74] *Jill had <u>also</u> attended the history seminar.*

[75] *You can <u>only</u> exit from this lane.*

Ces deux fonctions concernent un nombre très restreint d'adverbes, dont la liste quasi-exhaustive est présentée ci-dessous, par ordre alphabétique :

- additif : *also*, *as well*, *even*, *too*,
- restrictif : *alone*, *but*, *exactly*, *exclusively*, *just*, *merely*, *only*, *precisely*, *purely*, *simply*, *solely*.

Parmi les modifieurs (ou subjoints) focalisants restrictifs, Quirk et al. distinguent les adverbes "exclusifs" (*exclusives*), qui restreignent l'application de l'énoncé exclusivement au focus du modifieur/subjoint, comme *only* et *exclusively*, et les adverbes "particularisants" (*particularizers*), qui indiquent simplement une prédominance du focus, comme les adverbes *especially* et *mostly* (op. cit. : 604). Ces derniers sont identifiés dans *CGE* comme des modifieurs restrictifs partiels (*partial restrictive focusing modifiers*, op. cit. : 592).

Nous verrons dans la section consacrée au placement des adverbes que la position privilégiée par les modifieurs focalisants est celle que l'on peut observer dans les exemples [74] et [75] ci-dessus, c'est-à-dire une position médiane. Or ce placement introduit une ambiguïté dans la portée sémantique de l'adverbe, celui-ci pouvant porter sur plusieurs éléments du GV, comme le verbe, les compléments directs et indirects, les adjoints, ou encore une partie seulement de ceux-ci. Pullum et Huddleston attirent l'attention sur le fait que le constituant que ces adverbes modifient d'un point de vue syntaxique n'est pas assimilable à leur "focus", ou cible (op. cit. : 596) : "In order to understand the meaning contribution of

only and *also* it is not sufficient to identify the syntactic head that they modify: one must know which element they apply to semantically". Cette ambiguïté n'est cependant pas plus problématique que celle qui est posée la plupart du temps par la négation verbale, qui dépend syntaxiquement du GV mais dont la portée peut être restreinte à certains éléments de celui-ci, sans que cela présente une gêne importante pour les personnes utilisant l'anglais grâce à la présence d'un contexte précis (Quirk et al., 1985 : 605-606). Biber et al. notent également que, même s'ils s'appliquent à une partie de la proposition uniquement, les modifieurs focalisants additifs jouent un rôle dans la cohésion du texte (Biber et al., 1999 : 780).

Les adverbiaux additifs et restrictifs, selon la terminologie de *LGSWE*, sont parmi les trois types d'adverbiaux circonstanciels les plus courants. L'étude de la distribution des termes instanciant ces fonctions montre une très faible diversité, celles-ci étant réalisées par un très petit nombre d'adverbes, avec *also* et *only* en tête pour les corpus écrits (Biber et al, 1999 : 795-796). *Also* est particulièrement fréquent dans les corpus de documents informatifs et descriptifs, comme les écrits universitaires.

e. Le placement des adverbes dans la proposition

L'objectif de cette section est de passer en revue les différents paramètres identifiés et les généralisations formulées dans les grammaires de référence, afin d'aboutir à une modélisation du placement des adverbes en fonction d'adjoint/adverbial. L'un des paramètres les plus importants étant le type sémantique de l'adverbe concerné, cette question est abordée en premier. Nous présentons ensuite les trois principales positions existantes, et en dernier lieu nous synthétisons les informations recueillies.

La question du placement des adverbes en séries n'est pas abordée, la variété des placements possibles rendant ce phénomène relativement rare : s'il est nécessaire d'inclure plusieurs adverbes dans une proposition, ceux-ci adopteront le plus souvent des placements différents. La question de l'influence de la négation n'est pas abordée non plus à ce stade mais elle est prise en compte dans la sous-partie 2.2.2 consacrée à la modélisation des phénomènes dans l'objectif de créer des règles de détection et de correction.

Les types sémantiques d'adverbe en fonction d'adjoint/adverbial

Comme indiqué dans les trois ouvrages, le type sémantique des adjoints ou adverbiaux a une grande influence sur leur placement dans la proposition (Quirk et al., 1985 : 491 ; Biber et al., 1999 : 773 ; Pullum et Huddleston, 2002 : 577). La question de la classification

sémantique des adverbes dans plusieurs langues a suscité l'attention de nombreux et nombreuses linguistes, depuis les années 1940 en particulier (Nølke, 1990 : 117). Parmi les tentatives de classification les plus rigoureuses pour les adverbes anglais, on peut citer les travaux de Greenbaum, qui est par ailleurs un des auteurs de *ACG* (1969), Jackendoff (1972) et Bellert (1977). Nous nous conformons cependant à la méthode énoncée en 2.1.1 en donnant la priorité aux synthèses présentées dans les trois grammaires de référence.

Les auteurs de *CGE* ont choisi de présenter séparément les types sémantiques des adverbes lorsqu'ils sont en fonction d'adjoint, mais ce n'est pas le cas des auteurs de *ACG* et *LGSWE* : Quirk et al. n'indiquent pas de types sémantiques spécifiquement pour les adverbes, et se reposent sur la catégorisation donnée pour les adverbiaux, tandis que Biber et al. classent les adverbes en types sémantiques s'appliquant à toutes les fonctions instanciées par ces derniers.

Pour cette raison, la synthèse des placements possibles pour les adverbes en fonction de leur type sémantique, présentée dans le tableau 15 dans la section suivante, repose en majeure partie sur la classification donnée par Pullum et Huddleston, complétée par des indications relevées dans les autres ouvrages. Afin de mettre en lumière les correspondances existant entre les différentes classifications, celles-ci sont illustrées dans le tableau suivant. Nous avons laissé de côté la classification de Quirk et al., puisqu'elle concerne l'ensemble des structures pouvant fonctionner comme adverbial. Le GP étant la structure la plus courante pour les adverbiaux, et non le GAdv, on ne peut présenter de correspondances directes.

Le tableau 15 présente les types sémantiques identifiés dans *CGE* dans la colonne de gauche, et les types correspondants dans *LGSWE* dans celle de droite. Nous indiquons les types supplémentaires identifiés par Biber et al. lorsque ceux-ci ne sont pas représentés chez Pullum et Huddleston. Les types sémantiques sont accompagnés d'exemples extraits des grammaires respectives, parfois tronqués pour éviter de surcharger le tableau. Dans certains cas, le terme utilisé dans l'exemple n'est pas identifié comme un adverbe dans *CGE*. Il n'existe pas toujours de correspondance directe entre les catégories de *CGE* et celles de *LGSWE*. Une catégorie dans une colonne peut être répartie sur plusieurs catégories dans la colonne opposée, ou bien ne pas y trouver d'équivalent du tout. Lorsqu'une catégorie admet plusieurs sous-types, ceux-ci sont énumérés individuellement, sauf dans le cas de la catégorie "Lieu", qui n'a pas de correspondance dans *CGE*. Les correspondances que nous établissons sont discutées dans les paragraphes suivants lorsque celles-ci sont problématiques.

CGE	***LGSWE***
	Lieu (position, direction, distance) [76] *It hopped backward among its companions.*
Manière [77] *She walked unsteadily to the door.*	**Manière** [78] *You can run fast but not here.*
Moyen (*means*) et instrument [79] *Planets can be detected radio-telescopically.*	**Moyen** [80] *The technical achievement of opening a vessel measured angiographically was similarly successful for both groups of patients.*
Lié à l'acte (*act-related*), évaluation subjective [81] *She has foolishly gone to the police.*	
Lié à l'acte (*act-related*), évaluation de l'intention [82] *They deliberately kept us waiting.*	
Degré [83] *The share price has increased enormously.*	**Degré** [84] *They thoroughly deserved a draw last night.*
Lieu temporel (*temporal location*) [85] *She subsequently left town.*	**Temps, position** [86] *She doesn't say go away very much now.*
Durée [87] *We were staying in a motel temporarily.*	**Temps, durée** [88] *She will remain a happy memory with us always.*
Aspect [89] *Some of the guests are already here.*	**Temps, relation** [90] *When they took the old one it was already in seven pieces!*
Fréquence [91] *Do you come here often?*	**Temps, fréquence** [92] *The product itself is often dull and unchanging.*
Chronologie (*serial order*) [93] *The play was next performed in 1901.*	**Temps, relation** [94] *Note that the store location accessed still contains a copy of the information.*
Domaine [95] *Politically, the country is always turbulent.*	
Modalité [96] *This is necessarily rather rare.*	**Positionnement, évaluation épistémique** [97] *No it's alright, I'll probably manage without it.*
Évaluation [98] *Fortunately this did not happen.*	**Positionnement, attitude** [99] *I lost the manual that goes with it, unfortunately.*
Lié à l'acte de parole [100] *Frankly, I'm just not interested.*	**Positionnement, style** [101] *Quite simply, life cannot be the same.*
Connecteur [102] *Moreover, he didn't even apologise.*	**Connecteur** [103] *Police, however, would not say where they were.*
Focalisation additive/restrictive [104] *Jill had also attended the history seminar.*	**Addition/restriction** [105] *The formula also shows the number of moles.*

Tableau 15. Types sémantiques des adverbes en fonction d'adjoint/adverbial

La catégorie "Lieu" dans la classification de Biber et al. ne trouve pas de correspondance chez Pullum et Huddleston en raison de la délimitation plus restreinte de la catégorie syntaxique des adverbes, qui en exclut l'essentiel des termes à dénotation spatiale au profit de la catégorie des prépositions. Les types sémantiques liés au temps ne sont pas gérés de la même façon : ceux-ci sont regroupés sous un intitulé général avec quatre sous-types dans *LGSWE*, alors que ces sous-types sont énumérés individuellement dans *CGE*. Par ailleurs, le sous-type "Temps, relation" de *LGSWE*, qui identifie les utilisations des adverbes indiquant les relations dans le temps entre deux évènements ou états, semble regrouper les types "Aspect" et "Chronologie" de *CGE*.

La catégorie "Domaine", qui concerne les adverbes servant à restreindre le domaine auquel la proposition s'applique, n'a pas d'équivalent dans la description de *LGSWE* ; Biber et al. mentionnent cependant dans leur description des fonctions sémantiques des adverbiaux une catégorie mineure intitulée "Respect" (op. cit. : 781-782), qui pourrait correspondre à celle de Pullum et Huddleston. Les exemples donnés, ainsi que la définition ne mentionnent cependant pas le rôle des adverbes.

La catégorie intitulée "Lié à l'acte" (proposition de traduction du terme *act-related*) concerne l'utilisation d'un adverbe pour véhiculer une évaluation sur le fait même que le sujet accomplisse l'action, plutôt que sur la manière dont l'action est menée. L'exemple [81] est paraphrasable de la façon suivante : "It was foolish of her to go to the police", et non "She has gone to the police in a foolish manner". La catégorie admet deux variations : l'évaluation subjective, que nous venons d'illustrer et dans laquelle la personne énonciatrice porte un jugement sur l'action menée et le sujet agentif, et l'évaluation de l'intention, qui concerne l'intention avec laquelle le sujet agentif accomplit l'action. Cette catégorie n'est présente que dans *CGE*, et elle ne semble pas avoir d'équivalent défini dans *LGSWE*. Il est possible que le premier sous-type soit assimilé au sous-type "Positionnement, attitude", dans une version élargie, et le second au type général "Manière".

Comme il a été mentionné en introduction à cette sous-partie, nous avons également ajouté dans la colonne *CGE* le type de la focalisation additive/restrictive que les auteurs n'intègrent pas à leur présentation des types sémantiques et préfèrent traiter séparément. Celui-ci ayant un équivalent facilement identifiable dans *LGSWE*, nous avons choisi de l'intégrer dans le tableau à ce stade.

La principale originalité de la description de Pullum et Huddleston consiste en leur prise en compte, dans leur classification sémantique, de la question de la portée de l'adjoint, et donc de

l'adverbe dans cette fonction. Les types sémantiques y sont donc divisés en deux ensembles, ceux dont la portée se limite au GV (*VP-oriented*), et ceux dont la portée s'étend à la proposition (*clause-oriented*). Les deux ensembles sont présentés dans le tableau suivant (op. cit. : 576) ; nous reprenons les exemples donnés dans le tableau précédent pour clarifier les différents types :

Types d'adjoint portant sur le GV	**Types d'adjoint portant sur la proposition**
Manière ex. : *She walked unsteadily to the door.*	**Domaine** ex. : *Politically, the country is always turbulent.*
Moyen ou instrument ex. : *Planets can be detected radio-telescopically.*	**Modalité** ex. : *This is necessarily rather rare.*
Lié à l'acte (évaluation subjective, évaluation de l'intention) ex. : *She has foolishly gone to the police.* ex. : *They deliberately kept us waiting.*	**Évaluation** ex. : *Fortunately this did not happen.*
Degré ex. : *The share price has increased enormously.*	**Lié à l'acte de parole** ex. : *Frankly, I'm just not interested.*
Lieu temporel ex. : *She subsequently left town.*	**Connecteur** ex. : *Moreover, he didn't even apologise.*
Durée ex. : *We were staying in a motel temporarily.*	
Aspect ex. : *Some of the guests are already here.*	
Fréquence ex. : *Do you come here often?*	
Chronologie ex. : *The play was next performed in 1901.*	

Tableau 16. Classification des types d'adjoint selon leur portée dans CGE

Les auteurs indiquent que cette distinction a pour objectif de permettre la formulation de généralisations concernant le placement des adverbes dans la proposition. Ils font les observations suivantes concernant la portée des différents types sémantiques d'adjoint (op. cit. : 576) :

> VP-oriented adjuncts denote modification of the details of the predicate of a clause: if the predication corresponds semantically to a type of action, adjuncts of these types tend to specify aspects such as the way in which the action was carried out, the time it took, the

> degree to which it was carried out, or the order in which it was done relative to other actions. […]
>
> Clause-oriented adjuncts represent modifications of the applicability of the clause content. That is, their semantic effect is to characterize how the propositional content of the clause relates to the world or the context. […] Clause-oriented adjuncts have meaning contributions that are much more external to the content of the proposition.

En parallèle de ces observations sémantiques, Pullum et Huddleston ajoutent que les GAdv qui instancient des adjoints portant sur le GV sont plus étroitement associés aux constituants du GV, et plus susceptibles d'être placés à l'intérieur ou à proximité de ce dernier, alors que ceux qui instancient des adjoints ayant portée sur toute la proposition manifestent le comportement inverse. Ceci les amène à conclure sur le rôle du type sémantique et de la portée des adverbes fonctionnant comme adjoints sur leur placement (op. cit. : 576) :

> Putting the syntactic and semantic observations together, we see that the closeness of the adjuncts in linear proximity to the predicator at the heart of a clause tends to correlate with the closeness of what the adjuncts express to the content of the proposition.

Malgré le poids des facteurs additionnels, que nous détaillons dans la sous-section suivante, la généralisation descriptive offerte par Pullum et Huddleston se révèle être un précieux outil prédictif lorsque l'on cherche à corriger automatiquement les erreurs de placement des adverbes.

Description des trois principaux placements des adverbes en fonction d'adjoint/adverbial

L'une des principales caractéristiques des adjoints (dans le sens de *CGE*) ou adverbiaux (dans le sens de *ACG* et *LGSWE*) est leur capacité à être positionnés à plusieurs endroits de la proposition. Cette caractéristique les distingue des autres éléments qui composent la proposition, dont le placement est relativement fixe, et autour desquels les adjoints/adverbiaux s'organisent (Quirk et al., 1985 : 490 ; Biber et al., 1998 : 770). Comme nous le verrons dans la sous-partie 2.2.2, cette flexibilité est source d'erreurs pour les personnes utilisatrices de l'anglais L2, notamment les francophones.

Considérant la diversité des placements et la variété des facteurs pesant sur leur sélection, Pullum et Huddleston soulignent prudemment la difficulté, voire l'impossibilité, d'énoncer des consignes strictes sur cette question (op. cit. : 576) :

> Only rather broad and approximate flexible generalisations about adjunct placement and sequence can be made. There is a great deal of variation in use, and features of context, style, prosody, and euphony play a role in some decisions.

Le champ du placement des adjoints/adverbiaux est cependant loin d'être inexploré, et les auteurs des trois grammaires de référence s'accordent sur l'existence de facteurs particulièrement déterminants (Quirk et al., 1985 : 491 ; Biber et al., 1998 : 773 ; Pullum et Huddleston, 2002 : 780). Les deux facteurs remportant l'unanimité sont d'une part le type de réalisation, c'est-à-dire la nature (ex. : GAdv, GP, GN) et la structure interne de l'adjoint/adverbial, et d'autre part sa catégorie sémantique. *LGSWE* mentionne également l'influence de la longueur du segment fonctionnant comme adjoint/adverbial, qui est liée à sa structure interne. Quirk et al. soulignent l'importance du type d'adverbial (adjoint, subjoint, disjoint ou conjoint) et de l'organisation de l'information dans la phrase. Ces différents facteurs ne sont bien sûr pas indépendants les uns des autres, la longueur d'un élément découlant de sa structure interne, et ayant une influence sur l'organisation de la phrase. D'après Quirk et al., il existe également une corrélation directe entre la réalisation des adjoints/adverbiaux et leur catégorie sémantique (op. cit. : 500). Nous reviendrons sur ces paramètres ainsi que sur leur application à l'utilisation des adverbes en fonction d'adjoint/adverbial.

Trois principaux types de placement sont identifiés dans la proposition. Ils sont relativement semblables dans les trois grammaires, mais font l'objet de choix de terminologie différents. Le tableau 17 en page suivant expose ces trois types de position, et leur définition accompagnée d'exemples tirés de *CGE*.

Pullum et Huddleston ont ainsi choisi d'utiliser les termes *Front* et *End* à la place des dénominations plus traditionnelles "Initial" et "Final" afin de mieux représenter le fait que l'adjoint peut être positionné en début ou en fin de proposition sans être nécessairement le premier ou le dernier élément, car on peut trouver plus d'un adjoint dans ces positions (op. cit. : 670). Notons cependant qu'il est relativement rare de trouver deux GAdv à la suite en début ou en fin de proposition, et en partie pour cette raison nous choisissons ici d'adopter la terminologie de Biber et. al., qui nous semble être la plus claire, notamment dans sa traduction française.

Les placements indiqués ci-dessous sont approximatifs, et deux d'entre eux (le placement médian et le placement en fin de proposition) admettent des variantes plus précises dans la

description de Quirk et al. Ces deux placements sont généralement ceux que privilégient les adverbes, nous rentrons donc dans le détail de leurs variantes.

ACG	***LGSWE***	***CGE***	**Définition et exemples**
Initial	Initial	En début (*Front*)	Avant le sujet, et éventuellement d'autres éléments de la proposition : [106] *Happily, they watched TV until dinner.*
Médian	Médian	Central	Entre le sujet et le verbe lexical, avec des variantes en fonction de la présence et du nombre d'auxiliaires : [107] *They happily watched TV until dinner.* [108] *They would probably watch TV for hours.*
En fin (*End*)	Final	En fin (*End*)	Après le verbe, et éventuellement certains ou la totalité de ses dépendants : [109] *They watched TV happily until dinner.*

Tableau 17. Les trois principaux placements des adverbes en fonction d'adjoint/adverbial

Le placement en position médiane est divisé en quatre sous-types, en raison de la possibilité de présence d'un ou plusieurs auxiliaires (op. cit. : 493-498) ; ces sous-types sont définis et illustrés dans le tableau suivant (exemples extraits de *ACG* ; nous reprenons également les abréviations de *ACG* sans adaptation vers le français) :

Sous-types	**Définition et exemples**
Médian-initial (*iM*)	Position entre le sujet et l'auxiliaire : [110] *She really had delighted her audience.*
Médian (*M*)	Position immédiatement après le sujet et, le cas échéant, le premier auxiliaire (BE et HAVE étant considérés comme des auxiliaires même lorsqu'ils sont les seuls verbes du GV) : [111] *The soprano was really at her best tonight.*
Médian-médian (*mM*)	Position immédiatement avant le dernier auxiliaire dans des GV en contenant trois ou plus : [112] *This bridge may have actually been designed by Brunel.*
Médian-final (*eM*)	Position immédiatement avant le verbe lexical : [113] *They will have seriously considered him for the post.*

Tableau 18. Les quatre sous-types du placement médian dans ACG

Avant toute chose, nous souhaitons souligner le fait, mentionné dans le tableau, que BE et HAVE sont considérés comme des auxiliaires (*operators* dans la terminologie de Quirk et al.) dans tous les contextes d'utilisation. Les adverbiaux (nous adoptons ici la terminologie de

Quirk et al. puisqu'il est question de leurs analyses) placés après ces derniers sont donc considérés comme étant positionnés en *M* et non en *iE* (*End*-initial, voir ci-dessous).

Dans certaines configurations, comme dans la phrase [107] où l'adverbe est placé entre le sujet et le verbe lexical seul, les frontières entre ces différents sous-types ne sont pas apparentes. Nous ne rentrons pas dans le détail des tests syntaxiques et sémantiques tout à fait convaincants présentés par Quirk et al. pour justifier l'identification de ces quatre sous-types, mais nous souhaitons ajouter quelques précisions pertinentes. En raison de la rupture que peut constituer le placement médian, ce dernier est le placement privilégié d'éléments légers, comme un GAdv fonctionnant comme adverbial, en particulier s'il est constitué d'un adverbe seul (op. cit. : 493). Parmi les quatre sous-types, le placement en *eM*, c'est-à-dire immédiatement avant le verbe lexical, est plus particulièrement associé à l'expression du degré et de la manière (op. cit. : 495). Le placement en *mM* est très rare, d'une part parce qu'il n'est apparent qu'en présence d'au moins trois auxiliaires dans le GV, et d'autres part parce qu'il ne constitue de toute façon pas un positionnement canonique pour les adverbes qui acceptent ce placement (op. cit. : 495).

Le placement en position finale admet quant à lui deux sous-types, en raison de la possibilité de placer les adverbiaux *après* le verbe mais *avant* d'autres éléments obligatoires (op. cit. : 498-500) ; ces deux sous-types sont définis et illustrés ci-dessous (exemples extraits de *ACG*) :

Sous-types	**Définition et exemples**
End-initial (*iE*)	Position avant un élément obligatoire en fin de proposition, directement après le verbe ou bien après le verbe et un autre élément obligatoire : [114] *She kept writing in a feverish rage long, violent letters of complaint.* [115] *He urged secretly that she be dismissed.* [116] *She placed the book offhandedly on the table.*
End (*E*)	Position après tous les éléments obligatoires dans la proposition : [117] *He urged her dismissal secretly.*

Tableau 19. Les deux sous-types du placement final dans ACG

Le choix du placement en *iE*, qui constitue une interruption de la continuité du GV, peut être le résultat de contraintes d'organisation de l'information (*end-weight* et *end-focus*), comme dans l'exemple [114], la longueur et/ou l'importance informationnelle donnée au complément du verbe amenant à privilégier un placement de celui-ci en fin de proposition. Ce placement non-canonique se révèle également nécessaire dans le cas où un complément ou un

autre adverbial est instancié par une proposition, comme dans l'exemple [115] : si l'adverbial est placé en *E*, soit après le GV de la proposition subordonnée, il risque d'être interprété comme appartenant à ce GV. Quirk et al. soulignent cependant que le placement en *iE* est en concurrence avec les placements médians, qui sont possibles pour la plupart des catégories sémantiques (op. cit. : 499-500). Lorsqu'il est rencontré dans les productions de personnes apprenantes, ce placement est fréquemment désigné par les initiales SVAO ; nous consacrons de nombreux paragraphes à la question des facteurs d'acceptabilité de ce placement dans la section 2.2.2a.

Pullum et Huddleston offrent une approche plus limitée de la question des variantes du placement médian, qu'ils nomment "central", indiquant simplement que ce placement regroupe les positions pré-verbale et post-auxiliaire (voir exemples [107] et [108]), cette dernière étant la plus fréquente lorsqu'un auxiliaire est présent (op. cit. : 782). Les adjoints peuvent néanmoins être placés avant le premier auxiliaire (voir exemple [110]). L'acceptabilité de cette variante dépend non seulement de la catégorie sémantique de l'adjoint, mais également de la structure et de l'intonation de la phrase : si les éléments post-auxiliaires font l'objet d'une ellipse, ou bien si l'auxiliaire est accentué, le placement pré-auxiliaire est généralement jugé comme acceptable (exemples empruntés à Pullum et Huddleston, op. cit. : 782) :

[118] *He <u>already</u> IS in the hospital.*

[119] *A: He should be in the hospital. B: He <u>already</u> is__.*

Par ailleurs, d'après Pullum et Huddleston, il est important de distinguer les éléments qui sont intégrés syntaxiquement et prosodiquement à la proposition, c'est-à-dire les modifieurs, de ceux qui en sont détachés, qu'ils nomment "suppléments" (voir définition du terme donnée en 2.2.1a). À l'écrit, le détachement prosodique (*prosodic detachment*, op. cit. : 577) des suppléments est le plus souvent signifié par des virgules, comme dans cet exemple :

[120] *Chris, <u>luckily</u>, had forgotten it.*

Il est particulièrement courant pour les adjoints (dans le sens de *CGE*) à l'initiale d'être détachés par une virgule. S'ils ne sont pas en début de proposition, les adjoints sont le plus souvent détachés lorsque la position qu'ils occupent n'est pas celle qui serait attendue pour leur catégorie sémantique. Même si des généralisations sont possibles concernant l'utilisation des adverbes comme suppléments dans la proposition, leur détachement prosodique leur

procure une très grande souplesse de placement (op. cit. : 577). Nous indiquons ces généralisations dans le tableau 21 présenté dans la sous-section suivante.

Avant d'évoquer la fréquence des adverbes dans les différents types de placement, nous souhaitons explorer l'avis des grammaires de référence concernant la grammaticalité du placement des adjoints/adverbiaux entre un verbe lexical et un GN complément. Ce placement est traditionnellement considéré comme erroné, du moins du point de vue de l'apprentissage et de l'enseignement de l'anglais comme langue seconde/étrangère (ex. : Le Prieult, 1996 : 86).

Le seul ouvrage, parmi les trois grammaires auxquelles nous nous référons, à mentionner clairement l'impossibilité de placer un modifieur entre un verbe et son objet est *CGE* (op. cit. : 535). Cette impossibilité provient de l'interdiction générale pesant sur la séparation du noyau d'un groupe de son GN complément (op. cit. : 575). Le fait que les auteurs n'hésitent pas à formuler cette impossibilité n'est pas surprenant étant donné la tendance de l'ouvrage, soulignée par Leech (2004 : 127), à privilégier les analyses tranchées. Pullum et Huddleston jugent donc les deux exemples suivants comme clairement agrammaticaux :

[121] **They watched happily TV until dinner.*

[122] **He didn't read carefully the report.*

Ce placement est cependant jugé comme possible et même souhaitable lorsque le GN objet est postposé en raison de contraintes informationnelles. *CGE* en donne l'exemple suivant :

[123] *I have read very carefully all the articles she has written.*

Comme l'ont souligné Quirk et al., ce placement est cependant en concurrence avec un placement juste avant le verbe lexical. Dans l'exemple donné, le GAdv qui rompt la continuité verbe-objet est lui-même relativement long (adverbe noyau de trois syllabes accompagné d'un modifieur) ; en conséquence, un placement avant le verbe lexical risque d'être jugé comme peu naturel :

[124] *I have very carefully read all the articles she has written.*

Par ailleurs, comme nous le verrons dans la section 2.2.2c, les études de corpus et les tests de jugement de grammaticalité montrent que les GAdv incluant un modifieur privilégient un placement final plutôt que médian. Pullum et Huddleston ne s'étendent pas sur cette question, mais il semble que la longueur du GAdv soit un paramètre à prendre en compte dans le jugement de l'acceptabilité des placements non-canoniques entre verbe et objet.

Comme nous venons de le voir, le placement SVAO correspond à la position *iE* dans le cadre descriptif de *ACG*. Quirk et al. mentionnent la question de la longueur, de la nature et de l'importance informationnelle du complément du verbe dans le choix de cette position. Cependant, contrairement à *CGE*, les auteurs n'évoquent aucune impossibilité grammaticale concernant ce placement. La phrase suivante est d'ailleurs utilisée comme exemple dans l'évaluation de son acceptabilité (op. cit. : 490) :

[125] *The book must have been placed by then on the shelf.*

Cette réticence à déclarer des segments comme étant agrammaticaux a d'ailleurs été relevée par F. Aarts dans sa critique de l'ouvrage. Si les auteurs ne s'expriment pas explicitement sur l'agrammaticalité d'un placement en *iE* qui ne serait pas le résultat des contraintes fortes que nous venons d'évoquer, leur appréciation est néanmoins visible en creux. Ainsi, l'introduction à la description du placement des adverbiaux indique que "aucun placement n'est inacceptable" (op. cit. : 490 ; notre traduction), et pourtant on ne trouve dans les analyses aucun exemple du placement d'un adverbial entre un verbe lexical et un GN objet qui ne soit pas long et apportant une information nouvelle. En outre, aucun exemple n'est donné de l'utilisation d'un adverbe dans cette configuration. L'exemple [125] ci-dessus n'est d'ailleurs pas un exemple de GAdv en fonction d'adverbial placé entre un verbe et un GN complément d'objet, mais un GP adverbial placé avant un GP complément de lieu.

Par ailleurs, leur présentation du placement en *iE* inclut la phrase suivante (op. cit. : 499 ; italiques ajoutés par nos soins) : "This position seems least disturbing of *normal order* when an adverbial comes between a direct object and obligatory adverbial", qui semble indiquer leur reconnaissance d'un ordre des mots "normal", et donc au moins de la maladresse d'un placement en *iE* qui ne soit pas fortement motivé.

La même analyse peut être faite pour *LGSWE*, qui suit de manière générale le cadre descriptif de *ACG*. Le placement des adverbiaux entre un verbe et un complément d'objet direct y est jugé possible (op. cit. : 771) : "Adverbials can also be placed after the main verb but preceding other obligatory elements of the clause, such as obligatory adverbials, subject predicatives, and direct objects". Pourtant, aucun des exemples cités ne présente cette configuration avec un GN objet court.

La conclusion que l'on peut apporter est donc que si certains auteurs ne se sont pas prononcés explicitement sur l'agrammaticalité de ce placement, en partie à cause de son acceptabilité dans certaines conditions dont la validité est subjective (longueur relative du GN

objet, importance relative donnée à l'information apportée), mais également parce que l'objet d'une grammaire descriptive n'est pas de faire la liste de toutes les configurations impossibles, celui-ci n'est pas fréquent en anglais L1, et risque d'être jugé comme erroné ou inacceptable par des locutrices et locuteurs natifs (comme non-natifs).

L'exploration de la syntaxe du placement des adjoints/adverbiaux n'est pas aussi poussée dans *LGSWE* que dans les deux autres ouvrages, mais ce dernier présente l'avantage de fournir des données chiffrées concernant la fréquence des trois placements identifiés. Le placement qui est de loin le plus fréquent est le placement en fin de proposition, notamment pour les adverbiaux circonstanciels (dans le sens de *LGSWE* et *ACG*), qui représentent par ailleurs la majorité des adverbiaux. Cette fréquence s'explique facilement par le fait que la catégorie sémantique de ces adverbiaux les amène à avoir une orientation verbale. Ils ont donc tendance à être positionnés à la suite du verbe, et non avant ce dernier, en position médiane ou initiale. Les adverbiaux de positionnement sont quant à eux généralement placés en position médiane, et les adverbiaux connecteurs se retrouvent, en toute logique, en début de proposition (op. cit. : 772-773).

Cependant, comme nous l'avons évoqué, la réalisation syntaxique de l'adverbial influe sur son placement, et malgré la tendance des adverbiaux circonstanciels à s'installer en fin de proposition, lorsqu'un adverbe est utilisé dans cette fonction, c'est la position médiane qui est favorisée (op. cit. : 807-808) :

> In all four registers, adverbs as circumstance adverbials are more likely to be used in medial positions than any other grammatical forms. This trend is particularly notable for the written expository registers. […] [B]ecause single adverbs are the most common one-word adverbials, it is not surprising that medial adverbials are usually realized as adverbs.

Ainsi, les adverbiaux de manière, qui adoptent le plus souvent un placement final, sont réalisés par des adverbes lorsqu'ils interviennent en position médiane.

Ces résultats sont corroborés par l'étude de corpus menée antérieurement par Quirk et al. à l'aide du corpus *Survey of English Usage* sur la fréquence des trois placements. Nous reproduisons ci-dessous les résultats de cette étude concernant les adverbes de classe ouverte (*open-class adverbs*), cet ensemble restant relativement stable même si l'on prend en compte la nouvelle délimitation de la catégorie proposée par Pullum et Huddleston. Les résultats ne concernent que la partie écrite du corpus *SEU* ; la colonne "Total" indique le nombre d'adverbiaux réalisés par ce type d'adverbe relevés dans le corpus, tandis que les colonnes "*I*" à "*E*" donnent le pourcentage d'adverbes présents dans ces positions (op. cit. : 501) :

	Total	***I***	***iM***	***M***	***m/eM***	***iE***	***E***
Adverbes, classe ouverte	462	15 %	1,5 %	47 %	1,5 %	7 %	28 %

Tableau 20. Fréquences de placement des adverbes dans l'étude de ACG

Le placement le plus fréquent pour les adverbes en fonction d'adverbial est donc le placement médian en M, c'est-à-dire entre le sujet et le verbe lexical si aucun auxiliaire n'est présent, ou bien juste après le premier auxiliaire, y compris BE et HAVE même s'ils sont les seuls verbes du GV.

Synthèse du placement des adverbes selon leur catégorie sémantique

Le tableau 21 présente les différentes positions typiques pour les adverbes en fonction d'adjoint/adverbial. Les catégories sémantiques sont celles identifiées par Pullum et Huddleston, puisqu'elles sont spécifiques aux adverbes dans ce type de fonction, contrairement à celles qui sont présentées dans *ACG* et *LGSWE* (voir la sous-section "Les types sémantiques d'adverbe en fonction d'adjoint/adverbial"). Nous empruntons la description des placements à Quirk et al. en raison de leur degré de précision, avec les ajustements suivants :

- les quatre sous-types de M sont réduits à deux seulement : les placements en *M* (position immédiatement après le sujet et, le cas échéant, le premier auxiliaire) et *eM* (position immédiatement avant le verbe lexical), sont assimilés (symbole *e/M*), tous les placements immédiatement avant le verbe lexical étant regroupés, et le placement en *mM* (position immédiatement avant le dernier auxiliaire dans des GV en contenant trois ou plus) est laissé de côté en raison de sa rareté ;
- les placements après BE et HAVE, lorsque ceux-ci sont les seuls verbes du GV, ne sont pas considérés comme médians mais finaux ;
- le placement en *iE* (position avant un élément obligatoire en fin de proposition, directement après le verbe ou bien après le verbe et un autre élément obligatoire) y est inclus mais correspond uniquement au placement des adverbes entre le verbe et un dépendant non-obligatoire.

Les jugements sont fondés sur les analyses de Pullum et Huddleston (op. cit. : 579-580), complétées par les informations supplémentaires offertes par *ACG* et *LGSWE* concernant la fréquence des placements.

Le symbole [+] indique l'acceptabilité d'un placement, et le symbole [-] son inacceptabilité. Le symbole [?] est utilisé par Pullum et Huddleston dans le cas des placements qui peuvent être acceptables dans certains contextes sémantiques et syntaxiques, ou bien lorsque le type d'adverbe est le plus souvent détaché, ce qui accorde une grande souplesse à son placement. La colonne de droite donne des informations supplémentaires concernant le placement privilégié lorsque plusieurs sont possibles, et la nécessité éventuelle pour l'adverbe de faire l'objet d'un détachement prosodique (DP) dans certaines ou toutes ses positions.

Malgré le fait que les trois ouvrages n'utilisent ni les mêmes termes ni exactement le même découpage des catégories sémantiques, leurs analyses des placements typiques des adverbes dans la proposition convergent la plupart du temps. *ACG* est l'ouvrage qui utilise le cadre le plus éloigné et donne les généralisations les moins précises, nous ferons donc référence en priorité à *CGE* et *LGSWE*.

Comme Pullum et Huddleston l'ont explicité, la distinction entre les types d'adjoint portant sur le GV et ceux portant sur la proposition en entier permet de faire apparaître des généralisations sur le placement. Le placement médian en *e/M*, soit avant le verbe lexical, que ce soit immédiatement après le sujet ou après un auxiliaire, est le placement privilégié pour la moitié des catégories sémantiques d'adjoints à portée verbale, et constitue une alternative possible pour l'autre moitié des adjoints de ce type. Le placement en fin de proposition est également une préférence fréquente, alors qu'un placement à l'initiale, qui est rarement bienvenu, nécessite que le segment soit détaché de la proposition (Pullum et Huddleston, 2002 : 578).

À l'inverse, les catégories d'adjoints portant sur la proposition tendent à privilégier un placement à l'initiale, un placement proche du GV en *e/M* restant tout de même une alternative possible. Le tableau 21 en page suivante fait d'ailleurs apparaître que le placement médian, contrairement aux quatre autres possibilités, est jugé comme plus ou moins acceptable pour tous les types sémantiques des deux ensembles, ce qui s'explique par la relative neutralité de cette position en termes de portée.

Catégorie	I	iM	e/M	iE	E	Informations supplémentaires
Types d'adjoint portant sur le GV						
Manière	-	-	+	+	+	Préférence pour les placements *iE* et *E*.
Moyen ou instrument	-	-	+	+	+	Préférence pour les placements *iE* et *E*.
Lié à l'acte, év. subjective	+	+	+	-	-	Placement final possible seulement en cas de DP.
Lié à l'acte, év. intention	?	?	+	+	+	Préférence pour le placement *e/M*.
Degré	-	-	+	-	+	Grandes variations en fonction de l'adverbe. L'acceptabilité du placement peut dépendre du verbe.
Lieu temporel	+	?	+	?	+	Préférence pour le placement *e/M*.
Durée	-	-	+	+	+	Préférence pour le placement *e/M*.
Aspect	?	?	+	+	+	Préférence pour le placement *e/M*.
Fréquence	?	?	+	?	+	Préférence pour le placement *e/M*.
Chronologie	-	?	+	+	-	Doit précéder l'adjoint temporel auquel il est associé.
Types d'adjoint portant sur la proposition						
Domaine	+	?	?	?	+	Préférence pour le placement *I.* DP requis en positions *iM* et *e/M*.
Modalité	+	+	+	+	+	Préférence pour le placement *e/M*. Généralement accompagné de DP.
Évaluation	+	+	+	?	?	Généralement accompagné de DP.
Lié à l'acte de parole	+	?	?	?	?	Généralement accompagné de DP.
Connecteur	+	+	+	-	-	Préférence pour le placement I. Généralement accompagné de DP. Grandes variations en fonction de l'adverbe.
Modifieur focalisant						
Additif	?	?	+	?	?	Placement dépend du focus. Acceptabilité varie en fonction de l'adverbe.
Restrictif	-	?	+	?	?	Placement dépend du focus. Acceptabilité varie en fonction de l'adverbe.

Tableau 21. Synthèse du placement des adverbes selon leur catégorie sémantique

Les adjoints véhiculant une évaluation ou un avis sur un élément du message (*CGE* : domaine, modalité, évaluation, lié à l'acte de parole ; *LGSWE* : adverbiaux de positionnement) font preuve d'une grande mobilité (Biber et al., 1998 : 773). Celle-ci s'explique d'une part par leur portée globale, et d'autre part par le fait qu'ils sont généralement en fonction de supplément, et donc détachés.

Les liens entre catégorie sémantique et placement ne sont pas à sens unique. Ainsi, si la catégorie sémantique d'un adverbe a une influence sur le choix de son placement en phase de production du message, ce placement détermine en partie le sens qui est donné à l'adverbe en phase de réception. En effet, une proportion importante d'adverbes admet des significations secondaires obtenues par glissement métaphorique (Biber et al., 1998 : 552). Le cas le plus visible est celui des adverbes dénotant le degré, dont le sens est souvent dérivé d'une signification première de manière, comme dans l'exemple suivant (Pullum et Huddleston, 2002 : 583) (notons en passant que même le premier de ces usages résulte d'un glissement sémantique) :

[126] *They behaved dreadfully.*

[127] *I'm dreadfully sorry.*

Les adverbes dénotant en premier lieu la manière peuvent également être utilisés pour instancier des adjoints des catégories sémantiques "Lié à l'acte", "Évaluation" et "Lié à l'acte de parole", comme on peut le voir dans les exemples suivants (dans l'exemple [128], l'adverbe est en fonction de complément et non d'adjoint) :

[128] *They behaved foolishly.*

[129] *Foolishly, they didn't think to lock the back door.*

[130] *He looked up hopefully.*

[131] *Hopefully, we'll manage to finish on time.*

[132] *Can I give my opinion frankly?*

[133] *Frankly, I'm not interested.*

En complément de son contexte sémantique d'utilisation (sens du verbe et des autres éléments), la position de l'adverbe dans la proposition est dans ce cas déterminante pour l'interprétation du sens qui lui est donné, les placements qui produisent les deux types d'interprétation étant radicalement différents. Dans le cas d'un placement médian, une incertitude subsiste concernant l'interprétation du sens de l'adjoint :

[134] *She had carefully ticked all of the boxes in the application form.*

Dans cet exemple, l'adverbe peut tout aussi bien être interprété comme dénotant le soin qui est apporté à l'action que la prudence manifestée dans le fait de cocher toutes les cases.

Les modifieurs focalisants additifs et restrictifs ont été inclus dans le tableau, mais il est difficile, et sans doute moins pertinent, de généraliser leur placement de la même façon que pour les autres types sémantiques, puisque celui-ci dépend en grande partie du focus de l'adverbe et ce qui est admis par l'adverbe lui-même. Une généralisation concernant le placement médian est cependant possible, celui-ci étant d'une part accepté pour la quasi-totalité de ces modifieurs (Pullum et Huddleston, 2002 : 593 ; Quirk et al., 1985 : 605), et d'autre part le plus fréquent lorsque le modifieur porte sur le GV ou la proposition dans son ensemble. Il est ainsi observé huit fois plus souvent que les placements en début et en fin de proposition (Biber et al., 1999 : 802).

Certains adverbes focalisants, comme *also*, peuvent être placés avant comme après l'élément sur lequel porte leur focalisation, alors que d'autres, comme *just*, sont obligatoirement placés avant cet élément (Pullum et Huddleston, 2002 : 590, 593). Une description détaillée du placement des modifieurs focalisants impliquerait donc de faire la liste des placements possibles pour chaque adverbe pouvant être utilisé comme modifieur focalisant, ce qui n'est pas nécessairement souhaitable ici. Cependant, comme nous le verrons, l'observation des erreurs relevées dans le corpus nous oriente vers une présentation plus approfondie du placement de l'adverbe *also* (voir section 2.2.2b).

f. Création d'une ressource lexico-sémantique d'adverbes

Nous venons de voir que l'un des premiers facteurs à prendre en compte dans le placement des adverbes est leur type sémantique. Il est donc nécessaire d'avoir des indications sur le ou les types sémantiques d'un adverbe afin de détecter une éventuelle erreur de placement et d'en prédire la ou les corrections. Nous postulons que ces indications peuvent être fournies par une ressource lexicale dédiée aux adverbes. La réalisation de cette ressource est cependant rendue plus difficile par le fait que de nombreux adverbes sont polysémiques à la suite de glissements métaphoriques successifs. Certains adverbes peuvent avoir jusqu'à trois emplois différents ; c'est par exemple le cas de l'adverbe *naturally*, qui peut avoir une emploi de manière, de moyen ou d'évaluation, selon les types issus de la classification de *CGE* présentée ci-dessus. Les exemples suivants illustrent cette polysémie :

[135] *She was speaking very naturally despite her fear.* (manière)

[136] *The child was not conceived naturally, but through IVF.* (moyen)

[137] *Naturally, you are free to leave whenever you want to.* (évaluation)

Il serait également intéressant d'avoir des indications concernant la fréquence d'utilisation des différents sens de l'adverbe concerné, afin de pouvoir affiner les règles de détection et de correction si nécessaire.

À notre connaissance, il n'existe pas de ressource lexico-sémantique spécifiquement dédiée aux adverbes anglais. Les adverbes sont bien entendu représentés dans d'autres ressources de ce type, à commencer par les dictionnaires et thésaurus, et notamment dans la ressource *WordNet*, que nous évoquons plus en détail dans les paragraphes suivants. Il nous semble pourtant qu'une telle ressource est importante dans le cadre de la détection et de la correction des erreurs liées aux adverbes, s'il l'on souhaite pouvoir identifier les différents types d'adverbes et proposer des règles adaptées à chaque type. Nous proposons ici une liste d'indications à intégrer dans cette ressource, ainsi qu'une méthode pour sa constitution. Nous avons utilisé cette méthode pour la création d'une ressource pilote composée d'une centaine d'adverbes parmi les plus courants.

Nous avons identifié quatre types d'indications à fournir pour chaque adverbe dans cette ressource :

- les différents types sémantiques de l'adverbe,
- la portée de l'adverbe en fonction du type sémantique,
- le type morphologique de l'adverbe (adjectif + *-ly*, composé, autre),
- la scalarité de l'adverbe en fonction de son type sémantique.

L'indication des différents types sémantiques de l'adverbe constitue bien entendu l'intérêt principal d'une telle ressource. Il s'agit également de la principale difficulté rencontrée lors de sa constitution, que nous proposons de surmonter grâce à la méthode et aux sources présentées ci-dessous. Nous avons choisi d'inclure également un rang de fréquence, de l'usage le plus fréquent au moins fréquent. Avant de donner plus d'informations sur ces quatre critères, il nous faut évoquer les sources utilisées pour la création de cette ressource. Les trois sources utilisées pour l'identification des types sémantiques et de leur fréquence sont *WordNet*, *Wiktionary* et *CGE*.

WordNet (Princeton University, 2010) est une base de données du lexique anglais, créée en 1985 à l'Université de Princeton. Cette ressource est en perpétuelle évolution et de nouvelles versions sont créées régulièrement (la version accessible en ligne en mai 2014 est la version 3.1). Sa principale particularité est l'organisation des mots du lexique en *synsets*, ou groupements de synonymes exprimant un concept défini. Les *synsets* sont ensuite liés entre

eux par des relations sémantico-conceptuelles, créant ainsi un lexique sous forme de réseau. *WordNet* est accessible en ligne et disponible gratuitement en téléchargement. Cette ressource dispose d'un outil de recherche par mot clé ; les informations données pour un mot incluent l'étiquette de partie du discours, les *synsets* qui sont associés au mot ainsi qu'une paraphrase du concept sous-jacent, parfois complétée par un exemple d'utilisation en contexte, et enfin la fréquence de chaque sens, indiquée par un nombre entre parenthèses. Voici par exemple les informations associées à l'adverbe *naturally* :

- (22)S: (adv) **naturally**, of course, course (as might be expected) *"naturally, the lawyer sent us a huge bill"*
- (7)S: (adv) **naturally** (according to nature; by natural means; without artificial help) *"naturally grown flowers"*
- (2)S: (adv) **naturally**, by nature (through inherent nature) *"he was naturally lazy"*
- (1)S: (adv) **naturally** (in a natural or normal manner) *"speak naturally and easily"*

Tableau 22. Présentation de l'entrée lexicale pour l'adverbe naturally *dans WordNet 3.1*

Nous avons choisi d'utiliser la ressource *WordNet* en particulier en raison de la présence d'informations concernant la fréquence d'utilisation des différents sens. Ceux-ci sont donnés du plus fréquent au moins fréquent, d'après les annotations sémantiques effectuées manuellement dans un corpus par les personnes ayant créé et développé *WordNet*. Lorsqu'un sens n'est pas repéré dans le corpus, aucune indication de fréquence n'est donnée, et si plusieurs sens sont concernés pour un mot, ceux-ci sont présentés dans un ordre aléatoire (Tengi, 1998 : 112).

Les informations recueillies dans *WordNet* sont complétées par des indications issues de *Wiktionary*, dictionnaire collaboratif en ligne pour l'anglais. Cette ressource inclut toutes les informations lexicales qu'on est en droit d'attendre d'un dictionnaire (étymologie et informations de morphologie, prononciation, définitions) et certaines autres moins courantes, comme des synonymes, des traductions des différents sens, et des indications quant à la scalarité du mot recherché. Cette dernière indication rend l'utilisation de *Wiktionary* particulièrement intéressante pour notre ressource, puisque nous avons souhaité y inclure le caractère scalaire ou non-scalaire des adverbes. La nature collaborative de *Wiktionary*, qui peut être vue comme une cause de méfiance quant au sérieux des informations données, est également gage de richesse et de dynamisme, la ressource ayant le potentiel d'évoluer aussi rapidement que les usages linguistiques. Croisées avec des ressources fiables telles que

WordNet et *CGE*, les informations tirées de *Wiktionary* se révèlent très utiles. Notons également que les informations y sont présentées de manière extrêmement concise, précise et claire, contrairement à d'autres dictionnaires en ligne dont le nom évoque un plus grand sérieux.

La dernière source utilisée pour la création du lexique d'adverbes n'est pas une ressource lexicale à proprement parler, puisque nous nous tournons simplement vers *CGE*, auquel nous empruntons les étiquettes des types sémantiques d'adverbes. Nous avons recours à cette grammaire en cas de contradictions entre nos deux ressources lexicales, pour compléter certaines informations morphologiques, ainsi que dans le choix des adverbes à inclure dans l'échantillon pilote.

Revenons à présent sur les critères inclus dans le lexique d'adverbes. Comme indiqué ci-dessus, les étiquettes des types sémantiques d'adverbes ainsi que les informations concernant la portée de l'adverbe pour chaque type sémantique identifié sont issues de *CGE*. Nous utilisons *WordNet* et *Wikitionary* pour déterminer quels sont les différents types sémantiques de l'adverbe concerné. Ceux-ci n'étant pas toujours directement transférables à la catégorisation de *CGE*, nous effectuons des recoupements. Les types sont indiqués dans une liste numérotée par ordre de fréquence dès que cela est possible, selon les indications données dans *WordNet*. Lorsque celles-ci ne permettent pas de trancher (ex. : nombre d'occurrences non données, très basses ou comparables), les types sont indiqués sous forme d'une liste non numérotée. Les informations morphologiques sont issues d'un croisement des indications de *Wiktionary*, de *CGE*, et de nos propres observations. L'indication de scalarité provient de *Wiktionary*, dont les informations sont parfois nuancées par nos observations, et est donnée pour chaque type sémantique identifié pour un adverbe donné. Le tableau 23 en page suivante est un extrait de dix adverbes dans la ressource.

Une première ressource pilote a été créée à partir d'un échantillon de 94 adverbes. L'échantillon a été choisi parmi les adverbes les plus courants, d'après les listes données dans *CGE* ainsi que dans des ressources en ligne (voir 3.2.1c). Nous avons également inclus des adverbes de types différents. La ressource pilote est disponible en entier en Annexe 2 sous forme de tableau.

Adverbe	Type sémantique	Portée	Scalaire	Type morpho.
actually	Modalité	Proposition	Non	Adj. *-ly*
also	Focalisant additif	Focus	Non	Autre (composé)
carefully	1. Manière 2. Lié à l'action (volition.)	GV Proposition	Oui Oui	Adj. *-ly*
legally	Moyen ou instrument Domaine	GV Proposition	Non Non	Adj.*-ly*
naturally	1. Évaluation 2. Moyen ou instrument 3. Manière	Proposition GV GV	Non Non Oui	Adj. *-ly*
nevertheless	Connecteur Concession	Proposition GV	Non Non	Autre (composé)
sadly	1. Évaluation 2. Manière	Proposition GV	Oui	Adj.*-ly*
simply	1. Restrictif, sensible au focus 2. Degré 3. Manière	Focus Adverbe ou adj. GV	Non Non Oui	Adj. *-ly*
thus	1. Connecteur 2. Manière	Proposition GV	Non Non	Autre (simple)
well	1. Manière 2. Degré 3. Modalité	GV GV Proposition	Oui Oui Oui	Autre (simple)

Tableau 23. Ressource lexicale sur les adverbes anglais : échantillon

Comme nous l'expliquons dans la sous-partie suivante, nous avons orienté nos recherches vers le traitement des adverbes de manière et de l'adverbe focalisant additif *also* pour le présent projet. L'utilisation de la présente ressource n'est donc pas indispensable pour cette partie du projet, mais sera importante si le traitement des erreurs de placement des adverbes est étendu à toutes les catégories sémantiques. Cependant, cette ressource a été jugée d'un grand intérêt pour la réalisation du projet *LELIE*, dédié à la création d'un "assistant intelligent pour l'analyse et la prévention des risques dans les processus industriels" (Site de présentation du projet, voir Bibliographie – Sites internet). La ressource a donc donné lieu à un développement complet par des membres de l'équipe de recherche constituée autour du projet *LELIE*, suivant la méthode et les critères définis par nos soins. Dans ce projet, l'identification du type sémantique des adverbes est utilisée afin de détecter la présence d'adverbes "flous" (*fuzzy*), c'est-à-dire sujets à des interprétations variées selon les personnes, et de proposer des corrections adaptées au type d'adverbe. La ressource a été augmentée afin d'intégrer des informations concernant le caractère flou ou non des adverbes, ainsi qu'une traduction en français, dans le but de rendre la ressource bilingue, puisque le projet *LELIE* pourra également être adapté au français à l'avenir (Kang et Saint-Dizier, 2014 : exemplaire non paginé).

La présentation de notre proposition pour une ressource lexicale sur les adverbes anglais clôt notre introduction à la syntaxe générale du placement des adverbes. Dans cette partie, nous avons présenté une synthèse des descriptions de la syntaxe des adverbes dans les trois grammaires de référence. On a pu observer la progression des analyses concernant la délimitation de cette catégorie, sa syntaxe interne et son comportement dans la proposition. Nous avons confronté les différents cadres descriptifs des trois grammaires de référence ayant trait aux différentes fonctions instanciées par les adverbes ainsi qu'à leur placement dans la proposition. Malgré des disparités parfois importantes entre ces cadres descriptifs, il a été possible de produire une synthèse exhaustive combinant les différents points forts des trois approches. Nous utilisons ainsi les types sémantiques identifiés par Pullum et Huddleston, cadre qui offre un équilibre entre simplicité et précision. Les types de placements utilisés pour notre synthèse ainsi que leurs étiquettes constituent une simplification des placements proposés par Quirk et al. Le recours à *LGSWE* nous permet d'asseoir cette synthèse sur des observations fiables concernant la fréquence des types d'adverbe et des types sémantiques d'adjoints, ainsi que sur les placements privilégiés de ces types. Cette synthèse a abouti à une modélisation générale du placement des adverbes en fonction d'adjoint, qui est présentée dans les paragraphes précédents sous forme de tableau. Elle constitue une base de travail indispensable, et nous a permis d'entrevoir et de formuler des généralisations sur le placement des adverbes.

Il ressort également de cette synthèse que le traitement des adverbes représente un défi pour la correction automatisée. Les adverbes en eux-mêmes constituent une catégorie "nébuleuse", dont les frontières ne font pas consensus. Nous avons abordé la proposition faite par les auteurs de CGE de redessiner les contours de la catégorie des adverbes, afin d'en exclure un certain nombre de termes devant être réinterprétés en tant que préposition ou pronom. Ces auteurs avancent des arguments tout à fait convaincants du point de vue de la linguistique pure ; cependant, on doit admettre que la recherche de la justesse linguistique de la délimitation de cette catégorie n'est pas absolument nécessaire dans le cadre de l'enseignement de l'anglais et de la correction automatisée, puisque certains des mots qui sont exclus de la catégorie des adverbes par *CGE* sont utilisés dans la même fonction que ceux-ci, et peuvent donc être sources des mêmes types d'erreur de placement, comme les termes pronominaux *yesterday* ou *here*. Nous verrons que l'utilisation de ces mots est incluse dans notre recensement des erreurs liées au placement des adverbes.

Par ailleurs, nous avons vu que les adverbes instancient non seulement des adjoints de sens très variés dans la proposition, mais aussi que leur placement a une influence sur l'interprétation de leur sens. Nous verrons dans la sous-partie 2.2.2 comment aborder cette question. En outre, la grande quantité de symboles [?] dans le tableau présenté plus haut montre que, malgré l'existence de tendances généralisables et l'identification de paramètres, des incertitudes subsistent concernant la grammaticalité de certains placements. Le placement d'un adverbe entre un verbe et son GN complément d'objet est un exemple de ces incertitudes : s'il est unanimement jugé comme plus aisément acceptable en cas de poids syntaxique et informationnel du GN objet, deux des trois grammaires ne le jugent pas explicitement comme agrammatical même si ces conditions ne sont pas présentes. Une lecture approfondie de ces ouvrages révèle cependant que si le placement d'un adverbe entre un verbe et un objet n'est pas rejeté, il n'est présenté dans aucun des exemples utilisés dans les parties consacrées aux adverbes et aux adverbiaux (dans le sens de *LGSWE* et *ACG*).

Le traitement du placement des adverbes adjoints pose également problème en raison de la fenêtre large qui est concernée dans la proposition, qui découle de leur grande flexibilité de placement. En effet, détecter un placement erroné et prédire un placement grammatical à l'aide de patrons de reconnaissance et d'instructions pour la réécriture implique de pouvoir reconnaître des groupes verbaux avec auxiliaires, des groupes nominaux et des groupes prépositionnels. Ces difficultés ne sont bien entendu pas insurmontables, et nous les abordons dans la sous-partie 2.2.2 ainsi que dans le Chapitre 3.

La modélisation que nous avons présentée est ainsi loin d'être assez approfondie pour permettre un passage direct à la création de règles de correction. Procéder de la sorte serait d'ailleurs peu efficace, puisque le placement des adverbes ne pose pas problème dans toutes ses configurations. Comme nous l'avons expliqué dans la section 2.1.1, nous procédons par étapes successives afin de recueillir les informations nécessaires sur ce type d'erreur et sa correction. La partie suivante propose d'approfondir notre analyse des erreurs concernées, ainsi que de formuler des prévisions informées par des recherches sur l'utilisation des adverbes par des francophones.

2.2.2 Vers la détection et la correction des erreurs liées au placement des adverbes dans la proposition

L'organisation de cette sous-partie reflète le fonctionnement de nos règles de traitement automatique. Comme nous l'avons déjà vu, celles-ci incluent une phase de détection et une

phase de correction. Pour être en mesure de détecter les erreurs, il est nécessaire dans un premier temps de pouvoir juger de la grammaticalité d'un segment, et dans un second temps de reconnaître les structures concernées. Obtenir des informations sur les placements erronés régulièrement employés par les personnes utilisatrices, ainsi que sur les configurations dans lesquelles ces placements se retrouvent le plus souvent, permet de limiter le champ sémantique et syntaxique de la détection de l'erreur.

Dans le cas du placement des adverbes, corriger une erreur revient à déplacer l'adverbe vers une position qui restaurera la grammaticalité de la phrase sans en altérer le sens. Cela implique de prédire les différentes positions disponibles pour l'adverbe concerné. Nous avons vu que le type sémantique de l'adverbe influence le choix de placement : il est par conséquent nécessaire de disposer d'informations sur le type sémantique de l'adverbe.

La première section de cette sous-partie est consacrée à l'étude des erreurs dans le placement des adverbes en général. Nous utilisons les résultats de travaux effectués dans le domaine de l'acquisition des langues secondes (ALS), ainsi que nos propres analyses des erreurs présentes dans notre corpus. Cette étude nous permet de formuler des généralisations sur les erreurs produites et sur les principaux types d'adverbes sur lesquels ces erreurs portent. Les deux autres sections, 2.2.2b et 2.2.2c, se focalisent ainsi sur le traitement des erreurs dans le placement de l'adverbe *also* et des adverbes de manière respectivement. À la fin de ces sections, nous présentons une modélisation complète de la correction de ces erreurs, permettant ainsi un passage facilité à l'implémentation.

a. Description des erreurs de placement des adverbes

L'acquisition du placement des adverbes

Dans un article consacré à l'exploration de certaines erreurs de personnes apprenantes francophones dans le placement des adjoints en anglais, Gledhill (2005 : 92) explique de la façon suivante la présence de ce type d'erreur dans leurs productions, dont nous avons observé la relative fréquence dans notre propre corpus :

> There are several reasons why adjunct placement may pose such a problem for learners. Firstly adjunct placement is not widely taught and so there is virtually no awareness of the problem. Secondly French can very freely place adjuncts in post-verbal position, and this happens to be one of the most identifiable differences between English and French word order.

Gledhill identifie ainsi deux causes pour la présence de ces erreurs dans les productions de francophones : une lacune dans l'enseignement de ces structures, et une zone de friction entre l'anglais et le français.

Nous avons déjà mentionné le fait que les contraintes grammaticales pesant sur le placement des adverbes en français et en anglais ne sont pas identiques, et que leurs différences peuvent être cause de difficultés pour les personnes apprenantes et utilisatrices de l'anglais. En particulier, les deux langues acceptent chacune un placement qui est agrammatical dans l'autre : les adverbes français peuvent intervenir entre un verbe et son objet (SVAO), alors que les adverbes anglais peuvent être placés immédiatement après le sujet (SAV) (White, 1990 : 339). Les exemples suivants illustrent ces configurations :

[138] *Bruno déballa soigneusement son cadeau.*

[139] **Bruno unwrapped carefully his present.*

[140] **Bruno soigneusement déballa son cadeau.*

[141] *Bruno carefully unwrapped his present.*

Par contre, les adverbes peuvent être placés en début comme en fin de proposition dans les deux langues, avec ou sans détachement prosodique, ainsi qu'entre un auxiliaire et un verbe lexical :

[142] *Soigneusement, Bruno déballa son cadeau.*

[143] *Carefully, Bruno unwrapped his present.*

[144] *Bruno déballa son cadeau soigneusement.*

[145] *Bruno unwrapped his present carefully.*

[146] *Bruno a soigneusement déballé son cadeau.*

[147] *Bruno is carefully unwrapping his present.*

Dans le cadre théorique de la grammaire universelle, cette distinction quant à l'acceptabilité des placements SAV et SVAO est considérée comme le résultat de l'existence de paramètres différents concernant la "montée" du verbe dans la proposition (*verb raising*) : les paramètres grammaticaux du français exigeraient que les verbes montent dans la structure syntaxique de la proposition, alors que ceux de l'anglais interdiraient cette montée (White, 1991 : 135-137 ; Ayoun, 2005 : 36). Ces hypothèses ont suscité une série d'études datant des années 1990 sur le placement des adverbes anglais par des francophones dans le domaine de

l'acquisition des langues secondes, et utilisant le cadre théorique chomskyen (White, 1990 ; White, 1991 ; Trahey et White, 1993 ; Trahey, 1996). L'objectif principal de ces études était de découvrir si les personnes apprenantes francophones étaient en mesure d'acquérir un paramétrage différent, c'est-à-dire d'une part de ne pas utiliser le placement SVAO et d'en reconnaître l'agrammaticalité, et d'autre part d'utiliser le placement SAV à la place.

Ces études se sont également penchées sur les moyens à utiliser pour faciliter l'acquisition du paramètre, et plus précisément sur l'efficacité de la présentation d'indications positives, comme la confrontation à des exemples de phrases bien formées en anglais, et d'indications négatives, c'est-à-dire le retour correctif sur les erreurs et des informations sur les structures erronées. Les recherches sur l'acquisition du placement des adverbes se sont également poursuivies dans les années 2000, notamment avec l'étude d'Ayoun (2005).

On remarque que les auteurs de ces recherches n'hésitent pas à qualifier d'agrammatical le placement d'un adverbe entre un verbe et son objet, contrairement aux auteurs de *ACG* et *LGSWE*. Par ailleurs, les auteurs de ces recherches manquent de noter que ce placement peut être rendu grammatical et même optimal face à des contraintes plus fortes, comme la longueur et le poids informationnel du GN objet, contraintes que nous avons déjà évoquées. Nous reviendrons sur ce point lors de notre exposé de l'étude d'Osborne (2008) dans les pages suivantes.

Bien que les questions de recherche et le cadre théorique de ces études ne soient pas directement en lien avec nos recherches, leurs résultats présentent néanmoins un intérêt pour ces dernières. Les recherches de White et Trahey (White, 1990 ; White, 1991 ; Trahey et White, 1993 ; Trahey, 1996) portent sur l'acquisition du placement des adverbes en anglais par des groupes d'élèves francophones âgés de dix à douze ans selon l'étude concernée. Ces quatre études s'intéressent en partie à l'efficacité des méthodes employées pour amener les sujets à cette acquisition. L'étude de White (1990) s'intéresse à la possibilité pour les élèves d'acquérir le placement des adverbes par le biais de l'enseignement de la formation des questions, qui concerne également le paramètre de montée du verbe. Dans l'étude de White (1991), les sujets reçoivent des indications grammaticales sur le placement des adverbes incluant des retours correctifs (preuves négatives), et sont confrontés à des phrases bien formées (preuves positives). Les résultats de cette étude peuvent être comparés à ceux de Trahey et White (1993), qui évaluent l'efficacité de l'utilisation de preuves positives uniquement. L'étude de Trahey (1996) tire des conclusions sur l'efficacité de cette méthode à long terme.

Les tests utilisés dans ces études comprennent généralement les tâches suivantes :

- une tâche de jugement de grammaticalité, pour laquelle les sujets doivent déplacer l'adverbe s'il est jugé mal placé,
- une tâche d'indication de préférence, pour laquelle les sujets doivent indiquer si des phrases sont correctes ou incorrectes,
- une tâche de manipulation de phrases, pour laquelle les sujets doivent former des phrases correctes avec des mots donnés sur des cartes.

Les sujets sont testés avant traitement, puis juste après le traitement, et enfin après un laps de temps plus ou moins long après le traitement, c'est-à-dire trois semaines (Trahey et White, 1993), cinq semaines (White, 1990 ; 1991) ou un an (Trahey, 1996). Les adverbes utilisés pour réaliser ces tests sont des adverbes de manière et des adverbes de fréquence, à parts égales.

Les conclusions de ces études concernent l'efficacité des méthodes utilisées pour favoriser l'acquisition de ce trait linguistique, ainsi que la durabilité de l'acquisition dans le temps. En plus de ces deux aspects, les chercheuses tirent des conclusions sur l'étendue de l'acquisition de ce trait en distinguant la reconnaissance du placement SAV comme étant grammatical de la reconnaissance du placement SVAO comme étant agrammatical, ce dernier point posant apparemment plus de difficultés.

Toutes les études concluent que les sujets utilisent le placement SVAO et le reconnaissent comme grammatical avant traitement, ce qui, d'une part, confirme la difficulté de l'acquisition de ce trait, et d'autre part semble indiquer la présence de transfert linguistique de la L1 à la L2.

White (1990 : 356-358) conclut que seule l'instruction directe sur les adverbes, et non sur une problématique connexe comme la formation des questions, mène à l'acquisition de ce trait. Cette acquisition est manifestée par l'utilisation et la reconnaissance de la grammaticalité du placement SAV, et la reconnaissance de l'agrammaticalité de SVAO. Cependant les résultats sont limités aux structures précises auxquelles les élèves ont été confrontés lors du traitement.

White (1991 : 158-160) indique que l'utilisation de preuves positives n'est pas suffisante pour favoriser la reconnaissance de l'agrammaticalité du placement SVAO. Par ailleurs, il semblerait qu'aucun des deux types de preuves ne favorisent l'acquisition à long terme, même si cela peut être expliqué dans cette étude par la rareté de l'utilisation des adverbes en classe et

par l'arrêt de l'attention aux erreurs dans le placement des adverbes une fois le traitement spécifique terminé. Les résultats de cette étude suggèrent également que les placements SAV et SVAO peuvent coexister dans l'interlangue d'un même sujet.

Ces résultats sont corroborés par l'étude de Trahey et White (1993), qui montre que la présentation intensive de phrases comportant des adverbes bien placés permet d'augmenter l'incidence du placement SAV mais pas de faire baisser celle du placement SVAO. Les résultats de Trahey (1996 : 133-136) indiquent que la présentation de preuves positives n'a pas non plus d'effets apparaissant à long terme, même si les résultats obtenus après le traitement sont maintenus. Concernant l'inefficacité de l'apport de preuves positives comme négatives sur l'acquisition à long terme, Trahey reprend l'hypothèse selon laquelle il est nécessaire d'inclure l'instruction grammaticale dans un contexte authentique. Précisons que cette approche est désormais largement répandue, et constitue même la base du cadre conceptuel du *Cadre Européen de Référence pour les Langues* (Conseil de l'Europe, 2001).

L'étude d'Ayoun, datant de 2005, diffère des recherches présentées ci-dessus en raison du choix de sujets et des résultats obtenus. Les sujets de l'étude sont de jeunes adultes francophones ayant un niveau intermédiaire à avancé en anglais (2005 : 35). Leur utilisation des adverbes est évaluée à travers une tâche de production écrite, une tâche de jugement de grammaticalité, et une tâche d'indication de préférence (2005 : 43). Aucune instruction ou présentation concernant le placement des adverbes n'est effectuée. Les résultats obtenus montrent une maîtrise du placement des adverbes anglais, ce qui indique la capacité d'acquérir le paramétrage de la L2 concernant le mouvement de verbe (2005 : 65). Ces résultats différents peuvent s'expliquer par l'âge des sujets et leurs compétences plus avancées et plus sophistiquées en anglais, ce qui indiquerait que l'acquisition du paramétrage se développe dans le temps (2005 : 65).

Les travaux que nous venons de présenter concernent l'étude du placement des adverbes anglais par des francophones en tant que confirmation ou infirmation des théories de la grammaire universelle sur le fonctionnement de l'acquisition des langues secondes. La question du placement des adverbes anglais a aussi été envisagée du point de vue de l'influence translinguistique, dans le but d'identifier les types de configurations produites par des personnes apprenantes de L1 diverses, les causes possibles de ces erreurs, et la présence éventuelle de transfert L1-L2.

Contrairement aux travaux présentés ci-dessus, qui font appel à des tâches de production et de jugement, les travaux d'Osborne (2008) reposent sur une étude de corpus. Ce dernier

compare le placement des adverbes dans les productions écrites de personnes ayant l'anglais comme langue maternelle, et celles de personnes utilisatrices de l'anglais à un niveau "post-intermédiaire", identifié comme allant de B2 à C2 (cf. *CECRL*). Les corpus d'anglais natif utilisés sont *Essay Bank* de l'Université de Savoie (anglais britannique), et *LOCNESS* du CECL de l'Université de Louvain (anglais britannique et américain). Les corpus d'interlangue sont *ICLE* (v. 1), qui inclut des productions de personnes de onze L1 différentes, et le *Chambéry Corpus*, composé de dissertations d'étudiantes et étudiants francophones. Notre compte-rendu de cette recherche est centré sur les résultats obtenus à partir des corpus francophones.

Pour commencer, les comparaisons de corpus qui sont présentées dans l'étude d'Osborne indiquent que les personnes apprenantes et utilisatrices de l'anglais à un niveau intermédiaire à avancé emploient des placements qui ne sont pas canoniques en anglais natif. Osborne classe les types de placement en quatre catégories standards : à l'initiale, avant le verbe (SAV), entre le verbe et l'objet (SVAO) et en fin de proposition. Il dénombre jusqu'à quatre fois plus de placements SVAO dans les productions de francophones que dans les productions natives (2008 : 134-135). Ces résultats semblent de prime abord entrer en contradiction avec ceux de Ayoun. On soulignera cependant que les études d'Ayoun et d'Osborne emploient des méthodes de recherche très différentes, et ont des perspectives inverses sur les résultats : Ayoun s'intéresse aux occurrences de placement correct, qui sont après tout majoritaires même dans l'étude d'Osborne, alors que ce dernier se concentre sur les placements non-canoniques. Étant donné le niveau d'anglais des sujets étudiés, il est logique que les utilisations correctes soient plus nombreuses que les utilisations erronées, quelle que ce soit la nature de la structure prise en compte.

La conclusion d'Osborne concernant l'utilisation du placement SVAO par des francophones s'accompagne d'une indication de la présence possible de transfert entre la L1 et la L2, puisque les personnes ayant le plus recours à ces placements sont locutrices natives du français, de l'espagnol et de l'italien, trois langues à montée du verbe et dans lesquelles ce placement est donc parfaitement grammatical (2008 : 142). Il semble ainsi utile de cibler précisément ce type de public lors de la création de règles de correction automatisée.

Par ailleurs, les travaux d'Osborne vont au-delà de la simple constatation chiffrée : ils présentent également une analyse des raisons de l'utilisation du placement SVAO en anglais natif et non-natif, ainsi qu'une comparaison de la prise en compte par ces deux groupes des paramètres qui le régissent. L'étude de corpus confirme que le placement SVAO, s'il n'est pas

canonique, est présent en anglais natif (2008 : 133), et est déclenché par deux paramètres : la longueur du GN complément d'objet du verbe, et l'existence d'un lien étroit entre verbe et adverbe, qu'Osborne décrit comme une collocation (2008 : 138).

La longueur et la "lourdeur" d'un GN ne sont pas des caractéristiques facilement quantifiables. Osborne utilise les paramètre du nombre de mots et de la complexité linguistique, les GN longs dépassant six mots et comportant généralement une proposition subordonnée ou d'autres types de compléments et modifieurs (2008 : 136). D'après les analyses sur des corpus d'anglais natif menées par Osborne, 48 % des cas de placement de type SVAO sont liés à la présence d'un GN objet de plus de six mots. Dans les corpus de francophones, cette configuration concerne seulement 22 % à 23 % des placements SVAO selon les corpus, alors que 39 % à 43 % de ces placements ont lieu avec des GN comportant de un à trois mots (2008 : 137).

Les cas de collocations mentionnés par Osborne concernent des associations telles que *take seriously, handle carefully, judge fairly, put clearly,* qui diffèreraient des autres combinaisons verbe+adverbe du fait que l'adverbe ne pourrait être placé avant le verbe (2008 : 138), contrairement à des expressions telles que *damage seriously* par exemple (phrases modifiées à partir des exemples d'Osborne) :

[148] *It would seriously damage the independent role of the National Rivers Authority.*

[149] **The University seriously takes its responsibility.*

D'après nous, les expressions citées par Osborne correspondent à des cas plus ou moins centraux d'adverbes fonctionnant non comme adjoint mais comme complément du verbe, en particulier dans le cas de *take seriously*, sur lequel Osborne s'étend le plus. Cette analyse est renforcée par le changement de sens induit par la suppression de l'adverbe :

[150] *?The University takes its responsibility.*

Le fait que ces adverbes soient en réalité des compléments du verbe expliquerait d'une part qu'ils ne puissent aisément être déplacés à la gauche du verbe, et d'autre part qu'ils entrent en compétition avec le GN objet pour un placement à proximité de celui-ci. Elle justifierait également l'impression exprimée par Osborne que, dans ces combinaisons, verbe et adverbe sont "liés l'un à l'autre d'une certaine manière" (2008 : 138 ; notre traduction), la co-dépendance des éléments étant une des principales caractéristiques qui distinguent la relation de complémentation de celle de modification. Osborne suggère que ce lien syntaxique et/ou

sémantique joue un rôle de facilitation dans le placement des adverbes à la suite du verbe en cas de présence d'un GN objet long.

D'après Osborne, l'analyse de 517 combinaisons [Verbe + Adv] de type SVAO repérées dans les corpus d'interlangue révèle une grande irrégularité, mais également deux configurations récurrentes : l'utilisation d'un verbe de cognition/perception avec un adverbe de manière ou de degré, et l'utilisation d'un adverbe pour préciser un verbe au sens générique, stratégie qui semble pallier un déficit lexical (2008 : 139-140). Les phrases suivantes illustrent ces deux configurations :

[151] **Why not trying to speak with each other and to explain clearly each other's opinions.*

[152] **The language affects badly the learning process.*

L'utilisation d'adverbes focalisant additifs en SVAO, et particulièrement *also*, est également remarquée par Osborne parmi les configurations récurrentes. Nous y reviendrons.

Ces analyses amènent Osborne à conclure qu'il n'est pas nécessairement pertinent de comparer le pourcentage d'utilisation du placement SVAO en anglais natif et en anglais L2, puisque les personnes utilisatrices et apprenantes tendent à ne pas suivre les paramètres justifiant ce placement non-canonique en anglais natif : "In other words, even when, quantitatively, learners do not produce more V-Adv-O sequences than natives, those that they do produce are likely to appear strange because they have non target-like characteristics" (2008 : 137).

Le passage en revue de travaux concernant l'acquisition du placement des adverbes anglais nous a permis de confirmer certaines intuitions et d'obtenir des informations précises sur les erreurs produites. Ces études indiquent avant tout l'intérêt porté à cette problématique du point de vue de l'acquisition des langues secondes, et laissent entrevoir un besoin en termes de traitement automatisé. Le placement en SVAO y est unanimement considéré comme non-canonique, et semble concentrer la majorité des erreurs produites par les francophones, à des niveaux intermédiaires ou avancés, enfants ou jeunes adultes. Ceci est imputable aux effets d'une influence translinguistique du français vers l'anglais. L'étude d'Osborne ne le mentionne pas, mais Ellis, citant Ringbom (2007), rappelle que le transfert est le produit des différences et des similarités des langues en présence : il est ainsi plus difficile d'acquérir un trait de la L2 différent mais *proche* d'un trait de la L1 qu'un trait complètement étranger à la L1 (Ellis, 2008 : 398-399). C'est le cas pour le placement des adverbes en français et en anglais, comme

nous l'avons vu dans les premiers paragraphes de cette sous-section. Cela semble expliquer que l'acquisition du placement des adverbes soit facilitée par l'instruction, mais aussi que la remédiation au placement SVAO soit difficile, et aux effets peu durables. Nous reviendrons sur cette question dans le Chapitre 3, lors de notre évaluation de l'utilité de l'accompagnement des corrections automatiques par des messages de retour correctif.

Nous sommes également mieux informés concernant les paramètres phraséologiques, syntaxiques et sémantiques de l'acceptabilité des placements SVAO, qui sont possibles en anglais natif lorsque le GN objet est long, c'est-à-dire supérieur à six mots d'après Osborne, et qu'il existe un lien de collocation ou de complémentation entre le verbe et l'adverbe, ce qui touche peu de combinaisons différentes. La sous-section suivante consiste en une analyse des erreurs relevées dans notre corpus.

Les erreurs de placement d'adverbes dans le corpus étudié

En raison de la taille de notre corpus et du nombre réduit de segments erronés relevés, les analyses présentées ici sont données à titre indicatif, et ne peuvent être utilisées comme base pour la formulation d'hypothèses concernant l'acquisition du placement des adverbes. Elles peuvent par contre corroborer certaines des hypothèses avancées dans la sous-section précédente. De plus, comme nous allons le voir, observer ces segments de près permet d'affiner la modélisation des erreurs et du placement des adverbes en général.

Le tableau suivant est un rappel de la distribution des erreurs de placement des adverbes dans notre corpus ; la colonne de gauche indique le pourcentage de ces erreurs par rapport au total des erreurs relevées pour le sous-corpus concerné :

Sous-corpus	Occurr.	% Total
Publications	13	6,2
ICLE	16	11,7
Courriels	1	0,4
Rapport	6	13,7
Total erreurs adv.	**36**	**5,4**

Tableau 24. Distribution des erreurs de placement des adverbes

Les segments relevés ont été classés dans la catégorie "Groupe verbal – Placement des modifieurs après le verbe – Adverbes", ce qui appelle quelques commentaires. Nous avons vu que les adverbes en fonction de modifieur relèvent du GV *ou* de la proposition. Nous n'avons

pourtant pas souhaité faire apparaître cette distinction, afin de conserver une unité dans les types d'erreurs, mais aussi parce que les erreurs de placement d'adverbes concernent le plus souvent un placement SVAO maladroit, qui intervient donc dans le GV. Par ailleurs, cette catégorie n'inclut pas les erreurs liées à la négation ou au lexique, qui ont leur propre catégorie. Enfin, on remarquera l'inclusion de segments concernant des termes qui ne sont pas étiquetés comme des adverbes, mais comme des prépositions selon la redéfinition de cette catégorie proposée dans *CGE* (ex. : *then*, *now*, *here*) ; ces segments présentant néanmoins les mêmes caractéristiques que ceux qui incluent des adverbes prototypiques, et ces termes étant sans doute envisagés de manière identique par les personnes utilisatrices, nous n'avons pas jugé utile de respecter cette distinction strictement.

Le tableau descriptif contenant tous les segments relatifs aux erreurs de placement des adverbes dans le corpus est disponible en Annexe 3. Les critères retenus pour le classement et l'analyse des segments sont les suivants :

- statut du segment : agrammatical ou inacceptable,
- type sémantique de l'adverbe concerné,
- type de placement : SVAO ou autre,
- correction(s) à apporter : mouvement (Mvt) vers un autre placement possible, ou autre correction (ex. : modification plus importante de la phrase, ajout de ponctuation pour indiquer un détachement prosodique).

Les jugements quant au statut du segment ont l'inconvénient d'être subjectifs, mais sont appuyés par les indications relevées concernant les possibilités de placement des adverbes. En particulier, nous avons vu que les auteurs de *ACG* et *LGSWE* étaient réticents à déclarer le placement en SVAO comme agrammatical, celui-ci pouvant se produire dans des énoncés en anglais L1. Les résultats de l'étude d'Osborne concernant le recours à ce placement semblent justifier cette réticence. En dépit de cela, les autres travaux que nous avons pu consulter à ce sujet, à commencer par *CGE*, relèvent l'agrammaticalité des placements SVAO, ou du moins leur très faible fréquence dans des corpus d'anglais L1. Comme nous l'avons vu, leur emploi est alors lié à des contraintes lexicales ou d'équilibre du GV. Dans *A Semantic Approach to English Grammar*, Dixon mentionne clairement la tendance de l'anglais et de ses locutrices et locuteurs à éviter ce placement (2005 : 386).

Pour l'ouvrage *Adverbial Positions in English* (1964), Jacobson a analysé manuellement le placement des adjoints relevés dans un corpus de 66 romans et autres ouvrages en anglais L1,

totalisant environ 3,5 millions de mots. Sur un total de 9207 adverbes fonctionnant comme adjoint, seulement 44 sont placés entre un verbe et un GN objet, soit 0,48 % des occurrences (1964 : 80). Jacobson ajoute également que dans presque tous les cas, l'objet est beaucoup plus "lourd" que l'adverbe (142). En raison de son infréquence et des déclarations des auteurs de grammaire à ce sujet, le placement SVAO est considéré comme non-standard dans le cadre de notre étude.

En dépit des frontières floues entre grammaticalité et acceptabilité, cette distinction est introduite dans le relevé des erreurs. Les phrases suivantes sont un exemple d'un segment jugé agrammatical, puis d'un segment jugé inacceptable :

[153] **Our system is able to derive automatically information for a large number of verbs.*

[154] *?The similarity between words depends then on the amount of normalized contexts they share.*

Les segments agrammaticaux sont principalement les segments présentant un placement SVAO non justifié par la présence d'un GN objet long (dans l'exemple [153], le GP *for a large number of verbs* est un dépendant du verbe et non du nom *information*). Les segments considérés comme inacceptables ne contiennent pas de structure allant à l'encontre des principes grammaticaux, mais apparaissent néanmoins comme maladroits et improbables.

Sur un ensemble de 36 erreurs relevées, 8 sont jugées inacceptables. Le jugement d'inacceptabilité concerne dans la moitié des cas des segments dans lesquels le verbe est complété par une proposition infinitive ou un GP, comme dans l'exemple ci-dessus ; ces segments, s'ils ne sont pas clairement agrammaticaux, sont suffisamment peu fréquents pour être considérés comme non-standards. Deux autres cas relèvent du placement inhabituel d'un adverbe connecteur sans détachement prosodique, comme dans le segment suivant :

[155] *?but exhibit <u>nevertheless</u> the dependency relationships observed in the source parse tree.*

Les types de placement déclenchant un jugement d'agrammaticalité ou d'inacceptabilité sont majoritairement les placements SVAO, qui concernent 26 erreurs sur 36. Le placement SVAO est ici défini de façon stricte, incluant uniquement un verbe lexical, simple ou suivi d'une préposition (*phrasal verb*), un adverbe, et un GN complément d'objet, incluant parfois des dépendants. Selon les critères d'Osborne (2008 : 136), 10 des GN objets sont courts (de 1 à 3 mots), 13 sont de longueur moyenne (4 à 6 mots), et seulement 3 sont longs (plus de 6 mots). Dans le cas de la présence d'un GN long, les segments sont jugés comme inacceptables

car le placement SVAO n'est pas déclenché par une collocation facilement idientifiable, ou faisant partie des collocations relevées dans les travaux respectifs d'Osborne, Gledhill et Dixon. Par ailleurs, des placements alternatifs sont disponibles dans chaque cas.

Le tableau suivant présente une synthétisation des différents schémas de placements observés dans le corpus d'erreurs, illustrés d'exemples, et accompagnés du nombre de segments concernés dans notre corpus (nous n'incluons pas les placements liés aux adverbes du comparatif) :

Schémas	**Exemple**	**Occ.**
(Aux.) + Verbe + ADV + GN	[156] *"Europe 92" won't change completely the life of its citizens.*	26
(Aux.) + Verbe + ADV + Prop. infinitive	[157] *The value of N needs also to be experimentally elaborated.*	4
(Aux.) + Verbe + Prep. + ADV + GN	[158] *in order to hang down exclusively family memories*	1
(Aux.) + Verbe + ADV + GP	[159] *[It] depends then on the amount of normalized contexts they share.*	1
Aux. + Aux. + ADV + Verbe + GP/GN	[160] *It can be also considered as restricted in terms of variant.*	1
(Aux.) + Verbe + ADV + ADV + GN	[161] *favorizing then perhaps more easily the student's future professional insertion*	1

Tableau 25. Schémas des erreurs de placement d'adverbes dans le corpus

Notons que dans le cas des schémas recueillant peu d'occurrences, le schéma concerné n'est pas généralisable à l'ensemble des adverbes ou à l'ensemble des configurations possibles. Par exemple, nous verrons dans la section suivante que l'adverbe *also* n'est pas généralement placé avant le verbe lexical lorsque deux auxiliaires sont présents, mais cette configuration est possible pour les adverbes de manière, comme le montre l'exemple suivant (Dixon, 2005 : 420), indiquant que le traitement doit être modifié selon le type d'adverbe :

[162] *John had been happily studying for the evaluation.*

Dans d'autres cas, même si une erreur est rendue visible par le placement de l'adverbe, agir sur celui-ci n'est pas nécessairement la correction la plus appropriée. Une correction optimale peut consister en une réécriture globale du segment. C'est par exemple le cas pour l'exemple [159], qui peut être corrigé par le remplacement de l'adverbe, dont l'utilisation semble être la conséquence d'un transfert de l'adverbe français "alors".

Cependant, dans la grande majorité des cas, la correction à apporter consiste à déplacer l'adverbe vers une position médiane (pour 17 occurrences) ou vers une position finale (pour 7

occurrences). Dans 5 cas sur 36, la correction à apporter est autre, et implique par exemple de modifier la phrase, ou bien de remplacer la combinaison [Verbe + Adv] par un verbe plus précis. Par ailleurs, 9 segments admettent plus d'une proposition de correction (classées par ordre de préférence dans le tableau en annexe), faisant généralement appel aux deux possibilités de placements courants (médian et final).

Les adverbes concernés par les erreurs relevées ici appartiennent à un ensemble restreint de types sémantiques. Les types sémantiques apparaissant plus d'une fois dans le corpus d'erreurs sont les suivants, présentés du plus fréquent au moins fréquent :

- focalisation additive (12),
- manière (6),
- degré (5),
- connecteur (4),
- lieu temporel (2),
- lieu (2).

Même si les erreurs les concernant sont incluses dans la liste donnée en annexe, nous ne prenons pas en compte les erreurs liées à l'expression *more and more*, puisque nous avons choisi de laisser de côté les utilisations comparatives des adverbes dans la présente étude.

Le type sémantique de focalisation additive concerne en réalité uniquement les utilisations de l'adverbe *also*, représentant le plus fréquent de cette catégorie (Biber et al., 1999 : 793). Les adverbes utilisés pour l'expression du degré ont parfois un sens premier de manière. Voir l'exemple suivant :

[163] **[They] were found to improve substantially the performance of either modality.*

Dans deux cas, l'adverbe *then* est analysé comme un connecteur plutôt que comme indiquant un lieu temporel. Les exemples suivants montrent ces deux emplois différents :

[164] *?The similarity between words depends then on the amount of normalized contexts they share.* (connecteur)

[165] **It had then its own evolution.* (lieu temporel)

Les types d'adverbe les plus fréquemment liés aux erreurs dans notre corpus sont donc les adverbes de manière, et l'adverbe focalisant additif *also*. Ce résultat est en cohérence avec les fréquences des types d'adverbes évaluées par Biber et al. dans *LGSWE,* et que nous avons

évoquées dans la section 2.2.1d (1999 : 787-788). Comme nous l'avons indiqué, le placement des adverbes dépend largement de leur sens. Cette information paraît donc essentielle à la création de règles de correction automatisée fondées sur des méthodes linguistiques. L'incorporation de la ressource lexicale présentée en 2.2.1f pourra être utilisée pour distinguer les différents types sémantiques d'adverbes ainsi qu'aborder les cas de polysémie. Notre objectif n'est cependant pas ici de développer un système complet de traitement des erreurs liées au placement des adverbes, mais de créer des règles de détection et de correction pouvant ensuite être amplifiées pour un traitement informatique du phénomène qui soit complet et robuste. Prenant cela en compte, il est judicieux de nous limiter au traitement d'un petit nombre d'adverbes pertinents. Les deux types d'adverbe les plus fréquents sont ainsi sélectionnés comme base pour l'élaboration des règles. Ces deux types, dont les traitements s'avèrent relativement distincts, sont à envisager comme des exemples introduisant une méthode : une fois la méthode élaborée et décrite, il est possible de transposer les règles à d'autres adverbes. Les deux types sélectionnés présentent d'ailleurs l'avantage d'appeler des traitements distincts : par exemple, si *also* est facilement identifiable et facilite des études de corpus ciblées, sa distribution est beaucoup plus variée que celle des adverbes de manière.

Le tableau ci-dessus, qui présente les schémas d'erreurs, constitue l'ébauche de la modélisation des patrons de détection et instructions de correction nécessaires pour le traitement des erreurs. L'observation globale des erreurs nous a permis en particulier de commencer à exclure certains schémas qui ne font pas partie des erreurs produites, et pour lesquels il n'est donc pas nécessaire de créer des règles de traitement. Par exemple, les placements abusifs à l'initiale ne semblent pas être une difficulté récurrente. Les deux sections suivantes, qui se concentrent uniquement sur le placement de l'adverbe *also* et des adverbes de manière en tant que modifieurs ou adjoints dans la proposition, font appel à d'autres sources linguistiques, qu'elles soient théoriques ou expérimentales, afin d'affiner la modélisation et permettre l'implémentation quasi-directe des règles.

b. *Modélisation du placement de l'adverbe* also

Dans la section 2.2.1e, nous avons présenté des généralisations pour le placement des adverbes en fonction de modifieur focalisant additif/restrictif, et indiqué que ce dernier dépendait d'une part du focus de l'adverbe, et d'autre part de l'adverbe lui-même. En raison de sa forte présence dans les erreurs observées, et de l'absence des autres types d'adverbes fonctionnant comme modifieur focalisant, nous nous concentrons ici sur le seul adverbe *also*.

Nous sommes alors en mesure de présenter ses tendances de placement en détail. Au-delà des lieux généraux auxquels l'adverbe est rencontré, et que nous présentons ci-dessous, il est également nécessaire de documenter les possibilités de placement d'*also* en rapport avec les auxilaires et les compléments du verbe autres que les GN.

Étude des positions adoptées par *also* dans des corpus d'anglais L1

CGE indique deux critères à prendre en compte lors de l'analyse du placement de ces adverbes par rapport à leur focus : leur capacité à être placés avant ou après l'élément focus, et à être placés à proximité ou à distance de celui-ci (op. cit. : 588). *Also* peut être rencontré dans toutes ces configurations, comme le montrent les exemples ci-dessous, dans lesquels le focus est entre crochets ; les exemples qui ne sont pas de notre propre création sont empruntés à *ACG* (op. cit. : 609) et *CGE* (op. cit. : 593) :

[166] *John has seen the fox <u>also</u> [near his back door].*

[167] *Sue <u>also</u> bought [a CD].*

[168] *We plan to visit [Paris] <u>also</u>.*

[169] *[Tempeh] is <u>also</u> produced in this soy product factory.*

Néanmoins, ces placements sont loin d'être courants, ou même possibles, avec tous les types de focus. Dans son étude de corpus consacrée au placement et à l'utilisation des adverbes *also* et *too*, Fjelkestam-Nilsson indique que le placement de *also* dépend de la fonction syntaxique de l'élément focus, et identifie trois fonctions : sujet (tout ou partie), "prédicat" (tout ou partie), et autres constituants (op. cit., 1983 : 28). Le prédicat y est considéré comme incluant les verbes lexicaux et d'éventuels auxiliaires, ainsi que les compléments prédicatifs, aussi appelés attributs du sujet. Les compléments d'objets et les adjoints sont classés parmi les autres constituants.

Les positionnements pris en compte par Fjelkestam-Nilsson sont les mêmes que ceux qu'identifient Pullum et Huddleston (nous utilisons nos propres abréviations) : pré-adjacent (Pré-A), post-adjacent (Post-A), détaché du focus et le précédant (Pré-D), détaché du focus et le suivant (Post-D). L'utilisation d'*also* portant sur la proposition en entier n'est pas prise en compte, et le fonctionnement d'*also* par rapport à la présence d'auxiliaires n'est malheureusement pas documenté dans l'étude de Fjelkestam-Nilsson.

Le tableau suivant synthétise les résultats de cette étude, indiquant pour chaque fonction du focus quels placements sont observés dans le corpus et avec quelle fréquence. Les fonctions

sont illustrées à l'aide d'extraits de corpus, parfois tronqués, fournis par Fjelkestam-Nilsson. Précisons que les jugements concernant l'attribution du focus sont ceux de Fjelkestam-Nilsson et ne sont pas toujours en accord avec les analyses que nous aurions faites de ces configurations.

Focus	**Placements**		**Exemples**
Sujet	**Pré-A**	Fréquent seulement avec *there*	[170] *There is also [a large Royal Air Force station in the Vale].*
	Post-A	Fréquent	[171] *[Her mother] also was a person of superior mind.*
	Pré-D	Inexistant	
	Post-D	Très fréquent	[172] *[Saudi Arabian troops] had also arrived.*
Prédicat	**Pré-A**	Très fréquent	[173] *He also [expanded and modernized] the radio system.*
	Post-A	Rare	[174] *They [export] also.*
	Pré-D	Rare	[175] *It also, I should add, [formed] an easy means of re-entry.*
	Post-D	Rare	[176] *Most people [would oppose] it also.*
Autres	**Pré-A**	Fréquent	[177] *We can look back also [to the ultimate error].*
	Post-A	Rare	[178] *[For them] also a time will come.*
	Pré-D	Très fréquent	[179] *It also points up [the non-representation of urbanism in literature].*
	Post-D	Inexistant	

*Tableau 26. Synthèse des placements d'*also *par rapport à l'élément focus*

Les résultats de l'étude des placements d'*also* relativement à son focus confirment les généralisations formulées par les grammaires concernant la fréquence du placement médian, puisque les placements les plus fréquents repérés par Fjelkestam-Nilsson sont des placements directement après un sujet, juste avant un verbe lexical ou un auxiliaire, entre un auxiliaire et un verbe lexical, ou encore entre BE et un complément prédicatif.

Une incertitude subsiste en ce qui concerne la grammaticalité et l'acceptabilité de certains placements, et en particulier des placements post-verbaux, sur lesquels l'étude de Fjelkestam-Nilsson reste vague. Elle indique la rareté de placements post-adjacents au verbe lorsque celui-ci est le focus d'*also*, mais aussi la fréquence des placements pré-adjacents au complément. Ce second cas pourrait correspondre à un ensemble de configurations très

diverses, et en l'absence des données utilisées pour cette étude, il est impossible d'effectuer des vérifications.

Afin de pallier ce manque, nous avons réalisé une étude indicative concernant les placements d'*also* dans des corpus d'anglais natif contemporain. Notre objectif était d'évaluer la fréquence d'utilisation des placements post-adjacents au verbe, et notamment en SVAO, ainsi que le placement d'*also* par rapport aux auxiliaires. Les corpus utilisés sont le *British National Corpus* et le *Corpus of Contemporary American English*, auxquels nous avons eu accès grâce à la plateforme mise en place par Davies pour Brigham Young University. Pour chaque corpus, 500 occurrences d'*also* ont été analysées. Ce chiffre a été choisi parce qu'il correspond globalement au nombre d'occurrences d'*also* relevées dans la section français L1 de *ICLE* (voir ci-dessous), et à la moitié des occurrences prises en compte dans l'étude de Fjelkestam-Nilsson. Ces occurrences sont sélectionnées au hasard et proviennent de tous les sous-ensembles des corpus, excepté les sous-ensembles d'anglais oral.

Les cas d'emploi d'*also* avant le verbe avec les utilisations lexicales de HAVE et BE sont classés sous la première catégorie ([ALSO + Vlex]), mais les occurrences d'*also* entre BE et un complément prédicatif ou locatif sont classées dans une catégorie à part ([BE + ALSO + Compl.]). Les compléments de BE peuvent être des GN, des GP, ou des GAdj. L'intervention d'autres modifieurs dans le GV n'a pas été prise en compte dans les schémas liés au verbe. Une entrée spécifique a été créée pour le schéma *but/and also* en raison de sa fréquence. De la même façon, nous avons isolé les emplois d'*also* immédiatement après le verbe *see*, parce que ce type de placement semble limité à ce verbe. Les schémas dans lesquels *also* est placé entre un verbe lexical une proposition infinitive ou introduite par *that* sont également présentés séparément. *Also* à l'initiale correspond aux utilisations de l'adverbe seul en début de proposition, détaché par une virgule ou non. Nous avons regroupé sous la catégorie "Autres" les cas dans lesquels *also* précède une forme non-finie du verbe (hors construction avec un auxiliaire), ou est placé avant ou après un GN focus en début de proposition ou en apposition, ainsi que d'autres configurations mineures. Les différentes configurations rencontrées sont illustrées par des exemples dans le tableau 27, et le tableau 28 présente le nombre d'occurrences et le pourcentage correspondant des utilisations de ces schémas dans les deux corpus. Les schémas y sont présentés du plus fréquent au moins fréquent dans la mesure du possible, les résultats n'étant pas exactement les mêmes dans les deux corpus.

Schémas	**Exemples**
1. ALSO + Vlex	[180] *This cooperative effort also includes key operators in India (COCA)*
2. Aux + ALSO + Vlex	[181] *Warning notifications can also be too subtle (COCA)*
3. BE + ALSO + Compl.	[182] *It is also a scale model of industry practice (COCA)*
4. But/And + ALSO	[183] *This is not only incorrect, but also a largy security and safety problem (COCA)*
5. Aux + ALSO + Aux + Vlex	[184] *Spirits, mixers and minerals will also be waiting (BNC)*
6. ALSO initial	[185] *Also I give out powerful signals that I'm desperate to get another husband (BNC)*
7. ALSO + Aux + Aux/Vlex	[186] *These memories also will dissolve in a month or a year*
8. "See also"	[187] *See also feature on page 35 (BNC)*
9. Vlex + ALSO + GP	[188] *SMI as an identity was measured also by one item (COCA)*
10. Vlex + ALSO + GN	[189] *… which embraces also the citizenship of entitlement (BNC)*
11. Vlex + ALSO + Prop.	[190] *Remember also that few other countries have the same level of free healthcare*
12. Autres	[191] *It would also by itself prevent many serious errors (COCA)*

*Tableau 27. Illustration des schemas d'emploi d'*also *dans* BNC *et* COCA

Schémas	***BNC***		***COCA***	
	Nbr	**%**	**Nbr**	**%**
1. ALSO + Vlex	153	30,6	187	37,4
2. Aux + ALSO + Vlex	117	23,4	117	23,4
3. BE + ALSO + Compl.	108	21,6	57	11,4
4. But/And + ALSO	41	8,2	48	9,6
5. Aux + ALSO + Aux + Vlex	26	5,2	18	3,6
6. ALSO initial	14	2,8	15	3
7. ALSO + Aux + Aux/Vlex	1	0,2	20	4
8. "See also"	6	1,2	6	1,2
9. Vlex + ALSO + GP	2	0,4	1	0,2
10. Vlex + ALSO + GN	4	0,8	0	0
11. Vlex + ALSO + Prop.	2	0,4	0	0
12. Autres	26	5,2	31	6,2
TOTAL	**500**	**100**	**500**	**100**

*Tableau 28. Fréquence des schémas d'emploi d'*also *dans* BNC *et* COCA

La distribution d'*also* dans ces deux corpus est globalement comparable, avec quelques différences notables. Parmi les douze schémas identifiés, les quatre premiers sont les plus fréquents dans les deux corpus, et selon le même ordre de fréquence. Ceux-ci concernent le placement d'*also* avant un verbe lexical (schémas 1 et 2), entre BE lexical et un complément (schéma 3), et dans l'expression *but/and also* (schéma 4). Les différences importantes sont commentées dans les paragraphes qui suivent.

Les résultats de cette étude confirment les tendances relevées par les grammairiens : *also* privilégie les placements médians, c'est-à-dire avant un verbe lexical, entre un auxiliaire et un verbe lexical, ou entre deux auxiliaires (schémas 1, 2 et 5). À l'exclusion de l'expression quasi-idiomatique *see also* (schéma 8), les placements entre un verbe et un GN complément ne sont fréquents que lorsque ce verbe est BE (schéma 3). Les placements SVAO avec d'autres verbes sont donc très rares, voire inexistants, selon le corpus (schéma 10). Les placements d'*also* entre un verbe et une autre nature de dépendant, GP ou proposition (schémas 9 et 11), sont également très rares. L'acceptabilité du placement SVAO ne semble pas être sensible à la longueur du complément.

Concernant les positions qu'*also* adopte en présence d'un ou de plusieurs auxiliaires, on remarque qu'ici également le placement après le premier auxiliaire est privilégié. On observe le plus souvent des pourcentages relativement proches dans les deux corpus, excepté en ce qui concerne le placement d'*also* avant les auxiliaires et leur placement après BE : le corpus d'anglais américain contient beaucoup plus d'occurrences d'*also* utilisé avant un ou plusieurs auxiliaires (schéma 7), ce schéma étant quasiment inexistant dans le *BNC*. Parallèlement, on y relève deux fois moins d'occurrences d'*also* entre BE et un complément (schéma 3). S'il existe une plus grande flexibilité quant au placement d'*also* avant les auxiliaires, il est possible qu'elle s'applique également à BE, et qu'*also* soit donc plus souvent placé avant BE en anglais américain qu'en anglais britannique. Nous n'avons pas vérifié ce phénomène, d'autant plus que BE et HAVE ont été groupés avec l'ensemble des verbes lexicaux lorsqu'ils n'avaient pas un usage strictement grammatical (formation des aspects et du passif). Nous notons alors qu'on relève un plus grand nombre d'occurrences d'*also* avant un verbe lexical en anglais américain qu'en anglais britannique : puisque les structures de type [ALSO + BE] sont incluses dans la catégorie [ALSO + Vlex], ce plus grand nombre d'occurrences pour le schéma 1 pourrait en effet être le signe d'un transfert du schéma 3 au schéma 1 dans le corpus *COCA*.

Si une étude de corpus ne permet pas d'avoir accès à la "grammaire" des personnes locutrices de la langue afin d'évaluer quelles structures sont considérées comme

grammaticales, inacceptables ou agrammaticales, elle nous permet néanmoins de distinguer les schémas qui sont fréquemment utilisés de ceux qui, par leur rareté, peuvent être considérés comme non-standards. L'approche descriptiviste se doit d'aller dans les deux sens : s'il est inconcevable de proscrire l'utilisation de certaines structures alors que leur utilisation est attestée, il n'est pas plus sensé de considérer comme parfaitement standard des usages dont la fréquence est négligeable. L'utilisation de schémas non-standards, qui peut être considérée comme marquée, risque d'être un frein à la communication et à la compréhension du message transmis, ce qui justifie de proposer une alternative à ces schémas par le biais d'une proposition de correction.

Modélisation des erreurs dans le placement d'*also*

Détecter les erreurs par le biais de patrons de détection et de règles de réécriture implique non seulement d'avoir des informations sur les schémas corrects, mais également d'avoir une connaissance précise, à la partie du discours près si possible, de la forme que prennent les erreurs. Dans un objectif d'efficacité, il convient également d'avoir un modèle des schémas des erreurs permettant de limiter les configurations à traiter à celles qui sont susceptibles d'être produites par les francophones utilisant l'anglais. Si les résultats de l'étude de corpus présentée ci-dessus et les généralisations formulées par les grammaires nous permettent de savoir quels sont les placements standards d'*also*, ils ne fournissent pas d'informations concernant le format des erreurs, qui peuvent en théorie être extrêmement variées.

Afin de compléter les descriptions des schémas relevés dans notre corpus, nous avons mené une analyse du corpus *ICLE*. Nous avons sélectionné les textes produits par des personnes francophones (généralement de nationalité belge), n'ayant pas d'autres L1 et n'utilisant que le français dans leur entourage. Cette sélection a pour résultat un corpus de 177000 mots, dans lequel on relève 597 occurrences d'*also*. L'objectif de cette étude est double : il s'agit d'une part de vérifier la présence de segments non-standards dans les productions de francophones utilisant l'anglais, et d'autre part d'obtenir des informations précises sur ces segments.

Les types de schémas rencontrés fréquemment dans l'étude de corpus natifs ci-dessus sont considérés comme standards, et ceux que l'on rencontre rarement sont considérés comme non-standards. En l'absence d'indications claires provenant de sources vérifiées, nous évitons d'utiliser les appellations de "grammatical" et "agrammatical" pour les placements autre que SVAO. Les schémas y sont décrits du point de vue des groupes syntaxiques et, lorsque cela

est possible, des parties du discours, plutôt que de la fonction syntaxique des éléments concernés. Précisons que les catégories incluant la suite [Vlex + ALSO] ne concernent pas l'expression quasi-idiomatique *see also*, qui n'est jamais rencontrée dans le corpus étudié. Dans la mesure du possible, les numéros donnés aux schémas sont les mêmes que pour l'étude similaire menée sur des corpus d'anglais natif, sauf pour les schémas 2 et 5 qui sont regroupés sous un seul schéma, et le schéma 11 qui est lui distingué en deux schémas.

Schémas	**Occ. d'*also***		**Exemples**
	Nbr	**%**	
Segments standards			
1. ALSO + Vlex	168	28,14	[192] *Women also received the right to express their political opinion.*
2/5. Aux + ALSO + Aux/Vlex	167	27,97	[193] *Television can also be seen as a way of alienating the masses.*
3. *be* + ALSO + Compl.	100	16,75	[194] *Beside linguistic problems, there are also economic […] ones.*
4. *But/and* + ALSO	98	16,42	[195] *Wars have changed the frontiers and also the languages.*
6. ALSO initial	7	1,17	[196] *Also, practising their new jobs, women have come to […]*
7. ALSO + Aux + Aux/Vlex	2	0,34	[197] *This also could represent a kind of fetichism*
Autres	23	3,85	[198] *Here also, difficulties appear.*
Sous-total	565	94,6	
Segments agrammaticaux ou non-standards			
10. Vlex + ALSO + GN	18	3,01	[199] **Solidarity plays also an important part.*
9. Vlex + ALSO + Prep. + GN	7	1,17	[200] *?The red lobster refers also to Martin's weaknesses.*
11. Vlex + ALSO + *That* + GN	3	0,5	[201] *?Free will implies also that man must make choices.*
11'. Vlex + ALSO + *To* + Aux/Vlex	2	0,34	[202] *?The E.U. wants also to be taken seriously.*
Erreurs lexicales	2	0,34	[203] **loving someone also dominant*
Sous-total	32	5,4	
TOTAL	**597**	**100**	

*Tableau 29. Schémas des placements d'*also *dans* ICLE *(section français L1)*

Les placements les plus fréquents d'*also* sont proches de ceux observés dans les corpus *BNC* et *COCA* (voir Tableau 28). On retrouve environ 28 % d'occurrences d'*also* avant un verbe lexical dans le corpus non-natif, pour 34 % en moyenne dans les corpus natifs. Dans les deux études, environ 28 % des placements d'*also* sont après un auxiliaire, et environ 17% entre BE et un complément. Concernant les placements considérés comme standards, la

principale différence est observée pour l'utilisation du schéma 4, [but/and + ALSO], qui est près de deux fois plus fréquent dans le corpus d'interlangue que dans les corpus natifs (16,42 % contre 8,9 % en moyenne) : cet écart peut s'expliquer par une richesse d'expression moindre chez le public apprenant, menant à l'emploi plus fréquent de certaines expressions.

Malgré les correspondances entre les deux études concernant les usages les plus courants d'*also*, les résultats de cette étude de corpus viennent confirmer la présence de segments non-standards dans les productions de francophones, à hauteur de 5,4 % des occurrences d'*also*. Pour avoir un point de comparaison, nous avons effectué la même étude sur la section du corpus consacrée au néerlandais L1 : on y relève 605 occurrences d'*also*, 0,66 % d'entre elles faisant partie des usages non-standards identifiés ci-dessus, ce qui se rapproche du taux relevé dans les productions d'anglais L1 (0,9 % en moyenne sur les deux corpus *BNC* et *COCA*).

Schémas des erreurs et des corrections pour le placement de l'adverbe *also*

L'observation des schémas d'emploi non-standard d'*also* dans le corpus *ICLE*, ainsi que dans les exemples de notre propre corpus, fournit des informations sur les configurations précises des erreurs, dont nous nous servons pour dresser le tableau ci-dessous. Celui-ci vient clore les analyses linguistiques liées au placement d'*also* et présente une modélisation des placements erronés et de leurs corrections possibles. Plusieurs facteurs influençant les possibilités de détection et correction sont pris en compte :

- le nombre d'auxiliaires présents dans le GV ;
- la nature du verbe lexical ;
- le type de complément du verbe.

La présence ou l'absence d'un seul auxiliaire ne modifie pas la correction à apporter, puisque *also* est placé avant le verbe lexical ; cependant, si deux auxiliaires sont présents, *also* sera placé après le premier. Le cas de la présence de trois auxiliaires n'est pas pris en compte en raison de sa rareté. Cette caractéristique doit être incluse dès la phase de détection, ce qui implique un dédoublement des patrons de détection. Nous prenons également soin d'exclure le verbe *see* des verbes lexicaux concernés, étant donné l'existence de l'expression quasi-idiomatique *see also.* BE est également exclu des verbes lexicaux puisque le schéma [BE + ALSO + GN] est tout à fait grammatical et même très fréquent. Le complément du verbe peut être un GN, un GP, une proposition infinitive ou une proposition introduite par *that*.

Le tableau suivant inclut ainsi les différentes configurations observées relevant du placement post-verbal de cet adverbe, ainsi qu'une configuration plus rare dans laquelle *also* est introduit entre un second auxiliaire et le verbe lexical. La partie correspondant au schéma d'erreur, qui sert à l'élaboration des patrons de détection, est présentée dans la partie gauche du tableau. La partie droite donne la correction du schéma, qui sert à l'élaboration des règles de réécriture du segment concerné. Nous n'incluons pas la présence de sujets, puisque ceux-ci ne semblent pas interférer avec le placement post- ou pré-verbal d'*also*. Cette absence permet également d'alléger les patrons et leur processus d'implémentation, ainsi que d'éviter un certain nombre de faux positifs et d'erreurs dans la détection et la correction en lien avec la forme du sujet. Dans le cas d'*also*, une seule proposition de correction est suffisante. Les configurations incluant un GN et un GP à la suite du verbe sont incluses dans les mêmes schémas en rendant la préposition optionnelle. La modélisation présentée dans ce tableau constitue la base des patrons de détection et des règles de réécriture, dont l'implémentation est présentée dans le chapitre 3. Nous donnons le détail de chaque patron dans les paragraphes qui suivent le tableau. Le signe [+] est ici utilisé pour indiquer la précédence immédiate linéaire.

Erreur		**Correction**
1a. Vlex + **ALSO** + (Prep) + GN	⇨	**ALSO** + Vlex + (Prep) + GN
1b. Aux + Aux + Vlex + **ALSO** + (Prep) + GN	⇨	Aux + **ALSO** + Aux + Vlex + (Prep) + GN
2a. Vlex + **ALSO** + TO + Vlex	⇨	**ALSO** + Vlex + TO + Vlex
2b. Aux + Aux + Vlex + **ALSO** + TO + Vlex	⇨	Aux + **ALSO** + Aux + Vlex + TO + Vlex
3a. Vlex + **ALSO** + THAT + GN	⇨	**ALSO** + Vlex + THAT + GN
3b. Aux + Aux + Vlex + **ALSO** + THAT + GN	⇨	Aux + **ALSO** + Aux + Vlex + THAT + GN
4. Aux + Aux + **ALSO** + Vlex	⇨	Aux + **ALSO** + Aux + Vlex

*Tableau 30. Modélisation des placements erronés d'*also *et de leurs corrections*

Les schémas 1a et 1b sont deux versions du placement d'*also* entre un verbe lexical et un complément. Le complément peut être un GN ou un GP ; afin de rendre la modélisation la plus concise possible et de limiter le nombre de patrons, cette possibilité est aménagée en rendant la préposition optionnelle. Le schéma 1a n'inclut pas d'auxilaire, et l'adverbe est simplement déplacé devant le verbe lexical. Ce schéma prend également en charge les segments incluant un seul auxiliaire : dans le cas de la présence d'un auxiliaire unique,

l'adverbe doit être repositionné juste avant le verbe, ce qui rend cette configuration identique au schéma 1a. Le schéma 1b a été créé pour le traitement des segments dans lesquels on trouve deux auxiliaires dans le GV ; dans ce cas, la correction diffère de celle proposée dans le schéma 1a, car *also* doit être déplacé entre le premier et le second auxiliaire, et non avant le verbe lexical.

Les schémas 2a et 2b sont également deux versions du même placement d'*also,* permettant les ajustements nécessaires à la présence de deux auxiliaires. Ces schémas permettent de repérer les placements d'*also* entre un verbe lexical et une proposition infinitive introduite pas *to*. Ici également, la correction consiste à placer *also* avant le verbe, soit juste avant le verbe lexical, soit entre les deux auxiliaires.

Dans les schémas 3a et 3b, dédoublés pour les mêmes raisons, *also* est placé entre un verbe lexical et une proposition introduite par *that*. La correction consiste à placer l'adverbe avant le verbe lexical ou entre les deux auxiliaires. Rappelons ici que le verbe BE est exclu des possibilités pour les verbes lexicaux, puisque les configurations des schémas 1, 2 et 3 sont grammaticalement correctes lorsque le verbe central est BE.

Le schéma 4 permet de corriger les segments dans lesquels *also* est placé à l'intérieur d'un groupe de verbes contenant deux auxiliaires et un verbe lexical, et plus précisément entre le deuxième auxiliaire et le verbe lexical. La correction consiste à placer l'adverbe entre les deux auxiliaires.

c. Modélisation du placement des adverbes de manière

La catégorie des adverbes de manière, fondée sur des critères sémantiques, est une des catégories les plus traditionnelles et prototypiques des adverbes : "Aucune classification sémantique n'omet de mentionner l'adverbe de manière. C'est l'une des seules classes que l'on retrouve uniformément chez tous les grammairiens ; c'est également l'une de celles qui ne semblent guère poser de problème de reconnaissance" (Guimier, 1988 : 143). L'auteur de cette remarque poursuit cependant en notant que la légitimité de la reconnaissance d'une telle catégorie est remise en question par certains linguistes, notamment McCawley, qui la qualifie de "pseudo-catégorie" ne pouvant se prêter à une analyse uniforme (1973 : 407 ; cité par Guimier). Ce dernier propose ainsi quatre classes d'adverbe de manière, en fonction de l'élément précis dont l'adverbe prédique une caractéristique d'un point de vue sémantique.

La réflexion autour de la classification des adverbes n'est bien entendu pas dénuée d'intérêt, même après des décennies de recherche consacrées à ce sujet ; cependant nous n'irons pas plus loin dans la présentation de cette remise en question. La catégorie des adverbes de manière est par ailleurs reconnue par les grammairiens dont nous empruntons les descriptions. Pullum et Huddleston, dont nous avons utilisé la catégorisation, font des distinctions fines entre les différentes portées syntaxiques et sémantiques des adverbes en contexte, ce qui les amène à séparer les adverbes de manière des adverbes liés à l'acte, ou encore des adverbes de domaine, et ainsi à rendre les contours de la catégorie "manière" beaucoup plus nets. Nous laissons à des recherches futures le soin d'explorer la puissance explicative d'éventuelles catégorisations alternatives des adverbes en ce qui concerne leur placement.

Le placement des adverbes de manière : grandes lignes et précisions

D'après les grammaires consultées, les adverbes de manière sont généralement placés en position médiane ou en position finale. Quirk et al. indiquent que la position *eM*, c'est-à-dire juste avant le verbe lexical, est associée aux adverbiaux (définition de *ACG*) de manière et de degré, et plus particulièrement lorsque les adverbiaux sont instanciés par des adverbes. Les résultats de la vaste étude de corpus utilisée dans *LGSWE* montrent que les adverbiaux de manière sont plus de cinq fois plus nombreux en position finale, avec 80 % des occurrences, qu'en position médiane, représentant seulement 15 % des occurrences (1999 : 802). Lorsqu'un adverbe remplit cette fonction, il a néanmoins plus de chances d'être placé en position médiane, surtout dans les textes déclaratifs, qui constituent notre cible (op. cit. : 807). Pullum et Huddleston mentionnent quant à eux que les placements en fin de proposition sont souvent privilégiés (2002 : 579).

Deux placements corrects sont donc disponibles pour un même adverbe, ce qui signifie que deux solutions de correction peuvent en théorie être proposées pour une erreur donnée. Afin de guider les personnes recevant ces propositions, celles-ci peuvent être hiérarchisées. Le placement final semble être le plus standard. Nous verrons dans la dernière sous-section que l'implémentation des règles nécessite parfois d'adapter les propositions de correction afin d'éviter d'éventuels faux-positifs et des corrections erronées. Cependant, afin de travailler à partir d'une base linguistique la plus précise possible, nous souhaitons d'une part confirmer cette tendance et d'autre part découvrir les différents facteurs qui justifient un écart de la norme.

Dans *Adverbial Positions in English* (1964), Jacobson utilise un corpus d'anglais britannique de 3,5 millions de mots afin de documenter le placement des adjoints (définition de *CGE*) de manière exhaustive. Il identifie trois positions principales (initiale, médiane et finale) et respectivement trois, huit et cinq sous-types pour chaque position, ces sous-types se déclinant également en plusieurs subdivisions. Jacobson distingue les différentes structures fonctionnant comme adjoint, et recense donc les schémas spécifiques aux adverbes. Il utilise cependant la délimitation traditionnelle des adverbes, incluant dans son étude des mots comme *down*. La catégorie des adverbes de manière regroupe une majorité d'adverbes prototypiques (adverbe dérivés d'adjectif avec le suffixe *-ly*), et Jacobson manipule un corpus d'une taille impressionnante (et qui laisse songeur en regard de la date de publication de son ouvrage) : cette délimitation large devrait ainsi avoir peu d'impact sur le compte des occurrences de ces adverbes.

Le tableau 31 en page suivante présente les différents sous-types identifiés par Jacobson accompagnés de leur définition (op. cit. : 61-66), celle-ci étant nécessaire pour la compréhension du tableau 32. Les définitions sont accompagnées d'exemples que nous avons sélectionnés dans les corpus *COCA* ou *BNC* à partir des indications de Jacobson, et parfois modifiés pour les configurations les plus rares. Les exemples ne concernent pas systématiquement les adverbes de manière, notamment pour les cas où leur usage est rare (voir tableau 32). Pour la position médiane, seulement quatre des huit sous-types sont présentés, les autres ne trouvant pas d'occurrence pour les adverbes de manière. Les abréviations sont celles de Jacobson. Il est malheureusement impossible de donner les équivalents exacts de ces placements en fonction de ceux utilisés par Quirk et al., en raison de l'éloignement des critères utilisés et du degré de détail.

Le tableau 31 est fondé sur les résultats des analyses manuelles effectuées par Jacobson. Il concerne uniquement les adverbes de manière fonctionnant comme adjoint, dont l'auteur de l'étude a relevé 1425 occurrences (op. cit. : 80). Les nombres d'occurrences donnés par Jacobson sont ici convertis en pourcentage du total des occurrences, afin d'améliorer la lisibilité des résultats. Pour les sous-types M2, M3 et M4, qui concernent le placement des adverbes avant un auxiliaire, entre un auxiliaire et un verbe lexical, et après au moins deux auxiliaires, les chiffres présentés en italique correspondent au pourcentage d'adverbes dans cette position relativement au total des situations pour lesquelles ce placement est possible.

Subdiv.	**Définition**
Position initiale (*front-position*)	
F1	L'adverbe est en début de proposition, avant tous les autres éléments, sauf les conjonctions, et n'est pas le seul élément optionnel placé avant le sujet. [204] *Ironically, today, they did two things. (*COCA*)*
F2	L'adverbe est le seul élément optionnel en début de proposition. [205] *Fortunately, Jim got a call on his cell phone (*COCA*).*
F3	L'adverbe est en début de proposition, après d'autres éléments, sauf les conjonctions, et n'est pas le seul élément optionnel placé avant le sujet. [206] *In this case, however, each set of bars represents a single drawing. (*COCA*)*
Position médiane (*mid-position*)	
M1	L'adverbe est entre le sujet et un verbe lexical ou un auxiliaire suivi d'une ellipse du reste du GV. [207] *She quickly made friends with the Reverend. (*BNC*)* [208] *Well, hopefully we already have. (*BNC*)*
M2	L'adverbe est entre le sujet et un auxiliaire (y compris BE), lorsque l'auxiliaire est suivi d'un autre auxiliaire ou d'un verbe lexical. [209] *It definitely would have helped in some degree. (*BNC*)*
M3	L'adverbe est entre un auxiliaire (y compris BE) et un verbe lexical, un autre auxiliaire, ou un complément de BE. [210] *The state has actively sought to develop tourism. (*BNC*)* [211] *The quality of life is undoubtedly much better down here. (*BNC*)*
M4	L'adverbe est après deux auxiliaires ou plus (y compris BE), et avant un verbe lexical ou un complément de BE. [212] *Some fool will have accidentally released the spirit. (*BNC*)* [213] *His conclusions might be frequently uncomfortable. (*BNC*)*
Position finale (*end-position*)	
E1	L'adverbe est en fin de proposition, avant d'autres éléments optionnels. [214] *Cecily pulled away from her sister, slowly, deliberately. (*COCA*)*
E2	L'adverbe est entre un verbe lexical et un GN objet. [215] *She heard clearly the cries of the ghosts alone on the plain in the wind and snow. (*COCA*)*
E3	L'adverbe est en fin de proposition, entre deux éléments optionnels. [216] *Cecily pulled away from her sister, slowly, deliberately, as if in fear. (*COCA*, modifié)*
E4	L'adverbe est le seul élément optionnel en fin de proposition. [217] *Geder licked his lips nervously. (*COCA*)*
E5	L'adverbe est le dernier élément de la proposition, après d'autres éléments optionnels. [218] *The jurors were determined to listen to that 911 call one more time, very carefully. (*COCA*)*

Tableau 31. Les types de placement identifiés par Jacobson

	Position initiale		**Position médiane**		**Position finale**	
	Sous-type	**%**	**Sous-type**	**%**	**Sous-type**	**%**
	F1	0,35	M1	9,12	E1	16,21
	F2	2,95	M2	*0,45*	E2	0,77
	F3	0,42	M3	*77,27*	E3	7,50
			M4	*73,08*	E4	32,14
					E5	14,53
Sous-total		**3,72**		**25,12**		**71,16**
Total	**100**					

Tableau 32. Distributions des adverbes de manière dans le corpus de Jacobson

Dans les grandes lignes, on retrouve des résultats similaires à ceux de Biber et al., et qui sont congruents avec les descriptions que nous avons déjà données : les adverbes de manière sont peu fréquents à l'initiale (la présence de détachement prosodique n'est pas documentée), ils sont courants en position médiane, et privilégient globalement un placement en fin de proposition. Concernant ce dernier cas, les résultats de Jacobson montrent que les adverbes sont le plus souvent placés juste après les éléments obligatoires du GV ; cette configuration, qui regroupe les placements en E1 et E4, représente 48,35 % du total des placements. Le placement de type SVAO est rare.

La catégorie du placement médian est dominée par le placement avant le verbe lexical du GV, qui regroupe les positions M1, M3 et M4, et représente 25,05 % des occurrences. Dans les GV présentant plus d'un auxiliaire, l'auteur n'introduit malheureusement pas de distinction entre le placement des adverbes de manière juste avant le verbe lexical et entre les deux auxiliaires, qui sont regroupés en M3. Il semble rare de trouver des adverbes de manière entre le sujet et un auxiliaire (M2). Pour l'adverbe *also*, nous avions rencontré plus d'occurrences de son placement dans cette position en anglais nord-américain, or Jacobson utilise un corpus d'anglais britannique uniquement, fait qu'il faut également garder à l'esprit lors de la comparaison de ces résultats avec ceux de Biber et al. et Dixon notamment.

De plus, Jacobson documente un facteur supplémentaire susceptible de modifier le placement des adverbes en général. Ainsi, il indique que les GAdv incluant un modifieur (ex. : *nearly*, *very*) sont plus souvent placés en position finale que les adverbes seuls : le pourcentage d'adverbe en position médiane passe de 50 % à 24 % lorsqu'ils sont modifiés, et de 43 % à 66 % pour la position finale (op. cit. : 107). Ce facteur peut donc accentuer la tendance des adverbes de manière à être placés en fin de proposition. Par ailleurs, les adverbes

de manière sont de bons candidats à la modification par des intensifieurs, ceux-ci pouvant souvent être gradués.

Dans *A Semantic Approach to English Grammar* (2005), Dixon n'hésite pas à trancher sur la question du placement des adverbes par rapport aux auxiliaires, jugeant que celui-ci dépend de la "fonction" de l'adverbe concerné. Dans le cadre théorique de Dixon, les adverbes peuvent adopter une "fonction de manière" (*manner-function*) ou une "fonction de phrase" (*sentential function*). L'appellation "manière" regroupe chez Dixon plusieurs catégories sémantiques traditionnelles, comme les adverbes de degré et de moyen. Il indique que les adverbes en fonction de phrase, comme l'adverbe *also*, sont généralement placés juste après le premier auxiliaire ; le placement est considéré comme de moins en moins standard et acceptable à mesure que l'adverbe se rapproche du verbe lexical (après le second auxiliaire, et éventuellement après le troisième) (2005 : 389). Les adverbes en fonction manière, qui portent sur le verbe, sont placés juste avant celui-ci. Ceux-ci disposent d'ailleurs uniquement de deux placements possibles : immédiatement avant le verbe lexical, et immédiatement après le verbe et ses compléments éventuels (op. cit. : 386).

Dixon souligne la possibilité pour les adverbes d'intervenir entre un verbe et une proposition complément (op. cit. : 248, 393). L'auteur nous renseigne également sur la possibilité d'introduire un adverbe de manière entre un verbe et un complément autre qu'un GN. D'après ce linguiste, lorsqu'un verbe est suivi d'un GP complément dont il sélectionne la préposition, comme dans l'exemple suivant emprunté à Dixon, un adverbe de manière peut être positionné entre le verbe et le GP :

[219] *They decided <u>most carefully</u> on a new chairperson.*

La description de Dixon présente néanmoins une image très simplifiée de l'interaction des adverbes avec les GP et les prépositions sélectionnées par les verbes. D'après Huddleston, on trouve trois principaux schémas (2002 : 272-290) (exemples empruntés à Huddleston) :

- les verbes prépositionnels : ces verbes sélectionnent la préposition noyau lorsqu'ils sont suivis d'un GP complément, les prépositions pouvant être fixes ou mobiles :

  [220] *I <u>referred to</u> her book.*

- les verbes à particule : ces verbes sont complétés par une unique particule, souvent instanciée par une préposition locative intransitive, qui intervient soit entre le verbe et un éventuel GN objet, soit après ce dernier ;

  [221] *She <u>took off</u> the label.*

[222] *She took the label off.*

- les idiomes verbaux : un verbe et une ou plusieurs prépositions (d'autres configurations sont également possibles) forment une unité lexicale.

[223] *He carried out his threat.*

[224] *We look forward to your visit.*

Le schéma décrit par Dixon correspond à celui des verbes prépositionnels. Les travaux dédiés à la syntaxe et au placement des adverbes n'abordent pas cette question en détail, mais Huddleston inclut des indications concernant le placement des adjoints par rapport à ces différents schémas.

Pour commencer, Huddleston confirme la possibilité de placer un adverbe de manière entre un verbe et un GP complément, en dehors des trois configurations citées ci-dessus (2002 : 276, 282). Par exemple, dans la phrase suivante il ne s'agit pas d'une combinaison entre un verbe et une préposition sélectionnée par le verbe, mais d'un verbe seul suivi d'un complément de lieu, le choix de la préposition étant déterminé par le sens donné au complément :

[225] *I ran quickly to the door.*

Les verbes complétés par une particule ne peuvent généralement être séparés de celle-ci que par un objet, et uniquement lorsque le verbe lui-même le permet. Ces éléments ne peuvent donc pas être séparés par un adverbe de manière (op. cit. : 282). Conformément à la tendance générale concernant le placement des adverbes en SVAO, un adverbe de manière ne peut pas non plus venir s'insérer entre un verbe suivi d'une particule et un GN objet. La présence d'une telle configuration dans le GV ne modifie donc en rien les possibilités de placement habituelles des adverbes de manière (exemples empruntés à Huddleston, op. cit. : 282) :

[226] **She took carefully off the label.*

[227] **She took off carefully the label.*

[228] *She took off the label carefully.*

[229] *She carefully took off the label.*

Dans le cas des idiomes verbaux contenant des prépositions intransitives (ex. : *break up*, *cut down*, *keep on*), la possibilité de placer un adjoint entre le verbe et la préposition est grandement réduite ; l'acceptabilité de ce placement semble être déterminée par le degré de lexicalisation de l'idiome, et jusqu'à quel point celui est non-compositionnel. Huddleston compare les deux phrases suivantes (op. cit. : 285) :

[230] *It faded gradually away.*

[231] **He passed gradually away.*

La première phrase est un exemple d'un idiome verbal autorisant le placement d'un adverbe de manière entre le verbe et la preposition, alors que la seconde illustre un cas d'impossibilité avec un idiome plus lexicalisé. Ce type de placement tend néanmoins à être peu fréquent, d'autant plus que d'autres positions sont le plus souvent disponibles.

La possibilité de placer un adverbe entre un verbe et un GP complément dont la préposition est sélectionnée par le verbe dépend du statut de celle-ci. S'il s'agit d'une préposition mobile, les possibilités de placement sont les mêmes que dans une configuration où la préposition ne dépend pas du verbe, c'est-à-dire que le GP complément peut être séparé du verbe par un adverbe ; si la préposition est fixe, elle ne peut être détachée du verbe. Cette différence est illustrée par les deux exemples qui suivent :

[232] *I referred politely to her book.*

[233] **I came somehow across these letters.*

Dans les deux cas, nous avons affaire à un verbe sélectionnant une préposition : celle-ci est mobile dans le premier exemple, et fixe dans le second. L'exemple donné par Dixon (*They decided most carefully on a new chairperson*) correspond au premier cas, ce qui explique son acceptabilité. Dans l'exemple suivant, un tel placement est impossible.

Le tableau 33 en page suivante est une synthèse des possibilités de placement des adverbes avec des verbes suivis de prépositions, du point de vue des éléments visibles en surface. L'abréviation AdvM désigne les adverbes de manière. Les GP sont scindés en Prep et GN afin de pouvoir représenter tous les types de configuration. Le symbole [✓] indique la grammaticalité d'un schéma, et le symbole [×] son agrammaticalité.

Type		**Schémas de surface**	**✓/×**
Verbe + GP complément		Vlex + AdvM + Prep + GN ex. : *I ran quickly to the door.*	✓
Verbe + particule		Vlex + AdvM + Prep + GN ex. : **She took carefully off the label.*	×
		Vlex + Prep + AdvM + GN ex. : **She took off carefully the label.*	×
Idiome verbal		Vlex + AdvM + Prep + GN ex. : *It faded gradually away.* ex. : **He passed gradually away.*	✓/×
Verbe prépositionnel	P. mobile	Vlex + AdvM + Prep + GN ex. : *I referred politely to her book.*	✓
	P. fixe	Vlex + AdvM + Prep + GN ex. : **I came somehow across those letters.*	×

Tableau 33. Acceptabilité du placement des adverbes dans les combinaisons Verbe + GP

Il apparaît que l'acceptabilité du même type de configuration de surface varie en fonction de critères lexicaux, comme la mobilité d'une préposition, ou l'existence d'un idiome. Cette variation est très problématique pour la reconnaissance des schémas d'erreur, d'autant que nous ne disposons pas d'informations concernant la fréquence de chaque configuration en anglais. Ainsi, même s'il existe plus de configurations n'acceptant pas le placement d'un adverbe entre un verbe et un GP, cela ne se traduit pas nécessairement par une majorité de segments de ce type dans les textes, certaines configurations étant sans doute plus fréquentes que d'autres. Nous expliquons dans la section 3.1.1c que les approches de la correction grammaticale automatisée à destination d'un public apprenant se doivent d'être extrêmement prudentes lorsqu'il s'agit de détecter les erreurs, afin d'éviter des détections erronées. Pour cette raison, nous laissons de côté pour l'instant le traitement des erreurs dans lesquelles un adverbe est placé entre un verbe et un GP dépendant du verbe, en raison du risque de faux positif qu'il porte. Des recherches futures sur les aspects lexicaux de ces structures pourront nous permettre d'y revenir. Par contre, nous prenons en compte dans les patrons de détection la présence des verbes à particule, qui adoptent le même comportement que les verbes sans particule. Cette configuration est visible dans le tableau synthétique présentant la modélisation du placement des adverbes (voir tableau 36 en page 211).

Exploration de facteurs supplémentaires à l'aide de tests de jugement de grammaticalité

Nous avons vu que les adverbes de manière tendent le plus souvent à être placés à la fin de la proposition, mais sont également parfois placés juste avant le verbe lexical. Les travaux

consultés jusqu'à présent n'évoquent cependant pas de facteurs sémantiques ou structurels amenant à préférer l'un ou l'autre de ces placements, hormis l'inclusion d'un modifieur interne au GAdv, qui le rendrait plus long et donc moins éligible à un placement en milieu de phrase. Afin d'explorer cette variation, nous avons mené des tests de jugement de grammaticalité auprès de locuteurs et locutrices anglophones.

Nous avons identifié deux facteurs structurels supplémentaires susceptibles d'influencer le placement : la présence/absence d'un GN objet à la suite du verbe, et la longueur du GAdv. Afin de tester ces hypothèses, nous avons soumis des tests de jugement de grammaticalité à quatre sujets anglophones. Trois sujets ont pour langue maternelle un anglais d'origine nord-américaine, et un sujet a pour langue maternelle une variété d'anglais britannique. Trois sujets ont un niveau de maîtrise du français évalué à C2+ et résident en France depuis plusieurs années, et un sujet apprend le français au niveau A1 et n'a jamais résidé en France. Les trois sujets résidant en France enseignent l'anglais au niveau universitaire. Nous ne présentons ici qu'une partie limitée du test, qui était à l'origine beaucoup plus étendu ; la longueur du test entier explique le nombre très réduit de personnes ayant participé à l'étude.

Le test est composé de dix ensembles de trois à cinq phrases déclaratives représentant les différentes positions plausibles pour un adverbe de manière sans détachement prosodique. Les ensembles incluent certains exemples agrammaticaux, mais uniquement un d'entre eux, dans lequel le GN objet est lourd, inclut un exemple de placement SVAO. On a demandé aux sujets d'évaluer la grammaticalité et l'acceptabilité des exemples de chaque ensemble, en indiquant si une phrase est grammaticale et acceptable, grammaticale mais inacceptable car peu naturelle, ou agrammaticale. Nous avons également incorporé au test de grammaticalité un test de préférence, les sujets devant cocher une case supplémentaire indiquant quelle phrase leur semblait la plus naturelle. Une consigne supplémentaire demande aux sujets de souligner l'adverbe lorsqu'un placement leur semble correct uniquement si l'adverbe fait l'objet d'une emphase. Le tableau 34 en page suivante est un exemple de ces ensembles de phrases. Le contenu complet du test ainsi que les consignes sont disponibles en annexe.

L'influence du poids du GAdv sur son placement est évaluée en comparant les choix de placement des sujets anglophones avec un adverbe court (*slowly*), un adverbe long (*erratically*), un adverbe court modifié par *very*, et un adverbe long modifié par *very*. Deux autres ensembles de phrases évaluent l'influence de la modification du GAdv lorsque le GV ne contient pas d'auxiliaire. L'influence de la présence, de l'absence et de la nature du GN objet est évaluée en comparant des phrases incluant un GV intransitif, un GV complété par un

pronom, et un GV complété par un GN long (six mots, ne contenant pas de proposition enchâssée). La liste des critères est visible dans le tableau 35, présenté après le tableau 34.

	Sentences	**OK**	**Incorrect**		**Best choice**
			Gram. but unnatural	**Ungram.**	
1.	Slowly she had opened the door to the second guestroom.				
2.	She slowly had opened the door to the second guestroom.				
3.	She had slowly opened the door to the second guestroom.				
4.	She had opened slowly the door to the second guestroom.				
5.	She had opened the door to the second guestroom slowly.				

Tableau 34. Test de jugements de grammaticalité : exemple d'un ensemble de cinq phrases

Type d'exemple	**I**	**iM**	**e/M**	**iE**	**E**	**Pref.**
a. Adverbe court	-	-	+		+	E
b. Adverbe long	-	-	+		+	E
c. Adverbe court avec prémodification	?	-	+		+	E
d. Adverbe long avec prémodification	-	-	?		+	E
e. Pas d'objet après le verbe	-	-	+		+	E
f. Pronom après le verbe	-	-	+		+	E/eM
g. GN long après le verbe	?	-	+	-	+	eM
h. GV sans auxiliaire, adverbe court	+		+		+	eM
i. GV sans auxiliaire, adverbe court avec prémodif.	+		?		+	E
j. GV avec auxiliaire, adverbe court	+	-	+		+	eM

Tableau 35. Résultats des tests de jugement de grammaticalité

Le tableau 35 est une synthèse des jugements donnés par les sujets. Les symboles [+] et [-] sont utilisés pour indiquer l'acceptation ou le rejet d'un placement, à partir d'une moyenne entre les jugements des sujets. Par exemple, si un exemple est jugé grammatical par une personne mais agrammatical par les trois autres, il est considéré comme agrammatical. Si un segment est considéré comme peu naturel par deux sujets et comme agrammatical par les deux autres, il reçoit un jugement global d'agrammaticalité. Le point d'interrogation, [?], indiquent que les jugements ne sont pas tranchés, par exemple si un exemple est jugé comme grammatical par deux personnes et comme peu naturel par les deux autres. Les cases barrées

représentent les positions qui ne sont pas documentées ou que la structure de la phrase rend indisponibles. Les cases grisées correspondent aux jugements unanimes. Les abréviations de placement sont celles que nous avons utilisées dans le tableau 21 ; le placement *iE* représente ici uniquement le placement entre un verbe et un objet direct. La colonne de gauche indique quel placement est jugé optimal, à l'unanimité (case grisée) ou à la majorité.

Le premier résultat remarquable est le manque d'unanimité dans les réponses des sujets anglophones, qui ne s'accordent sur l'optimalité d'un placement que dans la moitié des cas. Comme l'avait déjà remarqué Jacobson, le fait que le GAdv contienne un modifieur tend à rendre la position médiane moins acceptable. La longueur de l'adverbe en lui-même ne semble pas avoir d'influence sur les préférences de placement. Concernant la présence et la longueur des GN objets, il semble que le fait que le verbe soit intransitif influe légèrement sur la capacité de l'adverbe à être placé avant celui-ci. La présence d'un GN long, au lieu de faciliter un placement en SVAO, déclenche une préférence pour le placement médian. De manière inattendue, mais pas forcément significative, l'absence d'un auxiliaire dans le GV a pour résultat une préférence pour le placement médian.

Ce type de test a cependant de nombreuses limites. Celles-ci sont bien résumées par Ayoun (2005 : 60) :

> Grammatical judgments are better viewed as one type of decision-making or judgment behavior among many others, and as such, subject to variation and extra-linguistic factors. For instance, participants may reject a sentence because they think that it is semantically or pragmatically odd; or they may end up rejecting a sentence they wanted to initially accept after over-analyzing it. Extra-linguistic factors also include the participants' state at the time of the data collection: They may have been tired, distracted, too eager, etc.; in other words a variety of conditions having nothing to do with their linguistic representations.

On peut ajouter à ces limites l'autorité linguistique allouée aux locutrices et locuteurs natifs, qui soulève les questions que nous avons évoquées dans notre présentation de l'approche de l'anglais comme *lingua franca*.

La plupart des facteurs extra-linguistiques évoqués par Ayoun sont applicables à notre étude : interprétation différente des consignes, perte de repères et sur-analyse des phrases, fatigue, etc. La significativité des résultats de notre étude est elle-même largement limitée par le faible nombre de sujets. Une étude à plus grand échelle pourrait donner une vision plus fiable des phénomènes. Par ailleurs, même si nous avons gardé les mêmes schémas et tenté de

conserver des phrases et des adverbes comparables, les résultats peuvent être largement influencés par le sens des phrases et les collocations pouvant exister entre les différents éléments. Ceci est visible dans le fait que l'exemple [j.], qui reproduit le même schéma et emploie le même adverbe que l'exemple [a.] fait pourtant l'objet de jugements différents. On peut également évoquer les possibles différences liées aux variétés de l'anglais utilisées, à l'influence de la connaissance du français pour les sujets concernés, et aux variations dans les idiolectes, facteurs à prendre en compte lorsqu'on s'intéresse à des phénomènes à grande flexibilité.

Une étude à plus grande échelle serait donc nécessaire pour dégager des tendances fiables, mais nous pouvons déjà retenir quelques indications, notamment sur la hiérarchie à adopter dans les propositions de placement :

- si un adverbe est modifié par un adverbe intensifieur, le GAdv doit être placé en priorité en fin de proposition ;
- l'absence de complément après le verbe et l'absence d'auxiliaire dans le GV assouplissent les possibilités de placement médian ;
- la présence d'un GN long rend le placement médian plus favorable que le placement final.

Ces indications servent de base à la création des schémas d'erreur et de correction présentés dans la sous-section suivante.

Schémas des erreurs et des corrections pour le placement des adverbes de manière

Afin de modéliser la correction des erreurs liées au placement d'*also*, nous avons pu observer un ensemble d'erreurs différentes grâce à une recherche ciblée dans un corpus d'anglais L2. Le type sémantique des adverbes ne faisant pas l'objet d'un étiquetage spécifique lors de l'annotation de ces corpus, il est impossible d'effectuer la même recherche pour les adverbes de manière. Il est certes possible d'étendre la recherche à tous les adverbes et de trier manuellement les différents types sémantiques, mais une étude d'une telle ampleur dépasse le cadre de nos travaux actuels. La modélisation des erreurs dans le placement des adverbes de manière repose donc sur des données plus limitées.

D'après les études sur le placement des adverbes anglais par des francophones, les erreurs concernent dans l'immense majorité des cas les placements SVAO. Les erreurs relevées dans notre corpus montrent la même tendance. En conséquence, les règles de détection et

correction se concentrent sur ce type d'erreur. Une étude de corpus d'interlangue à grande échelle serait nécessaire afin de dégager d'autres tendances moins évidentes.

Les sous-sections précédentes ont fait le détail des facteurs à prendre en compte dans la détection et la correction des erreurs. Les principaux éléments sont les suivants :

- le choix du lieu de placement (final vs. médian),
- le placement de l'adverbe en rapport avec les auxiliaires,
- les règles à adopter concernant les adverbes placés dans des GV incluant un GP dépendant,
- la prise en compte des GN objets longs.

Nous présentons les réponses apportées pour la prise en compte de ces quatre facteurs dans les paragraphes suivants. Les patrons sont ensuite détaillés précisément après leur exposition schématique dans le tableau 36.

Dans la grande majorité des cas, les positions finale et médiane sont toutes deux disponibles, et deux solutions de correction peuvent donc être proposées. La position finale étant plus fréquente, la logique voudrait que celle-ci soit proposée en premier. Cependant, même si cette étude porte uniquement sur les adverbes de manière, nous devons prêter attention aux difficultés posées par la polysémie des adverbes. Du point de vue d'un traitement automatisé, le placement médian a ceci d'intéressant qu'il est accepté par presque tous les types sémantiques d'adverbes (voir sous-section 2.2.1e "Synthèse du placement des adverbes selon leur catégorie sémantique") ; il permet ainsi de conserver une éventuelle ambiguïté sur le sens de l'adverbe, ce qui est utile lorsque nous ne sommes pas en mesure de savoir quel sens la personne concernée souhaite donner à l'adverbe.

Qui plus est, placer les adverbes en position finale implique de pouvoir évaluer les frontières du GN objet. Le risque de fournir une solution de correction plus problématique que le segment original est donc plus élevé avec ce placement. Comme nous le verrons lorsque nous aborderons la question de la correction automatisée pour l'enseignement des langues assisté par ordinateur (cf. 3.1.1c), il est primordial de présenter au public apprenant les informations les plus justes possibles. En effet, ce public particulier ne dispose pas des connaissances linguistiques nécessaires pour nuancer les indications données par la machine. Ceci est d'autant plus important lorsqu'on a affaire à des phénomènes pour lesquels les règles de construction et d'utilisation reposent sur un ensemble de variables complexes. Privilégier la position médiane dans les solutions de correction permet également de rester proche du

segment original, dans lequel l'adverbe a été placé à proximité du verbe. Pour ces raisons, nous avons opté pour une présentation des solutions de correction qui propose en premier le placement médian, et en second le placement final lorsque deux solutions sont possibles. Lorsqu'un placement médian n'est pas acceptable, comme par exemple lorsque l'adverbe de manière est modifié par un adverbe de degré, rendant le GAdv trop long pour un placement médian, le placement final est privilégié.

En ce qui concerne le placement des adverbes par rapport aux auxiliaires, les descriptions linguistiques font apparaître clairement que les adverbes de manière en position médiane sont placés juste avant le verbe lexical. Ce placement nous permet de ne pas inclure les auxiliaires dans les patrons. Ceci constitue une simplification bienvenue, que ce soit en termes de rapidité du système, qu'en ce qui concerne la gestion de la négation dans le GV. En effet, les patrons et règles s'appliquent ainsi tout aussi bien à une phrase déclarative affirmative que négative.

Toujours dans l'objectif de limiter les faux positifs, le verbe BE a été exclu des patrons après que des tests préliminaires ont révélé la production de faux positifs en lien avec ce verbe (ex. : *This is clearly a good idea*, qui affiche la structure de surface [Verbe + AdvM + GN]). Cette exclusion n'a pas d'effets negatifs sur la qualité de la detection, puisque, dans son usage lexical, BE véhicule un sens statique, ce qui rend rare, voire impossible, sa co-occurrence avec un adverbe de manière, fondamentalement dynamique. La mise en place de cette exclusion, ainsi que d'autres ajustements mineurs, est expliquée en détail dans la section 3.2.1b, consacrée à la présentation technique des patrons et règles.

Le placement des adverbes dans des GV incluant des dépendants prépositionnels pose des difficultés, en particulier dans la phase de détection. En effet, les mêmes schémas de surface peuvent être jugés corrects comme erronés en fonction de critères lexicaux. Le fait qu'il existe un plus grand nombre de configurations refusant le placement d'un adverbe entre un verbe et un GP que de configurations acceptant ce placement ne garantit en aucun cas que les premières soient plus fréquentes dans les textes. Malheureusement, nous ne disposons pas encore des ressources nécessaires à l'identification des différents paramètres lexicaux liés aux combinaisons [Vlex + GP], et nous avons évoqué ailleurs les difficultés que pose la correction automatisée des erreurs liées aux prépositions. On remarque néanmoins que les configurations de type [Vlex + Prep. + GN ou GP], correspondant aux verbes à particules suivis d'un dépendant sous forme de GN ou de GP, n'acceptent jamais le placement d'un adverbe entre la particule et le dépendant qui la suit (voir tableau 33 ci-dessus). Si l'on considère que la

particule est un élément attenant au verbe, ce type de placement correspond globalement au placement SVAO. Ce schéma est donc inclus parmi les patrons d'erreur à relever dans le système.

Le dernier point à prendre en compte est la présence de GN objets lourds. Osborne montre que deux critères doivent être respectés afin que le placement d'un adverbe entre un verbe et un GN objet soit acceptable : le GN doit être particulièrement long, et une collocation doit exister entre le verbe et l'adverbe (2008 : 136-138). D'après ses résultats, ce second critère est est rarement reconnu et respecté par les francophones (op. cit. : 143). Par ailleurs, la présence d'un GN objet long peut empêcher un placement final. En conséquence, les schémas présentant un placement SVAO incluant un GN objet long se voient proposer une seule solution de correction prenant la forme d'un déplacement vers la position médiane (voir schéma 6).

À ces principaux points s'ajoutent des facteurs dont la gestion est plus aisée. L'intensifieur *very* étant l'un des plus courants, la possibilité de sa présence dans le GAdv est prise en compte dans la modélisation. Il faut cependant aussi prendre en compte le fait que la présence de *very* décourage le placement médian, sauf dans le cas où le GN objet est long, cas dans lequel le GAdv est plus avantageusement placé avant le verbe. Les schémas peuvent être adaptés à d'autres intensifieurs, avec les modifications qui s'imposent (ex. : la présence de *too* rend agrammatical le placement médian, ex. *I very quickly forgot everything she had told me that fateful day* vs. **I too quickly forgot everything she had told me that fateful day*).

Finalement, un des schémas est dédié au traitement de l'adverbe *well*. Celui-ci est un adverbe de manière qui a un comportement spécifique, en ceci qu'il n'accepte un placement médian que lorsqu'il suit les auxiliaires BE ou MAY ; par ailleurs, avec MAY l'adverbe n'a plus un sens de manière mais d'évaluation :

[234] *The dance was well executed.*

[235] *He may well leave town if we let him.*

Le placement final est donc la seule solution de correction possible pour cet adverbe. On peut inclure dans ce schéma la présence optionnelle de *very*. Les autres schémas ne s'appliquent pas à cet adverbe.

Les schémas d'erreur accompagnés des propositions de correction sont présentés dans le tableau 36. L'inclusion d'éléments optionnels et la fusion de schémas similaires permettent de

limiter le nombre de schémas à trois, déclinés avec et sans l'adverbe de degré *very* pour le schéma 5. La numérotation des schémas reprend la suite des schémas proposés pour *also*.

Erreur		**Correction**
5a. Vlex + (Prep) + **AdvM** + GN court	⇨	**AdvM** + Vlex + (Prep) + GN court Vlex + (Prep) + GN court + **AdvM**
5b. Vlex + (Prep) + **Very + AdvM** + GN court	⇨	Vlex + (Prep) + GN court + **Very + AdvM**
6. Vlex + (Prep) + **(Very)** + **AdvM** + GN long	⇨	**(Very) + AdvM** + Vlex + (Prep) + GN long
7. Vlex + (Prep) + **(Very)** + **WELL** + GN court	⇨	Vlex + (Prep) + GN court + **(Very)** + **WELL**

Tableau 36. Modélisation des placements erronés des AdvM et de leurs corrections

Les schémas 5a et 5b correspondent à la configuration la plus simple, qui repère les adverbes de manière placés entre un verbe et un GN. Pour le schéma 5a, dans lequel l'adverbe n'est pas modifié par adverbe de degré, deux corrections sont proposées : le placement de l'adverbe juste avant le verbe lexical, et le placement de l'adverbe après le GN. Dans le schéma 5b, le GAdv est composé d'un adverbe noyau modifié par l'adverbe *very* ; la longueur du GAdv appelle dans ce cas un placement final plutôt que médian. Une préposition optionnelle est introduite après le verbe afin de permettre le traitement des structures incluant un verbe à particule. L'inclusion de la préposition s'applique à tous les schémas.

Le schéma 6 permet d'adapter la correction des erreurs de placement d'adverbe en fonction de la longueur du GN complément. En effet, nous avons vu que dans le cas où le GN complément est long, il n'est pas opportun de placer l'adverbe après celui-ci. Les raisons de ce choix sont linguistiques et techniques. D'un point de vue linguistique, un placement final risque de perturber l'équilibre de la distribution de l'information dans la phrase. D'un point de vue technique, le GN long risque de ne pas être reconnu en entier, ce qui amènerait l'adverbe à être placé au milieu de celui-ci, et produirait une phrase erronée. Le fait que la correction consiste à placer le GAdv en position médiane nous permet d'éviter la création d'un schéma supplémentaire pour l'inclusion de *very*, qui est ici inclus de manière optionnelle.

Le schéma 7 correspond au cas particulier de l'adverbe *well*, qui n'accepte pas la position médiane. La correction proposée est donc un placement final dans le cas où le GN complément est court. L'adverbe *very* est également introduit dans le schéma de manière optionnelle. Notons que nous ne proposons pas de correction pour le cas de la présence d'un GN long : le placement final étant problématique et le placement médian impossible, la conservation de la structure de départ semble être la meilleure option.

Ce dernier tableau conclut la présentation des différents aspects de la syntaxe externe des adverbes et du traitement des erreurs qui lui sont liées. Le caractère nébuleux de cette catégorie est visible à plusieurs niveaux, et est la source d'analyses différentes concernant sa délimitation, ses différents types sémantiques, et ses fonctions dans le GV et la proposition. La grammaticalité de ses placements, notamment le placement SVAO, donne également lieu à des interrogations et des opinions divergentes, au moins en apparence. Toutes ces caractéristiques composent le défi que cette catégorie représente pour le traitement automatisé de l'anglais, comme pour son enseignement. Ainsi, les recherches en acquisition des langues secondes au sujet du placement des adverbes montrent que les francophones rencontrent des difficultés d'utilisation et d'acquisition des placements canoniques des adverbes en anglais. Les adverbes de manière et l'adverbe *also* sont particulièrement concernés par les erreurs, notamment en raison de leur fréquence d'emploi.

Afin de tenter de répondre à ces difficultés par un traitement automatisé, nous avons synthétisé les descriptions du placement de ces adverbes à partir de grammaires et de travaux ponctuels. Lorsque cela a été possible, notamment dans le cas d'*also*, nous avons complété ces descriptions par des études de corpus d'anglais L1 et L2, afin, d'une part, de découvrir la fréquence d'utilisation de certains schémas, et d'autre part de cibler les solutions de correction proposées. Des tests expérimentaux, sous la forme de jugement de grammaticalité par des anglophones, sont également venus compléter le tableau que nous avons peint du placement de ces adverbes. Malgré un certain nombre de limites que nous n'avons pas manqué de souligner, il a été possible d'aboutir à des hypothèses informées concernant les placements à privilégier en fonction d'un ensemble de facteurs tels que la présence de compléments non-nominaux ou les GN objets longs. Nous avons également vu que certains compromis doivent être faits afin d'adapter les règles à un public d'apprenant, en nous assurant que les corrections proposées soient aussi fiables que possible.

Le résultat de ce travail consiste en une modélisation des schémas d'erreur et de correction, pour laquelle nous avons recherché un équilibre entre précision, simplicité, et pragmatisme quant aux limites imposées par le cadre pratique. Nous avons abouti à un ensemble de quatre schémas d'erreur et de correction pour le placement d'*also*, et de trois schémas pour le placement des adverbes de manière. Parmi ces sept schémas, quatre sont dédoublés afin de prendre en compte les ajustements nécessaires à la présence d'auxiliaires dans le cas d'*also*, et à la modification de l'adverbe noyau par l'adverbe de degré *very* pour les adverbes. Les schémas fonctionnent également avec des exclusions, comme l'exclusion du verbe lexical *see*

dans les schémas intégrant *also*, puisque notre étude sur des corpus natifs a révélé la fréquence de l'expression ["See also" + GN]. Ces exclusions sont présentées de manière exhaustive lors de la présentation technique des patrons et règles dans le Chapitre 3.

Le nombre limité de schémas proposés est lié à la recherche de fiabilité dans la détection, toutes les configurations présentant un risque élévé de provoquer des détections érronées et présentant un faible intérêt pour la correction ayant été écartées. La recherche de concision est également une des raisons de leur faible nombre : à chaque fois que cela est possible, des éléments sont inclus dans les schémas de manière optionnelle, par exemple dans le cas des prépositions à la suite des verbes. Ceci permet d'éviter la multiplication de schémas, leur faible nombre constituant un avantage pour la transition vers la phase d'implémentation technique. Avant de présenter les modalités de cette transition, nous nous penchons dans la partie suivante sur le second type d'erreur traité ici, qui présente des défis encore plus conséquents.

2.3 L'utilisation des noms en fonction de dépendant dans le groupe nominal

Des études de corpus diachroniques ont montré que l'emploi des séquences N+N a fortement augmenté depuis les années 1950 (Biber et Clark, 2002 : 53), notamment dans les domaines informationnels (ex. : journaux, magazines, publications scientifiques). Cette augmentation ne semble pas encore avoir été prise en compte dans l'enseignement de l'anglais, puisqu'elle n'occupe qu'une place marginale dans le traitement des groupes nominaux, ces structures étant souvent assimilées à des noms composés (ex. : Larreya et Rivière, 2005 : 236). Cette lacune se traduit par la présence d'une proportion significative d'erreurs liées à leur emploi dans les écrits de personnes utilisatrices de l'anglais, en particulier dans les productions scientifiques.

La première sous-partie est consacrée à la présentation de la syntaxe et de l'emploi de ces structures, et se conclut sur l'identification de contraintes pesant sur leur formation, dans l'espoir de dégager des critères précis permettant d'évaluer l'acceptabilité des segments et de fournir des corrections adaptées. Ces critères sont mis à l'œuvre dans la seconde partie, qui présente l'analyse des erreurs relevées dans le corpus et révèle une grande diversité dans les formats et les types d'erreur que l'on peut identifier. Ces erreurs regroupent certains des obstacles les plus difficiles à surmonter dans le domaine du traitement des langues, comme l'interprétation sémantique ou le choix lexical ; nous verrons que si certaines erreurs bien

délimitées se prêtent à un traitement automatisé préliminaire, d'autres constituent un défi encore non résolu.

2.3.1 Syntaxe et emploi des structures N+N

Plusieurs expressions ont été utilisées pour faire référence aux suites de deux noms formant un GN, telles que "séquences N+N", "structures N+N" ou "composés nom-nom". Lorsque ces séquences sont analysées comme des constructions syntaxiques, le nom qui n'est pas en fonction de noyau peut être qualifié de "modifieur nominal", "nom épithète", ou "nom adjectival" selon les sources. Payne et Huddleston utilisent l'expression "nominal composite" pour décrire les GN incluant un nom en fonction de dépendant. Suivant l'exemple de Pastor-Gómez, auteure d'une monographie récente sur ce thème, et en raison de la neutralité relative de ces expressions, nous utilisons les termes "structure N+N" ou "séquence N+N" de manière interchangeable pour renvoyer aux GN composés d'un noyau et d'au moins un nom fonctionnant comme dépendant, et pouvant lui-même admettre des dépendants. Les abréviations N_1 et N_2 renvoient respectivement au nom en fonction de dépendant et au nom en fonction de noyau. Par ailleurs, le terme de "dépendant" est utilisé avec la définition donnée dans *CGE*, et permet de renvoyer simultanément aux noms en fonction de modifieur et de complément.

Cette partie est consacrée au passage en revue de problématiques importantes ayant trait à la syntaxe et à l'emploi des structures N+N. La première section traite du schéma global de la syntaxe interne des GN comme elle est décrite dans *ACG*, *LGSWE* et *CGE*, avec une attention particulière aux dépendants placés avant le noyau. Les arguments en faveur de la distinction entre structures syntaxiques et compositions lexicales sont introduits dans la deuxième section, tandis que la troisième relate les résultats d'études de corpus concernant l'emploi des structures N+N en anglais contemporain. La dernière section constitue une transition vers la seconde sous-partie, en identifiant les paramètres formels, syntaxiques et pragmatiques qui permettent d'évaluer l'acceptabilité d'une structure N+N donnée.

a. Syntaxe interne du groupe nominal

Comme dans le cas du placement des adverbes, la description de l'emploi et de la construction des structures N+N implique d'évoquer la syntaxe interne du GN dans son ensemble. Pour des raisons évidentes, cette partie ne vise pas l'exhaustivité, et dans la mesure du possible elle se limite à l'exposé des aspects qui sont pertinents pour l'étude des structures

N+N. Cependant, pour garantir la clarté des descriptions, un passage en revue rapide de la terminologie et des analyses du GN utilisées dans les trois ouvrages de référence s'avère nécessaire.

En effet, les termes utilisés sont semblables, mais ceux-ci n'ont pas les mêmes définitions ou les mêmes référents. Les termes clés dans la description de la syntaxe du GN sont présentés dans le tableau 37 en page suivante, accompagnés de leurs définitions d'après les cadres descriptifs d'*ACG*, *LGSWE* et *CGE*. Comme pour la syntaxe des adverbes, les descriptions de Biber et al. sont en partie calquées sur celles de Quirk et al., et celles-ci sont donc fusionnées dans la même colonne. Elles diffèrent en certains points, qui sont indiqués dans le tableau. Les cases grisées indiquent qu'un terme n'est pas utilisé dans la grammaire correspondante. Les définitions s'entendent dans le cadre de la description du GN, certains des termes étant utilisés pour faire référence au même type de fonction à d'autres niveaux syntaxiques (ex. : modifieur, complément). Des explications supplémentaires concernant les termes du tableau sont données dans les paragraphes qui suivent.

L'expression "Groupe Nominal", ou *Noun Phrase*, n'est pas incluse dans le tableau en raison de l'homogénéité des définitions qui en sont données dans les trois ouvrages. Nous adoptons le même type de définition, selon laquelle un GN est un groupe syntaxique dont le noyau est un nom, et qui peut fonctionner comme complément (sujet, objet ou complément prédicatif) dans la proposition (Huddleston, 2002 : 54).

L'inversion des référents des termes *determiner*/*determinative* dans *ACG* et *CGE*, remarquée d'ailleurs avec une certaine irritation par Leech dans sa critique de l'ouvrage de Huddleston et Pullum et al. (2004 : 130), est l'une des trois grandes différences à noter dans le cadre descriptif du GN dans ces trois grammaires. Les termes *determinative* et *determiner* sont laissés en anglais dans le tableau afin de conserver les nuances démontrées dans les grammaires, sans postuler de choix de traduction qui ne serait peut-être pas adéquat. Dans *ACG*, le terme de *determiner* est utilisé pour faire référence à la catégorie des déterminants, alors que *determinative* fait référence à la fonction de détermination dans le GN ; c'est le contraire dans *CGE*. Dans la suite du document, nous utilisons les termes "déterminant" et "détermineur" pour renvoyer à la catégorie et à la fonction respectivement, sauf lorsque nous relatons les descriptions données par Quirk et al.

	ACG* et *LGSWE	***CGE***
Nominal	Tous les constituants qui apparaissent dans des **fonctions** le plus souvent instanciées par des **GN**, mais qui ne sont pas nécessairement de nature nominale.	**Fonction** intermédiaire entre le GN et le nom noyau. (ex. : *the red apple*)
Determinative	***ACG* : Fonction** le plus souvent instanciée par les mots faisant partie de la catégorie des *determiners*.	**Catégorie** de mots instanciant le plus souvent la fonction de *determiner* dans le GN.
	Pas d'utilisation de ce terme dans LGSWE	
Determiner	***ACG* : Catégorie** de mots instanciant le plus souvent la fonction de *determinative* dans le GN.	**Fonction** le plus souvent instanciée par les mots faisant partie de la catégorie des *determinatives*.
	***LGSWE* : Catégorie** et **fonction** des mots utilisés pour la détermination dans le GN.	
Dépendant	*Pas d'utilisation de ce terme dans* LGSWE *et* ACG	Élément remplissant la **fonction** d'un constituant externe ou interne du GN. Terme regroupant les **compléments** et les **modifieurs**.
Modifieur	Élément remplissant la **fonction** de **modification** dans le GN. Tous les éléments du GN, hors noyau et *determinative,* sont considérés comme des modifieurs.	Élément remplissant la **fonction** de **modification** dans le GN. Diffère des compléments.
Complément	*Pas d'utilisation de ce terme dans* LGSWE	Élément remplissant une **fonction** de **complémentation** sélectionné par le noyau du GN, et pouvant être avant ou après celui-ci. Diffère des modifieurs.
	LGSWE : Élément remplissant une **fonction** de **complémentation** sélectionné par le noyau du GN et placé après celui-ci. Diffère des modifieurs.	

Tableau 37. Comparaison des définitions liées à la syntaxe du GN

La seconde différence à noter entre les deux cadres descriptifs est la place importante donnée à la fonction de complément dans le GN dans *CGE*, cette fonction n'étant pas distinguée de celle de la modification dans *ACG*. Ainsi, dans l'exemple suivant, le segment souligné est considéré comme un modifieur dans *ACG*, et comme un complément dans *CGE* :

[236] *Do you have [the key to her apartment]?*

La troisième différence réside en la reconnaissance dans *CGE* d'un niveau syntaxique intermédiaire entre le GN et le nom noyau, appelé "nominal". Dans les deux autres sources, ce terme désigne les constituants adoptant des fonctions généralement instanciées par les GN. Ces distinctions, ainsi que d'autres, sont abordées dans les paragraphes suivants.

D'après Quirk et al., le GN est composé de quatre types de constituant (1985 : 1238-1239) :

- le noyau,
- le déterminant (fonction), qui inclut les pré-déterminants, les déterminants centraux et les post-déterminants,

- les pré-modifieurs,
- les post-modifieurs.

Le rôle sémantique de la modification dans le GN est décrit à l'aide de deux dichotomies : la modification peut être restrictive ou non-restrictive, et temporaire ou permanente. Elle est restrictive lorsque le modifieur concerné est nécessaire à l'identification du référent du nom noyau ; elle est en revanche non-restrictive lorsque son rôle se limite à apporter des informations supplémentaires non-essentielles à l'identification du référent (op. cit. : 1239). Les modifieurs les plus restrictifs sont généralement placés après le nom noyau. Les deux exemples suivants illustrent ces deux possibilités, avec le même adjectif servant de modifieur restrictif dans le premier cas, et non-restrictif dans le second (le GN est placé entre crochets et le modifieur est souligné) :

[237] *Which vase did you throw at him? – Oh, just [the ugly one].*

[238] *I don't want him to put [his ugly nose] into my house again.*

La seconde dichotomie est également en lien avec la position des modifieurs, en ceci que les modifieurs qui sont placés avant le noyau sont envisagés comme renvoyant à des traits plus permanents, ou du moins plus caractéristiques, que ceux qui sont placés après le noyau, et qui sont associés à des caractéristiques temporaires ou créées en discours. Cette distinction est visible dans la différence de sens qui est automatiquement créée lorsqu'un adjectif en fonction de pré-modifieur d'un nom se retrouve en fonction d'attribut du sujet instancié par ce même nom (exemples tirés de Quirk et al. : 1243) :

[239] *The big toe.*

[240] *The toe is big.*

Cette caractéristique joue notamment un rôle dans le choix d'employer une structure N+N plutôt qu'une structure avec post-modification. Nous reviendrons sur cette notion.

Le schéma de base des GN donné par Biber et al. est similaire à celui de Quirk et al., à une distinction importante près : *LGSWE* inclut une différenciation entre les modifieurs et les compléments parmi les éléments qui sont placés après le noyau. Ces compléments sont le plus souvent des propositions en *that* ou *to*, comme dans l'exemple suivant (op. cit. : 575 ; le complément est souligné) :

[241] *the idea that he was completely cold and unemotional*

La description de Payne et Huddleston est encore plus éloignée de celle de Quirk et al., et repose notamment sur l'introduction du niveau du "nominal" entre le GN et le nom. Le

nominal est considéré comme le noyau du GN, le nom devenant le "noyau ultime" (*ultimate head*, op. cit., 330). L'existence du nominal est justifiée par le fait que l'on retrouve dans certaines fonctions une structure qui n'est ni un nom noyau seul, ni un GN complet puisque que le déterminant n'est pas présent. Par exemple, dans les structures N+N, le dépendant pré-noyau (*pre-head*) est considéré comme un nominal plutôt que comme un nom fonctionnant seul, puisqu'il peut lui-même être accompagné de dépendants, reflétant la distinction entre GAdj et adjectif. L'exemple suivant illustre ce type de fonction (op. cit. : 329 ; le dépendant est souligné) ; le nominal utilisé comme modifieur de *officials* est lui-même composé d'un noyau (*Ministry*) accompagné d'un dépendant (*of Defence*):

[242] *those Ministry of Defence officials*

[243] **those the Ministry of Defence officials*

La reconnaissance d'un niveau intermédiaire fait également apparaître un parallèle entre la structure du GN et la hiérarchie à trois niveaux que forment la proposition, le GV et le verbe. Par souci de simplicité, nous utilisons le terme "nom" pour renvoyer à ces deux niveaux lorsque nous évoquons sa fonction de dépendant dans les structures N+N.

Liée au concept de nominal, la dichotomie dépendants externes/dépendants internes est également une spécificité de la description de la syntaxe du GN donnée dans *CGE* ; les dépendants internes sont les constituants immédiats du nominal et non du GN. Dans le segment suivant, le nominal est indiqué entre crochet et le dépendant interne est souligné :

[244] *the [price of the other one]*

Le détermineur (fonction) est le plus important des dépendants externes ; celui-ci peut être instancié par un déterminant (catégorie) ou un GN au cas génitif. Les modifieurs placés avant le détermineur, ou pré-déterminateurs, et les modifieurs périphériques (*peripheral modifiers*) sont également des dépendants externes du GN. Les phrases suivantes donnent un exemple de chacune de ces deux fonctions ; le GN y est mis entre crochets et le modifieur est souligné (op. cit. : 331) :

[245] *It's [two thirds the price of the other one]*. (pré-déterminateur)

[246] *We couldn't manage with [the car alone]*. (modifieur périphérique)

Les dépendants internes peuvent se trouver avant ou après le nom noyau, et peuvent être en fonction de modifieur ou de complément dans ces deux positions. La distinction entre

modifieur et complément dans le GN est une autre spécificité du cadre descriptif de *CGE*. Celle-ci est cependant difficile à identifier, du propre aveu des auteurs (op. cit. : 439-440) :

> The distinction between these two kinds of dependent is essentially the same as in clause structure, but in the NP they are not as clearly differentiated syntactically. Note in particular that while complements of a verb are not infrequently recognisable as such by virtue of being obligatory, the contrast between obligatory and optional elements is of virtually no relevance to distinguishing complements from modifiers in NP structure.

Payne et Huddleston proposent donc de distinguer les compléments des modifieurs dans le GN sur la base de huit critères, inspirés des distinctions identifiées dans le GV (op. cit. 440-443). Nous regroupons les trois derniers critères, liés à des aspects sémantiques, pour obtenir les cinq critères suivants :

- la sélection par le nom noyau : les compléments qui sont placés après le noyau contiennent une préposition dont la sélection dépend du nom ; la préposition *of* est l'option par défaut, et elle est également souvent la seule option possible,
- la portée de l'anaphore : généralement, on ne peut remplacer le nom noyau par *one* si celui-ci est accompagné d'un complément, mais c'est possible s'il est accompagné d'un modifieur (op. cit. : 441) :

  [247] *?I've told my history tutor, but I can't find [my <u>French</u> one]*. (complément)

  [248] *I don't want a British nanny, I want [a <u>French</u> one]*. (modifieur)

- corrélation avec la catégorie : les adjectifs sont généralement utilisés en fonction de modifieur, alors que les noms sont souvent en fonction de complément,
- mobilité : les compléments sont placés le plus proches possibles du nom noyau, après tous les modifieurs éventuels,
- les compléments expriment des arguments sémantiques du nom noyau ; en conséquence, les rôles sémantiques des compléments sont sélectionnés par le nom noyau, et font l'objet de restrictions.

Les GN suivants illustrent la différence entre ces deux fonctions (op. cit. : 442, 446, 439, 444) :

[249] *the rays <u>of the sun</u>* (complément)

[250] *a woman <u>of great wisdom</u>* (modifieur)

[251] *a <u>legal</u> adviser* (complément)

[252] *many very angry farmers* (modifieur)

Les deux premiers exemples illustrent la distinction entre complément et modifieur à la suite du nom noyau, les dépendants prenant la forme d'un GP ; les deux autres exemples montrent quant à eux cette distinction pour les dépendants placés avant le noyau et prenant la forme d'un adjectif. Les adjectifs sont cependant rarement utilisés comme complément ; les compléments placés avant le noyau sont généralement instanciés par des noms ou nominaux, comme dans les deux GN suivants (op. cit. : 439) :

[253] *a linguistics student*

[254] *an income tax adviser*

Les noms ou nominaux peuvent cependant aussi être modifieurs (op. cit. : 444) :

[255] *its entertainment value*

[256] *those Egyptian cotton shirts*

Le tableau 38 constitue une synthèse de la structure du GN élaborée à partir des descriptions de Payne et Huddleston. Les catégories pouvant instancier les fonctions concernées sont indiquées dans la dernière ligne. Cette présentation synthétique n'est bien sûr pas exhaustive, certains aspects ayant été laissés de côté, tels les compléments indirects (ex. : *a larger galaxy than initial measurements suggested*, op. cit. : 443). Le format du tableau ne permet pas de présenter plusieurs modifieurs à la suite, mais ceci est bien entendu possible en réalité. Par ailleurs, les constituants des GN utilisés comme illustration sont le plus souvent simples, alors qu'ils pourraient admettre des dépendants. En vertu de son degré de détail, le cadre descriptif de *CGE* est adopté dans la suite des analyses. La terminologie utilisée est en cohérence avec ce choix de cadre.

GROUPE NOMINAL									
			NOMINAL (noyau intermédiaire)						
Dépendants pré-noyau					**NOM (noyau ultime)**	**Dépendants post-noyau**			
Dépendants externes			**Dépendants internes**			**Dépendants internes**			**Dép. externe**
Mod. périph.	**Pré-déterm.**	**Déterminateur**	**Modifieur**	**Complément**		**Complément**	**Modifieur**	**Mod. apposé**	**Mod. périph.**
even	all	the	preposterous		salary	from Lloyd's	that Bill gets		
		the school's	welcome		ban	on smoking			
only		his	famous		opera			'Carmen'	
		the		public relations	adviser		from Harvard		herself
	both	those	cotton		dresses		on the shelf		
	half	the		tax	year			of 2004	
		an	accurate		depiction	of the situation			
Adv., GP, Pronom réfl.	*Déterminant, GN, GAdj*	*GDét, GN au génitif*	*GAdj, GV, Nominal*	*Nominal, Adj.*	*Nom commun, Nom propre*	*GP, Proposition*	*GDét, GAdj, GN, GP, Proposition*	*Nom propre, GP avec* of	*Adv., GP, Pronom réfl.*

Tableau 38. Composition du groupe nominal d'après CGE

b. Le statut des structures N+N : noms composés ou nominaux composites ?

D'après *CGE,* les noms, ou plus précisément les nominaux, peuvent être employés comme dépendant interne pré-noyau dans le GN, en fonction de complément ou de modifieur (voir exemples ci-dessus). Leur utilisation dans le GN semble donc relever de la construction syntaxique, à l'instar de l'utilisation des adjectifs dans les mêmes fonctions. Selon cette approche, l'expression *chocolate pudding* est un GN composé d'un nom noyau modifié par un autre nom. Néanmoins, l'existence d'expressions telles que *ice-cream* vient mettre en doute cette analyse : *ice-cream* est reconnu comme un nom composé, dans lequel le terme *ice* n'est pas un modifieur mais une base (Payne et Huddleston, 2002 : 448). Quelle est donc la différence entre *ice-cream* et *chocolate pudding* ? Les séquences N+N ne devraient-elles pas plutôt être analysées comme des créations lexicales ?

En dehors de son intérêt linguistique, cette question est particulièrement importante lorsque cette problématique est abordée avec l'objectif de pouvoir détecter et corriger les erreurs liées à l'utilisation des structures N+N. Si ce type de séquence relève avant tout de la création lexicale, alors il est envisageable, en théorie, de détecter les erreurs en les confrontant à un lexique de noms composés. En pratique, il suffit d'observer un corpus d'anglais contemporain pour se rendre compte que cette solution est une fausse piste, ou du moins, qu'elle est loin d'être un raccourci. En effet, la fréquence des GN comportant une séquence de deux noms ou plus dans les productions écrites est extrêmement élevée, tout comme leur diversité. Ces deux observations mettent en doute d'une part la faisabilité de la création d'un tel lexique, et d'autre part le statut purement lexical des séquences N+N.

Deux problématiques sont à différencier : la question de l'origine, lexicale ou syntaxique, des séquences N+N dans leur ensemble, et l'identification du statut, GN complexe ou nom composé, d'une séquence donnée en anglais contemporain. Même si la première de ces problématiques est évoquée dans les paragraphes qui suivent, c'est plutôt la seconde qui attire notre attention dans cette section, en vertu des conséquences que la définition de leur statut peut avoir sur la détection des erreurs qui sont liées à l'emploi de ces séquences. Il va sans dire que nous n'épuisons ici ni le sujet, ni les études existantes le concernant. L'ambiguïté qui existe entre noms composés et structures syntaxiques concerne plus particulièrement les cas où le premier terme fonctionnerait comme complément dans le GN concerné, si la structure était analysée comme telle (Pastor-Gómez, 2011 : 47) ; les noms étant la catégorie la plus

fréquente dans cette fonction, cette indication ne permet cependant pas une réduction significative du champ d'étude.

Pour ce qui est de l'origine des séquences N+N, toutes les approches semblent coexister. Certains linguistes, comme Benczes, partent du principe que toutes ces séquences sont générées en lexique, considérant donc la séquence *university teaching award committee member* comme un nom composé (2006 : 7). D'autres, comme Pastor-Gómez, adoptent l'approche inverse, selon laquelle ces séquences sont générées syntaxiquement, et subissent dans certains cas un processus de lexicalisation (2011 : 14). Pour d'autres encore, les séquences N+N ont une origine soit lexicale, soit syntaxique selon certaines caractéristiques, notamment le type de relation sémantique qu'entretiennent les termes (Giegerich, 2004 : 1).

Aucune supposition n'est faite explicitement concernant l'origine de ces séquences dans les trois grammaires de référence. Quirk et al. (1330), Biber et al. (590) et Payne et Huddleston (451) s'accordent sur la coexistence en anglais contemporain de noms composés comportant une séquence de deux noms et de GN incluant un dépendant instancié par un nom. Ils notent également la difficulté de les distinguer et l'ambiguïté du statut de certaines séquences. Payne et Huddleston introduisent le terme de "nominal composite" pour faire référence aux structures N+N générées syntaxiquement (op. cit. : 448). Sans surprise, la distinction la plus stricte entre ces structures et les noms composés est à trouver dans *CGE*, les deux autres ouvrages adoptant une approche moins tranchée de cette problématique. Par exemple, *LGSWE* mentionne l'existence d'un continuum (*cline*, op. cit. : 590) entre ces deux types de séquence. Payne et Huddleston préférent à ce concept celui de "cas limite" (*borderline cases*, op. cit. : 451). Comme indiqué dans l'introduction à la présente partie, nous utilisons l'expression "structure N+N" pour faire référence exclusivement aux structures syntaxiques [Dépendant interne + Nom noyau] de type *chocolate pudding*.

Identifier la nature de ces séquences implique de faire appel à des critères et tests précis. Un ensemble de ceux-ci est utilisé dans les trois grammaires comme dans d'autres travaux consacrés plus particulièrement à ces séquences. L'un des critères les plus utilisés est celui de l'accentuation : les composés sont généralement accentués sur le premier terme, et les structures syntaxiques sont accentuées plus fortement sur le nom noyau (le premier nom reçoit alors l'accentuation standard d'un dépendant). Ce critère est mentionné par Quirk et al. (op. cit. : 1330), qui indiquent que les conditions d'attribution des différents schémas d'accentuation sont liées au degré d'"institutionalisation" d'une séquence. Nous revenons sur ce concept et sur celui de la lexicalisation dans notre présentation des travaux de Pastor-

Gómez. Par ailleurs, le critère de l'accentuation a été utilisé dans l'étude de Giegerich mentionnée plus haut, et dont la conclusion est que les séquences N+N dans lesquelles le N_1 est un complément sont générées lexicalement, alors que celles dont le N_1 est un modifieur sont construites syntaxiquement (2004 : 1). Nous reviendrons sur la distinction entre N_1 modifieur et N_1 complément dans la seconde sous-partie.

Biber et al. mentionnent le critère de l'accentuation mais préfèrent se fier au critère de la séparation orthographique des séquences, partant de l'observation selon laquelle les noms composés les plus figés ne forment qu'un seul mot (ex. : *seaweed*), ou bien sont liés par un tiret (ex. : *ice-cream*), alors que les structures N+N sont présentées en plusieurs mots distincts (op. cit. : 590).

Payne et Huddleston distinguent les critères syntaxiques des critères non-syntaxiques. Ces derniers sont au nombre de quatre et incluent les deux critères que nous venons d'évoquer (accentuation et orthographe), les deux autres étant la transparence sémantique et la productivité. Ainsi, les structures N+N sont compositionnelles et transparentes, alors que les composés sont souvent non-compositionnels, ou bien adoptent une signification spécifique (par exemple, le mot composé *blackbird* fait bien référence à un type d'oiseau de couleur sombre, mais par le biais de la référence à une espèce et non à un individu, tous les individus de cette espèce n'étant pas de couleure noire). Le critère de productivité concerne quant à lui le fait que le N_1 d'une structure N+N peut être remplacé par un adjectif compatible d'un point de vue sémantique, alors que la forme d'un composé est figée (op. cit. 451).

Les critères syntaxiques concernent les possibilités de coordination et de modification des termes de la séquence en question (exemples tirés de *CGE*, op. cit. : 449, 451) :

[257] *London and Oxford colleges*

[258] **ice- and custard-creams*

[259] *south London colleges*

[260] **crushed ice-cream*

Les deux premiers exemples illustrent l'impossibilité pour le premier terme du nom composé *ice-cream* d'entrer en coordination avec un autre terme ; les deux GN suivants sont des exemples du même type d'impossibilité appliqué à la modification. Dans les nominaux composites, les deux éléments sont des constituants d'un GN, et peuvent donc chacun être modifiés ou coordonnés à un autre terme selon les règles standards de la syntaxe interne du GN. Ceci n'est pas le cas des éléments d'un nom composé, qui, comme nous l'avons

mentionné, sont considérés comme des bases. Les auteurs indiquent cependant que les critères non-syntaxiques ne sont pas toujours en cohérence avec les critères syntaxiques ; considérant que la structure à délimiter est une construction syntaxique, Payne et Huddleston donnent la primauté à ces critères lorsque les résultats divergent (op. cit. : 451).

Pastor-Gómez consacre une partie de sa monographie sur les structures N+N à la frontière entre syntaxe et morphologie dans l'identification de leur statut, et passe en revue l'essentiel des travaux concernant cette distinction, examinant chacun des critères que nous venons de citer. Confrontée à certaines incohérences dans les résultats obtenus à l'application de ces critères, elle aboutit à la conclusion suivante, qui rejoint celle de Biber et al. (2011 : 88) :

> The only solution is to assume that the distinction between morphological compounds and free syntactic phrases is not binary but gradient. If this is a valid assumption, then we would be able to explain why some N+N structures satisfy some criteria but not others.

La problématique est réorientée vers l'évaluation du degré d'institutionnalisation et de lexicalisation des structures N+N. Les critères utilisés pour cette évaluation sont une synthèse de ceux que nous avons déjà mentionnés. Pastor-Gómez les regroupe en deux catégories (op. cit. : 150) :

- les critères principaux : acceptation de la modification, acceptation de la coordination et transparence sémantique;
- les critères annexes : accentuation et orthographe.

Sur un total de 7466 séquences relevées dans un corpus d'anglais écrit et oral de 600 000 mots, près de 95 % des structures N+N sont identifiées comme non-lexicalisées (op. cit. : 151). Plus de détails sont donnés sur les résultats de l'étude de corpus menée par Pastor-Gómez dans la section suivante.

c. L'emploi des structures N+N

Le choix d'utiliser un nom en tant que dépendant interne pré-noyau plutôt que post-noyau fait disparaître les éléments grammaticaux, tels que les prépositions, qui indiquent le type de relation liant le noyau au dépendant (Pastor-Gómez, 2011 : 37 ; Biber et al., 1999 : 590). On peut voir cette différence dans les deux GN suivants :

[261] *Cooking apples*

[262] *Apples for cooking*

Ceci peut avoir pour résultat une perte de clarté, et faire basculer dans l'implicite une partie du sens transmis par le GN, posant un risque d'interprétation erronée (Quirk et al., 1985 : 1330). Leur utilisation constitue une surcharge cognitive pour la ou le destinataire du message, à qui il revient de déduire la relation qui unit les termes du GN (Biber et al., 1999 : 593). En revanche, ce type de structure permet de présenter l'information sous une forme condensée, nécessitant moins de mots qu'un dépendant post-noyau prenant la forme d'un GP ou d'une proposition. Ces deux conséquences sont applicables à tous les éléments fonctionnant comme dépendant pré-noyau, mais elles sont particulièrement apparentes dans le cas des structures N+N (Biber et al., 1999 : 588). L'utilisation de ces structures repose donc sur un équilibre subtil entre deux injonctions communicatives : la précision du message d'une part et l'efficacité de son expression d'autre part (Biber et al., 1999 : 590).

D'après Pastor-Gómez, le choix d'utiliser un dépendant pré-noyau plutôt que post-noyau dépend avant tout de la sélection d'objectifs de communication différents, ces deux structures n'ayant pas la même fonction communicative. Dans un contexte journalistique, la récursivité des structures N+N permet de délivrer des "paquets d'information compacts" (*compact packages of information*, op. cit. : 40). Elles véhiculent également un ton neutre et impersonnel, donnant l'impression que les informations relayées sont objectives. (op. cit. : 41). En dehors de leur intérêt en termes d'économie de langage, elles ont également un intérêt spécifique pour les textes en langues de spécialité. Comme il a déjà été mentionné, les dépendants pré-noyau tendent à faire référence à des caractéristiques permanentes et statiques ; cette propriété est mise à l'œuvre dans l'utilisation des structures N+N dans les domaines scientifiques, car elles permettent d'une part de faire apparaître des classes (exemples tirés de notre corpus : *music detection*, *jingle detection*, *speech detection*), et d'autre part d'amorcer ou de perpétuer la standardisation de la terminologie utilisée (Varantola, 1993, cité par Pastor-Gómez, 2011 : 41). L'utilisation des noms plutôt que des adjectifs dans cette fonction est avantageuse car elle permet de faire référence à un ensemble de caractéristiques propres à ce nom, et qui ne sont pas nécessairement disponibles sous la forme d'un adjectif. Le rôle global des noms dans cette fonction est présenté de manière synthétique par Pastor- Gómez (op cit. : 47) :

> Nouns as dependent accompany the head of the noun phrase, providing it with those specific qualities that are singular to the given entity. They specify and characterize the head noun, and by providing additional information become essential ingredients in the process of exchanging information within a given communicative process.

Ainsi, les résultats des analyses de corpus de *LGSWE* indiquent que l'utilisation de dépendants pré-noyau est particulièrement fréquente dans les corpus d'articles de journaux et d'écrits scientifiques et universitaires, avec environ 30 % d'occurrences en plus que dans le corpus de fiction. Après celle de l'adjectif, la catégorie du nom est la deuxième catégorie la plus fréquemment observée dans cette fonction, représentant 40 % des occurrences dans les textes journalistiques et 30 % dans les écrits universitaires, soit deux à trois fois plus que dans les textes de fiction (op. cit. : 589). Pastor-Gómez, qui compare des corpus de textes narratifs (littérature) et non-narratifs (presse, productions scientifiques) de taille égale, relève également une proportion plus importante de structures N+N dans ces derniers, avec 843 occurrences dans le sous-corpus de fiction, contre près du double dans le corpus d'écrits scientifiques (1592 occurrences), et du triple dans le corpus de reportages (2345 occurrences) (op. cit., 2009 : 151). Les résultats de cette étude montrent également que l'utilisation des structures N+N est plus fréquente dans les écrits en variété d'anglais américain que d'anglais britannique (op. cit. : 156).

Comme il a déjà été évoqué, Pastor-Gómez a différencié les structures N+N lexicalisées des structures non-lexicalisées. Sur l'ensemble des corpus utilisés, seulement 5 % de ces structures remplissent les critères de lexicalisation énoncés. Ce pourcentage est encore plus bas dans le corpus d'écrits scientifiques, pour lequel il descend à 1,88 %. En conséquence, on peut s'attendre à ce que la majorité des segments erronés relevés dans notre corpus concerne des structures N+N non-lexicalisées.

D'après Biber et al., certains noms se révèlent très productifs dans la construction de séquences N+N, pouvant être retrouvés en fonction de dépendant en combinaison avec plus de cent noms noyaux différents. Pour le corpus d'écrits universitaires et scientifiques, ces noms sont les suivants : *cell, class, community, computer, family, government, group, information, language, library, research, school, state, surface, system, test, time, work*. Les combinaisons incluant ces noms sont également fréquentes, avec plus de cinquante occurrences par million de mots (op. cit. : 593). Cependant, on peut supposer que la présence de certains noms plutôt que d'autres dépend principalement du domaine scientifique concerné.

Ces résultats et observations viennent apporter une explication au pourcentage relativement élevé d'erreurs liées aux structures N+N rencontrées dans notre corpus. Ce pourcentage est surprenant au premier abord, puisque, à notre connaissance, ce type d'erreur ne fait pas l'objet d'études spécifiques en TAL ou ELAO, et n'apparaît pas dans les listes d'erreurs courantes qui sont régulièrement constituées (cf. Leacock et al., 2010 : 15-18). Il est également possible que

l'invisibilité de ces erreurs soit attribuable à l'utilisation de systèmes de catégorisation différents, certaines de ces erreurs pouvant être incluses dans des catégories "parapluies", telles que l'utilisation des collocations, les erreurs lexicales ou la construction des GN.

L'une des originalités de notre recherche est le recours à un corpus d'utilisation de l'anglais à partir de sources variées plutôt qu'à un corpus d'interlangue. Celui-ci intègre des écrits scientifiques, représentant 40 % du corpus global. Il est donc probable que la présence d'erreurs liées aux structures N+N soit attribuable simplement à l'utilisation plus fréquente de ces structures dans le type de production qui constitue une part importante de notre corpus. On retrouve par ailleurs une proportion bien plus faible dans les productions tirées du corpus *ICLE*, avec seulement 2,9 % de l'ensemble des erreurs relevées dans ce sous-corpus, contre 15,7 % dans le sous-corpus d'articles scientifiques. Nous donnons plus de détails sur la distribution des erreurs dans la sous-partie suivante.

Du point de vue de l'acquisition des langues, ces erreurs diffèrent des erreurs liées au placement des adverbes, qui peuvent être en partie expliquées par l'existence d'un transfert du français à l'anglais. Les erreurs liées aux structures N+N semblent être la conséquence d'au moins deux phénomènes. Le premier concerne la surgénéralisation des conditions d'utilisation des structures N+N. Après avoir été confrontée à de nombreux exemples de structures N+N, celles-ci étant fréquentes en anglais comme on a pu le voir, une personne apprenante peut procéder par analogie et en déduire que ces structures sont acceptables, mais sans en maîtriser toutes les contraintes. Les erreurs peuvent donc être liées à un manque d'informations données aux personnes apprenantes quant aux règles sémantiques, syntaxiques et pragmatiques qui régissent la construction et l'utilisation des structures N+N.

Le second phénomène est celui de la simplification, qui peut également être envisagée comme une stratégie d'évitement. D'un point de vue formel, la construction d'une séquence N+N est d'une grande simplicité, puisqu'il suffit d'associer deux noms ; en revanche, la construction d'un GN comprenant un dépendant post-noyau implique le plus souvent de sélectionner une préposition, et on sait les difficultés que cette tâche pose aux personnes apprenantes et utilisatrices de l'anglais.

Ces deux règles peuvent également entrer en jeu simultanément : le souhait d'éviter d'utiliser une préposition dans un dépendant mène à la surproduction de ces structures, et le manque d'informations quant à leurs contraintes d'emploi a pour résultat la production de segments erronés. Ces contraintes sont en effet difficiles à identifier à l'œil nu pour qui n'a pas l'anglais comme L1 ; la section suivante propose d'en délimiter les contours.

d. Règles de construction des structures N+N

Nous avons identifié trois types de paramètres à prendre en compte dans la construction des structures N+N, et donc dans la détection des erreurs qui leur sont associées. Ceux-ci sont d'ordre formel, sémantique et pragmatique. Ils sont examinés l'un après l'autre dans les paragraphes qui suivent.

En ce qui concerne la forme des structures N+N, un nombre important de noms en fonction de dépendant avant le noyau peut rendre une structure inacceptable. D'un point de vue strictement syntaxique, le GN anglais est récursif à l'infini, et peut donc admettre un nombre infini de dépendants pré-noyau, admettant eux-mêmes des dépendants, dans la limite de ce que permet la syntaxe. En pratique, nous avons vu que l'utilisation de ces structures alourdit la charge cognitive du lecteur ou de la lectrice. Ce phénomène ne peut qu'être amplifié par l'ajout de noms dépendants supplémentaires, qu'ils dépendent du noyau ou d'un autre dépendant, puisque le nombre de relations sémantiques possibles est multiplié à chaque ajout (Biber et al., 1999 : 598). De telles structures nécessitent également une analyse des relations syntaxiques dans le GN de la part de la personne destinataire (complémentation, modification empilée, modification enchâssée).

Ainsi, si l'expression *university teaching award committee member*, citée plus haut, est acceptable d'un point de vue sémantique et syntaxique, il y a fort à parier qu'il aura fallu à nos lecteurs et lectrices quelques secondes de réflexion pour en identifier un sens plausible. Par conséquent, parmi tous les GN comprenant au moins un dépendant pré-noyau dans le corpus de *LGSWE*, les GN comprenant trois ou quatre de ces dépendants ne représentent que 2 % du total. Les GN avec deux dépendants sont présents à hauteur de 20 %, et ceux ne comprenant qu'un seul dépendant sont en majorité, avec environ 80 % des occurrences (op. cit. : 597).

Lorsque plusieurs dépendants pré-noyau sont présents, l'ordre de ces éléments est également un paramètre à prendre en compte. Comme nous l'avons vu, les noms sont le plus souvent en fonction de complément du noyau ; en raison du lien étroit existant entre noyau et complément en anglais, les éléments dans cette fonction doivent être placés juste avant le nom, selon une contrainte rigide (Payne et Huddleston, 2002 : 452). Dans les deux GN suivants, l'adjectif est en fonction de modifieur du nom *detection*, et le nom est complément :

[263] *accurate music detection*

[264] **music accurate detection*

Dans le premier exemple, le complement est placé à proximité du nom, et le modifieur adopte la position la plus éloignée du noyau. Dans le second, cet ordre est inversé, et l'éloignement du complément et du nom noyau rend le GN agrammatical.

Dans les deux exemples qui suivent, les deux dépendants sont des modifieurs du nom noyau ; l'inacceptabilité relative du second segment provient cette fois-ci de contraintes labiles selon lesquelles, dans le présent cas, les modifieurs indiquant le matériau précèdent ceux indiquant le type du référent (op. cit. : 453). Le second exemple peut cependant être jugé acceptable selon les contextes :

[265] *silk evening dress*

[266] *?evening silk dress*

Malgré le fait qu'ils ne documentent pas la fonction de complémentation des noms dépendants, Quirk et al. indiquent que ceux-ci sont généralement placés immédiatement avant le noyau, dans la zone réservée aux dépendants "les moins adjectivaux et les plus nominaux", rejoignant ainsi les descriptions de Payne et Huddleston par un autre chemin (op. cit. : 1339)

Le dernier paramètre formel qui sera évoqué ici concerne l'emploi d'un N_1 au pluriel, comme dans les exemples suivants, extraits de *LGSWE* (op. cit. : 595) :

[267] *the customs officer*

[268] *a baked beans can*

D'après Quirk et al., les N_1 ont normalement un nombre "neutre", c'est-à-dire superficiellement singulier (op. cit. : 1333). Biber et al. nuancent cette affirmation, indiquant qu'il est possible de trouver des N_1 au pluriel, mais que leur distribution est beaucoup plus limitée et leur présence plus rare que celles des N_1 au singulier. Ils sont ainsi bien plus fréquents dans les productions journalistiques, et sont utilisés plus couramment en anglais britannique qu'en anglais américain (op. cit. : 594), une tendance qui est déjà soulignée par Quirk et al. (op. cit. 1333). Leur présence dans les écrits de presse peut s'expliquer par le recours fréquent à des GN incluant des dépendants pré-noyau nombreux et complexes ; l'utilisation de N_1 au pluriel permettrait ainsi de guider l'interprétation du lectorat (op. cit. : 595).

Biber et al. identifient cinq facteurs ayant une influence sur l'utilisation des N_1 au pluriel :

- le sens associé au N_1 au pluriel est différent de celui du N_1 singulier ; c'est le cas de l'exemple [267] ci-dessus,

- le N_1 est lui-même complexe, et le noyau est normalement au pluriel ; c'est le cas de l'exemple [268] ci-dessus,
- le N_1 ou bien le GN en entier est un nom propre :

  [269] *the FBI Exhibits Section*
- le N_1 est une citation :

  [270] *the "operations" group*
- le N_1 fait partie d'un titre d'article de journal :

  [271] *Armagh car-parts theft.*

Les facteurs ci-dessus s'appliquent plus particulièrement aux séquences N+N relevées dans les productions journalistiques. Quirk et al. citent deux facteurs supplémentaires qui semblent décrire les types d'utilisation de noms au pluriel fréquemment rencontrés dans les productions scientifiques (op. cit. : 1334) :

- l'utilisation d'un nom pluriel dénote la variété des items auxquels le nom dépendant fait référence, en particulier si ce nom est un hyperonyme :

  [272] *appliances manufacturer*
- une séquence N+N avec un nom pluriel est utilisée après l'introduction du concept par le biais d'un GN avec des dépendants post-noyau, la relation sémantique étant ainsi institutionnalisée de manière temporaire :

  [273] *the idea of levels*

  [274] *the levels idea*

Dans ce second cas, on note que les exemples donnés par Quirk et al. ont pour noyau des noms devant être complétés par un "contenu" (ex. : *issue, idea, concept, affair*).

Le type de relation sémantique pouvant être exprimé par le biais d'une structure N+N est également un des facteurs importants du jugement d'acceptabilité d'un tel segment. Quirk et al. relient cette problématique à celle de la permanence du trait exprimé par le N_1 (op. cit. : 1331) :

> One noteworthy constraint against using nouns from postmodifying phrases as premodifiers is the relative impermanence of the modification in question. […] However, this is not a property of the lexical item […], but of the semantic relation.

La description et l'analyse des relations sémantiques qui sont exprimées dans les structures N+N a été le point focal de la première tranche des recherches consacrées à ces structures dans la deuxième moitié du 20ème siècle (Pastor-Gómez, 2011 : 107). Elles y étaient alors le plus souvent considérées comme des composés. Parmi les travaux les plus influents sur ce thème figurent ceux de Downing (1977), Levi (1978) et Warren (1978).

Tout en remarquant que les structures N+N sont utilisées pour exprimer une gamme déconcertante de relations sémantiques différentes, Biber et al. proposent un ensemble de quinze relations, qui nous reproduisons dans le tableau 39 (tous les exemples sont extraits de *LGSWE*, op. cit. : 590-591).

Relation	**Définition**	**Exemple**
Composition	N_2 est fait à partir de N_1 ; N_2 est constitué de N_1	[275] *word classes*
But	N_2 est fait dans l'objectif de N_1; N_2 est utilisé pour N_1	[276] *chess board*
Identité	N_2 a le même référent que N_1, mais le classifie en termes d'attributs différents	[277] *women algebraists*
Contenu	N_2 est à propos de N_1; N_2 concerne N_1	[278] *interest group*
Source	N_2 provient de N_1	[279] *Pentagon proposals*
Objet, type 1	N_1 est l'objet du procès décrit par N_2, ou de l'action effectuée par l'agent décrit par N_2	[280] *computer users*
Objet, type 2	N_2 est l'objet du procès décrit par N_1	[281] *substitute forms*
Sujet, type 1	N_1 est le sujet du procès décrit par N_2; N_2 est dérivé d'un verbe intransitif	[282] *eye movement*
Sujet, type 2	N_2 est le sujet du procès décrit par N_1	[283] *labor force*
Temps	N_2 se trouve/se produit au moment indiqué par N_1	[284] *summer conditions*
Lieu, type 1	N_2 se trouve/se produit à l'endroit indiqué par N_1	[285] *Paris conference*
Lieu, type 2	N_1 se trouve à l'endroit indiqué par N_2	[286] *notice board*
Institution	N_2 identifie une institution consacrée à N_1	[287] *insurance companies*
Partition	N_2 identifie une partie de N_1	[288] *cat legs*
Spécialisation	N_1 identifie une aire de spécialisation pour la personne ou le métier décrit en N_1 (N_2 est animé)	[289] *Education Secretary*

Tableau 39. Relations sémantiques dans les structures N+N d'après LGSWE

Les auteurs précisent d'une part qu'un même GN peut être classé dans plusieurs catégories, et d'autre part que quantité de structures N+N ne correspondent à aucune de ces relations sémantiques. Ces catégories s'appliquent sans distinction aux noms en fonction de modifieur

comme de complément. On peut également souligner qu'en raison de la récursivité existant dans le GN, plusieurs relations sémantiques peuvent exister à l'intérieur d'une même structure N+N intégrant plusieurs noms en position pré-noyau.

Le dernier critère d'acceptabilité des structures N+N est celui de leur recevabilité pragmatique. Une des caractéristiques de ce type de GN est le fait que leur bonne compréhension par un lectorat dépend du partage de connaissances implicites permettant d'assembler correctement le puzzle linguistique qu'elles constituent. Une structure N+N présentant un contenu que la personne destinataire ne peut interpréter sera jugée comme irrecevable. On peut s'assurer de l'existence des connaissances nécessaires à l'interprétation correcte d'une telle structure par le biais de plusieurs stratégies. Le plus souvent, les structures N+N font appel à des connaissances universelles ou culturelles largement partagées ; l'exemple [272] ci-dessus, *Pentagon proposals*, repose sur des connaissances concernant la nature de l'organe de gouvernement américain appelé Pentagone, et sur sa fonction de conseil et de prise de décision.

La stratégie qui est souvent à l'œuvre dans les écrits scientifiques est du même ordre, mais a recours à des connaissances précises liées au domaine de spécialité de la production ; celles-ci ont de grandes chances d'être partagées par le lectorat de cette production. Les écrits journalistiques font quant à eux appel à des connaissances culturelles partagées mais éphémères, ayant une pertinence sur un temps court : "[N]ews relies primarily on premodifying nouns from those semantic domains most commonly associated with current events (such as government, business, education, the media, and sports)" (Biber et al. 1999 : 593). Une troisième stratégie consiste à créer de toute pièce une référence partagée au cours du texte, en employant des mentions explicites du même item en amont de l'utilisation d'une structure N+N, créant un réseau de références comparable à celui qui sous-tend l'utilisation des pronoms dans un texte. Ainsi, comme l'indique Pastor-Gómez, le choix d'utiliser une structure N+N est effectué avec une attention particulière au contexte de son emploi (op. cit. : 38).

Que ce soit au niveau de leur statut, de leur interprétation, ou de leur construction, les questionnements autour des structures N+N révèlent une complexité que leur apparente simplicité de forme ne laisse pas soupçonner. Nous avons confronté les descriptions et analyses de la syntaxe interne du GN données dans les trois grammaires de référence, en nous attardant plus particulièrement sur les dépendants situés avant le noyau. Cette synthèse aboutit à une représentation synoptique des principaux éléments du GN, fondée plus particulièrement

sur le cadre de *CGE*. La question du statut des séquences N+N a été abordée, celle-ci ayant une influence sur le choix de règles à adopter pour le traitement automatique de ces séquences. Il apparait que, malgré l'existence de critères rigoureux permettant en théorie de distinguer composés lexicaux et constructions syntaxiques, les séquences N+N dans leur ensemble se prêtent mal à la séparation en deux groupes distincts. Ces structures seraient plutôt à situer sur un gradient entre les deux pôles de la syntaxe et du lexique, un petit nombre d'entre elles devenant progressivement lexicalisé.

L'apparente simplicité des structures N+N, ajoutée à l'intérêt de leur utilisation dans des contextes où la densité du contenu informationnel et la concision du message sont des caractéristiques recherchées, semble avoir eu pour conséquence l'augmentation de leur utilisation depuis la moitié du 20ème siècle (cf. Pastor-Gómez, 2011). Elles sont particulièrement fréquentes dans les textes non-narratifs, que les personnes utilisatrices de l'anglais sont le plus susceptibles de produire (cf. 1.2.1b). Par ailleurs, les contraintes pesant sur leur construction et leur utilisation sont difficiles à identifier, puisqu'elles relèvent souvent plus de tendances que de règles de grammaire strictes. Nous avons cependant tenté de clarifier les règles de construction formelles, sémantiques et pragmatiques qui les régissent. Du point de vue formel, il semble que les structures comportant plus de deux noms dépendants soient problématiques. Les relations sémantiques pouvant exister entre les deux noms ont fait l'objet de plusieurs études, mais l'ensemble de celles-ci est très large et ne fait pas encore consensus.

Si la capacité des structures N+N à exprimer une large gamme de relations sémantiques constitue un avantage lors de la production de textes, elle pose des difficultés pour les personnes non-natives ainsi que pour la correction automatisée. D'après les résultats de notre étude de corpus, ces différents paramètres ont pour résultat une présence significative d'erreurs, résultats de la rencontre de règles d'évitement et d'un phénomène de surgénéralisation. La sous-partie suivante opère une catégorisation de ces erreurs, et amorce une réflexion concernant les meilleures solutions de correction à apporter.

2.3.2 Détection et correction des erreurs liées aux structures N+N : solutions et perspectives de recherche

En accord avec la méthode mixte que nous avons présentée dans la première partie de ce chapitre, la création de règles de détection et correction des séquences N+N est informée par l'analyse linguistique et grammaticale de la structure concernée, et par l'observation des erreurs produites. Les séquences N+N ont été beaucoup moins étudiées que le placement des

adverbes dans le domaine de l'acquisition des langues, et font preuve d'une très grande diversité formelle et lexicale ; pour ces deux raisons, les outils expérimentaux que nous avons utilisés pour l'étude des erreurs de placement d'adverbes ne peuvent être employés dans cette partie.

La première section présente une analyse classique des erreurs, proposant une étude de leurs formats et une classification en cinq catégories. Dans la deuxième section, nous proposons des règles de détection et correction pour deux de ces catégories, ainsi que des pistes de recherche pour les trois autres. On observe que le traitement des erreurs dans l'emploi des structures N+N est un concentré des principales problématiques actuelles dans le domaine de la correction automatisée.

a. Modélisation des erreurs liées aux structures N+N

Pour commencer, le tableau suivant rappelle la distribution des erreurs de ce type dans notre corpus ; la colonne de droite indique le pourcentage de ces erreurs par rapport au total des erreurs relevées dans le sous-corpus concerné :

Sous-corpus	**Nbr.**	**% Total**
Publications	33	15,7
ICLE	4	2,9
Courriels	4	1,5
Rapport	4	9,1
Total erreurs N+N.	**45**	**6,8**

Tableau 40. Distribution des erreurs liées aux structures N+N

Les segments relevés dans le corpus correspondent au noyau du GN et à tous ses dépendants pré-noyau. Les dépendants post-noyau, s'ils existent, peuvent facilement être récupérés par le biais d'une recherche sur la plateforme *Sketch Engine* (cf. 1.2.1b). Comme on peut le voir, ces erreurs sont beaucoup plus fréquentes dans le sous-corpus de publications scientifiques. Des suggestions d'explication de ce phénomène ont été présentées dans la section 2.3.1c concernant l'emploi des structures N+N.

Détection, catégorisation et correction des structures N+N erronées

Dans la section précédente consacrée aux contraintes pesant sur l'utilisation et la construction de ces structures, nous avons évoqué le fait que leur compréhension, et donc leur

acceptabilité, est étroitement liée à la somme de connaissances partagées entre auteur et destinataire. Le corpus que nous utilisons incorpore une proportion importante de publications scientifiques, dont certaines concernent des domaines qui sont soit légèrement, soit tout à fait en dehors du champ des connaissances de la personne ayant relevé les erreurs. Il existe donc un risque que certaines structures N+N soient jugées comme inacceptables uniquement en raison d'un manque de connaissances sur le domaine.

Afin de prévenir ce risque de manière simple, nous avons utilisé une comparaison des résultats donnés par des moteurs de recherches. Chacun des segments jugés comme étant erronés a fait l'objet d'une recherche sur le moteur de recherche Google, puis d'une recherche sur la version de Google recensant uniquement les productions scientifiques, Google Scholar. Le plus souvent, la forme exacte des segments est utilisée ; le déterminant et/ou certains modifieurs sont parfois exclus s'il semble que la structure N+N puisse fonctionner seule. Les segments obtenant plus de 50 000 résultats sur Google et/ou plus de 100 résultats sur Google Scholar sont considérés comme acceptables. Ces deux seuils ont été sélectionnés après avoir appliqué cette méthode à un échantillon de structures N+N jugées plus ou moins acceptables selon les cas. Les retours des moteurs de recherche sont passés en revue afin de s'assurer que les résultats correspondent bien au segment (ex. : pas d'ajout de ponctuation, noms présents dans le même GN). Toutes les requêtes ont été effectuées en avril 2012. Cette méthode a ainsi permis de ne pas classer parmi les erreurs des segments tels que *metaphor understanding*, *video sequence detection* ou *media treatment*, alors que l'inacceptabilité de segments tels que ?*oblivion policy*, ?*society domain* et ?*ghettos sickness* est confirmée suite à l'observation de leur quasi-absence dans les résultats de requêtes.

Les obstacles rencontrés lors du relevé, du classement et de la correction de ces segments sont autant de fenêtres sur les difficultés liées à l'automatisation future de ces processus. Comme nous l'avons vu, malgré l'existence de tendances fortes, il n'existe pas de règle grammaticale stricte portant sur la construction de ces séquences. En raison de la diversité des constructions, qui est comparable à celle des GN avec pré-modification en général, il n'est pas pertinent de mener une étude de corpus d'ampleur modeste, comme celles que nous avons menées concernant les adverbes. Même si nous avons suivi les indications des grammaires exposées dans la section précédente, puis mis en place un système de vérification simple grâce à des moteurs de recherche, les jugements d'inacceptabilité restent subjectifs. Une annotation des erreurs par au moins deux personnes qualifiées serait nécessaire à l'amélioration de la fiabilité du processus de détection.

Pour ces raisons, aucun segment n'est jugé comme strictement agrammatical. Cependant, il existe un gradient dans l'acceptabilité des segments, certains d'entre eux étant clairement inacceptables (ex. : ?*our friend relationship*), et d'autres étant plus proches de la limite de l'acceptabilité, s'ils ne la franchissent pas totalement pour certains locuteurs ou locutrices (ex. : ?*the language knowledge acquisition bottlenecks*). Par souci de simplicité, nous faisons cependant référence aux segments comme étant "erronés". Proposer une correction s'avère également très délicat, particulièrement dans le cas des erreurs liées à la relation sémantique existant entre les deux termes, ou bien dans le cas d'empilement de noms et d'adjectifs. Nous reprenons ce sujet dans notre présentation de chaque catégorie puisque le type de correction à apporter est en lien avec le type d'erreur. Comme nous allons le voir, la catégorie générale des erreurs liées aux structures N+N cache en réalité plusieurs types d'erreurs nécessitant un traitement adapté.

Les segments erronés relevés sont classés en cinq catégories principales, selon l'aspect de la structure qui pose problème : "Relation sémantique", "N_1 défini", "N_1 avec *-s*", "Empilement", et "Lexique". Des recoupements peuvent exister, mais les segments sont généralement dominés par une caractéristique d'erreur principale. Le tableau 41 présente ces cinq catégories accompagnées d'un exemple, des propositions de correction pour cet exemple, et du nombre d'occurrences concernées.

Catégorie	Segment	Proposition de correction	Occ.
Relation sémantique	*?the society domain*	1. the domain of society 2. the societal domain 3. the social domain	**23**
Empilement	*?information system security strategies heterogeneity*	1. heterogeneity of information system security strategies 2. heterogeneity in strategies for information system security	**9**
N_1 défini	*?the State official discourse*	1. the State's official discourse 2. the official discourse of the State	**7**
N_1 génitif	*?the ghettos sickness*	1. the ghettos' sickness 2. the ghetto's sickness 3. the sickness of the ghettos 4. the sickness of the ghetto	**5**
Lexique	*?our friend relationship*	1. our friendship 2. our friendly relationship	**5**

Tableau 41. Extrait de la catégorisation des séquences N+N erronées relevées dans le corpus

Un tableau similaire contenant toutes les erreurs concernant les séquences N+N est disponible en Annexe 5. Comme mentionné plus haut, il existe des recoupements entre les catégories : les erreurs qui sont présentes dans deux catégories sont suivies d'un croisillon (#) dans le tableau en annexe. Des segments sont également présents dans le corpus plusieurs fois sous la même forme ; le nombre d'occurrences est alors indiqué entre parenthèses à la suite du segment concerné. Les paragraphes suivants donnent le détail de la constitution des catégories, et des critères utilisés pour la détection.

La catégorie "Lexique" est une des plus marginales. Elle concerne les erreurs de sélection d'un des mots de la séquence, généralement le N_1. Dans certains cas, la structure est améliorée par la substitution au nom de l'adjectif correspondant, comme le segment *?money inequality*, qui peut être corrigé en *financial inequality*. Dans d'autres cas, comme celui qui est donné dans le tableau pour cette catégorie, la meilleure correction consiste à substituer un nom unique exprimant la même chose que la séquence N+N. Ces segments pourraient être classés dans la catégorie "Lexique" du domaine Groupe Nominal dans la catégorisation globale du corpus d'erreurs ; cependant, ils forment un sous-ensemble suffisamment cohérent pour être considérés comme un type d'erreur lié aux structures N+N.

Les segments classés dans la catégorie "N_1 génitif" ont le format [The + N_1 + N_2] et contiennent un nom dépendant portant le morphème *-s*. Dans la sous-partie précédente, nous avons présenté les contraintes énoncées par Quirk et al. et Biber et al. concernant l'emploi de noms au pluriel en fonction de dépendant ; les segments inclus dans cette catégorie ne semblent pourtant pas relever des différentes situations décrites. L'intitulé de cette catégorie n'est pas "N_1 au pluriel" en raison de la façon dont le nom portant le morphème *-s* est employé dans ces segments ; nous formulons l'hypothèse que celui-ci puisse soit provenir d'une erreur dans l'emploi du génitif (ex. : *the ghetto's sickness*), soit être la marque intentionelle d'un pluriel, sa présence provoquant l'omission de la marque de ponctuation du génitif (ex. : *the ghettos' sickness*). Dans ce cas, la structure n'aurait que la forme de surface d'une séquence N+N. Cette hypothèse est renforcée par le fait que ces segments incluent un déterminant défini, en la forme de l'article défini *the* ou du déterminant possessif *their*, indiquant la présence d'une structure sous-jacente de type [[The N_1's] N_2] plutôt que [The [N_1 N_2]].

D'après Quirk et al., les relations sémantiques qui peuvent être exprimées dans une séquence N+N sont des relations permanentes (op. cit. : 1331). Leur emploi dépend donc de la possibilité pour le N_1 d'avoir un référent générique, et non défini. Les segments de la catégorie "N_1 défini" ne semblent pas remplir cette condition, notamment lorsque le N_1 est le

nom d'une personne : *?Maurice Barrès principles*. Cette analyse est également fondée sur l'observation des corrections à apporter qui nous semblent les mieux adaptées ; à l'exception du cas particulier que nous venons de citer, et dans lequel le premier terme est par nature défini puisqu'il s'agit d'un nom propre, celles-ci nécessitent à chaque fois l'introduction de l'article défini *the* devant le N_1 lors de la réorganisation du segment.

La catégorie "Empilement" regroupe les séquences comprenant au moins trois termes fonctionnant comme dépendant. Précisons que le terme "empilement" n'est pas utilisé ici avec une signification syntaxique, et qu'il peut s'appliquer à des structures syntaxiquement empilées comme enchâssées. Les erreurs N+N "Empilement" ne peuvent être considérées comme des erreurs grammaticales, si toutefois elles sont bien construites d'un point de vue syntaxique, mais même ce paramètre s'avère difficile à juger. Elles sont considérées comme maladroites en vertu des difficultés de compréhension qu'elles sont susceptibles de poser, en particulier si certaines des relations sémantiques exprimées sont inattendues. Leur emploi entre notamment en contradiction avec les préconisations du domaine de l'anglais comme *lingua franca*, invitant à l'utilisation de formules aussi intelligibles que possible. La limite de trois termes est déterminée à partir des observations de corpus présentées dans Biber et al. (1999 : 597) et citées plus haut. Pour rappel, celles-ci indiquent que les structures à deux termes dépendants sont fréquentes avec 20 % des occurrences, la fréquence tombe à 2 % pour les structures incluant plus de deux dépendants.

La catégorie regroupant le plus de segments est aussi la plus difficile à définir. L'emploi approprié d'une séquence N+N dépend en grande partie de la relation sémantique que l'on souhaite exprimer : celle-ci doit être permanente et stable, et reposer sur des connaissances partagées, afin que la séquence puisse être facilement interprétable. Les recherches à ce sujet indiquent que l'éventail des relations sémantiques pouvant être exprimées est très vaste (Biber et al. : 590), et les grammaires de référence ne donnent bien sûr pas d'indications sur les relations qui sont impossibles. La grammaire pédagogique de Larreya et Rivière isolent cependant deux types de relation pour lesquelles l'emploi d'une structure N+N n'est pas adéquat (les auteurs font référence aux noms composés, mais leurs explications indiquent qu'elles s'appliquent en grande partie aux structures N+N) (2005 : 248-249 ; exemples tirés de l'ouvrage) :

- relation délimitant une partie d'un ensemble ou d'une masse indénombrable :

  [290] **a toast slice*

- relation de type effet-cause :

[291] **a relief sigh*

Certains des segments que nous avons classés dans la catégorie "Relation sémantique" correspondent en effet à ces critères (ex. : *?admiration cries, ?the memorial laws effect*). Par contre, d'autres expriment des relations sémantiques qui sont indiquées comme acceptables par Biber et al. :

[292] *?the paper reader*

[293] *?television information*

Le premier segment semble correspondre à la relation Objet type 1, "N_1 est l'objet du procès décrit par N_2, ou de l'action effectuée par l'agent décrit par N_2" (cf. 2.3.1d). Le second correspond à la relation Source, "N_2 provient de N_1". Ces deux segments sont néanmoins maladroits, et semblent peu susceptibles d'être utilisés par des personnes locutrices natives de l'anglais, sans qu'il soit possible de fournir une explication simple à cette constatation. La catégorie "Relation sémantique" inclut également des segments plus clairement inacceptables en raison de leur construction ou de l'expression d'une relation sémantique trop vague ou très difficilement interprétable (ex. *?the cooking recipe extract, ?texts candidates, ?insignificant size segments*).

Globalement, il semble que les aspects des structures N+N posant le plus de problèmes soient la sélection de relations sémantiques permanentes et le caractère générique du nom dépendant. Un défaut d'attention à l'intelligibilité de la séquence peut aussi être considéré comme une cause de production de séquences maladroites.

Même en ayant accès au contexte des segments, il est parfois difficile de proposer une correction pour une séquence N+N erronée. Dans le cas des segments présentant un empilement de termes, proposer une correction implique de choisir une analyse de la structure du GN, ce qui comporte nécessairement une part d'arbitraire. Dans d'autres cas, la correction envisagée n'est pas meilleure, ou est plus obscure et maladroite que le segment original, comble de l'exercice. Ceci se produit notamment lorsque l'erreur concerne la relation sémantique exprimée ou bien provient d'une erreur de choix lexical. Dans ce cas, une réécriture globale du GN ou de la proposition semble plus adaptée. Biber et al. ont souligné cette difficulté, présente même lorsque les segments sont correctement construits d'un point de vue syntaxique et sémantique (op. cit. : 588) :

> [T]he rephrasing of noun premodifiers is not at all straightforward, because noun+noun sequences can represent many different meaning relations, with no overt indication of which meaning is intended in any given case. […] In fact, such structures often represent more than one possible meaning relation.

La difficulté de reformuler les structures N+N erronées a plusieurs conséquences pour la correction des segments. Tout d'abord, comme dans le cas des erreurs de placement des adverbes, plusieurs corrections sont proposées pour le même segment. Le nombre de corrections possibles varie ainsi de une à quatre. Quinze segments sur 45 reçoivent plus d'une proposition de correction. De plus, les corrections proposées pour les erreurs identifiées dans notre étude peuvent prendre huit formes différentes, dont le tableau 42 présente la liste accompagnée d'exemples. Les types de correction sont présentés par ordre de fréquence :

Type de correction	**Exemple**	
	Segment erroné	**Correction**
1. Réorg. + Insertion Prep. (Det)	*?this meaning transposition*	this transposition of meaning
2. N_1 vers Adjectif	*?the society domain*	the social domain
3. Ns vers Ns'/N's	*?the objects properties*	the objects' properties
4. Réorg. + Insertion V-*ing*	*?the width peak method*	the method using width peak
5. Réorganisation	*?the length excess*	the excess length
6. N_1 N_2 vers N_1's N_2	*?the State official discourse*	the State's official discourse
7. Insertion Prep. (Det)	*?the meaning utterance*	the meaning of the utterance
8. N_1 N_2 vers N	*?our friend relationship*	our friendship

Tableau 42. Types de corrections pour les séquences N+N erronées

Les types de correction pour chaque erreur sont donnés dans la liste des erreurs présentées en annexe (Annexe 5). Les types 1, 4 et 5 consiste en une réorganisation du GN ; il peut s'agir simplement d'inverser les deux termes, comme pour le type 5, ou bien l'ajout de terme est nécessaire, comme pour le type 1 (ajout d'une préposition et parfois d'un déterminant) et le type 4 (ajout d'un verbe dans sa forme V-*ing*). Inversement, la correction peut également consister à insérer une préposition sans changer l'ordre des noms, comme pour le type 7. Dans le cas des types 2, 3 et 6, le nom dépendant n'est pas déplacé mais est modifié : il peut être remplacé par un adjectif du même champ sémantique (type 2), ou porter la marque du génitif

au singulier ou au pluriel (types 3 et 6). Enfin, la correction peut consister en une modification de la séquence N+N en un seul nom exprimant la même idée (type 8).

Sur les 79 propositions de correction, la correction la plus fréquente est de loin la réorganisation de la séquence couplée à l'ajout d'une préposition et éventuellement d'un déterminant, avec 48 propositions de correction. La modification du N_1 vers un adjectif ou sa modification vers une forme génitive sont également fréquents, avec 10 propositions de correction pour chacune de ces modifications. La réorganisation de la séquence avec inclusion d'un verbe en *–ing* et la réorganisation seule sont très peu fréquentes, avec 4 et 3 occurrrences respectivement. Enfin, les trois derniers types sont marginaux, avec une ou deux occurrences (ajout du génitif, inclusion d'une préposition sans réorganisation, remplacement des deux noms par un seul).

Le tableau suivant reprend les cinq types d'erreur identifiés et donne les formats de correction qui sont utilisés pour ces erreurs, par ordre de fréquence :

Type d'erreur	**Types de correction**
Relation sémantique	Réorganisation + Insertion Préposition (Det.) Réorganisation + Insertion V-*ing* N_1 vers Adjectif Réorganisation
Empilement	Réorganisation + Insertion Préposition (Det.) Réorganisation + Insertion V-*ing*
N_1 défini	Réorganisation + Insertion Préposition (Det.) N_1 N_2 vers N_1's N_2 Insertion Prep. (Det) N_1 vers Adjectif
N_1 génitif	Réorganisation + Insertion Préposition (Det.) Ns vers Ns'/N's
Lexique	N_1 vers Adjectif Réorganisation + Insertion Préposition (Det.) Réorganisation N_1 N_2 vers N

Tableau 43. Corrections mobilisées pour chaque type d'erreur dans les structures N+N

Ce tableau illustre le fait que les types d'erreur ne correspondent pas nécessairement à un modèle de correction déterminé. On remarque tout de même certaines tendances, notamment le fait que les erreurs d'empilement soient ici toutes corrigées par le biais d'une réorganisation du GN, et que celles dans lesquelles le N_1 est un génitif sont corrigées par le biais d'une

réorganisation et ajout de préposition/déterminant ainsi que par une modification morphologique faisant apparaître la forme au génitif.

La section suivante traite de la modélisation des règles de correction pour certaines catégories d'erreurs, et offre des pistes de réflexion pour les erreurs que nous ne sommes pas encore en mesure de corriger automatiquement. Nous verrons quelles solutions de correction peuvent être mises en places parmi celles que nous venons d'évoquer. Avant de passer à cette section, nous présentons une synthèse des formats de surface adoptés par les structures N+N erronées.

Format de surface des structures N+N erronées

Jusqu'à présent, nous avons utilisé le terme "N+N" pour faire référence symboliquement à l'ensemble de ces structures. Comme nous l'avons vu dans les paragraphes précédents, leurs formes sont bien entendu beaucoup plus variées. Une synthèse de tous les formats que prennent les structures N+N erronées relevées dans notre corpus est donnée dans le tableau 44. Le nom noyau est indiqué en gras. La colonne de droite indique le nombre d'occurrences pour chaque schéma. Les noms au singulier sont distingués des noms avec le morphème *-s* (Ns) uniquement au niveau des dépendants ; l'occurrence de noms propres (Npr) est également documentée, mais ne concerne qu'un seul segment. Sont présentées en premier les structures n'ayant pas de déterminant, puis les schémas incluant l'article *the*, et enfin les schémas incluant des déterminants autres que *the*. Cette distinction est rendue visible dans les schémas parce que, comme nous venons de le voir, la présence de l'article *the* est liée à une certaine catégorie d'erreur.

Ce tableau ne prend en compte que les formes de surface, et non la structure syntaxique des segments. La présentation de cette dernière serait hasardeuse, pour au moins deux raisons. Tout d'abord, puisque les segments sont erronés, ils ne sont pas toujours bien formés d'un point de vue syntaxique ou sémantique, ce qui rend caduque toute analyse : dans *?the meaning utterance*, par exemple, il est difficile, et pas nécessairement pertinent, de définir le rôle syntaxique de *meaning* par rapport à *utterance*, étant donné les problèmes de construction du segment. D'autre part, dans certains cas, l'inacceptabilité du segment provient justement du fait que celui-ci contient un nombre important de dépendants, ce qui rend son découpage partiellement arbitraire. De plus, si l'on combine les descriptions formelles et les analyses syntaxiques, on décompte presque autant de descriptions que de segments. Afin de voir apparaître des tendances significatives, l'observation d'un corpus beaucoup plus large est

inévitable. Globalement, il semble que les noms dépendants soient le plus souvent en fonction de complément, ce qui confirme les observations de plusieurs linguistes (ex. : Payne et Huddleston, 2002 : 439).

Format des segments											**Occ.**
									N	**N**	**7**
									Npr	**N**	**1**
									Ns	**N**	**1**
								N	N	**N**	**1**
								N	Ns	**N**	**1**
							N	N	N	**N**	**3**
						N	N	N	Ns	**N**	**1**
					Adj				N	**N**	**1**
					Adj			N	N	**N**	**1**
					Adj			N	Ns	**N**	**3**
The									N	**N**	**8**
The									Ns	**N**	**6**
The								N	N	**N**	**2**
The							N	N	N	**N**	**1**
The					Adj				Ns	**N**	**1**
The				Adj	Adj			N	N	**N**	**1**
The			Adj	Adj	Adj				N	**N**	**1**
The		N			Adj					**N**	**1**
The	V-*ing*								N	**N**	**1**
Det									N	**N**	**3**
Total											**45**

Tableau 44. Formats des séquences N+N erronées dans le corpus

La première observation à faire à la suite de cette synthèse concerne la grande diversité de l'échantillon recueilli. On recense en effet 20 schémas différents ; ce chiffre tombe cependant à 14 si les distinctions entre les différents déterminants et les noms au singulier et au pluriel ne sont pas prises en compte. Il est probable que cette grande diversité soit liée à la taille de notre corpus, suffisamment important pour recueillir un large échantillon de formats différents, mais pas assez pour voir apparaître des tendances significatives. Toutefois, la majorité des occurrences adopte un format relativement simple : 26 segments sur 45 sont de la forme [The/Det + N(s) + N]. Neuf segments incluent au moins un adjectif, et 14 comportent plus de deux noms.

En second lieu, on remarque que la forme seule des segments n'est pas toujours une bonne indication de la présence d'une erreur, puisque pour certaines d'entre elles les schémas adoptés peuvent être rencontrés dans des textes en anglais L1, même s'ils sont rares. C'est l'apparition de tendances contraires à celles de l'anglais L1 qui permet de mettre à jour la possibilité d'un schéma d'erreur récurrent : comme le montre également la catégorisation des erreurs dans ces structures, on retrouve par exemple cinq cas de N_1 avec une possible interprétation de cas génitif dans quatre sources différentes provenant du sous-corpus de publications et d'extraits de rapports techniques.

b. Propositions de solutions pour la détection et la correction des erreurs dans les structures N+N

Les cinq catégories d'erreur introduites ci-dessus représentent des difficultés différentes dans la maîtrise de l'emploi des structures N+N. En conséquence, elles appellent des traitements distincts et directement liés au type d'erreur, même s'il existe des recoupements. Trois de ces catégories nécessitent des recherches beaucoup plus approfondies que celles que nous avons été en mesure d'effectuer dans notre cadre, si l'on souhaite aboutir à l'implémentation de la correction. Les solutions proposées constituent donc des pistes informées pour des recherches futures. Les deux autres catégories font l'objet de solutions concrètes, même si elles ont également des limites importantes. Les deux sous-sections suivantes présentent les règles de traitement proposées pour ces deux groupes.

Catégories "Lexique", "N_1 défini" et "Relation sémantique"

Les difficultés posées par les catégories "Lexique", "N_1 défini" et "Relation sémantique". constituent des pierres d'achoppement pour le domaine du traitement des langues en général. Nous remarquons d'ailleurs que la mise en place de la détection et de la correction de ces erreurs s'inscrit dans le champ du traitement automatisé des langues plutôt que dans celui de la linguistique, puisqu'elles reposent notamment sur le développement et l'utilisation de ressources lexicales et sémantiques utilisables dans des systèmes automatisés, ainsi que sur des méthodes statistiques. Les paragraphes suivants présentent les solutions de détection et de correction envisagées pour chaque type, ainsi que les limites de ces solutions.

De manière générale, les erreurs de choix lexical sont les plus difficiles à traiter automatiquement, ce qui explique que, malgré leur très haute fréquence dans les corpus d'interlangue, il existe peu de recherches à leur sujet en dehors des celles portant sur les collocations (Leacock et al., 2010 : 102). La détection et la correction des erreurs lexicales

dans les structures N+N n'échappent pas à cette règle puisque, comme nous l'avons vu, il est impossible de créer des lexiques exhaustifs faisant la liste de toutes les structures acceptables, qu'elles soient considérées commes des noms composés ou non. La détection pose particulièrement problème en raison de la similitude des formats des structures erronées et des structures bien formées. Une détection fondée sur la forme du segment n'est donc pas envisageable. Puisque l'on ne peut pas se reposer sur la forme de surface pour détecter l'erreur, et que nous n'avons pas les ressources lexicales nécessaires permettant de confronter les segments présents dans les textes à l'ensemble des segments acceptés, l'estimation de leur acceptabilité peut être effectuée en observant leur fréquence d'utilisation. La solution envisagée pour la détection est ainsi similaire à la méthode que nous avons utilisée afin d'estimer l'acceptabilité des structures N+N rencontrées dans le corpus (voir 2.3.2a "Détection, catéorisation et correction des structures N+N erronées").

Cette méthode de détection consiste, après identification des structures N+N dans un texte, à les rechercher automatiquement sur internet à l'aide d'un moteur de recherche, dans le but d'utiliser le nombre de retours de ce segment pour en évaluer la popularité. Un nombre de retours inférieur à une quantité définie pourrait déclencher un avertissement quand à la présence possible d'une erreur. Puisque les erreurs dans les structures N+N sont plus fréquentes dans les documents informatifs, notamment les publications scientifiques, la recherche pourrait être limitée à un moteur dédié à ce type de texte, tel que Google Scholar. La propagation de l'utilisation d'une structure N+N pouvant être rapide, le recours à un moteur de recherche est intéressant car il permet d'évaluer leur présence dans un "corpus" en expansion et renouvellement perpétuels. En raison des limites de l'utilisation d'un moteur de recherche (ex. : fiabilité du nombre des retours et des sources utilisées), le développement de cette méthode de détection devrait cependant aussi évaluer la pertinence de l'utilisation de corpus existants, tels que la section journalistique du *COCA*, ou bien le corpus *TenTen* pour l'anglais, qui rassemble et trie des documents issus d'internet, et dont la seconde version comprend 4,65 milliards de mots (Jakubíček et al., 2013).

Une fois la détection effectuée par ce biais, la correction pourrait reposer sur plusieurs stratégies différentes, présentées ici de la plus satisfaisante à la moins satisfaisante. Dans toutes les occurrences de ces erreurs dans le corpus, la correction peut être effectuée en substituant un adjectif au N_1. La première solution de correction pourrait donc être de proposer cette modification. La sélection de l'adjectif pourrait être effectuée par le biais de l'utilisation ou du développement de ressources lexicales faisant correspondre noms et

adjectifs du même champ. La majorité des erreurs de cette catégorie accepte également une correction prenant la forme d'une réorganisation du segment, avec l'ajout éventuel d'une préposition et d'un déterminant. Cette seconde solution pourrait ainsi être proposée. Notons cependant qu'elle pose par nature de nombreuses difficultés, comme le choix de la préposition et le choix de l'intégration ou non d'un déterminant. La dernière solution n'en est pas vraiment une : en raison du trop grand risque d'erreur dans la proposition d'une correction, on peut choisir de s'en tenir à la détection, en proposant à la personne utilisatrice de revoir le segment jugé comme erroné, éventuellement à partir d'exemples (les modalités de la création de messages de retours correctifs sont exposées dans le Chapitre 3, section 3.3.2).

En ce qui concerne la catégorie "N_1 défini", on dispose d'une information quant à la forme de surface de l'erreur puisque les GN concernés sont déterminés par l'article *the*. La présence de ce déterminant est nécessaire au jugement d'erreur, mais elle n'est pas suffisante pour pouvoir classer de tels segments dans cette catégorie, ou même pour juger le segment comme erroné : certains segments erronés avec *the* ne sont pas jugés comme faisant partie de la catégorie "N_1 défini", et certaines séquences N+N acceptables incluent également cet article lorsque c'est le nom noyau qui est défini :

[294] **the society domain* (catégories "Lexique"/"Relation sémantique")

[295] *the success rate (of this method)*

On ne peut donc pas se reposer uniquement sur le format de surface des segments pour le repérage des erreurs. Cette étape peut cependant être effectuée à l'aide de la même méthode de recherche de fréquence sur internet ou dans de grands corpus. Les informations quant à la présence de *the* peuvent être utilisées par la suite afin d'orienter la correction.

Parmi les erreurs de la catégorie "N_1 défini", toutes sauf une peuvent être corrigées par une réorganisation du segment avec ajout d'une préposition et éventuellement d'un déterminant. La seule erreur appelant une autre correction est **the meaning utterance*, qui est marginale. Comme pour la catégorie "Lexique", on peut aussi choisir de ne pas proposer de correction et de présenter à la personne utilisatrice des aides indirectes, comme des exemples ou des explications linguistiques.

Comme pour les deux catégories précédentes, la détection est également le premier obstacle important au traitement des erreurs de la catégorie "Relation sémantique". Ici également, la structure de surface ne peut pas être utilisée comme indication de la présence d'une erreur. La détection de ces erreurs par des méthodes linguistiques impliquerait donc de

pouvoir interpréter le sens du segment. Des tâches proches de celle-ci sont entreprises dans le domaine du traitement automatisé des langues consacré à l'interprétation des composés et des groupes nominaux complexes (ex. : Fabre, 1996 ; Buckeridge et Sutcliffe, 2002 ; Costello et al., 2006 ; Butnariu et al., 2009 ; Hendrickx et al., 2013), domaine dans lequel elles sont considérées comme particulièrement difficiles et loin d'être abouties (Hendrickx et al., 2013 : 143). Les difficultés les plus fréquemment citées sont la rareté des données (*data sparseness*) pouvant servir de base à l'interprétation du sens des GN, la diversité des formes à interpréter et des relations sémantiques, et la nécessité de se limiter à un domaine précis.

Les recherches dans le domaine de l'interprétation des composés et des GN ont pour objectif de faciliter la recherche de données dans des textes, la traduction automatique, les résultats d'études de question-réponse ainsi que d'autres tâches de traitement automatisé des langues. Ces objectifs diffèrent de ceux de recherches éventuelles sur cette erreur en détection et correction automatisées, pour lesquelles l'interprétation sémantique de la structure ne constitue qu'une étape de la phase de détection. De plus, ces travaux ne sont que partiellement transférables au traitement des erreurs parce qu'ils utilisent comme base des textes considérés comme entièrement corrects, alors que, par définition, nous nous intéressons uniquement aux textes comportant des erreurs. L'utilisation des techniques développées dans ce domaine impliquerait de vérifier leurs performances lorsque les relations sémantiques exprimées sont imprévisibles. L'interprétation par le biais de paraphrases, qui est une des tâches que les recherches dans ce domaine cherchent à accomplir, pourrait alors se révéler utile dans la phase de correction.

L'utilisation de la ressource *FrameNet* est également une piste intéressante à explorer. Les types de travaux en reconnaissance sémantique évoqués ci-dessus sont d'ailleurs souvent fondés sur l'utilisation de *FrameNet*. Le projet *FrameNet*, amorcé à Berkeley en 1997, a pour objectif le développement d'une ressource lexicale pour l'anglais disponible en ligne et utilisable directement par des personnes ou dans le cadre de projets de traitement automatisé des langues (Ruppenhofer et al., 2006 : 5 ; "About *FrameNet*", site internet du projet, voir bibliographie). Comme son nom l'indique, *FrameNet* est fondé sur la Sémantique des cadres, théorie élaborée par Charles J. Fillmore dans les années 1970, et selon laquelle la compréhension du sens d'un mot dépend de la connaissance des concepts qui sous-tendent sa signification. Les mots s'inscriraient donc dans des cadres sémantiques, qui prennent la forme d'une description de la relation, de l'évènement ou de l'entité concernée, ainsi que de ses participants. Ruppenhofer et al. donnent l'exemple de l'action de cuisiner (*cooking*), qui fait

partie du cadre "Apply_heat", qui implique les participants COOK, FOOD, CONTAINER et HEATING_INSTRUMENT (mise en forme dans l'original). Ces participants sont les éléments de cadre (*frame elements*), et les termes qui évoquent ce cadre, tels que *fry*, *boil* ou *grill*, sont les unités lexicales du cadre (*lexical units*) (op. cit. : 5). La ressource *FrameNet* contient 10 000 unités lexicales accompagnées d'exemples annotés, ainsi que 170 000 phrases dans lesquelles les rôles sémantiques ont été annotés à la main. Les annotations sont effectuées à partir de l'observation de l'usage des mots dans des textes en langue naturelle.

FrameNet est une ressource dans laquelle toutes les informations concernant les rôles sémantiques ont été entrées manuellement. Il existe cependant des projets d'automatisation de l'étiquetage des rôles sémantiques ayant débouché sur des analyseurs sémantiques disponibles gratuitement (projets *Shalmaneser*, *LTH* et *Semaphor* ; voir bibliographie pour l'adresse de leurs sites internet). *FrameNet* étant fondé sur l'annotation des rôles sémantiques dans des phrases, il est également nécessaire d'étudier l'adéquation d'une telle ressource au traitement des GN. Par ailleurs, l'annotation automatique de rôles sémantiques n'est pas directement transférable à la détection d'erreurs dans ces relations : en effet, si un système est performant dans la détermination de relations sémantiques correctes, cela ne signifie pas nécessairement qu'il soit facilement adaptable à la recherche de relations sémantiques erronées. Nous rencontrons là le problème classique du traitement des productions d'interlangue, pour lesquelles on ne peut mobiliser les mêmes stratégies que pour des textes bien formés. Une étude précise de cette ressource et des outils d'automatisation disponibles devrait donc être menée afin d'explorer leur potentiel pour la détection des erreurs de relation sémantique dans les séquences N+N.

L'ensemble du texte dans lequel la séquence N+N se trouve peut aussi être pris en considération dans la phase de détection. Nous avons vu que l'acceptabilité des séquences N+N dépend autant du type de relation exprimée que de la permanence de celle-ci, un certain degré d'institutionnalisation pouvant être atteint par le biais de références au concept concerné en amont de l'utilisation d'une structure N+N. Une stratégie de détection possible pourrait consister en l'analyse du co-texte afin de vérifier si les termes de la structure ont déjà été utilisés à proximité l'un de l'autre, indiquant la stabilisation progressive de la relation. Enfin, la méthode de recherche sur internet que nous avons évoquée plus haut pourrait aussi s'avérer utile dans la détection des erreurs de cette catégorie.

La majorité des erreurs de la catégorie "Relation sémantique" peuvent être corrigées par le biais d'une réorganisation des termes du GN, incluant l'ajout d'une préposition ou d'un V-*ing*.

Cette solution de correction pourrait donc être appliquée, avec les difficultés que nous avons citées plus haut. Notons que si les erreurs sont détectées à partir d'informations sémantiques, celles-ci pourront être utilisées pour aider à la sélection de la préposition à ajouter. Enfin, ici également, il est possible de s'en tenir à une détection guidée sans proposer de correction dans le but d'éviter les corrections erronées, risque très important en raison des difficultés posées par la réorganisation du GN.

Pour résumer, la méthode de recherche des segments sur internet est une solution de détection qui peut convenir à ces trois catégories d'erreur, avec certains ajustements. De plus, si l'on choisit de proposer des corrections, la solution prenant la forme de la réorganisation du GN peut également être appliquée aux trois erreurs. Malgré l'attrait de la possibilité de traiter les erreurs de ces catégories avec les mêmes stratégies, cette solution unique est problématique. En effet, l'utilisation des mêmes méthodes a pour conséquence l'effacement des distinctions entre les trois catégories d'erreur, rendant la production de messages correctifs ciblés impossible. Par ailleurs, la méthode de détection proposée s'inscrit dans le domaine du traitement des langues plutôt dans celui de la linguistique, ce qui la place quelque peu en dehors du cadre de nos recherches. Nos recherches futures pourront cependant porter sur l'exploration linguistique de l'utilisation d'annotations de rôles sémantiques pour la correction des erreurs de la catégorie "Relation sémantique", qui sont d'ailleurs les plus fréquentes pour les séquences N+N dans notre corpus. Le développement de règles fondées sur ces informations nécessite des recherches dédiées et très approfondies. Les deux autres catégories d'erreurs, qui présentent des formes de surface plus distinctives, se prêtent d'avantage à un traitement dans notre cadre. Celui-ci s'accompagne néanmoins de difficultés.

Catégories "Empilement" et "N_1 génitif"

La structure des erreurs d'empilement est de loin la plus facile à détecter, puisqu'elle a une structure distinctive qui implique la présence de plusieurs noms, et éventuellement d'adjectifs. Ces structures longues et complexes ne sont pas pour autant agrammaticales en tant que telles ; elles constituent néanmoins une charge cognitive importante pour le lectorat, et peuvent provoquer une rupture dans la compréhension si certaines des relations sémantiques exprimées ne sont pas facilement interprétables par ailleurs.

Proposer une réécriture des séquences N+N erronées en des GN incluant des dépendants placés après le noyau constitue une stratégie possible. Les deux segments suivant sont un

exemple d'une erreur d'empilement corrigée par le biais d'une telle réorganisation du segment :

[296] **heterogeneous information sources cooperation*

[297] *cooperation between heterogeneous information sources*

La réorganisation des structures N+N est cependant extrêmement hasardeuse, même lorsque la tâche est effectuée à la main, comme la sous-section précédente l'a évoqué. Cette tâche nécessite de faire des prédictions informées sur les paramètres suivants :

- la structure syntaxique globale du GN, c'est-à-dire la façon dont il doit être "découpé" et réorganisé en fonction des relations de complémentation et de modification existant entre les différents éléments,
- la sélection d'une préposition adéquate introduisant le nouveau dépendant post-noyau,
- le choix de l'introduction d'un déterminant pour le nominal dépendant, qui est transformé en GN complet, et la sélection de ce déterminant,
- le choix de l'introduction d'un déterminant pour le GN initial, le nom noyau étant désormais le premier terme.

Le fait que ce processus implique l'automatisation d'analyses syntaxiques et de la sélection de prépositions et de déterminants est représentatif de la difficulté de sa mise en place, puisqu'il combine les principales difficultés rencontrées dans le domaine de la correction automatisée.

Pour ces raisons, nous proposons pour l'instant d'adopter une des solutions mentionnées dans la sous-section précédente, c'est-à-dire de relever ces segments sans leur adosser de proposition de correction, ni de jugement d'agrammaticalité ou d'inacceptabilité. Ils peuvent par contre être accompagnés de messages correctifs informant la personne utilisatrice des possibles problèmes posés par l'utilisation de ces structures. Les modalités de la création de ces messages sont présentées dans le Chapitre 3. Étant donné la diversité potentielle de ces segments, il n'est pas possible de créer des patrons permettant de tous les relever ; nous utilisons la description des formats des segments effectuée dans la section précédente comme base des patrons pour réduire le champ des possibilités. La modélisation de ces segments est présentée dans le tableau suivant (la numérotation des schémas reprend la suite des schémas proposés pour les adverbes) :

Segment
8. (Det) + N + N + N + **N**
9. (Det) + N + N + N + N + **N**
10. (Det) + Adj + N + N + **N**
11. (Det) + Adj + Adj + N + N + **N**

Tableau 45. Modélisation des séquences N+N avec empilement

Les schémas incluent systématiquement des nominaux comportant au moins quatre éléments, avec un déterminant optionnel. Ce nombre a été choisi après évaluation de la limite à partir de laquelle le nombre de dépendants oblige le destinataire à un effort conscient d'interprétation de la séquence. Ce choix repose aussi sur le résultat des études de corpus présentées dans *LGSWE* à ce sujet, et qui soulignaient la rareté des GN comportant plus de deux dépendants, donc plus de trois éléments avec le nom noyau (Biber et al., 1999 : 597).

Dans notre catégorisation, au moins trois de ces élements sont des noms, puisqu'une structure comportant uniquement deux noms ne peut pas être considérée comme un empilement problématique. Si le segment ne comporte que trois noms, ceux-ci doivent être précédés d'au moins un adjectif. Des segments comportant plus d'éléments peuvent exister mais ils sont rares. Le fait d'étendre les schémas ferait aussi augmenter le risque de relever des séquences de termes appartenant à plusieurs GN situés côte à côte dans une proposition. Notons que les patrons n'admettent pas de virgules entre les adjectifs, puisque ces signes de ponctuation peuvent servir d'outil de clarification du GN et le rendent donc plus facile à comprendre.

Le format particulier des erreurs de la catégorie "N_1 génitif" simplifie leur détection. Les segments sont caractérisés par la présence d'un déterminant défini précédant un nom au pluriel. D'autres erreurs de ce type peuvent en théorie être produites avec un schéma différent, le corpus utilisé étant trop réduit pour épuiser tous les schémas possibles. Par ailleurs, les indications de surface ne permettent pas d'éviter que certaines structures correctes soient relevées. Ce risque est abordé dans la partie du Chapitre 3 consacrée à l'évaluation.

Le tableau 46 montre la modélisation des segments et des propositions de correction ; les choix de règles sont expliqués dans les paragraphes qui le suivent.

Segment		**Propositions de correction**
12. Det défini + (Adj) + N_1s + **N_2**	⇒	Det défini + (Adj) + N_1s + ['] + **N_2** Det défini + (Adj) + N_1 + ['s] + **N_2** THE + **N_2** + OF + Dét défini + (Adj) + N_1s THE + **N_2** + OF + Dét défini + (Adj) + N_1

Tableau 46. Modélisation des segments "N_1 génitif" et des propositions de correction

Les cinq segments suivants illustrent l'application des schémas de correction à un segment erroné :

[298] **the objects properties*

[299] *the objects' properties*

[300] *the object's properties*

[301] *the properties of the objects*

[302] *the properties of the object*

Deux principaux types de correction sont proposés : la modification du GN pour faire apparaître un cas génitif ([299] et [300]), et la réorganisation du GN ([301] et [302]).

Dans le premier cas, la correction consiste soit à introduire un signe de ponctuation ([299]), soit à modifier le nombre du N_1 pour le faire passer du pluriel vers le singulier, tout en ajoutant la marque du génitif ([300]). En effet, la nature du format des segments déclenche une ambiguïté concernant le nombre du N_1. Si l'on postule que ce segment devrait inclure un génitif, on ne peut savoir si la personne l'ayant produit envisage un N_1 pluriel ou singulier, puisque la marque du génitif masque cette distinction, du moins à l'oral. Pour cette raison, deux propositions sont données pour chaque solution, ce qui donne quatre propositions de correction pour un même segment.

La réorganisation du segment est ici plus simple que dans les cas que nous avons évoqués précédemment, chacun des paramètres étant déjà connu : la relation sémantique est une relation d'appartenance (réelle ou métaphorique), la préposition à sélectionner est donc *of*, et le N_1 déplacé après le noyau conserve le déterminant de départ. Le type de structure appelle l'ajout d'un déterminant défini devant le nouveau GN ; l'article *the* est sélectionné par défaut.

Le schéma de détection ainsi que les corrections proposées incluent un adjectif optionnel modifiant le N_1, car cette structure a été observée pour une des erreurs de ce type relevées dans le corpus (**the semantic reasoners maturity*). L'adjectif est placé avant le N_1, et conserve

ce placement dans les corrections. Il ne s'agit que d'un type de structure parmi de nombreux autres qui pourraient exister et qui ne sont pas schématisés ici, cependant ce type représente une configuration courante, c'est-à-dire la modification d'un nom par un adjectif. Pour le moins, ce choix est en accord avec notre approche ascendante, qui consiste à utiliser l'observation d'erreurs réelles pour informer les schémas de détection.

Dans cette partie consacrée à la modélisation linguistique des erreurs liées aux séquences N+N, et de leurs corrections possibles, nous avons commencé par soulever le problème de la détection de ces erreurs, qui peut être très subjective et dépendre des connaissances lexicales de la personne relevant les erreurs dans le domaine sur lequel porte la production d'interlangue (ex. : publications scientifiques). La solution que nous proposons pour ce problème consiste à vérifier la fréquence des séquences dans un moteur de recherche, spécialisé ou non, en partant du principe que, s'il est difficile de déterminer l'acceptabilité d'une séquence de manière objective, on peut au moins en vérifier la présence dans des productions existantes.

Nous avons ensuite présenté une catégorisation des erreurs relevées dans le corpus. Cinq catégories ont été identifiées, incluant les erreurs liées à la relation sémantique exprimée par la séquence N+N, les cas d'empilement de noms avant le noyau pouvant gêner la compréhension, les séquences dissimulant un cas génitif, celles qui incluent un N_1 défini plutôt que générique, et enfin les erreurs lexicales. Parmi ces cinq catégories, la première est de loin la plus fréquente, avec 23 erreurs sur les 49 relevées (certaines erreurs appartenant à plusieurs catégories). Elle est également la plus difficile à identifier étant donné le nombre de relations sémantiques pouvant exister entre les deux noms des séquences N+N, ainsi que le manque de consensus concernant l'identification des relations acceptables.

Huit types de correction sont également identifiés. La correction la plus fréquente, et qui peut être utilisée pour tous les types d'erreur, consiste en la réorganisation de la séquence N+N en un GN dans lequel le nom est modifié ou complété par un GP. On remarque que les formes de surface des erreurs sont très variées, allant de structures très simples incluant deux noms sans déterminant, jusqu'à des structures lourdes avec un déterminant, deux ou trois adjectifs, et jusqu'à quatre noms.

Pour trois des types d'erreur, il est impossible d'utiliser les formes de surface pour la détection, mais une solution unique se dégage sous la forme de l'évaluation de la fréquence des structures dans des productions sur internet. Une solution unique est également envisageable pour la correction des erreurs de ces trois types, qui pourrait consister en une

réorganisation de la séquence N+N. Dans le cadre d'une recherche utilisant des méthodes linguistiques, l'utilisation de cette solution unique ne semble pas satisfaisante, puisqu'elle ne donne pas de réel éclairage linguistique sur les erreurs. Nous proposons d'autres pistes pour la détection fondées sur des informations linguistiques, notamment l'identification des relations sémantiques existant entre les deux noms, qui pourra faire l'objet de recherches futures. Il semble toutefois que le traitement complet de ces erreurs, en plus de nécessiter des recherches dédiées approfondies sur des problèmes complexes, nécessite également un compromis entre méthodes linguistiques et stratégies informatiques.

Les deux autres types d'erreur présentent des formats distinctifs, et se prêtent donc bien à la détection par le biais de patrons, avec cependant l'existence d'un risque de détection erronée. Quatre schémas d'erreur sont proposés pour les erreurs d'empilement ; la correction la plus adaptée est la réorganisation du segment, mais le risque important de corrections insatisfaisantes mène à l'abandon de l'étape de la correction afin de conserver la qualité des corrections proposées au public utilisateur. Ici également, un compromis s'avère nécessaire. Les erreurs incluant des cas génitif sont quant à elles traitées par le biais d'un seul schéma, associé à quatre propositions de correction.

Conclusion

L'objectif principal de ce chapitre est la modélisation linguistique des erreurs et des corrections concernant le placement des adverbes en fonction d'adjoint et l'utilisation des structures N+N. Ces modélisations doivent mener à l'implémentation de règles de correction permettant de détecter et de corriger ces erreurs automatiquement. Afin d'atteindre cet objectif, nous avons appliqué une méthode mixte fondée d'une part sur la synthèse des descriptions linguistiques des phénomènes, et d'autre part sur l'observation précise des erreurs. Cette seconde étape permet d'affiner les descriptions et de cibler les configurations qui posent le plus de difficultés aux personnes utilisatrices de l'anglais. Un des défis majeurs de cette recherche fut d'établir des critères fiables auxquels se référer afin de juger de l'acceptabilité et/ou de la grammaticalité d'un segment donné.

La première partie de ce chapitre a introduit la méthode utilisée, ainsi que les trois grammaires de référence qui nous servent de sources principales. Celles-ci sont complémentaires dans leurs approches, *ACG* adoptant une approche nuancée du jugement de la grammaticalité des exemples, alors que *CGE* présente des analyses plus tranchées des phénomènes. Le recours à *LGSWE* nous a permis de compléter les descriptions par des

résultats d'analyses portant sur des corpus de plusieurs millions de mots en anglais L1. Nous utilisons la plupart du temps le cadre descriptif et la terminologie de *CGE*, en raison de sa précision et des innovations apportées.

Dans la deuxième partie, consacrée au placement des adverbes, l'observation des erreurs produites nous a amenée à nous concentrer sur le placement des adverbes de manière et de l'adverbe *also*, et plus particulièrement à leur placement entre un verbe lexical et son complément. Il a été établi que ce type de placement, s'il n'est pas toujours jugé comme agrammatical, est du moins très peu utilisé par des locuteurs et locutrices de l'anglais L1.

Dans cette partie, nos travaux apportent les contributions suivantes à la présente recherche : une synthèse du placement des adverbes dans la proposition et le GV en fonction de leur type sémantique, l'identification de facteurs précis influençant le placement des adverbes, tels que la forme du complément et la lourdeur du groupe adverbial, et enfin une étude du placement d'*also* dans un corpus d'anglais L1, que nous comparons à leur placement dans un corpus d'anglais L2 produit par des francophones, et qui permet d'indiquer la présence d'un transfert entre le français L1 et l'anglais L2.

Enfin, la contribution la plus importante consiste en la modélisation des erreurs et des corrections pour l'adverbe *also* et les adverbes de manière, permettant l'implémentation des règles de correction dans le chapitre suivant. La méthode utilisée peut être transposée à l'étude d'autres types d'adverbes.

Dans la dernière partie, nous présentons les aspects principaux de la syntaxe des séquences N+N, qui sont considérées comme des GN complexes. La description des erreurs liées à ces structures révèle l'existence de plusieurs catégories distinctes nécessitant des traitements adaptés. Malgré l'existence de segments considérés comme inacceptables dans notre corpus, l'étude des contraintes pesant sur leur emploi et leur construction n'a pas révélé de critères d'agrammaticalité, ces contraintes étant principalement d'ordre sémantique et pragmatique. En conséquence, nous menons une réflexion sur les raisons de l'apparente inacceptabilité des segments relevés dans notre corpus, et sur les façons de les améliorer. Nous proposons une modélisation des erreurs des catégories "Empilement" et "N_1 génitif", et offrons des pistes de recherches pour la correction des trois autres catégories, qui concentrent des problématiques non résolues dans le domaine du TAL.

L'automatisation des processus de détection et de correction fonctionne mieux à partir de certitudes, mais les difficultés rencontrées quant à l'établissement de critères non ambigus

pour la détection et la correction nous ont amenée à adapter les règles afin d'obtenir un équilibre entre efficacité et fidélité à la langue. Les règles incluent donc dans certains cas plusieurs propositions de correction. Dans le cas des structures N+N en particulier, la détection est envisagée comme une indication de la possibilité de la présence d'une erreur, et la correction comme un guide fourni à la personne utilisatrice du système. La génération de messages correctifs, présentée dans le chapitre suivant, joue un rôle important dans l'accompagnement de la correction.

Chapitre 3
Implémentation des règles de détection et de correction automatisées

Introduction

Dans ce chapitre, nous abordons le troisième et dernier volet de nos travaux, qui en constitue également l'application. Après avoir identifié deux types d'erreur présents dans les productions de personnes utilisatrices de l'anglais à un niveau intermédiaire à avancé, et décrit les schémas de ces erreurs et les possibilités de correction, nous présentons la façon dont nous avons mobilisé ces informations linguistiques afin de détecter et corriger ces erreurs automatiquement dans une plateforme fonctionnelle. Nous définissons la détection comme l'étape de marquage spécifique d'un segment de texte signalant la présence d'une erreur, et la correction comme la proposition d'un segment pouvant directement remplacer le segment erroné afin de le rendre grammatical ou acceptable.

La première partie de ce chapitre est consacrée au domaine de la correction grammaticale automatisée. Nous présentons les caractéristiques du domaine, l'évolution des différentes technologies utilisées, ainsi que leurs avantages et limites. Le public cible de nos règles est un public apprenant/utilisateur de l'anglais, ce qui donne une orientation spécifique à la création de règles de correction automatisée ; nous abordons donc la façon dont le fonctionnement de la correction automatisée doit être adapté à ce type de profil. Le second volet de cette partie est consacré à l'évaluation indicative des capacités actuelles d'une sélection de correcteurs grammaticaux.

La deuxième partie de ce chapitre constitue la présentation de l'implémentation des règles de détection et correction dans la plateforme <TextCoop>. Nous détaillons les caractéristiques générales de cette plateforme, avant de présenter le format des règles de correction que nous avons mises en place, ainsi que les ressources lexicales et grammaticales mobilisées. Cette partie inclut également les résultats de l'évaluation des règles sur un corpus d'anglais L1 et un corpus modifié d'anglais L2. Cette évaluation nous amène à identifier les points forts comme les limites des règles utilisées, et à proposer des pistes pour l'amélioration future des corrections.

Enfin, la troisième partie aborde l'accompagnement des corrections, étape qui est très souvent intégrée aux correcteurs grammaticaux, en particulier lorsque ceux-ci s'adressent à un public apprenant. Une présentation des résultats de travaux de recherche en acquisition des langues secondes (ALS) et en enseignement des langues assisté par ordinateur (ELAO) nous aide à dresser une liste de consignes guidant la création de messages correctifs associés à nos

règles de correction. Ceci aboutit à la création d'un canevas en cinq étapes, pouvant être adapté au profil des personnes utilisatrices du système.

3.1 La correction grammaticale automatisée

L'expression "correction grammaticale automatisée" (CGA), ou *automatic grammar checking*, fait référence à l'utilisation de programmes informatiques pour la détection et la correction d'erreurs principalement morphosyntaxiques et syntaxiques dans des textes écrits. Certaines erreurs lexicales sont parfois également traitées, la dénomination de "grammaticale" étant envisagée dans une acception large (ex. : Tschichold et al., 1997). Elle est souvent associée à la correction orthographique et à la correction du style dans les systèmes à visée commerciale, bien que les technologies soient développées séparément. Pour simplifier, nous utilisons cette expression pour renvoyer aux étapes de la correction et de la détection des erreurs, même si celles-ci sont généralement distinctes dans les règles utilisées. Les recherches sur la correction automatisée ont débuté dans les années 1970, et les logiciels qui en sont issus font partie des technologies d'informatique linguistique les plus connues du grand public, notamment en raison de leur intégration à des logiciels de traitement de texte (Haswell, 2006 [2005] : 1). Les capacités des correcteurs commerciaux sont cependant jugées comme étant en-deçà des attentes du grand public et des professionnels de la langue (ex. Chen, 2009 : 175), ce qui amène Heift et Schulze à souligner, dans leur ouvrage consacré au traitement automatique des erreurs en ELAO, l'existence d'un véritable besoin de correcteurs grammaticaux dans le domaine de l'apprentissage des langues (2007 : 87).

La première sous-partie est consacrée à la présentation des différentes méthodes existantes dans le domaine de la CGA. Nous commençons par aborder les caractéristiques des types de travaux menés dans ce domaine, qui permettent de le délimiter et de classer ces travaux selon un ensemble de paramètres. Ensuite, nous présentons l'évolution des technologies utilisées pour la détection et la correction, avant de présenter les spécificités de la correction grammaticale automatisée pour l'ELAO. La seconde sous-partie consiste en une évaluation des capacités d'une sélection de correcteurs grammaticaux disponibles actuellement. Après avoir présenté notre méthode et les critères de sélection utilisés pour ces programmes, nous analysons les résultats obtenus sur un échantillon de notre corpus d'erreurs.

3.1.1 Les méthodes utilisées dans le domaine de la correction grammaticale automatisée

a. Caractéristiques et délimitation du domaine de la CGA

Les travaux menés dans le but de détecter et de corriger automatiquement des erreurs dans des textes écrits s'inscrivent dans un ensemble de cadres différents, et peuvent être classés selon plusieurs variables. Nous en identifions au moins sept ; la liste de ces variables est présentée ci-dessous, et elles sont détaillées dans les paragraphes qui suivent :

- objectif des travaux,
- phase du traitement de l'erreur concernée par les travaux,
- types d'erreur traités,
- choix de langue cible,
- choix du type de public cible,
- choix de langue source,
- méthodes et technologies utilisées.

Ces variables sont inter-dépendantes, le type d'erreur traité dépendant par exemple de l'objectif des travaux, du type de public et du couple langue source/langue cible choisi. Il existe trois objectifs principaux pour les travaux en CGA. L'objectif le plus évident, mais pas le plus courant, est la création d'un correcteur grammatical automatique "complet", qu'il soit à vocation commerciale ou non. C'est par exemple le cas des travaux de Tschichold et al. (1997), Naber (2003), et Gamon et al. (2009). Un autre objectif, proche du premier, consiste à créer un correcteur grammatical ou certaines règles de correction dans le but de les intégrer à un système d'ELAO, tel un tuteur intelligent (*Intelligent Language Tutoring System*). C'est l'objectif des travaux sur la CGA dans le projet *ICICLE* (McCoy et al., 1996 ; Michaud et al., 2000 ; Greenbacker et McCoy, 2008), ainsi que des travaux de Krüger et Hamilton (1997) et Bender et al. (2004). Enfin, certains travaux se consacrent à l'exploration de nouvelles règles de CGA, dans certains cas sans objectif défini quant à leur intégration dans des systèmes complets, du moins d'après ce qui est dévoilé par les auteurs des recherches. On peut citer les travaux de Brockett et al. (2006), Meurers et Metcalf (2006), De Felice (2008), Hermet et Désilets (2009) et Madnani et al. (2012). Nos travaux s'inscrivent dans cette troisième catégorie.

Nous avons choisi une approche totale du traitement des erreurs, de la sélection des erreurs à traiter dans un corpus original à la création de messages correctifs. Cette approche implique de prévoir des règles de détection comme de correction des erreurs. Si la plupart des travaux en CGA concernent la détection et la correction des erreurs en même temps, tous les travaux n'adoptent pas ce type d'approche, en particulier ceux de la troisième des catégories que nous venons d'identifier. Certains travaux se concentrent sur la détection des erreurs, comme par exemple ceux de Chodorow et al. (2007). Par ailleurs, il existe des correcteurs grammaticaux disponibles en ligne, gratuits ou payants, qui détectent les erreurs mais ne proposent pas systématiquement de correction ; c'est le cas des correcteurs *SpellCheckPlus* et *ProofWriter* (adresse des sites internet dans la bibliographie, catégorie "Sites internet et systèmes informatiques"). L'intérêt de se concentrer sur la détection des erreurs est parfois d'éviter de proposer une correction inadaptée ou pire que l'original. D'autres recherches s'attèlent à la correction d'erreurs déjà détectées, afin d'explorer des stratégies innovantes pour cette tâche sans avoir à prendre en charge la détection en amont. Parmi ceux-ci, on peut citer les recherches de Hermet et Désilets (2009) et Madnani et al. (2012). Ils sont cependant moins nombreux, puisqu'il est obligatoire de les coupler à des règles de détection afin de les rendre utilisables dans des situations authentiques.

Lorsque l'objectif des travaux en CGA est la création d'un correcteur grammatical ou leur intégration dans un tuteur de langue, on cherche généralement à traiter une majorité de types d'erreur, le plus souvent parmi les types les plus fréquents. Par exemple, le système *ESL Assistant* de Microsoft, dont la création est documentée dans Gamon et al. (2009), cible les erreurs suivantes : présence et choix des articles définis et indéfinis, présence et choix des prépositions, nombre des noms, confusion entre infinitif et V-*ing*, présence et choix des auxiliaires, confusion entre nom et adjectif, ordre des mots localisé, régularisation erronée des conjugaisons verbales (op. cit. : 493). D'autres études, comme la nôtre, se concentrent sur un type d'erreur en particulier. Par exemple, nous avons vu que les erreurs dans l'utilisation des articles et des prépositions, particulièrement fréquentes dans les productions de personnes apprenant l'anglais, font l'objet d'un nombre important de travaux.

La plupart des travaux conduits actuellement en CGA adoptent l'anglais comme langue cible, en raison du statut international de cette langue. Il existe cependant des travaux similaires sur d'autres langues, comme par exemple les travaux de Dansuwan et al. sur le thaï (2001), de Nagata sur le japonais (2002), de Eeg-Olofsson et Knutsson sur le suédois (2003), de Vandeventer sur le français (2003), de Delmonte sur l'allemand (2003).

La sélection du public visé et le choix de langue source vont de pair. Si le public visé est un public natif de la langue choisie comme langue cible, la question de la sélection de la langue source ne se pose pas. Traditionnellement, les correcteurs grammaticaux commerciaux ont ciblé les locutrices et locuteurs de la langue traitée, mais c'est de moins en moins le cas pour l'anglais. Nous verrons dans la section 3.1.1c que la correction grammaticale automatisée à destination des personnes apprenantes doit se plier à plusieurs contraintes qui ne concernent pas la correction à destination des personnes natives. Si le public visé est un public apprenant/utilisateur, les auteurs des études peuvent choisir soit de ne pas sélectionner de langue source définie, soit de cibler une famille de langues, soit de cibler une seule langue source. Dans le cas de la CGA appliquée à l'anglais, la frontière entre les systèmes à destination d'un public natif et à destination d'un public apprenant est floue : la majorité des personnes utilisant et apprenant l'anglais ne l'ont pas comme langue maternelle, ce qui veut dire que ces systèmes ont de fortes chances d'être utilisés par ces publics. Cependant, cela ne signifie pas forcément que ceux-ci leur soient adaptés, les erreurs produites par une personne native et une personne non-native étant le plus souvent très différentes.

Choisir une langue source permet de travailler à partir d'un corpus d'erreurs plus homogène, et d'utiliser les indications d'influence trans-linguistique, alors que le choix de laisser cette option ouverte élargit le public cible potentiel. Dans les travaux de Bender et al. (2004) et De Felice (2008), dont les règles de correction sont à destination de publics non-natifs de l'anglais, aucune langue source n'est définie. Les travaux liés à la création du *ESL Assistant* de Microsoft ciblent les langues asiatiques, notamment le chinois et le japonais (Gamon et al., 2009). Le projet *ICICLE* concerne uniquement les personnes sourdes locutrices de la langue des signes américaine (Greenbacker et McCoy, 2008). La sélection d'une langue source, le français, pour des corrections à destination d'un public utilisateur de l'anglais est un des points centraux de la recherche présentée ici.

La dernière variable identifiée concerne le type de méthode utilisée pour la mise en place de la détection et de la correction. La section suivante présente les différentes méthodes existantes, leur évolution, ainsi que les technologies émergentes.

b. Évolution des technologies utilisées en CGA

Dans *Automated Grammatical Error Detection for Language Learners* (2010), Leacock et al. font un compte-rendu de l'évolution des méthodes utilisées par les correcteurs

grammaticaux depuis la création des premiers d'entre eux. Nous nous fondons en partie sur ce compte-rendu pour le résumé qui suit.

Les premiers correcteurs grammaticaux utilisaient des algorithmes de recherche de sous-chaîne (*string matching*), recherchant des suites précises de caractères ayant été codées par les auteurs du programme concerné (Heift et Schulze, 2007 : 23). Cette technologie a cependant été rapidement supplantée par d'autres règles plus "intelligentes" (cf. Dodigovic, 2005 : 108), faisant appel à l'étiquetage morpho-syntaxique (*part-of-speech tagging*) et à l'analyse syntaxique (*parsing*). Les analyseurs syntaxiques reposent sur des grammaires formelles, qui peuvent être créées manuellement ou par apprentissage automatique. Les grammaires formelles les plus utilisées sont *Head-driven Phrase Structure Grammar*, la grammaire lexicale-fonctionnelle, la théorie du gouvernement et du liage, le modèle des principes et paramètres, les grammaires logiques, et enfin les grammaires à clauses définies (Heift et Schulze, 2007 : 30). D'après Leacock et al., tous les systèmes de CGA disponibles actuellement utilisent un certain degré d'analyse linguistique automatisée (op. cit. : 5). Dans un troisième temps, la création de corpus annotés à l'aide de ces mêmes technologies a permis le développement de techniques statistiques, notoirement très gourmandes en données.

L'analyse syntaxique et l'étiquetage morpho-syntaxique sont utilisés pour d'autres applications en traitement automatisé des langues, comme le traitement de corpus ou l'exploration de données. Les tâches de détection et de correction des erreurs présentent cependant une caractéristique unique, en ceci qu'elles concernent le traitement d'un texte présentant des erreurs grammaticales. Les technologies mobilisées doivent donc être capables de fonctionner à partir de données qui ne sont pas congruentes avec la grammaire formelle utilisée, puisqu'un analyseur qui omettrait de traiter des segments en raison de leur agrammaticalité ne serait d'aucune utilité pour la tâche de correction. Cette particularité est exacerbée dans la gestion de textes non-natifs, en raison de la quantité et de la variété des erreurs. Un analyseur syntaxique capable de traiter ce type de données est qualifié de "robuste". Plusieurs règles sont utilisées, les deux principales méthodes étant la création de "mal-règles" (*malrules*), aussi appelées "grammaires d'erreurs", et la méthode de gestion des contraintes (Heift et Schulze, 2007 : 39-43).

Cette dernière regroupe un ensemble de règles alternatives, dans lesquelles certaines des contraintes des règles de la grammaire sont relâchées afin de pouvoir générer plusieurs analyses lorsqu'une erreur est rencontrée ; l'analyse comprenant le moins de violation de contraintes est alors considérée comme étant la plus plausible. Si cette technique semble être

la plus efficace dans la détection des erreurs (Heift et Schulze, 2007 : 43), elle obtient des taux de précision jugés trop bas (la précision étant une mesure de la proportion d'erreurs correctement identifiées par rapport au total de segments relevés) (Vandeventer, 2001 : 116). Par ailleurs, elle ne constitue pas une facilitation de la mise en place de la phase de correction (Milton et Cheng, 2010 : 34).

L'utilisation d'une grammaire d'erreurs consiste à anticiper les erreurs qui peuvent être produites, soit en observant un corpus de productions d'interlangue, soit par analyse des possibilités d'influence trans-linguistique. Ces erreurs sont ensuite converties en règles et inscrites manuellement dans l'analyseur syntaxique afin lui permettre de générer des analyses des structures erronées. Cette méthode est par exemple utilisée dans le projet *ICICLE* (Michaud et McCoy, 2000), et dans les travaux de Eeg-Olofsson et Knutsson (2003). Cette approche a plusieurs avantages et inconvénients. Tout d'abord, son utilisation mène au choix d'une seule L1 source, afin de limiter le nombre de règles d'erreur à créer, ce qui réduit aussi la taille du public cible. En outre, les erreurs imprévisibles, tout comme les erreurs peu fréquentes pour lesquelles aucune règle n'a été créée, ne peuvent pas être détectées. Cependant, aucune approche, utilisant cette méthode ou non, ne permet d'analyser toutes les structures erronées, surtout si elles sont particulièrement imprévisibles. Par ailleurs, l'utilisation d'une grammaire des erreurs permet d'avoir des informations précises sur celles-ci, préparant ainsi la génération de corrections et de messages correctifs détaillés, s'il y a lieu (Heift et Schulze, 2007 : 38). De plus, elle est adaptée au travail sur des erreurs ciblées, ce qui peut se révèler utile lorsque le système est utilisé dans un cadre pédagogique.

Comme nous l'avons vu quelques paragraphes plus haut, l'évolution la plus récente dans les technologies de détection et correction automatisées repose sur un apprentissage automatique à partir de corpus annotés de très grande taille. Ces méthodes sont dites "statistiques" ou "dirigées par les données" (*data-driven*), et peuvent relever de deux principales approches : l'utilisation d'un classifieur, ou la création d'un modèle de langue. Dans le premier cas, un classifieur est entraîné sur de larges corpus de texte bien formé d'un point de vue grammatical, et apprend les caractéristiques de l'utilisation de certaines parties du discours, utilisant une fenêtre de trois à sept termes en général. Lorsqu'il est confronté à du texte comprenant des erreurs sur le type de mot pour lequel il a été entraîné, le classifieur détecte les utilisations de ce mot qui ne correspondent pas à ce qu'il a rencontré auparavant, et qui peuvent donc être des erreurs. Les travaux de De Felice et Pulman (2008) sont un exemple de l'utilisation d'un

classifieur à entropie maximale pour la détection et la correction des erreurs liées à l'utilisation des prépositions et des déterminants.

Les modèles de langue sont fréquemment utilisés en traitement automatisé des langues, et consistent à assigner une probabilité, ou score, à une séquence de mots appelée "n-gramme", souvent au nombre de deux ou trois, alors appelés "bi-gramme" et "tri-gramme". Dans le cadre de l'utilisation de modèles de langue pour la détection des erreurs, on considère qu'un score faible peut être indicatif de la présence d'une erreur (Leacock et al., 2010 : 12). Cette méthode fait partie des règles mobilisées dans une des études de Gamon et al. (2009) pour la création du *ESL Assistant* de Microsoft.

Les méthodes statistiques présentent l'avantage d'éliminer la tâche de création de règles formelles que nécessite l'utilisation de l'analyse syntaxique. Leur principal inconvénient est qu'elles requièrent l'accès à des données linguistiques en grand nombre, sous la forme de corpus annotés de plusieurs centaines de millions de mots ; ceux-ci sont même parfois encore insuffisants pour permettre un apprentissage concernant certaines structures. En conséquence, les taux de rappel (proportion d'erreurs détectées sur l'ensemble d'erreurs existantes) sont généralement bas. Dans l'idéal, le processus d'apprentissage automatique devrait se faire sur des corpus d'interlangue, afin de permettre la prise en compte de la fréquence des erreurs ou bien de l'influence de la L1, mais ceux-ci n'existent pas encore en quantité et qualité satisfaisantes (Leacock et al., 2010 : 91). De notre point de vue, elles ont également l'inconvénient de ne pas être explicatives, en ceci qu'elles ne permettent pas d'appréhender les causes des erreurs ou d'avoir des informations grammaticales sur les formes concernées.

Leacock et al. identifient cinq innovations technologiques récentes dans le traitement des erreurs, la plupart étant liées au type de données utilisées pour les méthodes statistiques. Elles incluent l'utilisation de corpus de taille gigantesque (plusieurs milliards de mots) générés automatiquement par l'extraction de documents sur internet, l'utilisation du corpus de n-grammes compilé par Google, et la comparaison du nombre de retours de recherches sur internet pour plusieurs variantes d'un segment erroné. D'autres innovations concernent le recours à un "aller-retour" de traduction automatisée pour obtenir des propositions de correction, et la recherche d'indications concernant la L1 des auteurs des productions d'interlangue grâce à la fonction de recherche "régionale" (limitée à certaines aires géographiques et linguistiques) des moteurs de recherche.

Pour finir, Leacock et al. mentionnent que toutes les erreurs ne nécessitent pas le même traitement, certaines étant avantageusement traitées à l'aide de règles et d'heuristiques, et

d'autres, comme les erreurs dans l'utilisation des articles et des prépositions, exigeant un traitement statistique (op. cit. : 74). Les auteurs ajoutent qu'un correcteur grammatical robuste est un correcteur qui utilise plusieurs méthodes de traitement (op. cit. : 13).

Comme nous l'avons déjà mentionné, nos règles de traitement des erreurs utilisent principalement des patrons de détection, une technique également nommée "filtrage par motif" (*pattern-matching*), complétés d'instructions pour la correction, ces deux étapes formant les règles de détection et correction. Nous nous appuyons donc sur une grammaire des erreurs construites à partir d'une analyse des erreurs et de recherches approfondies sur les caractéristiques grammaticales des structures concernées. L'utilisation de patrons diffère de celle des algorithmes de recherche de sous-chaîne (*string-matching*), en ceci qu'elle se fonde sur les parties du discours et non sur des chaînes de caractères précises. La plateforme utilisée, <TextCoop>, intègre une fonction de recherche de structures lexico-grammaticales à l'aide de lexiques et de grammaires simples. Une grande partie des ressources nécessaires au fonctionnement des règles dans <TextCoop> dans notre recherche a été constituée par nos soins. Notre choix initial a été d'éviter l'utilisation d'un analyseur syntaxique, qui alourdirait le traitement de manière importante. La fenêtre de recherche des patrons est aussi précise que possible afin de limiter les besoins en analyse sytaxique automatique, et de permettre de faire abstraction de la complexité globale de la phrase dans laquelle le segment erroné se trouve. Nous revenons en détail sur la constitution des règles dans la partie 3.2.

c. La correction grammaticale automatisée pour l'ELAO

Avant de passer à l'étude des capacités des correcteurs grammaticaux, nous souhaitons résumer les indications données par des experts concernant les paramètres à prendre en compte lors de la création de correcteurs grammaticaux pour l'ELAO, et plus largement pour des locutrices et locuteurs non-natifs (LNN). Notre objectif ici n'est pas la création d'un système complet ; cependant, notre choix d'explorer des règles de correction à destination d'un public utilisateur et apprenant à partir d'informations grammaticales, et de proposer des messages correctifs pouvant favoriser un apprentissage, nous place dans une optique similaire à celle de l'ELAO. Nous développons cet aspect dans la dernière partie de ce chapitre.

Les correcteurs grammaticaux créés pour les locutrices et locuteurs natifs (LN) ne seraient pas adaptés au traitement de productions interlangues, d'une part en raison de leur faible taux de détection des erreurs des personnes apprenantes, et d'autre part à cause du risque de faux positif, c'est-à-dire l'indication de la présence d'une erreur alors que le segment est correct. En

effet, si l'ensemble des erreurs produites par les LNN peut englober celles qui sont produites par les LN, une grande partie des erreurs dans les productions interlangue ne concernent pas les productions en langue maternelle : une apprenante pourra écrire **I would of done it*, erreur courante pour les anglophones, mais un anglophone ne produira jamais la phrase **I eat often apples*. Ces erreurs seraient donc invisibles pour les correcteurs grammaticaux "standards". En outre, ceux-ci génèrent souvent des faux positifs. Même si elle peut être agaçante, cette faiblesse n'est pas réellement problématique pour les personnes natives, qui ont les compétences nécessaires pour trier le bon grain de l'ivraie ; en revanche, ils constituent des obstacles importants pour les personnes apprenantes (Tschichold, 1999 : exemplaire non paginé) :

> Such feedback is clearly inadequate for non-proficient writers, because they cannot rely on their intuitions as much as native speakers and are therefore disconcerted much more easily by such messages that are not only superfluous, but could also induce the student to introduce an error in a previously correct text passage. In the worst case, the student might even wrongly assume that the rule displayed by such a grammar checker is valid in all contexts and learn a rule which later has to be "unlearned" again.

La creation de correcteurs grammaticaux à destination spécifiquement des LNN est donc nécessaire. Tschichold donne un ensemble de consignes pour leur élaboration, autour de deux concepts centraux : l'augmentation de la précision du système, et la favorisation de la prise d'autonomie de la personne apprenante. Ce second concept amène à repenser le rôle du correcteur grammatical dans l'apprentissage, dont Tschichold considère qu'il doit se limiter à apporter une aide formelle aux personnes apprenantes, sans prendre en charge totalement la correction de leurs productions. Cette aide est avantageusement fournie par l'inclusion de ressources supplémentaires, telles des dictionnaires, thésaurus, ou concordanciers. La favorisation de l'autonomie du public apprenant passe également par la création de messages correctifs adaptés. Nous reviendrons sur les consignes données par Tschichold à ce sujet dans la dernière partie de ce chapitre, qui est consacrée à la génération de ces messages.

L'augmentation de la précision du système passe par la prise en compte de la L1 des utilisateurs et utilisatrices. Pour Tschichold, ce paramètre est primordial afin de pallier certaines imperfections du correcteur, et permet de fournir des messages correctifs en langue source. Un correcteur grammatical idéal devrait également être adapté au niveau en L2 du public utilisateur, comme à leurs connaissances préalables en grammaire simple (identification du sujet, du verbe, etc.). Enfin, la consigne la plus stricte concerne la nécessité

de réduire drastiquement, voire d'éliminer totalement les faux positifs, même au prix d'une baisse du nombre d'erreurs correctement détectées. Ceci peut également être favorisé par la réduction des dictionnaires utilisés, afin d'éliminer des termes problématiques ou bien de les limiter au vocabulaire que le public apprenant est susceptible d'utiliser pour la tâche prévue.

Dans la sous-partie suivante, nous aurons le loisir de comparer les performances de correcteurs grammaticaux prenant en compte certaines de ces caractéristiques à celles de systèmes "tout-venant".

3.1.2 Les capacités des correcteurs grammaticaux automatiques : une étude indicative

a. Présentation de la méthode d'évaluation des correcteurs grammaticaux automatiques

L'objectif de cette évaluation dans notre projet n'est pas de faire un état des lieux exhaustif des compétences de ces programmes. Nous souhaitons donner un aperçu de leurs capacités, en tentant d'identifier les types d'erreurs qui semblent non problématiques et ceux qui ne sont pas encore pris en compte. En particulier, une seconde étude plus approfondie est menée sur les deux types d'erreur qui nous occupent ici.

Des consignes concernant l'évaluation des correcteurs grammaticaux sont données dans Starlander et Popescu-Belis (2002) et Tschichold (1994). L'étude de Starlander et Popescu-Belis a non seulement pour objectif d'évaluer deux correcteurs grammaticaux et orthographiques pour le français, mais espère également favoriser la standardisation des méthodes d'évaluation de ces systèmes en proposant des consignes fondées sur les normes ISO. Ces normes impliquent notamment la définition d'exigences de qualité pour l'outil évalué. Cependant, si une partie des consignes données par Starlander et Popescu-Belis est pertinente pour notre étude, la plupart ne sont pas directement applicables à celle-ci en raison des exigences particulières du public apprenant, les correcteurs évalués dans leur étude étant à destination de personnes locutrices natives du français. De la même façon, l'étude longitudinale de Kies (2008) concernant les capacités de sept correcteurs pour l'anglais porte sur la correction d'erreurs en L1.

En revanche, l'article de Tschichold traite précisément de l'évaluation des correcteurs grammaticaux à destination des personnes apprenantes. Elle y prévoit l'évaluation d'aspects non-linguistiques relatifs à l'ergonomie et à l'accessibilité des systèmes, ainsi que certains aspects liés à leur orientation didactique, comme l'inclusion de dictionnaires et d'aides variées.

Nous mettons ces aspects de côté pour nous concentrer ici sur la qualité de la détection et de la correction ; la question du contenu des messages correctifs est abordée dans la troisième partie de ce chapitre. Pour l'évaluation des tâches de détection et de correction, Tschichold recommande d'utiliser des erreurs authentiques, et d'inclure non seulement différentes classes d'erreurs, mais aussi des erreurs similaires dans des contextes variés (op. cit. : 198). Elle prévoit également la prise en compte des erreurs manquées et des faux positifs (op. cit. : 193), ceux-ci étant particulièrement problématiques pour les personnes apprenantes. Ces consignes recoupent en partie celles de Starlander et Popescu-Belis (2002 : 271).

Comme Tschichold et Starlander et Popescu-Belis, nous adoptons une approche de type "boîte noire" (*black box*), ce qui signifie que le fonctionnement interne du système n'est pas pris en compte. Les exigences de qualité formulées pour notre évaluation sont simples : les systèmes évalués doivent détecter les segments erronés, et uniquement ceux-ci, et proposer des corrections qui permettent d'améliorer le texte d'origine en restorant sa grammaticalité ou bien en le rendant plus naturel.

Dans un premier temps, l'évaluation est menée sur un échantillon d'erreurs extraites de notre corpus d'erreurs, provenant de toutes les catégories. De une à trois erreurs différentes sont utilisées pour chaque catégorie, en fonction du nombre d'erreurs présentes dans la catégorie. L'échantillon utilisé comprend ainsi 55 erreurs et est disponible en annexe (Annexe 6). Dans un second temps, une évaluation est menée sur dix erreurs issues de chacune des deux catégories traitées dans notre recherche, avec une répartition proportionnelle sur les différentes sous-classes qu'elles admettent. Les erreurs sont présentées dans leur contexte d'origine, comprenant toute la phrase ou au moins tout le groupe syntaxique qui les contient, afin de présenter aux correcteurs des entrées aussi authentiques et complètes que possible. Des modifications mineures sont parfois introduites afin de simplifier les phrases. En raison de la nature de notre échantillon de test, composé quasiment uniquement de segments erronés, il est moins probable de retrouver de nombreux faux positifs que dans le cas de l'utilisation de textes entiers ; cette possibilité est cependant prise en compte, puisqu'il est possible que les systèmes donnent une indication d'erreur différente de l'erreur réelle. Les résultats de la détection et de la correction sont évalués séparément. Trois possibilités sont envisagées pour ces deux tâches :

- la détection peut être correcte, inexistante, ou erronée, lorsqu'elle signale une erreur sur une portion correcte d'un segment, ou un autre type d'erreur,

- la correction peut être adéquate, inexistante ou ne pas constituer une amélioration de l'original.

Notons que les faux positifs (détection/correction erronée) sont considérés comme pires que l'absence de détection et de correction. Certaines combinaisons sont inévitables : par exemple, si un segment correct est détecté comme étant erroné, la correction est forcément inadéquate. Nous ne prenons pas en compte la détection des erreurs d'orthographe ou de style.

Tous les correcteurs grammaticaux ayant des fonctionnements différents, il est nécessaire de définir ce que nous considérons comme les indications de "détection" et les indications de "correction". Une erreur est considérée comme ayant été détectée par le système lorsque celui-ci attribue un marquage spécifique au segment ou mot concerné (soulignage, surlignage, coloris contrastant, etc.). Nous adoptons une définition stricte de la correction, en distinguant celle-ci du retour correctif ; ainsi, on considère que le système fournit une correction lorsqu'il indique clairement une alternative précise au segment original. Lorsque plusieurs solutions sont proposées et qu'au moins une solution correcte y figure, nous considérons la correction comme effective. Dans certains cas, le correcteur fournit une règle ou une indication d'erreur suivie d'un exemple similaire et de sa correction : ces instances ne sont pas considérées comme des corrections dans le cadre de notre évaluation. La plupart des correcteurs fournissent des messages correctifs, incluant parfois des explications ; ils sont évoqués plus en détail dans la sous-section 3.3.2.

Pour des raisons pratiques, l'évaluation est menée sur un ensemble de systèmes de correction disponibles à peu de frais, qu'ils soient gratuits, ou disposant de versions d'essai. La plupart d'entre eux sont accessibles en ligne. Comme il a déjà été indiqué, les correcteurs grammaticaux qui ne sont pas créés spécifiquement pour un public apprenant, ou au moins adaptés à leurs besoins, sont souvent jugés comme inefficaces dans la détection des erreurs produites par ce type de public. En théorie, il n'est donc pas pertinent d'évaluer les capacités de ces systèmes ici. Cependant, il existe très peu de systèmes identifiés comme adaptés aux apprenants et facilement accessibles en ligne ; *Criterion* est par exemple un système d'évaluation automatique des productions écrites (*Automated Writing Evaluation*) très connu et semblant donner de bons résultats (cf. Chodorow et al., 2010 ; Lee et Hegelheimer, 2012), mais il n'est accessible qu'à des établissements pédagogiques. Le système *ESL Assistant* de Microsoft a quant à lui été désactivé en 2010 et n'est plus disponible. Pour cette raison, nous avons également évalué des systèmes qui ne sont pas explicitement élaborés à destination de ce public, et qui ne donnent généralement pas d'indication particulière concernant le type de

public ciblé ("Non précisé" dans le tableau). Ceci a l'avantage de nous permettre d'apprécier la différence effective entre les systèmes "standards" et les systèmes pour personnes apprenantes. Une première version de cette évaluation a été effectuée en 2010 ; les résultats présentés ci-dessous sont ceux d'une seconde évaluation menée en août 2013.

Les systèmes évalués sont *SpellCheckPlus*, *Grammarly*, *Ginger*, *Correct English*, *Proofwriter*, *Language Tool*, et le correcteur grammatical disponible dans Word 2007. Leurs caractéristiques principales sont données dans le tableau suivant :

Système	Source	Accessibilité	Public cible	Messages correctifs
SpellCheckPlus	T. Nadasdi et S. Sinclair	Gratuit	Angloph. et appren.	Oui
Grammarly	Grammarly Inc.	Payant – version d'essai	Angloph. et appren.	Oui
Ginger	Ginger Software	Gratuit	Angloph. et appren.	Oui
Correct English	Vantage Linguistics	Payant – version d'essai	Angloph. et appren.	Oui
Proofwriter	Educational Testing Service	Payant	Non précisé	Oui
Language Tool	D. Naber et M. Milkowski	Gratuit	Non précisé	Oui
Word 2007	Microsoft	Gratuit avec Word	Non précisé	Oui

Tableau 47. Caractéristiques des correcteurs grammaticaux évalués

b. Résultats et discussion de l'étude indicative

Le tableau suivant donne les taux de précision et de rappel pour chaque système, pour les étapes de détection et de correction concernant l'échantillon principal. Pour la détection, la précision indique le pourcentage d'erreurs détectées parmi l'ensemble des segments marqués comme erronés, et donne donc une information concernant la quantité de faux positifs. Le rappel concerne le pourcentage d'erreurs détectées parmi toutes les erreurs présentes dans l'échantillon, représentant ainsi l'étendue de la couverture du système. Les taux de précision et de rappel concernant la correction sont étroitement liés à ceux de la détection. La précision indique le pourcentage de corrections adéquates parmi l'ensemble des corrections proposées. Le rappel représente ici le pourcentage de corrections adéquates par rapport au nombre d'erreurs existantes. La comparaison de ces taux nous informe sur le décalage entre les capacités respectives de détection et de correction des systèmes.

Il est nécessaire de signaler que la taille de l'échantillon utilisé, surtout considérant le nombre de catégories d'erreur, n'est pas suffisant pour donner des résultats définitifs

concernant la couverture des correcteurs grammaticaux. Les résultats ont cependant un intérêt comparatif, chaque système ayant été évalué sur le même échantillon représentatif d'erreurs extraites de productions authentiques.

Correcteur	Détection		Correction	
	Précision %	Rappel %	Précision %	Rappel %
SpellCheckPlus	65	20	20	4
Grammarly	46	11	36	7
Ginger	100	31	71	22
Correct English	87	24	73	20
Proofwriter	43	18		
Language Tool	60	5	75	5
Word 2011	75	11	60	5

Tableau 48. Résultats de l'évaluation des correcteurs grammaticaux

Dans l'ensemble, les systèmes détectent et corrigent peu d'erreurs de notre échantillon, ce qui est indiqué par les faibles taux de rappel. Dans le meilleur des cas, moins d'un tiers des erreurs sont détectées, et moins d'un quart d'entre elles sont corrigées efficacement. Les taux de précision sont parfois très faibles, ce qui est préoccupant lorsque les systèmes s'adressent à des personnes non natives ; dans le cas du système *Grammarly* par exemple, dont la présentation dit qu'il s'adresse aux personnes locutrices natives comme aux non-natives, plus d'une erreur signalée sur deux est un faux positif. On ne voit pas apparaître de disctinctions claires entre des outils ne s'adressant pas spécifiquement aux LNN, comme le correcteur de *Word*, et des outils déclarés comme étant adaptés à ce type de public, comme *Grammarly*. Pour l'échantillon évalué, *Ginger* ne présente aucun faux positif en détection, et a le taux de corrections adéquates le plus élevé, même si celui-ci reste faible. Ce résultat global est similaire à celui que nous avions obtenu lors de la première version de l'évaluation menée en 2010. *Ginger* semble ainsi être le mieux adapté aux apprenants, si l'on s'en tient aux recommandations de Tschichold, qui indique que la réduction du nombre de faux positifs doit primer sur la quantité d'erreurs détectées et corrigées (1999 ; exemplaire non paginé).

Nous avons déjà évoqué le fait que les erreurs morphosyntaxiques (accords, construction du groupe verbal) sont les plus largement traitées par les correcteurs grammaticaux automatiques ; cela est vérifié ici, puisque les erreurs qui sont corrigées par au moins quatre

systèmes sur sept sont majoritairement des erreurs d'accord, et pour deux d'entre elles des erreurs lexicales liées à la construction du groupe verbal :

[303] *I hope you'll like *these file.*

[304] *I *can lost a love.*

[305] *Systems are robust and could *be improve on every corpus.*

[306] *please *say [] him "Bonjour" from me.*

[307] *Why *should we loss our identity?*

[308] *I'm returning you* the old system which have never functioned.*

Les erreurs liées au choix des prépositions sont quant à elles rarement relevées. Le placement des modifieurs et les problèmes de construction et d'organisation de la phrase passent inaperçus également, tout comme les erreurs dans le choix des temps et des aspects, beaucoup plus complexes que les erreurs de conjugaison.

Les erreurs dans le placement des adverbes de manière et de l'adverbe *also* sont un angle mort pour tous les correcteurs grammaticaux que nous avons testés, sauf un : *Correct English* a détecté et corrigé adéquatement un adverbe de manière placé en SVAO. Cependant, les autres erreurs du même type, bien que similaires d'un point de vue formel, n'ont pas été relevées par ce correcteur.

Les erreurs liées à l'utilisation des structures N+N sont mieux loties, même si leur couverture n'est ni cohérente ni complète. *Correct English* détecte et corrige les erreurs du type **the ghettos sickness* ; deux autres correcteurs, *Ginger* et *Language Tool*, ont détecté et corrigé une de ces erreurs, mais pas les autres. Certaines erreurs d'empilement de noms sont détectées par *Grammarly* et *Language Tool*, mais d'une part ces correcteurs ne signalent qu'une partie du GN concerné, et d'autre part aucune correction n'est suggérée. D'autres erreurs d'empilement passent inaperçues. Une erreur sémantique (**the concept meaning*) est détectée et corrigée efficacement par *Ginger*. Notons que lors de l'évaluation menée en 2010, aucune erreur concernant les structures N+N n'était corrigée par les correcteurs grammaticaux étudiés ; il semblerait que, malgré le fait que cette problématique n'ait pas fait l'objet de nombreuses recherches rendues publiques, elle ait été remarquée par les équipes développant des systèmes commerciaux. La plupart de ces erreurs ne sont cependant pas systématiquement corrigées, même dans les deux catégories évoquées ici.

Il semble donc que les capacités des correcteurs grammaticaux automatiques justifient les critiques qui sont formulées à leur égard. En outre, on note que les correcteurs créés à destination des personnes non natives ne sont globalement pas plus performants sur ces erreurs que ceux qui ciblent un public natif. Pour conclure, notons que si les capacités des systèmes sont amalgamées, on obtient un taux de détection de 58 % et un taux de correction de 35 %, alors que les taux maximums atteints par les correcteurs pris individuellement sont de 31 % et 22 %. Une solution possible aux faibles capacités des correcteurs tient ainsi peut-être à la mise en place de collaborations plutôt qu'au maintien de la compétition (notamment économique) sur ce créneau.

La première partie de ce chapitre était consacrée à un passage en revue de différents aspects du domaine de la CGA. Nous avons en particulier évoqué les avantages et les inconvénients des technologies utilisées, ainsi que les ajustements nécessaires à la création de règles à destination de personnes utilisatrices de l'anglais. Les méthodes et technologies utilisées dans notre étude ont également été présentées. Certaines des erreurs que nous ciblons sont désormais traitées par des correcteurs existants, mais de manière partielle et inégale. La partie suivant présente dans un premier temps l'implémentation des modèles de détection et de correction détaillés dans le chapitre précédent, et l'ensemble des ressources lexicales et grammaticales mises en place pour cette implémentation. L'évaluation des règles est présentée dans un second temps.

3.2 Implémentation des règles de détection et de correction

Dans cette partie, nous reprenons les modélisations des erreurs et de leurs corrections présentée dans le Chapitre 2 afin de présenter les différentes facettes de leur implémentation dans une plateforme initialement créée pour le traitement du discours, mais pouvant être adaptée pour servir à d'autres tâches de traitement linguistique. La première sous-partie est consacrée à la description de la réalisation technique de l'implémentation des règles, et aborde les caractéristiques de la plateforme <TextCoop>, avant de présenter en détail le format des règles que nous avons implémentées. Nous donnons également la liste des ressources grammaticales et lexicales utilisées. La seconde sous-partie concerne l'évaluation des règles. Nous présentons notre méthode d'évaluation, ainsi qu'un système de classement des différents retours du système pouvant être adapté à l'évaluation d'autres règles ou systèmes de correction. Cette sous-partie donne les résultats de l'évaluation, ouvrant la discussion sur les

limites et points forts de notre choix de règles, ainsi que sur les améliorations qui peuvent y être apportées dans des recherches futures.

3.2.1 Réalisation technique des règles de détection et de correction automatisées

a. Présentation de <TextCoop> et DisLog

Les caractéristiques de la plateforme <TextCoop> sont décrites dans les chapitres 4 et 5 de l'ouvrage *Challenges of Discourse Processing: The Case of Technical Documents* (Saint-Dizier, 2014) ainsi que dans les articles "<TextCoop> : un analyseur de discours basé sur des grammaires logiques" (Saint-Dizier, 2011) et "Processing natural language arguments with the <TextCoop> platform" (Saint-Dizier, 2012). Nous utilisons ces sources dans la présentation qui suit.

<TextCoop> est décrit comme un "environnement basé sur les grammaires logiques dédié à l'analyse de structures discursives" (Saint-Dizier, 2011 : exemplaire non paginé). DisLog est le langage de programmation des règles et spécifications qui les accompagnent dans <TextCoop>, et son nom renvoie notamment à *discourse in logic* (op. cit.). <TextCoop> a été développé entre 2005 et 2008 dans l'équipe ILPL à l'IRIT, Toulouse. Une version de la plateforme a ensuite servi de base aux recherches menées dans le cadre du projet *LELIE*, dont le développement a couru de 2010 à 2013, et a exploité les capacités de la plateforme pour la correction des erreurs dans la documentation technique. Le projet *LELIE* est actuellement dans sa phase de valorisation en collaboration avec de grandes entreprises françaises et étrangères (le nom de ces entreprises n'est pas divulgué pour des raisons de confidentialité).

La plateforme <TextCoop> est parfaitement adaptée à l'analyse des structures qui composent le discours, notamment des phrases ou des portions de phrases, et peut donc être avantageusement utilisée dans notre cadre. <TextCoop> permet ainsi la reconnaissance et l'étiquetage de structures par le biais de patrons, mais également leur modification, ou correction, par le biais d'instructions de réécriture. C'est de cette fonctionnalité dont nous tirons profit avec l'utilisation de la plateforme <TextCoop>.

Nous présentons tout d'abord l'architecture de <TextCoop>, puis donnons le format des règles dans DisLog. Nous nous concentrons sur les caractéristiques utilisées dans notre implémentation des règles, mais il faut noter que <TextCoop> et DisLog ont d'autres fonctionnalités que nous ne mentionnons pas.

L'environnement <TextCoop> est organisé en cinq modules différents :

- un module de patrons et de règles programmées en DisLog,
- un ou plusieurs modules consacrés aux ressources lexicales,
- un module consacré au traitement morphologique,
- un module pour la gestion du système, incluant la gestion des contraintes et les spécifications des cascades,
- le moteur de <TextCoop>.

Le moteur de <TextCoop> et ses ressources lexicales sont implémentés en SWI Prolog, utilisant la syntaxe standard de cette version de Prolog mais aucune des librairies supplémentaires, notamment pour garantir la portabilité de la plateforme. Il s'agit donc d'une application qui peut être utilisée seule (*stand alone*). Le code utilisé pour le moteur de <TextCoop> est de taille réduite et consiste en un interpréteur en Prolog.

Les règles sont organisées en "paquets de règles" (*rule clusters*) : les règles relatives à l'identification d'une structure donnée sont identifiées par un symbole placé en partie gauche de chaque règle. Les règles d'un même paquet sont exécutées selon un ordre de lecture déterminé en amont, ce qui signifie qu'elles sont activées dans l'ordre précis indiqué dans une spécification du moteur, sans retour arrière. Si une règle fonctionne sur un segment de texte, alors les autres règles ne sont pas appelées pour ce même segment. Ce système permet de gérer les problèmes de concurrence entre les règles, celles-ci étant de préférence organisées des plus exceptionnelles au plus courantes. La forme générale des règles dans <TextCoop> est la suivante (Saint-Dizier, 2011) :

```
forme(Identifiant, Entrée, Sortie, Patron, Contraintes, Résultat).
```

Les différents éléments de cette forme sont explicités lors de la présentation de nos propres règles dans la sous-section suivante. Avant de passer à cette présentation, donnons quelques informations sur DisLog, le langage de programmation de <TextCoop>. DisLog utilise les formalismes de grammaires logiques, et les adapte au traitement du discours. Ce langage intègre également des extensions aux expressions régulières, format qui est souvent utilisé dans les outils de traitement automatisé des langues (Saint-Dizier, 2014 : 77). Il permet par ailleurs d'inclure dans les patrons un processus de raisonnement, capable de lever certaines ambiguités ou d'effectuer des calculs. Un des objectifs de la plateforme <TextCoop> est l'identification de structures du discours, qui sont constituées d'éléments de différents niveaux et types, comme les éléments lexicaux, marques morphologiques, marques de ponctuation ou

de typographie ; pour cette raison, ces différents éléments peuvent être présents en même temps dans les règles programmées en DisLog, ce qui constitue une innovation parmi les outils fondés sur des grammaires logiques (op. cit.). Pour cette même raison, cet environnement est adapté à notre objectif, puisque nos propres règles exigent l'introduction d'éléments morphologiques et lexicaux.

b. Présentation technique des règles de détection et correction

Pour la création et l'organisation des règles, nous avons suivi les consignes explicitées par Saint-Dizier (2014 : 112-115). La première de ces consignes concerne le choix des règles, qui doivent avoir un degré d'abstraction équilibré : si les règles sont trop précises, elles risquent de ne concerner que des cas exceptionnels, et les règles d'un haut degré d'abstraction englobent trop de structures différentes. Cette consigne a déjà été prise en compte dans la création des schémas d'erreur et de leurs corrections présentés dans le Chapitre 2, pour lesquels nous avons recherché un équilibre entre largeur du spectre de la détection et précision linguistique.

La seconde consigne concerne l'organisation des règles dans les paquets, les règles les plus contraintes, donc moins fréquentes, devant être examinées en premier dans l'ordre de lecture à déterminer. Cette précaution permet aux cas particuliers, qui correspondront donc aux règles examinées en premier, de ne pas être rendus invisibles par l'activation de règles plus générales. Par exemple, dans notre cas, les règles correspondant au placement des adverbes entre un verbe et un GN long doivent être placées en premier ; si les règles concernant les GN courts sont placées avant celles-ci, le moteur risque de déclencher l'utilisation de la règle "GN court" dans tous les cas, sans faire appel à la règle "GN long".

La troisième consigne pertinente dans notre cadre concerne la gestion du recoupement des règles entre elles, qui est à éviter absolument en raison du risque d'ambiguïtés. Le contenu des règles doit être suffisamment précis pour que celles-ci puissent être clairement distinguées les unes des autres ; dans le cas où le recoupement est inévitable, Saint-Dizier préconise d'établir un ordre de préférence. Le problème posé par cette approche est l'alourdissement du fonctionnement du système, qui doit ainsi appeler toutes les règles concernant des cas exceptionnels avant d'arriver aux cas les plus courants, provoquant ainsi un alongement de la durée de réponse du système. Ceci n'est pas un problème important dans notre cas, puisque nous avions connaissance de ce problème et avons anticipé en limitant le nombre de règles mobilisées. Par ailleurs, l'avantage des correcteurs grammaticaux automatiques est qu'ils

traitent le texte à mesure que celui-ci est rédigé, ce qui permet de tolérer un temps de réponse moins court que pour des systèmes prévus pour traiter des textes après rédaction (ex. : systèmes de traduction automatique). Saint-Dizier indique cependant qu'il est possible de commencer par les règles les plus courantes si celles-ci sont très fréquemment mobilisées, et si leurs premiers termes sont très précis et ne sont pas présents dans les autres règles.

Nous présentons dans les paragraphes suivants les aspects techniques de l'écriture des règles de correction dans notre système. Comme nous l'avons vu, les règles utilisées dans <TextCoop> ont une forme définie. Rappelons la forme "vide" de ces règles :

```
forme(Identifiant, Entrée, Sortie, Patron, Contraintes, Résultat).
```

La forme générale dédiée à l'identification et à la correction des erreurs de placement des adverbes de manière porte l'identifiant corr-adv, qui permet de distinguer ces règles de celles créées pour la gestion d'autres erreurs. Voici donc la forme de départ des règles de détection et correction pour ces erreurs, dans laquelle le patron, les contraintes et le résultat ne sont pas encore spécifiés :

```
forme(corr-adv, E, S, [], [], []).
```

Les espaces entre crochets accueillent la description du patron, des contraintes pesant sur le patron, et du résultat. Dans notre cas, le patron correspond au schéma de détection, le résultat à la correction de l'erreur, et les contraintes sont les exclusions ou précisions liées au schéma de détection. Les éléments de la description prennent la forme de "prédicats", dont la première partie est appelée "foncteur de prédicat", et dont la seconde partie donne les "arguments" (terme utilisé en Prolog avec le même sens que celui donné aux "arguments" d'un prédicat en linguistique). Voici un exemple de prédicat :

```
adv(ADV,_,E,E1)
```

Le foncteur est adv, et renvoie au lexique d'adverbes utilisé dans le système. Les arguments sont (ADV,_,E,E1). Le premier argument est une variable, reconnaissable comme telle parce que l'argument est donné en lettres majuscules. Cet argument sert à identifier l'élément concerné de manière symbolique, afin de pouvoir le reprendre dans le résultat, c'est-à-dire la partie dédiée à la correction. Le second argument est également une variable ; elle est laissée vide ici, mais nous verrons qu'elle peut être utilisée pour limiter les éléments repérés à un sous-ensemble des éléments représentés par le foncteur, par exemple pour limiter le repérage aux adverbes de manière.

Les deux derniers arguments (E, E1) servent à indiquer l'ordre dans lequel les éléments doivent se trouver dans le texte afin de correspondre à la structure à détecter. En effet, la syntaxe des langages de programmation logique est fondée sur des expressions logiques, ce qui signifie qu'une "suite" de prédicats n'est justement pas envisagée comme une suite, mais comme un ensemble d'éléments qui doivent être présents afin que la règle soit reconnue comme juste par le programme, et appliquée le cas échéant. Si l'on omet d'indiquer un ordre dans les arguments des prédicats formant nos patrons, le système relèvera des segments de texte incluant tous les éléments correspondant aux prédicats, mais ne respectant pas nécessairement l'ordre que nous recherchons (ex. : *[quickly] [opened] [the door]*, *[opened] [quickly] [the door]*, *[opened] [the door] [quickly]*). Ceci n'est évidemment pas souhaitable lorsque nous nous intéressons aux erreurs de placement. Afin d'éviter cette situation, les couples d'arguments de type (E, E1) indiquent le rang que l'élément concerné adopte dans la structure.

Le premier exemple que nous allons utiliser est le schéma d'erreur le plus central, c'est-à-dire le placement d'un adverbe entre un verbe et un GN objet, en nous intéressant aux adverbes de manière pour commencer. Le schéma linguistique simplifié donné pour cette erreur est le suivant :

Vlex + AdvM + GN

Cette description de l'erreur correspond au patron de la structure à relever, et est donc intégrée à l'emplacement du patron (NP, *Noun Phrase*, est mis pour GN). De manière simplifiée, la forme est la suivante (les zones de contraintes et de résultat sont laissées vides pour l'instant) :

forme(corr-adv, E, S, **[verb, adv, np]**, [], []).

Les éléments de la description doivent cependant respecter le format de programmation ; ils y sont donc intégrés sous forme de prédicats :

forme(corr-adv, E, S, **[verb(V,_,E,E1), adv(ADV,_,E1,E2), np(NP,_,E2,S)]**, [], []).

Cette première partie entre crochets permet l'identification de la structure erronée, c'est-à-dire la détection de l'erreur. On peut imaginer le patron comme une demande : existe-t-il dans le document soumis à l'analyse une séquence telle que celle décrite dans le patron ? Cependant, pour que le système puisse identifier une séquence dans un document en langue naturelle comme étant identique à la séquence donnée dans le patron, il faut qu'il puisse "reconnaître" les mots du texte, ainsi que leur catégorie. "Reconnaître" un mot signifie associer une chaîne de caractères à un terme défini dans un des lexiques intégrés au système (voir la section suivante pour la présentation des lexiques). Le type de mot recherché est

défini par le foncteur du prédicat (ex. : **adv**(ADV,_,E,E1)). Ainsi, si l'on souhaite qu'un patron recherche les adverbes dans un document, il faudra que le système ait accès à un lexique d'adverbes lui permettant de reconnaître certains mots comme des adverbes. Par ailleurs, si par exemple l'adverbe *unabashedly* ne fait pas partie du lexique des adverbes, il ne pourra pas être reconnu comme un adverbe par le système. Si ce terme inconnu fait partie d'une séquence, la séquence entière ne pourra être relevée.

Après avoir complété la zone du patron, il faut maintenant remplir la partie permettant de corriger l'erreur, qui correspond à la zone de résultat. Pour cet exemple, nous nous limitons à une seule proposition de correction consistant à placer l'adverbe avant le verbe. Le schéma de correction est le suivant :

Vlex + Adv + GN => Adv + Vlex + GN

Même si un programme informatique de correction automatisée semble annoter un texte original pour y identifier des erreurs, ce n'en est pas le fonctionnel réel : le programme identifie une structure en entrée, puis la remplace par une autre structure en sortie, dans laquelle l'erreur est annotée. Si la structure de départ est reproduite à l'identique, le remplacement est invisible, en dehors des indications qui ont été ajoutées. Une structure erronée peut également être corrigée directement, par exemple lorsqu'un traitement de texte corrige les erreurs lors de la frappe. La zone de patron permet l'identification d'une séquence erronée, et c'est la zone de résultat qui permet de lui substituer une séquence corrigée. Dans nos règles de correction, la correction ne remplace pas le segment à corriger, mais fournit une annotation d'erreur, suivit d'une proposition de correction également annotée comme telle. Le contenu de la zone de résultat devra donc avoir ce type de structure, indiquée ici en langage naturel :

Erreur[Vlex + Adv + GN], Correction[Adv + Vlex + GN]

Puisque l'on cherche à reproduire les mots exacts relevés dans le texte, la zone de résultat intègre les variables données dans le premier argument des prédicats présents dans le patron (ex. : ADV, V, NP). Voici la règle complétée, sans les annotations pour l'instant :

forme(corr-adv, E, S, [verb(V,_,E,E1), adv(ADV,_,E1,E2), np(NP,_,E2,S)],

[],

[V, ADV, NP, ADV, V, NP]).

Cette règle fonctionne, mais elle ne permet pas de distinguer l'erreur de la proposition de correction. Les annotations que nous avons choisies utilisent le format XML et sont des balises fermantes et ouvrantes servant à délimiter l'erreur et sa correction. Notons que ces

annotations sont indiquées entre guillemets simples, un signe de programmation en Prolog qui permet d'inclure n'importe quelle chaîne de caractères dans les règles ; ces annotations pourraient donc prendre des formes variées. Voici la règle, annotations incluses :

```
forme(corr-adv, E, S, [verb(V,_,E,E1), adv(ADV,_,E1,E2), np(NP,_,E2,S)],
[],
['<erreur>', V, ADV, NP, '</erreur>', '<correct>', ADV, V, NP, '</correct>']).
```

Grâce à ces marques supplémentaires, le segment erroné et la correction proposée sont clairement identifiés dans le texte original. Pour illustration, voici à quoi ressemblent les sorties du système :

He **<erreur> opened quickly the door </erreur>**
<correct> quickly opened the door </correct>.

Nous nous servons également de ces annotations pour indiquer le numéro du schéma concerné, en suivant les numéros donnés dans la modélisation linguistique.

Avant de présenter le contenu de la zone de contraintes, que nous avons laissée vide pour l'instant, revenons sur la façon dont le système reconnaît des séquences de mots dans un texte, et leur substitue la séquence de termes donnés dans la zone de résultat. Dans un premier temps, le système repère une correspondance entre une chaîne de caractères du texte et un des termes du lexique correspondant à la catégorie précisée par le foncteur du prédicat. Par exemple, dans la phrase **He opened slowly the door* correspondant à la règle utilisée comme illustration, *opened* est reconnu comme une des formes du verbe *open*. Les différentes formes des verbes sont spécifiées dans des fichiers différents, et un élément du programme permet d'associer ces formes à la forme de base du verbe. Le système fait donc correspondre *opened* avec la forme de base *open*. *Slowly* et *the door* sont reconnus comme un adverbe de manière et un GN respectivement. Les variables identifiées pour le segment du texte sont donc les suivantes :

```
V = 'open'
ADV = 'slowly'
NP = 'the door'.
```

Dans un second temps, le système construit le segment de sortie à partir de ces variables, en les faisant correspondre aux variables précisées dans la zone de résultat. Le segment de sortie vient remplacer le segment original. Si l'on applique la règle donnée plus haut, la sortie du système pour la phrase **He **opened** quickly the door* sera la suivante :

He <erreur> **open** quickly the door </erreur>

<correct> quickly **open** the door </correct>.

La variable identifiée pour *opened* étant la base du verbe, la forme au prétérit est remplacée par celle-ci. Il n'est cependant pas acceptable que les retours du système modifient de la sorte le texte original, d'autant plus que ce problème se pose pour tous les termes recevant des suffixes flexionnels, comme les noms. La forme à utiliser est spécifiée dans la zone de contraintes, qui contient les limites, spécifications ou exclusions s'appliquant sur le patron. Dans nos règles, la zone de contraintes est notamment utilisée pour créer un retour sur le patron, qui substitue à la variable V, par exemple, une autre variable correspondant à la chaîne de caractères située à l'emplacement spécifié (dans l'exemple qui suit, entre les bornes E1 et E de la phrase). L'exemple que nous avons utilisé jusqu'à présent comporte un verbe et un GN, il faut donc inclure ces deux contraintes :

forme(corr-adv, E, S, [verb(V,_,E,E1), adv(ADV,_,E1,E2), np(NP,_,E2,S)],
[conc(VE1,E1,E), conc(NP11,S,E2)],
['<erreur>', V, ADV, NP, '</erreur>', '<correct>', ADV, V, NP, '</correct>']).

Les nouvelles variables prennent ici la forme de VE1 et NP11, les variables V1 et NP1 étant utilisées dans le patron lorsque plusieurs verbes ou GN sont présents. Elles sont incluses dans la zone de contraintes grâce au prédicat incluant le foncteur conc. Pour que les termes apparaissent dans la sortie, il faut cependant remplacer les variables de base par celles-ci dans la zone de résultat :

forme(corr-adv, E, S, [verb(V,_,E,E1), adv(ADV,_,E1,E2), np(NP,_,E2,S)],
[conc(VE1,E1,E), conc(NP11,S,E2)],
['<erreur>', **VE1, ADV, NP11**, '</erreur>', '<correct>', **ADV, VE1, NP11**, '</correct>']).

La règle ci-dessus est la base des règles que nous utilisons pour la correction du placement des adverbes. Cette base peut ensuite être modifiée selon les besoins de la détection et de la correction. Les paragraphes ci-dessous présentent tous les ajouts effectués dans cette règle. Notons que les règles citées ci-dessous sont parfois simplifiées pour mieux focaliser la présentation sur un de leurs aspects précis. L'ensemble des règles que nous avons créées et qui sont utilisées pour la détection et la correction est présenté dans un tableau récapitulatif en Annexe 7. Les explications données ici devraient en simplifier la consultation.

Le premier ajustement important consiste à spécifier le type sémantique des adverbes que l'on cherche à repérer. La règle présentée ci-dessus repère tous les adverbes, puisque le second argument du prédicat de l'adverbe est laissé libre (adv(ADV,_,E1,E2)). L'ajout de la mention manner pour cet argument permet de limiter le repérage uniquement aux éléments dans le

lexique d'adverbes portant également l'argument manner, c'est-à-dire les adverbes identifiés comme des adverbes de manière d'après notre annotation manuelle du lexique :

forme(corr-adv, E, S, [verb(V,_,E,E1), adv(ADV,**manner**,E1,E2), np(NP,_,E2,S)],

[conc(VE1,E1,E), conc(NP11,S,E2)],

['<erreur>', VE1, ADV, NP11, '</erreur>', '<correct>', ADV, VE1, NP11, '</correct>']).

Pour la correction des erreurs liées à *also* et à *well,* la règle ne doit pas limiter le repérage à un type d'adverbe, mais à un seul adverbe. Dans ce cas, la variable donnée dans le premier argument du prédicat de l'adverbe dans le patron indique la forme exacte de l'adverbe recherché. Cette forme doit également être reproduite à l'identique dans la zone de résultat :

forme(corr-adv, E, S, [verb(V,_,E,E1), adv(**[also]**,_,E1,E2), np(NP,_,E2,S)],

[conc(VE1,E1,E), conc(NP11,S,E2)],

['<erreur>', VE1, **[also]**, NP11, '</erreur>', '<correct>', **[also]**, VE1, NP11, '</correct>']).

Cette solution s'applique également à l'ajout de l'adverbe *very* dans les schémas d'erreur, pour permettre le repérage des segments contenant un GAdv avec ce modifieur. Un prédicat supplémentaire est ajouté dans la zone de patron, dans lequel le premier argument est la forme exacte de l'adverbe. Ce nouvel élément doit bien sûr être également intégré à la zone de résultat :

forme(corr-adv, E, S, [verb(V,_,E,E1), adv(**[very]**,_,E1,E2), adv(ADV,manner,E2,E3), np(NP,_,E3,S)],

[conc(VE1,E1,E), conc(NP11,S,E3)],

['<erreur>', VE1, **[very]**, ADV, NP11, '</erreur>', '<correct>', VE1, NP11, **[very]**, ADV, '</correct>']).

Les règles dans <TextCoop> admettent également la presence d'éléments optionnels, ce qui permet d'éviter la création de plusieurs règles se distinguant uniquement par un élément de la zone de patron. Nous avons mis cette possibilité à profit dans plusieurs cas, notamment par l'inclusion de prépositions optionnelles permettant de relever les segments incluant des verbes à particule. Les prédicats optionnels sont inclus dans le patron avec un foncteur supplémentaire indiquant leur statut. Ils doivent également être inclus dans la zone de résultat, mais leur statut optionnel n'a pas besoin d'y figurer, car s'ils ne sont pas présents, la variable est vide :

forme(corr-adv, E, S, [verb(V,_,E,E1), **opt(prep(P,_,E1,E2))**, adv(ADV,manner,E2,E3), np(NP,_,E3,S)],

[conc(VE1,E1,E), conc(NP11,S,E3)],

['<erreur>', VE1, **P**, ADV, NP11, '</erreur>', '<correct>', ADV, VE1, **P**, NP11, '</correct>']).

Cette fonction est également utilisée pour permettre à la même règle de relever les segments incluant un GN ou un GP après le verbe, simplement en incluant une préposition optionnelle avant le GN :

forme(corr-adv, E, S, [verb(V,_,E,E1), adv([also],_,E1,E2), **opt(prep(P,_,E2,E3))**, np(NP,_,E3,S)],

[conc(VE1,E1,E), conc(NP11,S,E3)],

['<erreur>', VE1, [also], **P**, NP11, '</erreur>', '<correct>', [also], VE1, **P**, NP11, '</correct>']).

Dans la modélisation linguistique, nous avons mentionné plusieurs fois la possibilité d'exclure certains mots des règles, notamment les verbes, certains schémas ne s'appliquant pas à tous les éléments d'une catégorie. Par exemple, nous avons vu qu'il était fréquent de trouver *also* entre le verbe *see* et un GN ou GP, alors que ce schéma n'est généralement pas accepté pour les autres verbes lexicaux. Afin d'éviter des faux positifs avec le verbe *see,* il est donc nécessaire de l'exclure des verbes reconnus par le patron. Ces exclusions sont spécifiées dans la zone de contraintes, à l'aide du prédicat portant le foncteur not, qui indique les termes ne devant pas faire partie des éléments reconnus par le patron. L'exclusion porte sur la forme finale et non uniquement sur la forme de base, le prédicat contient donc la variable précisée juste avant dans la zone de contrainte :

forme(corr-adv, E, S, [verb(V,_,E,E1), adv([also],_,E1,E2), np(NP,_,E2,S)],

[conc(VE1,E1,E), conc(NP11,S,E2), **not(VE1 = [see])**],

['<erreur>', VE1, [also], NP11, '</erreur>', '<correct>', [also], VE1, NP11, '</correct>']).

Dans le cas de *see*, la configuration mentionnée ci-dessus n'est observée qu'avec la forme de base du verbe (ex. : *See also on page 2*), et non avec toutes ses formes possibles. Il s'est également avéré nécessaire d'intégrer des exclusions pour le verbe BE, notamment dans le traitement des erreurs liées à *also* pour éviter de relever les formes correctes du type *It is also a good idea to leave early*. Contrairement au cas de see, BE peut en théorie être présent dans toutes ses formes flexionnelles (ex. : *am, is, are, was, being*, etc.), ainsi que sous ses formes réduites incluant la négation (ex. : *isn't, wasn't*, etc.). Le défaut du recours à la variable finale, s'il est inévitable, est qu'il oblige à faire la liste de toutes les formes exclues dans la zone de contraintes :

forme(corr-adv, E, S, [verb(V,_,E,E1), adv([also],_,E1,E2), np(NP,_,E2,S)],

[conc(VE1,E1,E), conc(NP11,S,E2),

not(VE1 = [be]), not(VE1 = [am]), not(VE1 = [are]), not(VE1 = [is]), not(VE1 = [was]), not(VE1 = [were]),

not(VE1 = [isn,simplequote,t]), not(VE1 = [aren,simplequote,t]), not(VE1 = [wasn,simplequote,t]), not(VE1 = [weren,simplequote,t]),

not(VE1 = [been]), not(VE1 = [being]),],

['<erreur>', VE1, [also], NP11, '</erreur>', '<correct>', [also], VE1, NP11, '</correct>']).

L'apostrophe présente dans les formes abrégées de BE doit être représentée par le terme simplequote, puis que ce symbole typographique a une fonction propre dans la programmation des règles dans <TextCoop>.

Certains des schémas reposent sur la possibilité de distinguer les GN courts des GN longs, car cette caractéristique influence le placement de l'adverbe. Cette distinction est effectuée par le biais d'une indication du nombre de mots présents dans le GN relevé par le patron. Le prédicat portant le foncteur card permet d'identifier le nombre de mots dans le GN (ici indiqué par la variable NP11), et celui-ci est couplé à l'indication d'un seuil numérique (ex. : (T > 4)). Ceci contraint le patron à ne relever que les segments comprenant des GN dont le nombre de mots est supérieur ou inférieur à un seuil déterminé :

forme(corr-adv, E, S, [verb(V,_,E,E1), adv(ADV,manner,E1,E2), np(NP,_,E2,S)],

[conc(VE1,E1,E), conc(NP11,S,E2),

card(NP11, T), (T > 4)],

['<erreur>', VE1, ADV, NP11, '</erreur>', '<correct>', ADV, VE1, NP11, '</correct>']).

La règle ci-dessus indique donc que le GN faisant partie du segment relevé doit être constitué de plus de quatre mots. La correction indiquée pour ce segment consiste à placer l'adverbe avant le verbe en raison de la longueur du GN. Le choix du seuil entre GN court et GN long est subordonné à la grammaire du GN que nous avons créée pour la reconnaissance de ces groupes, et qui est relativement simple et non-exhaustive. La partie consacrée à l'évaluation des règles revient sur les limites de cette grammaire. Les GN longs sont identifiés comme incluant au moins cinq termes, de façon à pouvoir avoir la forme minimale suivante : [Det + N + Prep + Det + N]. Ils peuvent cependant avoir une forme différente. Les GN identifiés comme courts comprennent au plus quatre termes.

La dernière caractéristique importante à exposer avant de regarder de plus près les règles créées pour le traitement des séquences N+N est la production de plusieurs propositions de correction. Il n'est pas possible d'inclure plusieurs zones de résultats dans la règle, mais il est possible d'intégrer les deux propositions de correction dans la même zone de résultat. Cette méthode permet d'éviter la répétition de règles ayant la même zone de patron, éliminant par la

même occasion le risque que les règles alternatives ne soient pas lues par le système, et que les autres corrections ne soient donc pas proposées à la personne utilisatrice. La première règle de correction donnée pour le placement des adverbes, qui se trouve d'ailleurs être la règle la plus fréquemment mobilisée lors de l'évaluation (voir section 3.2.2b "Evaluation des règles de correction pour le placement des adverbes"), intègre deux possibilités de correction. L'intégration de ces deux corrections dans la zone de résultat est observable ci-dessous :

```
forme(corr-adv, E, S, [verb(V,_,E,E1), opt(prep(P,_,E1,E2)), adv(ADV,manner,E2,E3),
np(NP,_,E3,S)],

[conc(VE1,E1,E), conc(NP11,S,E3), card(NP11, T), (T < 5),

not(VE1 = [be]), not(VE1 = [am]), not(VE1 = [are]), not(VE1 = [is]), not(VE1 = [was]),
not(VE1 = [were]), not(VE1 = [isn,simplequote,t]), not(VE1 = [aren,simplequote,t]),
not(VE1 = [wasn,simplequote,t]), not(VE1 = [weren,simplequote,t]), not(VE1 = [been]),
not(VE1 = [being]),

not(ADV = [fast])],

['<erreur>', VE1, P, ADV, NP11, '</erreur>',

'<correct1>', ADV, VE1, P, NP11, '</correct1>',

'<correct2>', VE1, P, NP11, ADV, '</correct2>']).
```

L'autre règle qui intègre plusieurs propositions de correction est consacrée à la correction des erreurs N+N de la catégorie "N_1 génitif", et propose quatre solutions de correction différentes dans la zone de résultat.

Les règles de correction pour les séquences N+N utilisent la même forme de base ainsi que les mêmes fonctionalités. Elles ont cependant quelques spécificités qu'il est nécessaire de mentionner. Pour commencer, les formes que nous avons présentées jusqu'à maintenant incluent des prédicats pour le relevé de GN, mais pas pour les noms seuls. Pour les règles de correction des séquences N+N, nous avons besoin d'inclure des prédicats pour des noms seuls. Le prédicat portant le foncteur nconj est utilisé afin de permettre le repérage des noms au singulier et au pluriel. Puisque les séquences N+N à repérer sont par définition composées de plusieurs noms, la variable donnée dans le premier argument du prédicat doit être numérotée afin de pouvoir distinguer les différents noms du patron. La règle créée pour la détection des segments N+N "empilés" ayant le schéma [(Det) + N + N + N + N] est la suivante, sans les zones de contraintes et de résultat pour l'instant :

```
forme(corr-nn, E, S, [opt(det(DET,_,E,E1)), nconj(NOM1,_,E1,E2),
nconj(NOM2,_,E2,E3), nconj(NOM3,_,E3,E4), nconj(NOM4,_,E4,S)],
[], []).
```

Notons aussi que le déterminant est intégré dans le patron de manière optionnelle. Ici également, la variable renvoie à la forme de base du nom, qui est au singulier. Pour que la sortie du système soit identique au texte original, il faut donc modifier cette variable dans la zone de contraintes :

```
forme(corr-nn, E, S, [opt(det(DET,_,E,E1)), nconj(NOM1,_,E1,E2),
nconj(NOM2,_,E2,E3), nconj(NOM3,_,E3,E4), nconj(NOM4,_,E4,S)],
[conc(N1,E2,E1), conc(N2,E3,E2), conc(N3,E4,E3), conc(N4,S,E4)],
[]).
```

Le traitement des erreurs d'empilement s'arrête à la détection des erreurs et ne propose pas de correction. La zone de résultat ne comporte donc que la partie consacrée à la délimitation de l'erreur. La partie consacrée à la correction est laissée vide :

```
forme(corr-nn, E, S, [opt(det(DET,_,E,E1)), nconj(NOM1,_,E1,E2),
nconj(NOM2,_,E2,E3), nconj(NOM3,_,E3,E4), nconj(NOM4,_,E4,S)],
[conc(N1,E2,E1), conc(N2,E3,E2), conc(N3,E4,E3), conc(N4,S,E4)],
['<erreur>', DET, N1, N2, N3, N4, '</erreur>', ' <correct>', '</correct>']).
```

Le traitement des erreurs de la catégorie "N_1 génitif" (ex. : **the objects properties*) est un peu plus complexe, puisque la règle concernée donne quatre corrections différentes, qui ne consistent pas seulement en la réorganisation du segment, mais nécessitent aussi l'ajout d'éléments extérieurs. Nous sommes ainsi amenée à modifier le nombre d'un nom et/ou à ajouter des éléments de ponctuation ou des prépositions. Les spécificités de chaque proposition de correction sont présentées ci-dessous.

La première proposition de correction pour des segments du schéma [Det + (Adj) + Ns + N] consiste à leur substituer la forme [Det + (Adj) + Ns + ['] + N]. La correction passe par l'inclusion d'un signe de ponctuation marquant le génitif. Cette modification simple est ajoutée dans la partie consacrée à la correction dans la zone de résultat, sous la forme de l'élément simplequote :

```
forme(corr-nn, E, S, [det(DET,def,E,E1), nconj(NOM1,plu,E1,E2), nconj(NOM2,_,E2,S)],
[conc(N1,E2,E1), conc(N2,S,E2)],
['<erreur>', DET, ADJ, N1, N2, '</erreur>',
'<correct>', DET, ADJ, N1, [simplequote], N2, '</correct>']).
```

La seconde proposition de correction pour cette erreur modifie le nombre du N_1, le faisant passer du pluriel au singulier. Elle consiste aussi en l'ajout d'une marque de ponctuation, complétée par un *s* seul afin de recréer la marque du génitif, puisque le nom est désormais au

singulier : [Det + (Adj) + N + ['s] + N]. La modification du nombre du N_1 est effectuée en s'appuyant sur les variables données dans la zone de résultat. La variable donnée dans la partie représentant l'erreur est celle qui est spécifiée dans les contraintes, c'est-à-dire le mot présent dans le texte, qui est ici la forme plurielle ; celle qui est donnée dans la partie représentant la correction est la variable présente pour ce nom dans la zone de patron, et qui est donc la forme de base du mot, ou forme au singulier. On obtient la règle suivante :

forme(corr-nn, E, S, [det(DET,def,E,E1), nconj(NOM1,plu,E1,E2), nconj(NOM2,_,E2,S)],
[conc(N1,E2,E1), conc(N2,S,E2)],
['<erreur>', DET, ADJ, **N1**, N2, '</erreur>',
'<correct>', DET, ADJ, **NOM1**, [simplequote], [s], N2, '</correct>']).

Les troisième et quatrième propositions de correction pour cette erreur consistent en une réorganisation du segment, faisant passer la structure N+N à une structure dans laquelle le nom noyau a un dépendant post-noyau prenant la forme d'un GP. En plus du déplacement du N_1, cette correction appelle plusieurs modifications : l'ajout d'une préposition (*of* est utilisée par défaut), et l'ajout de l'article *the* en début de GN, le déterminant défini du segment de départ étant déplacé avec le N_1. Le N_1 est mis au singulier dans la troisième proposition, et la forme au pluriel est conservée dans la quatrième proposition. Les deux règles correspondantes sont présentées ci-dessous :

forme(corr-nnd, E, S, [det(DET,_,E,E1), nconj(NOM1,plu,E1,E2), nconj(NOM2,_,E2,S)],
[conc(N1,E2,E1), conc(N2,S,E2)],
['<erreur>', DET, N1, N2, '</erreur>',' **<correct>', [the], N2, [of], DET, N1, '</correct>'**]).

forme(corr-nne, E, S, [det(DET,_,E,E1), nconj(NOM1,plu,E1,E2), nconj(NOM2,_,E2,S)],
[conc(N1,E2,E1), conc(N2,S,E2)],
['<erreur>', DET, N1, N2, '</erreur>','<**correct>', [the], N2, [of], DET, NOM1, '</correct>'**]).

Ces deux règles concluent notre présentation de l'implémentation de règles de détection et correction dans <TextCoop> à partir de schémas linguistiques. L'ensemble des règles est donné en Annexe 7. La section suivante est consacrée à l'exposé des ressources utilisées pour le fonctionnement des règles de correction.

c. Détail des ressources utilisées pour l'implémentation des règles

Les règles présentées dans la section précédente donnent déjà un aperçu des ressources nécessaires pour la correction. Ces ressources sont de deux types : les ressources lexicales, et

les grammaires locales. Les ressources lexicales concernent quasiment la totalité des catégories linguistiques, puisque les deux erreurs touchent des domaines distincts et centraux dans la proposition, c'est-à-dire le GV et le GN.

Les ressources lexicales se présentent sous la forme de listes de mots programmées en Prolog afin qu'elles puissent être lues par <TextCoop>. Aux listes de mots se rajoutent des informations morphologiques permettant la reconnaissance des différentes formes des mots soumis à la flexion. Dans certains cas, ces informations sont ajoutées par le biais de listes dupliquées, comme pour les verbes : quatre listes créées à partir des mêmes bases verbales donnent les formes finies et non-finies des verbes (V-*ed*/*-en*, V-*ing*, V-*s*, infinitif), et une cinquième liste regroupe les formes irrégulières des verbes concernés (V-*ed* et V-*en*). Dans d'autres cas, des règles ont été créées et intégrées au moteur du système afin de générer les différentes formes de certains éléments. C'est le cas des noms, dont les formes au pluriel sont reconnues automatiquement dans les textes grâce à l'inclusion de règles de reconnaissance dans le moteur. Nous revenons sur ces modifications dans les paragraphes qui suivent le tableau récapitulatif des ressources lexicales utilisées dans le système.

La forme suivante est un exemple d'entrée lexicale pour les noms dans <TextCoop> :

```
n([abolition], [nc,nd]) --> [abolition].
```

La partie de gauche est un prédicat donnant des informations sur le terme concerné. La catégorie de ce mot est indiquée par le foncteur du prédicat, c'est-à-dire la lettre située avant la parenthèse ; il s'agit ici d'un nom. Les informations incluses dans la parenthèse sont présentées comme des constantes, c'est-à-dire entre crochets. Le premier élément donne la forme terminale du terme ([abolition]), alors que le second élément donne des attributs élémentaires du terme ; ici ils indiquent que le nom est un nom commun (nc) et qu'il est indénombrable (nd). La partie de droite fait correspondre à ce prédicat le nom concerné. Pour les noms, elle est identique à la forme présente dans la partie de gauche car, comme nous l'avons dit, les modifications morphologiques sont traitées par des règles pour cette catégorie. Dans le cas des verbes, pour lesquels il existe plusieurs listes, la seconde partie de la forme porte les indications morphologiques terminales :

```
verb-engS(like, troispers) --> [likes].
```

L'ensemble des ressources lexicales est donné dans le tableau qui suit. Chaque type de lexique est accompagné de son format dans <TextCoop>, de la taille de chaque ressource en nombre de mots, et de la source initiale de la liste de termes (dont les références sont données dans notre bibliographie). Il faut noter que ces sources, libres de droits, ont simplement servi

de base. Nous avons en effet examiné chaque liste afin d'en éliminer les membres peu probables afin d'alléger le système (ex. : *trapezoid* (nom), *acetify* (verbe), *spiffy* (adjectif)), puis nous avons ajouté des entrées manquantes au fur et à mesure du développement des ressources et du système. Dans certains cas, il a également été nécessaire de "nettoyer" les listes utilisées afin de retirer des termes ne faisant pas partie de la catégorie concernée. Nous avons également ajouté des attributs aux éléments lexicaux, par exemple pour les adverbes et les noms.

Certaines ressources proviennent directement de la version de <TextCoop> que nous utilisons, c'est donc cette plateforme qui est indiquée comme source. Dans le cas des déterminants, nous avons constitué nous-mêmes la liste de termes. Les paragraphes suivant le tableau donnent des informations supplémentaires concernant le traitement morphologique de chaque catégorie linguistique incluse dans nos ressources, ainsi que les attributs que nous avons ajoutés.

Catégorie	**Format**	**Taille**	**Source**
Noms	n([abolition], [nc,nd]) --> [abolition].	2439	Site internet : *Moms Who Think* Site internet : *Desi Quintans*
Pronoms	pro([you],_) --> [you].	35	
Adjectifs	adj_eng([vegan], _) --> [vegan].	1098	Site internet : *Moms Who Think*
Prépositions	prep([of], _) --> [of].	83	Site internet : *Wikipedia*
Déterminants	det([a], ndef) --> [a].	55	
Verbes	verb_eng(like, nonfinite) --> [like]. verb_engING(like, ing) --> [liking]. verb-engS(like, troispers) --> [likes]. verb_engED(like, preterit) --> [liked]. verbIrr(write, wrote, written).	3142	<TextCoop>
Adverbes	adv([nicely], manner) --> [nicely].	1242	Site internet : *Paul Noll* Site internet : *Moms Who Think*
Auxiliaires	aux([can],_) --> [can].	35	
TOTAL	**8129**		

Tableau 49. Liste des ressources lexicales utilisées

Les entrées du lexique de noms incluent des attributs indiquant si le terme est un nom propre ([p]) ou un nom commun ([nc]), et si le nom est dénombrable ([den]) ou indénombrable ([nd]). Ces informations ont été entrées manuellement. Elles ne sont pas mobilisées dans la présente version de nos règles de correction, mais ont été utilisées dans des

versions précédentes. Le pluriel des noms est généré automatiquement grâce à un fichier supplémentaire dans <TextCoop>, créé à partir d'informations linguistiques concernant les pluriels exceptionnels (ex. : *mouse, mice*) et les modifications de la base liées à la lettre de fin (ex. : *knife, knives* / *butterfly, butterflies*). Le lexique des déterminants indique si ceux-ci sont définis ([def]) ou indéfinis ([ndef]).

Le lexique de verbes est en réalité composé de cinq fichiers différents. Le premier est le fichier de base, et trois d'entre eux sont des duplications de ce fichier. Ils représentent les quatre formes possibles pour un verbe, indiquées dans les attributs : infinitif ([nonfinite]), verbe avec suffixe *-ing* ([ing]), verbe avec suffixe *-s* ([troispers]), verbe avec suffixe *-ed* ([preterit]). Certaines modifications ont été effectuées automatiquement (ex. : omission du *-e* final à l'ajout du suffixe *-ing*, redoublement de consonnes finales), et les listes ont été finalisées manuellement. Le cinquième fichier est une liste de verbes irréguliers donnant les formes du prétérit et du participe passé de ces verbes, et dont les entrées sont légèrement différentes (voir tableau).

Le lexique d'adverbes contient des informations concernant le type sémantique de l'adverbe concerné. Nous avons pour l'instant inclus cette annotation uniquement pour les adverbes de manière. Comme nous l'avons vu dans le Chapitre 2, le type sémantique des adverbes a une grande influence sur son placement ; un système de correction ayant pour ambition de détecter les erreurs de placement concernant tous les adverbes doit donc inclure ce type d'information sémantique. L'intégration des informations liées à la ressource lexico-sémantique d'adverbes anglais présentée dans le Chapitre 2 est un des objectifs de développement de notre système. Ces informations ont été utilisées pour la spécification des adverbes de manière dans la version actuelle du système.

Les ressources grammaticales utilisées par les règles se limitent à une grammaire locale du GN, directement intégrée au système. Celle-ci sert à la reconnaissance des GN dans les règles de détection et correction liées aux adverbes, et nous évite d'avoir à décrire les GN à l'intérieur des règles. Une grammaire élémentaire du GAdj nous permet de la même manière d'utiliser la récursivité des règles dans <TextCoop> pour intégrer des GAdj dans les GN sans avoir à les décrire. L'avantage de l'utilisation de ces types de ressource est leur compatibilité avec la plateforme utilisée, ainsi que la possibilité d'y faire appel directement sans avoir à passer par un outil d'analyse extérieur. Elles ont cependant des inconvénients, à commencer par le fait qu'elles doivent être préparées manuellement. Dans notre cas, cela signifie que la grammaire locale du GN est loin de décrire toutes les configurations possibles dans la syntaxe

du GN. Cette grammaire décrit 16 configurations différentes, intégrant des GAdj, de un à quatre noms et des GP dépendants introduits par *of*. La règle ci-dessous est un exemple des règles utilisées pour cette grammaire :

```
np([A,B,C,D,F],_,E,S) --> det(A,_,E,S1), adjp(B,_,S1,S2), nconj(C,_,S2,S3),
prep_of(D,_,S3,S4), np(F,_,S4,S).
```

Cette règle permet de repérer des GN semblables à *the amazing depth of his knowledge* ou *a very pretty shade of minty green*. Les règles de la grammaire du GN utilisent la même base de formation et suivent les mêmes consignes d'organisation que le système des règles de correction exposé dans la section précédente. La partie gauche de la règle est un prédicat dont le foncteur est `np` ; le premier argument décrit une suite de quatre éléments, qui sont spécifiés dans la partie droite de la règle. Les règles étant récursives, les GN relevés peuvent inclure plus de quatre éléments, comme les exemples cités ci-dessus.

Les limites des ressources grammaticales et lexicales utilisées dans notre système sont exposées dans la sous-partie suivante, consacrée à l'évaluation des règles de détection et correction. Nous verrons que l'étendue de ces ressources a un impact important sur la qualité de la détection et de la correction. Comme les autres modules de notre système, ceux-ci seront bien sûr amenés à évoluer lors de recherches futures.

3.2.2 Évaluation des règles de détection et correction automatisées dans notre recherche

a. Méthode pour l'évaluation des règles

L'objectif de l'évaluation dans notre recherche est double. Tout d'abord, tester l'application des règles sur un nouveau corpus est indispensable afin d'en éprouver les performances. Ensuite, l'évaluation permet de révèler les angles morts des règles, et ainsi d'améliorer ces dernières.

Afin d'être pertinent, un système de correction grammaticale automatisée doit repérer et corriger efficacement la plupart des erreurs, éventuellement d'un seul type, tout en ne signalant pas d'erreur sur des segments corrects, c'est-à-dire en ne générant pas de faux positifs. Le second aspect de l'évaluation n'a certes pas été pris en compte dans l'étude des capacités des correcteurs grammaticaux présentée en section 3.1.2, puisque nous souhaitions avant tout évaluer leurs réactions aux erreurs relevées dans notre corpus. L'évaluation de la précision est cependant l'aspect le plus important de l'évaluation de nos règles : en effet, comme l'ont notamment souligné Tschichold (1999) et Heift et Schulze (2007), limiter le

nombre de faux positifs est une priorité pour les systèmes de correction à destination de publics non-natifs.

Idéalement, un système devrait être évalué sur le même type de productions que celles à destination desquelles il a été conçu ; dans notre cas, il s'agit de textes produits par des francophones en anglais à un niveau intermédiaire à avancé (B2-C1). Cette exigence est cependant très difficile à satisfaire dans notre cadre, pour trois raisons principales. Tout d'abord, les corpus d'interlangue de ce type exact sont rares, et nous ne disposons ainsi que d'une quantité limitée de documents. Cette difficulté est cependant surmontable, puisque le sous-corpus de productions de francophones du corpus *ICLE* contient près de 230 000 mots, dont nous n'avons utilisé que 50 000 pour notre corpus exploratoire. Deux autres aspects s'avèrent plus problématiques : d'une part, les vérifications et l'analyse des retours du système sont effectuées manuellement, et d'autre part les textes cibles ont des taux d'erreurs relativement bas, comme notre analyse des erreurs relevées dans le corpus exploratoire l'a montré. Ce dernier paramètre rend nécessaire l'utilisation de très grandes quantités de corpus pertinents si l'on souhaite obtenir des résultats non-anecdotiques. En association aux deux autres paramètres, cette exigence nous oblige à rechercher d'autres moyens d'évaluer nos règles de correction.

Notre méthode d'évaluation est fondée sur l'utilisation complémentaire de corpus d'anglais L1 et L2. Dans un premier temps, afin d'évaluer la précision des règles, c'est-à-dire la présence de faux positifs, nous avons constitué un corpus d'environ 80 000 mots de productions en anglais L1 issues d'articles de presse, d'articles scientifiques et de notes de blogs disponibles sur internet, et dont les auteurs sont strictement anglophones d'après les informations dont nous disposons. Ces types de documents sont relativement similaires aux types de production utilisés dans notre corpus exploratoire, et susceptibles d'être produites par des personnes utilisatrices de l'anglais (nous n'incluons pas de textes de fiction, par exemple). Les variétés d'anglais présentes dans ce corpus sont les variétés britannique et nord-américaine. Le tableau 50 en donne la composition synthétique ; les sources exactes sont données en Annexe 8.

Type de document	Variété	Nombre de mots
Articles scientifiques	*n/d*	27 891
Articles de presse	Brit.	10 724
	Amér.	14 803
Notes de blogs	Brit.	8 688
	Amér.	19 027
Total		**81 133**

Tableau 50. Détail du corpus d'anglais L1 pour l'évaluation des règles

La rareté des corpus de productions d'interlangue utilisables pour la génération automatique de règles de détection/correction ainsi que pour l'évaluation de tels systèmes est un problème récurrent dans le domaine de la correction grammaticale automatisée. Ceci a amené certains chercheurs à utiliser des ressources ou des corpus modifiés de façon à les faire ressembler à des productions en interlangue. C'est notamment le cas dans les travaux de Bigert et al. (2004), Brockett et al. (2006) et Wagner et al. (2007). Plus particulièrement, Foster et Andersen (2009) décrivent le développement du système *GenERRate*, un outil de génération automatique d'erreurs pouvant être paramétré pour modifier un corpus afin d'y introduire des erreurs ressemblant à celles qui sont produites par les personnes apprenantes.

Nous avons opté pour une solution similaire, sans toutefois automatiser le processus de modification. Le corpus d'anglais L2 utilisé pour les tests est composé d'un extrait du corpus *ICLE* (8804 mots), ainsi qu'un corpus de courriels similaires à ceux utilisés dans notre corpus exploratoire (1146 mots). Des erreurs ont été introduites manuellement dans ce corpus par une ingénieure de recherche. Cette personne avait pour consigne d'introduire des erreurs correspondant globalement aux types pris en compte dans nos travaux, mais en utilisant également les termes dans des structures correctes, avec un ratio d'environ une erreur pour deux usages corrects. La consigne la plus importante concernait l'aspect naturel du texte produit, les erreurs devant correspondre à des phénomènes plausibles pour le niveau de maîtrise de l'anglais et la langue source des auteurs des textes originaux (ex. : **I have also good news,* ou *?I like going also in the summer* mais pas **I like going in the also summer*). La personne ayant effectué cette tâche est francophone avec un niveau avancé en anglais (C2), et n'est pas familiarisée avec le contenu de nos règles. Cette méthode a été appliquée à un corpus d'environ 10 000 mots, incluant les types de documents indiqués plus hauts. Le corpus servant de base aux modifications est le même pour les différents types d'erreur évalués. Une fois le corpus modifié, celui-ci a été annoté afin de signaler les erreurs en amont de la phase

d'évaluation des règles de détection et de correction, et ainsi d'éviter certains biais au moment de l'évaluation. Afin de tester les limites des règles, le signalement des erreurs concerne également les segments non-standards, mais ne constitue en aucun cas un jugement strict sur leur statut de grammaticalité ou d'acceptabilité.

Nous utilisons le même type de classement que dans notre étude des correcteurs grammaticaux, en distinguant les étapes de la détection et de la correction, avec quelques modifications. Pour l'évaluation de la précision menée sur le corpus d'anglais L1, nous adoptons une approche stricte, en considérant que toute détection déclenchée par le système constitue un faux positif, même si la correction proposée constitue une alternative possible au segment original (ex. : *I hope also that this year…* / correction proposée par le système : *I also hope that this year…*). Cette approche est motivée par le fait que, comme le note Tschichold (1999 ; exemplaire non paginé), les personnes apprenantes utilisatrices des correcteurs grammaticaux ne possèdent pas nécessairement les compétences requises pour interpréter les retours de ces systèmes. Même à un niveau intermédiaire, un diagnostic d'erreur sur un segment correct, que la correction proposée soit acceptable ou non, constitue une indication négative concernant le texte original, et pourrait amener la personne apprenante à mettre en doute son utilisation de la structure en question. Les cas de faux positifs produits par le système sont analysés afin d'en déceler les causes et de pouvoir y remédier dans des travaux futurs.

Nous avons élaboré une classification des types de réponse du système qui est adaptée à notre cadre technologique, mais peut être utilisée pour d'autres systèmes, notamment ceux qui utilisent des méthodes non-statistiques. Il présente l'intérêt d'inclure une analyse des causes probables des échecs du système évalué. Les phases de détection et de correction sont liées, puisque la qualité potentielle d'une correction est subordonnée à la précision de la détection préalable ; ce classement concerne donc ces deux phases et reflète leur hiérarchie. À chaque retour du système sont associées une indication sur la qualité de la détection (correcte, inexistante ou erronée), et une indication concernant la correction si cela est pertinent (correction adéquate, n'améliore pas l'original, ou est pire que l'original). Des informations supplémentaires concernant les causes probables d'échec sont données le cas échéant. Nous avons attribué des codes aux différentes configurations possibles ; ceux-ci peuvent servir à l'annotation des retours du système. Le tableau 52 présente l'ensemble des indications utilisées ainsi que les codes qui leur sont associés.

Détection		**Correction**		**Analyse**	
Code	**Définition**	**Code**	**Définition**	**Code**	**Définition**
ED	Erreur détectée	C	Correction adéquate		
		CNB	C. n'améliore pas l'original		
		CW	Correction pire que l'original		
END	Erreur non détectée			PAT	Insuffisance dans les règles
				GR	Insuffisance dans la grammaire
				LEX	Insuffisance dans le lexique
F	Faux positif			TXT	Problème dans le co-texte
				UNK	Raison inconnue

Tableau 51. Système de classement pour l'évaluation de règles de correction automatisée

Quatre causes principales d'échec sont identifiées, et s'appliquent aussi bien à la non-détection qu'à la présence de faux positifs. Les trois premières sont en lien direct avec la qualité des règles de correction et des ressources mobilisées, alors que la dernière concerne le texte sur lequel la détection s'applique. En voici le détail :

- une insuffisance dans les règles, c'est-à-dire une règle inexistante, par exemple pour des erreurs dont la configuration est rare, ou bien une règle qui est activée sur un segment correct, indiquant que celle-ci devrait être affinée ;
- une insuffisance dans les ressources grammaticales du système, par exemple dans la délimitation des GN ; les problèmes posés par des homonymies (ex. : *purchase* (n.)/*purchase* (v.)) sont regroupées dans cette catégorie puisque seule l'analyse grammaticale permet de différencier les homonymes ;
- une insuffisance dans le lexique du système, c'est-à-dire lorsqu'une erreur est mal ou non-détectée en raison d'un mot manquant dans les ressources lexicales de la plateforme ;
- un problème dans le co-texte de l'erreur, par exemple lorsque plusieurs erreurs sont superposées, rendant l'erreur traitée invisible au système.

Nous avons également intégré une cinquième catégorie pour les cas de non-détection ou de faux positifs aux causes obscures. Notons que toutes les catégories ne sont pas représentées dans les résultats des tests, mais cette catégorisation exhaustive peut être appliquée à d'autres tests dans des recherches futures.

D'après ce système, la réussite des règles est exprimée par le code [ED – C], indiquant une détection et une correction adaptées. Dans notre approche stricte, tous les autres cas correspondent à des échecs, puisqu'ils ne permettent pas d'améliorer le segment original et ont donc peu d'intérêt pour un public apprenant. La distinction entre segments erronés et acceptables a été effectuée à partir des analyses linguistiques présentées dans le Chapitre 2. Notons par ailleurs que l'évaluation a été effectuée par une seule personne, ce qui constitue une limite importante dans la représentativité des résultats.

Les deux premières colonnes de ce tableau ne s'appliquent pas à la première étape d'évaluation, c'est-à-dire sur l'évaluation des règles sur un corpus d'anglais L1, puisqu'il ne peut y avoir ni détection ni correction adéquate possible, mais certaines des catégories de la dernière colonne sont utilisables pour la qualification des faux positifs relevés.

En raison de son caractère artificiel, l'utilisation d'un corpus modifié ne permet pas de tirer de conclusions définitives concernant l'adéquation des règles et leur efficacité sur des corpus naturels, mais peut servir à les affiner en rendant visibles des difficultés que nous n'aurions pas anticipées. Aucun autre élément du texte n'ayant été retouché, ces tests sont tout de même indicatifs de l'application des règles à des textes naturels pour des aspects qui sont liés indirectement au placement des adverbes, comme la gestion de la reconnaissance des GN. Les résultats de l'évaluation sont ainsi analysés en détail afin d'en extraire les informations nécessaires à l'amélioration future des règles.

Seules les règles de correction du placement d'*also* et des adverbes de manière ont été soumises à une évaluation sur un corpus d'anglais L2. L'évaluation sur corpus d'anglais L1 est par contre effectuée pour l'ensemble des règles. Comme nous le verrons ci-dessous, les règles de détection des erreurs liées à des empilements et des formes génitives sous-jacentes dans les structures N+N donnent lieu à un nombre important de faux positifs dans les corpus en anglais natifs. Par ailleurs, la méthode décrite ci-dessous permettant de créer un corpus de test n'est pas adaptée à l'évaluation de ces erreurs. L'introduction d'adverbes dans des phrases existantes est relativement aisée, puisqu'il ne s'agit que d'ajouter une information supplémentaire simple et contenue en un mot. Par contre, l'introduction d'erreurs liées aux structures N+N implique la modification importante de GN existants ou l'invention de GN non-existants, ce qui constitue une tâche très lourde pour la personne modifiant le corpus. De plus, les erreurs liées aux structures N+N sont courantes dans les publications, représentant près de 16 % des erreurs dans ce sous-corpus, il semble donc envisageable d'utiliser un corpus non-modifié de publications scientifiques pour ces tests. Enfin, nous avons donné dans la

section 2.3.2b un ensemble de solutions à mettre en place pour la détection des erreurs dans les structures N+N des autres catégories, notamment celles liées aux relations sémantiques, qui sont les plus fréquentes. Nous attendons les résultats de ces recherches futures afin de mener une évaluation plus complète.

La section suivante présente les résultats de l'évaluation des règles et l'analyse de ses résultats, ainsi que des pistes de remédiation lorsque les résultats ne sont pas jugés satisfaisants.

b. Résultats de l'évaluation des règles et discussion des résultats

Évaluation des règles liées aux structures N+N : erreurs d'empilement et de N+N génitif

Les règles de correction de ces deux sous-types d'erreur n'ont été évaluées que sur le corpus d'anglais L1. En effet, cette évaluation a révélé l'existence d'un nombre élevé de faux positifs. Dans ces conditions, il paraît plus efficace d'affiner les règles avant de les tester sur un corpus d'anglais L2 dans des recherches futures. Le tableau suivant donne le détail de ces faux positifs avec des exemples représentatifs, sous la forme de segments du texte d'origine, dans lesquels la portion jugée erronée par le système est soulignée si elle ne correspond pas à la totalité du segment (l'abréviation "Occ." renvoie au terme "occurrences") :

Règle concernée	Code	Exemples	Occ.
Empilement	GR	*Last month security forces…* *Hearing children learning English…*	4
	PAT	*An air force flight surgeon* *A Saturday morning cartoon group shot*	5
N_1 génitif	GR	*These examples show…* *A site that features properties…* *Upon these conditions children's happiness*	18
	PAT	*The games industry* *The sales manager*	3

Tableau 52. Erreurs N+N : analyse des faux positifs dans le corpus anglais L1

Les règles ont engendré neuf faux positifs concernant les structures empilées, et 21 pour les N+N génitif. Ces chiffres sont donnés en valeurs absolues, et non relatives comme ce sera le cas pour les adverbes : étant donné la variété des structures concernées, c'est-à-dire les GN incluant plusieurs noms, il ne nous a pas été possible de confronter les faux positifs à

l'ensemble des structures du même genre présentes dans le corpus, ce qui aurait nécessité de les relever manuellement.

Comme le montre le tableau, les faux positifs sont dus à des insuffisances dans les ressources grammaticales du système et au contenu des règles de correction pour les deux types d'erreurs. Les faux positifs dus aux ressources grammaticales sont particulièrement fréquents pour les erreurs "N_1 génitif". Ceci s'explique par le fait que les patrons de détection pour ces erreurs sont beaucoup plus simples que pour les erreurs d'empilement, et font remonter plus de segments.

Pour les erreurs d'empilement, les faux positifs dus aux ressources grammaticales sont le résultat soit d'erreurs dans la délimitation des frontières de plusieurs GN, soit d'homonymies menant à l'interprétation de formes verbales comme des noms. Les segments donnés dans le tableau sont des exemples de ces deux cas. Le segment *last month security forces* est un extrait de la phrase *Last month security forces raided the home of a BBC Persian employee's relative in Tehran*, et contient deux GN : le premier, *last month*, fonctionne comme un adjoint temporel dans la phrase, et le second, *security forces*, en est le sujet. Le système n'est pas capable en l'état de repérer les frontières de ces deux GN, et, puisqu'ils forment ensemble une structure de type [Adjectif + Nom + Nom + Nom], ils sont identifiés comme un seul GN comportant une erreur d'empilement. Le segment *hearing children learning English* fait partie du GN *studies involving hearing children learning English as a second language* ; dans ce segment, *learning* n'est pas un nom mais un verbe à la forme V-*ing* ayant pour object *English* […] et pour sujet *hearing children*. Puisque *learning* est également un nom, ce segment est reconnu comme un empilement de la forme [Adjectif + Nom + Nom + Nom], ou bien [Nom + Nom + Nom + Nom], si *hearing* est interprété comme un nom par le système. Ces deux problèmes peuvent également survenir en même temps, comme dans l'exemple *one day way back*. Ici, le segment reconnu comme un GN est en réalité constitué de deux groupes en fonction d'adjoints, l'un étant un GN (*one day*) et l'autre un GP (*way back*) ; par ailleurs, la préposition *back* est interprétée comme le nom *back*. La détection erronée est donc le résultat de deux problèmes conjoints.

Le problème de la délimitation des GN est également présent dans le traitement des erreurs "N_1 génitif", et est visible dans le segment *these conditions children*, extrait de la phrase *Upon these conditions children's happiness and fulfillment are seen to depend*. Cependant la plupart des faux positifs dus aux ressources grammaticales pour ce type d'erreur découlent d'homonymies entre les noms et les verbes, ou parfois d'autres formes. Ainsi, sur 20 faux

positifs liés aux ressources grammaticales, quatre concernent la délimitation des GN, alors que 16 sont dus à des homonymies. Tous concernent l'interprétation d'une forme verbale comme un nom (ex. : *bear*, *show*, *account*). Pour une majorité de cas (11 sur 16), l'erreur d'interprétation porte sur le N_2, comme dans l'exemple *a few of his friends find it cool* ; dans ces configurations, un nom au pluriel portant un déterminant défini est le sujet ou la fin du sujet d'un verbe au présent, qui est interprété comme un nom. En surface le segment se confond alors avec la structure [Déterminant Défini + Nom-*s* + Nom]. Dans quatre cas, l'erreur d'interpréation porte sur le N_1, comme dans l'exemple *a site that features properties*, configuration incluant une proposition relative dont le sujet du verbe est le pronom relatif *that*, dont le verbe est au présent et a un complément commençant par un nom. *That* est alors interprété comme un déterminant défini, et le verbe comme un nom, s'il s'y prête.

L'intégration de l'analyse syntaxique automatique dans notre utilisation de la plateforme <TextCoop> permettrait de limiter les faux positifs liés aux ressources grammaticales, l'activation des règles pouvant être subordonnée à l'identification préalable des GN, dans la limite des capacités de l'outil utilisé.

Cependant, tous les faux positifs ne sont pas imputables aux ressources grammaticales. Ainsi, les segments *a Saturday morning cartoon group shot* et *the games industry* ont été relevés comme des erreurs, conformément aux structures codées dans les patrons. On ne peut nier que ces segments soient en réalité corrects, mais il est troublant de voir que des segments tels que *?information system security strategies heterogeneity* et *?the objects properties* ne sont pas aussi naturels, alors qu'ils semblent présenter les mêmes caractéristiques. Nous nous confrontons ici de nouveau au cœur du problème : même si les règles de correction pouvaient être facilement mises en place d'un point de vue technique, la plus grande difficulté persiste, puisqu'il faut encore pouvoir décider de ce qui constitue une erreur. Nous pourrions interpréter le faux positif cité ci-dessus comme un signe que ces segments ne sont jamais erronés ou même inacceptables. Cependant, nous postulons plutôt qu'il existe d'autres facteurs à prendre en compte, et en particulier la possibilité d'erreurs conjointes, associant empilement et relations sémantiques inhabituelles.

Avant que des recherches supplémentaires puissent non seulement venir confirmer cette hypothèse, mais également apporter des solutions de correction fonctionnelles, il est fort probable que la seule solution équilibrée à adopter pour l'instant soit le signalement discret des possibles problèmes d'intelligibilité posés par le segment concerné. Un regard sur le contexte peut également permettre d'identifier des mentions de certaines parties du GN,

indiquant que le sens de la structure empilée est construit au fil du texte et peut ainsi être facilement reconstruit. Cependant, pour l'exemple donné ci-dessus, *a Saturday morning cartoon group shot*, il n'existe aucune mention préalable des parties du GN dans le reste du texte.

Dans le cas des structures N+N pouvant être des constructions avec un génitif déguisé, la précision de la détection pourrait être améliorée par la prise en compte des facteurs sémantiques que nous avons déjà évoqués dans la section 2.3.2b, comme l'identification des relations d'appartenance entre les deux éléments du GN. En effet, la différence flagrante entre *the objects properties* et *the games industry* est la relation sémantique qui est exprimée dans ces deux GN : dans le premier, le N_2 est un élément du N_1, correspondant à la relation "Partition" identifiée par Biber et al. (1999 : 590-591), alors que le second inclut une relation du type "Institution", dans laquelle le N_2 est une institution consacrée au N_1 (op. cit. : 590-591). Il semble donc que l'économie de la prise en compte des relations sémantiques à l'intérieur des structures N+N soit impossible, même pour les erreurs qui ne découlent pas uniquement d'un problème dans le choix des relations exprimées.

Évaluation des règles liées au placement des adverbes : adverbe *also*

Les règles de détection et de correction des erreurs de placement d'*also* ont été évaluées sur le corpus de productions en anglais L1 et sur le corpus modifié de productions en anglais L2. Pour l'évaluation sur le corpus d'anglais L1, une formule de recherche intégrée dans le fichier de patrons et de règles nous permet d'obtenir le nombre total des occurrences de l'adverbe *also* dans ce corpus. On relève ainsi 0,8 % de faux positifs pour les règles liées à *also*, soit un faux positif sur 129 occurrences de cet adverbe. Le tableau suivant présente cette occurrence :

Règle concernée	Code	Exemples	Occ.
Also	PAT	[309] *I hope also that this year will be…*	1

Tableau 53. Erreurs adverbe also *: analyse des faux positifs dans le corpus anglais L1*

La phrase concernée est classée dans la catégorie des détections erronées dues à des insuffisances dans les règles. Ce cas de faux positif est particulièrement intéressant : cette phrase correspond en effet à un placement de l'adverbe *also* que nous avons inclus dans les schémas d'erreur parce qu'il n'apparaissait que deux fois pour 500 occurrences de l'adverbe dans le corpus *BNC*, et n'avait aucune occurrence sur 500 dans le corpus *COCA* (voir 2.2.2b). La phrase dans laquelle on peut observer ce placement est la suivante : *I hope also that this*

jubilee year will be a time to give thanks. Elle est extraite d'un article du *Guardian* qui retranscrit un message écrit transmis au public par la Reine d'Angleterre Élisabeth II à l'occasion de son jubilé de diamant.

Nous avons indiqué précédemment que tous les retours du système de détection et correction sur des productions en anglais natif sont considérés comme des faux positifs ; toutefois, aussi controversé qu'il puisse paraître de reprendre la Reine sur son utilisation de l'anglais, nous pensons que cette configuration a tout de même sa place parmi les schémas de détection, en raison de son absence dans les corpus natifs étudiés. L'observation du contexte de cette utilisation particulière révèle par ailleurs la présence d'une phrase introduite par la formule *I hope that* dans le paragraphe précédent. Nous postulons que le placement d'*also* a été choisi afin de conserver l'intégrité de la formule *I hope*, créant ainsi une anaphore stylistique. Ainsi, si cette configuration ne peut être considérée comme une erreur grammaticale, il semble néanmoins important de signaler sa rareté à des personnes apprenantes, d'autant plus que la correction proposée est tout aussi adaptée (*I also hope that*…).

En raison de ces résultats satisfaisants en termes de précision, le système de détection et correction pour ces erreurs a été testé sur le corpus modifié d'anglais L2. Les résultats de cette évaluation sont présentés dans le tableau 54 en page suivante, qui utilise le système de classement expliqué plus haut.

Avant de donner notre interprétation des résultats, rappelons la définition des termes "précision" et "rappel". Nous prendrons l'exemple des retours des règles de détection, mais ces calculs peuvent être appliqués à la détection comme à la correction. Dans notre cadre, la précision mesure la "justesse" des retours du système, et est calculée en divisant le nombre d'erreurs correctement détectées par le nombre d'erreurs détectées en tout, c'est-à-dire incluant les fausses détections (aussi appelés faux positifs, terme que nous avons utilisé jusqu'à présent). Le rappel mesure la "couverture" ou l'"étendue" du système, et est calculé en divisant le nombre d'erreurs correctement détectées par le nombre d'erreurs présentes dans le document. La précision et le rappel peuvent être présentés sous forme d'un nombre entre 0 et 1, 1 correspondant à une précision ou un rappel parfait, mais pour plus de clarté nous présentons ces mesures sous forme de pourcentages, 100 % représentant la perfection de la précision ou du rappel.

Détection	Correction	Analyse	Occ.
Erreur détectée	Correction adéquate		**36**
	C. n'améliore pas l'original		**-**
	C. pire que l'original		**-**
Erreur non détectée		Insuffisance dans les règles	**22**
		Insuffisance dans la grammaire	**-**
		Insuffisance dans le lexique	**8**
		Problème dans le co-texte	**-**
		Raison inconnue	**-**
Faux positif		Insuffisance dans les règles	**-**
		Insuffisance dans la grammaire	**-**
		Insuffisance dans le lexique	**-**
		Problème dans le co-texte	**-**
		Raison inconnue	**-**
Nombre d'erreurs présentes dans le corpus			**66**
Nbre d'occ. d'*also*			**157**

Tableau 54. Placement de l'adverbe also *: résultats de l'évaluation sur un corpus d'anglais L2*

L'évaluation des règles de détection et de correction pour l'adverbe *also* sur un corpus d'anglais L2 modifié confirme les résultats de l'évaluation sur un corpus d'anglais L1, la précision du système atteignant 100 %. Ce résultat signifie que le système n'a produit aucune détection erronée sur ce corpus. Si l'on se réfère aux chiffres exacts donnés dans le tableau, cela signifie que les 36 retours de détection s'appliquent à des erreurs avérées. La précision de la correction est également de 100 %, puisque le système n'a proposé aucune correction qui soit moins bonne ou pire que le segment original. Ces taux élevés sont cependant très probablement imputables à la taille réduite du corpus de test, ainsi qu'au fait que le corpus soit modifié. Il y a fort à parier que la précision sera moins élevée lors de tests sur un corpus plus grand et non modifié.

Les règles ont permis la détection de 36 erreurs sur les 66 erreurs repérées manuellement dans le corpus en amont de l'évaluation. Ceci correspond à un rappel de 54,5 %. Parmi les 30 erreurs qui n'ont pas été repérées, huit ne l'ont pas été en raison d'insuffisances dans les ressources lexicales du système. Cela signifie simplement que les règles n'ont pas repéré l'erreur car un des mots du segment ne faisait pas partie des lexiques auxquels les règles font appel. Après ajout des mots concernés dans les lexiques, toutes les erreurs de cette catégorie

ont été détectées, amenant le taux de rappel à 66,7 %. Si les règles venaient à être intégrées à un système de correction global disposant de lexiques plus développés, ce problème serait beaucoup moins fréquent.

Les non-détections les plus problématiques sont celles qui sont causées par des insuffisances dans les règles, c'est-à-dire des manques dans les patrons de détection qui les empêche de détecter certaines erreurs. C'est le cas pour 22 erreurs dans ce corpus ; parmi celles-ci, nous identifions quatre ensembles de causes, regroupant pour la plupart plusieurs configurations. Nous verrons que certains segments ne sont pas jugés comme strictement agrammaticaux, et correspondent plutôt à des placements non standards, soit en tant que tels, soit dans l'utilisation précise qui est en faite dans le corpus.

Le premier ensemble regroupe les cas de non-détection liés à un placement d'*also* dans une position que nous n'avons pas prise en compte. Il concerne 10 des 22 erreurs non-détectées. En voici le détail, accompagné d'exemples extraits du corpus modifié :

- placement d'*also* en début d'énoncé (qui n'est pas une phrase ici – un seul cas) :

  [310] *?Also the modern man.*

- placement d'*also* en fin de proposition (trois cas) :

  [311] *?[It] defines what I mean also.*

- placement d'*also* entre deux parties d'un même groupe syntaxique placé après un verbe (six cas) :

  [312] **It is a system also which imposes its standards of beauty.*

  [313] *?People should be conscious also of their own value.*

Les deux premières configurations ne sont pas considérées comme strictement agrammaticales, ces placements étant théoriquement possibles. Leur utilisation dans ces exemples est cependant maladroite. La dernière configuration ne respecte pas les consignes qui avaient été données pour la modification du corpus, et qui demandaient de ne pas placer les adverbes à l'intérieur de groupes syntaxiques non verbaux. Cependant, dans les cas concernés, *also* intervient à un point de rupture relativement naturel des GN ou GAdj, c'est-à-dire entre le noyau et un dépendant post-noyau. Ce type de configuration n'a pas été observé dans les erreurs, mais il se peut qu'il se produise marginalement. Le deuxième exemple ne reçoit pas de jugement d'agrammaticalité car il est également possible que le focus d'*also* soit *of their own value*, dans une configuration stylistique particulière.

Le deuxième ensemble concerne six cas et regroupe les placements d'*also* entre un verbe et un type de dépendant dont la forme n'a pas été prise en compte dans les patrons. En voici le détail :

- le complément du verbe est une proposition introduite par un mot en *wh-* (deux cas) :

  [314] **We can wonder also what the results actually are.*

- le verbe est suivi d'un GAdv modifieur (trois cas) :

  [315] *?He cared also too much.*

- le complément est un GAdj attribut du sujet (un seul cas) :

  [316] **We get also lazy.*

Le premier cas n'a pas été intégré dans les schémas d'erreur car il n'a pas été rencontré dans le corpus. Dans le deuxième cas, *too much* est un GAdv modifieur exprimant le degré, et non un complément du verbe, mais le placement d'*also* entre le verbe et ce modifieur semble cependant inacceptable. N'ayant pas observé cette configuration dans les corpus d'anglais L1 ou dans le corpus de productions d'interlangue, cette possibilité n'a pas été prise en compte. Le troisième cas concerne la présence d'un GAdj en fonction de complément du verbe, configuration généralement limitée aux verbes d'état. Nous avons exclu BE des schémas afin d'éviter de signaler comme erronées des structures correctes du type *She is also very tall*, mais l'impossibilité de placer *also* entre un GAdj et un autre verbe d'état que BE n'a pas été prise en compte.

Le troisième ensemble concerne quatre cas et regroupe les placements d'also entre deux dépendants du verbe, qu'ils soient tous deux compléments, ou bien respectivement complément et adjoint. En voici deux exemples :

[317] *?To distract itself also from the stress […]*

[318] *?It prevents us also from communicating with each other.*

Ces segments ne semblent pas être strictement agrammaticaux mais présentent des configurations rares. Il est probable que le focus d'*also* soit le second dépendant (*from communicating with each other*), et l'adverbe pourrait donc être placé à proximité du focus pour des raisons de prosodie ou de clarté. Rappelons cependant que l'étude de Fjelkestam-Nilsson sur les placements de l'adverbe *also* par rapport à son focus indique que, quel que soit la fonction du focus dans la proposition, le placement "central" avant le verbe lexical est le plus fréquent (1983, voir section 2.2.2b).

Le dernier ensemble regroupe deux cas semblables, dans lesquels *also* est placé entre la base verbale de BE à la suite d'un auxiliaire et un complément

[319] *?Their neighbour might be also in the same situation.*

[320] *?We would be also naïve if we thought that […]*

Les schémas d'erreur pour *also* prennent en compte son placement entre un verbe et un GP dépendant ; cependant, BE étant exclu des verbes potentiels pour ces schémas, les configurations des exemples ci-dessus ne sont pas détectées. Nous n'avons pas envisagé que le fait que BE soit précédé d'un auxiliaire dans le groupe verbal puisse changer l'acceptabilité du placement d'*also* entre ce verbe et son complément, mais il semblerait que ce soit le cas. Des recherches plus approfondies sont nécessaires pour nous en assurer et découvrir les causes de cette différence.

Ayant présenté les insuffisances dans les règles donnant lieu à des échecs de détection, nous pouvons à présent identifier les pistes d'amélioration les plus fiables. Au-delà de leur identification, celles-ci devront dans tous les cas faire l'objet d'un retour aux grammaires et aux études de corpus afin de s'assurer de leur stabilité avant intégration dans les schémas.

Certaines configurations sont erronées dans le contexte du corpus modifié mais représentent en réalité des placements possibles, comme le placement d'*also* en début et en fin de proposition. Ces configurations n'ont donc pas leur place dans les schémas d'erreur. Le placement d'*also* entre deux dépendants après le verbe (exemples [316] et [317]) n'est pas courant mais n'est pas nécessairement erroné ; du point de vue de la détection, détecter ces erreurs de placement pose des difficultés importantes, puisque cela implique de délimiter les groupes syntaxiques qui entourent l'adverbe et d'en déterminer la fonction par rapport au verbe. Ceci n'est bien sûr pas infaisable avec l'intégration d'une analyse syntaxique automatique, mais le bénéfice à en retirer en termes de détection est maigre, pour des résultats qui ne sont pas nécessairement fiables.

La détection du placement erroné d'*also* à l'intérieur d'un groupe syntaxique situé après un verbe (exemples [311] et [312]) implique de reconnaître la présence d'un groupe syntaxique entier au lieu de deux groupes différents, comme dans le cas de la présence de deux dépendants du verbe, et d'y déceler automatiquement une rupture. Ceci peut être source de nombreuses complications et de faux positifs. Par ailleurs, il est possible que ces erreurs soient la conséquence d'un biais lié à la modification du corpus. Il faudrait vérifier la présence de ces erreurs dans les productions de personnes apprenantes.

Les configurations qui sont les meilleure candidates à l'intégration dans les schémas de détection sont la prise en compte des compléments prenant la forme d'un GAdj (exemple [315]) ou d'une proposition introduite par un mot en *wh-* (exemple [313]), le placement d'*also* entre le verbe et un GAdv modifieur du verbe (exemple [314]), et entre BE et son complément lorsque BE est dans sa forme infinitive (exemples [318] et [319]). La prise en compte de ces configurations implique des ajustements aux schémas de détection, mais ces ajustements peuvent être effectués rapidement, sans ajout de ressources lexicales ou sémantiques, ou d'outil d'analyse automatique supplémentaire. Si la prise en compte de ces configurations est validée par des recherches futures, elles permettraient de repérer six erreurs supplémentaires, dans le corpus utilisé, et donc de faire passer le taux de rappel de 66,7 % (après prise en compte des cas de non détection dûs à des insuffisances lexicales) à 78,8 %.

Comme nous l'avons déjà souligné, cette évaluation ne peut avoir qu'une valeur indicative, en raison du type de corpus utilisé et de sa taille limitée, et du fait que l'évaluation a été effectuée par une seule personne. Les résultats obtenus sont cependant encourageants et en accord avec l'exigence de précision élevée qui s'applique aux systèmes à destination de publics utilisateurs ou apprenants. Les retours de l'évaluation nous aident à ouvrir la voie vers l'augmentation de la couverture du système.

Évaluation des règles liées au placement des adverbes : adverbes de manière

Les règles de détection et de correction des erreurs de placement des adverbes de manière ont été évaluées sur le corpus de productions en anglais L1 et sur le corpus modifié de productions en anglais L2. Pour l'évaluation sur le corpus d'anglais L1, une formule de recherche intégrée dans le fichier de patrons et de règles nous permet d'obtenir le nombre total des occurrences d'adverbes de manière dans ce corpus. On relève ainsi 0,6 % de faux positifs pour les règles liées aux adverbes de manière, soit deux faux positifs sur 333 occurrences de ces adverbes. Le tableau suivant présente ces occurrences :

Règle concernée	Code	Exemples	Occ.
Adverbes de manière	GR	[321] *Please kindly check it.* [322] *They know very well that their policy will not be tolerable*	2

Tableau 55. Erreurs AdvM : analyse des faux positifs dans le corpus anglais L1

Les deux faux positifs sont dûs à des insuffisances dans les ressources grammaticales. Dans le premier cas, on observe deux confusions : l'adverbe *please* est interprété comme le verbe *please*, et le verbe *check* comme le nom *check*. La suite est ainsi confondue avec le

schéma d'erreur [Vlex + AdvM + GN]. Dans le second cas, la conjonction *that* est interprétée comme le pronom *that* ; cette confusion a pour conséquence l'analyse de la suite *know very well that* comme ce même schéma d'erreur.

En raison de ces résultats satisfaisants en termes de précision sur un corpus d'anglais natif, le système de détection et correction pour ces erreurs a été testé sur le corpus modifié d'anglais L2. Rappelons que le schéma d'erreur le plus central pour le placement des adverbes de manière donne deux propositions de correction (voir Tableau 36, schéma 5a). Dans de nombreux cas, les erreurs relevées reçoivent deux propositions de correction. Le tableau suivant présente les résultats de l'évaluation pour la première, ou unique correction selon les cas. La qualité des solutions de correction alternatives est évaluée dans un second temps.

Détection	**Correction**	**Analyse**	**Occ.**
Erreur détectée	Correction adéquate		**56**
	C. n'améliore pas l'original		**-**
	C. pire que l'original		**1**
Erreur non détectée		Insuffisance dans les règles	**2**
		Insuffisance dans la grammaire	**5**
		Insuffisance dans le lexique	**18**
		Problème dans le co-texte	**-**
		Raison inconnue	**-**
Faux positif		Insuffisance dans les règles	**-**
		Insuffisance dans la grammaire	**2**
		Insuffisance dans le lexique	**-**
		Problème dans le co-texte	**-**
		Raison inconnue	**-**
Nombre d'erreurs présentes dans le corpus			**82**
Nbre d'occ. d'adverbes de manière			**186**

Tableau 56. Placement des adverbes de manière : résultats de l'évaluation sur un corpus d'anglais L2

La précision de la détection est ici légèrement moins élevée que celle que nous avons pu observer pour le traitement des erreurs de placement d'*also*, puisque le système produit deux faux positifs, dûs à des insuffisances dans les ressources grammaticales. Le taux de précision pour la détection est donc de 96,6 %. Si l'on calcule la précision de la correction uniquement à partir des erreurs qui sont déjà correctement détectées, ce taux est de 98,2 %, puisque le système a proposé une correction inadaptée parmi les erreurs correctement détectées. Si ce

taux est calculé en prenant en compte les faux positifs pour la correction également, il est alors de 94,9 %.

Les faux positifs concernent les deux segments suivants :

[323] *He knew very well that his troops would win.*

[324] *We will have to seriously face problems.*

Le premier cas est semblable à un des cas de détection erronée repérés dans le corpus d'anglais L1, dans lequel la conjonction *that* est interprétée comme le pronom *that*. Le second cas de faux positif est également lié à des problèmes d'homonymie, le verbe *face* étant interprété comme le nom *face*, déclenchant le repérage du schéma [Vlex + AdvM + GN].

La confusion liée à *that* ayant causé le premier de ces faux positifs a été reproduite sur cinq autres segments, mais ceux-ci ne sont pas comptabilisés comme des faux positifs car l'emploi de l'adverbe est moins standard que la correction proposée. En voici deux exemples :

[325] *?[…] if I say boldly that equality is something very important.*

[326] *?We can understand easily that an armed intervention in Kuwait […]*

Il apparaît cependant que placement des adverbes dans ces phrases est loin d'être problématique, et pourrait être produit dans certaines conditions dans des textes en anglais L1. Ainsi, même si les corrections proposées ne sont pas erronées dans la plupart des cas de confusion liés à *that*, le système devrait être affiné afin d'éviter cette situation, que nous considérons comme une faille plutôt qu'un hasard heureux.

La précision de la correction est affaiblie par le fait que le système produit dans un cas une correction qui est plus erronée que le segment original :

[327] *?Everybody knows perfectly that all men are equal.*

[328] **Everybody perfectly knows that all men are equal.* (correction proposée)

Dans ce cas, le segment original paraît inacceptable, mais il semble que cela soit plutôt dû au choix d'adverbe ; le GAdv *perfectly well* semble plus adapté. Étant donné la nature plus lexicale que syntaxique de l'erreur, elle ne peut pas être correctement traitée par nos règles.

Pour ce qui est de la couverture du système, les règles ont permis la détection de 57 erreurs sur les 82 erreurs repérées manuellement dans le corpus en amont de l'évaluation. Ceci correspond à un rappel de 69,5 %. La majorité des cas de non-détection est due à des insuffisances dans le lexique, avec 18 erreurs concernées pour 25 non détectées au total.

Après ajout des mots concernés dans les lexiques, toutes les erreurs de cette catégorie ont été détectées, amenant le taux de rappel à 91,5 %. Le taux de précision se trouve également modifié à la suite de ces ajouts, puisque le nombre d'erreurs correctement détectées augmentent alors que le nombre de faux positifs reste stable. Le taux de précision atteindrait ainsi 97,4 %. Devant ces taux très élevés, nous nous devons de rappeler les limites de cette évaluation, qui sont la taille et le type de corpus utilisé, ainsi que l'absence d'une évaluation extérieure neutre.

Les cas de non-détection sont attribuables à des insuffisances dans les ressources grammaticales ainsi que dans les règles. Concernant les ressources grammaticales, les insuffisances dans la description du GN sont à la source des cinq cas dans cette catégorie. Rappelons que la délimitation des GN est effectuée à partir d'une grammaire simple intégrée aux ressources grammaticales de <TextCoop>. Elle nous permet surtout d'utiliser la récursivité des règles dans <TextCoop> pour éviter d'avoir à décrire les GN en entier dans les règles de détection et correction. Cette grammaire est élémentaire, et n'inclut pas toutes les configurations possibles pour les GN. Les cas de non-détection sont plus particulièrement liés à la variation des configurations des dépendants pré-noyau : si le début du GN n'est pas reconnaissable par le système, celui-ci abandonnera la reconnaissance avant d'arriver au nom noyau. Les variations des dépendants post-noyau ne posent pas de problème pour la reconnaissance de la présence d'un GN, mais sont par contre très problématiques pour sa délimitation, comme nous allons le voir. Les cas de non-détection liés à la grammaire du GN concernent donc des segments dans lesquels un adjectif placé avant le nom a une forme qui n'est pas intégrée à la grammaire (forme comparative, adjectif participial), ou bien est coordonné avec un autre adjectif. Voici des exemples de ces cas (l'adjectif ou la coordination d'adjectifs est soulignée):

[329] *Dreaming and imagination occupy unconsciously a <u>larger</u> place […] than we realize.*

[330] *He experiences unexpectedly a <u>physical and spiritual</u> revival.*

Les deux cas de non-détection causés par des insuffisances dans les règles concernent le placement d'un adverbe entre un verbe et un complément prenant la forme d'une proposition introduite par un mot en *wh-*, situation que nous avons déjà constatée dans le traitement des erreurs de placement d'*also*. Le segment suivant illustre ce cas :

[331] *The French phrase […] <u>defines perfectly what I mean</u>.*

Pour une grande majorité des détections (50 sur un total de 57), deux solutions de correction sont proposées. Cette situation se produit lorsque le schéma 5a est activé, soit le placement d'un adverbe seul entre un verbe et un complément nominal ou prépositionnel. La première correction consiste à placer l'adverbe avant le verbe, alors que la seconde place l'adverbe après le complément. Ceci montre également que ce schéma est mobilisé beaucoup plus souvent que les autres. Vingt-trois de ces solutions alternatives, soit presque la moitié d'entre elles, proposent des segments moins grammaticaux ou acceptables que le segment original. Dans 20 cas, les corrections erronées sont liées à des insuffisances dans les ressources grammaticales, qui empêchent la reconnaissance de groupes entiers tels que les GN (pour 15 cas) et les propositions en *that* (pour cinq cas). Voici des exemples des corrections qui ont été proposées (l'adverbe est souligné, et des crochets vides indiquent la position de l'adverbe dans le segment de départ) :

[332] *The increasing number of factories *has spoiled [] the landscape's fatally attraction.*

[333] *Television has replaced [] the traditional dream rapidly which was offered by literature.*

[334] *We can understand [] that easily an armed intervention in Kuwait […]*

Dans trois autres cas, la correction place l'adverbe entre deux dépendants à la suite du verbe. Ce placement n'est pas agrammatical pour les adverbes de manière, mais selon la nature des dépendants, il peut provoquer des modifications de sens, qui doivent être évitées dans les systèmes de correction automatisée. Les deux phrases suivantes sont un exemple de ce phénomène ; la première est le segment original, la seconde est la première correction proposée, et la troisième donne la seconde solution de correction :

[335] **It is easy to sit in our armchairs watching quietly other people working.*

[336] *It is easy to sit in our armchairs quietly watching other people working.*

[337] *It is easy to sit in our armchairs watching other people quietly working.*

Cependant, nous avons également observé des cas dans lesquels le placement entre deux dependants est parfaitement acceptable, voire meilleure que la première solution de correction, tant grammaticalement que pour le sens de la phrase :

[338] **Teachers can use skillfully television to illustrate their lessons.*

[339] *Teachers can skillfully use television to illustrate their lessons.*

[340] *Teachers can use television skillfully to illustrate their lessons.*

Dans la plupart des cas de placement de l'adverbe au milieu d'un GN, le fait qu'une deuxième solution de correction soit proposée découle déjà d'une insuffisance dans les ressources grammaticales : si le GN était correctement repéré comme étant long, le schéma mobilisé ne proposerait pas de seconde correction plaçant l'adverbe après le complément.

L'évaluation des règles nous mène à identifier des pistes d'amélioration du système. La première amélioration à apporter consiste à ajouter des spécifications à la reconnaissance de *that*. Avant d'introduire des analyses syntaxiques automatiques, il est possible de remédier à certaines erreurs de détection en limitant les types de mot pouvant être rencontrés après *that*, réduisant par la même occasion les chances que ce terme soit interprété comme un pronom au lieu d'une conjonction.

Une seconde amélioration relativement simple concerne l'expansion de la grammaire du GN créée manuellement, afin d'y introduire la reconnaissance de dépendants pré-noyau de formes plus variées. Ceci permettrait d'une part de répérer certaines des erreurs n'ayant pas été détectées avec le système existant. On remarque également que la délimitation des GN pose problème, surtout lorsqu'une seconde correction est proposée. Il ne paraît pas intéressant de supprimer cette option, même en raison des fausses corrections qu'elle déclenche, car elle est aussi génératrice de corrections appropriées. Avant de recourir à un outil extérieur au système et permettant la délimitation des GN, il faut explorer la possibilité d'augmenter la richesse des dépendants post-noyau également, afin d'améliorer la précision de l'idenfication des GN longs, ce qui permettrait de limiter le placement intempestif d'adverbes de manière au milieu de GN. Le problème du placement de l'adverbe entre deux dépendants et modifiant le sens du segment nécessite par contre une analyse des fonctions des éléments dans la phrase, qui pourrait être fournie par un outil d'analyse automatique supplémentaire et extérieur au système de correction.

Pour finir, toutes les règles de correction consacrées aux adverbes pourrait être rendues plus précises par l'inclusion d'informations lexicales concernant les verbes pouvant admettre des adverbes compléments, et les adverbes pouvant fonctionner comme compléments. En effet, Osborne indique que le placement des adverbes en SVAO est facilité par l'existence d'une collocation entre le verbe et l'adverbe, mais nous avons évoqué la possibilité que ce lien soit en réalité un lien syntaxique de complémentation (ex. *to treat fairly*, *to handle carefully*). Des informations supplémentaires sur cette possibilité permettraient d'éviter certains faux positifs.

c. Limites et points forts des règles de détection et correction

Des pistes d'amélioration des règles ont déjà été données dans les parties consacrées à l'évaluation des règles pour chaque type d'erreur. Cette section revient sur des limites et points forts généraux du système.

L'évaluation a montré qu'une des limites du système est le fait qu'il n'inclut pas d'analyse syntaxique automatique. Si certaines difficultés peuvent être surmontées en augmentant la précision des règles et la couverture de la grammaire du GN, l'analyse syntaxique automatique pourrait permettre de résoudre partiellement d'autres problèmes. En particulier, une des difficultés les plus importantes du traitement de textes en anglais est la gestion de l'homonymie, qui est extrêmement répandue dans cette langue, étant le résultat de certains modes de dérivation lexicale et de la rareté des marques flexionnelles permettant de distinguer les catégories. Le recours à un analyseur syntaxique pourrait permettre de distinguer les homonymes sur la base de leur fonction dans la proposition, et ainsi rendre plus précise la détection des structures erronées. L'inclusion d'un tel outil dans <TextCoop> est envisageable, mais les limites de ces outils doivent également être prises en compte, puisqu'elle peuvent alourdir le fonctionnement du système et ne sont pas infaillibles dans leurs analyses.

Le fait d'utiliser des patrons et règles de correction pose une contrainte sur le texte de départ, qui doit être globalement bien construit afin que les erreurs soient repérées. Cette contrainte existe cependant de manière plus ou moins forte pour tous les systèmes de correction automatisée, des productions arbitrairement agrammaticales ne pouvant être traitées par de tels systèmes (Leacock et al., 2010 : 7). De plus, le fait d'utiliser des parties du discours à des formes neutres et non des formes exactes pour les règles permet d'introduire une certaine marge dans ce qui peut être détecté. L'exactitude de la conjugaison et des accords dans le groupe verbal n'est ainsi pas nécessaire pour qu'une règle fonctionne. Par ailleurs, notre public cible est composé de personnes ayant un niveau intermédiaire à avancé, et qui produisent donc généralement des phrases grammaticales dans leur ensemble.

Nous avons vu qu'un nombre important de cas de non-détection sont dus à des insuffisances dans les ressources lexicales. En effet, les lexiques utilisés sont loin d'être exhaustifs et la non-reconnaissance de certains mots bloquent les corrections. S'il est nécessaire de faire l'inventaire des ressources lexicales à inclure dans le système ainsi que de développer des ressources nous permettant de résoudre certains problèmes (cf. ressource lexico-sémantique pour les adverbes), l'inclusion de lexiques exhaustifs est une étape qui

appartient à la phase de développement industriel du système, qui n'est évidemment pas ce dont il question dans cette recherche.

L'utilisation de patrons pose également le problème de leur création, puisque celle-ci doit être effectuée à la main, chaque cas exceptionnel nécessitant une règle propre ou bien des aménagements particuliers, ce qui peut amener à devoir créer un nombre de règles très important. Cependant, nous avons vu que l'application d'une méthode progressive d'exploration linguistique des segments erronés et du phénomène concerné permet de balayer la plupart des configurations fréquentes. La phase d'évaluation est également utilisée pour compléter l'ensemble des règles. La méthode de création des règles permet par ailleurs de pouvoir les affiner autant que nécessaire, grâce aux fonctions d'optionalité ou d'exclusion que nous avons utilisées.

Notons aussi comme point fort de l'utilisation de règles de correction dans <TextCoop> la facilité d'implémentation technique après modélisation des phénomènes : en raison du fonctionnement logique de Prolog, il est très rapide de passer d'une description linguistique à une implémentation technique. De plus, la création de nouvelles règles ou l'adaptation de règles existantes peut être effectuée par des personnes n'ayant pas de compétences et connaissances préalables en programmation, et recevant une formation rapide au fonctionnement du système et de Dislog. Le module de détection et correction des erreurs est également applicable à d'autres L1 pertinentes, telles que les autres langues romanes.

Concernant les résultats de l'évaluation, nous avons déjà mentionné les limites de la méthode et du type de corpus utilisés pour évaluer les règles, et qui sont probablement partiellement responsables des taux élevés de précision et de rappel obtenus pour le traitement des erreurs de placement des adverbes. Ces taux élevés sont cependant encourageants pour la suite éventuelle du développement du système. Nous nous sommes efforcés de suivre les recommandations concernant la création de règles de correction automatisée à l'usage de publics apprenants en privilégiant la précision de la détection et de la correction au détriment de la couverture du système. Ceci est particulièrement visible dans le cas de la correction du placement d'*also*, pour laquelle le taux de rappel est nettement inférieur à la précision. Cette approche doit être préservée lors de futures recherches pour l'amélioration des règles.

Enfin, le dernier point fort de l'utilisation des règles de correction au niveau pratique concerne la précision des messages correctifs qui peuvent venir compléter ces règles, celles-ci ciblant des constructions déterminées. De plus, les recherches et synthèses linguistiques ayant mené à leur création permettent d'avoir une base solide à partir de laquelle des explications

peuvent être fournies à la personne utilisant le système de correction. La partie suivante est ainsi consacrée à l'élaboration d'un canevas pour la création de messages explicatifs à associer aux règles de correction.

3.3 Accompagner la correction automatisée

Le terme de "rétroaction" (*feedback*) est utilisé dans différents domaines pour faire référence au retour d'un effet sur sa propre cause, que ce soit au sein d'un système intégrant un module d'autocontrôle, ou bien dans le domaine des interactions humaines (Heift et Schulze, 2007 : 115). La "rétroaction corrective" (*corrective feedback*) est ainsi un retour d'information sur une action ou production inadéquate, et ayant pour but de susciter une action ou production adéquate en réaction à ce retour. En acquisition des langues, la rétroaction corrective a pour objectif de permettre aux apprenants et apprenantes de modifier leur interlangue (Ellis, 2008 : 958), c'est-à-dire de procéder à la correction de leurs productions, et dans l'idéal à la modification des savoirs linguistiques appris précédemment, et qui s'avèrent être erronés.

Au sens strict du terme, le simple fait de donner une indication sur la présence d'une erreur constitue une rétroaction. La correction grammaticale automatisée repose donc en elle-même sur la création de retours correctifs, qu'il s'agisse simplement de détecter une erreur, de proposer une correction, ou encore que la rétroaction soit étoffée d'explications grammaticales et d'exemples. L'inclusion de modules d'accompagnement de la correction dans les correcteurs automatiques est très courante. Celle-ci peut prendre la forme de messages donnant des indications sur l'erreur, ou bien renvoyer à des ressources grammaticales et lexicales présentes dans le système. Puisque notre système fournit déjà une détection des erreurs, ainsi que des propositions de correction dans la plupart des cas, nous nous intéressons ici aux messages correctifs accompagnant la proposition d'une correction, ou la détection seule. L'objectif de l'inclusion de ces messages dans notre système est de permettre aux personnes l'utilisant de remédier aux erreurs futures ou même de les prévenir, et ainsi de gagner en autonomie.

Toutefois, avant de nous pencher sur la génération de messages correctifs pour accompagner les corrections proposées par notre système, il sera nécessaire de poser les deux questions suivantes :

- L'inclusion de messages accompagnant la détection/correction est-elle utile ?

- Si c'est le cas, quelle forme ces messages doivent-ils adopter pour maximiser leur pertinence ?

La première sous-partie est consacrée à la recherche de réponses à ces questions, à partir des résultats de travaux sur ce thème dans le domaine général de l'acquisition des langues secondes, puis plus précisément en enseignement des langues assisté par ordinateur et dans le domaine de la correction automatisée. Dans un second temps, nous présentons notre application de ces résultats à la création de messages correctifs et à leur adaptation à des profils d'utilisatrices et utilisateurs différents, sous la forme d'un canevas modulable en cinq étapes. La question de la mise en œuvre pratique de cet accompagnement des corrections dans notre utilisation de <TextCoop> est également abordée.

3.3.1 État des recherches en rétroaction corrective

a. La rétroaction corrective en acquisition des langues secondes

Corriger les erreurs des personnes apprenantes semble à première vue être une étape banale, mais également essentielle, de l'enseignement et de l'acquisition des langues secondes. Les recherches sur la rétroaction corrective révèlent cependant une absence relative de consensus concernant l'utilité de cette pratique en général et de ses différents types en particulier. Les types de rétroaction corrective s'organisent selon au moins quatre axes :

- écrit / oral,
- implicite / explicite : la rétroaction corrective est jugée implicite lorsque la forme correcte est fournie sans mentionner la présence d'une erreur, et explicite lorsque l'attention est attirée sur la présence d'une erreur,
- direct / indirect : dans une rétroaction directe, la correction est donnée, alors qu'elle n'est pas présente dans une rétroaction indirecte, dans laquelle l'erreur peut par exemple être soulignée,
- ciblé / non ciblé : une rétroaction corrective ciblée ne concerne qu'un type d'erreur, mais une rétroaction corrective non ciblée prend en compte tous les types d'erreur.

Tous ces axes ne sont pas applicables à notre étude : la distinction entre rétroaction ciblée et non-ciblée n'est pas pertinente puisque l'étude ne concerne que deux types d'erreur. Par ailleurs, les messages utilisent le médium de l'écrit, puisqu'il s'agit de la correction automatique de textes ; on peut imaginer des systèmes fournissant un retour oral, mais la production de messages écrits reste la stratégie la plus simple et la plus courante. De plus, s'il

n'est pas impossible de fournir une rétroaction implicite dans un correcteur grammatical, c'est-à-dire fournir des corrections sans indiquer la présence d'erreurs, l'objectif de ce type de système est néanmoins de faire remarquer les erreurs, rendant la rétroaction nécessairement explicite. Nous nous focalisons donc particulièrement sur l'utilisation de la rétroaction corrective par écrit et explicite, et portant plus particulièrement sur les formes grammaticales. Nous verrons qu'il est possible d'adapter le système pour qu'il fournisse des messages de rétroaction directe comme indirecte, c'est-à-dire incluant la correction de l'erreur, ou bien uniquement des informations supplémentaires permettant à la personne utilisatrice de corriger l'erreur elle-même.

Malgré la place centrale qu'occupe la problématique de la rétroaction corrective dans le domaine de l'acquisition des langues secondes, on y déplore encore à ce jour le manque de résultats définitifs concernant son utilité en général, et l'intérêt des différents types de rétroaction corrective en particulier (Russell et Spada, 2006 : 156 ; Li, 2010 : 348). Nous avons mentionné l'absence relative de consensus au sujet de la pertinence de la pratique de la rétroaction corrective. Celle-ci est due en partie à la difficulté d'obtenir des résultats de recherches fiables, qui s'explique notamment par la diversité des facteurs et des variables à prendre en compte, comme le format des rétroactions, le type d'erreur concerné et les caractéristiques internes aux personnes apprenantes (Heift et Schulze, 2007 : 154). Elle est également liée aux théories concurrentes de l'acquisition des langues secondes, et notamment à leur position sur l'importance des connaissances implicites et explicites pour l'acquisition et la possibilité de l'existence d'une interface entre ces modes de connaissance : la rétroaction corrective tendant à agir sur les connaissances explicites, les théories qui remettent en question le rôle de ces connaissances sur l'acquisition jettent donc le doute sur l'utilité de cette pratique (Cornillie et al., 2012 : 260). Parallèlement, son utilisation est mise en avant dans le cadre de la "Noticing Hypothesis" (Schmidt, 1990), qui suggère que l'acquisition est facilitée lorsque la personne apprenante est amenée à remarquer les formes.

Dans un article devenu célèbre, Truscott (1996) a remis en question la pertinence de l'utilisation de la rétroaction corrective sous toutes ses formes, appelant à l'abandon de son utilisation. Cet article a suscité de nombreuses réactions en défense de la rétroaction corrective (ex. : Ferris, 1999), menant d'ailleurs à une augmentation du nombre et de la qualité des recherches menées sur cette question.

Malgré le manque de données comparables disponibles sur l'effet de l'utilisation de la rétroaction corrective sur l'acquisition d'une L2, la communauté de recherche dans ce domaine

tend à considérer cette pratique comme globalement bénéfique pour l'acquisition (Cornillie et al., 2012 : 260). Russell et Spada (2006) ont évalué l'effet de la rétroaction corrective sur l'acquisition de la grammaire en L2 à travers une méta-analyse. Celle-ci conclut à l'existence d'un effet positif de cette pratique sur l'acquisition de formes grammaticales. Les auteurs indiquent que les effets semblent également être relativement durables (op. cit. : 152). Concernant l'efficacité de certains types de rétroaction par rapport à d'autres (ex. : implicite vs. explicite), Russell et Spada concluent qu'il n'existe pas encore assez de preuves de la supériorité d'un type en particulier, mais qu'il semblerait que les élèves tirent plus de bénéfices de rétroactions les poussant à retravailler leurs productions sans leur fournir de correction directe (op. cit. : 154). Dans le cadre de notre recherche, il est également important de noter que la rétroaction corrective est vue comme un moyen d'éviter la fossilisation des formes erronées chez les personnes apprenantes adultes (Ferris, 2004 : 56).

b. La rétroaction corrective en ELAO

D'après certaines études, la rétroaction corrective est plus efficace lorsqu'elle est délivrée immédiatement (Kulik et Kulik, 1988 ; cité par Heift et Schulze, 2007 : 121), rendant son intégration à des systèmes d'enseignement des langues assisté par ordinateur particulièrement pertinente, puisque ceux-ci sont adaptés à la génération de messages personnalisés (Nagata et Swisher, 1995 : 339). Malgré ce potentiel, Heift et Schulze déplorent la pauvreté des messages correctifs traditionnellement utilisés dans les systèmes d'ELAO (op. cit. : 146). Les recherches consacrées à cette pratique dans l'enseignement assisté par ordinateur sont encore moins nombreuses qu'en acquisition des langues secondes, et peu d'entre elles concernent spécifiquement la correction grammaticale automatisée. Il est cependant possible de dégager quelques pistes intéressantes pour la création de messages correctifs dans ce cadre.

S'appuyant sur une étude effectuée par Lyster et Ranta (1997, cité par Heift, 2004 : 418) visant à catégoriser les types de rétroaction utilisés par le personnel enseignant lors d'un enseignement en présentiel, Heift (2004) définit les types de retours pertinents dans le cadre de l'ELAO. Les types identifiés par Lyster et Ranta sont les suivants :

- correction explicite de l'erreur,
- reformulation du segment erroné,
- demande de clarification,
- rétroaction métalinguistique, c'est-à-dire donnant des informations sur la langue (ex. : règle de grammaire)

- élicitation d'une forme correcte, par exemple en utilisant une ellipse (ex. : *"To listen [] music?"*)
- répétition de la forme erronée avec interrogation.

Les messages correctifs ne pouvant bien entendu pas prendre le même format dans des classes et dans des programmes d'ELAO, Heift restreint le champ des possibilités à quatre types :

- correction donnée directement (correspondant à la correction explicite et à la reformulation),
- encouragement à proposer une autre réponse (correspondant à une demande de clarification),
- explication du type d'erreur, ou rétroaction métalinguistique,
- marquage de l'erreur (*highlighting*, correspondant à l'élicitation des formes correctes et à la répétition).

Heift a ensuite comparé les réactions d'autocorrection d'un public d'apprenant par rapport aux trois derniers types, et à certaines combinaisons de ces types, dans le cadre d'exercices de grammaire et de vocabulaire effectués dans un tuteur de langue intelligent pour l'acquisition de l'allemand. Les résultats obtenus indiquent que les types de rétroaction les plus susceptibles de mener à des réactions d'autocorrection sont ceux qui sont les plus explicites et donnent des informations précises sur les erreurs, c'est-à-dire les messages de rétroaction métalinguistique, en combinaison avec le surlignage des erreurs (op. cit. : 428-429). Cette étude sur la réaction des personnes apprenantes par rapport au type de rétroaction en ELAO vient compléter d'autres études focalisées sur l'efficacité des différentes règles de rétroaction en termes d'acquisition. Les résultats des travaux de Nagata à ce sujet indiquent que les messages donnant des informations précises sur les erreurs sont plus bénéfiques pour l'acquisition qu'une simple indication de la présence d'une erreur (1993 : 337).

Les résultats de ces recherches, ainsi que des consignes données précédemment au sujet de la correction automatisée, nous amènent à proposer une liste d'exigences pour la création de messages correctifs dans un système de correction automatisée destiné à des personnes utilisatrices de l'anglais. Nous identifions quatre exigences :

- les messages correctifs doivent consister en un marquage du segment erroné, associé à une rétroaction métalinguistique, c'est-à-dire donnant des informations sur l'erreur et ses causes,

- les messages présentés doivent cependant être concis et utiliser un lexique simple (cf. van der Linden, 1993)
- comme la détection, le contenu du message doit être exact,
- les messages correctifs doivent être adaptés, et adaptables, aux différents profils d'utilisation.

La sous-partie suivante présente la mise en œuvre de ces consignes appliquées à la création de messages correctifs pour les erreurs de placement des adverbes dans nos travaux.

3.3.2 Production de messages correctifs : proposition d'un canevas en cinq étapes

Comme il a déjà été mentionné, la plupart des correcteurs grammaticaux automatiques incluent désormais des messages d'accompagnement des propositions de correction. C'est par exemple le cas de tous les correcteurs utilisés pour les tests que nous avons présentés dans la sous-partie 3.1.2. Avant de présenter le canevas envisagé pour la génération des messages correctifs dans nos travaux, nous avons passé en revue les messages proposés par ces systèmes. Le tableau 57 en page suivante présente les messages correctifs pour l'erreur qui a été détectée par tous ces systèmes (**I can lost a love*). Les messages sont donnés avec leur format exact. La colonne de droite indique les types de rétroaction utilisés ; nous utilisons les types identifiés par Heift (2001 ; correction, explication, marquage de l'erreur), ainsi que des types supplémentaires (instruction de remédiation, illustration). La rétroaction "encouragement", qui incite la personne utilisant le système à proposer une autre réponse, n'est pas présente dans ces systèmes, puisqu'il s'agit avant tout de correcteurs utilisés sur des productions spontanées, et non de tuteurs de langue incluant des exercices. Ce type de rétroaction pourrait cependant faire partie d'un correcteur élaboré spécifiquement pour l'apprentissage.

Le marquage dans le texte de départ, effectué par le biais du soulignement ou de la mise en couleur du texte, ne peut pas être reproduit ici, mais il est bien présent pour tous les systèmes.

Système	Message	Type
SpellCheckPlus	You should probably use the bare form of the verb here, e.g.: *It can **change** your life*. Another possibility would be to add *be* or *have*, e.g.: *The soup should **be** done, You should never **have** given away the secret*.	Instruction de reméd. Illustration
Grammarly	Perfect tense verb used with modal verb not in proper form.	Explication
Ginger	I love books where I can lose myself.	Illustration
Correct English	Consider [lose] instead of 'lost'. Use the base form of a verb for a command, if it follows 'to', or if it is used with a modal verb such as 'could', 'can', 'should', 'will', or 'might'.	Correction Instruction de reméd. Explication
Proofwriter	Verbs show what a subject is doing or express the relationship between a subject and an object. Proofread this sentence to make sure you have used the correct form of the highlighted verb.	Explication Instruction de reméd.
Language Tool	The verb 'can' requires base form of the verb: 'lose' Correction: lose	Explication Correction
Word 2007	lose After certain auxiliary verbs such as "can" or "may," use the base form of the verb. Instead of: Is it true that the broker cannot rewriting the contract? Consider: Is it true that the broker cannot rewrite the contract?	Correction Instruction de reméd. Illustration

Tableau 57. Messages de rétroaction dans les correcteurs grammaticaux

Les schémas à partir desquels les messages correctifs sont construits sont extrêmement variés, tout comme les différentes formulations utilisées. Pour la remédiation, ces formulations peuvent consister en des instructions à l'impératif plus ou moins directes suivant le verbe utilisé (ex. : comparer *Consider using…* et *Proofread this sentence*), ou bien en des conseils utilisant des auxiliaires modaux (ex. : *should*). En dehors du marquage de l'erreur, qui intervient systématiquement en amont du message, il est difficile de dégager un ordre standard d'organisation de celui-ci. On retrouve cependant les étapes suivantes, dont nous indiquons les définitions :

- marquage de l'erreur : l'erreur est signalée dans le texte d'origine par des caractères gras, du souligné, une police et/ou un fond de couleur différente,
- correction : le système propose une version corrigée du segment d'origine,
- explication : le système présente des informations permettant d'expliquer l'erreur et/ou la règle de grammaire concernée (voir ci-dessous),
- instruction de remédiation : le système donne des indications permettant de corriger le segment,

- illustration : le message inclut un exemple d'utilisation correcte de la même structure.

Certaines de ces étapes, comme le marquage et l'illustration, sont simples à élaborer et à mettre en place une fois que l'erreur a été correctement identifiée et corrigée. D'autres étapes sont plus problématiques en elles-mêmes. C'est le cas de la rétroaction métalinguistique, dont les contours précis peuvent être difficiles à définir, et qui se confond parfois avec les instructions de remédiation et les diagnostics d'erreur. Nous avons donc mené une étude plus approfondie concernant l'utilisation de la rétroaction métalinguistique et de l'explication dans les messages correctifs (Garnier, 2011).

Parmi les systèmes évalués dans la sous-partie 3.1.2, le correcteur grammatical de *Word* 2007 semble présenter les messages correctifs les plus complets. Par ailleurs, ce système intègre une progression qui permet de recevoir des messages correctifs sur demande, répondant ainsi aux exigences de concision et d'adaptation aux personnes utilisatrices. L'étude est donc fondée sur l'observation de 36 messages différents obtenus sur un corpus d'interlangue de 7 200 mots. Nous n'avons pas pris en compte la pertinence des corrections, puisque notre objectif était de "récolter" des messages afin d'en analyser le contenu. Malgré une grande diversité dans la gestion de la rétroaction métalinguistique dans les messages correctifs fournis par ce correcteur, ceux-ci incluent systématiquement les étapes suivantes : marquage de l'erreur, titre du message (ex. : *Number agreement*, *Connecting words*), rétroaction métalinguistique, illustration. Partant du principe que l'explication fait partie des procédés rhétoriques, nous avons analysé et annoté cette étape dans ces messages en utilisant les étiquettes de la *Rhetorical Structure Theory* (Mann et Thompson, 1988). Nous avons identifié trois schémas, dont les différences résident dans la nature du noyau rhétorique, qui peut prendre la forme d'une instruction de remédiation, d'une règle de grammaire ou d'un diagnostic d'erreur. Ces noyaux peuvent être accompagnés d'autres éléments, comme des concessions, des élaborations, des indications concernant l'objectif de la remédiation. Voici un exemple de chacun de ces trois cas :

- le noyau prend la forme d'une instruction de remédiation :

> [**concession** Although "a lot of" or "lots of" may be used informally], **substitute "many" or "much"** [**objectif** for a more formal or traditional tone.]

- le noyau prend la forme d'une règle de grammaire :

> **A noun and the words that modify that noun must agree in number.** [**élaboration** ["Many" and "few" modify plural nouns. "Much" and "less" modify nouns that cannot be counted or divided such as "much oil," "less happiness". In addition, the phrase "one of" must modify a plural noun.]

- le noyau prend la forme d'un diagnostic d'erreur :

Only one of the marked words is necessary to signal that a noun follows.

Cette étude nous amène à identifier deux questions supplémentaires ayant trait à l'élaboration de l'explication dans un correcteur grammatical à visée pédagogique :

- À quel type de questionnement l'explication doit-elle répondre ?
- Qu'est-ce qui constitue une explication dans une rétroaction corrective ?

Nos propositions de réponses à ces questions viennent compléter les consignes définies plus haut pour la création de messages correctifs efficaces. Il nous semble que l'explication devrait au moins fournir des informations sur les raisons qui ont mené à l'identification d'une erreur. Cette étape consiste à poser un diagnostic d'erreur, qui prend la forme d'une déclaration concernant la forme du segment erroné (ex. : "Dans cette phrase, l'adverbe *also* est placé entre le verbe *have* et son complément *good news*"). Avec le diagnostic d'erreur, l'étape de rétroaction métalinguistique fait partie de l'explication ; elle consiste en un énoncé synthétique de la règle de grammaire ou d'usage concernée (ex. : "En anglais, il est rare de placer les adverbes entre les verbes et leurs compléments") et est à distinguer de la remédiation, qui prend la forme d'une instruction ou d'un conseil (ex. : "Vous pouvez corriger cette phrase de la façon suivante…").

Synthétisant ces différentes conclusions, nous avons abouti au canevas suivant, qui comprend cinq étapes distinctes :

1. Marquage de l'erreur dans le texte,
2. Diagnostic d'erreur reprenant les termes du texte d'origine
3. Rétroaction métalinguistique (ou informations grammaticales),
4. Instruction de remédiation,
5. Illustration à l'aide d'un exemple correct.

Les messages sont fournis dans la L1 des personnes utilisatrices, c'est-à-dire en français. Voici un exemple d'un message complet pour l'erreur **I have also good news* :

1. I <u>have also good news</u>.
2. Dans cette phrase, l'adverbe *also* est placé entre le verbe *have* et son complément *good news*.
3. En anglais, il est rare de placer les adverbes entre les verbes et leurs compléments.
4. Vous pouvez corriger cette phrase en plaçant l'adverbe *also* avant le verbe.

5. ex. : *I also like apples*.

L'intérêt de distinguer ces cinq étapes est de permettre de les combiner différemment en fonction du public utilisant le système, et de leurs différentes attentes. Nous identifions quatre profils possibles, auxquels correspond une combinaison spécifique dans les éléments du message :

- profil "apprenant/e" : pour ce profil, nous suivons les recommandations des recherche sur la rétroaction corrective en acquisition des langues secondes et en ELAO, et prévoyons d'inclure toutes les étapes à l'exception notable de la correction, dans le but de favoriser l'auto-correction ;
- profil "curieux/se" : ce profil est celui des personnes qui souhaitent avoir le plus d'informations possibles sur le segment erroné, nous prévoyons donc de fournir toutes les étapes, correction incluse ;
- profil "prudent/e" : les personnes correspondant à ce profil sont celles qui souhaitent vérifier les causes de l'erreur, afin de décider si le texte doit être modifié ou non ; uniquement les étapes 1, 2 et 5 peuvent être présentées ;
- profil "pressé/e" : les personnes correspondant à ce profil souhaitent uniquement obtenir une correction ; aucun message au-delà du marquage de l'erreur (et de la correction) n'est inclus.

Nous envisageons deux moyens d'identifier les profils, et donc les messages à fournir. Ceux-ci pourraient être sélectionnés en amont par les personnes utilisant le système ; cette approche implique cependant que les besoins soient stables, alors que ceux-ci peuvent changer selon les types d'erreur. Il est également possible d'adopter la stratégie visible dans le correcteur de *Word* 2007, qui fournit des messages de plus en plus complets selon les demandes de la personne utilisatrice ; il faut ainsi cliquer au moins deux fois, et sur deux éléments différents, pour obtenir le message complet. Le problème posé par cette approche est la relative invisibilité des messages correctifs, qui passent inaperçus pour la plupart des personnes utilisant *Word*. Cependant, si elle venait à être utilisée dans un système d'apprentissage, la présence de messages correctifs pourrait être rendue plus visible en raison de leur pertinence directe pour la tâche à accomplir.

Le contenu d'un petit ensemble de messages a déjà été préparé, mais ceux-ci n'ont pas été implémentés. Une étude exploratoire a montré qu'il était possible de permettre à la personne utilisatrice d'accéder à ces messsages grâce à l'ouverture d'une fenêtre supplémentaire au clic.

Nous n'avons a fortiori pas évalué les réactions de publics authentiques à ces messages, ainsi que l'efficacité réelle de ceux-ci, ce qui constituerait une étape obigatoire du développement d'un tel système de rétroaction corrective.

Conclusion

Une part importante de ce dernier chapitre est consacrée à la présentation de l'implémentation des règles de correction, dans lesquelles la détection des segments erronés est effectuée grâce à des patrons, une technique également appelée "filtrage par motif". Cette approche se distingue des méthodes statistiques souvent utilisées dans le domaine de la correction grammaticale automatisée. Les règles ont été élaborées à partir d'une grammaire des erreurs créée manuellement d'après l'analyse des erreurs et des phénomènes linguistiques concernés présentés dans le Chapitre 2. Elles sont implémentées dans la plateforme <TextCoop>, un analyseur de discours programmé en Prolog. <TextCoop> utilise le langage Dislog, qui est adapté au traitement du discours et est donc utilisable pour la correction automatisée dans des textes. Le fonctionnement des règles de correction dans notre système est expliqué en détail, ce qui permet d'en documenter la création. Les ressources utilisées incluent des lexiques à hauteur de 8129 mots, regroupant la quasi-totalité des catégories syntaxiques (noms, verbes, auxiliaires, adjectifs, adverbes, déterminants, prépositions). Les listes de départ ont été nettoyées et modifiées pour y inclure des informations morphologiques et lexico-sémantiques. Les ressources incluent également une grammaire locale pour le GN créée pour améliorer l'efficacité de la création et de l'utilisation des règles.

La correction automatique à destination d'un public apprenant doit viser à maximiser la précision de la détection et de la correction, même au détriment de la couverture des erreurs. Une étude indicative des performances de sept correcteurs grammaticaux sur un extrait représentatif de notre corpus d'erreurs montre que le plus efficace d'entre eux, *Ginger*, atteint en effet un taux de précision de 100 %, mais avec un taux de seulement 31 % pour le rappel. Cependant, certains correcteurs se trompent dans la détection plus d'une fois sur deux. Nous n'avons pas remarqué de différence entre les performances des correcteurs annoncés comme étant adaptés aux personnes non-natives et celles de correcteurs à destination d'anglophones natifs. Les erreurs liées au placement des adverbes ne sont pas détectées par les systèmes existants, sauf pour des cas marginaux. Certaines des catégories des erreurs liées à l'utilisation des structures N+N sont détectées et corrigées par le système *CorrectEnglish*, et de manière plus éparse par d'autres correcteurs. On remarque un manque de systématicité dans la

correction des erreurs de ces deux types, mais il semblerait que le traitement des erreurs N+N constitue un domaine en expansion.

Nos règles de correction pour les types d'erreurs traités ont d'abord été évaluées sur un corpus d'anglais L1, afin de vérifier qu'elles ne causaient pas de faux positifs dans des textes considérés *a priori* comme bien formés. Les quantités élevées de faux positifs dans le cas des structures N+N ont mené à limiter l'évaluation sur un corpus d'anglais L2 aux règles pour la correction des erreurs de placement des adverbes, en attendant le développement de stratégies plus précises et plus étendues pour ces erreurs. Les principales causes de la production de faux positifs sont des erreurs dans la délimitation des GN, et l'homonymie des noms et des verbes. L'intégration de l'action d'un analyseur syntaxique en amont de l'activation des règles de correction est proposée comme solution à ces difficultés.

Les erreurs de placement des adverbes sont évaluées sur un corpus modifié d'anglais L2. La précision de la détection et de la correction pour les deux types d'erreur atteint des taux dépassant 95 %, mais ceux-ci sont à tempérer en raison de limites dans la méthode d'évaluation. Les taux de rappel initiaux sont plus bas, avec 54,5 % pour les erreurs liées à *also*, et 69,5 % pour les erreurs liées aux adverbes de manière. L'une des principales causes de non-détection est l'absence de certains termes dans le lexique ; des lexiques plus complets permettraient d'augmenter significativement les taux de rappel pour ces deux erreurs. L'extension de la grammaire locale du GN permettrait également de proposer des solutions de correction alternative plus justes. Les insuffisances dans les règles sont également une des causes principales de non-détection, en particulier pour le traitement des erreurs liées à *also*. Les pistes d'amélioration proposées pour les deux types d'erreur regroupent la prise en compte de configurations plus variées dans les dépendants suivant le verbe, et l'inclusion de limites pour la reconnaissance de certains termes.

Cependant, l'observation la plus importante à faire à la suite de l'évaluation des règles concerne l'identification des erreurs. Dans certains cas, la présence de faux positifs ne provient pas d'insuffisances dans les lexiques ou les ressources grammaticales, mais plutôt d'un "excès de zèle" dans la détection des erreurs. En effet, certains segments corrects sont relevés comme des erreurs car ils correspondent exactement aux schémas d'erreur à partir desquels les règles ont été créées. Ce problème apparaît plus particulièrement lors de l'évaluation sur le corpus d'anglais L1 des règles pour la correction des erreurs N+N. Nous en concluons que ces contradictions apparentes sont le résultat de l'influence de facteurs

supplémentaires qui n'ont pas encore été pris en compte, comme par exemple l'association de plusieurs causes d'erreur sur le même segment.

Dans le cadre de la correction automatisée à destination de personnes n'ayant pas l'anglais comme L1, nous envisageons l'accompagnement des corrections par le biais de messages de rétroaction corrective permettant aux personnes utilisatrices de remédier aux erreurs à plus ou moins long terme. Après avoir passé en revue les résultats des recherches sur ce thème dans les domaines de l'acquisition des langues secondes et de l'ELAO, nous prévoyons la création de messages correctifs incluant cinq étapes (marquage de l'erreur, diagnostic, rétroaction métalinguistique, instruction de remédiation, illustration) en plus de la correction, ces cinq étapes indépendantes pouvant être modulées en fonction des profils d'utilisation. Un petit ensemble de messages a été constitué à la main, mais ce module nécessite un développement plus complet, ainsi que l'évaluation de son utilité et de son efficacité.

Conclusion générale

Rappel de la problématique

L'objectif de la recherche présentée dans ce document est la création de règles pour la détection et la correction automatisée d'erreurs produites en anglais par des francophones ayant un niveau de maîtrise intermédiaire à avancé. La recherche devait pouvoir aboutir à un système fonctionnel et implémenté dans une plateforme informatique. Pour atteindre l'objectif fixé, nous avons dû trouver des solutions à trois ensembles de problèmes.

La sélection des erreurs à traiter constituait le premier problème. En effet, nous avions posé trois exigences pour les types d'erreur sélectionnés : ces erreurs devaient correspondre à une difficulté linguistique réelle pour le public apprenant et utilisateur, elles devaient représenter une innovation pour la correction grammaticale automatisée, et enfin leur traitement devait être faisable à l'intérieur de notre cadre pratique et technologique.

La question du cadre technologique constituait justement le deuxième ensemble de questions auxquelles nous souhaitions apporter des réponses. Notre recherche est ancrée dans le domaine linguistique, et les méthodes de traitement automatique utilisées découlent de ce choix préalable. Il s'agissait donc d'identifier les limites et les points forts de ces méthodes, et d'évaluer leurs capacités de traitement face aux types d'erreurs choisies. La question de leur adéquation à la correction d'erreurs pour un public apprenant a également été soulevée.

À ces deux questions, présentes dès le début de la recherche, est venue s'ajouter une problématique ayant émergé à la suite de la sélection des erreurs. Il est devenu évident que cette question est celle qui est au centre de nos travaux, et qui a motivé nos choix de méthodes avant même d'avoir été formulée avec précision. Les deux types d'erreur sur lesquels nous avons travaillé ont en effet en commun le fait d'interroger les normes de grammaticalité et d'acceptabilité, semblant décourager les jugements tranchés et les généralisations. La détection et la correction automatisées ne peuvent toujours pas reproduire la finesse des réactions d'une personne spécialiste de l'anglais confrontée à ces structures, et nécessite, pour fonctionner efficacement, que les phénomènes soient au moins partiellement généralisables. Les deux erreurs sélectionnées pour notre recherche s'opposent ainsi à d'autres types d'erreur pour lesquels l'identification et la correction sont facilitées par l'existence de règles claires, à l'intérieur d'une variété donnée, comme par exemple les erreurs morphosyntaxiques. Confrontés à cette difficulté, nous nous sommes demandé quelles étaient les conséquences de ces normes complexes pour notre objectif, pour la détection comme pour la correction des erreurs.

Détecter les erreurs implique de pouvoir distinguer les segments erronés des formes correctes. Notre recherche a donc porté sur l'identification du seuil entre les segments considérés comme grammaticaux et les segments considérés comme agrammaticaux pour les deux types d'erreur, en prenant également en compte le critère de l'acceptabilité. L'identification de ce seuil s'accompagne de la détermination des facteurs faisant basculer les segments d'un côté ou de l'autre de celui-ci, ainsi que de l'exploration de la possibilité de transposer ces conclusions à un système automatisé. La modélisation linguistique résultant de cette recherche doit éviter deux risques connexes : la production de fausses détections, c'est-à-dire l'identification de formes correctes comme des erreurs, et le biais prescriptif, qui consisterait en un défaut de prise en compte de la variété des expressions possibles, et donc de la complexité de la langue.

Nous présentons les réponses qui ont été apportées à ces problématiques ainsi que les difficultés rencontrées en chemin dans le résumé des chapitres. Les perspectives de recherche envisagées sont exposées dans une section à part à la suite du résumé des chapitres.

Résumé des chapitres et réponse à la problématique

Chapitre 1

La sélection des erreurs à traiter était l'objectif principal du premier chapitre. Afin de faire cette sélection, nous nous sommes appuyés sur la méthode de l'analyse des erreurs, héritée du domaine de l'acquisition des langues secondes, utilisée ici comme méthode de recueil et de traitement des données nécessaires à la sélection des erreurs. Analyser les erreurs implique de définir ce qui constitue une erreur. Dans notre cadre, une erreur est définie comme l'écart pouvant exister entre une production d'interlangue et une production native. Il peut s'agir d'un écart dans la grammaticalité ou l'acceptabilité du segment concerné, le premier critère étant théoriquement objectif et non tributaire du contexte, alors que le second est lié à un ensemble d'usages reconnus par une communauté linguistique ayant la langue cible comme langue maternelle et dépend du contexte de production.

Le relevé des erreurs est effectué dans un corpus d'anglais L2 constitué spécifiquement pour cette recherche. Ce corpus d'environ 100 000 mots est représentatif de l'utilisation de l'anglais par notre public cible, et regroupe quatre types différents de production écrite (publications scientifiques, productions d'interlangue, courriels professionnels et semi-professionnels, rapports techniques industriels). Il s'agit à notre connaissance du premier corpus d'anglais L2 envisagé spécifiquement comme un corpus d'utilisation de l'anglais écrit

par des francophones, c'est-à-dire incluant des textes produits en dehors d'un contexte d'apprentissage. Les 661 erreurs relevées manuellement dans ce corpus sont ensuite catégorisées selon un système de 31 catégories donnant des informations sur le domaine linguistique de l'erreur. Nous avons également créé un schéma d'annotation utilisant le format XML et permettant d'inclure des informations nuancées sur les erreurs et les corrections, comme l'évaluation de l'intelligibilité du segment et la confiance de la personne annotant les erreurs concernant les corrections proposées. À ce jour, ce schéma a été appliqué à un échantillon de 31 erreurs.

Les deux types d'erreur répondant aux exigences posées pour cette recherche font partie des six types les plus fréquemment rencontrés dans notre corpus, avec 6,8 % des erreurs pour l'utilisation des N+N et 5,4 % pour le placement des adverbes dans la proposition ou le groupe verbal. La proportion relativement importante d'erreurs N+N, qui, à notre connaissance, est spécifique à notre recherche, s'explique par l'inclusion dans notre corpus de productions à caractère informatif dans lesquelles l'utilisation de ces structures est courante. Ces deux erreurs n'ont pas fait l'objet de recherches approfondies dans le domaine de la correction automatisée, bien que certaines erreurs N+N soient relevées marginalement par un correcteur grammatical disponible en ligne parmi les sept correcteurs évalués dans le Chapitre 3. Ces deux types répondent parfaitement à deux de nos trois exigences de départ concernant les erreurs à traiter en priorité.

L'exigence de faisabilité n'est cependant pas respectée avec le choix de traiter les erreurs N+N, puisque nous avons vu dans les Chapitres 2 et 3 que la correction automatique de la plupart de ces erreurs nécessite des ressources supplémentaires que nous n'avons pas été en mesure de développer dans le cadre de cette recherche, notamment en terme de connaissances sémantiques et d'analyse syntaxique automatique. Ce choix répond cependant tout à fait à l'exigence de pertinence linguistique, puisque notre recherche a permis de mettre en lumière la fréquence, la distribution et la typologie des erreurs relatives aux structures N+N, structures dont l'utilisation est en expansion, et de proposer des pistes de correction automatisée à explorer.

L'autre limite importante à souligner concerne la détection manuelle des erreurs dans le corpus, qui a été effectuée uniquement par l'auteure de la recherche. Les résultats de cette étape sont donc subjectifs, en particulier pour les formes régies par des normes complexes, ce qui est justement l'objet de notre recherche. Afin d'obtenir des résultats plus fiables, cette tâche doit être effectuée par au moins deux personnes présentant les mêmes compétences. Il

s'agit cependant d'une tâche cognitivement lourde, chronophage, et nécessitant un niveau de formation en anglais très élevé. Le problème posé par les tâches complexes nécessitant une intervention humaine existe pour une majorité de projets en traitement automatisé des langues, et des solutions originales ont été explorées, comme le recours au *Mechanial Turk* d'Amazon, que nous avons déjà évoqué. Dans notre cas, il est peu réaliste d'envisager de pouvoir soumettre la tâche de détection sur le corpus entier à une ou deux personnes supplémentaires correspondant exactement à nos exigences, surtout si la taille du corpus vient à être augmentée. Des solutions alternatives peuvent néanmoins être testées, comme la possibilité de diviser le corpus en parties plus courtes et de soumettre la détection à plus de personnes. Une autre solution consisterait à effectuer une première détection, pour soumettre aux personnes un document pré-annoté et ainsi alléger la tâche demandée.

Chapitre 2

Le second chapitre avait pour objectif de fournir une modélisation linguistique des erreurs et de leurs corrections permettant le passage à l'implémentation de règles de correction dans une plateforme informatique. C'est donc dans ce chapitre que se pose plus particulièrement le problème de l'identification des normes linguistiques et des facteurs intervenant dans les deux types d'erreurs sélectionnés dans le premier chapitre. La méthode utilisée afin de déterminer le seuil de grammaticalité ou d'acceptabilité pour le placement des adverbes et l'emploi des structures N+N combine la synthèse des descriptions linguistiques disponibles pour ces phénomènes à l'analyse de la forme et des causes linguistiques précises des erreurs produites par notre public cible. Les sources utilisées pour la synthèse linguistique sont principalement les trois grammaires de référence les plus récentes pour l'anglais, dont les approches sont complémentaires. En pratique, le cadre descriptif retenu dans la plupart des cas est celui de *The Cambridge Grammar of the English Language*, en vertu de la précision des descriptions pour les deux phénomènes qui nous préoccupent. Par ailleurs, cet ouvrage plus récent a l'avantage de pouvoir utiliser les descriptions et conclusions des deux grammaires précédentes pour les approfondir, et bénéficie de plusieurs décennies supplémentaires de recherches en grammaire. Les informations présentées dans les trois ouvrages sont complétées ponctuellement par d'autres études portant sur des aspects précis des phénomènes, comme celles de Fjelkestam-Nilsson pour *also* (1983) et Pastor-Gómez pour les structures N+N (2010, 2011).

Les recherches effectuées dans le domaine de l'acquisition des langues confirment la difficulté posée par le placement des adverbes en anglais pour les francophones. La synthèse

du placement des adverbes effectuée à partir des trois grammaires de référence montre la multiplicité des placements possibles ainsi que des facteurs régissant ce placement. Ceux-ci sont d'ordre sémantique, syntaxique, lexical, et prosodique. Le facteur le plus déterminant dans ce placement est le type sémantique de l'adverbe, associé à la portée du GAdv correspondant, en lien avec le type sémantique. Afin de détecter et de corriger les erreurs dans le placement des adverbes, il est donc primordial de pouvoir reconnaître le type de l'adverbe concerné ; c'est pourquoi nous avons créé une méthode pour la constitution d'une ressource lexico-sémantique d'adverbes anglais pouvant être intégrée à un système informatique. Malgré un manque de consensus apparent, le placement d'un adverbe entre un verbe lexical et un groupe fonctionnant comme complément d'objet du verbe est reconnu comme généralement inacceptable. Cette impossibilité est cependant levée en cas de présence d'un GN complément long, en particulier lorsqu'une collocation semble exister entre le verbe et l'adverbe, comme souligné par Osborne dans son étude comparative (2008). On postule cependant que ce lien peut aussi être interprété comme un lien de complémentation. L'analyse des erreurs relevées dans notre corpus indique que les adverbes le plus souvent concernés par les erreurs sont les adverbes de manière et l'adverbe focalisant *also*. La modélisation des erreurs et des corrections associées est donc limitée à ces deux types.

L'étude des placements corrects et erronés d'*also* est facilitée par le fait qu'il s'agit d'un adverbe unique fréquemment utilisé. Nous sommes donc en mesure d'observer son placement dans des corpus d'anglais L1 et d'anglais L2. Les résultats de ces études de corpus, complétés par les descriptions faites par Fjelkestam-Nilsson, permettent de modéliser les placements standards et non-standards de cet adverbe, en prenant en compte les spécificités des deux variétés étudiées (variété américaine et variété britannique). Sept schémas d'erreur et de correction sont identifiés. Ceux-ci prennent en compte la présence de dépendants du verbe prenant la forme d'un GN ou d'un GP, ou encore d'une proposition infinitive ou introduite par *that*. La présence d'auxiliaires dans le GV influence les corrections à apporter et est donc également intégrée aux schémas. Un seul des schémas ne concerne pas le placement SVAO.

La variété des adverbes de manière et le fait qu'il n'existe pas de corpus d'anglais L1 ou L2 permettant de rechercher les adverbes par type sémantique nous obligent à adopter une approche différente pour la description des placements acceptables et inacceptables pour ces adverbes. Nous utilisons donc les résultats de l'étude de corpus menée par Jacobson (1964), ainsi que les descriptions sémantiques et syntaxiques proposées par Dixon (2005), qui viennent compléter les descriptions des placements acceptables données dans les grammaires

de référence. Nous poussons l'exploration des facteurs déterminant le placement des adverbes de manière à travers une étude des jugements de grammaticalité donnés par des anglophones. La synthèse des résultats de ces études différentes mène à la création de quatre schémas d'erreur accompagnés de leurs corrections. Ceux-ci prennent en compte uniquement les erreurs de placement des adverbes entre un verbe et un groupe nominal. La longueur du groupe nominal est prise en compte dans les schémas. Les schémas incluent également des variations en cas de présence d'un verbe à particule et de l'adverbe *very* en fonction de modifieur dans le groupe adverbial concerné.

Dans le cas du traitement des erreurs dans le placement des adverbes, la difficulté posée par la complexité des normes et des facteurs régissant ce phénomène est partiellement surmontée en multipliant les angles d'approche. Les indications de grammaticalité ou d'acceptabilité provenant de grammaires de référence ou de tests effectués par des anglophones sont complétées par des études de corpus donnant des indications de fréquence d'utilisation des structures. Il apparaît cependant que la recherche de jugements absolus concernant la grammaticalité et l'acceptabilité des placements est vaine, et que la prise en compte de la complexité de l'anglais dans ce cas passe par l'acceptation de l'existence d'une marge de variabilité dont la correction grammaticale automatisée doit s'accommoder. Il semble que la présence d'un léger biais prescriptif soit inévitable, mais celui-ci est sans doute d'avantage l'apanage du processus d'enseignement et d'apprentissage d'une L2 que du recours à un outil de correction automatique. Ces questions se posent également pour le traitement des erreurs N+N, mais leur traitement dans notre projet s'est heurté à des difficultés supplémentaires.

Bien que les structures N+N soient envisagées comme des nominaux composites, c'est-à-dire des constructions en syntaxe, elles répondent pourtant à bien peu de règles syntaxiques. Leur simplicité de forme est même probablement une des raisons de leur attrait pour les francophones écrivant des textes informatifs, et la complexité des autres facteurs qui les régissent, une des raisons des nombreux échecs d'utilisation observés dans le corpus. Nous identifions cinq sous-types d'erreur à partir des segments problématiques relevés dans le corpus. Ces cinq sous-types représentent de manière indirecte les facteurs à prendre en compte pour leur correction. Ces facteurs sont d'ordre formel, sémantique et pragmatique. Du point de vue de la forme des segments, on relève les cas d'empilement de plus de deux noms en fonction de dépendant, qui présentent un risque pour l'intelligibilité du segment. De plus, dans certains segments dont le N_1 est au pluriel, ce qui est possible dans un ensemble de cas

définis, le segment correspond à une forme au génitif mal représentée. Ces deux types de segments sont facilement reconnaissables en raison de leurs formes distinctives. Ils font donc l'objet d'une modélisation en vue de la création de règles de correction, avec quatre configurations proposées pour les segments empilés, mais pas de correction, et un seul schéma pour le cas des N_1 au génitif, associé à quatre possibilités de correction.

Les trois autres sous-types identifiés sont liés aux relations sémantiques exprimées par la séquence N+N, à la présence d'un N_1 défini et non générique, et enfin à des erreurs de choix lexical. La détection et la correction de ces erreurs ne peut pas passer par leur modélisation syntaxique. Nous proposons des pistes pour la détection de ces erreurs à partir de relevés de fréquence obtenus lors de requêtes sur un moteur de recherche en ligne. Pour les erreurs sémantiques, qui sont les plus fréquentes, la détection pourrait être mise en place à partir de l'identification des relations existant entre les noms, leur confrontation aux relations possibles, et la proposition d'une correction dans laquelle le choix de la préposition découle de la relation sémantique identifiée. Ces tâches nécessitent des recherches supplémentaires approfondies.

La mention de ces erreurs nous permet d'aborder une limite importante de notre approche de la détection et de la correction de ces deux types d'erreur. Notre approche dès le départ a été majoritairement syntaxique et grammaticale, mais l'analyse détaillée de la formation des séquences N+N et des erreurs produites a révélé l'obligation de prendre en compte les aspects sémantiques et lexicaux. Les recherches qui pourraient permettre de corriger ces erreurs, et sur lesquelles nous avons ouvert quelques portes, nécessitent qu'un projet de recherche entier leur soit consacré, en raison de la complexité de la tâche que constitue l'interprétation du sens des nominaux composites, et encore plus des erreurs dans leur combinaison sémantique. Cette recherche pourra s'appuyer sur les travaux existants en linguistique sur la sémantique des groupes nominaux, et sur la désambiguïsation des groupes nominaux dans le domaine du traitement automatisé des langues. Concernant les adverbes, même si leurs types sémantiques sont bien pris en compte dans la correction des erreurs dans leur placement, il nous semble que leur traitement automatique pourrait être nettement amélioré par la prise en compte d'aspects lexicaux, comme l'identification des verbes et des adverbes pouvant entrer dans une relation de complémentation.

Chapitre 3

Le troisième chapitre présente l'implémentation des règles de correction. Ce chapitre débute par un passage en revue de l'évolution des méthodes informatiques utilisées pour la correction grammaticale automatisée. Les méthodes les plus utilisées à ce jour reposent majoritairement sur l'utilisation de modèles de langue nécessitant l'entraînement de systèmes sur des corpus gigantesques permettant la production de probabilités quant aux formes pouvant être rencontrées dans les textes bien formés. La méthode utilisée dans notre projet s'apparente plutôt aux méthodes fondées sur la description des erreurs permettant leur reconnaissance dans les textes. Par ailleurs, nous reprenons les exigences pesant sur les systèmes de correction automatisée à destination des apprenants énoncées par Tschichold. Ceux-ci concernent l'importance de privilégier la précision de la détection et de la correction, et de favoriser l'autonomie des personnes utilisatrices. L'évaluation d'une sélection de correcteurs grammaticaux disponibles montre que les deux types d'erreur sur lesquels porte notre recherche font l'objet de détections sporadiques. Ceci est vrai en particulier pour les erreurs N+N, qui semblent susciter un intérêt croissant.

Les règles de correction sur lesquelles le système final repose sont écrites dans le langage Dislog, qui fonctionne dans la plateforme <TextCoop> programmée en Prolog. Les ressources lexicales et grammaticales qui ont dû être créées ou rassemblées pour permettre le fonctionnement des règles concernent une grande partie des catégories grammaticales, et incluent également les formes flexionnelles des termes faisant partie des lexiques. Le système de correction est évalué grâce à une méthode permettant également de dégager des pistes d'amélioration. La première étape consiste à évaluer le système sur un corpus d'anglais L1. Les règles créées pour le traitement des erreurs N+N ne réussissent pas ce test, donnant lieu à de trop nombreux faux positifs. Ces faux positifs sont d'autant plus problématiques qu'ils remettent en question les schémas d'erreur identifiés. Les règles de correction pour le placement des adverbes de manière et d'*also*, dont les résultats sont jugés satisfaisants sur le corpus d'anglais L1, donnent également des résultats relativement satisfaisants lorsqu'elles sont évaluées sur un corpus modifié d'anglais L2. Le traitement des erreurs liées à *also* atteint un taux de précision de 100 %, ce qui répond à l'exigence de haute précision posée pour les systèmes créés à destination d'un public apprenant. Le taux de précision élevé est cependant contrebalancé par un taux de rappel pouvant aller de 54,5 % à 66,7 % selon les cas de non-détection pris en compte. Le traitement des erreurs liées aux adverbes de manière obtient également des taux de précision élevés pour la détection comme pour la correction, dépassant

les 95 %. Le rappel est bien meilleur pour les règles concernant les adverbes de manière, avec des taux allant de 69,5 % à 91,5 % selon les cas de non détection pris en compte.

En dépit des limites à prendre en compte dans la considération des résultats, et que nous exposons ci-dessous, ceux-ci semblent indiquer que l'utilisation de méthodes linguistiques est une piste productive et instructive pour la correction grammaticale automatisée, au moins pour la correction des erreurs relatives au placement des adverbes. La pertinence des méthodes linguistiques dans le traitement des erreurs N+N peut sembler moins évidente en raison des résultats obtenus ici, mais nous avons vu que la correction de ces erreurs nécessite surtout des recherches en linguistique, notamment dans le but d'identifier les relations sémantiques existant entre les noms. Le manque de résultats n'est ainsi pas dû à l'impossibilité de corriger ces erreurs avec des méthodes linguistiques, mais plutôt à la nécessité d'orienter les recherches linguistiques dans une autre direction.

Concernant les erreurs dans le placement des adverbes, une étude approfondie des phénomènes, associée à une approche prudente et progressive menant à la modélisation des erreurs et des corrections, permet de corriger une large proportion d'erreurs réalistes, sans créer un nombre de faux positifs alarmant. Cette approche permet également de coller au plus près des besoins des personnes apprenantes dans ce domaine, avec un public cible très précis. On remarque cependant que le recours à des outils d'analyse automatique extérieurs, notamment pour la résolution des problèmes d'homonymie, ou bien à des stratégies statistiques, comme le relevé de la fréquence des retours des segments dans un moteur de recherche, pourrait venir compléter de manière avantageuse les règles fondées sur des analyses linguistiques. Comme dans de nombreux domaines, il semble que la réponse appropriée aux défis posés par la correction grammaticale automatisée consiste en l'utilisation combinée de plusieurs approches et outils, visant à tirer profit des points forts de chaque élément, tout en conservant l'analyse linguistique des phénomènes au cœur des stratégies.

La dernière étape du traitement des erreurs consiste à fournir des messages correctifs. L'objectif de cette étape est de favoriser l'autonomie des personnes utilisant le système de correction automatisée. Cet intérêt a été bien compris par les équipes ayant conçu les correcteurs existants, puisqu'ils proposent tous des messages accompagnant les corrections. Il est vrai que l'utilité de la rétroaction corrective a été longuement débattue dans le domaine de l'acquisition des langues ; cependant, il semblerait que celle-ci soit désormais reconnue même si elle a fait l'objet de peu d'études dans le domaine spécifique de l'enseignement des langues assisté par ordinateur. Nous utilisons les résultats de recherche et les recommandations

provenant de ce domaine pour informer la création de messages correctifs accompagnant les corrections de notre système. Les messages correctifs suivent un canevas en cinq étapes, incluant le marquage de l'erreur, la correction, l'explication, les instructions de remédiation et enfin une ou plusieurs illustrations. Ce canevas est modulable afin de correspondre au contexte d'utilisation du système ainsi qu'au profil de la personne utilisant le système.

La première limite importante à mentionner concernant les travaux présentés dans ce chapitre porte sur les modalités de l'évaluation, qui est effectuée sur des corpus d'anglais L2 modifiés par la personne ayant créé les règles de correction. Il est sans doute peu réaliste d'envisager de pouvoir évaluer les stratégies sur des corpus complètement naturels en raison du manque de corpus d'interlangue adaptés. Cependant, il est évident que les règles devraient être évaluées par des personnes extérieures, et sur des corpus beaucoup plus conséquents. Il sera particulièrement avantageux de conduire cette évaluation pour une version prochaine du système, après la prise en compte et l'implémentation des améliorations indiquées par l'analyse des retours de la présente évaluation. De la même façon, il est nécessaire d'évaluer le fonctionnement et la pertinence des messages correctifs, qui existent à l'état d'échantillons à la fin de notre projet.

Perspectives de recherche

À court terme, le développement le plus évident pour notre recherche est l'amélioration des règles de correction à partir des indications données à la suite de l'évaluation du système. Celle-ci s'applique particulièrement au traitement des erreurs relatives au placement des adverbes. Certaines améliorations sont ponctuelles et clairement définies, comme l'intégration de schémas plus variés pour les compléments à la suite du verbe. D'autres améliorations impliquent des modifications plus importantes du système, comme l'ajout d'un outil supplémentaire permettant l'analyse syntaxique automatique des groupes nominaux. Comme nous l'avons déjà mentionné, ces améliorations doivent être faites uniquement après que des recherches linguistiques supplémentaires, comme des études de corpus, nous permettent de préciser les modalités de leur implémentation, afin d'éviter que celles-ci déclenchent des détections ou des corrections erronées.

Le développement du système de correction pour les adverbes passe également par l'ajout d'informations lexicales concernant les collocations verbe/adverbe, ainsi que par le développement de la reconnaissance des différents types sémantiques des adverbes. Cette amélioration implique le développement de la ressource lexico-sémantique d'adverbes pour

notre recherche et son implémentation technique dans le système. Ce développement pourrait également avoir pour résultat l'ouverture du système à la prise en compte de types sémantiques plus variés, pour lesquels des problématiques spécifiques doivent encore être identifiées.

Comme il a été mentionné, le traitement des erreurs relatives aux structures N+N est loin d'être abouti, mais des portes intéressantes ont été ouvertes vers des recherches sur l'identification automatique de relations sémantiques basée uniquement sur les noms et ne pouvant se reposer sur la présence de prépositions. Il sera nécessaire de revenir aux erreurs pour remettre en question certains des schémas relevés, et identifier des causes plus fines liées aux associations de plusieurs facteurs formels, sémantiques et lexicaux dans la production de segments maladroits. Il faudra également s'intéresser de plus près au développement de l'utilisation des segments N+N, notamment dans les publications scientifiques, afin d'identifier d'éventuelles tendances émergentes dans leur utilisation. Dans cette optique, il sera sans doute nécessaire de limiter la recherche à un domaine scientifique précis.

La proposition de messages correctifs dans notre projet a pour objectif de favoriser l'autonomie des personnes utilisatrices, comme préconisé dans les recherches en enseignement des langues assisté par ordinateur. L'accompagnement des corrections peut cependant prendre une nature différente. Nous avons vu que pour certaines règles, dont la règle qui est la plus souvent mobilisée pour les adverbes de manière d'après les résultats de l'évaluation, plusieurs corrections sont proposées. Les personnes apprenantes utilisant des correcteurs grammaticaux automatisés n'ont pas forcément les connaissances requises pour évaluer les différences entre les propositions de correction, et l'adéquation de celles-ci avec leurs objectifs de communication. Le schéma d'annotation des erreurs que nous avons présenté dans le Chapitre 1 inclut des attributs accompagnés de leur valeur (de 0 à 2 le plus souvent). Ceux-ci peuvent être utilisés comme base informative pour la génération de messages argumentatifs destinés à aider la personne utilisant le système à faire un choix entre au moins deux propositions de correction. Les annotations incluent des informations sur la modification possible du sens du message et de sa longueur, et sur la certitude de l'annotateur quant à la présence d'une erreur. Ces annotations peuvent être incluses dans les règles de correction, et utilisées pour la génération de messages argumentatifs grâce à un processus de décision automatisé qui reste à définir. Le processus de décision peut donner lieu à la génération de textes argumentatifs à partir d'une banque d'énoncés, de manière semblable à la méthode de génération des messages correctifs. De nombreux paramètres techniques et

théoriques devront être définis afin d'explorer cette piste d'accompagnement des corrections de manière approfondie.

À moyen terme, on peut envisager l'utilisation des règles issues de nos recherches dans des systèmes plus complets. Les erreurs relatives au placement des adverbes sont présentes dans des proportions relativement proches dans les sous-corpus de publications, de productions interlangue et de rapports techniques, indiquant que l'on retrouve ces erreurs dans un ensemble varié de textes pouvant être produits par des personnes apprenantes. Il peut ainsi être profitable de les intégrer à une plateforme d'apprentissage. Cette intégration peut prendre plusieurs formes. Il peut s'agir simplement d'un correcteur grammatical automatique à destination d'un public francophone apprenant l'anglais, ou bien les informations grammaticales rassemblées et les règles de corrections peuvent servir à créer des exercices ou tâches d'apprentissage liées à l'utilisation des adverbes. L'intérêt de l'utilisation des règles de correction que nous avons élaborées est la possibilité de traiter du texte libre, ce qui permet d'offrir une alternative aux exercices à trous ou à choix multiples pouvant être jugés répétitifs ou trop faciles pour les publics maîtrisant l'anglais à un niveau B2-C1. Le développement des recherches en ce sens nécessite de recenser les différents programmes existants afin d'évaluer la pertinence et d'étudier les modalités de l'introduction de nos règles de correction dans ces systèmes.

Les règles de détection et correction du placement des adverbes peuvent également intéresser le domaine de la rédaction de documents techniques. En effet, la rédaction de ces documents spécifiques correspond aux domaines dans lesquels des francophones sont amenés à rédiger des textes en anglais dans un cadre professionnel, sans nécessairement bénéficier de corrections de leurs productions. La justesse et la précision de ces documents est néanmoins essentielle, puisqu'ils décrivent souvent des processus à risques, et sont lus par d'autres personnes non-anglophones. Au-delà de la correction des erreurs de placement, les améliorations envisagées pour le système, comme la prise en compte des aspects sémantiques et lexicaux, devraient être adaptés spécifiquement au traitement des documents techniques, afin d'en améliorer l'efficacité. Le développement de ces recherches nécessite de s'intéresser de près à ces documents très spécifiques et d'étudier les besoins réels existant dans ce domaine.

Les erreurs dans la création et l'utilisation des segments N+N se retrouvent quasiment uniquement dans les publications scientifiques, et comme nous l'avons vu, elles relèvent plus de l'inacceptabilité et des problèmes d'intelligibilité que de règles grammaticales. Pour ces

raisons, il ne paraît pas opportun de les intégrer à un système d'apprentissage à destination de publics apprenants larges. Une fois améliorées selon les pistes données plus haut, les règles pourront par contre être intégrées à des plateformes d'aide à la rédaction permettant de vérifier la fluidité du style et l'intelligibilité des structures, en particulier pour la rédaction de publications scientifiques, éventuellement limitées à un certain domaine scientifique. Il sera nécessaire d'identifier les plateformes ou correcteurs existants déjà pour cette tâche. L'intégration des règles pour les structures N+N à des aides automatiques pour la rédaction de documentation technique est également une possibilité intéressante.

La dernière perspective de recherche envisagée nous renvoie à la première étape de recherche, pour lui faire adopter un chemin différent. Comme nous l'avons indiqué précédemment, le corpus constitué pour notre recherche est à notre connaissance le premier corpus de productions écrites en anglais spécifiquement envisagé comme un corpus représentatif de l'utilisation de l'anglais, et non de son apprentissage. Il s'apparente donc aux corpus ayant été créés pour l'observation des spécificités de l'anglais comme *lingua franca*, mais ces corpus regroupaient des transcriptions de productions orales. On peut se demander si les productions écrites en anglais *lingua franca* se distinguent des productions d'interlangue ou des productions natives, et si oui, de quelle(s) façons(s). Le corpus peut donc être développé afin d'atteindre la plus grande représentativité possible, et permettre ainsi des observations fiables.

La négociation du sens dans des contextes de discussions spontanées étant une des spécificités observées dans l'utilisation de l'anglais comme *lingua franca*, il serait intéressant d'ajouter au corpus des textes issus des nouveaux médias incorporant des discussions écrites. Les commentaires laissés sur des forums ou des blogs constitueraient un ajout précieux pour observer cette négociation, et sont une des formes de communication en expansion sur internet entre personnes ayant des L1 différentes.

Enfin, nous envisageons également l'observation "positive" des données du corpus, par le biais de l'étude des spécificités de cet anglais émergent. Après avoir consacré une longue recherche à la détection et à la correction des erreurs que peuvent produire des personnes utilisant l'anglais comme L2 ou comme langue véhiculaire, il semble tout aussi scientifiquement justifié qu'intellectuellement rafraîchissant de s'intéresser aux contributions originales de ces personnes à la continuité de l'évolution de l'anglais.

Bibliographie

Ouvrages et articles

AARTS, Flor. 1988. "*A Comprehensive Grammar of the English Language*: the Great Tradition continued". *English Studies*, vol. 2. 163-173.

AARTS, Bas. 2004. "Grammatici certant". *Journal of Linguistics*, vol. 40, n°2. 365-382.

ALONSO, Maria Rosa. 2002. *The Role of Transfer in Second Language Acquisition*. Vigo : University of Vigo Press.

AYOUN, Dalila. 2005. "Verb movement in the L2 acquisition of English by adult native speakers of French". *EuroSLA Yearbook*, vol. 5. 35-76.

BELLERT, Irena. 1977. "On the semantic and distributional properties of sentential adverbs". *Linguistic Inquiry*, vol. 8. 337-351.

BENCZES, Réka. 2006. *Creative compounding in English: the semantics of metaphorical and metonymical noun-noun combinations*. Amsterdam : John Benjamins.

BENDER, Emily M., FLICKINGER, Dan, OEPEN, Stephan, WALSH, Annemarie et BALDWIN, Timothy. 2004. "Arboretum: Using a precision grammar for grammar checking in CALL". *Proceedings of InSTIL/ICALL Symposium, 17-19 juin 2004, Venise*. 83-86.

BIBER, Douglas, JOHANSSON, Stig, LEECH, Geoffrey, CONRAD, Susan et FINEGAN, Edward. 1999. *Longman Grammar of Spoken and Written English*. Harlow : Pearson Education.

BIBER, Douglas et Clark, Victoria. 2002. "Historical shifts in modification patterns with complex noun phrase structures". *English Hisorical Syntax and Morphology*. Éds. Teresa Fanego, María-José López-Couso, Javier Pérez-Guerra. Amsterdam : John Benjamins. 43-66.

BIGERT, Johnny, SJÖBERGH, Jonas, KNUTSSON, Ola et SAHLGREN, Magnus. 2005. "Unsupervised evaluation of parser robustness". *Proceedings of CICling, 13-19 février, Mexico*. 142-154.

BOUCHER, Paul, DANNA, Frédéric et SÉBILLOT, Pascale. 1993. "Compounds: an intelligent tutoring system for learning to use compounds in English". *Computer Assisted Language Learning*, vol. 6, n°3. 249-272.

BROCKETT, Chris, DOLAN, William B. et GAMON, Michael. 2006. "Correcting ESL errors using phrasal SMT techniques". *Proceedings of the 21st International Conference on Computational Linguistics, 17-21 juillet 2006, Sydney*. 249-256.

BROWN. Keith. 2006. *Encyclopedia of Language and Linguistics*. 2ème éd. Amsterdam : Elsevier.

BRUMFIT, Christopher. 2001. *Individual Freedom in Language Teaching: Helping Learners to Develop a Dialect of their Own*. Oxford : Oxford University Press.

BUCKERIDGE, Alan M. et SUTCLIFFE, Richard F. E. 2002. "Disambiguating noun compounds with latent semantic indexing". *Proceedings of the 2nd International Workshop on Computational Terminology, COLING, 23 août 2002, Taipei*. 1-7.

BURT, Marina K. 1975. "Error analysis in the adult EFL classroom". *TESOL Quaterly*, n°9. 53-63.

BUTNARIU, Cristina, KIM, Su Nam, NAKOV, Preslav, Ó SÉAGHDHA, Diarmuid, SZPAKOWICZ, Stan, VEALE, Tony. 2009. "SemEval-2010 Task 9: The Interpretation of Noun Compounds Using Paraphrasing Verbs and Prepositions". *Proceedings of the NAACL HLT Workshop on Semantic Evaluations, juin 2009, Boulder*. 100–105.

CHANG, Yu-Chia, CHANG, Jason S., CHEN, Hao-Jan et LIOU, Hsien-Chin. 2008. "An automatic collocation writing assistant for Taiwanese EFL learners: A case of corpus-based NLP technology". *Computer Assisted Language Learning*, vol. 21, n°3. 283-299.

CHEN, Hao-Jen Howard. 2009. "Evaluating Two Web-based Grammar Checkers - Microsoft ESL Assistant and NTNU Statistical Grammar Checker". *Computational Linguistics and Chinese Language Processing*, vol. 14, n°2. 161-180.

CHODOROW, Martin, TETREAULT, Joel et HAN, Na-Rae. 2007. "Detection of grammatical errors involving prepositions". *Proceedings of the 4th ACL-SIGSEM Workshop on Prepositions, 28 juin 2007, Prague.* 25-30

CHODOROW, Martin, GAMON, Michael et TETREAULT, Joel. 2010. "The utility of article and preposition error correction systems for English language learners: Feedback and assessment". *Language Testing*, vol. 27, n°3. 419-436.

CHOMSKY, Noam. 1957. *Syntactic Structures*. La Haye, Paris : Mouton.

COGO, Alessia et DEWEY, Martin. 2006. "Efficiency in ELF communication: from pragmatic motives to lexico-grammatical innovation". *Nordic Journal of English Studies*, vol. 5, n°2. 59-93.

CONLON, Sumali Pin-Ngern et EVENS, Martha. 1992. "Can computers handle adverbs?". *Proceedings of the 15th International Conference on Computational Linguistics, 23-28 août 1992, Nantes*. 1192-1196.

CONSEIL DE L'EUROPE. 2001. *Un cadre européen commun de référence pour les langues: apprendre, enseigner, évaluer*. Paris : Didier.

COOK, Vivian. 1993. *Linguistics and Second Language Acquisition*. Basingstoke : McMillan.

COOK, Vivian. 2002. "Background to the L2 User". *Portraits of the L2 User*. Éd. V. Cook. Clevedon : Multilingual Matters. 1-28.

CORDER, Stephen Pit. 1981. *Error Analysis and Interlanguage*. Oxford : Oxford University Press.

CORNILLIE, Frederik, CLAREBOUT, Géraldine et DESMET, Piet. 2012. "Between learning and playing? Exploring learners' perceptions of corrective feedback in an immersive game for English pragmatics". *ReCALL*, vol. 24, n°3, 257-278.

COSTELLO, Fintan J., VEALE, Tony et DUNNE, Simon. 2006. "Using WordNet to automatically deduce relations between words in Noun-Noun compounds". *Proceedings of the 21st International Conference on Computational Linguistics, 17-21 juillet 2006, Sydney*. 160-167.

CRYSTAL, David. 2003. *The Cambridge Encyclopedia of the English Language*. 2ème éd. Cambridge : Cambridge University Press.

CULICOVER, Peter W. 2004. "Review article: The Cambridge Grammar of the English Language". *Language*, vol. 80, n°1. 127-141.

CUSHING, Ellen. 2012. "Dawn of the digital sweatshop". *East Bay Express*, 1 août 2012. 6 mars 2013. http://www.eastbayexpress.com/oakland/dawn-of-the-digital-sweatshop/Content?oid=3301022&showFullText=true

DAGNEAUX, Estelle, DENNESS, Sharon et GRANGER, Sylviane. 1998. "Computer-aided error analysis". *System*, n°26. 163-174.

DALGISH, Gerard M. 1991. "Computer-assisted error analysis and courseware design: applications for ESL in the Swedish context". *CALICO Journal*, vol. 9, n°2. 36-56.

DANSUWAN, Suyada, NISHINA, Kikuko, AKAHORI, Kanji et SHIMIZU, Yasutaka. 2001. "Development and evaluation of a Thai learning system on the web using Natural Language Processing". *CALICO Journal*, vol. 19, n°1. 67-88.

DE FELICE, Rachele. 2008. "Automatic error detection in non-native English". Thèse. University of Oxford.

DE FELICE, Rachele et PULMAN, Stephen. 2009. "Automatic detection of preposition errors in learner writing". *CALICO Journal*, vol. 26, n°3. 512-528.

DELFITTO, Denis. 2006. "Adverb classes and adverb placement". *The Blackwell Companion to Syntax*. Vol. 1. Éds. M. Everaert, H. Van Riemsdijk, R. Goedesmans, B. Holledbrandse. Wiley-Blackwell. 83-120.

DELMONTE, Rodolfo. 2003. "Linguistic knowledge and reasoning for error diagnosis and feedback generation". *CALICO Journal*, vol. 20, n°3. 513-532.

DIAB, Nuwar. 1997. "The transfer of Arabic in the English writings of Lebanese students". *The ESPecialist*, vol. 18, n° 1. 71-83.

DÍAZ-NEGRILLO, Ana et FERNÀNDEZ-DOMÍNGUEZ, Jesus. 2006. "Error-tagging systems for learner corpora". *RESLA*, n°19. 83-102.

DIXON, R. M. W. 2005. *A Semantic Approach to English Grammar*. 2ème édition. Oxford : Oxford University Press.

DOWNING, Pamela. 1977. "On the creation and use of English compound nouns". *Language* vol. 53, n°4.810-842.

DULAY, Heidi C., BURT, Marina K., KRASHEN, Stephen D. 1982. *Language Two*. New York, Oxford : Oxford University Press.

DUSKOVA, Libuse. 1969. "On sources of errors in foreign language learning." *International Review of Applied Linguistics*, vol. 7, n°1. 11-36.

EEG-OLOFSSON, Jens et KNUTSSON. 2003. "Automatic grammar checking for second language learners – the use of prepositions". *Proceedings of the 14th Nordic Conference in Computational Linguistics, 30-31 mai 2003, Reykjavik.*

ELLIS, Rod. 2008. *The Study of Second Language Acquisition*. 2ème éd. Oxford : Oxford University Press.

FABRE, Cécile. 1996. "Interpretation of nominal compounds: combining domain-independent and domain-specific information". *Proceedings of the 16th International Conference on Computational Linguistics, 5-9 août 1996, Copenhague.* 364-369.

FERRIS, Dana. 1999. "The Case for Grammar Correction in L2 Writing Classes: A Response to Truscott (1996)". *Journal of Second Language Writing*, vol. 8, n° 1. 1-11.

FERRIS, Dana. 2004. "The "Grammar Correction" Debate in L2 Writing: Where are we, and where do we go from here? (and what do we do in the meantime ...?)". *Journal of Second Language Writing*, vol. 13, n° 1. 49-62.

FITIKIDES, T. J. 1936. *Common Mistakes in English*. Londres : Longman.

FITZPATRICK, Eileen et SEEGMILLER, M.S. 2004. "The Montclair Electronic Language Database Project". *Applied Corpus Linguistics. A Multidimensional Perspective*. Éds. U. Connor et T. A. Upton. Amsterdam : Rodopi. 223-237.

FJELKESTAM-NILSSON, Brita. 1983. *Also and Too: A Corpus-Based Study of their Frequency and Use in Modern English*. Stockholm : Almqvist & Wiksell.

FLICK, William C. 1979. "A multiple component approach to research in second language acquisition". *The Acquisition and Use of Spanish and English as First and Second Languages*. Éd. R. Andersen. Washington D.C. : TESOL.

FOSTER, Jennifer et ANDERSEN, Oistein. 2009. "GenERRate: Generating errors for use in grammatical error detection". *Proceedings of the 4th Workshop on Innovative Use of NLP for Building Educational Applicatons, NAACL, 5 juin 2009, Boulder.*

FOWLER, Henry Watson. 1906. *The King's English*. Oxford : Clarendon Press.

GAMON, Michael, LEACOCK, Claudia, BROCKETT, Chris, DOLAN, William B., GAO, Jianfeng, BELENKO, Dmitriy et KLEMENTIEV, Alexandre. 2009. "Using statistical techniques and Web searches to correct ESL errors". *CALICO Journal*, vol. 26, n°3. 491-511.

GAMON, Michael. 2010. "Using mostly native data to correct errors in learners' writing". *Proceedings of the 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics, 1-6 juin 2010, Los Angeles.* 163-171.

GARNIER, Marie. 2011. "Explanation and corrective feedback in grammar checking systems". *Proceedings of the Explanation-Aware Computing Workshop, International Joint Conference on Artificial Intelligence, 16-17 juillet 2011*. 81-90.

GARNIER, Marie, RYKNER, Arnaud et SAINT-DIZIER, Patrick. 2009. "Correcting errors using the framework of argumentation: towards generating argumentative correction propositions from error annotation schemas". *Proceedings of the Pacific Asia Conference on Language, Information and Computation, 3-5 décembre 2009, Hong-Kong*. 140-149.

GIEGERICH, Heinz. 2004. "Compound or phrase? English noun-plus-noun constructions and the stress criterion". *English Language and Linguistics*, vol. 8, n°1. 1-24.

GLEDHILL, Chris. 2005. "Problems of adverbial placement in Learner English and the British National Corpus". *Linguistics, Language Learning and Language Teaching*. Éds. D. Allerton, C. Tschichold, J. Wieser. Basel : Schwabe. 85-104.

GRANGER, Sylviane. 2003. "Error-tagged learner corpora and CALL: a promising synergy". *CALICO Journal*, vol. 20, n°3. 465-480.

GREENBACKER, Charlie et MCCOY Kathleen F. 2008. "The ICICLE Project: An overview". Poster. *First Annual Computer Science Research Day.* Department of Computer and Information Sciences, University of Delaware.

GREENBAUM, Sidney. 1969. *Studies in English Adverbial Usage*. Londres : Longman.

GUIMIER, Claude. 1988. *Syntaxe de l'adverbe anglais*. Lille : Presses Universitaires de Lille.

HASWELL, Richard. 2006 [2005]. "Automated Text-Checkers: A Chronology and a Bibliography of Commentary". *Computers and Composition Online*, automne 2005.

HEIFT, Trude. 2004. "Corrective feedback and learner uptake in CALL". *ReCALL*, vol. 16, n° 2. 416-431.

HEIFT, Trude et SCHULZE, Mathias. 2007. *Errors and Intelligence in Computer-Assisted Language Learning: Parsers and Pedagogues*. New York, Londres : Routledge.

HENDRICKX, Iris, KOZAREVA, Zornitsa, NAKOV, Preslav, Ó SÉAGHDHA, Diarmuid, SZPAKOWICZ, Stan, VEALE, Tony. 2013. "SemEval-2013 Task 4: Free Paraphrases of Noun Compounds". *Proceedings of the *SEM Workshop on Semantic Evaluation, Atlanta, 14-15 juin 2013*. 138–143.

HERMET, Matthieu et DÉSILETS, Alain. 2009. "Using first and second language models to correct preposition errors in second language authoring". *Proceedings of the 4th Workshop on Innovative Use of NLP for Building Educational Applications, NAACL, 5 juin 2009, Boulder*. 64-72.

HIRST, Graeme. 2001. "Review of *The Longman Grammar of Spoken and Written English*". *Computational Linguistics*, vol. 7, n°1. 132-139.

HUDDLESTON, Rodney. 2002a. "Syntactic overview". Huddleston et Pullum et al. 43-70.

HUDDLESTON, Rodney. 2002b. "The verb". Huddleston et Pullum et al. 71-212.

HUDDLESTON, Rodney. 2002c. "The clause: complements". Huddleston et Pullum et al. 213-322.

HUDDLESTON, Rodney. 2002d. "Non-finite and verbless clauses". Huddleston et Pullum et al. 1171-1272.

HUDDLESTON, Rodney et PULLUM, Geoffrey K. et al. 2002. *The Cambridge Grammar of the English Language*. Cambridge : Cambridge University Press.

IZUMI, Emi, UCHIMOTO Kiyotaka et ISAHARA, Hitoshi. 2005. "Error annotation for corpus of Japanese learner English". *Proceedings of the 6th International Workshop on Linguistically Interpreted Corpora, 15 octobre 2005, Jesu*. 71-80.

JACKENDOFF, Ray S. 1972. *Semantic Interpretation in Generative Grammar*. Cambridge, Massachussetts : MIT Press.

JACOBSON, Sven. 1964. *Adverbial Positions in English*. Uppsala : AB Studentbok.

JACOBSON, Sven. 1980. "Contextual influences on adverb placement in English". *Studia Linguistica*, vol. 34, n°2. 135-140.

JAKUBÍČEK, Miloš, KILGARRIFF, Adam, KOVÁŘ, Vojtěch, RYCHLÝ, Pavel et SUCHOMEL, Vít. 2013. "The TenTen Corpus Family". *Proceedings of the 7th International Conference on Corpus Linguistics, 23-26 juillet 2013, Lancaster, RU*. Exemplaire non paginé.

JAMES, Carl. 1990. "Learner language". *Language Teaching*, n°23. 205-213.

JAMES, Carl. 1998. *Errors in Language Learning and Use: Exploring Error Analysis*. Londres : Longman.

JARVIS, Scott et PAVLENKO, Aneta. 2007. *Crosslinguistic Influence in Language and Cognition*. Londres, New York : Routledge.

JESPERSEN, Otto. 2006 [1933]. *Essentials of English Grammar*. Londres : Routledge.

JOHANSSON, Stig. 1973. "The identification and evaluation of errors in foreign languages: a functional approach". *Errata: Papers in Error Analysis*. Éd. J. Svartvik. Lund : CWK Gleerup.

KACHRU, Braj. B. 1985. "Standards, codification and sociolinguistic realism: the English language in the outer circle". *English in the World: Teaching and Learning the Languages and Literatures*. Éds. R. Quirk et H. G. Widdowson. Cambridge : Cambridge University Press. 11-30.

KACHRU, Braj. B. 1996. "The paradigms of marginality". *World Englishes*, n°15. 241-255.

KANG, Juyeon et SAINT-DIZIER, Patrick. 2014. "Towards an error correction memory to enhance technical texts authoring in LELIE". *Proceedings of the 4th Workshop on Controlled Natural Languages, 20-22 août 2014, Galway*. Non paginé.

KIES, Daniel. 2008. "Evaluating Grammar Checkers: A Comparative Ten-Year Study". *Proceedings of the 6th International Conference on Education and Information Systems, Technologies and Applications, 29 juin 2008.* Non paginé.

KULIK, James A., KULIK, Chen-Lin C. 1988. "Timing of feedback and verbal learning". *Review of Educational Research*, vol. 58, n°1. 79-97.

KRÜGER, Anja et HAMILTON, Simon. 1997. "RECALL: individual language tutoring through intelligent error diagnosis". *ReCALL*, vol. 9, n°2. 51-58.

LARREYA, Paul et RIVIÈRE, Claude. 2005. *Grammaire explicative de l'anglais*. 3ème éd. Paris : Pearson Longman.

LEACOCK, Claudia, CHODOROW, Martin, GAMON, Michael et TETREAULT, Joel. 2010. *Automated Grammatical Error Detection for Language Learners.* Seattle : Morgan Claypool.

LEE, John et SENEFF, Stephanie. 2006. "Automatic grammar correction for second-language learners". *Proceedings of the 9th International Conference on Spoken Language Processing (Interspeech), 17-21 septembre 2006, Pittsburgh*. 1978-1981.

LEE, Jooyoung et HEGELHEIMER, Volker. 2012. "A hybrid use of teacher feedback and Criterion in process writing". Communication. *EuroCALL, 22-25 août, Gothenburg*.

LEECH, Geoffrey. 2004. "A new Gray's Anatomy of English grammar". *English Language and Linguistics*, vol. 8, n° 1. 121-147.

LENNON, Paul. 1991. "Error: some problems of definition, identification and distinction". *Applied Linguistics*, vol. 12, n°2. 180-196.

LE PRIEULT, Henri. 1996. *Grammaire progressive de l'anglais*. Paris : Belin.

LEVI, Judith N. 1978. *The syntax and semantics of complex nominals*. New York : Academic Press.

L'HAIRE, Sébastien et VANDEVENTER FALTIN, Anne. 2003. "Diagnostic d'erreur dans le projet FreeText". *ALSIC*, vol. 6, n°2. 21-37.

LI, Shaofeng. 2010. "The effectiveness of corrective feedback in SLA: A meta-analysis". *Language Learning*, vol. 60, n° 2. 309-365.

LIU, Ann Li-E, WIBLE, David et TSAO, Nai-Lung. 2009. "Automated suggestions for miscollocations". *Proceedings of the 4th Workshop on Innovative Use of NLP for Building Educational Applications, NAACL, 5 juin 2009, Boulder*. 47-50.

LÜDELING, Anke, WALTER, Maik, KROYMANN, Emil et ADOLPHS, Peter. 2005. "Multi-level error annotation in learner corpora". *Proceedings of the Corpus Linguistics Conference, 14-17 juillet 2005, Birmingham*.

LYONS, John. 1968. *Introduction to Theoretical Linguistics*. Cambridge : Cambridge University Press.

LYSTER, Roy et RANTA, Leila. 1997. "Corrective feedback and learner uptake: negotiation of form in communicative classrooms". *Studies in Second Language Acquisition*, vol. 19, n° 1. 37–66.

MADNANI, Nitin, TETREAULT, Joel et CHODOROW, Martin. 2012. "Exploring Grammatical Error Correction with Not-So-Crummy Machine Translation". *Proceedings of the 7th Workshop on Innovative Use of NLP for Building Educational Applicatons, NAACL,7 juin 2012, Montréal.* 44-53.

MANN, William. C. et THOMPSON, Sandra A. 1988. "Rhetorical Structure Theory: Towards a Functional Theory of Text Organization". *Text - Interdisciplinary Journal for the Study of Discourse*, vol. 8, n°3. 243-281.

MCCAWLEY, James D. 1973. "Fodor on where the action is". *The Monist*, vol. 57, n°3. 396-407.

MCCOY, Kathleen F., PENNINGTON, Christopher A. et SURI, Linda Z. 1996. "English error correction: A syntactic user model based on principled 'mal-rule' scoring". *Proceedings of the Fifth International Conference on User Modeling, janvier 1996, Hawaï.* 59-66.

Merriam-Webster's Dictionary of Contemporary English Usage. 1994. Springfield, MA : Merriam-Webster.

METCALF, Vanessa et MEURERS, Detmar. 2006. "Towards a treatment of word order errors: When to use deep processing – and when not to". Communication. *NLP in CALL Workshop, CALICO, 16 mai 2006, Hawaï.*

MEURERS, Detmar et METCALF, Vanessa. 2006. "Towards a treatment of word order errors in Computer-Assisted Language Learning". Communication. *Large-Scale Grammar Development and Grammar Engineering Research Workshop of the Israel Science Foundation, 25-28 juin 2006, Haifa.*

MEURERS, Detmar. 2010. "On linguistically analyzing learner language". Communication., *NaTAL Workshop on NLP and CALL, 18 juin 2010, Nancy.*

MICHAUD, Lisa N., MCCOY, Kathleen F., et PENNINGTON, Christopher A. 2000. "An intelligent tutoring system for deaf learners of written English". *Proceedings of the 4th International Association for Computing Machinery Conference on Assistive Technologies, 13-15 novembre 2000, Arlington*. 92-100.

MILROY, James et MILROY, Lesley. 1985. *Authority in Language: Investigating Language Perception and Standardization*. Londres : Routledge.

MILTON, John et CHENG, Vivying S.Y. 2010. "A toolkit to assist L2 learners become independent writers". *Proceedings of the NAACL HLT 2010 Workshop on Computational Linguistics and Writing, juin 2010, Los Angeles*. 33-41.

MITTWOCH, Anita, HUDDLESTON, Rodney, et COLLINS, Peter. 2002. "The clause: adjuncts". Huddleston et Pullum et al. 663-784.

NABER, Daniel. 2003. "A rule-based style and grammar checker". Mémoire, Universität Bielefeld.

NADASDI, Terry et SINCLAIR, Stéfan. 2007. "Anything I can do, CP(U) can do better: a comparison of human and computer grammar correction for L2 writing using BonPatron.com". 20 novembre 2011. http://www.ualberta.ca/~tnadasdi/Dublin.htm.

NAGATA, Noriko. 1993. "Intelligent computer feedback for second language instruction". *The Modern Language Journal*, vol. 77, n°3. 330-339.

NAGATA, Noriko. 2002. "BANZAI : An application of Natural Language Processing to web-based language learning". *CALICO Journal*, vol. 19, n°3. 583-599.

NAGATA, Noriko et SWISHER, M. Virginia. 1995. "A Study of Consciousness-Raising by Computer: The Effect of Metalinguistic Feedback on Second Language Learning". *Foreign Language Annals*, vol. 28, n°3. 337-347.

NAKAMURA, Wataru. 1997. "A cognitive approach to English adverbs". *Linguistics*, vol. 35. 247-287.

NAKOV, Preslav Ivanov. 2007. "Using the Web as an implicit training set : application to noun compound syntax and semantics". Thèse, University of California, Berkeley.

NICHOLLS, Diane. 2003. "The Cambridge Learner Corpus – error coding and analysis for lexicography and ELT". *Proceedings of the Corpus Linguistics Conference, 28-31 mars 2003, Lancaster*. 571-582.

NØLKE, Henning. 1990. "Recherche sur les adverbes : bref aperçu historique des travaux de classification". *Langue Française*, n°88. 117-127.

ODLIN, Terence. 1989. *Language Transfer: Cross-linguistic influence in language learning*. Cambridge : Cambridge University Press.

OSBORNE, John. 2008. "Adverb placement in post-intermediate learner English: a contrastive study of learner corpora". *Linking up contrastive linguistics and interlanguage research*. Éds. S. Papp, B. Díez et G. Gilquin. Atlanta : Rodopi. 127-146.

PASTOR-GOMEZ, Iria. 2010. "Nominal modifiers in noun phrase structure: Evidence from Contemporary English". Thèse, Universidade de Santiago de Compostela.

PASTOR-GOMEZ, Iria. 2011. *The Status and Development of N+N Sequences in Contemporary English Noun Phrases*. Bern : Peter Lang.

PAYNE, John et HUDDLESTON, Rodney. 2002. "Nouns and noun phrases". Huddleston et Pullum et al. 323-524.

PRODROMOU, Luke. 2008. *English as a Lingua Franca: A Corpus-Based Analysis*. Londres : Continuum.

PULLUM, Geoffrey K. et HUDDLESTON, Rodney. 2002. "Preliminaries". Huddleston et Pullum et al. 1-42.

PULLUM, Geoffrey K. et HUDDLESTON, Rodney. 2002. "Adjectives and adverbs". Huddleston et Pullum et al. 525-596.

PULLUM, Geoffrey K. et HUDDLESTON, Rodney. 2002. "Prepositions and preposition phrases". Huddleston et Pullum et al. 597-662.

QUIRK, Randolph, GREENBAUM, Sydney, LEECH, Goeffrey et SVARTVIK, Jan. 1972. *A Grammar of Contemporary English*. Londres : Longman.

QUIRK, Randolph, GREENBAUM, Sydney, LEECH, Goeffrey et SVARTVIK, Jan. 1985. *A Comprehensive Grammar of the English Language*. Londres : Longman.

RICHARDS, Jack. C. 1974 [1971]. "A non-contrastive approach to error analysis". *English Language Teaching* vol. 25, n°3. *Error Analysis: Perspectives on Second Language Acquisition*. Éd. J. C. Richards. Londres : Longman. 172-188.

RINGBOM, Håkan. 2007. *The Importance of Cross-linguistic Similarity in Foreign Language Learning: Comprehension, Learning and Production*. Clevedon : Multilingual Matters.

RUPPENHOFER, Josef, ELLSWORTH, Michael, PETRUCK, Miriam R. L., JOHNSON, Christopher R., SCHEFFCZYK, Jan. 2010. *FrameNet II: Extended Theory and Practice*. Berkeley : International Computer Science Institute. https://framenet2.icsi.berkeley.edu/docs/r1.5/book.pdf

RUSSELL, Jane et SPADA, Nina. 2006. "The effectiveness of corrective feedback for the acquisition of L2 grammar". *Synthesizing research on language learning and teaching*. Éds. J. Norris & L. Ortega. Amsterdam : Benjamins.133–164.

SAINT-DIZIER, Patrick. 2011. "<TextCoop> : un analyseur de discours basé sur des grammaires logiques". *Actes du colloque TALN, 27 juin-1er juillet 2011, Montpellier*. Exemplaire non paginé.

SAINT-DIZIER, Patrick. 2012. "Processing natural language arguments with the <TextCoop> platform". *Argument and Computation*. vol. 3, n° 1. 49-82.

SAINT-DIZIER, Patrick. 2014. *Challenges of Discourse Processing: The Case of Technical Documents*. Cambridge : Cambridge Scholars.

SCHACHTER, Jacquelyn et CELCE-MURCIA, Marianne. 1977. "Some reservations concerning error analysis". *TESOL Quaterly*, n°11. 141-151.

SCHACHTER, Jacquelyn. 1974. "An error in error analysis". *Language Learning*, vol. 24, n°2. 205-214.

SCHMID, Hans-Jörg. 2003. "Book review: *Longman Grammar of Spoken and Written English*". *Journal of Pragmatics*, vol. 35. 1265-1269.

SCHMIDT, Richard. 1990. "The Role of Consciousness in Second Language Learning". *Applied Linguistics*, vol. 11. 129-158.

SEIDLHOFER, Barbara. 2003. "Autour du concept d'anglais international : de l''anglais authentique' à l''anglais réaliste' ?". Conseil de l'Europe.

SEIDLHOFER, Barbara. 2011. *Understanding English as a Lingua Franca*. Oxford : Oxford University Press.

SELINKER, Larry. 1969. "Language transfer". *General Linguistics*, n°9. 67-92.

SELINKER, Larry. 1974 [1972]. "Interlanguage". *International Review of Applied Linguistics*, vol. 10, n°3. 209-301. *Error Analysis: Perspectives on Second Language Acquisition*. 1974. Éd. J. C. Richards. Londres : Longman. 31-54.

STARLANDER, Marianne et POPESCU-BELIS, Andrei. 2002. "Corpus-based evaluation of a French spelling and grammar checker". *Proceedings of LREC, 29-31 mai 2002, Las Palmas.* 268-274.

SWAN, Michael et SMITH, Bernard. 2001. *Learner English: A Teachers's Guide to Interference and Other Problems*. 2ème éd. Cambridge : Cambridge University Press.

TARONE, Elaine. 2006 [1994]. "Interlanguage". *Encyclopedia of Language and Linguistics*, 2ème éd., vol. 5. Éd. K. Brown. 2006. Amsterdam : Elsevier. 747-752.

TENGI, Randee I. 1998. "Design and implementation of the WordNet Lexical Database and searching software". *WordNet: An Electronic Lexical Database*. Éd. C. Fellbaum. 1998. Cambridge : MIT Press. 105-128.

TETREAULT, Joel et CHODOROW, Martin. 2008. "Native judgments of non-native usage: experiments in preposition error detection". *Proceedings of the Workshop on Human Judgments in Computational Linguistics, COLING, 23 août 2008, Manchester*. 24-32.

TETREAULT, Joel et CHODOROW, Martin. 2008. "The ups and downs of preposition error detection in ESL writing". *Proceedings of the 22nd International Conference on Computational Linguistics (COLING), 18-22 août 2008, Manchester.* 865-872.

TETREAULT, Joel, FILATOVA, Elena et CHODOROW, Martin. 2010. "Rethinking grammatical error annotation and evaluation with the Amazon Mechanical Turk". *Proceedings of the 5th Workshop on Innovative Use of NLP for Building Educational Applicatons, NAACL, 5 juin 2010, Los Angeles.* 45-48.

THOUËSNY, Sylvie. 2011. "Modeling second language learners' interlanguage and its variablity: A computer-based dynamic assessment approach to distinguishing between errors and mistakes". Thèse, Dublin City University.

TORLAKOVIC, Edina et DEUGO, Dwight. 2004. "Application of a CALL system in the acquisition of adverbs in English". *Computer-Assisted Language Learning*, vol. 17, n° 2. 203-235.

TRAHEY, Martha. 1996. "Positive evidence in language acquisition: some long-term effects". *Second Language Research*, vol. 12, n° 2. 111-139.

TRAHEY, Martha et WHITE, Lydia. 1993. "Positive evidence and preemption in the second language classroom". *Studies in Second Language Acquisition*, vol. 15, n° 2. 181-204.

TRUSCOTT, John. 1996. "The case against grammar correction in L2 writing classes". *Language Learning*, n° de juin. 327-369.

TSCHICHOLD, Cornelia, BODMER, Franck, CORNU, Etienne, GROSJEAN, François, GROSJEAN, Lysiane, KÜBLER, Natalie, LEWY, Nicolas et TSCHUMI, Corinne. "Developing a new grammar checker for Englis as a second language". *From Research to Commercial Applications: Making Natural Language Processing Work in Practice. Proceedings of a Workshop sponsored by the Association for Computational Linguistics, 12 juillet 1997, Madrid.* Éd. Jill BURSTEIN et Claudia LEACOCK. 7-12.

TSCHICHOLD, Cornelia. 1994. "Evaluating second language grammar checkers". *TRANEL (Travaux neufchâtelois de linguistique)*, n°21. 195-204.

TSCHICHOLD, Cornelia. 1999. "Grammar checking for CALL: Strategies for improving foreign language grammar checkers". *CALL: Media, Design and Applications*. Éd. K. Cameron. Lisse : Swets et Zeitlinger. 203-222.

UCL – Learner Corpora Around the World. Centre for English Corpus Linguistics, Université Catholique de Louvain. Éd. Faculté de Philosophie, Arts et Lettres. 4 mars 2013. https://www.uclouvain.be/en-cecl-lcworld.html

VANDEVENTER, Anne. 2001. "Creating a grammar checker for CALL using constraint relaxation: a feasibility study". *ReCALL*, vol. 13, n°1. 110-120.

VANDEVENTER FALTIN, Anne. 2003. "Syntactic error diagnosis in the context of Computer Assisted Language Learning". Thèse. Université de Genève.

VAN DER LINDEN, Elisabeth. 1993. "Does feedback enhance computer-assisted language learning?". *Computers and Education*, vol. 21, n° 1-2. 61-65.

WAGNER, Joachim, FOSTER, Jennifer, et VAN GEN-ABITH, Josef. 2007. "A comparative evaluation of deep and shallow approaches to the automatic detection of common

grammatical errors". *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the Conference on Computational Natural Language Learning, 28-30 juin, Prague*. 112-121.

WARREN, Beatrice. 1978. *Semantic Patterns of Noun-Noun Compounds*. Göteborg : Acta Universitatis Götheburgensis.

WHITE, Lydia. 1989. "The adjacency position on case assignment: do learners observe the Subset Principle". *Linguistic Perspectives on Second Language Acquisition*. Éd. S. Gass et J. Schachter. Cambridge : Cambridge University Press. 134-156.

WHITE, Lydia. 1990. "The verb-movement parameter in second language acquisition". *Language Acquisition*, vol. 1, n°4. 337-360.

WHITE, Lydia. 1991. "Adverb placement in second language acquisition: some effects of positive and negative evidence in the classroom". *Second Language Research*, vol. 7, n° 2 : 133-161.

WIBLE, David, KWO, Chin-Hwa, TSAO, Nai-Lung, LIU, Anne et LIN, Hsiu-Ling. 2003. "Bootstrapping in a language learning environment". *Journal of Computer Assisted Learning*, vol. 19, n°4. 90-102.

WIDDOWSON, Henry. 2003. *Defining Issues in English Language Teaching*. Oxford : Oxford University Press.

YI, Xing, GAO, Jianfeng et DOLAN, William B. 2008. "A web-based English proofing system for English as a second language users". *Proceedings of the International Joint Conference on Natural Language Processing (IJCNLP), 7-12 janvier 2008, Hyderabad*. 619-624.

Sites internet et systèmes informatiques

Amazon Mechanical Turk. Amazon.com, Inc. 6 mars 2013. https://www.mturk.com/mturk/welcome

Anglais facile. Laurent Camus. 22 juin 2014. http://www.anglaisfacile.com/.

Correct English. Vantage Linguistics. 24 août 2013, http://www.correctenglish.com/

CyberTeachers. TeleLangue. 22 juin 2014. http://learning.telelangue.com/elearning/

Ginger. Ginger Software. 24 août 2013, http://www.gingersoftware.com/grammarcheck

Grammarly. Grammarly, Inc. 24 août 2013, http://www.grammarly.com/

Language Tool. Daniel NABER. 24 août 2013, http://www.languagetool.org/

Lélie – Risk Analysis and Prevention. Patrick Saint-Dizier. 5 juin 2014. http://www.irit.fr/recherches/ILPL/lelie/accueil.html

Proofwriter. Educational Testing Service. 24 août 2013, https://b2b.proofwriter.ets.org/proof.php

Rosetta Stone. Rosetta Stone Ltd. 22 juin 2014. http://www.rosettastone.fr/

Sketch Engine. Lexical Computing. 4 mars 2013. http://www.sketchengine.co.uk/

Spell Check Plus. Nadaclair Language Technologies. Éd. T. Nadasdi et S. Sinclair. 24 août 2013. http://spellcheckplus.com/

Word 2011. Microsoft Corporation.

Corpus et ressources lexicales

Australian Corpus of English. 1986. Macquarie University. 11 septembre 2013. http://www.ausnc.org.au/corpora/ace

Cambridge Learner Corpus. Cambridge University Press, University of Cambridge ESOL examinations. 4 Mars 2013. http://www.cambridge.org/fr/elt/catalogue/subject/custom/item3646603/Cambridge-English-Corpus-Cambridge-Learner-Corpus/?site_locale=fr_FR

DAVIES, Mark. 2004. *BYU-BNC*. (Basé sur le *British National Corpus*, Oxford University Press). 9 juillet 2013. http://corpus.byu.edu/bnc/.

DAVIES, Mark. 2008. *The Corpus of Contemporary American English: 450 million words, 1990-present*. 9 juillet 2013. http://corpus.byu.edu/coca/.

ELFA 2008. *The Corpus of English as a Lingua Franca in Academic Settings*. Dirigé par A. Mauranen. 3 mars 2013. http://www.helsinki.fi/elfa/elfacorpus.

English TenTen Web Corpus. Lexical Computing. 8 juin 2014. http://www.sketchengine.co.uk/documentation/wiki/Corpora/enTenTen

FrameNet. International Computer Science Institute. 9 juin 2014. https://framenet.icsi.berkeley.edu/fndrupal/home

FRANCIS, W. Nelson et KUCERA, Henry. 1979 [1964, 1971]. *The Brown Corpus: A Standard Corpus of Present-Day Edited American English*. Providence, Rhode Island.

GRANGER Sylviane, DAGNEAUX Estelle, MEUNIER Fanny et PAQUOT Magali. 2009. *The International Corpus of Learner English. Version 2. Handbook and CD-Rom*. Louvain-la-Neuve : Presses Universitaires de Louvain.

Dictionnaires de français. Larousse. 24 juin 2014. http://www.larousse.fr/dictionnaires/francais/faute/33042?q=faute#32960

LEECH, Geoffrey, JOHANSSON, Stig, GARSIDE, Roger, HOFLAND, Knut. 1981-1986. *The LOB Corpus, POS-tagged version*. University of Bergen.

LONGDALE. Longitudinal Database of Learner English. F. Meunier, S. Granger, D. Littré, M. Paquot. Centre for English Corpus Linguistics, Université Catholique de Louvain. 4 mars 2013. http://www.uclouvain.be/en-cecl-longdale.html.

Longman Learners' Corpus. Pearson Longman. 4 mars 2013. http://www.pearsonlongman.com/dictionaries/corpus/learners.html

Longman Spoken and Written English Corpus. Pearson Longman. 11 septembre 2013. http://www.pearsonlongman.com/dictionaries/corpus/

Moms Who Think. 19 septembre 2013. http://www.momswhothink.com/reading/list-of-adverbs.html

Moms Who Think. 19 septembre 2013. http://www.momswhothink.com/reading/list-of-adjectives.html

Paul and Bernice Noll's Window on the World. Paul Noll. 19 septembre 2013. http://www.paulnoll.com/Books/Clear-English/English-adverbs.html

Princeton University "About WordNet." *WordNet*. Princeton University. 2010. http://wordnet.princeton.edu

QUINTANS, Desi. *Paper Tiger*. 19 septembre 2013. http://www.desiquintans.com/downloads/nounlist.txt

Scientext. Agence Nationale pour la Recherche, LIDILEM, Université de Grenoble 3, LLS, Université de Savoie, Licorn, Université de Bretagne Sud. 4 mars 2013. http://scientext.msh-alpes.fr/scientext-site-en/spip.php?article1

Survey of English Usage. University College London. 12 mai 2013. http://www.ucl.ac.uk/english-usage/resources/sales.htm.

VESPA. Varieties of English for Special Purposes Database. M. Paquot, S. de Cock, S. Granger, L. Valentin. Centre for English Corpus Linguistics, Université Catholique de Louvain. 4 mars 2013. http://www.uclouvain.be/en-cecl-vespa.html.

VOICE. 2013. *The Vienna-Oxford International Corpus of English* (version 2.0 XML). Dirigé par B. Seidlhofer. Recherche : A. Breiteneder, T. Klimpfinger, S. Majewski, R. Osimk-Teasdale, M.-L. Pitzl, M. Radeka. 3 mars 2013. http://www.univie.ac.at/voice/page/index.php

Wall Street Journal corpus section. 1993. Association for Computational Linguistics / Data collection initiative. Linguistic Data Consortium, Philadelphia.

Wikipedia. "List of English prepositions". 19 septembre 2013. http://en.wikipedia.org/wiki/List_of_English_prepositions

Annexes

Annexe 1 : Échantillon d'erreurs annotées

N°	Erreur annotée
1.	***the intricate and tricky problem of *<u>the</u> feminism*** <error comprehension="**1**" grammaticality="**1**" categ="**gn-det**"> **the feminism** <correction meaning="**no**" length="**yes**" qualif="**2**" correct="**feminism**"> </correction> </error>
2.	***a <u>way decent</u>*** <error comprehension="**1**" grammaticality="**0**" categ="**gn-moda**"> **a way decent** <correction meaning="**no**" length="**no**" qualif="**2**" correct="**a decent way**"> </correction> </error>
3.	***<u>information extraction technology results</u>*** <error comprehension="**0**" grammaticality="**1**" categ="**gn-modn**"> information extraction technology results <correction meaning="**no**" length="**yes**" qualif="**1**" correct="**results from information extraction technology**"> </correction> <correction meaning="**no**" length="**yes**" qualif="**1**" correct="**results from the technology of information extraction**"> </correction> </error>
4.	***a *rise <u>of</u> nationalism*** <error comprehension="**1**" grammaticality="**1**" categ="**gn-prep**"> **rise of nationalism** <correction meaning="**no**" length="**no**" qualif="**2**" correct="**rise in nationalism**"> </correction> </error>
5.	***but also *<u>this</u> of a new political configuration*** <error comprehension="**0**" grammaticality="**0**" categ="**gn-pro**"> **this of a new political configuration** <correction meaning="**no**" length="**no**" qualif="**2**" correct="**that of a new political configuration**"> </correction> </error>
6.	***a folder containing *<u>three list</u> of files*** <error comprehension="**2**" grammaticality="**0**" categ="**gn-agr**"> **three list** <correction meaning="**no**" length="**no**" qualif="**2**" correct="**three lists**"> </correction> </error>
7.	***to ease *the <u>fulfill</u> of some enriched aircraft documentation structures*** <error comprehension="**1**" grammaticality="**0**" categ="**gn-lex**"> **the fulfill** <correction meaning="**no**" length="**no**" qualif="**2**" correct="**the fulfillment**"> </correction> </error>
8.	***<u>the Pagnol's dialogue</u>*** <error comprehension="**0**" grammaticality="**0**" categ="**gn-o**"> ***the Pagnol's dialogue** <correction meaning="**no**" length="**yes**" qualif="**2**" correct="**the dialogue by Pagnol**">

	</correction> <correction meaning="**no**" length="**yes**" qualif="**1**" correct="**Pagnol's dialogue**"> </correction> </error>
9.	***The competition coming from Japan is a lot *more stronger*** <error comprehension="**2**" grammaticality="**0**" categ="**ga-comp**"> **more stronger** <correction meaning="**no**" length="**yes**" qualif="**2**" correct=" **stronger**"> </correction> </error>
10.	***This is *typical for the 20th century*** <error comprehension="**2**" grammaticality="**2**" categ="**ga-prep**"> **typical for the 20th century** <correction meaning="**no**" length="**yes**" qualif="**2**" correct="**typical of the 20th century**"> </correction> </error>
11.	***One trait of her personality is *very much striking*** <error comprehension="**2**" grammaticality="**0**" categ="**ga-lex**"> **very much striking** <correction meaning="**no**" length="**yes**" qualif="**2**" correct="**very striking**"> </correction> </error>
12.	***Since a few years*** <error comprehension="**1**" grammaticality="**0**" categ="**gprep**"> **since a few years** <correction meaning="**no**" length="**no**" qualif="**2**" correct="**for a few years**"> </correction> </error>
13.	***To *index efficiently the soundtrack of multimedia documents*** <error comprehension="**2**" grammaticality="**1**" categ="**gv-modadv**"> **index efficiently the soundtrack of multimedia documents** <correction meaning="**no**" length="**no**" qualif="**2**" correct="**index the soundtrack of multimedia documents efficiently**"> </correction> <correction meaning="**no**" length="**no**" qualif="**2**" correct="**efficiently index the soundtrack of multimedia documents**"> </correction> </error>
14.	***we might *change a bit the statement proposed*** <error comprehension="**2**" grammaticality="**1**" categ="**gv-modo**"> **change a bit the statement proposed** <correction meaning="**no**" length="**no**" qualif="**2**" correct="**change the statement proposed a bit**"> </correction> </error>
15.	***I will *put on my website some mp3 files*** <error comprehension="**1**" grammaticality="**0**" categ="**gv-comp**"> **put on my website some mp3 files** <correction meaning="**no**" length="**no**" qualif="**2**" correct="**put some mp3 files on my website**"> </correction> </error>
16.	***I *will sent all the information required*** <error comprehension="**2**" grammaticality="**0**" categ="**gv-morph**"> **will sent** <correction meaning="**no**" length="**no**" qualif="**2**" correct="**will send**"> </correction> </error>

17.	***Have you never opened a newspaper?*** <error comprehension="**1**" grammaticality="**2**" categ="**gv-neg**"> **have you never** <correction meaning="**yes**" length="**no**" qualif="**1**" correct="**have you ever**"> </correction> </error>
18.	***this system *results of the study*** <error comprehension="**1**" grammaticality="**0**" categ="**gv-prep**"> **results of the study** <correction meaning="**no**" length="**no**" qualif="**2**" correct="**results from the study**"> </correction> <correction meaning="**yes**" length="**no**" qualif="**0**" correct="**results in the study**"> </correction> </error>
19.	***to *remind a document*** <error comprehension="**1**" grammaticality="**1**" categ="**gv-lex**"> **remind a document** <correction meaning="**no**" length="**no**" qualif="**2**" correct="**remember a document**"> </correction> </error>
20.	***he should also *feel inhabitant of the European nation*** <error comprehension="**1**" grammaticality="**0**" categ="**gv-o**"> **feel inhabitant of the European nation** <correction meaning="**yes**" length="**yes**" qualif="**2**" correct="**feel like he is an inhabitant of the European nation**"> </correction> </error>
21.	***We are going to become Europeans very soon. *What implies this?*** <error comprehension="**1**" grammaticality="**0**" categ="**p-int**"> **what implies this?** <correction meaning="**no**" length="**yes**" qualif="**2**" correct="**what does this imply?**"> </correction> </error>
22.	***This introduction roughly defines what a procedure is, *what is its structure in linguistic and conceptual terms.*** <error comprehension="**1**" grammaticality="**2**" categ="**p-subf**"> **what is its structure** <correction meaning="**no**" length="**no**" qualif="**2**" correct="**what its structure is**"> </correction> </error>
23.	***who has *imagined to divide the world into three categories of men*** <error comprehension="**1**" grammaticality="**1**" categ="**p-subn**"> **imagined to divide the world** <correction meaning="**no**" length="**yes**" qualif="**2**" correct="**imagined dividing the world**"> </correction> <correction meaning="**no**" length="**yes**" qualif="**1**" correct="**imagined the division of the world**"> </correction> </error>
24.	***the *people who enjoys the long summer*** <error comprehension="**2**" grammaticality="**0**" categ="**p-agr**"> **people who enjoys** <correction meaning="**no**" length="**no**" qualif="**2**" correct="**people who enjoy**"> </correction> </error>
25.	***Can you tell me quickly if I *receive a parcel this Thursday?*** <error comprehension="**1**" grammaticality="**2**" categ="**p-ten**"> **if I receive a parcel this Thursday** <correction meaning="**no**" length="**yes**" qualif="**2**" correct="**if I will receive a parcel this Thursday**">

	</correction> <correction meaning="**yes**" length="**no**" qualif="**0**" correct="**if I received a parcel this Thursday**"> </correction> </error>
26.	***They say that man has no opportunity to dream or imagine and that he *<u>becomes</u> a robot*** <error comprehension="**1**" grammaticality="**1**" categ="**p-asp**"> **he becomes a robot** <correction meaning="**no**" length="**yes**" qualif="**2**" correct="**he is becoming a robot**"> </correction> </error>
27.	***the idea according to which there <u>would</u> be a specific problem*** <error comprehension="**1**" grammaticality="**0**" categ="**p-aux**"> **the idea according to which there <u>would</u> be a specific problem** <correction meaning="**no**" length="**yes**" qualif="**2**" correct="**the idea according to which there is a specific problem**"> </correction> <correction meaning="**yes**" length="**no**" qualif="**1**" correct="**the idea according to which there could be a specific problem**"> </correction> </error>
28.	***Belgium is the capital of Europe and <u>she</u> is federal*** <error comprehension="**1**" grammaticality="**0**" categ="**p-pro**"> **Belgium is the capital of Europe and she is federal** <correction meaning="**no**" length="**no**" qualif="**2**" correct="**Belgium is the capital of Europe and it is federal**"> </correction> </error>
29.	***for your culture *august 15 in France, <u>it is</u> Assumption day*** <error comprehension="**1**" grammaticality="**0**" categ="**p-inf**"> **august 15 in France, <u>it is</u> Assumption day** <correction meaning="**no**" length="**yes**" qualif="**2**" correct="**august 15 in France is Assumption day**"> </correction> </error>
30.	***<u>Least</u>, but not <u>last</u>*** <error comprehension="**2**" grammaticality="**2**" categ="**p-lex**"> **Least but not last** <correction meaning="**no**" length="**no**" qualif="**2**" correct="**last but not least**"> </correction> </error>
31.	***it is already <u>of habit</u>*** <error comprehension="**0**" grammaticality="**1**" categ="**p-o**"> **it is already of habit** <correction meaning="**no**" length="**yes**" qualif="**2**" correct="**it is customary**"> </correction> </error>

Annexe 2 : Ressource lexicale pour les adverbes anglais, étude pilote

Adverbe	Type sémantique	Portée	Scalaire	Type morpho.
actually	Modalité	Proposition	Non	Adj. *-ly*
admittedly	Lié à l'acte de parole	Proposition	Non	Adj.*-ly*
again	Ordre dans une série	GV	Non	Autre (simple)
almost	Degré	GV	Non	Autre (composé)
already	Aspectuel	GV	Non	Autre (composé)
also	Focalisant additif	Focus	Non	Autre (composé)
always	1. Durée 2. Fréquence	GV GV	Non	Autre (composé)
apparently	Modalité	Proposition	Non	Adj. *-ly*
approximately	Degré	GV	Non	Adj. -ly
briefly	1. Durée 2. Manière 3. Lié à l'acte de parole	GV GV Proposition	Oui Oui Oui	Adj.*-ly*
carefully	1. Manière 2. Lié à l'action (volition.)	GV Proposition	Oui Oui	Adj. *-ly*
certainly	Modalité	Proposition	Oui	Adj.*-ly*
clearly	1. Modalité 2. Manière	Proposition GV	Oui Oui	Adj. *-ly*
cleverly	1. Manière 2. Lié à l'action (subjectif)	GV GV	Oui Oui	Adj.*-ly*
closely	Manière	GV	Oui	Adj. *-ly*
completely	Degré	GV	Non	Adj. *-ly*
daily	Fréquence	GV	Non	Nom *-ly*
deliberately	1. Lié à l'action (volition.) 2. Manière	GV GV	Non Non	Adj.*-ly*
directly	Manière	GV	Oui	Adj. *-ly*
earlier	Lieu temporel	GV	Non	Adj. *-ly* + comp.
easily	1. Manière 2. Degré 3. Modalité	GV GV Proposition	Oui Oui Oui	Adj. *-ly*
enough	Degré	GV	Non	Autre (simple)
entirely	Degré	GV	Non	Adj. *-ly*
equally	1. Degré 2. Manière	GV GV	Non Non	Adj. *-ly*
especially	1. Restrictif partiel, sensible au focus 2. Degré	Focus GV	Non Non	Autre (composé)
even	Additif, sensible au focus	Focus	Non	Autre (simple)
ever	Fréquence	GV	Non	Autre (simple)

exactly	Restrictif, sensible au focus	Focus	Non	Adj. *-ly*
fairly	1. Degré 2. Manière	Adverbe ou adj. GV	Non Oui	Adj. *-ly*
fast	Manière	GV	Non	Autre (adjectif)
finally	1. Connectif 2. Lieu temporel	Proposition GV	Non Non	Adj.*-ly*
first	1. Connectif 2. Ordre dans une série	Proposition GV	Non Non	Autre (simple)
frequently	Fréquence	GV	Oui	Adj. *-ly*
fully	Degré	GV	Oui	Adj. *-ly*
generally	1. Fréquence 2. Manière	GV GV	Non Oui	Adj. *-ly*
greatly	Degré	GV	Oui	Adj. *-ly*
happily	1. Manière 2. Evaluation	GV Proposition	Oui Non	Adj.*-ly*
hardly	Degré	GV	Non	Adj. *-ly*
heavily	1. Degré 2. Manière	GV GV	Oui Oui	Adj. *-ly*
highly	Degré Manière	GV GV	Oui Oui	Adj. *-ly*
immediately	1. Lieu temporel 2. Manière	GV GV	Non Non	Adj. *-ly*
indeed	1. Modalité 2. Degré 3. Connecteur	Proposition GV Proposition	Non Non Non	Autre (composé)
just	1. Restrictif, sensible au focus 2. Aspectuel 3. Degré	Focus GV GV	Non Non Non	Autre (simple)
largely	1. Degré 2. Manière	GV GV	Oui Oui	Adj. *-ly*
last	1. Ordre dans une série 2. Connecteur	Proposition Proposition	Non Non	Autre (simple)
later	Lieu temporel	GV	Non	Adj. comparatif
legally	Moyen ou instrument Domaine	GV Proposition	Non Non	Adj.*-ly*
maybe	Modalité	Proposition	Non	Autre (composé)
maybe	Modalité	Proposition	Non	Autre (composé)
merely	Restrictif, sensible au focus	Focus	Non	Adj. *-ly*
moreover	Connecteur	Proposition	Non	Autre (composé)
naturally	1. Evaluation 2. Moyen ou instrument 3. Manière	Proposition GV GV	Non Non Oui	Adj. *-ly*
nearly	1. Degré 2. Manière	GV GV	Oui Oui	Adj. *-ly*
never	Fréquence	GV	Non	Autre (simple)
nevertheless	Connecteur Concession	Proposition GV	Non Non	Autre (composé)

next	1. Ordre dans une série 2. Connecteur	GV Proposition	Non Non	Autre (simple)
obviously	Modalité	Proposition	Non	Adj. *-ly*
often	Fréquence	GV	Oui	Autre (simple)
once	1. Fréquence 2. Lieu temporel	GV GV	Non Non	Autre (simple)
only	Restrictif, sensible au focus	Focus	Non	Autre *-ly*
otherwise	1. Condition 2. Manière 3. Connecteur	GV GV Proposition	Non Non Non	Autre (composé)
particularly	1. Degré 2. Restrictif partiel, sensible au focus	GV Focus	Non Oui	Adj. *-ly*
perhaps	Modalité	Proposition	Non	Autre (composé)
personally	1. Lié à l'acte de parole 2. Manière	Proposition GV	Non Oui	Adj.*-ly*
possibly	Modalité	Proposition	Non	Adj. *-ly*
previously	Lieu temporel	GV	Non	Adj. *-ly*
primarily	Restrictif partiel, sensible au focus	Focus	Non	Adj. *-ly*
probably	Modalité	Proposition	Non	Adj. *-ly*
quickly	1. Manière 2. Lieu temporel	GV GV	Oui Oui	Adj. *-ly*
quite	Degré	Adverbe ou adj.	Non	Autre (simple)
rapidly	Manière	GV	Oui	Adj. *-ly*
rather	1. Connecteur 2. Degré	Proposition GV	Non Non	Autre (simple)
really	1. Evaluation 2. Degré	Proposition GV	Non	Adj.*-ly*
recently	Lieu temporel	GV	Oui	Adj. *-ly*
relatively	1. Degré 2. Manière	GV GV	Non Non	Adj. *-ly*
sadly	1. Evaluation 2. Manière	Proposition GV	Oui	Adj.*-ly*
simply	1. Restrictif, sensible au focus 2. Degré 3. Manière	Focus Adverbe ou adj. GV	Non Non Oui	Adj. *-ly*
slightly	1. Degré 2. Manière	GV GV	Oui Oui	Adj. *-ly*
slowly	Manière	GV	Oui	Adj.*-ly*
somehow	Manière	GV	Non	Autre (composé)
sometimes	Fréquence	GV	Non	Autre (composé)
somewhat	Degré	GV	Non	Autre (composé)
soon	Lieu temporel	GV	Oui	Autre (simple)
suddenly	Manière	GV	Oui	Adj. *-ly*
then	1. Connectif 2. Lieu temporel	Proposition GV	Non	Autre (simple)

theoretically	Domaine Moyen ou instrument	Proposition GV	Non Non	Adj.-*ly*
therefore	Connecteur	Proposition	Non	Autre (composé)
thus	1. Connecteur 2. Manière	Proposition GV	Non Non	Autre (simple)
truly	1. Modalité 2. Degré	Proposition GV	Oui Oui	Adj. -*ly*
twice	Fréquence	GV	Non	Autre (simple)
unfortunately	Evaluation	Proposition	Non	Adj.-*ly*
usually	Fréquence	GV	Non	Adj. -*ly*
very	Degré	Adverbe ou adj.	Non	Autre (simple)
well	1. Manière 2. Degré 3. Modalité	GV GV Proposition	Oui Oui Oui	Autre (simple)

Annexe 3 : Liste et catégorisation des erreurs liées au placement des adverbes

Type sém.	Segment	Placement	Correction
Manière	**to remind collectively a genealogy*	SVAO	Mvt vers E Mvt vers M
	**we have tested separately all the parameters*	SVAO	Mvt vers E
	**to index efficiently the soundtrack of multimedia documents*	SVAO	Mvt vers M Mvt vers E
	**our system is able to derive automatically information for a large number of verbs*	SVAO	Mvt vers E Mvt vers M
	**perhaps he cannot show very well the way he feels*	SVAO	Autre Mvt vers E
	**favorizing then perhaps more easily the student's future professional insertion*	SVAO	Autre Mvt vers E
Degré/Manière	**[they] were found to improve substantially the performance of either modality*	SVAO	Mvt vers M
	**I understand very well this point of view*	SVAO	Mvt vers E
	**"Europe 92" won't change completely the life of its citizens*	SVAO	Mvt vers E Mvt vers M
	**his father resembles strongly his own character*	SVAO	Mvt vers M Mvt vers E
Lieu temporel	**it had then its own evolution*	SVAO	Mvt vers E
	**it has now the status of a sense*	SVAO	Mvt vers M
	**the prototype is now evolving to become shortly a software component implemented in Java*	SVAO	Mvt vers M
Lieu	**the educational range […] seems to have here to play its role*	Autre	Mvt vers E
	**we have here an advice and a warning*	SVAO	Mvt vers E
Connecteur	*?but exhibit nevertheless the dependency relationships observed in the source parse tree*	SVAO	Mvt vers M
	?everything is possible but anyway will happen so quickly	Autre	Mvt vers E Autre
	**favorizing then perhaps more easily the student's future professional insertion*	Autre	Autre
	?the similarity between words depends then on the amount of normalized contexts they share	Autre	Mvt vers M
Modalité	**the input documents can be a priori any type of Web page*	SVAO	Autre Mvt vers M Mvt vers E
Focalisation additive	**the treatment of this official day exemplifies also an answer to associations*	SVAO	Mvt vers M
	**that has also its effect on the atmosphere and relationship between people*	SVAO	Mvt vers M
	**Europe has also weak points*	SVAO	Mvt vers M
	**it means also a cultural richness conveyed by these dialects*	SVAO	Mvt vers M
	**the Community forms also a mosaic of ethnic cultures*	SVAO	Mvt vers M
	**I have also a bad news.*	SVAO	Mvt vers M
	**LFG got also several implementations*	SVAO	Mvt vers M

	**considering <u>also</u> specifications taking into account by smart assistants*	SVAO	Mvt vers M
	?it can be <u>also</u> considered as restricted in terms of variants	Autre	Mvt vers M
	?the value of N needs <u>also</u> to be experimentally elaborated	Autre	Mvt vers M
	?the European community wants <u>also</u> to be taken seriously	Autre	Mvt vers M
	?it has <u>also</u> to provide political answers	Autre	Mvt vers M
Focalisation restrictive	*?in order to hang down <u>exclusively</u> family memories*	SVAO	Mvt vers E
Comparatif	**the latent harmony will <u>more and more</u> increase*	Autre	Mvt vers E Autre
	**international business will <u>more and more</u> play a major role in the world*	Autre	Autre

Annexe 4 : Tests de jugement de grammaticalité

Tests on English adverb placement

Guidelines:

- You do not need any specific linguistic skill or knowledge other than your competence as a native speaker of English in order to complete these tests.
- The tests are made up of sequences or four or five sentences which are identical except for the placement of the adverb. For each sentence, you should indicate (by ticking the appropriate box) whether you judge that:
    - the sentence is correct in every way (grammar/naturalness/meaning),
    - the placement of the adverb in this sentence is grammatically correct but unnatural, and/or yields a very different meaning from that of the other sentences (i.e. not slight variations in focus),
    - the placement of the adverb in this sentence makes it ungrammatical (and, incidentally, it is unnatural).
- In the fourth column, you are asked to indicate which of the four or five possibilities is your "best choice", i.e. the most idiomatic, the least marked, or the one you would most readily use.
- Moreover, you might find sentences in which the placement of the adverb is only correct or natural if the word is given specific emphasis, in which case you should simply underline the adverb to indicate that it receives special stress.
- As you will notice, the adverbs in these tests are always integrated into the sentence, i.e. they are not detached by the use of commas or pauses in speech. You should always bear this in mind when evaluating the grammaticality of a sentence, as the integration/detachment of the adverb may have an effect on the acceptability and meaning of the sentence.
- The sentences which make up these tests have been designed specifically for that purpose and are therefore artificial and often repetitive. We hope you will be able to look past these aesthetic defects.
- We suggest that you complete the tests in several sittings, in order to avoid "adverb fatigue" and ensure the accuracy of the results. Also, do not spend too much time on each group of sentences: your evaluations should reflect what, as a native speaker, you spontaneously judge to be correct.
- A blank sheet is provided at the end of the tests, in case you have any remarks or additional information to give us. You are very welcome to write directly on the test next to the sequences if you feel that you need to add something important.

	Sentences	**OK**	**Incorrect**		**Best choice**
			gram. but unnatural	*ungram., etc..*	

a.

1.	Slowly she had opened the door.				
2.	She slowly had opened the door.				
3.	She had slowly opened the door.				
4.	She had opened the door slowly.				

b.

1.	Erratically she would tell her story.				
2.	She erratically would tell her story.				
3.	She would erratically tell her story.				
4.	She would tell her story erratically.				

c.

1.	Very slowly she has opened the door.				
2.	She very slowly has opened the door.				
3.	She has very slowly opened the door.				
4.	She has opened the door very slowly.				

d.

1.	Very erratically she would tell her story.				
2.	She very erratically would tell her story.				
3.	She would very erratically tell her story.				
4.	She would tell her story very erratically.				

e.

1.	Slowly she was eating.				
2.	She slowly was eating.				
3.	She was slowly eating.				
4.	She was eating slowly.				

f.

1.	Slowly she had opened it.				
2.	She slowly had opened it.				
3.	She had slowly opened it.				
4.	She had opened it slowly.				

g.

1.	Slowly she had opened the door to the second guestroom.				
2.	She slowly had opened the door to the second guestroom.				

3.	She had slowly opened the door to the second guestroom.				
4.	She had opened slowly the door to the second guestroom.				
5.	She had opened the door to the second guestroom slowly.				

h.

1.	Slowly he closed the box.				
2.	He slowly closed the box.				
3.	He closed the box slowly.				

i.

1.	Very slowly he closed the box.				
2.	He very slowly closed the box.				
3.	He closed the box very slowly.				

j.

1.	Slowly she was losing her mind.				
2.	She slowly was losing her mind.				
3.	She was slowly losing her mind.				
4.	She was losing her mind slowly.				

Annexe 5 : Liste et catégorisation des erreurs liées aux structures N+N

Catégorie	Segment	Proposition de correction	Type de correction
Relation sémantique	*?the integration tradition*	the tradition of implementing integration	Réorg. + Insertion V-*ing*
	?an oblivion policy	a policy in favor of oblivion	Réorg. + Insertion Prep (Det)
	?the width peak method	the method using width peak	Réorg. + Insertion V-*ing*
	?insignificant size segments	segments of an insignificant size	Réorg. + Insertion Prep (Det)
	?this meaning transposition	this transposition of meaning	Réorg. + Insertion Prep (Det)
	?annotations reliability computing	computing of annotations reliability	Réorg. + Insertion Prep (Det)
	?data models discrepancies	discrepancies in data models	Réorg. + Insertion Prep (Det)
	?addition goal	goal for addition	Réorg. + Insertion Prep (Det)
	?evaluation consideration	1. the consideration of evaluation 2. considerations concerning evaluation	Réorg. + Insertion Prep (Det) Réorg. + Insertion V-*ing*
	?the paper reader	the reader of the paper	Réorg. + Insertion Prep (Det)
	?the cooking recipe extract #	the extract from the cooking recipe	Réorg. + Insertion Prep (Det)
	?the society domain #	1. the domain of society 2. the societal domain 3. the social domain	Réorg. + Insertion Prep (Det) N1 vers adj. N1 vers adj.
	?admiration cries	cries of admiration	Réorg. + Insertion Prep (Det)
	?texts candidates	candidate texts	Réorganisation
	?the darkness element	the element of darkness	Réorg. + Insertion Prep (Det)
	?money inequality #	1. inequality in money 2. financial inequality	Réorg. + Insertion Prep (Det) N1 vers adj.
	?television information	information from the television	Réorg. + Insertion Prep (Det)
	?the length excess	the excess length	Réorganisation
	?partner equivalence table	table for partner equivalence	Réorg. + Insertion Prep (Det)
	?the architecture study process	the process of architecture study	Réorg. + Insertion Prep (Det)

	?different access administration concept	different concept for access administration	Réorg. + Insertion Prep (Det)
	?the memorial laws effect	the effect of memorial laws	Réorg. + Insertion Prep (Det)
	?the original French overseas population motivation #	the original motivation of French overseas population	Réorg. + Insertion Prep (Det)
Empilement	*?heterogeneous information sources cooperation*	cooperation of heterogeneous information sources	Réorg. + Insertion Prep (Det)
	?information system security strategies heterogeneity	1. heterogeneity of information system security strategies 2. heterogeneity in strategies for information system security	Réorg. + Insertion Prep (Det) Réorg. + Insertion Prep (Det)
	?security concept model granularity	the granularity of security concept models	Réorg. + Insertion Prep (Det)
	?speech music classification tool	tool for speech and music classification	Réorg. + Insertion Prep (Det)
	?local data entities structure (2 occ.)	1. the structure of local data entities 2. the local structure of data entities	Réorg. + Insertion Prep (Det) Réorg. + Insertion Prep (Det)
	?information extraction technology results	results of information extraction technology	Réorg. + Insertion Prep (Det)
	?the original French overseas population motivation #	the original motivation of French overseas population	Réorg. + Insertion Prep (Det)
	?the semi-structured procedural text analysis challenge	the challenge of semi-structured procedural text analysis	Réorg. + Insertion Prep (Det)
	?the language knowledge acquisition bottlenecks	the bottlenecks affecting language knowledge acquisition	Réorg. + Insertion V-*ing*
N_1 défini	*?the meaning utterance*	the meaning of the utterance	Insertion Prep (Det)
	?the concept meaning	the meaning of the concept	Réorg. + Insertion Prep (Det)
	?the State official discourse	1. the official discourse of the State 2. the State's official discourse	Réorg. + Insertion Prep (Det) N1 N2 vers N1's N2
	?Maurice Barrès principles	1. the principles of Maurice Barrès 2. Maurice Barrès's principles	Réorg. + Insertion Prep (Det) N1 N2 vers N1's N2
	?the domain specificities	the specificities of the domain	Réorg. + Insertion Prep (Det)
	?the cooking recipe extract #	the extract from the cooking recipe	Réorg. + Insertion Prep (Det)
	?the indication context #	1. the context of the indication	Réorg. + Insertion Prep (Det)

		2. the indicative context	N1 vers adj.
N_1 génitif	*?the ghettos sickness*	1. the ghettos' sickness 2. the ghetto's sickness 3. the sickness of the ghettos 4. the sickness of the ghetto	Ns vers Ns'/N's Ns vers Ns'/N's Réorg. + Insertion Prep (Det) Réorg. + Insertion Prep (Det)
	?the objects properties	1. the objects' properties 2. the object's properties 3. the properties of the objects 4. the properties of the object	Ns vers Ns'/N's Ns vers Ns'/N's Réorg. + Insertion Prep (Det) Réorg. + Insertion Prep (Det)
	?the annotations reliability	1. the annotations' reliability 2. the annotation's reliability 3. the reliability of the annotations 4. the reliability of the annotation	Ns vers Ns'/N's Ns vers Ns'/N's Réorg. + Insertion Prep (Det) Réorg. + Insertion Prep (Det)
	?the semantic reasoners maturity	1. the semantic reasoners' maturity 2. the semantic reasoner's maturity 3. the maturity of the semantic reasoners 4. the maturity of the semantic reasoner	Ns vers Ns'/N's Ns vers Ns'/N's Réorg. + Insertion Prep (Det) Réorg. + Insertion Prep (Det)
	?their contents reliability	1. their contents' reliability 2. their content's reliability 3. the reliability of their contents 4. the reliability of their content	Ns vers Ns'/N's Ns vers Ns'/N's Réorg. + Insertion Prep (Det) Réorg. + Insertion Prep (Det)
Lexique	*?our friend relationship*	1. our friendly relationship 2. our friendship	N1 vers adj. N+N vers N
	?the society domain #	1. the domain of society 2. the societal domain 3. the social domain	Réorg. + Insertion Prep (Det) N1 vers adj. N1 vers adj.
	?money inequality #	1. inequality in money 2. financial inequality	Réorg. + Insertion Prep (Det) N1 vers adj.
	?the indication context #	1. the context of the indication 2. the indicative context	Réorg. + Insertion Prep (Det) N1 vers adj.
	?the brothers slaves	1. the slave brothers 2. the fellow slaves	Réorganisation N1 vers adj.

Annexe 6 : Liste des erreurs utilisées pour l'évaluation des correcteurs grammaticaux automatiques

Catégorie		**Segments erronés**
Groupe Nominal		
Détermination		[It] may bring a valuable information they have to switch [] television off Total equality between the human beings nevertheless belongs to the realm of utopia.
Modification	**Adjectif**	We can continue to operate our patients in a way decent I have excelling relations
	Nom	*this meaning transposition is the major point addressed in this paper* *[It was] accompanied by admiration cries* *This explains a little bit the length excess* *information system security strategies heterogeneity* *the language knowledge acquisition bottlenecks [] could be overcome in the future* *Most of the concept meaning is not altered* *the taking into account of the domain specificities* *the ghettos sickness* *They are often based on the objects properties* *we can pursue our friend relationship as before*
Utilisation des prepositions		I think all the participants to this task are currently very busy your request of help from the French embassy
Utilisation des pronoms		they play at soldiers, at killing each others
Accord		I hope you'll like these file
Lexique		we owe them to progresses in technology you will transmit my apologize to her
Groupe Adjectival		
Comparatif et superlatif		the more and more quick ways of transport and communication
Choix des prépositions		it is by definition dependent from a given corpus
Autres		one trait of her personality is very much striking
Groupe Prépositionnel		
Présence des prépositions		Since a few years now people in all parts of the European continent
Groupe Verbal		
Modifieurs	**Adverbes**	*[they] give the opportunity to remind collectively a genealogy* *To index efficiently the soundtrack of multimedia documents, it is necessary to extract [segments]* *perhaps he cannot show very well the way he feels* *his father resembles strongly his own character* *our system is able to derive automatically information for a large number of verbs* *Europe has also weak points* *the Community forms also a mosaic of ethnic cultures* *LFG got also several implementations* *that has also its effect on the atmosphere and relationship between people*

		the value of N needs also to be experimentally elaborate
	Autres	I still play a lot the trombone the word "birth" in the title is according to me ironical
Compléments		It will wake up me.
Morphologie du GV		I can lost a love a few applied works shown that it is possible to simplify the problem Systems are robust and could be improve on every corpus
Utilisation de la négation		You do not have forgotten us
Utilisation des prépositions		this system results of the study of two detection subsystems this tendency can come at the surface please say [] him "Bonjour" from me
Lexique		Why should we loss our identity? we precise the two different feature spaces I am in a hurry to be there
Autres		It allows to implement inner links at the local level
Proposition et phrase		
Phrases interrogatives		What implies this?
Prop. subord. verbe fini		Let's first have a look at what is Europe actually. When appeared the Question Noire, there were many claims about the history of slavery James Baldwin, which characteristic is to be an American writer in France
Prop. subord. verbe non fini		It has the advantage to present long periods of speech Thanks for your kindness to read my bad English
Sujet-verbe		it is labour-intensive and require to have a practical application available they see the world in a two-valued orientation, just like television do I'm returning you the old system which have never functioned
Choix du temps		she proposed and I accept
Choix de l'aspect		Mr Wimbush butts in again, forces him to listen to what he says I'm writing this mail for one hour now
Choix de l'auxiliaire modal		patients who suffer from FRS would be unable to correctly monitor their actions the transfer could be quickly done, as soon as we will get those numbers.
Planif. de l'information		Remote territories, out of the metropolitan borders, it is firstly with overseas departments that France discovers its identity for your culture august 15 in France, it is Assumption day
Cohérence des pronoms		Belgium is the capital of Europe and she is federal
Lexique		Although, such an excess of mental effort should be reduced at all costs
Autres		Though it has been many years that the first rocket has been launched into space

Annexe 7 : Liste des règles de détection et de correction dans <TextCoop>

Type de règle	Forme
Also	**4. Aux + Aux + ALSO + Vlex => Aux + ALSO + Aux + Vlex**
	forme(corr-adva, E, S, [aux(AUX1,_,E,E1), aux(AUX2,_,E1,E2), adv([also],_,E2,E3), verb(V,_,E3,S)], [conc(VE1,S,E3)], ['<erreur4>', AUX1, AUX2, [also], VE1, '</erreur>', '<correct>', AUX1, [also], AUX2, VE1, '</correct>']).
	3b. Aux + Aux + Vlex + ALSO + THAT + GN => Aux + ALSO + Aux + Vlex + THAT + GN
	forme(corr-adva, E, S, [aux(AUX1,_,E,E1), aux(AUX2,_,E1,E2), verb(V,_,E2,E3), adv([also],_,E3,E4), conj([that],_,E4,E5), np(NP,_,E5,S)], [conc(VE1,E3,E2), conc(NP11,S,E5), not(VE1 = [be]), not(VE1 = [am]), not(VE1 = [are]), not(VE1 = [is]), not(VE1 = [was]), not(VE1 = [were]), not(VE1 = [isn,simplequote,t]), not(VE1 = [aren,simplequote,t]), not(VE1 = [wasn,simplequote,t]), not(VE1 = [weren,simplequote,t]), not(VE1 = [been]), not(VE1 = [being]), not(VE1 = [see])], ['<erreur3b>', AUX1, AUX2, VE1, [also], [that], NP11, '</erreur>', '<correct>', AUX1, [also], AUX2, VE1, [that], NP11, '</correct>']).
	3a. Vlex + ALSO + THAT + GN => ALSO + Vlex + THAT + GN
	forme(corr-adva, E, S, [verb(V,_,E,E1), adv([also],_,E1,E2), conj([that],_,E2,E3), np(NP,_,E3,S)], [conc(VE1,E1,E), conc(NP11,S,E3), not(VE1 = [be]), not(VE1 = [am]), not(VE1 = [are]), not(VE1 = [is]), not(VE1 = [was]), not(VE1 = [were]), not(VE1 = [isn,simplequote,t]), not(VE1 = [aren,simplequote,t]), not(VE1 = [wasn,simplequote,t]), not(VE1 = [weren,simplequote,t]), not(VE1 = [been]), not(VE1 = [being]), not(VE1 = [see])], ['<erreur3a>', VE1, [also], [that], NP11, '</erreur>', '<correct>', [also], VE1, [that], NP11, '</correct>']).
	2b. Aux + Aux + Vlex + ALSO + TO + Vlex => Aux + ALSO + Aux + Vlex + TO + Vlex
	forme(corr-adva, E, S, [aux(AUX1,_,E,E1), aux(AUX2,_,E1,E2), verb(V1,_,E2,E3), adv([also],_,E3,E4), prep([to],_,E4,E5), verb(V2,_,E5,S)], [conc(VE1,E3,E2), conc(VE2,S,E5), not(VE1 = [be]), not(VE1 = [am]), not(VE1 = [are]), not(VE1 = [is]), not(VE1 = [was]), not(VE1 = [were]), not(VE1 = [isn,simplequote,t]), not(VE1 = [aren,simplequote,t]), not(VE1 = [wasn,simplequote,t]), not(VE1 = [weren,simplequote,t]), not(VE1 = [been]), not(VE1 = [being])], ['<erreur2b>', AUX1, AUX2, VE1, [also], [to], VE2, '</erreur>', '<correct>', AUX1, [also], AUX2, VE1, [to], VE2, '</correct>']).
	2a. Vlex + ALSO + TO + Vlex => ALSO + Vlex + TO + Vlex
	forme(corr-adva, E, S, [verb(V1,_,E,E1), adv([also],_,E1,E2), prep([to],_,E2,E3), verb(V2,_,E3,S)], [conc(VE1,E1,E), conc(VE2,S,E3), not(VE1 = [be]), not(VE1 = [am]), not(VE1 = [are]), not(VE1 = [is]), not(VE1 = [was]), not(VE1 =

	[were]), not(VE1 = [isn,simplequote,t]), not(VE1 = [aren,simplequote,t]), not(VE1 = [wasn,simplequote,t]), not(VE1 = [weren,simplequote,t]), not(VE1 = [been]), not(VE1 = [being])], ['<erreur2a>', VE1, [also], [to], VE2, '</erreur>', '<correct>', [also], VE1, [to], VE2, '</correct>']).
	1b. Aux + Aux + Vlex + ALSO + (Prep) + GN => Aux + ALSO + Aux + Vlex + (Prep) + GN
	forme(corr-adva, E, S, [aux(AUX1,_,E,E1), aux(AUX2,_,E1,E2), verb(V,_,E2,E3), adv([also],_,E3,E4), opt(prep(P,_,E4,E5)), np(NP,_,E5,S)], [conc(VE1,E3,E2), conc(NP11,S,E5), not(VE1 = [be]), not(VE1 = [am]), not(VE1 = [are]), not(VE1 = [is]), not(VE1 = [was]), not(VE1 = [were]), not(VE1 = [isn,simplequote,t]), not(VE1 = [aren,simplequote,t]), not(VE1 = [wasn,simplequote,t]), not(VE1 = [weren,simplequote,t]), not(VE1 = [been]), not(VE1 = [being]), not(VE1 = [see])], ['<erreur1b>', AUX1, AUX2, VE1, [also], P, NP11, '</erreur>', '<correct>', AUX1, [also], AUX2, VE1, P, NP11, '</correct>']).
	1a. Vlex + ALSO + (Prep) + GN => ALSO + Vlex + (Prep) + GN
	forme(corr-adva, E, S, [verb(V,_,E,E1), adv([also],_,E1,E2), opt(prep(P,_,E2,E3)), np(NP,_,E3,S)], [conc(VE1,E1,E), conc(NP11,S,E3), not(VE1 = [be]), not(VE1 = [am]), not(VE1 = [are]), not(VE1 = [is]), not(VE1 = [was]), not(VE1 = [were]), not(VE1 = [isn,simplequote,t]), not(VE1 = [aren,simplequote,t]), not(VE1 = [wasn,simplequote,t]), not(VE1 = [weren,simplequote,t]), not(VE1 = [been]), not(VE1 = [being]), not(VE1 = [see])], ['<erreur1a>', VE1, [also], P, NP11, '</erreur>', '<correct>', [also], VE1, P, NP11, '</correct>']).
Adv. manière	**6. Vlex + (Prep) + (Very) + AdvM + GN long => (Very) + AdvM + Vlex + (Prep) + GN long**
	forme(corr-advd, E, S, [verb(V,_,E,E1), opt(prep(P,_,E1,E2)), opt(adv([very],_,E2,E3)), adv(ADV,manner,E3,E4), np(NP,_,E4,S)], [conc(VE1,E1,E), conc(NP11,S,E4), card(NP11, T), (T > 4), not(VE1 = [be]), not(VE1 = [am]), not(VE1 = [are]), not(VE1 = [is]), not(VE1 = [was]), not(VE1 = [were]), not(VE1 = [isn,simplequote,t]), not(VE1 = [aren,simplequote,t]), not(VE1 = [wasn,simplequote,t]), not(VE1 = [weren,simplequote,t]), not(VE1 = [been]), not(VE1 = [being]), not(ADV = [fast])], ['<erreur6>', VE1, P, [very], ADV, NP11, '</erreur>', '<correct>', [very], ADV, VE1, P, NP11, '</correct>']).
	7. Vlex + (Prep) + (Very) + WELL + GN court => Vlex + (Prep) + GN court + (Very) + WELL
	forme(corr-adva, E, S, [verb(V,_,E,E1), opt(prep(P,_,E1,E2)), opt(adv([very],_,E2,E3)), adv([well],well,E3,E4), np(NP,_,E4,S)], [conc(VE1,E1,E), conc(NP11,S,E4), not(VE1 = [be]), not(VE1 = [am]), not(VE1 = [are]), not(VE1 = [is]), not(VE1 = [was]), not(VE1 = [were]), not(VE1 = [isn,simplequote,t]), not(VE1 = [aren,simplequote,t]), not(VE1 = [wasn,simplequote,t]), not(VE1 = [weren,simplequote,t]), not(VE1 = [been]), not(VE1 = [being]),], ['<erreur7>', VE1, P, [very], [well], NP11, '</erreur>', '<correct>', VE1, P, NP11, [very], [well], '</correct>']).
	5b. Vlex + (Prep) + Very + AdvM + GN court => Vlex + (Prep) + GN court + Very + AdvM

	forme(corr-advd, E, S, [verb(V,_,E,E1), opt(prep(P,_,E1,E2)), adv([very],_,E2,E3), adv(ADV,manner,E3,E4), np(NP,_,E4,S)], [conc(VE1,E1,E), conc(NP11,S,E4), card(NP11, T), (T < 5), not(VE1 = [be]), not(VE1 = [am]), not(VE1 = [are]), not(VE1 = [is]), not(VE1 = [was]), not(VE1 = [were]), not(VE1 = [isn,simplequote,t]), not(VE1 = [aren,simplequote,t]), not(VE1 = [wasn,simplequote,t]), not(VE1 = [weren,simplequote,t]), not(VE1 = [been]), not(VE1 = [being]), not(ADV = [fast])], ['<erreur5b>', VE1, P, [very], ADV, NP11, '</erreur>', '<correct>', VE1, P, NP11, [very], ADV, '</correct>']).
	5a. Vlex + (Prep) + AdvM + GN court => 1. AdvM + Vlex + (Prep) + GN court **2. Vlex + (Prep) + GN court + AdvM**
	forme(corr-adva, E, S, [verb(V,_,E,E1), opt(prep(P,_,E1,E2)), adv(ADV,manner,E2,E3), np(NP,_,E3,S)], [conc(VE1,E1,E), conc(NP11,S,E3), card(NP11, T), (T < 5), not(VE1 = [be]), not(VE1 = [am]), not(VE1 = [are]), not(VE1 = [is]), not(VE1 = [was]), not(VE1 = [were]), not(VE1 = [isn,simplequote,t]), not(VE1 = [aren,simplequote,t]), not(VE1 = [wasn,simplequote,t]), not(VE1 = [weren,simplequote,t]), not(VE1 = [been]), not(VE1 = [being]), not(ADV = [fast])], ['<erreur5a>', VE1, P, ADV, NP11, '</erreur>', '<correct1>', ADV, VE1, P, NP11, '</correct1>', '<correct2>', VE1, P, NP11, ADV, '</correct2>']).
N+N : Empilement	**11. (Det) + Adj + Adj + N + N + N**
	forme(corr-nna, E, S, [opt(det(DET,_,E,E1)), adj(ADJ1,_,E1,E2), adj(ADJ2,_,E2,E3), nconj(NOM1,_,E3,E4), nconj(NOM2,_,E4,E5), nconj(NOM3,_,E5,S)], [conc(N1,E4,E3), conc(N2,E5,E4), conc(N3,S,E5)], ['<erreur11>', DET, ADJ1, ADJ2, N1, N2, N3, '</erreur>', '<correct>', '</correct>']).
	10. (Det) + Adj + N + N + N
	forme(corr-nna, E, S, [opt(det(DET,_,E,E1)), adj(ADJ,_,E1,E2), nconj(NOM1,_,E2,E3), nconj(NOM2,_,E3,E4), nconj(NOM3,_,E4,S)], [conc(N1,E3,E2), conc(N2,E4,E3), conc(N3,S,E4)], ['<erreur10>', DET, ADJ, N1, N2, N3, '</erreur>', '<correct>', '</correct>']).
	9. (Det) + N + N + N + N + N
	forme(corr-nna, E, S, [opt(det(DET,_,E,E1)), nconj(NOM1,_,E1,E2), nconj(NOM2,_,E2,E3), nconj(NOM3,_,E3,E4), nconj(NOM4,_,E4,E5), nconj(NOM5,_,E5,S)], [conc(N1,E2,E1), conc(N2,E3,E2), conc(N3,E4,E3), conc(N4,E5,E4), conc(N5,S,E5)], ['<erreur9>', DET, N1, N2, N3, N4, N5, '</erreur>', '<correct>', '</correct>']).
	8. (Det) + N + N + N + N
	forme(corr-nna, E, S, [opt(det(DET,_,E,E1)), nconj(NOM1,_,E1,E2), nconj(NOM2,_,E2,E3), nconj(NOM3,_,E3,E4), nconj(NOM4,_,E4,S)],

	[conc(N1,E2,E1), conc(N2,E3,E2), conc(N3,E4,E3), conc(N4,S,E4)], ['<erreur8>', DET, N1, N2, N3, N4, '</erreur>', '<correct>', '</correct>']).
N+N : Génitif	**12. Det défini + (Adj) + N_1s + N_2 => 1a. Det défini + (Adj) + N_1s + ['] + N_2** **1b. Det défini + (Adj) + N_1 + ['s] + N_2** **2a. THE + N_2 + OF + Dét défini + (Adj) + N_1s** **2a. THE + N_2 + OF + Dét défini + (Adj) + N_1**
	forme(corr-nna, E, S, [det(DET,def,E,E1), opt(adj(ADJ,_,E1,E2)), nconj(NOM1,plu,E2,E3), nconj(NOM2,_,E3,S)], [conc(N1,E3,E2), conc(N2,S,E3)], ['<erreur12a>', DET, ADJ, N1, N2, '</erreur>', '<correct1a>', DET, ADJ, N1, [simplequote], N2, '</correct1a>', '<correct1b>', DET, ADJ, NOM1, [simplequote], [s], N2, '</correct1b>', <correct2a>', [the], N2, [of], DET, ADJ, N1, '</correct2a>', '<correct2b>', [the], N2, [of], DET, ADJ, NOM1, '</correct2b>']).

Annexe 8 : Documents utilisés pour l'évaluation des règles sur un corpus d'anglais L1

N°	Sources	Nbre de mots
1	**Articles scientifiques**	**27891**
	a. L. Michaud et K. McCoy, "An intelligent tutoring system for deaf learners of written English"	2970
	b. M. Chodorow, C. Leacock, "An Unsupervised Method for Detecting Grammatical Errors"	1450
	c. D. Philps, "Reconsidering phonæsthemes: Submorphemic invariance in English 'sn-words'"	1586
	d. M. Brattain, "Forgetting the South and Southern Strategy"	2683
	e. P. Summerfield, "Conflict, Power and Gender in Women's Memories of the Second World War"	2163
	f. Jude Brown, Simon Deakin, Frank Wilkison. Capabilities, Social Rights and European Market Integration, ESRC Centre for Business Research, University of Cambridge, Working Paper 253 (2002)	3846
	g. Kirsten I. Taylor, Barry J. Devereux & Lorraine K. Tyler (2011): Conceptual Structure: Towards an integrated neurocognitive account, Language and Cognitive Processes, 26:9, 1368-1401	2909
	h. Leigh Shaw-Taylor, "Labourers, Cows, Common Rights and Parliamentary Enclosure: The Evidence of Contemporary Comment c. 1760-1810"	3926
	i. Tiffany Stern, "A Small-Beer Health to His Second Day": Playwrights, Prologues, and First Performances in the Early Modern Theater	2269
	j. Boyden, Jo. "Children under Fire: Challenging Assumptions about Children's Resilience." *Children, Youth and Environments 13(1),* Spring 2003	4089
2	**Articles de journaux en ligne**	**25527**
	a. The Guardian (UK) http://www.theguardian.com/politics/2012/feb/03/chris-huhne-expected-resign-charges-speeding http://www.theguardian.com/books/2012/feb/03/national-library-day-year-protests http://www.theguardian.com/world/2012/feb/03/bbc-persian-staff-iranian-intimidation http://www.theguardian.com/business/2012/feb/05/transport-secretary-vote-network-rail-bonus http://www.theguardian.com/law/2012/feb/06/ken-clarke-divorced-fathers-rights http://www.theguardian.com/uk/2012/feb/06/queen-elizabeth-diamond-jubilee-year	4358
	b. The Independent (UK) http://www.independent.co.uk/news/uk/home-news/pm-under-pressure-over-gay-marriage-7844369.html http://www.independent.co.uk/news/uk/home-news/a-month-and-a-halfs-rain-in-36-hours-40-flood-alerts-but-theres-still-a-drought-7843639.html	4364

	http://www.independent.co.uk/life-style/health-and-families/features/after-the-dukan-get-the-skinny-on-the-omg-diet-7836832.html http://www.independent.co.uk/news/world/middle-east/iran-designing-nuclear-submarine-claims-news-agency-7843960.html http://www.independent.co.uk/sport/football/worldcup/uefa-to-investigate-alleged-racism-directed-at-mario-balotelli-7844403.html http://www.independent.co.uk/news/world/europe/angela-merkel-insists-europeans-must-press-ahead-with-reforms-7844676.html http://www.independent.co.uk/news/world/middle-east/police-hunt-hardline-jews-after-pronazi-graffiti-attack-after-holocaust-museum-defiled-7837056.html	
	c. The Express (UK) http://www.express.co.uk/news/uk/326150/Falkland-Islands-to-hold-referendum-to-tell-Argentina-we-are-British http://www.express.co.uk/news/uk/326151/Child-abuse-rife-across-England http://www.express.co.uk/news/uk/326146/Britain-stands-firm-as-EU-tries-to-take-control-of-our-banks http://www.express.co.uk/news/uk/326075/Now-80-demand-vote-to-quit-EU	2002
	d. The New York Times (US) http://www.nytimes.com/2012/02/03/technology/from-earliest-days-zuckerberg-focused-on-controlling-facebook.html?pagewanted=all&_r=0 http://www.nytimes.com/2012/02/03/us/komen-foundation-urged-to-restore-planned-parenthood-funds.html http://www.nytimes.com/2012/02/06/business/mortgage-relief-plan-is-closer-to-winning-support-of-2-key-states.html?pagewanted=all http://www.nytimes.com/2012/02/06/us/politics/for-ron-paul-a-distinctive-worldview-of-long-standing.html?pagewanted=all	5805
	e. The International Herald Tribune (US) http://query.nytimes.com/gst/fullpage.html?res=9407E6DA1239F931A25755C0A9649D8B63 http://www.nytimes.com/2012/06/12/arts/design/archaeologists-say-greek-antiquities-threatened-by-austerity.html?pagewanted=all http://www.startribune.com/158546415.html http://www.nytimes.com/2012/06/12/world/middleeast/china-not-issued-waiver-for-oil-trade-with-iran.html http://www.nytimes.com/2012/06/12/world/asia/after-32-years-coroner-confirms-dingo-killed-australian-baby.html http://www.nytimes.com/2012/06/12/world/africa/tensions-at-manouba-university-mirror-turbulence-in-tunisia.html?pagewanted=all	6031
	f. The Huffington Post (US) http://www.huffingtonpost.com/2012/06/12/texas-dolphins-stranded-dead-noaa_n_1591645.html http://www.huffingtonpost.com/2012/06/13/household-wealth-drop-dems-gop_n_1592783.html http://www.huffingtonpost.com/2012/06/12/national-mortgage-settlement-_n_1589499.html	2967
3	**Notes de blogs**	**27715**
	a. Oh She Glows, blog de cuisine (US) http://ohsheglows.com/2012/06/12/layered-raw-taco-salad-for-two/	2373

	b. The Everywhereist, blog de voyage (US) http://www.everywhereist.com/pitch-be-crazy-or-how-i-respond-to-pr-emails/	4234
	c. The Truth about Cars, blog automobile (US) http://www.thetruthaboutcars.com/2012/06/mahindra-xuv500-receives-4-star-ancap-rating/ http://www.thetruthaboutcars.com/2012/06/i-have-a-sports-car-again-one-ponycar-purchase-experience/	3135
	d. The Hairpin, blog de mode (US) http://thehairpin.com/2012/06/why-buying-from-emerging-fashion-designers-costs-more-money-and-why-thats-okay	2733
	e. Get Rich Slowly, blog de finance pour particuliers (US) http://www.getrichslowly.org/blog/2009/09/26/furniture-shopping-secrets-how-to-tell-superior-from-shoddy/ http://www.getrichslowly.org/blog/2012/06/08/ask-the-readers-how-much-do-you-spend-on-fun/	2821
	f. Kill Screen, blog de jeux videos (US) http://killscreendaily.com/articles/reviews/no-theyre-not-malfunctioning-theyre-alive-ethics-vessel/ http://killscreendaily.com/articles/articles/boomshakalaka/emergent-play-sports-games/	3731
	g. The Scary duck, blog d'humour (UK) http://scaryduck.blogspot.fr/2012/06/confessions-of-degenerate-gambler.html http://scaryduck.blogspot.fr/2012/05/on-hospital-waiting-rooms-and-cursed.html	2585
	h. Charlie Goes Raw, blog de cuisine et de santé (UK) http://www.charlielagoa.co.uk/fresh-juice-every-day/ http://www.charlielagoa.co.uk/winter-time-nourishment-for-pms/	1208
	i. Pickled Politics, blog politique (UK) http://www.pickledpolitics.com/archives/14067 http://www.pickledpolitics.com/archives/13967 http://www.pickledpolitics.com/archives/13931 http://www.pickledpolitics.com/archives/13856	1812
	j. Leninology, blog géopolitique (UK) http://beforeitsnews.com/middle-east/2012/06/greece-the-historical-bloc-and-populism-2242057.html	3083
	TOTAL	**81183**

Annexe 9 : Liste complète des erreurs relevées dans le corpus

NB : Le soulignement est utilisé dans les segments à titre purement indicatif, souvent pour indiquer sur quel mot précis l'erreur porte, et n'a pas de signification particulière pour la catégorisation des erreurs.

Groupe Nominal		**238**
Détermination	PUBLICATIONS - Concerning [] history of slavery in France - [] History of slavery constitutes a double trauma - and [] memory of slavery a fault line for French identity - the capacity of integration of [] French republic - VS is the personification of [] abolition of slavery - [] Jingle is characterized by - Note that [] training of these models was performed - on [] personal database - a margin of [] half second - [] Evaluation of the automatic comparison has been made - 14 seconds extracted from [] TV movie - an audio document of [] few seconds - segments of [] length lower than 20ms - MC, [] specialist of slavery matters - dealing with [] history of the French Antilles - in all [] corpus - a valuable information - the first national commemoration of the slavery - borrowing as well from the history, historiography and sociology - the construction of the social identity - it belongs to the political language - the principle of the modern political equality - The regional cultures are officially recognized by the decree of 1982 - The figure 3 gives an example of speech/music classification - The figure 1 gives a simple example - The figure 3 presents the results - The figure 4 presents an example of indexation - the speech detection - the music detection - the jingle localization - the slavery abolition - They are compared using an Euclidean distance - an history (33) INTERLANGUE - he ends up [] owner of several bakeries - I (un)fortunately cannot predict [] future. - The worse realisations are unfortunately to be found on [] social level - we are [] witness of a sort of empty conversation - the products of [] human mind - The best solution for them to get rid of the children is to switch [] TV on - they have to switch [] television off - the members of [] European parliament - On the contrary, [] computer, with all its possibilities, enables the user to create new programs - it's there that [] idea of a European Community was born - Total equality between the human beings - the role that women played in the society in general - the family may also be rejected by the society - When people are rejected by the society - the intricate and tricky problem of the feminism - The language has an enormous influence on the culture of the country - That is the case of the farmers, the customs officers…	94

	- That is the case of the farmers, the customs officers… - the "Europe 92" won't change completely the life of its citizens - I am quite sure that life would be impossible if the man stopped dreaming and imagining - industrialisation does not apparently favour the imagination - The identity of a country covers three main aspects: the independence of decision, the way of living in general, and the moral and cultural values in particular - The identity of a country covers three main aspects: the independence of decision, the way of living in general, and the moral and cultural values in particular. - the more difficult one being the Foreign Policy - mental degradation caused by material interest, represented in the novel by the silver - But aren't the films sometimes more violent than the reality? - But aren't the films sometimes more violent than the reality? - religion and television can be seen as the one and the same thing - All the year long - an European country - an unified Europe - an European nation - an horrifying effect (33) COURRIELS - it is my last day before [] holidays - This document takes into account [] agreement we had - during [] meeting on the 29th of March - each time Jira catch [] message on server - each time Jira catch message on [] server - I've passed an evening at [] friend's party - Francois made a arrangement for [] symphonic orchestra - I was looking for you [] next week - each person in [] participants' team - As I said [] few hours ago to Martin - give me [] vacation - with [] few days of delay - in [] few words - I try to show how you use the Van Eyck's picture - the Van Eyck's mirror - to have an access of the documentation - an purchase order - Francois made a arrangement for symphonic orchestra - a bad news (19) RAPPORT - before having [] result of the automatic enrichment quality - while enabling [] author to stay at a natural modularization level - the functional mock-up activity that this state of [] art prepares - [] Next point is to understand if the restricted set of pattern types… - In all these cases, [] structures concerned can be a priori detected - Interviews revealed that [] Initial goal concerning annotation of technical Documentation… - [] Definition of a complex process is difficult to demonstrate - [] Use of the diacritic on a frontier node indicates that it is a substitution node - at [] different genericity level (9)	

Modification	**Adjectif**	PUBLICATIONS - Which must be developed according to the aimed real application - no tasks ready where to plug the acquired SCFs - the national big family - this reducing approach - most contracts have very applied goals - a system able to detect these components (6) INTERLANGUE - we will get all the time wished to dream - the searched harmony - a too scientific mentality (3) COURRIELS - I will turn over your document signed - in a way decent - it will automatically generate some desynchronisation always difficult to handle - there are some letters misplaced in my mail - Not a so nice name - your quickly answer - I have excelling relations - automatics system - a Belgium composer - I have to keep the windows and store close - But for the others group I do not know them at all (11) RAPPORT	20
	Nom	PUBLICATIONS - the ghettos sickness - the brothers slaves - heterogeneous information sources cooperation - the objects properties - the meaning utterance - the annotations reliability - annotations reliability computing - texts candidates - local data entities structure (x2) - data models discrepancies - information system security strategies heterogeneity - the State official discourse - Maurice Barrès principles - the integration tradition - the indication context - an oblivion policy - the original French overseas population motivation - the width peak method - insignificant size segments - this meaning transposition - security concept model granularity - the concept meaning - the memorial laws effect - different access administration concept - addition goal - evaluation consideration - speech music classification tool - the paper reader - the cooking recipe extract - the society domain - the semantic reasoners maturity - their contents reliability (33)	45

		INTERLANGUE - admiration cries - the darkness element - money inequality - television information (4) COURRIELS - the length excess - our friend relationship - partner equivalence table - the architecture study process (4) RAPPORT - information extraction technology results - the semi-structured procedural text analysis challenge - the language knowledge acquisition bottlenecks - the domain specificities (4)	
Choix et utilisation des prépositions		PUBLICATIONS - a greater place <u>of</u> this history in school programs - the descendants <u>from</u> slavery - participation <u>to</u> these trials - the ritual is <u>in</u> the center of the social life - determine the right tag(s) <u>of</u> each word - the millions [] Africans victims of the slave trade (6) INTERLANGUE - there is a rise <u>of</u> nationalism everywhere - This may also be emphasized by a certain tendency <u>of</u> universalism and internationalization - Love and religion have been the subject <u>for</u> many passionate writings - Mr Wimbush interrupts that attempt <u>to</u> conversation - There is a certain pride <u>of</u> being Belgian - one student coming from a working class family does not have the same advantages [] a student <u></u>whose parents are doctors or lawyers - most <u>of</u> houses - most <u>of</u> people - This 19th century quotation [] Victor Hugo is, I think, still true today - the first part deals with the question [] whether or not the expedition to… (10) COURRIELS - your request <u>of</u> help from the French embassy - to have an access <u>of</u> the documentation - a training course <u>of</u> robotics and/or computer sciences - participants <u>to</u> this task - all the participants <u>to</u> this task - all the participants <u>to</u> the evaluation - your rate <u>of</u> year 2007 - I need some explanation <u>on</u> where is XX - it is less than 3 hours <u>far</u> from Toulouse by car (9) RAPPORT - we have strong doubts <u>on</u> its usability (1)	26
Choix et utilisation des pronoms		PUBLICATIONS INTERLANGUE - The stake of the objective "92" is mainly an economic one but also <u>this</u> of a new political configuration - they play at soldiers, at killing <u>each others</u> (2)	2

	COURRIELS RAPPORT	
Accord	PUBLICATIONS - the millions Africans victims of the slave trade - four very different corpus (2) INTERLANGUE COURRIELS - these cost - this goods - these file - I have so few close friend - these friend - I have plenty of thing to do - two others addresses - a folder containing three list of files - a lot of great concert (9) RAPPORT - at different genericity level (1)	12
Lexique	PUBLICATIONS - the adding of video parts - the using of the Gaussian mixture models - for further informations - advices (x 7) - French medias - Evening gathering gives the opportunity to remind a genealogy (12) INTERLANGUE - progresses in technology (1) COURRIELS - it doesn't accept file bigger than 10Mo - there is enough spaces for sleeping - you will transmit my apologize to her - I can't present this document to my Direction without explanation such an inflation of price - other informations - automatics system - a Belgium composer - in my ownership - I am in courses - It is the end of the courses (10) RAPPORT - to ease the fulfill of some enriched aircraft documentation structures - advices (x 2) - not their fully integration as linguistic resources (4)	27
Autres erreurs	PUBLICATIONS - as many as possible of incorrect analyses - as less alterations as possible - the memory of slavery's requests - image has become one of the main vectors in the writing of history - Among the populations concerned by the law, the Caribbean reacted very violently (5)	12

	INTERLANGUE - Most of the parents have confused "happiness" with "success" - It is important for most of the people to have a task to fulfill - Book fairs have [] success - It will enlarge the European's views (4) COURRIELS - some mp3 file - because of thief - it doesn't accept file bigger than 10Mo (3) RAPPORT	
Groupe Adjectival		16
Construction du comparatif et du superlatif	PUBLICATIONS INTERLANGUE - the more and more quick ways of transport and communication - The more varied, the more rich our culture will be - The competition coming from Japan is a lot more stronger - as many as possible of incorrect analyses - similar feature than forum's one - similar feature to forum organization (6) COURRIELS - The better will be to pay the difference (1) RAPPORT	7
Choix des prépositions	PUBLICATIONS - dependent from the verb - dependent from a given corpus - dependent from the domain or the text genre - what elements are dependent from a given predicate (4) INTERLANGUE - This is typical for the 20th century - They are not concerned for any possible loss of identity but to earn their daily bread (2) COURRIELS - I'm not aware on how this database server works - submission will be due to three weeks later (2) RAPPORT	8
Autres	PUBLICATIONS INTERLANGUE - one trait of her personality is very much striking (1) COURRIELS RAPPORT	1
Groupe Prépositionnel		11
Choix et présence des prépositions	PUBLICATIONS - in March 16th, 2005 - it is also, and before all, a way to conceptualise a domain	

		- they are exchanged and read on their electronic form (3) INTERLANGUE - these are five pieces of land if you look at them for "above" - Since a few years now people in all parts of the European continent - The man painted and repaired her apartment during almost fifty hours - Besides of this (4) COURRIELS - To your question: How much time would you need to evaluate the submissions? - you could answer me on this address - I've spent my last night in a train - From now [] (4) RAPPORT	
Groupe Verbal			218
Placement des modifieurs après le verbe	**Adverbes**	PUBLICATIONS - to remember collectively a genealogy - in order to hang down exclusively family memories - the treatment of this official day exemplifies also an answer to associations - We have tested separately all the parameters - To index efficiently the soundtrack of multimedia documents - [they] were found to improve substantially the performance of either modality - our system is able to derive automatically information for a large number of verbs - it had then its own evolution - [it] has now the status of a sense - the similarity between words depends then on the amount of normalized contexts they share - the educational range […] seems to have here to play its role - the value of N needs also to be elaborated - but exhibit nevertheless the dependency relationships observed in the source parse tree (13) INTERLANGUE - I understand very well this point of view - "Europe 92" won't change completely the life of its citizens - Perhaps he cannot show very well the way he feels - favorizing then perhaps more easily the student's future professional insertion - favorizing then perhaps more easily the student's future professional insertion - his father resembles strongly his own character - That has also its effect on the atmosphere and relationship between people - the latent harmony will more and more increase - international business will more and more play a major role in the world - such a person […] can't adapt completely its way of thinking and being - it has also to provide political answers - Europe has also weak points - the European community wants also to be taken seriously - it means also a cultural richness conveyed by these dialects - the Community forms also a mosaic of ethnic cultures - Everything is possible but anyway will happen so quickly (16) COURRIELS - I have also a bad news (1)	36

		RAPPORT - we have here an advice and a warning - it can be also considered as restricted in terms of variants - the prototype is now evolving to become shortly a software component - The input documents can be a priori any type of Web page - LFG got also several implementations - considering also specifications taking into account by smart assistants (6)	
	Autres	PUBLICATIONS - Ontological domains include in our view objects, their properties and relations - do not accommodate in their descriptions any metaphors - borrowing as well from history, historiography and sociology than from political philosophy - we have designed ourselves scales to structure complexity - We have developed for French the description of 1700 verb sense (5) INTERLANGUE - they have indeed to cope with their job - the word "birth" in the title is according to me ironical - we might change a bit the statement proposed - when Europe is for good under way - European community leaders want to create by the end of 1992 a single market of goods (5) COURRIELS - I am going one week to Poland - I still play a lot the trombone (2) RAPPORT - which is in the field of language technologies supported by various Evaluation Campaigns (1)	13
Placement des compléments après le verbe		PUBLICATIONS - The SCF acquisition system takes as input a large corpus (1) INTERLANGUE COURRIELS - I will put on my website some mp3 files - it will wake up me - I do not doubt that you will be able to formulate us solutions - please say him "Bonjour" from me - I hope to render comprehensible myself - That's why I'm sending to you some important information - I would like to remind [] that the problem is not solved today. (7) RAPPORT	8
Morphologie du GV		PUBLICATIONS - Systems are robust and could be improve on every corpus - These two values have been determinate after many experiments - does this other kind of evaluation correlates with intrinsic evaluation? (3) INTERLANGUE COURRIELS - You do not have takes action - you will not provided me satisfactory explanations - I will sent all the information required - I've to go to bed	31

	- I can lost a love - I can't lost a friend - I may planned to go to Prague - Must have to be great - I never had has to pay these costs - Nobody have answer me - Thank you for having calling me - I've promise you - you've listen to some great music - I've just check the prices - I hope you have spend a good week-end - it going to be too hot - one is become my sister's boyfriend - I forgotten to attach the file - something must be clear up - I am not suppose to attend it. - To be discuss later. - I am sorry, but he didn't used - As I did not heard about this phone call - I didn't managed (24) RAPPORT - It can takes a long time - considering also specifications taking into account by smart assistants - a few applied works shown that it is possible to simplify the problem - without any help to collect and formalized samples (4)	
Utilisation et construction de la négation	PUBLICATIONS - interventriculaire and pont do not belong any longer to the neighbors of artère - they are not only constrained to the author's point of view anymore (2) INTERLANGUE - people do no longer dream - Not only because we are no children any longer - We are no more ready to make sacrifices - Have you never opened a newspaper? (4) COURRIELS - You do not have forgotten us - You do not have takes action - I just not know when - I still not know the content - My employer have not to give me a vacation (5) RAPPORT	11
Choix et utilisation des prépositions	PUBLICATIONS - a category of people designated like Black - this system results of the study - the system is divided in two subsystems - based on a cepstral analysis for speech and of a linear spectral analysis for music - participates at various degrees to the success of the action - these scores depend from the gold standard - 1848 was presented like a gift from France - without knowing of what it was exactly - it was necessary to wait [] the nineteen sixties - after analyzing on this program - the majority of the jingles is classified [] music - (or inherits of) the property - the decision is made regarding to the maximum likelihood (13)	66

	INTERLANGUE - They are not concerned for any possible loss of identity but to earn their daily bread - But if we look deeper in this novel - this essay does not strive to exhaustiveness - This common culture leads us to wish [] a united Europe - this tendency can come at the surface - we could start on the first part of the statement - to work [] a few years in another European country (7) COURRIELS - I can't present this document at my Direction - If we decided to go down at the subgenre level - will participate to the TRECVID 2009 - Each whole program will be labeled by a specific genre - and maybe by a subgenre - the agreement needs to be translated in French - who will participate to the annotation - I am going in Poland - I am going in a conference - If you can't answer to my questions - to discuss about the tests - all these subsets do not have of the same size - I will not discuss on Jira - I will come back at home - when I come back at home - Speak/see to you soon - I don't wish to be taken as an hostage of your resentments - Thank you of your call - what should I do of this defective material? - The cake may bake [] around 20 minutes - You might ask somebody who is Ms SQL Powered to know if he knows [] that issue - we leave [] the 6th of January - I am going [] one week to Poland - If you arrive [] a week-end I can rent a car - I'm sure they will split [] soon - Could you be so kind [] to tell me - the products that I usually order [] you - I wait [] your answer - I didn't ask you [] these goods - I have some requests to formulate [] you - I regret to acknowledge [] you that I carried out an ordering of 10 probes - so that we can continue to operate [] our patients - I do not doubt that you will be able to formulate [] us solutions - Can you explain [] me how Qan FTL are made in Alt CM ? - The people in charge of these addresses will confirm [] you the telex reception - I wrote [] Silvia to book an apartment - Can you confirm [] me that everything's OK ? - Can you send [] me before Friday? - So I write [] you because I am looking for a training course - please say [] him "Bonjour" from me (40) RAPPORT - in order to reason on it depending of the context - some new triple set depending of the context - to directly refer [] the right access platform - It allows to the definition - it is important to distinguish [] Natural Language Technologies, Statistical	

	Text Mining and Text Semantics. - We end up this section by the presentation of two projects (6)	
Lexique	PUBLICATIONS - we precise the two different feature spaces - These two values have been determinate after many experiments - We can determinate the beginning of a jingle - it is mainly made by hand - The jingle is completely recovered by speech - this section processes in two successive phases - Its negation signifies an abnormal development - to remind a document - 14 seconds extracted from TV movie (9) INTERLANGUE - Why should we loss our identity? - The second point to be precised is the fact that - people are decided to keep their identity - This particular type of studies is very much attended to by students - they must get aware of the differences (5) COURRIELS - I've looked for you to come in Toulouse - We meet a problem with PKZ - I've already made a training in 2006 in NEWI - Now I am making a Master - test data will be comprised of program extracts - Have I directly to go through Lisa? - Are you agree? - I can respond to questions - to try to achieve the version of the evaluation guidelines for the task 10.2 - I passed one very good day off - I am in a hurry to be there - We will take again these transactions - would you have kindness from now on to mail your requests and/or questions to… - could you do the necessary on your side - I've passed an evening at a friend's party - I have to let you - I've passed my last night in a train - I'll have to let you - I hope you have passed a good week-end - I await your news - I await your response - you will turn over your document signed (22) RAPPORT	36
Autres erreurs	PUBLICATIONS - The combination of these approaches allows to raise the accuracy rate - A commemoration permits to associate the individual with the group - our system permits to detect any reference jingle - This important delay permits to process the data - This approach permits to better characterize each component - a phase allows to concatenate neighboring frames - It allows to implement - a RDF library enables to reason in situations - This property will allow to go further - a pattern which enable to identify many variants - Those standards enable to…(11)	17

	INTERLANGUE - Many young people feel themselves completely lost - he should also feel inhabitant of the European nation (2) COURRIELS - I always endeavour myself - I was sure that we were understood - to keep me busy - I don't care this much about the source code (4) RAPPORT	
Proposition et Phrase		178
Construction des phrases interrogatives	PUBLICATIONS INTERLANGUE - We are going to become Europeans very soon. What implies this? (1) COURRIELS - It is possible to receive the parcel for the end of August? - I would ask him how he wants to organize the return? - what a good partitioning would be? (3) RAPPORT - how to underlined words can be replaced? (1)	5
Construction des propositions subordonnées à verbe fini	PUBLICATIONS - it is necessary to know [...] what is their role in the action expressed by the predicate - in case where Ai is an advice - James Baldwin, which characteristic is to be an American writer in France - Evening gathering, whose ritual is at the center of the social life - those for who Vichy rhymes with slavery - We called this approach the ** approach to emphasize the fact it is necessary to … - We observe there is always a manual frontier - all the feelings that another is controlling the patient's thoughts - to the extent it was an emanation of the speech - When appeared the Question Noire, there were many claims about the history of slavery (10) INTERLANGUE - Let's first have a look at what is Europe actually. - we could ask ourselves what is this attempt towards harmony going to be after Mrs Ramsay's death - there are some factors [] justify both aspects of the problem - which is precisely characterized by that what is not material - It would be such a great pity [] you miss the opportunity (5) COURRIELS - what are the problems - who is your phone contact - I need some explanation on where is XX - If you can't answer my questions (that I would understand) - some laboratory on NEWI which works on this domain - Do you want that I return you the old probes RF? - I regret to acknowledge to you which I carried out an ordering of 10 probes (7) RAPPORT - This introduction roughly defines what a procedure is, what is its structure	25

	in linguistic and conceptual terms. - the way it can be evaluated - Advanced programming experience and considerable programming time may be necessary, that are not available in Web services or similar scenarios (3)	
Construction des propositions subordonnées à verbe non-fini	PUBLICATIONS - The President of the Republic wishes to honoured the memory of the… - A jingle has the particularity to contain speech - It has the advantage to present long periods of speech - which consists in detect the two basic components - a prior partitioning which consists in detect the components - respond to annotations for criticizing them - explore the thread for evaluating - The third topic means the parenthesis closing - no tasks ready where to plug the acquired SCFs (9) INTERLANGUE - and who has imagined to divide the world into three categories of men - a country without money cannot help feeding its population - They are therefore considered of being of very little value - But those areas becoming specialized on a world scale in the manufacture of particular products, there will be migrations of specialized workers - This little magic sentence sounds like belonging to an old past (5) COURRIELS - Thanks for your kindness to read my bad English - thank you to take into account my requests - thank you to indicate a total price - why not making another one this year? - No problem to have deleted the ACK ! - Calgary sounds to be a great place to work and study (6) RAPPORT - CRF will offer perspectives like to be trained (1)	21
Accord sujet-verbe	PUBLICATIONS - A year later, riot have started in Clichy-sous-Bois - Evening gathering give the opportunity to… - This is especially true when one work on a new language and do her first experiments - it is labour-intensive and require to have a practical application available - some properties that makes it a valid metaphor - when advice are executed - a pattern which enable to identify many variants (7) INTERLANGUE - Other react less strongly and puts limits - unity among the people who enjoys the long summer with her - Ever since the industrial revolution which have radically marked the transition - they see the world in a two-valued orientation, just like television do (4) COURRIELS - I'm returning you the old system which have never functioned - This have to be discussed - I'm a French student who have a placement in NEWI - the network work well - I think it is not a good solution but it work - I'm not aware on how this database server work	35

	- the XML file which describe the rules - each time Jira catch a message - the XML file read each minute - That lead me to another point - It work too - The plugin process only multipart mails - The send function which work - it work on my computer - there is enough spaces for sleeping - We arrives at 10:30 am - we leaves on the 6th of January - The week end was good, this evening quite bad and make me sad - My employer have not to give me vacations - The temperature are about 21°C -my train leave at 7pm - Nobody have answer me (22) RAPPORT - The causal structure that focus on the goal of an action - a way which comply with the formats (2)	
Choix du temps	PUBLICATIONS INTERLANGUE - a European Community is born (1) COURRIELS - [Do I have to do this] so that was simpler? - Can you tell me quickly if I receive a parcel this Thursday - This Friday, November 9 I have a meeting with Dr. K - I talk about your products - I fall in love with a guy who already has a girlfriend - You might find with this mail the package I promise to send to you - The week end was good, this evening quite bad and make me sad - she proposed and I accept - I just test a recipe from Marion - I believed it was ok, but now I realised that not at all (10) RAPPORT	11
Choix de l'aspect	PUBLICATIONS - the claims concerning the history of slavery are deriving from a special context - during the evaluation phase, we have studied the precision of the detection - terms that are not corresponding to actual local data entities - It is absolutely not dealing with slavery (4) INTERLANGUE - They say that man has no opportunity to dream or imagine and that he becomes a robot - Mr Wimbush butts in again, forces him to listen to what he says (2) COURRIELS - I see with the publisher for the quality. - So I try again - I return you the old system - when one speaks in French close to you - It looks like we don't receive any FT from/to DP for flights - I send you 3 of the PFS we built this morning	21

	- We discuss this point wih M - we still do not receive any FT from your side - I attach 2 papers for your consideration - That's why I send to you some important information - So I write to you because I am looking for a training course - I'm writing this mail for one hour now - I contact you to know how it is possible to get this license - We meet a problem with PKZ (14) RAPPORT - In this document, we will only focus on variants of NLP and Text semantics approaches that emerged through standardization efforts for a few years (1)	
Présence et choix de l'auxiliaire modal	PUBLICATIONS - the idea according to which there would be a specific problem - patients who suffer from FRS would be unable to correctly monitor their actions (2) INTERLANGUE - We would better say (1) COURRIELS - I would ask him how he wants to organize the return - I hope that at the end of July I could tell your more about this point. - You might find with this mail the package I promise to send to you - You might find in the attachment a summary from smartwings in PDF - The cake may bake for around 20 minutes - the transfer could be quickly done, as soon as we will get those numbers. - I will confirm these operations to you, as soon as they will be done on our side. - we will wait until your credit card facilities will be on line. - Scoring tools will be developed once the metrics will be discussed - I will convert them in MP3 as soon as I will have enough time (10) RAPPORT	13
Planification de l'information	PUBLICATIONS - Remote territories, out of the metropolitan borders, it is firstly with overseas departments that France discovers its identity - The first one is a comment of a championship of figure skating of 30mn - Recent doctoral theses from this group are the work of X and Y - A general definition of metaphors can be the following - To commemorate the abolition of slavery in France, it is to speak about humanity and Republican values (5) INTERLANGUE - it can influence some regimes for instance to change their policy (1) COURRIELS - The week end was good, this evening quite bad and make me sad - for your culture august 15 in France, it is Assumption day - I believed it was ok, but now I realised that not at all - So the days after, we will be reachable only by mobile phone (4) RAPPORT	10
Cohérence des pronoms	PUBLICATIONS - Each other can express its feedback and questions - Each other can express its feedback and questions (2) INTERLANGUE - That has also its effect on the atmosphere and relationship between	10

	people - we need money for anything - as long as it doesn't disadvantage himself - one cannot decide oneself that somebody is "like this" or "like that" - Belgium is the capital of Europe and she is federal (5) COURRIELS - It is essential to be honest and loyal on both sides so that functions - Your document doesn't have nothing like it - I still play a lot the trombone to keep me busy (3) RAPPORT	
Lexique	PUBLICATIONS - Although, such an excess of mental effort should be reduced at all costs - Although, accurate representations of predicted states... - For as much, if the use of expressions such as… - syntactic parsing is even the only way of ... - More, the acoustic signal is divided into four types (5) INTERLANGUE COURRIELS - Unless, it was raining today. - As soon, I return you the old system - unhappily for a small audience - Still one time - because of I still not know the content of my new internship - Prague misses me - Least, but not last - Well cordially (8) RAPPORT	13
Autres erreurs	PUBLICATIONS - similarly to a rusty tool may malfunction - it is already of habit - quite often of type +animate - the argument is of type psychological - As well as I do not ignore the historical importance of this day, I estimate that (5) INTERLANGUE - Though it has been many years that the first rocket has been launched into space - I think dream isn't lost at all - Dream has indeed always been a means to escape reality (3) COURRIELS - on the bottom and the right of the picture - I promise to make efforts with regard to my English - Must have be great as if I remember, they play very well - we will just care that your paper is reviewed independently - Thank very much - I cannot before (6) RAPPORT	14
TOTAL		661