Diagnostic et remédiation orientés vers le lexique en compréhension aurale de l'anglais

Marie-Pierre Jouannaud

N° d'ordre NNT : 2021LYSE2004

THESE de DOCTORAT DE L'UNIVERSITÉ DE LYON

Opérée au sein de

L'UNIVERSITÉ LUMIÈRE LYON 2

École Doctorale : ED 484

Lettres Langue Linguistique Arts

Discipline : Lexico Terminologie

Soutenue publiquement le 13 janvier 2021 par :

Marie-Pierre JOUANNAUD

Diagnostic et remédiation orientés vers le lexique en compréhension aurale de l'anglais

Devant le jury composé de :

Vincent RENNER, Professeur des universités, Université Lumière Lyon 2, Président

Shona WHYTE, Professeure des universités, Université Côte d'Azur, Rapporteure

Claire TARDIEU-GARNIER, Professeure des universités, Université Sorbonne Nouvelle - Paris 3, Rapporteure

Denis JAMET, Professeur des universités, Université Jean Moulin Lyon 3, Examinateur

Heather HILTON, Professeure des universités, Université Lumière Lyon 2, Directrice de thèse

Contrat de diffusion

UNIVERSITÉ DE LYON

École doctorale Lettres, Langues, Linguistique et Arts –ED 484

Laboratoire CeRLA (ex-CRTT), Université Lyon 2

Diagnostic et remédiation orientés vers le lexique en compréhension aurale de l'anglais

Marie-Pierre JOUANNAUD

Thèse pour l'obtention du grade de Docteure en didactique de l'anglais

Sous la direction de Heather HILTON

Date de soutenance : 13 janvier 2020

Composition du jury :

Vincent Renner	Professeur, Université Lyon 2, président
Claire Tardieu	Professeur, Université Sorbonne Nouvelle, rapporteuse
Shona Whyte	Professeur, Université Côte d'Azur, rapporteuse
Denis Jamet	Professeur, Université Lyon 3, examinateur
Heather Hilton	Professeur, Université Lyon 2, directrice de thèse

Table des matières

Remerciements

Je tiens tout d'abord à remercier Heather Hilton, Monica Masperi, Coralie Payre-Ficout et Annick Gibaud, qui, chacune à leur façon, ont beaucoup compté dans mon parcours de chercheuse. Merci en particulier à Heather, à sa direction avisée, son soutien sans faille et son enthousiasme contagieux ! Je n'aurais absolument pas pu mener à bien ce travail sans son aide et ses conseils.

Merci aux membres du jury, Claire Tardieu, Shona Whyte, Denis Jamet et Vincent Renner, qui ont accepté de lire et d'évaluer un manuscrit de plus en cette période chargée.

Enfin, merci à tous mes collègues enseignantes et enseignants, techniciennes et techniciens, chercheuses, chercheurs et administratifs (souvent un peu de tout cela à la fois), de l'UFR de Langues Etrangères de l'UGA, du projet Innovalangues, du Lansad/ Service des Langues, ainsi que du Lidilem à l'UGA et du CRTT à Lyon, qui m'ont accompagnée et aidée de près ou de loin, et qui m'ont tout simplement permis de travailler dans une atmosphère très conviviale, et linguistiquement et culturellement très riche. Merci également aux collègues rencontrés uniquement virtuellement, et qui ont partagé généreusement leurs articles, thèses, idées, et même stimuli, et m'ont permis de les utiliser. Merci, enfin, à mon relecteur préféré…

Introduction générale

En 1928, Paul Rankin constatait que la compréhension de l'oral (CO) était l'activité langagière la plus courante de la vie quotidienne, devant la production orale, et loin devant les compétences écrites. On a vérifié depuis que cette constatation était d'autant plus vraie en contexte scolaire (Imhof, 2008) ou universitaire (Barker et al., 1980), et en particulier dans les cours de langues étrangères (Chaudron, 1988, p. 51), où l'écoute est reconnue depuis les années 1970 comme une source essentielle de données langagières (*input*) qui sous-tendent les processus d'acquisition de la langue étrangère (par ex. S. M. Gass & Madden, 1985; Krashen, 1981). Cependant, si les élèves et étudiants passent la majeure partie de leur temps en cours de langue à écouter, la question est de savoir s'ils saisissent vraiment ce qu'ils entendent (Feyten, 1989), et surtout ce que nous pouvons faire pour les aider à mieux comprendre.

Chez les étudiants français apprenant l'anglais, qui nous concerneront dans cette étude, les tests de positionnement à l'entrée à l'université en filière LLCER (langues, littératures et civilisations étrangères et régionales) montrent que la compréhension de l'oral est la compétence où ils sont les plus faibles : moins de 25% des primo-inscrits en licence d'anglais à Grenoble ont le niveau B2 en compréhension de l'oral attendu en fin de lycée, alors qu'ils sont presque 30% en expression écrite, et entre 40 et 50% en compréhension de l'écrit et production/ interaction orale (Payre-Ficout, 2011). Pourtant, la compréhension de l'oral est essentielle pour ces étudiants dans la mesure où la majorité des enseignements en licence d'anglais est dispensée dans cette langue, et en particulier les cours magistraux. D'autres études (Terrier, 2011) confirment ces difficultés, cette fois chez un public LANSAD (LANgues pour les Spécialistes d'Autres Disciplines).

Cet état de fait est antérieur à l'université : selon l'enquête européenne Surveylang (Commission Européenne, 2012), 40% des collégiens de troisième (de 15 ans environ) n'ont pas encore atteint le niveau A1 (niveau introductif) du CECRL (Cadre Européen Commun de Référence pour les Langues) en compréhension de l'oral en anglais, alors que le niveau A1 est théoriquement attendu en fin de primaire (Ministère de l'Education Nationale, 2007). Ils ne

sont « que » 28% en compréhension de l'écrit et 24% en production écrite à ne pas atteindre ce niveau[1].

Le but de cette étude sera d'élaborer des tests diagnostiques en compréhension de l'oral en anglais pour les étudiants à l'entrée à l'université en France, qui permettent d'identifier le plus tôt possible dans leur parcours universitaire leurs lacunes et leurs acquis, et qui leur proposent des pistes de remédiation. Il existe déjà un certain nombre de tests diagnostiques en anglais, qui couvrent entre autres la compréhension de l'oral (ces tests seront décrits en détail dans le troisième chapitre). Cependant, les tests existants se focalisent sur la performance des candidats, et les résultats donnés portent toujours sur des micro-habiletés (par exemple, compréhension détaillée, ou compréhension de textes journalistiques) qui ne sont pas mises en relation avec une théorie du développement de la compétence, ce qui fait qu'on retrouve les mêmes micro-habiletés à tous les niveaux. Nous prendrons une autre voie et suivrons les recommandations de Field (2008a), pour qui une approche diagnostique doit s'intéresser au processus de l'écoute et non pas simplement à son résultat, la compréhension. Laveault et Grégoire exposent très clairement cette idée (ici dans le domaine mathématique) :

> [Le] but [d'un test diagnostique] est de comprendre le sens d'une performance. Par exemple, il ne s'agit plus, comme avec un test certificatif, de simplement vérifier si un élève peut additionner correctement deux nombres décimaux, mais de comprendre pourquoi certains élèves présentent des difficultés pour réaliser de telles additions. L'information que l'on désire recueillir ne se limite plus à la performance, mais concerne les capacités cognitives sous-jacentes à ces performances. Pour atteindre cet objectif, il est nécessaire d'utiliser un test qui s'appuie sur un modèle des processus mis en jeu pour réaliser des additions avec des décimaux. Un tel modèle permet d'éclairer les difficultés rencontrées par les élèves et, le cas échéant, de mettre en oeuvre des actions remédiatives. (Laveault & Grégoire, 2014, p. 10)

La démarche est la même dans le domaine des langues étrangères : ici non plus, il ne s'agit pas seulement de constater que les étudiants ont des difficultés à comprendre tel ou tel texte, mais bien d'arriver à comprendre les raisons de ces difficultés. Il faudra donc identifier, à partir d'une exploration des mécanismes cognitifs sous-jacents, les facteurs qui contribuent à la compétence de compréhension de l'oral. Cette exploration nous conduira à la description d'un modèle du processus de compréhension, sur lequel nous nous appuierons pour concevoir nos tests diagnostiques. Ces tests nous aideront, à leur tour, à mettre en lumière les étapes du

[1] Cependant, une étude récente de la Direction de l'Evaluation, de la Prospective et de la Performance du ministère de l'éducation nationale (DEPP, 2017) donne des résultats plus encourageants pour les élèves en fin de collège-, avec presque 60% des élèves ayant un niveau satisfaisant en CO en 2016.

processus qui sont problématiques pour certains étudiants, afin de leur proposer des activités de remédiation.

Un des facteurs qui joue un rôle essentiel dans la compréhension aurale, déjà bien identifié et étudié, est celui des connaissances lexicales (par ex., Hilton, 2006). D'autres facteurs ont également été proposés, comme les connaissances grammaticales, la capacité à utiliser l'accentuation pour le découpage lexical, ou la mémoire de travail. Cependant, ce travail de thèse se démarque des précédentes études à deux égards. Premièrement, elle a l'ambition d'analyser l'effet de plusieurs facteurs en même temps, alors que beaucoup d'études antérieures portent sur le rôle d'un seul facteur (par exemple l'influence des connaissances lexicales sur la compréhension de l'oral), ou la comparaison de deux facteurs (par exemple, les connaissances lexicales et grammaticales et leur effet comparé sur la CO). Nous essaierons au contraire d'inclure un nombre important de facteurs parmi ceux identifiés dans la revue de littérature - en particulier les connaissances phraséologiques, dont le rôle en compréhension de l'oral n'a encore été que très peu étudié. Pour identifier les niveaux atteints en compréhension de l'oral, nous utiliserons le test SELF, développé précédemment dans un autre cadre (Cervini et al., 2013).

Cette étude a également un objectif pratique, celui de développer des instruments qui soient utilisables directement par les enseignants qui accueillent des étudiants qui doivent étudier l'anglais ou en anglais, et pour qui la compréhension de l'oral est un atout essentiel. C'est particulièrement le cas pour les étudiants qui s'inscrivent en licence d'anglais, où l'objet d'étude est l'anglais et où la plupart des cours sont en anglais, mais peut concerner aussi les étudiants LANSAD, dont les cours de langue sont également dispensés en anglais, et qui peuvent également suivre une partie de leurs autres cours dans cette langue (EMILE, ou Enseignement d'une Matière par l'Intégration d'une Langue Etrangère, ou en anglais *CLIL*, *Content and Language Integrated Learning*). Les deux populations seront utilisées dans cette étude. Ce cadre réaliste et écologique impose quelques contraintes, notamment celle du mode (informatique, pour que ce soit possible à distance et que les résultats puissent être traités plus facilement) et du temps de passation (une durée d'une heure environ) des tests diagnostiques à développer.

Un autre objectif de cette thèse est de proposer des activités de remédiation correspondant aux facteurs identifiés comme importants. C'est pourquoi nous nous focaliserons essentiellement sur les facteurs qui peuvent être modifiés ou améliorés par l'instruction (et non sur le niveau

d'études des parents ou la mémoire de travail, par exemple). Nous nous pencherons en particulier sur le cas des étudiants qui n'ont pas atteint le niveau B2 du CECRL (Cadre Européen Commun de Référence en Langues) en compréhension de l'oral. Le niveau attendu actuellement en fin de scolarité secondaire est de B2 (utilisateur indépendant) pour la première langue vivante étrangère pour le bac général. C'est aussi le niveau qui est attendu en entrée de licence de langue, et le niveau B2 du CECR sera donc la frontière qui nous permettra d'identifier les étudiants ayant besoin de remédiation, c'est-à-dire essentiellement ceux de niveau A2 et B1. Cette remédiation, comme les tests eux-mêmes, est prévue en ligne, et en sus des cours, pour que les étudiants soient libres de le faire quand ils le veulent ou peuvent et en prenant le temps nécessaire, qui peut varier d'une personne à l'autre. Pour des contraintes de temps, ces propositions n'ont pas été expérimentées à grande échelle et seront uniquement présentées et justifiées dans la dernière partie.

Plan de la thèse

Par son accent sur la compréhension de l'oral en langue étrangère, cette thèse se positionne au croisement des études en acquisition du langage (y compris en psycholinguistique), et en didactique des langues étrangères. Par son intérêt pour l'évaluation diagnostique, elle se place dans le *testing* (un domaine de la linguistique appliquée qui est peu développé en France, mais qui est représenté par un journal prestigieux, *Language Testing*). Elle est divisée en trois parties : une première partie théorique, une deuxième partie sur la validation des instruments utilisés, et une dernière sur l'analyse des relations entre les variables explorées et les conséquences pédagogiques qui découlent des analyses effectuées.

La première partie, composée de trois chapitres, présentera le cadre théorique de l'étude. Bien que le sujet principal soit l'acquisition d'une langue étrangère (L2), le premier chapitre sera consacré à une revue de littérature sur les modèles de la compréhension de l'oral en langue maternelle (L1). En effet, comme nous le verrons, le consensus scientifique actuel postule que les processus sont les mêmes en L1 et en L2, et que seuls l'état initial et les données traitées par le système diffèrent. Le deuxième chapitre enchaînera plus précisément sur la compréhension aurale en anglais langue étrangère, et détaillera les problèmes spécifiques aux apprenants francophones. Enfin, le troisième chapitre conclura sur l'analyse des tests existants, et sur une présentation des outils du *testing* diagnostique que nous utiliserons ensuite pour élaborer les tests.

La deuxième partie, composée de six chapitres, décrira en détail l'élaboration de chacun des tests diagnostiques, et appliquera pour chacun d'entre eux le processus de validation décrit au chapitre trois. Le chapitre quatre présentera le plan de l'expérimentation et la logique de l'enchaînement de nos démarches (groupes de sujets, choix de tests statistiques, etc.). Les chapitres suivants décriront le développement et la validation de chacun des tests diagnostiques utilisés : le test de sensibilité prosodique (chapitre cinq), le test de discrimination phonémique (chapitre six), le test de reconnaissance aurale du lexique (chapitre sept), et le jugement aural de grammaticalité (chapitre huit). Nous présenterons dans le chapitre neuf les autres tests utilisés, mais qui n'ont pas été développés spécifiquement dans le cadre de ce travail, à savoir le test de connaissances phraséologiques, ainsi que le test de compréhension de l'oral, qui sera utilisé comme variable à expliquer dans la dernière partie.

Enfin, la troisième partie de cette thèse sera consacrée d'une part à l'étude corrélatoire entre les différents tests élaborés et analysés dans la deuxième partie, et à l'analyse de tous nos résultats. Elle se conclura par la présentation des conclusions et préconisations didactiques que nous pourrons formuler à partir de ces résultats, et sur des propositions d'activités de remédiation qui en découlent.

Précisions terminologiques

Comme Duchet et Paillard (1985) ou Hilton (2009), nous avons choisi de ne pas utiliser le terme « compréhension orale » pour parler de la réception d'un message oral. Nous utiliserons soit le terme « compréhension aurale » (puisque la réception passe par les oreilles), soit « compréhension de l'oral », qui a l'avantage de s'abréger en CO, une abréviation qui est couramment utilisée dans les matériels pédagogiques.

Etant donné que cette thèse sur l'acquisition de l'anglais a été écrite en français par une francophone enseignante d'anglais, et utilisant des sources en grande partie anglophones, de nombreux termes seront présentés de façon bilingue. Cela permettra, nous l'espérons, aux lecteurs des deux langues de repérer facilement des termes qu'ils connaissent mieux dans l'une ou l'autre langue, et éventuellement d'apprendre celui qu'ils ne connaissaient pas. Pour plus de facilité de lecture, les italiques seront réservées aux termes anglais (et à tout texte en anglais de façon générale). Les autres termes d'origine étrangère (en particulier latine et grecque) utilisés couramment en français ne seront pas en italiques.

Enfin, cette thèse utilise dans sa dernière partie des tests statistiques, un domaine dans lequel les enseignants de langue ne sont en général pas formés. Nous avons donc essayé de détailler le plus possible ces analyses, avec le risque d'ennuyer le lecteur chevronné, mais en espérant pouvoir être utile aux chercheurs débutants, semblable à celle que j'étais au moment de commencer ce travail. Le séparateur décimal sera la virgule dans le texte rédigé, mais le point dans la plupart des tableaux de résultats, qui ont été générés à l'aide du logiciel *R* (R Development Core Team, 2005).

Chapitre 1

Modèles de la compréhension de l'oral

Comment fonctionne la compréhension de l'oral? Comment se fait-il que cette transformation quasi instantanée de sons produits par l'appareil vocal d'un congénère en sens immédiatement compréhensible, qui nous semble si naturelle dans notre langue maternelle, nous paraisse si difficile dans une langue étrangère? Les locuteurs d'une langue étrangère donnent l'impression de parler vite, de ne pas s'arrêter entre les mots, de ne pas bien articuler. Toutes ces impressions sont fondées, mais ces phénomènes sont également constatés dans notre langue maternelle, comme nous allons le voir dans le début de ce chapitre.

Nous poursuivrons ensuite avec une présentation de deux modèles fondateurs de la compréhension de l'oral en L1, avant d'étudier plus en détail certaines des caractéristiques des processus qui les composent : activation en parallèle, interactivité des processus descendants et ascendants, automatisation et fonctionnement prédictif.

1.1. Caractéristiques du langage parlé

Même en langue maternelle, la compréhension aurale constitue un ensemble de processus complexes, posant de nombreux défis sur le plan cognitif. D'après Weber et Scharenborg (2012, p. 387), la difficulté de la compréhension de l'oral, et en particulier de l'identification des mots dans le flux de la langue parlée, est due à trois facteurs. Le premier est la ressemblance des mots entre eux, découlant du fait qu'ils sont tous créés à partir d'un nombre très limité de phonèmes. Le deuxième est leur variabilité : selon les locuteurs et le contexte linguistique et extralinguistique, la réalisation d'un même phonème, d'un même mot ou d'une même phrase peut être très différente. La dernière complication (qui recouvre en fait deux caractéristiques) vient du caractère transitoire et continu de la parole, qui rend difficiles non seulement la récupération des sons qu'on n'aurait pas bien entendus, mais aussi la

segmentation en mots (et groupes de mots), dans la mesure où ces derniers ne sont pas clairement séparés. Nous commencerons par exposer plus en détail les difficultés liées au caractère continu de la parole, avant de nous intéresser à sa variabilité.

1.1.1. Continuité de la parole

Comme le rappelle Harley (2007, p. 236), nous pouvons produire, reconnaître ou comprendre jusqu'à 20 phonèmes par seconde, soit 4 ou 5 mots par seconde, ce qui confirme l'impression de vitesse souvent ressentie à l'écoute d'une langue étrangère. Ces mots ne sont en général pas séparés par des pauses ni des marqueurs de frontière aisément reconnaissables. La Figure 1.1 présente un exemple de spectrogramme pour la phrase *Take the yellow shoes* (prononcée assez lentement par un locuteur masculin d'anglais britannique, à destination d'un public débutant en anglais langue étrangère, et enregistrée à l'occasion de la conception d'un jeu pour des élèves d'école primaire), analysée par le logiciel Praat (Boersma & Heuven, 2001). Le panneau supérieur permet de visualiser l'onde sonore (en particulier la variation de son intensité au cours du temps), et la fenêtre médiane contient un spectrogramme de fréquence, qui permet de visualiser les variations de fréquence hertzienne. Les zones plus foncées sont les formants, qui indiquent à quelles fréquences est concentrée l'énergie à chaque instant. Ces formants et leurs mouvements permettent de caractériser les voyelles ainsi que les transitions entre voyelles et consonnes.

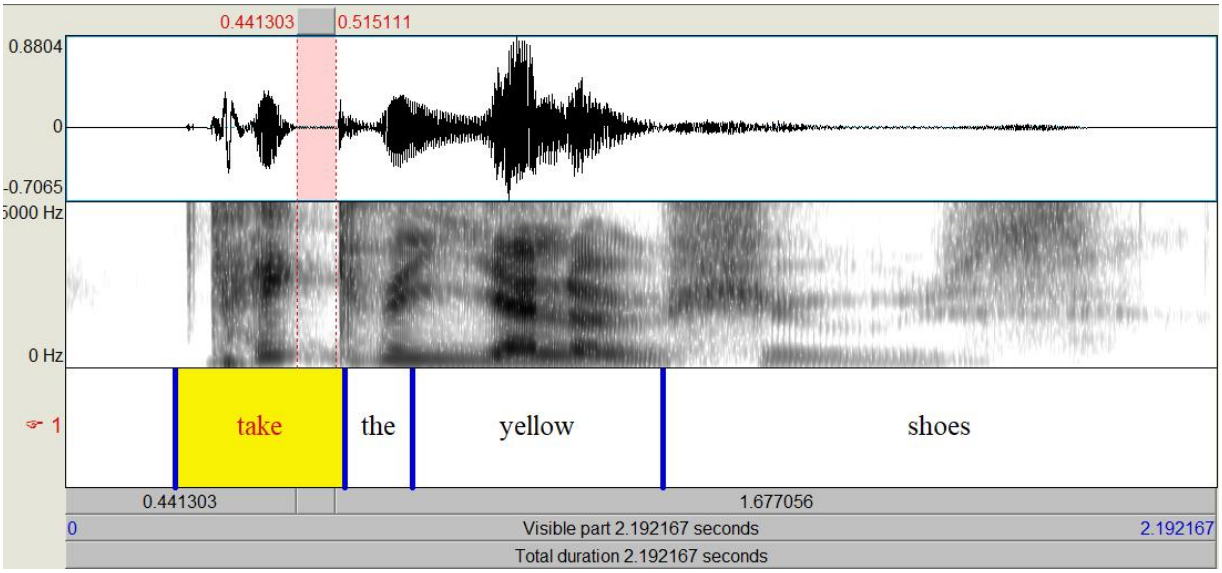


Figure 1.1 - intensité et spectrogramme de fréquence de *Take the yellow shoes* avec découpage en mots (logiciel *Praat*)

En l'espace d'un peu moins de 2,2 secondes (le temps total de l'enregistrement est indiqué dans la barre inférieure), 12 phonèmes sont prononcés : /t/ /eɪ/ /k/ /ð/ /ə/ /j/ /e/ /l/ /əʊ/ /ʃ/ /u:/

/z/ (il s'agit donc d'une vitesse assez lente, 6 phonèmes par seconde). On peut constater, grâce à l'annotation en mots rajoutée sous le spectrogramme de la Figure 1.1, d'une part que les mots ne sont pas séparés par des silences (par exemple entre *the*, *yellow* et *shoes*), et d'autre part que les silences (ou du moins les zones moins sonores) se trouvent parfois à l'intérieur des mots. A l'intérieur du mot *take*, par exemple, la plosion finale du /k/ est précédée d'une période de sept centièmes de seconde de silence (colorée en rose dans le haut du panneau) qui correspond au blocage de l'air à l'intérieur de la bouche (ou occlusion) avant la désocclusion: l'air est alors soudainement relâché et produit le son caractéristique du /k/, immédiatement suivi, sans silence cette fois, du début du /ð/ de *the*.

1.1.2. Continuité revisitée : indices suprasegmentaux et infrasegmentaux de frontières de mots

Cette absence apparente de frontières lexicales cache cependant une autre réalité enfouie dans le flux du signal sonore. Il existe bien des indices de frontières entre les mots, mais celles-ci ne sont pas matérialisées par des silences. Ces indices prennent différentes formes, ou sont plus au moins saillants, selon les langues. En anglais, il s'agit tout d'abord de l'accentuation : Cutler et Carter (1987) constatent par exemple que, dans un corpus de conversations, 90% des mots lexicaux commencent par une syllabe accentuée (c'est d'ailleurs le cas pour les mots lexicaux de notre exemple, *take*, *yellow* et *shoes*). L'accentuation est donc un premier indice (suprasegmental) relativement fiable de frontière lexicale. Dilley et McAuley (2008) montrent d'ailleurs que cet effet de la prosodie ne provient pas uniquement du contexte immédiat, mais aussi du rythme de la phrase dans son ensemble, c'est-à-dire de l'alternance de syllabes accentuées ou non dans les syllabes précédentes. Plus généralement, la prosodie donne également des informations sur la structure de la phrase : les coupures syntaxiques sont souvent marquées par une pause, et précédées de mouvements assez amples de la fréquence fondamentale (F0), qui correspond à l'intonation perçue (Cutler et al., 1997).

Un deuxième indice de frontière des mots provient des régularités phonotactiques de la langue considérée, c'est-à-dire des combinaisons de phonèmes qui sont acceptables, ou plus ou moins fréquentes, à l'intérieur des mots. Par exemple, la suite de deux consonnes /tk/ est rarement utilisée dans le lexique anglais (quelle que soit sa position dans le mot : L. White et al., 2012). Si un locuteur entend ces deux phonèmes à la suite, par exemple /hɒtkænən/, c'est donc qu'il y a très probablement une frontière de mot entre les deux (*hot cannon*). De même pour notre exemple, la suite /kð/ (*take the*) enjambe forcément une frontière de mot. Cette

information est appelée probabilité transitionnelle (*transitional probability*, ou *TP*). Les probabilités transitionnelles peuvent être calculées entre les phonèmes ou entre les syllabes d'une langue, et correspondent à la probabilité qu'un phonème (ou une syllabe) soit suivi(e) d'un(e) autre (Saffran, Newport, et al., 1996, p. 610). Les suites de phonèmes ou syllabes utilisé(e)s à l'intérieur des mots ont en général une probabilité transitionnelle plus élevée que celles qui enjambent deux mots. Par exemple, la probabilité que la syllabe /beɪ/ soit suivie de /bi/ est plus élevée que celle qu'elle soit suivie de /tu:/, parce que /beɪ bi/ forme le mot fréquent *baby*, tandis que /beɪ tu:/ n'est pas un mot. C'est donc une suite de syllabes moins fréquente, qu'on ne rencontrera que quand un mot se terminant par /beɪ/ sera suivi d'un mot commençant par /tu:/. Plusieurs études ont montré que les auditeurs d'une langue nouvelle (en général une langue artificielle créée en laboratoire pour les besoins de l'expérience) savaient rapidement utiliser ces informations pour repérer de nouveaux mots : ils arrivent à repérer les suites récurrentes de syllabes qu'ils ont entendues dans l'input qui leur était proposé (Saffran, Newport, et al., 1996). Jenny Saffran et ses collaborateurs ont également montré que ce mécanisme était utilisé par des enfants de huit mois en cours d'acquisition de leur langue maternelle (Saffran, Aslin, et al., 1996).

Le troisième indice que nous utilisons pour segmenter la parole découle du fait que les phonèmes n'ont pas tout à fait la même réalisation selon la position qu'ils occupent à l'intérieur des mots et des syllabes, mais peuvent être représentés par des variantes allophoniques[2]. Ce phénomène d'allophonie positionnelle peut aider à la segmentation du signal, une idée que défend Church (1987, p. 55) : « *I argue that allophonic variation is useful. When I find a flap, I know that it is foot internal (not initial in a stressed syllable); […] and when I find an aspirated stop, I know it probably starts a syllable.* ». Ainsi, une plosive (sourde) aspirée comme [t^h] sera probablement en début de mot (ou du moins de syllabe accentuée), et permettra de faire la différence entre *keeps talking*, où le /t/ initial de *talking* est aspiré, et *keep stalking*, où le /t/ est placé après /s/ et non en position initiale, et n'est donc pas aspiré. Par contre, un battement alvéolaire (*flap*) sera forcément à l'intérieur d'un mot, comme dans *later* où le /t/ est souvent prononcé /ɾ/ (du moins en anglais américain), ce qui ne sera jamais le cas en début de mot. D'autres exemples d'utilisation d'allophones positionnels en anglais sont le contraste entre /l/ clair en début de syllabe et /l/ sombre ([ɬ]) en fin de mot, et l'utilisation d'un coup de glotte avant une voyelle initiale (Nakatani & Dukes, 1977).

[2] En phonologie, les allophones sont les réalisations sonores possibles d'un phonème, qui peuvent varier selon le contexte (Lass, 1984).

Le quatrième indice de segmentation, qui est une sorte de variation allophonique, est l'allongement de la consonne initiale d'une part, et, d'autre part, l'allongement de la syllabe finale des mots, qui se traduit essentiellement par un allongement de la voyelle finale (D. K. Oller, 1973). Salverda et ses collègues ont montré que les locuteurs étaient capables de prendre en compte cette information, et qu'ils font la différence entre la syllabe *ham* du mot *ham* et celle du mot *hamster* (Salverda et al., 2003, sur le néerlandais). Ces phénomènes d'allongement semblent d'ailleurs universaux et non spécifiques à l'anglais, et correspondent peut-être à une tendance générale (qui n'est pas propre à la parole) de ralentir à la fin d'un mouvement musculaire (Vaissière, 1983).

Le signal sonore, même s'il est continu, contient donc bien des informations qui aident à le segmenter : le schéma accentuel, les probabilités transitionnelles entre phonèmes ou syllabes, et les phénomènes d'allophonie positionnelle (entre autres). Cependant, même si des études ont montré que les locuteurs sont capables de les exploiter dans des conditions d'écoute contrôlées (par ex., Tyler & Cutler, 2009), aucun de ces indices n'est complètement fiable, et ils ne suffisent pas à eux seuls pour découper le signal en mots. L'allongement de la voyelle, par exemple, est un indice non seulement de fin de mot, comme nous l'avons vu, mais aussi d'accentuation (donc plutôt un indice de début de mot en anglais), ou encore de voisement ou non de la consonne qui suit (une plosive sourde comme /t/ est précédée d'une voyelle moins longue qu'une plosive sonore comme /d/, ce qu'on peut constater dans le contraste *bat* / *bad*). Les indices que nous avons décrits sont ainsi la plupart du temps probabilistes et non univoques, et il n'est pas forcément facile d'interpréter l'information qu'ils apportent.

1.1.3. Variabilité de la parole

Outre le fait qu'il soit souvent difficile de séparer les mots, il existe une autre source de difficulté de décodage du langage parlé. D'après Johnson (2004), dans la conversation courante, plus de la moitié des mots subissent des modifications de phonèmes par rapport à leur forme de citation (la forme du mot quand il est prononcé seul et bien articulé). A partir d'un corpus de productions spontanées de 14 locuteurs américains, Johnson montre que près de 8% des mots lexicaux et 5% des mots grammaticaux perdent une syllabe. Par exemple, le mot *because* dans la suite *because if* peut être prononcé [k^hz] et passer de deux syllabes dans sa forme de citation, [bi'kɒz], à aucune syllabe, devenant ainsi un clitique servant d'attaque au mot suivant, *if*. Quant aux modifications de phonèmes par rapport à la forme de citation, 20% des mots lexicaux et 40% des mots grammaticaux en subissent au moins une (qu'il s'agisse ou

non d'une réduction). Au final, 60% des mots (*tokens*) de leur corpus ne sont pas produits sous leur forme de citation. Greenberg (1999) avait auparavant trouvé des chiffres similaires dans le corpus de conversation *Switchboard* : 22% des phonèmes y subissent une modification (le pourcentage de mots intacts n'est pas donné), le mot grammatical *and* est prononcé de 80 façons différentes, montrant ainsi une grande variabilité, et la variante de prononciation de *that* la plus courante (avec un /æ/ mais sans production de la consonne finale) ne couvre que 11% des instances dans le corpus (*tokens*).

Une des raisons de la variabilité de la parole provient des différences individuelles de l'appareil phonatoire des locuteurs, comme la taille du conduit vocal ou la densité des cordes vocales (Jusczyk, 1997, p. 7), qui produisent par exemple des voix plus ou moins graves. Un même locuteur peut également produire des instances assez différentes de la même syllabe, même dans des contextes similaires (Newman et al., 2001). Une autre source de variabilité est la vitesse d'élocution (*speech rate*). Liberman et ses collaborateurs ont montré dès 1956 qu'un ralentissement du tempo (de la transition entre la consonne et la voyelle) pouvait transformer un /bɛ/ en /wɛ/ (Liberman et al., 1956). Un même segment pourra donc être interprété différemment selon la vitesse du discours dans lequel il est intégré. L'influence de la vitesse est particulièrement sensible dans un contexte de production suivie et non de mots isolés : Klatt et Stevens (1973) soulignent que les segments y sont plus courts, que certains sons se téléscopent, ne sont pas réalisés pleinement ou disparaissent. Enfin, le phénomène de coarticulation est une source constante de variabilité. En effet, les sons de la langue ne sont pas prononcés séparément et l'appareil phonatoire est en train de finir la production du son précédent quand il commence la prononciation du suivant. Liberman et ses collègues notent par exemple que dans la parole continue normale, « *the acoustic signal at no point corresponds to the vowel alone, but rather shows, at any instant, the merged influences of the preceding or following consonant* » (Liberman et al., 1967, p. 440). Le même phonème a donc de nombreuses réalisations différentes en fonction des sons qui le précèdent et le suivent.

Cependant, ces phénomènes universaux ne nous empêchent pas de comprendre notre L1, à laquelle nous avons été exposés pendant des milliers d'heures[3]. Ce n'est qu'en L2 qu'ils nous gênent, faute d'expérience ou de connaissances suffisantes, et du fait de l'influence de notre

[3] Dans leur étude longitudinale du développement lexical d'enfants américains avant 3 ans, par exemple, Hart & Risley (2003, p. 8) considèrent qu'une année d'exposition à la langue d'un enfant correspond à 5200 heures; Hilton (2005, p. 13) cite 20 000 heures pour les 6 premières années de vie.

L1. Afin de tenter de séparer les sources de difficulté pour les locuteurs L2, d'identifier celles qui posent principalement problème, et donc d'y faire porter la majorité des efforts de remédiation, il est nécessaire d'avoir une vision globale du processus de compréhension de l'oral. C'est pourquoi nous allons décrire brièvement les différentes étapes proposées par les modèles actuels de la compréhension de l'oral en langue maternelle, avant de leur faire correspondre dans les chapitres suivants les outils diagnostiques et les pistes de remédiation en L2.

1.2. Deux modèles de la compréhension aurale en L1

Nous commencerons par présenter deux modèles consensuels et assez proches, ceux d'Anderson (1995), et de Cutler et Clifton (1999). Ces modèles sont déjà anciens, mais ce sont à notre connaissance les modèles les plus récents qui décrivent l'ensemble du processus de compréhension de l'oral, depuis la perception des sons de la langue jusqu'à la construction du sens du texte. D'autres modèles, moins complets mais plus spécialisés, se concentrent sur une seule des étapes du processus, comme la reconnaissance lexicale ou la construction d'un modèle de situation. Certains de ces modèles seront mentionnés par la suite afin de compléter ceux par lesquels nous commencerons. Enfin, ces modèles ayant été conçus pour expliquer le processus de compréhension de l'oral chez des natifs (L1), nous les illustrerons avec des études faites sur des populations natives. La compréhension aurale en L2 suivant les mêmes processus (par ex., Hopp, 2016, p. 25), nous les illustrerons dans le chapitre suivant avec des exemples tirés de la recherche en acquisition des langues étrangères, mais nous commencerons dès ce chapitre à réfléchir brièvement aux conséquences pour les apprenants L2 de certains mécanismes au fur et à mesure que nous les décrirons.

1.2.1. Le modèle d'Anderson (1995)

Le modèle de John Anderson (1995), qualifié par Lynch de « *dominant paradigm in listening comprehension* » (Lynch, 2010, p. 76), est un modèle assez général, adapté à la fois à la compréhension de l'écrit et de l'oral, et qui se divise en trois étapes : la perception (*perception*), l'intégration/ analyse syntaxique (*parsing*), et l'utilisation (*utilization*). La perception, dont les étapes sont peu détaillées pour l'oral, correspond aux opérations grâce auxquelles le cerveau parvient à extraire l'information du signal (sonore pour nous) et à y reconnaître des objets (des sons et des mots). L'intégration (qui comprend l'analyse syntaxique) consiste à créer une représentation mentale du sens combiné des mots reconnus à

l'étape de perception. Enfin, l'étape d'utilisation permet de faire les inférences nécessaires à une bonne compréhension du texte et à la création d'un modèle de situation (Kintsch & Van Dijk, 1978; Zwaan & Radvansky, 1998), qui est une représentation structurée de ce que l'auditeur (ou le lecteur) a compris. C'est ici aussi que les réactions éventuelles (dans le cas d'une interaction) sont planifiées.

Un modèle de situation s'affranchit des mots précis utilisés (représentation de surface) ainsi que du sens propositionnel des phrases, pour ne garder qu'une représentation schématique des événements importants, effectivement mentionnés ou inférés lors du processus de compréhension, et de leurs relations. Kintsch et ses collègues montrent par exemple que les lecteurs oublient très rapidement les phrases exactes qu'ils ont lues (après 40 minutes, ils ne savent plus distinguer entre les phrases exactes et leurs paraphrases), et que, après quelques jours, ils se rappellent assez bien du sens des phrases d'un texte mais en oublient tout de même la moitié. Par contre, ils se souviennent encore très bien du modèle de situation, c'est-à-dire des éléments principaux de l'histoire qu'ils ont lue (Kintsch et al., 1990). Zwaan et Radvansky (1998) expliquent qu'un modèle de situation comprend non seulement ces événements principaux, mais également leurs relations logiques et temporelles, ainsi que les acteurs et leurs buts. Pour eux, la finalité de la compréhension n'est pas d'arriver à une analyse linguistique complète des phrases qui constituent le texte, mais bien de comprendre un message : « *language is now seen as a set of processing instructions on how to construct a mental representation of the described situation* » (ibid., p.162). C'est pourquoi ce dont se souviennent les locuteurs est bien le modèle de situation et non les phrases précises utilisées pour y arriver. De fait, la langue est simplement l'outil que nous utilisons pour communiquer ces instructions de construction du sens.

Nous avons trouvé une schématisation intéressante du modèle d'Anderson (Nowrouzi et al., 2015, p. 264), que nous avons traduite en français ci-dessous (Figure 1.2). Cette représentation montre bien l'imbrication des opérations nécessaires. Comme nous le verrons plus loin, la perception (en particulier la segmentation lexicale) dépend en partie de l'intégration (l'analyse syntaxique) et même de l'utilisation (contexte pragmatique). Cependant, ce graphique n'est pas entièrement satisfaisant en ce qu'il ne permet pas de hiérarchiser les différents niveaux, alors qu'ils sont très clairement présentés comme (partiellement) successifs chez Anderson (1995, p.313) : « *These three stages—perception, parsing, and utilization—are by necessity partly ordered in time* ». En effet, la perception est

l'étape d'entrée dans le système et l'utilisation est celle de sortie, ce qui n'apparaît pas dans cette représentation, qui permet cependant une première tentative de modélisation graphique.

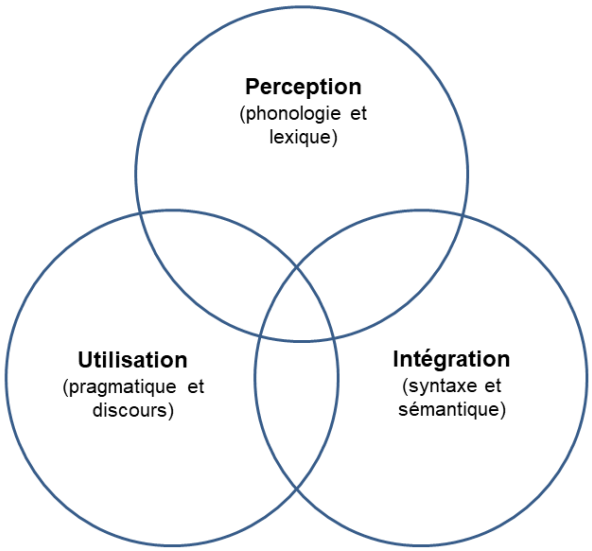


Figure 1.2 - tentative de représentation graphique du modèle d'Anderson (1995), d'après Nowrouski et al. (2015)

Ce premier modèle, conçu par un psychologue cognitiviste, reste un peu trop général (en particulier à l'étape de perception) pour répondre à notre objectif, qui est d'identifier les étapes du processus de compréhension qui risquent de poser des problèmes particuliers à nos apprenants. D'autre part, il est très influencé par la recherche sur la lecture et n'est pas particulièrement orienté vers la modalité orale. C'est pourquoi nous nous tournerons vers un modèle plus détaillé, et conçu spécifiquement pour l'oral par deux psycholinguistiques dont l'une (Anne Cutler) est plutôt spécialiste de reconnaissance lexicale et l'autre (Charles Clifton) de traitement des phrases.

1.2.2. Le modèle de Cutler et Clifton (1999)

Commençons par reproduire (Figure 1.3) le schéma proposé par Cutler et Clifton (1999) pour représenter ce qu'ils nomment dans le titre de leur chapitre *(A) Blueprint of the Listener* (nous avons réutilisé l'adaptation française de Hilton, 2009, p. 67). Ce modèle du processus de compréhension chez l'auditeur distingue quatre étapes, dont la première est le décodage (qui transforme le flux de la parole, après l'avoir séparé du fond sonore, en représentation abstraite, par exemple phonémique), et la deuxième la segmentation (qui découpe la suite de phonèmes en mots). Il comporte donc une étape supplémentaire par rapport à celui d'Anderson (1995), dans la mesure où il sépare le décodage de la segmentation, qui font tous

les deux partie de la perception chez Anderson. C'est, à notre avis, une façon de prendre en compte la spécificité de la compréhension de l'oral par rapport à celle de l'écrit, puisque, comme nous l'avons vu (section 1.1), les mots ne sont pas séparés à l'oral dans le flux de la parole alors qu'ils le sont à l'écrit (dans les écritures alphabétiques modernes). L'étape de segmentation présente donc des obstacles spécifiques qui requièrent des opérations particulières qui contribueront à la complexité de l'ensemble.

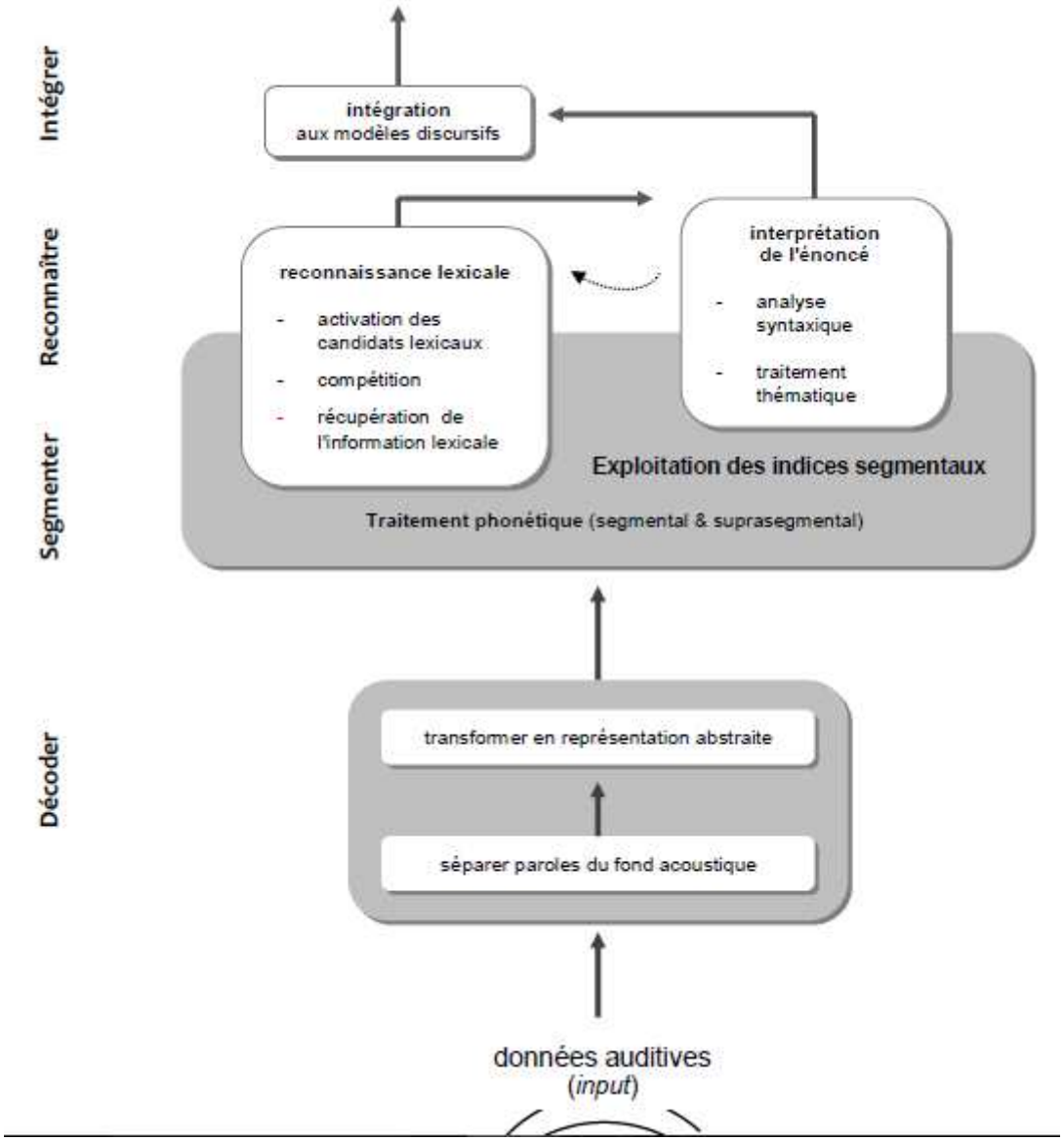


Figure 1.3 - modèle de la compréhension de l'oral, d'après Cutler et Clifton (1999), dans Hilton (2009, p.67)

La troisième étape du modèle de Cutler et Clifton est celle de la reconnaissance, celle du sens des mots et surtout de celui des phrases (elle correspond donc grossièrement à l'étape *parsing* d'Anderson), avant la quatrième étape d'intégration dans le modèle discursif (le modèle de situation, comme chez Anderson). L'étape de l'intégration discursive n'est pas détaillée chez

Cutler et Clifton – ils mentionnent simplement que les processus d'interprétation (et d'intégration) de l'énoncé sont les mêmes pour l'écrit et pour l'oral.

On constate que les étapes de ce modèle, qui utilisent les différents niveaux de traitement identifiés par les psycholinguistes (des sons aux mots, des mots aux phrases et des phrases au texte), convergent également avec les niveaux d'analyse linguistique traditionnels : phonétique et phonologie, puis morphologie, syntaxe et sémantique, et enfin pragmatique. Cependant, cette représentation (Figure 1.3) traduit aussi visuellement la difficulté à séparer les étapes de traitement : l'étape de segmentation, qui utilise en entrée la représentation abstraite obtenue en sortie de l'étape de décodage, a également besoin des informations de l'étape de reconnaissance lexicale (et même de l'analyse syntaxique, comme nous le verrons plus loin) pour fonctionner de façon satisfaisante. En effet, si l'auditeur peut s'aider pour la segmentation d'indices purement acoustiques venant du traitement du signal, c'est souvent insuffisant, et c'est parfois le fait d'avoir reconnu un mot qui indique que le suivant commence juste après : « *word segmentation may very well be a natural by-product of the recognition process itself* » (Pisoni & Luce, 1987, p. 39).

Cette intrication des niveaux de traitement est liée à deux phénomènes que nous examinerons tour à tour dans les paragraphes qui suivent : d'une part, le fait que ces étapes fonctionnent en parallèle et non de façon sérielle, et d'autre part, le fait que les différents niveaux interagissent en permanence.

1.3. Modèles sériels et modèles parallèles

Les modèles sériels du processus de compréhension aurale dans les années 1970 envisageaient une série d'opérations dont chacune devait être terminée avant que la suivante ne commence. Dans un tel modèle, tous les phonèmes doivent être reconnus avant que le mot le soit à son tour, il faut que le mot soit reconnu avant que son sens ne soit activé, tous les mots de la phrase doivent avoir été reconnus pour que l'analyse syntaxique commence, etc...: Forster (1976) supposait par exemple que « *the sentence cannot be processed until all the words have been accessed* ». Or, il a été clair assez tôt qu'un tel système ne pouvait pas fonctionner, comme nous allons le voir dans les paragraphes qui suivent.

1.3.1. Mots reconnus avant tous leurs phonèmes

Dans les années 1970, William Marslen-Wilson a commencé par montrer que les mots étaient souvent reconnus avant que l'interlocuteur ait fini de les prononcer, donc avant que tous les phonèmes aient été produits et reconnus. Il utilisait la technique du *shadowing* (répétition simultanée), où l'on demande à des locuteurs de répéter une phrase en même temps qu'ils l'entendent. Quand la phrase a un sens et que le contexte aide, il n'est pas rare que les locuteurs « répètent » un mot avant qu'il ait été complètement articulé, avec seulement un quart de seconde (250 ms) de latence entre le début d'articulation du mot entendu et celui du même mot répété (Marslen-Wilson, 1975). Ceci est d'autant plus vrai que son « point d'unicité » (à partir duquel il se distingue de tous les autres mots possibles commençant par les mêmes phonèmes) arrive tôt. Ces expériences ont ainsi montré qu'il n'était pas nécessaire d'avoir entendu tous les sons composant un mot avant de le reconnaître. Comment est-ce possible? L'hypothèse qui a fini par s'imposer est que dès le début de l'écoute, tous les mots compatibles avec ce début sont activés, sans attendre la fin du mot. S'il ne reste plus qu'un mot activé, il « gagne » le processus de sélection, même si sa production n'est pas terminée. Ce résultat a été confirmé par la suite par des études utilisant la méthode de l'oculométrie (*eye-tracking*) à partir des années 1990 (Allopenna et al., 1998).

1.3.2. Mot suivant reconnu avant ou en même temps que le précédent

Même si la communauté scientifique a rapidement compris qu'un mot pouvait être reconnu avant sa fin, on pensait toujours que la reconnaissance de chaque mot dépendait de celle du mot précédent. Dans la première version du modèle de reconnaissance lexicale développé par Marslen-Wilson, *Cohort*, par exemple, c'est le début du mot qui enclenche le processus d'activation lexicale nécessaire à la reconnaissance (Marslen-Wilson & Tyler, 1980). Si le moment où le mot commence n'est pas connu (parce que le mot précédent n'ayant pas encore été reconnu, on ne sait pas où il se termine), ou si le début n'a pas été bien entendu, le modèle ne peut pas fonctionner et le mot n'est pas reconnu.

C'est François Grosjean qui a montré dans les années 1980 (Grosjean, 1980, 1985) qu'un mot pouvait tout à fait être reconnu sans que le précédent le soit. La technique qu'il utilise, le *gating* (le paradigme du dévoilement successif par « portes », ou paliers temporels), consiste à faire entendre un début de mot ou phrase, à demander au sujet d'écrire le(s) mot(s) qu'il pense avoir entendu(s) ainsi que son degré de certitude, et à continuer le procédé en rajoutant

quelques secondes (une nouvelle « porte ») à chaque fois. Grosjean (1985) a trouvé que bon nombre de mots monosyllabiques, surtout quand ils n'ont pas de point d'unicité (c'est-à-dire qu'ils sont contenus comme début dans d'autres mots, comme *plum* dans *plumber* ou *plummet*), sont reconnus après leur fin (leur « point d'isolation » se trouve après leur coda), en même temps que le mot suivant est reconnu (voire deux mots plus tard, pour *doe* par exemple). La reconnaissance lexicale dans ces cas-là n'est pas séquentielle mais bien simultanée. Deux mots sont reconnus d'un coup quand on se rend compte que ce qu'on pensait être un seul mot ne donne pas une interprétation possible. Dans une étude de la même époque, Luce (1986a) montre d'ailleurs que seuls 40% des mots ont un point d'unicité avant leur fin[4]. Le phénomène mis en lumière par François Grosjean n'est donc pas rare.

1.3.3. Analyse d'une phrase avant sa fin

Si l'activation lexicale n'attend pas la reconnaissance de tous les phonèmes du mot avant de se déclencher, et si la reconnaissance lexicale de plusieurs mots peut se faire en même temps (celle du suivant n'attendant pas celle du précédent), il est intéressant de se demander ce qu'il en est de l'analyse syntaxique. En même temps qu'elles montraient que les mots pouvaient être reconnus avant leur fin, les études de Marslen-Wilson ont également fait comprendre que l'analyse syntaxique se faisait au fur et à mesure de l'écoute. Dans une de ses premières expériences (Marslen-Wilson, 1975), certaines des phrases qui devaient être répétées contenaient une anomalie lexico-syntaxique (*He's afraid he forgot to put a stamp on the already before he posted it* ; avec *already* au lieu de *envelope*), et certaines de ces anomalies étaient corrigées en ligne lors du *shadowing* par les sujets testés, montrant que l'analyse syntaxique avait commencé et était en cours bien avant la fin de la phrase.

Ce résultat a ensuite été confirmé par d'autres études sur les phrases ambiguës de type *garden path* (dont l'exemple classique est *The horse raced past the barn fell*), où la relative réduite (*raced past the barn*) est d'abord analysée comme la continuation de la principale : *The horse* (S) *raced* (V) *past the barn* (complément circonstanciel), avant de devoir être réanalysée au moment où le verbe principal *fell* est entendu (ou lu). Speer et al. (1996) ont montré qu'à l'oral, c'est la prosodie qui aide à la désambiguïsation. Ils utilisent des phrases comme *Whenever the guard checks the door it's/is locked*, où le deuxième groupe nominal (*the door*)

[4] : Il s'agit cependant d'une étude à partir des mots isolés du dictionnaire, sans tenir compte des indices de frontières de mots contenus dans le signal sonore et décrits en début de chapitre, ni des effets potentiellement facilitateurs du contexte.

peut être le sujet du verbe *is* (*Whenever the guard checks, the door is locked)*, ou bien l'objet du verbe *checks* (*Whenever the guard checks the door, it's locked*). Ils montrent, avec une expérience de dénomination intermodale oral/écrit, que si le schéma prosodique est compatible avec la syntaxe (par exemple, si les sujets entendent *Whenever the guard checks % the door* avec un allongement de la syllabe de *checks* pour indiquer la fin de la subordonnée initiale), le mot qui suit la partie potentiellement ambiguë (*is* dans notre exemple), présenté à l'écrit, est lu significativement plus rapidement que si le schéma prosodique est incompatible (allongement de la syllabe de *door*). L'analyse syntaxique, aidée par la prosodie, est donc déjà en cours avant la fin de la phrase

1.3.4. Analyse du schéma intonatif de la phrase

Le schéma intonatif de la phrase n'est cependant pas toujours lié à sa structure syntaxique. Encore une fois, les premières études supposaient que les intonations signalant des inférences pragmatiques étant difficiles à traiter pour l'auditeur, celles-ci prenaient du temps et étaient gardées pour la fin, alors que les implicatures conversationnelles « classiques » étaient pré-traitées avant la fin de la phrase (Levinson, 2000). C'est ainsi qu'encore récemment, Dennison et Schafer (2010) concluent à propos du traitement d'un accent contrastif sur le verbe en anglais (*Lisa had* vs. *HAD the bell*, selon qu'elle l'a peut-être encore ou non) que ce n'est qu'à la fin de la phrase que le sens négatif est compris.

Cependant, Kurumada et ses collaborateurs ont montré, en utilisant une expérience d'oculométrie de type « univers visuel » (*visual world*, où les référents possibles sont représentés à différents endroits de l'écran), qu'au contraire une intonation « pragmatique » pouvait être prise en compte immédiatement (Kurumada et al., 2014). Selon l'accent utilisé dans *It looks like a zebra*, on peut comprendre soit que c'est probablement un zèbre (intonation normale avec accent sur le nom en fin de phrase), soit au contraire que ça y ressemble mais n'en est en fait pas un (accent montant sur le verbe : *It LOOKS like a zebra (…but isn't)*). Dans ce deuxième cas et pas dans le premier, les auditeurs commencent tout de suite après avoir entendu le verbe, et donc avant d'avoir entendu le nom *zebra*, à porter leurs regards vers la paire d'animaux semblables (par exemple un zèbre et un okapi) plutôt que vers les deux autres animaux qui ne se ressemblent pas. Ils prennent donc tout de suite en compte, sans attendre la fin de la phrase, l'information apportée indirectement par l'intonation inhabituelle sur le verbe *looks (like)*, à savoir qu'il faudra se méfier parce que cela ressemble à quelque chose mais n'en est pas vraiment une.

1.3.5. Analyse de plusieurs propositions

Après les phonèmes, les mots et les phrases, la question de la temporalité du traitement s'est posée dans les années 1990 à propos de l'intégration du sens de plusieurs propositions. Millis et Just (1994), par exemple, proposaient, en contexte de lecture, que l'analyse de deux propositions reliées par *because* ne pouvait avoir lieu qu'après la fin de la deuxième proposition, entre autres parce que le temps de lecture augmente en général à la fin des propositions et des phrases et qu'ils supposaient que c'est là que se faisait l'intégration du sens. Cependant, les études d'oculométrie portant sur la compréhension en lecture (Traxler et al., 1997) ont au contraire montré que l'intégration se faisait en ligne, avant la fin de la deuxième proposition. A partir de la constatation que les phrases où *because* a un sens causatif (*Heidi felt very proud and happy because she won first prize at the art show*) sont plus faciles à comprendre (et plus courantes) que les phrases avec un *because* de déduction logique, qu'ils appellent « diagnostique » (qui s'apparente à *since*: *Heidi could imagine and create things because she won first prize at the art show*), ils ont montré que le retard pouvait être observé peu après la conjonction *because*, vers le milieu de la subordonnée, et donc avant la fin de la phrase.

Ces observations rejoignent les intuitions de Marslen-Wilson et Tyler, qui pensaient déjà que les résultats de leurs expériences plaidaient en faveur de l'existence d'un système de traitement en ligne : « *an on-line interactive language processing theory, in which lexical, structural (syntactic), and interpretative knowledge sources communicate and interact during processing in an optimally efficient and accurate manner* » (Marslen-Wilson & Tyler, 1980, p. 1). Les bases d'un système de traitement dynamique et interactif étaient donc posées, combinant toutes sortes d'informations dès qu'elles sont disponibles, qu'elles soient lexicales, syntaxiques ou sémantiques.

1.3.6. Avantages de l'analyse en parallèle

Quel peut être l'avantage pour le fonctionnement du cerveau d'un fonctionnement en parallèle plutôt que sériel ? Le fait d'activer en même temps toutes sortes de possibilités dont la plupart vont se révéler fausses s'apparente à première vue à un énorme gaspillage de ressources cognitives.

Cependant, Libben et Jarema (2007) démontrent très bien (à propos du traitement des mots composés) l'intérêt d'une telle architecture qui a mis du temps à s'imposer dans la communauté scientifique. Dans les modèles « élégants » recherchés à partir des années 1970, la redondance était bannie : tout ce qui pouvait être généralisé par une règle devait l'être. C'est encore le principe suivi par l'école générativiste et notamment par Steven Pinker. Dans son ouvrage *Words and Rules* (Pinker, 1999), le fonctionnement des verbes réguliers, explicable par une règle simple, doit faire l'objet d'un traitement séparé des verbes irréguliers, dont les formes sont apprises et reconnues séparément. Dans cette lignée de théorisation « élégante », on peut imaginer deux modes de fonctionnement opposés. Le premier serait un fonctionnement qui maximiserait l'efficacité de stockage : dans ce cas-là, on stocke le moins d'éléments possibles (uniquement les morphèmes, non décomposables en unités de sens plus petites), mais il faut recalculer à chaque fois tout ce qui est multimorphémique, d'où un coût élevé en calcul (c'est l'option de Steven Pinker, chez qui le passé des verbes réguliers n'est pas stocké, et est donc recalculé à chaque fois). A l'opposé, on peut imaginer un fonctionnement qui maximise l'efficacité de calcul : on ne calcule que ce qui est absolument nécessaire (par exemple ce qui n'a jamais été rencontré), et tout le reste est stocké, d'où un coût élevé en stockage.

Le fonctionnement qui semble prévaloir est différent : le cerveau privilégie non pas l'absence de gaspillage de place ou de calcul, mais au contraire stocke tout ce qui peut l'être et calcule également tout ce qui peut l'être, selon un principe de redondance maximale :« *activating everything possible appears to be the easiest and most generally applicable mental architecture [...] under the conditions of uncertainty that characterize word recognition* » (Libben & Jarema, 2007, p. 10). Cette redondance confère au système deux avantages : une grande vitesse de traitement, et une grande robustesse (« *the system remains relatively crash-proof by being able to extract all that it can* », ibid.., p.12).

Dans un tel système, tous les indices, quel que soit leur niveau, sont exploitables, et le cerveau n'a aucune « décision » à prendre à l'avance sur ce qu'il va rencontrer : le système est prêt à traiter tout ce qui lui arrive, sans avoir besoin de préconçu sur ce qui lui sera le plus utile pour décoder. Il y a rarement besoin de revenir en arrière en cas d'erreur de reconnaissance ou d'interprétation. En effet, les autres possibilités plus ou moins compatibles avec le signal ou le contexte sont toujours plus ou moins activées, même si elles le sont moins que la solution préférée à un instant donné. Quand cette solution préférée s'avère finalement peu probable,

son activation peut diminuer et celle des autres possibilités peut remonter avec une perte de temps minimale, sans qu'il y ait besoin de recommencer l'analyse à zéro.

1.3.7. Conséquences pour la compréhension en L2

Quelles pourraient être les conséquences d'une telle cascade d'opérations quasi simultanées pour les apprenants de L2 en situation de compréhension? On peut s'attendre à ce que cette architecture leur pose deux problèmes principaux : d'un côté, la quantité d'informations à traiter, et de l'autre, la coordination entre les différents niveaux de traitement qui s'activent (ou devraient s'activer) en cascade.

La quantité d'éléments à traiter peut s'avérer d'autant plus coûteuse que les apprenants de langue ont déjà tendance à garder activées trop de possibilités, même quand elles ne correspondent pas avec le signal. En effet, ils sont moins certains que les natifs que leurs interprétations sont correctes, et par ailleurs, ils tendent à s'accrocher à leur première interprétation : il leur faut beaucoup d'indices allant dans un autre sens avant qu'ils acceptent de l'abandonner (J. Field, 2004). Ce caractère coûteux permet peut-être d'expliquer la difficulté qu'ont certains apprenants à coordonner plusieurs niveaux à la fois (par exemple, la reconnaissance lexicale et l'analyse syntaxique). Dans certaines études où les apprenants tiennent un journal d'apprentissage (Goh, 2000), il arrive qu'ils aient l'impression de reconnaître les mots (niveau de la reconnaissance lexicale), mais de ne pas pouvoir construire en même temps une représentation syntaxique : « *When I was listening to an English song tape, I could catch most words. But I could not put all the words into a full sentence to get a full idea.* » (ibid., p.64). Certains sujets dans la thèse de Nawel Zoghlami (2015) ont exprimé le même ressenti lors d'un protocole de réflexion à haute-voix (*think aloud*). Une des compreneuses faibles de l'étude arrive à reconnaître les mots de l'extrait qu'elle entend, mais des problèmes de saturation de sa mémoire de travail l'ampêchent d'aller plus loin et de construire le sens de la phrase (ibid., pp. 229-230).

1.4. Interactivité entre processus ascendants et descendants

Outre la question du déroulement temporel de l'accès aux différents niveaux de traitement du signal sonore (dont nous venons de voir qu'il se fait en cascade dès que les informations sont disponibles), se pose la question de la relation et de l'interaction entre les différents niveaux, ainsi que de la direction de la transmission des informations d'un niveau à l'autre. Il semble

clair que les niveaux « inférieurs » (c'est-à-dire l'étape de perception chez Anderson, 1995, ou de décodage chez Cutler et Clifton, 1999) transmettent leurs informations aux niveaux « supérieurs » (respectivement, l'étape d'intégration/analyse syntaxique ou celle de segmentation). Ainsi, le niveau de traitement lexical (reconnaissance des mots et segmentation) a besoin des informations du niveau de traitement du signal sonore (reconnaissance des sons) afin de fonctionner. Ce type de transmission est appelé « ascendant » selon une métaphore qui associe la simplicité/ petitesse (les sons) au « bas » et la complexité/grandeur (les mots, puis les phrases, puis les textes) au « haut » (J. Field, 2004).

De nombreuses interrogations et recherches portent sur l'existence et le rôle des processus descendants, c'est-à-dire l'utilisation d'informations provenant des niveaux « supérieurs » (le texte, la phrase, le mot), pour aider à décoder un niveau « inférieur » (par exemple, les sons que l'on est en train d'entendre), ainsi que sur l'interaction entre les deux directions de transmission des informations. Nous commencerons par clarifier l'utilisation des termes « processus ascendants » et « descendants », avant d'exposer les résultats des expériences de psycholinguistique qui ont permis de mettre en évidence l'existence de processus descendants. Nous terminerons par les connaissances actuelles sur l'interaction entre les deux types de processus.

1.4.1. Processus descendants et ascendants (*top-down* et *bottom-up*)

Bien que ce premier chapitre soit essentiellement consacré à la compréhension en L1, nous ferons ici un détour par une acception des termes *bottom-up* et *top-down* assez répandue chez les didacticiens des langues étrangères, que nous contrasterons ensuite avec celle proposée par les psycholinguistes. Vandergrift et Goh (2012) considèrent que les processus ascendants (*bottom-up*) font essentiellement appel aux connaissances linguistiques :

> *This component of listening [bottom-up processing], seen as a decoding process, assumes that the comprehension process begins with information in the sound stream, with minimal contribution of information from the listener's prior knowledge of the world. Listeners draw primarily on linguistic knowledge, which includes phonological knowledge (phonemes, stress, intonation, and other sound adjustments made by speakers to facilitate speech production), lexical knowledge, and syntactic knowledge (grammar) of the target language.* (Vandergrift & Goh, 2012, p. 18)

Lors d'un traitement descendant, par contre, ce sont plutôt les connaissances extralinguistiques qui sont utilisées :

> *Listeners who approach a comprehension task in a top-down manner use their knowledge of the context of the listening event or the topic of a listening text to activate a conceptual framework for understanding the message. […]This top-down component of listening, seen as an interpretation process, assumes that comprehension begins with listener expectations about information in the text and subsequent application of appropriate knowledge sources to comprehend the sound stream.* (ibid., p.18)

Ici, c'est donc essentiellement le type de connaissances utilisées qui distingue *bottom-up* de *top-down* : connaissances linguistiques pour les processus ascendants et autres connaissances (situation d'énonciation et connaissances du monde) pour les processus descendants. Nous notons également l'association entre processus descendants et attentes de l'auditeur, nées ici des schémas préalablement acquis.

Field (2004) propose une définition peut-être plus rigoureuse de ces termes, et mieux alignée sur la recherche en psycholinguistique, dans la mesure où il considère que tout processus qui fait appel à un niveau supérieur (des unités plus grandes) est un processus descendant. Le fait d'utiliser ses connaissances lexicales pour compenser la mauvaise perception de certains sons est donc un exemple de processus descendant. Pour Vandergrift et Goh (2012), par contre, ce serait un processus ascendant (*bottom-up*), dans la mesure où il s'agit de l'utilisation de connaissances linguistiques. Dans cette conception plus psycholinguistique illustrée par Field, c'est moins le type de connaissances qui est important que la distinction entre informations contenues dans le signal, et informations venant des connaissances du sujet (qu'elles soient linguistiques ou encyclopédiques). C'est au moment de l'intégration de ces deux sources d'information que la compréhension a lieu :

> *The matching process can take its point of departure either in the input or in the recipient's knowledge. In the first case, information extracted from the input is integrated with increasingly complex knowledge systems (input-driven or bottom-up processing). In the second case, possible meaning is predicted on the basis of prior experience, the input being interpreted in the light of such expectations (knowledge-driven or top-down processing). In both cases, contextual information is utilized to support the comprehension process*. (Færch & Kasper, 1986, p. 264);

> *Perception occurs when bottom-up and top-down knowledge sources bind into a stable state.* (Goldinger & Azuma, 2003, p. 307)

C'est avec cette acception que nous utiliserons les termes *bottom-up* et *top-down* dans cette thèse. Nous utiliserons également les termes « processus de bas niveau » et « processus de haut niveau » avec le sens que leur donne Field (2004). Il propose en effet d'utiliser *lower-level processing* pour le décodage de ce qui se trouve dans le flux de la parole (le premier

niveau chez Cutler et Clifton), et *higher-level processing* pour ce qui a trait à la construction et l'interprétation du sens. On peut penser que la reconnaissance lexicale sert de pont entre les deux niveaux de processus : c'est le lieu où se rejoignent la forme (perçue auditivement) et le sens. Les processus de bas niveau sont parfois appelés « processus formels » justement parce qu'ils ont trait au traitement de la forme (Hilton, 2019).

1.4.2. Exemples de processus descendants (*top-down*)

1.4.2.1. influence des mots et phrases sur la reconnaissance des phonèmes

En psycholinguistique, de nombreuses recherches ont porté sur l'influence du niveau lexical sur la perception des sons. L'un des premiers résultats constatés a été le phénomène de « récupération de phonème » (*phoneme restoration effect*), mis en évidence par Warren et Warren (1970). Confrontés à la phrase *The state governors met with their respective legi*latures convening in the capital city*, où le /s/ de *legislatures* a été remplacé par un bruit de toux, les auditeurs sont incapables d'indiquer le son qui manque (ils entendent la toux, mais ne savent dire à quel endroit précis elle a lieu). Ils ont ainsi recréé, ou « récupéré », le /s/ manquant, grâce au contexte lexical : en effet, le contexte /ˈlɛdʒɪ--leɪtʃər/ est totalement contraint dans cet énoncé, et la seule possibilité pour le son manquant est /s/. Ce phénomène est observé même quand le contexte désambiguïsant se trouve plus tard dans la phrase : quand les sujets entendent la phrase *It was found that the *eel was on the_____ axle/ shoe/ orange/ table*, ils entendent respectivement (comme phonème initial du mot finissant par /-i:l/), /w/, /h/, /p/ et /m/, pour former les mots *wheel, heel, peel* et *meal*. L'absence d'un son passe encore une fois inaperçue grâce au contexte lexical environnant.

Un autre effet lexical est le *word superiority effect* qui fait qu'un phonème est détecté plus rapidement s'il est dans un contexte lexical. Ce phénomène avait déjà été démontré pour la lecture (une lettre est reconnue plus rapidement si elle se trouve dans le contexte d'un mot, Wheeler, 1970). En 1976, Philip Rubin et ses collègues montrent que c'est également vrai pour les phonèmes, du moins à l'initiale (Rubin et al., 1976). Fort et al. (2010) étendent ce résultat au contexte audiovisuel : dans une tâche de détection de phonème (*phoneme monitoring*), /p/ est détecté plus vite dans « chapeau » que dans « chapu », non seulement quand le mot est entendu, mais encore plus si on voit quelqu'un le prononcer. Ici encore, le contexte lexical semble aider à la reconnaissance d'une unité plus petite (le phonème).

En 1980, William Ganong constate que dans un contexte de consonne initiale ambiguë entre voisée et non voisée, les auditeurs ont une préférence pour une interprétation lexicale : ils entendront *task* plutôt que *dask* si ils sont confrontés à un stimulus qui est acoustiquement à mi-chemin entre les deux formes (c'est-à-dire si le voisement de la consonne initiale commence au milieu de l'intervalle entre le début typique pour un /d/ et le début typique pour un /t/). Le cerveau semble, en quelque sorte, préférer entendre un mot plutôt qu'une suite de sons qui n'a pas de sens. Comme le remarque Johnson (2011, p. 133), « *Listeners are inexorably drawn into hearing words even when the communication process fails. This makes a great deal of sense, considering that our goal in speech communication is to understand what the other person is saying, and words […] are the units we trade with each other when we speak* ». Cependant, quand les données acoustiques sont claires, par exemple si *dask* est prononcé de façon non ambiguë, avec un voisement qui commence tôt, c'est ce que les auditeurs disent entendre (Ganong, 1980). Les informations descendantes ne jouent ainsi que dans les cas où il existe une incertitude, et non quand les informations du signal sont claires.

1.4.2.2. influence du contexte phrastique sur la reconnaissance des mots

Un autre exemple d'effet descendant (*top-down*), après l'effet des connaissances lexicales sur la reconnaissance des sons, est celui du contexte de la phrase sur la reconnaissance lexicale, qui a été démontré avec trois types de tâches. Dès 1951, George Miller et ses collègues remarquaient qu'il est plus facile de comprendre des mots dans une phrase (en particulier en présence d'un bruit de fond) que les mêmes mots assemblés de façon aléatoire (Miller et al., 1951). Marslen-Wilson et Tyler (1980) trouvèrent ensuite le même résultat avec une tâche de *word monitoring* (repérage d'un mot) : nous détectons beaucoup plus rapidement un mot quand il se trouve dans une phrase plutôt que dans une suite de mots dans le désordre.

C'est aussi en 1980 que François Grosjean montre, en utilisant la technique expérimentale du *gating* présentée en 1.3.2, qu'un mot hors contexte (*camel*) est reconnu moins rapidement que quand il est inséré dans un contexte court (*the kids rode on the… camel*). Un contexte long (*At the zoo, the kids rode on the… camel*) conduit à un temps de reconnaissance encore plus court. Nous avons également déjà décrit les résultats de Grosjean (1985) qui montraient que certains mots n'étaient reconnus qu'en même temps que le mot suivant, ou même après lui, grâce au contexte supplémentaire apporté par ce(s) mot(s). Bard et al. (1988) ont ensuite étendu ce résultat avec des stimuli naturels tirés d'un corpus de productions spontanées. 20% des mots ne sont alors reconnus qu'après le suivant. Les mots grammaticaux sont

particulièrement susceptibles de ne pas être reconnus tout de suite, probablement parce qu'en anglais (en plus d'être courts et non accentués) ils précèdent les mots sur lesquels ils portent. Ainsi, après le verbe *eat* (contexte gauche), on peut trouver toutes sortes de prépositions, mais si le mot lexical suivant est *(a) spoon*, cela restreint les possibilités à *with*, ou peut-être *without*, ou *from*. Il n'est donc pas surprenant que le contexte droit aide à la reconnaissance.

Le contexte phrastique influence également la reconnaissance des morphèmes grammaticaux. Tuinman et al. (2014) ont montré, avec des sujets néerlandophones, que le suffixe verbal néerlandais de troisième personne /t/ était plus facilement reconnu si le verbe était accompagné d'un pronom de troisième personne. Ce pronom peut être placé avant ou après le verbe, ce qui fait qu'encore une fois, cela peut être le contexte droit qui aide à la reconnaissance.

1.4.2.3. influence d'informations extralinguistiques

Terminons par des exemples d'effets descendants qui correspondent probablement plus à la conception que nous avons qualifiée de « didactique » des procédés *top-down*, c'est-à-dire l'utilisation de connaissances extralinguistiques (visuelles, culturelles, etc…) pour aider à la compréhension.

Au niveau textuel (à l'écrit), Kintsch et Greene (1978, p. 2) ont montré l'importance des schémas culturels dans la compréhension des textes. Ces schémas amènent les locuteurs à anticiper la structure des textes qu'on leur donne à comprendre. Par exemple, dans un conte de tradition européenne, on sait que « *the events in the story must be causally and temporally related; […] stories contain episodes, each consisting of three story categories, which frequently are called exposition, complication, and resolution* ». Un conte qui s'inscrit dans cette tradition (un conte de Grimm, par exemple) sera ainsi mieux compris et retenu qu'un autre provenant d'une autre tradition (un conte traditionnel d'Alaska dans leur exemple). Le même procédé est à l'œuvre dans d'autres genres textuels bien définis culturellement, comme l'exposé scientifique (Kintsch & Van Dijk, 1978).

Cependant, les informations d'origine extralinguistique ne sont pas utilisées uniquement aux hauts niveaux du processus de compréhension. Les informations visuelles, en particulier, peuvent intervenir dès l'étape de décodage. Strand et Johnson (1996) montrent par exemple qu'une même syllabe, prononcée par un même locuteur, peut être interprétée comme commençant par un /s/ si elle est accompagnée d'un visage d'homme, mais comme un /ʃ/ si

c'est un visage de femme. En effet, les hommes ont en général une voix plus grave (utilisant des fréquences moins élevées) que les femmes. Le son /s/ ayant son énergie concentrée à des fréquences plus élevées que celle de /ʃ/, un son ambigu entre les deux (utilisant des fréquences intermédiaires) pourra être considéré comme « aigu » pour un homme (et donc correspondre à un /s/), mais « grave » pour une femme (un /ʃ/).

Le contexte paralinguistique et les informations visuelles peuvent aussi jouer au niveau de la reconnaissance lexicale : Casasanto (2008) montre que l'ethnicité du locuteur peut influencer la reconnaissance des mots anglais prononcés avec ou sans simplification du groupe consonantique final (*mass* vs. *mast*). Arnold et ses collaborateurs constatent qu'une hésitation avant un nom conduit les auditeurs (dans une expérience de monde visuel) à supposer que le locuteur se réfère à un nouveau référent, et c'est sur ce nouveau référent que se porte leur regard, dès le tout début de production du mot (Arnold et al., 2004). La même équipe (Arnold et al., 2007) montre ensuite que si l'on informe les auditeurs que le locuteur a un problème cognitif qui le conduit à chercher ses mots, cet effet disparaît : les auditeurs n'attribuent plus à cette hésitation la fonction de ciblage de nouveauté, s'attendant autant à entendre un ancien qu'un nouveau référent après la pause. Encore une fois, les sujets sont capables de prendre en compte toutes les informations, linguistiques ou non, présentes dans la situation, pour interpréter un message de façon efficace.

1.4.3. Exemples d'interaction de contraintes

Etant donné que les locuteurs L1 utilisent à la fois des processus ascendants et descendants lors de la compréhension de l'oral, on peut se demander quel est le rôle respectif de ces deux types de processus. C'est la question à laquelle Sven Mattys et son équipe (Mattys et al., 2005; L. White et al., 2010, 2012) ont essayé de répondre dans le cadre de la segmentation lexicale, c'est-à-dire le découpage du flux sonore en mots. La segmentation est la deuxième étape du modèle de Cutler et Clifton (1999), entre le décodage et la reconnaissance, c'est une étape qui est spécifique au traitement de la parole, et c'est également celle qu'il est le plus difficile de séparer des autres étapes, en particulier des étapes qui lui font suite. C'est donc un terrain privilégié pour observer l'interaction des différentes sources d'information.

Mattys et ses collaborateurs mettent en regard deux types de modèles de segmentation et de reconnaissance lexicale : d'une part, les modèles qui, outre les informations phonémiques, n'utilisent que les informations lexicales, et d'autre part, ceux qui insistent sur les

informations présentes dans le signal (en plus des informations phonémiques) qui peuvent aider à la segmentation (Mattys et al., 2005). Les premiers, parfois appelés *segmentation by lexical substraction*, supposent que la segmentation est le résultat du processus de compétition lexicale, qui met en concurrence toutes les possibilités de découpage du signal. Confrontés à la suite de phonèmes /hikɔ:ldɪmi:diətli/, par exemple, un auditeur anglophone trouvera le découpage *he called immediately* sans avoir besoin d'utiliser d'indices infra-phonémiques, simplement parce que c'est la seule façon possible de rendre compte de tous les phonèmes qui la composent en utilisant des mots existants. Un exemple de ce genre de modèle est le modèle implémenté TRACE (McClelland & Elman, 1986). Dans TRACE, les différents mots ou suites de mots compatibles avec le signal sont activés en parallèle jusqu'à ce que l'un deux « gagne » le processus de sélection parce que son activation est la plus haute ou parce que c'est le seul qui reste à la fin du processus. Le deuxième type de modèle (par exemple, Christiansen et al., 1998) utilise plus directement les indices présents dans le signal que nous avons présentés au début de ce chapitre, à savoir les indices prosodiques, les indices phonotactiques et les indices allophoniques (infra-phonémiques). Dans ce cas-là, il n'est pas forcément nécessaire de connaître les mots pour être capable de découper le signal ; il suffit de reconnaître les indices de segmentation. Ces modèles sont donc plus ascendants (*bottom-up*), alors que les premiers sont d'inspiration plus descendante (*top-down*).

White et al. (2012) comparent l'utilisation des connaissances lexicales (processus descendants) et des informations phonotactiques provenant du signal (processus ascendants) et constatent la primauté des informations lexicales. Lors d'une expérience d'amorçage intermodal, leurs sujets reconnaissent plus vite, et donc segmentent plus facilement, le mot *bag* dans le mot composé *plastic bag* (un composé fréquent) que dans *garlic bag* (non attesté), alors que les indices phonotactiques sont identiques (les deux mots sont séparés à la frontière /kb/). Par contre, ils ne reconnaissent pas plus rapidement *lipstick* dans *cream lipstick* (où la frontière de mots est marquée par le diphone /ml/, pratiquement jamais rencontré à l'intérieur d'un mot, ce qui devrait aider à la segmentation) que dans *drab lipstick*, où on trouve la suite /bl/ à la frontière, alors qu'elle est beaucoup plus courante à l'intérieur des mots du lexique anglais (ce qui devrait gêner la segmentation). D'après cette étude, ce sont donc essentiellement les connaissances lexicales (ici la connaissance des noms composés fréquents) qui sont exploitées, de préférence aux informations plus subtiles provenant du signal.

Mattys et al. (2005), avec des productions moins authentiques (lues et non produites lors d'interactions spontanées), avaient déjà montré que les indices venant du signal acoustique avaient moins de poids que les informations lexicales. Plus précisément, leur étude montrait que les informations prosodiques avaient moins de poids que les indices phonotactiques et allophoniques, qui eux-mêmes avaient moins de poids que les indices lexicaux. Par contre, l'information prosodique résistait beaucoup mieux à la dégradation des conditions d'écoute : en conditions bruyantes, par exemple, les indices prosodiques (l'accent lexical dans leur étude) sont plus robustes que les autres et sont donc utilisés en priorité. C'est ce qu'on voit sur la Figure 1.4 ci-dessous, qui représente visuellement (par la largeur du triangle gris qui diminue en descendant) la hiérarchie des indices utilisés en compréhension : pour reconnaître un mot, l'utilisation du contexte syntaxico-sémantique prime (comme cette hypothèse n'a pas été testée dans leur étude, nous avons grisé cette partie du schéma), suivi des connaissances lexicales. Les indices phonotactiques et allophoniques d'une part, et prosodiques d'autre part, ont un rôle moindre à jouer, sauf quand le contexte est appauvri, soit parce que le mot est inconnu, soit parce qu'il y a trop de bruit pour entendre distinctement les segments, auquel cas la prosodie reste l'indice le plus fiable.

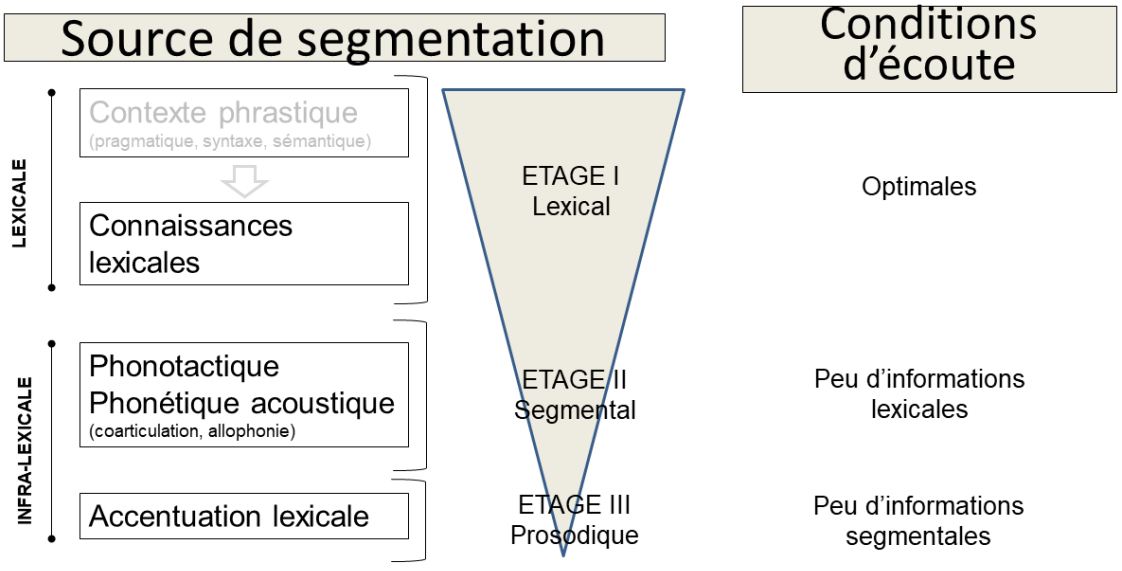


Figure 1.4 – Sources possibles d'information pour la segmentation lexicale, d'après Mattys et al. (2005). L'importance relative des différents indices est représentée par la largeur du triangle inversé.

Bien qu'un grand nombre d'études montrent que les locuteurs sont capables d'exploiter des indices acoustiques venant du signal, qu'ils soient prosodiques (par ex. Cutler & Norris, 1988; Slowiaczek, 1990), infra-phonémiques (M. H. Davis et al., 2002; Salverda et al., 2003), ou phonotactiques (McQueen, 1998), il semble que les informations sémantiques (et en particulier lexicales) gardent toujours un rôle prépondérant.

Notons cependant que les études de Mattys et de son équipe n'étudient pas directement le rôle respectif de la reconnaissance des phonèmes et des connaissances lexicales. Comme ils supposent les phonèmes reconnus, ils comparent en fait les indices supra- et infra-phonémiques, sans que le rôle des indices phonémiques proprement dit soit analysé en détail. D'autres études montrent d'ailleurs que quand le signal est clair, les non-mots, qui n'apportent aucune information lexicale, sont facilement reconnus (Ganong, 1980).

1.4.4. Processus descendants et prédiction

Un nombre croissant de psycholinguistes, et plus généralement, de psychologues cognitivistes s'intéresse au rôle des prédictions dans le fonctionnement de la compréhension et du cerveau en général. Dans son article de 2013, *Whatever next? Predictive brains, situated agents, and the future of cognitive science*, par exemple, le psychologue Andy Clark qualifie le cerveau de « *prediction machine* » (Clark, 2013, p. 181), et Kuperberg et Jaeger (2016, p. 33) affirment que « *in real-world communicative situations, the use of prediction to comprehend language is the norm* ».

Plusieurs linguistes que nous avons cités mentionnent également, lorsqu'ils parlent des processus descendants, l'importance des attentes du locuteur : « *listener expectations about information in the text* » (Vandergrift & Goh, 2012, p. 18), ou « *the input being interpreted in the light of such expectations* » (Færch & Kasper, 1986, p. 264). Ces attentes mènent à des prédictions (« *possible meaning is predicted* », ibid., p.264) qui sont ou non confirmées par le signal. Dans les deux articles ci-dessus, les prédictions des auditeurs ne sont mentionnées que pour les hauts niveaux de traitement, c'est-à-dire une fois que le sens (« *information »/ « meaning »*) intervient. Les auteurs font probablement référence, entre autres, aux études que nous avons citées sur l'utilisation de schémas culturels pour la compréhension des textes (les travaux de Kintsch et son équipe par exemple depuis les années 1970), ou à l'utilisation de scénarios stéréotypés (*scripts*) qui facilitent la compréhension de situations de la vie courante (Long, 1989). Cependant, tous les exemples de traitements descendants que nous avons mentionnés, y compris ceux intervenant à l'étape de décodage, peuvent être réinterprétés comme des exemples d'utilisation des prédictions.

1.4.4.1. exemples de traitement prédictif

Au niveau du décodage, si l'on prend l'exemple de la récupération de phonème (Warren & Warren, 1970), c'est parce que nos connaissances lexicales nous informent que la suite de

phonèmes /ˈledʒɪ/ ne peut être suivie que d'un /s/, que nous nous attendons à l'entendre et ne remarquons même pas qu'il est absent. C'est aussi parce qu'on s'attend à ce qu'un homme ait une voix plus grave et une femme une voix plus aiguë qu'un signal ambigu entre un /s/ et un /ʃ/ sera interprété comme un /s/ dans un cas, et un /ʃ/ dans l'autre (Strand & Johnson, 1996).

Si l'on « monte » à présent d'un niveau et que l'on se concentre sur la reconnaissance lexicale, nous avons également donné des exemples où le contexte syntaxico-sémantique conduisait les locuteurs à anticiper un mot, qui était alors reconnu plus rapidement. Les expériences de Grosjean (1980) montrent ainsi que plus le contexte qui précède est long et contraignant, mieux le mot final est reconnu, et celles de Marslen-Wilson (1975) qu'un mot dans un contexte contraint peut être répété avant qu'il ait fini d'être prononcé, parce que l'auditeur anticipe la fin du mot. D'autres expériences d'oculométrie comme celles d'Altmann et Kamide (1999) montrent que dès que les auditeurs entendent le verbe *eat*, par exemple, leur regard se porte sur les objets comestibles représentés et négligent les autres. Dans une étude plus récente utilisant les potentiels évoqués (*ERP* ou *evoked response potentials*, c'est-à-dire la captation des signaux électriques envoyés par les neurones, grâce à des électrodes placées sur le crâne des sujets), Delong et al. (2005) montrent que les prédictions sémantiques venues du contexte peuvent avoir une traduction au niveau du traitement du signal (écrit, dans leur cas). Les lecteurs s'attendent à ce que le début de phrase *The day was breezy so the boy went outside to fly …* se continue avec le mot *kite*. Cette prédiction se traduit, dès l'article qui précède, par un traitement différent qui donne lieu à un marqueur de surprise (N400) si l'article qui suit le verbe *fly* est *an* (qui ne va pas avec le nom *kite*) plutôt que *a*. Il est donc clair que les locuteurs ont fait une prédiction assez spécifique sur la forme du nom qui va suivre et, partant, sur celle de l'article qui le précède. Cela s'applique en particulier dans les situations de dialogue. D'après Magyari et de Ruiter (2012), et Pickering et Garrod (2013), la vitesse d'échange des tours de parole s'explique par l'utilisation de prédictions : les interlocuteurs prédisent quand le tour de parole de leur partenaire va se terminer et ce que ces derniers vont dire, afin d'être prêts à répondre quand leur tour viendra.

Même un traitement qui est normalement analysé comme découlant d'un processus ascendant (exploitation d'informations de bas niveau contenues dans le signal) peut aussi être vu comme un exemple d'utilisation de prédictions. Nous avons déjà évoqué le phénomène de coarticulation qui fait qu'on commence à articuler le son suivant avant d'avoir fini le précédent. Cette information est utilisée par les auditeurs pour anticiper le son qui va suivre.

Dahan et al. (2001) ont ainsi montré, avec une expérience d'oculométrie, que des sujets qui entendent le mot *net* dont les consonne et voyelle initiales viennent du mot *neck* commencent par regarder l'image d'un cou (*neck*), alors que la consonne /k/ n'apparaît pas dans le signal. C'est donc qu'ils ont utilisé l'information présente dans la voyelle /e/ pour prédire la consonne qui aurait dû suivre. Quand leur prédiction ne se trouve pas confirmée, la consonne suivante étant /t/, ils renoncent finalement à leur hypothèse initiale mais mettent ainsi plus de temps à reconnaître le mot-cible (*net*).

Nous constatons donc que les prédictions peuvent intervenir à tous les niveaux, depuis la reconnaissance des sons jusqu'à la compréhension des relations entre les phrases d'un texte et la construction d'un modèle de situation. Elles ne sont donc pas spécifiques aux niveaux supérieurs porteurs de sens. Certaines de ces prédictions sont clairement d'origine linguistique (prise en compte des effets de coarticulation, connaissances lexicales ou grammaticales), mais d'autres viennent du contexte (visuel, culturel, ou autre). Il n'est d'ailleurs pas toujours facile de les distinguer. En effet, comme le remarque Casasanto (2008), même si la source d'information proprement dite est extralinguistique (par exemple, la taille, le sexe ou l'ethnicité de notre interlocuteur), les connaissances auxquelles elle a abouti sont, elles, linguistiques :

> *Just as listeners might predict that t/d deletion is more likely before a consonant than before a vowel, they are predicting that it is more likely from a black speaker than from a white speaker. The similarity of these predictive processes makes it unsatisfying to classify the socially based phenomenon as stemming from outside the language system.* (Casasanto, 2008, p. 803)

C'est pourquoi nous avions remarqué plus haut qu'il ne nous paraissait pas satisfaisant de classer les processus ascendants ou descendants en fonction de l'origine des connaissances utilisées. En dernière analyse, ces connaissances ont toujours une traduction linguistique.

1.4.4.2. origine des prédictions : fréquence et apprentissage statistique

Si nous sommes capables de prédire quel son va suivre, quel mot va suivre, quels types de phrases vont suivre (à l'intérieur d'un genre textuel contraint), c'est que nous sommes sensibles à la fréquence des événements auxquels nous assistons. Les linguistes savent depuis longtemps que la fréquence des unités linguistiques joue un rôle important dans le traitement du langage. Dès 1957, Howes montrait que les mots plus fréquents étaient mieux reconnus à l'oral que les mots moins fréquents (le résultat était connu pour l'écrit depuis au moins 20 ans : Preston, 1935). En 1986, Paul Luce constata que les mots fréquents étaient également

reconnus plus rapidement - sans doute s'attend-on plus à entendre un mot fréquent qu'un mot rare : « *A word's frequency represents its prior probability and hence constitutes a prediction as to how likely the word is to appear in linguistic experience* » (Norris et al., 2016, p. 4). Cependant, ce n'est qu'en 1996 que Jenny Saffran et son équipe ont démontré que la mémorisation de la fréquence des unités linguistiques (sous forme de probabilités transitionnelles entre les syllabes) joue un rôle dès le début de l'acquisition d'une langue. Nous avons déjà résumé en début de chapitre ces expériences, conduites avec des enfants de huit mois et des adultes, qui sont capables de repérer des nouveaux « mots » (groupes phonologiques récurrents) après avoir écouté pendant deux minutes des suites de syllabes (sans intonation) du type *bidakupadotigolabubidaku*. Ces capacités s'exercent à la fois sur le long terme (pour estimer la fréquence d'un mot, il faut une grande quantité de données sur lesquelles l'estimation de la fréquence s'est peu à peu affinée), mais également sur le court terme. Les locuteurs sont capables de s'adapter rapidement aux caractéristiques changeantes de l'input (Fine et al., 2013), même quand il s'agit d'une langue qu'ils ne connaissent pas : dans les expériences de Saffran, les suites de syllabes à segmenter ne durent que quelques minutes.

Cette sensibilité à la fréquence des événements dont nous faisons l'expérience n'est pas limitée à la sphère linguistique : l'apprentissage statistique chez les êtres humains[5] existe aussi pour l'apprentissage des séquences de symboles visuels, par exemple (Arciuli, 2018). Il s'agit en fait d'une capacité générale à détecter les régularités dans le monde qui nous entoure.

Nick Ellis, dans son article fondateur paru en 2002, *Frequency Effects in Language Processing*, et dans d'autres écrits, fait le tour des implications de cette sensibilité à la fréquence :

> *What's the next letter in a sentence beginning* T… *? Native English speakers know it is much more likely to be* h *or a vowel than it is* z *or other consonants, and that it could not be* q. *But they are never taught this. What is the first word in that sentence? We are likely to opt for* the*, or* that*, rather than* thinks *or* theosophy. *If* The*… begins the sentence, how does it continue? "With an adjective or noun," might be the reply. And, if the sentences starts with* The cat*… , then what? And then again, how should we complete* The cat sat on the*… ? Fluent native speakers know a tremendous amount about the sequences of language at all grains. We know how letters tend to co-occur (common bigrams, trigrams, and other orthographic regularities). Likewise, we know the phonotactics of our tongue and its phrase structure regularities. We know thousands of concrete collocations, and we know abstract generalizations that derive*

[5] et chez d'autres espèces animales (Hauser et al., 2001)

> *from them.*[…] *Psycholinguistic experiments show that we are tuned to these regularities in that we process faster and most easily language which accords with the expectations that have come from our unconscious analysis of the serial probabilities in our lifelong history of input.* (N. C. Ellis, 2003, p. 75)

La connaissance des fréquences des éléments linguistiques et de leur cooccurrence intervient donc à tous les niveaux, et est essentielle pour la prédiction des éléments qui vont suivre lors du traitement de l'input : « *the way that a rational comprehender can maximize the probability of accurately recognizing new linguistic input is to use all her stored probabilistic knowledge, in combination with the preceding context, to process this input* » (Kuperberg & Jaeger, 2016, p. 37).

1.4.4.3. les modèles bayésiens

La théorie bayésienne est l'une des façons d'expliquer comment le cerveau peut tirer parti de la fréquence des événements observés afin de formuler des prédictions. Elle permet de comprendre comment il est possible de tirer des conclusions et de faire des prédictions à partir des informations partielles et souvent ambiguës fournies par notre environnement en général et nos capteurs sensoriels en particulier. Regier et Gahl (2004) reprennent le raisonnement original du mathématicien français Laplace (1749-1827), qui a posé les bases des statistiques qui ont ensuite été qualifiées de bayésiennes. Nous voyons le soleil se lever tous les matins de notre existence. Pouvons-nous en tirer la conclusion que le soleil se lèvera certainement demain (hypothèse 1) ? Il est possible que le soleil n'ait que 50% de chances de se lever chaque jour (hypothèse 2), et que le fait qu'il se soit pour l'instant toujours levé soit le fruit du hasard. Cependant, si cette hypothèse 2 était vraie, je m'attendrais tout de même à avoir été témoin de jours où il ne se lève pas. Plus j'accumule d'observations, moins cette hypothèse paraît donc plausible (et inversement, plus l'hypothèse 1 l'est). Le fait de ne pas observer quelque chose (qu'on attendrait) apporte ainsi des informations en soi, et chaque nouvelle observation nous permet d'augmenter (ou de diminuer) la probabilité que notre hypothèse initiale soit vraie, et de la mettre ainsi à jour.

Norris et McQueen (2008) appliquent ce raisonnement à la reconnaissance aurale des mots dans leur modèle *Shortlist B*. Comme le signal n'est jamais sans ambiguïté, la reconnaissance sera toujours probabiliste. Si nous prédisons l'arrivée d'un mot (hypothèse initiale), le calcul de cette probabilité est basé sur la fréquence du mot (on aura plus de chance de rencontrer un mot fréquent qu'un mot rare), ainsi que sur le contexte, qui réduit l'éventail des mots

possibles[6]. Cette probabilité ainsi estimée est la probabilité a priori de notre prédiction, antérieure au contact avec les données acoustiques. Une fois que les données acoustiques commencent à arriver, nous mettons à jour notre hypothèse en calculant la probabilité que ces données soient observées étant donné notre hypothèse. Je m'apprête à entendre le mot *cat* et j'entends le son /k/, cela est compatible avec mon hypothèse ; si j'entends /g/, c'est moins plausible mais encore possible ; si c'est /f/, il y a peu de chances que mon hypothèse soit confirmée. Nous mettons ainsi à jour nos hypothèses initiales au fur et à mesure (et en parallèle puisqu'il y a souvent plusieurs hypothèses possibles), jusqu'à ce qu'un mot soit reconnu.

L'utilisation des prédictions permet ainsi de mieux expliquer l'interaction des processus descendants et ascendants qui fonctionnent en tandem. Les informations déjà connues (fréquence, contexte gauche) sont utilisées pour anticiper ce qui va venir (processus descendants) et faciliter le traitement du signal (processus ascendants).

1.4.4.4. intérêt des prédictions

Nous avons décrit en détail le mécanisme de prédiction sans nous poser la question de son intérêt. Pourquoi nos connaissances de la fréquence des unités linguistiques auraient-elles un rôle à jouer *avant* le traitement de la nouvelle information (et non pendant, par exemple) ? La réponse généralement apportée par les chercheurs est que ce mécanisme de prédiction permet une plus grande rapidité de traitement. Si l'on est déjà prêt à entendre un son ou un mot, celui-ci peut être traité plus rapidement que quand l'on s'attend à entendre autre chose. Dans ce dernier cas, nous sommes « surpris » par ce que nous entendons et nous perdons du temps à revenir de notre surprise. C'est ainsi que Levy (Levy, 2008, p. 1128) considère que « *surprisal serves as a causal bottleneck between the linguistic representations constructed during sentence comprehension and the processing difficulty incurred at a given word within a sentence* ». Plus un mot est attendu, plus sa reconnaissance est rapide (par exemple, dans une langue où le verbe est en position finale, plus la phrase s'allonge et plus l'arrivée du verbe est probable). Nous avons d'ailleurs mentionné plus haut le fait que les mots fréquents (Luce, 1986b) ou placés dans un contexte facilitateur (Grosjean, 1980) étaient reconnus plus vite que les autres. Le phénomène d'amorçage (*priming*) peut également être réanalysé dans ce sens (Fine et al., 2013). L'amorçage est un paradigme expérimental où « la présentation rapide d'un mot conduit à faciliter le traitement d'un autre mot présenté juste après s'il existe un lien

[6] Le modèle simplifié qui est présenté dans l'article se base uniquement sur la fréquence.

sémantique entre les deux mots successifs » (Gaonac'h, 2005, p. 225). Quand on présente un nouveau mot qui est relié au premier par le sens, ou une phrase qui utilise la même structure syntaxique, le temps de réaction pour reconnaître ou traiter ce nouvel élément est réduit par rapport à d'autres éléments contrôles qui ne sont pas liés au mot ou à la phrase de départ. Fine et ses collègues considèrent que l'apparition du premier mot ou phrase nous conduit à changer notre estimation de la fréquence de cet élément (il est plus fréquent que prévu), et donc à nous attendre à l'entendre plus souvent. Quand il est présenté de nouveau, nous réagissons donc plus rapidement.

Outre un traitement plus rapide, l'utilisation des prédictions permet également de fonctionner à partir de moins de données acoustiques pour confirmer l'interprétation (nous pouvons relier ceci au fait que les mots fréquents sont mieux reconnus que les autres dans un contexte bruyant, Howes, 1957). Tulving et Gold (1963) constatent ainsi que plus une hypothèse est fortement activée, moins elle requiert d'information pour être confirmée : « *the greater the strength of the hypothesis, the less the amount of appropriate information necessary to confirm it* » (Tulving & Gold 1963, p.327, cités par Van Petten & Luka, 2012). En effet, les différentes sources d'information exploitées par les locuteurs sont complémentaires, et si le contexte apporte une importante quantité d'information, ils auront moins besoin des informations contenues dans le signal. Quand le contexte est fortement prédictif, il permet ainsi la compréhension dans des conditions non optimales. La compréhension en conditions réelles peut en effet être assez différente des conditions qui caractérisent les expériences psycholinguistiques qui ont lieu dans un environnement contrôlé où les conditions d'écoute sont en général optimales. Dans la vie réelle, le signal est souvent de mauvaise qualité, du fait du bruit environnant en particulier : « *we communicate in noisy and uncertain environments — there is always uncertainty about the bottom-up input* » (Kuperberg & Jaeger, 2016, p. 35). Cette incertitude née de la mauvaise qualité ou, plus généralement, de la variabilité du signal pourrait être paralysante en conditions réelles si un mécanisme compensatoire n'intervenait pas.

1.4.5. Conséquences pour la compréhension en L2

Nous avons vu dans les paragraphes qui précèdent que la compréhension repose en partie sur l'utilisation de prédictions qui permettent de traiter le signal plus rapidement et plus efficacement en conditions naturelles d'écoute (bruit environnant, style d'élocution de l'interlocuteur et plus généralement variabilité du signal). Ces prédictions supposent une

connaissance fine (acquise de façon implicite) de la fréquence d'occurrence des unités linguistiques de différents niveaux, ainsi que de leur cooccurrence. Dans la compréhension en langue étrangère, cela peut poser deux types de problèmes. D'une part, on peut supposer que les apprenants n'ont pas été exposés à suffisamment de données linguistiques en langue étrangère pour avoir acquis une connaissance suffisante des fréquences d'occurrence et de cooccurrence des unités linguistiques. D'autre part, la prédiction est une opération supplémentaire qui se rajoute aux processus déjà complexes de la compréhension, qui nécessitent l'activation en cascade de nombreux niveaux de traitement.

1.5. Processus automatiques et processus attentionnels

1.5.1. Caractérisation des processus automatiques

Nous avons souligné la rapidité et la robustesse du traitement linguistique grâce à l'utilisation des prédictions et la connaissance des fréquences. Ces caractéristiques sont celles des traitements cognitifs automatiques. D'après Hilton (2019, p. 6, citant le grand psychologue américain William James), l'automatisation est « absolument fondamentale à tout comportement humain. Dans toutes nos interactions avec l'environnement social et physique, une majorité des réactions du système nerveux auront lieu automatiquement – c'est-à-dire, sans effort conscient de notre part ». Hasher et Zachs (1984) font l'inventaire d'un certain nombre de caractéristiques des processus automatiques (dans leur cas, à propos de l'enregistrement automatique des fréquences d'événements). Tout d'abord, la caractéristique définitoire d'un processus automatique est l'absence de conscience (*awareness*) des sujets de ce processus, et le fait que la performance n'est pas meilleure (au contraire) si le processus devait être géré de façon intentionnelle (non automatique, donc). Le grand avantage des processus automatiques est ainsi qu'ils ne requièrent pas d'efforts particuliers et qu'ils ne consomment pas de ressources attentionnelles (qui sont limitées), et permettent donc des traitements conscients (également appelés « contrôlés ») en même temps. Un processus automatique « *has little impact on one's ability to simultaneously attend to other aspects of a situation, such as the interpretation of an ongoing conversation* » (Hasher & Chromiak, 1977, cités par Ellis, 2002, p. 146). Hasher et Chromiak soulignent d'autres caractéristiques qui découlent de cette propriété fondamentale : tout le monde en est capable, quels que soient l'âge ou le niveau d'éducation (les processus automatiques sont peu sensibles aux différences individuelles), et l'utilisation de stratégies conscientes n'améliore pas les performances. Logan (1988) ajoute que les processus automatisés ne sont pas contrôlables (les qualifiant de

« *ballistic* ») puisqu'ils ne font pas appel à l'attention : ils ont lieu qu'on le veuille ou non. C'est ce qu'a prouvé Ridley Stroop en 1935 à l'occasion d'une étude sur l'interférence, en montrant qu'il est très difficile de nommer la couleur des lettres d'un mot si ce mot désigne lui-même une autre couleur (par exemple, le mot « rouge » écrit en bleu). Notre lecture des mots est tellement automatique qu'il est difficile de l'inhiber quand la tâche requiert de ne pas lire.

Certains aspects d'un processus complexe peuvent être automatisés par la pratique (J. R. Anderson & Schunn, 2000; Logan, 1988). Dans la théorie ACT (*Adaptive Control of Thought*) d'Anderson, l'automatisation est le passage d'un traitement contrôlé, gouverné par des connaissances déclaratives (conscientes), à un traitement routinier où les opérations, regroupées par *chunks*, sont beaucoup plus rapides. Dans la théorie des instances de Logan, l'automatisation d'un processus est le fruit de la mémorisation de nombreuses instances du résultat du processus. Dans les deux théories, la rapidité liée à l'automatisation résulte de l'entraînement et de la pratique. Logan utilise l'exemple des opérations de base en arithmétique pour illustrer ses propos : les enfants comptent sur les doigts (processus attentionnel assez coûteux) pour faire des additions jusqu'à ce qu'ils aient suffisamment bien mémorisé les tables d'addition pour que les résultats soient activés automatiquement en mémoire à long terme en une seule étape. L'accélération qui en découle peut être caractérisée par une loi de puissance, selon laquelle les gains de vitesse sont importants au début mais diminuent petit à petit. A mesure que notre performance augmente, les gains de productivité associés à plus de pratique diminuent, et il devient de plus en plus difficile de s'améliorer (Logan, 1988, p. 97).

1.5.2. Processus automatisés en compréhension de l'oral

Puisque l'automatisation conduit à des performances plus rapides et moins sujettes à variation, il paraît essentiel que l'ensemble complexe des processus impliqués dans la compréhension de l'oral soient automatisés dans la mesure du possible. Nous avons pour l'instant travaillé à partir de deux modèles classiques hiérarchiques de la compréhension de l'oral, constitués de trois (Anderson) ou de quatre (Cutler et Clifton) niveaux successifs s'activant en cascade. Les auteurs de ces modèles n'ont pas ouvertement détaillé la nature automatique ou non des processus impliqués. D'autres modèles permettent cependant de statuer sur la part des traitements automatiques lors de la compréhension de l'oral. Le modèle proposé par Field (2008a), par exemple, postule deux opérations générales seulement : le

décodage (*decoding*) couvre non seulement la reconnaissance des sons et des mots, mais également des propositions et leur traduction en un sens qu'il appelle « littéral » (sens propositionnel). La construction du sens (*meaning building*) correspond à l'étape d'utilisation d'Anderson ou d'intégration de Cutler et Clifton, c'est-à-dire la construction du sens du texte et du modèle de situation (y compris les inférences) : « *adding to the bare meaning provided by decoding and relating it to what has been said before* » (J. Field, 2008a, p. 125).

Cette division bipartite peut sembler un peu grossière dans la mesure où chacun des deux niveaux recouvre nécessairement beaucoup d'opérations différentes. Le décodage, en particulier, correspond aux trois premiers niveaux du modèle de Cutler et Clifton et permet difficilement de mettre en avant la spécificité de la compréhension à partir d'un matériau oral. Cependant, ce modèle est intéressant dans la mesure où la distinction entre les deux opérations est faite en grande partie parce que, chez les auditeurs chevronnés (et en particulier chez les natifs), ce que Field appelle le « décodage » est automatique, alors que les opérations de construction du sens au niveau textuel font appel à des processus qui sont plus contrôlés : « *highly automatic [processes] in the case of decoding and more rational ones in the case of meaning building* » (J. Field, 2008a, p. 126). Il nous semble que l'objectif d'un enseignement de la compréhension de l'oral doit justement être de parvenir à un traitement automatique chez les apprenants de la partie « décodage » (au sens de Field) du processus, afin de libérer des capacités d'attention pour les opérations de haut niveau.

Cette distinction entre processus automatiques et processus faisant appel à l'attention nous paraît de plus correspondre à celle que fait Hilton :

> Dans une situation de compréhension routinière, ces processus linguistiques (traitements phonologiques et prosodiques, reconnaissance lexicale et grammaticale) ont lieu automatiquement – sans que l'on y consacre de ressources attentionnelles, notre effort conscient étant focalisé sur l'interprétation et l'intégration des aspects sémantiques et sociaux du message, la gestion du sens : cohérence des idées exprimées par rapport au discours qui précède, à notre représentation du monde et de notre interlocuteur, aux schémas de nos représentations sociales. Ce n'est qu'assez rarement en L1 qu'on « fait attention » à tel détail formel dans le discours (une syllabe, un mot, une forme grammaticale) – notamment et surtout si ce détail « cloche » avec ce qu'on est habitué à entendre. (Hilton, 2019, p. 7)

Ici encore, ce n'est qu'au niveau sémantique/ discursif (processus de « haut niveau ») que la compréhension de l'oral experte fait appel à des processus non automatisés. Aux niveaux

inférieurs, jusqu'au traitement grammatical compris, les processus linguistiques ne sollicitent pas l'attention, sauf en cas de problème intempestif.

1.5.3. Le rôle des stratégies

Si certains processus mentaux sont automatisés, d'autres ne le sont pas et nécessitent l'intervention de l'attention. L'utilisation de l'attention peut également être requise même pour des processus normalement automatisés, quand un problème inhabituel survient (comme le souligne Hilton dans la citation ci-dessus, un détail qui « cloche » est toujours possible). C'est dans ces moments que les stratégies peuvent entrer en jeu. Beaucoup d'études ont été menées sur les stratégies de compréhension aurale en L2, mais il en existe moins en L1, sans doute parce que la compréhension de l'oral en L1 est hautement automatisée et ne pose en général pas de problèmes particuliers (quoique son rôle dans l'acquisition de la lecture commence à être étudié plus en détail : par ex. Bianco, 2016). Les stratégies sont des actions conscientes mises en œuvre pour résoudre un problème : « *conscious mental behavior or action that is applied to achieve an explicit goal in the target language* » (A. D. Cohen, 1998, cité par Zoghlami, 2015, p. 47). Il est plus courant, dans la littérature psycholinguistique sur la L1, de parler de stratégies à propos de l'acquisition de la lecture, définies comme des processus métalinguistiques déployés consciemment pour résoudre un problème de décodage ou d'activation du sens : « *deliberate, goal-directed attempts to control and modify the reader's efforts to decode text, understand words, and construct meanings of text.* » (Afflerbach et al., 2008, p. 368). Les deux caractéristiques importantes sont que ces stratégies sont conscientes (et donc peuvent être verbalisées, par exemple lors d'une tâche de *think aloud*), et qu'elles sont tournées vers un but (de reconstruction du sens).

Nous reviendrons sur l'utilisation des stratégies dans la partie sur la compréhension an anglais L2, mais nous voudrions souligner tout de suite le contraste entre les processus automatisés, rapides, efficaces, et les processus plus complexes pouvant être améliorés par l'utilisation de stratégies soumises au contrôle de l'auditeur.

1.5.4. Passage d'un niveau à l'autre de traitement et *chunking* : le modèle *Chunk-and-Pass*

Nous voudrions terminer ce chapitre par la présentation d'un modèle de compréhension du langage qui tente d'expliquer la rapidité des processus en jeu que nous avons déjà soulignée

dans les paragraphes sur l'automatisation. Ce modèle est décrit dans un article de 2016 (Christiansen & Chater, 2016), *The Now-or-Never bottleneck: A fundamental constraint on language*. L'entonnoir (*bottleneck*) dont il est question dans le titre est la difficulté qu'a notre cerveau à traiter plus d'une poignée d'éléments à la fois (7±2 selon Miller, 1956, ou 4±1 selon Cowan, 2010), du fait de la capacité limitée de notre mémoire de travail. Pour expliquer le fonctionnement du modèle, Christiansen et Chater font le parallèle avec les techniques de mémorisation de longues suites de chiffres et donnent l'exemple du sujet S.F., étudié par Ericsson et al (1980). Au début de l'expérience, S.F., qui ne connaît aucune technique de mémorisation, échoue à mémoriser des suites de plus de 7 chiffres. Il apprend ensuite à grouper les chiffres par groupes de 3 ou 4, et peut alors mémoriser au moins 16 chiffres (en première approximation, 3 groupes de 4 chiffres et 4 chiffres isolés). Il arrive ensuite petit à petit à former des « super-groupes » de 3 ou 4 groupes, puis à structurer ces super-groupes, de sorte qu'il arrive à la fin de l'expérience (qui dure environ 230 heures sur un an et demi) à mémoriser près de 80 chiffres qui lui sont lus à haute voix à raison de 1 chiffre par seconde.

Christiansen et Chater font l'hypothèse que les locuteurs font la même chose pour le langage et appellent leur modèle *Chunk-and-Pass* : dès qu'un ou plusieurs phonèmes (ou syllabes) sont reconnus dans le flux acoustique, ils sont regroupés (*chunked*) et passent au niveau supérieur (celui des groupements phonologiques). Dès qu'une suite de phonèmes ou syllabes est reconnue comme un mot, elle est regroupée et passe au niveau lexical, et dès qu'un groupe de mots est reconnu, il passe au niveau phrastique, et ainsi de suite. Cette capacité du cerveau à grouper des éléments pour créer des unités plus grandes qui permettent *in fine* de mémoriser plus d'éléments de base n'est pas une découverte récente : Ellis (2003) rappelle que le terme de *chunk* vient de l'article fondateur de George Miller (1956), *The magical number seven plus or minus two: some limits on our capacity for processing information*. Dans cet article, Miller montre que la limite de 7 éléments qui peuvent être conservés en mémoire de travail peut être levée à condition de « regrouper » ces éléments en suites d'éléments qui contiennent plus d'information. Ce regroupement (*chunking*) donne lieu naturellement à une organisation hiérarchisée :

> *[A] chunk is a unit of memory organization, formed by bringing together a set of already formed elements […] in memory and welding them together into a larger unit. Chunking implies the ability to build up such structures recursively, thus leading to a hierarchical organization of memory.* (Newell 1990, p.7, cité par N. C. Ellis, 2003, p. 76)

L'intérêt de ce recodage constant des niveaux inférieurs aux niveaux supérieurs est triple. Premièrement, il permet de limiter l'interférence entre les éléments déjà entendus (les phonèmes, par exemple) et ceux qui continuent à arriver à une vitesse soutenue (jusqu'à 20 phonèmes par seconde, cf. section 1.1.1). Si les groupes de phonèmes sont regroupés et passés au niveau lexical au fur et à mesure de leur reconnaissance, cela permet de libérer de l'attention pour traiter les nouveaux phonèmes qui arrivent. La nécessité de faire face à ce flux constant « oblige » en quelque sorte à adopter cette organisation hiérarchique. Deuxièmement, cela explique comment peut fonctionner l'analyse en parallèle des différents niveaux de traitement décrite plus haut : l'intégration des informations phonémiques, lexicales, syntaxiques, etc. se fait dès que ces informations sont disponibles, et le traitement d'une phrase ou d'un texte se produit par incréments successifs (et non une fois seulement que toutes les informations sont disponibles). Loin de compliquer les choses, cette analyse en cascade à différents niveaux est une façon de faire face à l'afflux constant de données acoustiques. Troisièmement, le regroupement d'unités permet d'expliquer en partie les variations individuelles en compréhension. La taille des groupes ainsi formés dépend en effet de l'expérience des sujets, dans la mesure où les unités fréquemment entendues ensemble ont plus tendance à être regroupées. Les auditeurs avec plus d'expérience auront donc tendance à utiliser de plus grands regroupements. McCauley et Christiansen (2015) ont d'ailleurs montré, avec des sujets adultes, que la capacité à regrouper était corrélée avec la compréhension de phrases complexes, et Jones (2012) a utilisé une modélisation informatique de données d'acquisition pour montrer que la capacité à regrouper (liée à l'exposition répétée) permettait d'expliquer ces données chez de jeunes enfants. Hilton (2009, p. 104) souligne d'ailleurs le lien entre ce processus de regroupement et la théorie des instances de Logan (1988), que nous avons déjà citée à propos de l'automatisation, et où les *chunks* sont récupérés directement en mémoire au lieu d'être décomposés à chaque fois.

Nous reproduisons en Figure 1.5 une représentation graphique du fonctionnement de ce modèle. On y retrouve l'organisation hiérarchique en niveaux successifs correspondant à des unités linguistiques de plus en plus grandes. La partie noircie de chacun des niveaux correspond à la fenêtre temporelle active à un instant donné : même si le nombre d'unités disponibles en mémoire de travail à chaque instant est sensiblement le même quel que soit le niveau, le fait que les unités soient plus grandes aux niveaux supérieurs implique qu'elles couvrent également une durée plus longue. Ceci correspond d'ailleurs aux résultats en imagerie cérébrale (Lerner et al., 2011), selon lesquels les fenêtres temporelles auxquelles les

zones corticales sont sensibles augmentent progressivement, depuis les zones qui réagissent aux premières informations sensorielles (moins d'une seconde) jusqu'aux zones qui répondent à l'écoute d'un paragraphe ou d'une histoire complète (de l'ordre d'une minute et plus).

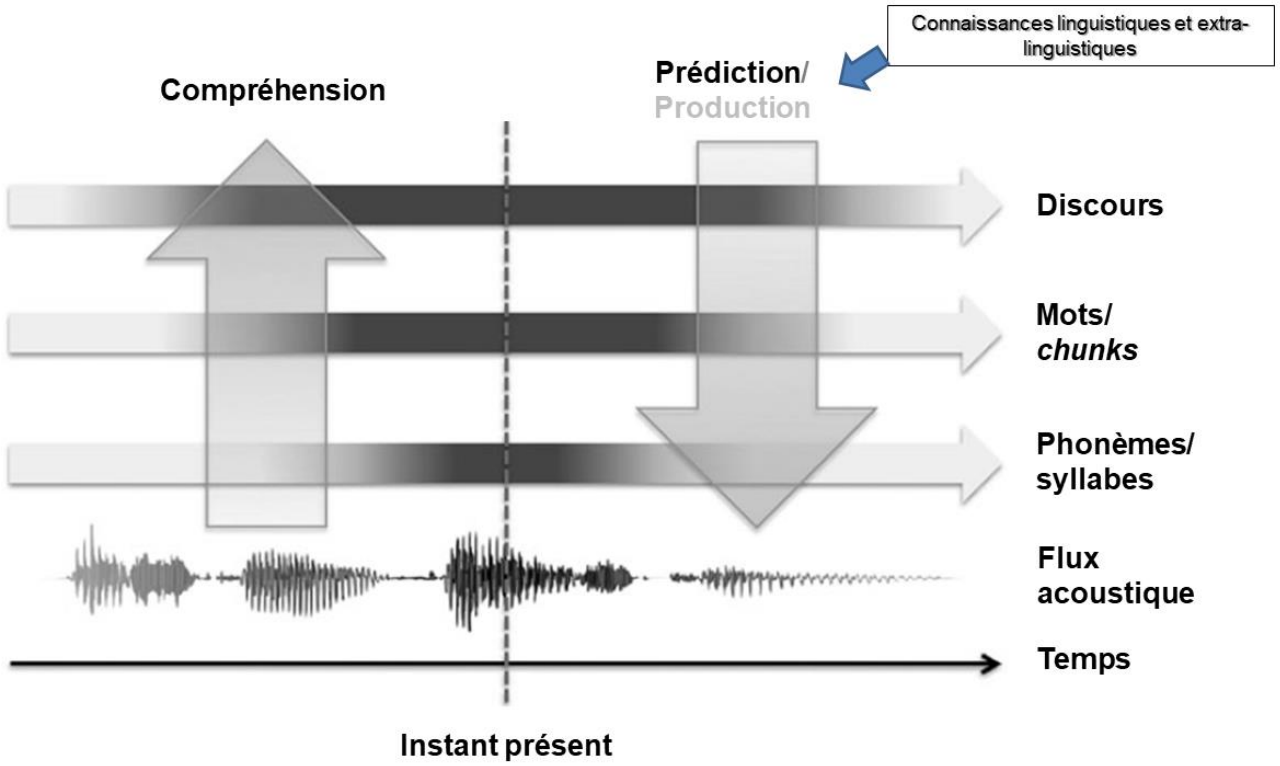


Figure 1.5 - modèle *Chunk and Pass* d'après Christiansen et al. (2016), représentant l'organisation hiérarchique du traitement linguistique, dans laquelle les unités linguistiques reconnues sont regroupées en unités de plus en plus grandes et de plus en plus abstraites, depuis le flux acoustique jusqu'au modèle de discours. La partie noircie de chacun des niveaux correspond à la fenêtre temporelle active à l'instant t, qui est de plus en plus étendue à mesure qu'on s'élève dans les niveaux. La partie droite du modèle montre que le système utilise ses connaissances linguistiques et extralinguistiques pour émettre des prédictions qui aident à traiter le flux entrant qui continue à arriver (le terme « Production » est grisé car cette partie du modèle ne nous concerne pas directement).

Ce modèle permet ainsi d'intégrer de nombreux résultats que nous avons décrits précédemment. Tout d'abord, nous y retrouvons l'organisation hiérarchique du traitement linguistique déjà présenté avec les modèles d'Anderson (1995) et surtout de Cutler et Clifton (1999). Seuls trois niveaux sont illustrés sur la représentation graphique du modèle, mais Christiansen et Chater se disent « agnostiques » (p.7) quant au nombre de niveaux effectivement utilisés. Ensuite, nous retrouvons l'idée que les informations sont traitées dès qu'elles sont disponibles, au fur et à mesure de leur arrivée - c'est d'ailleurs justement ce qui motive l'organisation hiérarchique. Par ailleurs, le processus de regroupement d'unités d'un niveau à l'autre entraîne une perte de certaines informations de détail du niveau inférieur (ce que Christiansen et Chater appellent *lossy chunking*, ou « regroupement avec perte d'informations »), ce qui explique par exemple qu'après avoir compris un discours (niveau du

modèle de situation), nous nous souvenons de son sens mais pas des mots exacts entendus (niveaux lexical ou syntaxique). Enfin, le rôle des prédictions est essentiel, parce qu'elles permettent un traitement beaucoup plus rapide de l'information qui continue à arriver. Le modèle illustre les phénomènes que nous avons décrits dans les sections précédentes : les prédictions ont lieu à tous les niveaux de traitement[7], du signal acoustique (prédictions sur le très court terme) jusqu'au niveau textuel (prédictions sur le long terme). Nous avons donc ajouté au modèle *Pass-and*-Chunk original une bulle (en haut à droite) pour préciser que les prédictions pouvaient être d'origine linguistique ou extralinguistique (visuelle, culturelle, encyclopédique, etc.).

1.6. Conclusion : un modèle parallèle et interactif pour la L2

Nous avons essayé dans ce chapitre de présenter le fonctionnement de la compréhension de l'oral de façon suffisamment détaillée pour que les difficultés qui risquent d'émerger en L2 puissent être identifiées précisément. Il nous semble que ces difficultés (que nous explorerons plus en détail dans le chapitre suivant) pourront être de deux ordres, que nous résumerons sous les étiquettes de « connaissances » et de « capacités de traitement ». Les connaissances sont celles qui sont nécessaires pour que chaque niveau du système hiérarchique fonctionne. Il s'agit de la connaissance des unités linguistiques : phonèmes, allophones, syllabes, lexique, grammaire, etc., connaissance des fréquences d'occurrence et de cooccurrence de ces unités, ainsi que connaissances extralinguistiques. Quant aux capacités de traitement, elles sont nécessaires pour que fonctionne le système complexe d'opérations en cascade qui caractérise la compréhension en ligne de l'oral.

Pour ce qui est des connaissances linguistiques, nous pouvons essayer de prévoir lesquelles sont les plus susceptibles de faire défaut à des apprenants L2 (quelle que soit la L2 pour l'instant ; nous nous concentrerons sur français L1/ anglais L2 dans le chapitre qui suit). Selon les niveaux hiérarchiques que nous avons identifiés, des problèmes peuvent se poser au niveau des unités sonores, puisque différentes langues utilisent des inventaires phonémiques différents. Cependant, une partie de ces inventaires est commun à un très grand nombre de langues. D'après la base de données *UPSID* (UCLA Phonological Segment Inventory Database, Maddieson & Disner, 1984), certains phonèmes consonantiques comme /m/, /k/, /j/ ou /p/, ou vocaliques comme /i/, /a/ ou /u/, sont communs à plus de 80% des 451 langues

[7] Dans leur modèle, les prédictions sont liées à la production, mais comme cela ne nous concerne pas directement, nous avons grisé le terme de « production ».

répertoriées dans la base. Quelles que soient les langues de départ ou d'arrivée, il sera très certainement possible à un niveau débutant de repérer quelques sons dans une langue étrangère, même si, comme nous le verrons, certaines distinctions phonémiques entre sons proches seront difficiles à percevoir.

Si deux langues ont toutes les chances d'avoir quelques phonèmes en commun, il n'en est pas de même pour le lexique, en particulier à l'oral. Des langues qui ne sont pas liées étymologiquement peuvent ne partager presque aucun des mots de leur vocabulaire (le français et le chinois, par exemple). Même en cas d'emprunts, l'adaptation au système phonologique de l'autre langue peut rendre ces mots méconnaissables à l'oral. Par exemple, le chinois mandarin a moins de 2% de mots empruntés (Haspelmath & Tadmor, 2009, p. 57), et ces mots sont souvent difficiles à reconnaître même pour les locuteurs de la langue d'emprunt (*kǎ* pour *card*, ou *sān míng zhì* pour *sandwich* par exemple pour des anglophones apprenant le mandarin). Même si le lexique commun au français et à l'anglais est beaucoup plus important, le problème de l'altération phonologique se posera également, comme nous le verrons. Par ailleurs, le nombre de mots d'une langue est sans commune mesure avec le nombre très limité de ses phonèmes, et impliquera donc des connaissances beaucoup plus étendues.

Si l'on passe à la structure grammaticale des phrases, deux choses peuvent faciliter leur compréhension. Le fait d'avoir reconnu les mots d'une phrase permet souvent d'avoir une vague idée du sens, surtout dans des contextes concrets. Par exemple, à partir de « avec Legourdin loi frapper à cravache une interdisait la Mademoiselle le de », on peut éventuellement reconstituer la relation « loi/ interdisait/ frapper/ cravache » (la phrase de départ étant « La loi interdisait à Mademoiselle Legourdin de le frapper avec une cravache »)[8]. D'autre part, l'ordre des mots suit en général des lois qui rendent très improbable celui de la phrase en désordre que nous venons de présenter : les éléments dont le sens se complète (« la » et « loi ») sont en général présentés côte à côte (principe du localisme, Christiansen & Chater, 2016), et il existe essentiellement trois ordres possibles pour les constituants principaux (sujet, verbe, objet) : l'ordre SVO (que suivent 36% des langues connues, par exemple le français ou l'anglais), l'ordre SOV (41% des langues connues, dont le japonais) et VSO (7% des langues, comme l'arabe classique, Dryer & Haspelmath, 2013). Il est donc probable que la syntaxe pose en compréhension moins de problèmes que le lexique.

[8] Il s'agit d'une phrase tirée de la traduction française de *Matilda*, de Roald Dahl.

Enfin, au niveau du discours, nous avons signalé à quel point les genres textuels étaient culturellement déterminés. Si l'on s'en tient à des genres bien définis dans la culture occidentale, on peut supposer que les schémas textuels seront similaires d'une langue à l'autre et ne poseront pas vraiment de problèmes de connaissances (ou du moins, pas plus dans la langue étrangère que dans la langue maternelle). Le véritable problème à ce niveau est différent, et c'est là que les capacités de traitement peuvent jouer un rôle.

Comme nous l'avons vu, la compréhension de l'oral met en jeu de nombreux processus qui doivent s'imbriquer de façon complexe pour que le tout fonctionne correctement, alors que les informations du signal arrivent continuellement. Beaucoup de ces processus sont automatisés en L1, et une partie de cette automatisation vient de la pratique et de l'exposition à la langue (par exemple à travers la création de *chunks* de plus en plus nombreux et de plus en plus grands). Même si ces processus impliquent des traitements de bas niveau au niveau de la perception des unités phonologiques, le but de la compréhension n'est pas de saisir ces unités. Elles sont simplement un moyen d'arriver à une fin, qui est la compréhension d'un message. C'est donc au niveau de la construction du sens du texte que tout se joue (en effet, décoder parfaitement les phonèmes ou les mots sans comprendre le sens du message ni les intentions du locuteur ne sert à rien). Or, nous avons vu qu'à ce niveau, les unités à traiter étaient beaucoup plus grandes, les processus étaient moins automatisés, et nécessitaient de l'attention. Même si les connaissances sur les genres textuels (similaires dans leur L1) ne font pas défaut aux auditeurs L2, c'est l'attention supplémentaire que requiert ce niveau qui pourra faire obstacle à la compréhension finale, et ce d'autant plus que le fonctionnement des niveaux inférieurs sera moins automatisé qu'en L1. Selon Field (2008a, p. 85-86), « *many meaning-building processes will certainly be fully established in the learner's native language, but they may not be applied in L2 listening because of the additional attention that has to be given to decoding unfamiliar sounds and words* ». Le problème véritable au niveau ultime est qu'il ne reste plus forcément assez d'attention aux auditeurs pour appliquer les stratégies nécessaires à la construction de la macrostructure du texte.

Nous proposons de résumer ces informations dans le tableau ci-dessous (Tableau 1.1), qui montre qu'on peut prévoir que les problèmes principaux en compréhension de l'oral risquent de surgir aux niveaux des connaissances phonologiques et lexicales (et particulièrement de ces dernières), et au niveau de la capacité de traitement en ligne du flux du matériau sonore.

Nous approfondirons ces points dans le chapitre qui suit consacré aux difficultés spécifiques des apprenants francophones apprenant l'anglais.

source de difficulté		**niveau de difficulté prévisible**	**éléments facilitateurs**	**obstacles potentiels**
connaissances	phonologiques et prosodiques	++	phonèmes communs avec L1	phonèmes et distinctions phonémiques inconnus ; structure prosodique différente
	lexicales	+++	mots transparents	étendue du lexique ; mots transparents difficiles à reconnaître à l'oral ; difficulté à segmenter
	grammaticales	+	sens phrastique partiellement dérivable du sens des mots	ordre des mots différent de L1 et dépendances non locales
	textuelles	-	structure commune L1-L2 des grands genres textuels	différences culturelles subtiles
capacités de traitement	attention	+++		manque d'automatisation aux niveaux inférieurs ; attention limitée pour le traitement du sens
	chunking	++		manque d'expérience (*input*) en L2

Tableau 1.1 - sources de difficulté prévisibles en compréhension de l'oral d'une langue étrangère

Chapitre 2

La compréhension de l'oral en anglais L2

Dans ce chapitre sur les difficultés spécifiques aux apprenants L2 (en particulier francophones) en compréhension aurale de l'anglais, nous reprendrons les modèles de la compréhension de l'oral en langue maternelle (L1) présentés dans le premier chapitre, car le consensus actuel chez les spécialistes de l'acquisition est que les mécanismes à l'œuvre sont les mêmes :

> *Although Anderson's (1995) three-phase model is based on first language (L1) comprehension, it is no less relevant to an understanding of second language comprehension (L2).* (Goh, 2000, p. 57);
>
> *The evidence to date gives no reason to suppose that second-language listening is in any fundamental way different from first-language listening. The same processes seem to apply.* (Buck, 2001, p. 48);
>
> *[M]any common processes underlie native and second-language syntactic processing.* (Frenck-Mestre, 2002, p. 217)

A propos des résultats d'une expérience sur l'utilisation de prédictions en L1 et en L2, Hopp conclut également que : « *these findings strongly argue against fundamental neurocognitive differences between native and adult non-native speakers* » (Hopp, 2016, p. 25). Les processus sont donc les mêmes, mais les connaissances linguistiques sont différentes du fait d'une expérience moindre en L2 et de l'influence de la L1. C'est justement parce qu'un même ensemble de processus est à l'oeuvre que l'interférence entre les deux langues peut survenir et que ces processus chez les apprenants L1 et L2 ne conduisent pas forcément à des résultats identiques :

> *Adult language learners are distinguished from infant L1 acquirers by the fact that they have previously devoted considerable resources to the estimation of the characteristics of another language – the native tongue in which they have considerable fluency (and any others subsequently acquired). Since they are using the*

> *same apparatus to survey their additional language too, their computations and induction are often affected by transfer.* (N. C. Ellis, 2009, p. 153)

Le modèle hiérarchique sur lequel nous nous appuyons fonctionne par niveaux successifs utilisant des unités de plus en plus grandes et s'appuyant sur des connaissances différentes à chaque niveau. Nous avons exposé en conclusion de premier chapitre lesquelles de ces connaissances sont susceptibles de poser à nos étudiants le plus de problèmes. Ce sont ces connaissances que nous étudierons en détail dans ce chapitre. Il s'agit en premier lieu des connaissances phonologiques, puis des connaissances lexicales et morphosyntaxiques. Pour les connaissances phonologiques, nous nous concentrerons sur la reconnaissance des phonèmes (à propos de laquelle nous mentionnerons également les unités syllabiques et les allophones), et la sensibilité accentuelle, dont nous avons montré qu'elle pouvait être importante pour la segmentation.

2.1. Le rôle des phonèmes

La première tâche de l'auditeur est de repérer des unités phonologiques dans le flux acoustique. Cela correspond, dans le modèle de Cutler et Clifton (1999) présenté dans le premier chapitre, à l'étape de décodage, qui permet de dégager des unités abstraites. Christophe et ses collaborateurs remarquent ainsi :

> Nous n'avons aucune difficulté à identifier un même mot prononcé, par un adulte ou par un enfant, murmuré ou crié, bien que le signal acoustique soit extrêmement différent selon les cas [...]. II semble donc raisonnable de postuler que nos représentations mentales des mots sont normalisées. Une illustration de ce que pourrait être une telle représentation nous est fournie par l'écriture : les mots écrits sont représentés à l'aide des lettres, et nous sommes capables de lire des écritures manuscrites très différentes. (Christophe et al., 1991, p. 62)

Cette étape de normalisation implique que le signal en entrée soit transformé en chaîne de phonèmes (ou autres unités phonologiques) en sortie, et que cette chaîne de phonèmes déclenche ensuite la reconnaissance lexicale.

2.1.1. Origine des difficultés de reconnaissance des phonèmes

Les difficultés de perception des phonèmes d'une langue étrangère et de ses contrastes phonémiques (paires minimales comme *sheep*/ *ship*) sont bien connus. On peut expliquer cet état de fait en remontant au phénomène d'aimant perceptuel mis en lumière dans les années 1990 par Patricia Kuhl (Kuhl, 1991; Kuhl et al., 1992) chez l'adulte et le très jeune enfant. Il

est plus difficile d'entendre la différence entre une voyelle prototypique et d'autres voyelles très proches qu'entre une voyelle non prototypique et des voyelles tout aussi proches. Une voyelle prototypique est obtenue en proposant à des locuteurs natifs d'écouter un certain nombre de variantes du phonème /i:/ (par exemple), et de noter sur une échelle de 1 à 7 à quel point chacune représente un bon exemple du phonème ; la variante avec la meilleure note est choisie comme prototype. La voyelle prototypique se comporte comme un « aimant perceptuel » qui attire la catégorisation (le même phénomène peut d'ailleurs être observé pour la vision des couleurs, Kay & Kempton, 1984). Cela permettrait aux jeunes enfants (qui perdent la capacité universelle à discriminer les sons entre 6 et 12 mois) de normaliser plus facilement la chaîne sonore qui leur est proposée : tous les sons ressemblant à l'attracteur sont catégorisés comme une instance de celui-ci. Ce phénomène est schématisé de la façon suivante par Kuhl et Iverson (1995, p. 124):

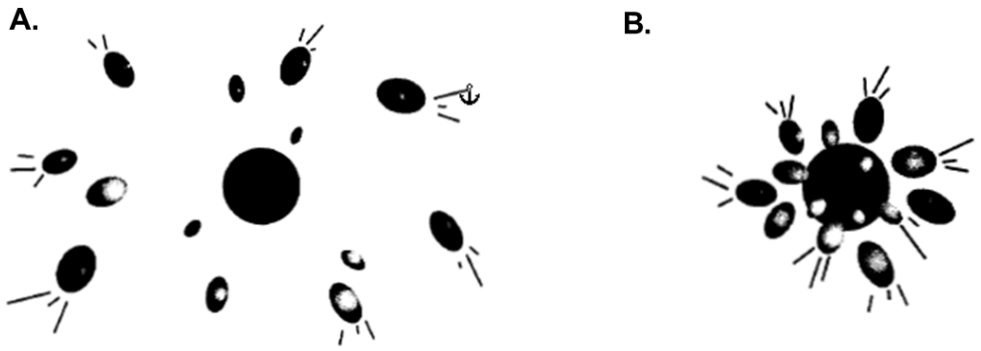


Figure 2.1 - représentation de la distance objective (A.) et de la distance perçue (B.) entre une voyelle prototypique (au centre) et des instances proches (d'après Kuhl & Iverson 1995)

Sur la Figure 2.1, le schéma A représente la distance objective entre un certain nombre d'instances du phonème /i:/ en anglais, et le schéma B, la façon dont ces instances sont perçues. On peut observer le comportement d'aimant perceptuel de la voyelle prototypique au centre.

Lors de l'utilisation d'une L2, ces aimants perceptuels empêchent l'apprenant d'entendre les différences entre des phonèmes proches et qui ressemblent tous les deux à un phonème unique de sa langue maternelle. C'est ce qui est expliqué par le *Perceptual Assimilation Model* (PAM) de Best (1995) et le *Speech Learning Model* (SLM) de Flege (Flege & MacKay,

2004), qui supposent que l'établissement des catégories phonémiques de la L2 est bloqué quand elles sont trop proches de la L1. Elles sont alors assimilées aux catégories de la L1, sans que les différences puissent être entendues par les apprenants. D'après Flege et MacKay, moins la distance perçue entre le son à acquérir de la L2 et le son le plus proche correspondant en L1 est élevée, moins il y aura de chances que l'apprenant crée une nouvelle catégorie pour le son de L2. Le PAM de Best s'intéresse plus particulièrement à l'acquisition des contrastes, et présente (entre autres) trois modes possibles d'appréhension des nouveaux phonèmes d'une L2. Le premier cas correspond à la situation où les deux phonèmes de la L2 sont proches de deux phonèmes différents de la L1. Dans ce cas, la discrimination ne pose pas de problème (c'est le cas appelé *Two-category assimilation*). Par contre, quand les deux phonèmes de la L2 sont proches d'un même phonème de la L1, mais qu'ils ne sont pas tous deux aussi proches de l'attracteur central du phonème (*Category Goodness Difference*), la discrimination est possible, mais plus difficile que dans le premier cas ; l'un des deux pourra peut-être être distingué de l'autre parce qu'il sera perçu comme un exemplaire « bizarre » du phonème. Enfin, quand les deux phonèmes L2 se rapprochent d'un unique phonème L1, mais qu'aucun des deux n'en est vraiment un bon exemple, la discrimination sera très difficile (*Single Category Assimilation*).

2.1.2. Phonèmes de l'anglais et apprenants francophones

A la lumière de ce que nous venons de voir, quels contrastes ou phonèmes risquent d'être difficiles à percevoir pour l'apprenant francophone? L'inventaire des phonèmes de l'anglais est un peu plus important que celui du français, avec entre 25 et 30% de phonèmes en plus (20 consonnes, 11 voyelles simples et 3 voyelles nasales en français, et 24 consonnes, 12 voyelles simples en anglais, plus 5 diphtongues en anglais américain- *General American* (*GA*) ou 8 en anglais britannique – *Received Pronunciation* ou *RP*). Nous examinerons séparément les consonnes et les voyelles, en commençant par des tableaux synthétiques, dans lesquels nous indiquerons en gras les phonèmes qui n'existent que dans une des deux langues, et soulignerons ceux qui sont propres à l'anglais, et donc susceptibles de poser des problèmes particuliers aux apprenants francophones.

Nous reprendrons ensuite brièvement l'analyse de Terrier (Terrier, 2011, p. 38), qui expose en détail ces problèmes pour chaque phonème, mais nous avons cherché à valider cette liste, qui s'inspire au départ essentiellement de problèmes connus en production, par les études que nous avons pu trouver sur la perception de ces phonèmes par des francophones.

2.1.2.1. Les consonnes

	Consonnes du français	Consonnes de l'anglais (*RP*)	Anglais (*GA*)
plosives	p b t d k g	id	id
fricatives	f v s z ʃ ʒ	id + **θ ð h**	id RP
affriquées		**tʃ dʒ**	id RP
nasales	m n **ɲ**	m n **ŋ**	id RP
liquides	l **ʁ** j w **ɥ**	l **r** j w	Id RP

Tableau 2.1 – tableau comparatif des consonnes du français et de l'anglais britannique (RP) et américain (GA), classées par mode d'articulation

Par convention, à l'intérieur de chaque cellule du Tableau 2.1, les consonnes sont classées par lieu d'articulation, depuis l'avant de la bouche (par exemple /p/ et /b/ qui sont bilabiales) à l'arrière de la bouche (/k/ et /g/ qui sont vélaires). Quand deux consonnes ont le même lieu d'articulation, la consonne non voisée est en premier (/p/), et la consonne voisée suit (/b/).

Nous constatons que les systèmes consonantiques sont peu différents au niveau phonémique. Les différences se situent essentiellement au niveau des fricatives, par lesquelles nous commencerons.

La fricative /h/ n'est proche d'aucun phonème du français et ne devrait donc pas subir le phénomène d'aimant perceptuel, ce qui devrait faciliter son traitement selon le SLM de Flege. Cependant, il faut que les apprenants francophones aient suffisamment d'exposition à l'anglais (et peut-être une aide explicite, qui est prévue dès les programmes de primaire du Ministère de l'Education Nationale) pour construire cette catégorie. Le fait que ce phonème soit toujours représenté à l'écrit, et qu'il y ait une correspondance bi-univoque phonème /h/ - consonne graphique <h> en début de mot (malgré une poignée d'exceptions comme *hour* ou *honor*) ne peut que les aider.

Les études sur la perception du /h/ par des francophones, par exemple Mah et al. (2016) et White et al. (2017), montrent que le problème éventuel ne se situe pas au niveau acoustique. Ces études utilisent les potentiels évoqués (*ERP*, cf. p. 33) pour montrer que les francophones sont parfaitement capables d'entendre le son /h/. Des locuteurs natifs et non-natifs écoutent une suite de syllabes ou de groupes consonantiques ne contenant pas le son /h/, parmi lesquels se trouve de temps en temps une syllabe le contenant (*oddball paradigm*, qui s'apparente à un jeu de « cherchez l'intrus »). Ils constatent que les natifs et les locuteurs L2 ont la même réaction à la présence inopinée de syllabes ou suites de consonnes commençant par /h/. Cette

absence de différence entre les réactions cérébrales des deux groupes de sujets montre que les non-natifs sont tout à fait capables d'entendre le /h/. Cependant, les deux études montrent également qu'en contexte linguistique (*um* vs. *hum* vs. *thumb* pour la première, et avec différents mots comme *happy* vs. *'appy*, *ugly* vs. *hugly* pour l'autre), les francophones de niveau faible n'arrivent plus à détecter la présence ou l'absence inattendues de ce phonème. Le problème se situerait donc probablement au niveau du traitement lexical et non au niveau acoustique.[9]

Les fricatives /θ/ et /ð/, autres consonnes absentes en français, sont également très régulières en anglais. Elles sont toujours notées par le digraphe <th>, qui inversement est presque toujours prononcé avec ces phonèmes mis à part quelques rares exceptions où la prononciation est /t/ (*thyme*, *Thomas*, *Thames*). Un certain nombre d'ouvrages (par exemple *La grammaire orale de l'anglais*, Huart, 2002, p. 71) mentionnent que ces deux fricatives dentales peuvent être assimilées à d'autres fricatives du français en compréhension, notamment /s/ et /z/, qui en sont assez proches, et un travail doctoral (Brannen, 2011, p. 176-177) confirme la difficulté pour les francophones (européens) de distinguer /θ/ de /s/ et /f/ dans des syllabes ouvertes (consonne + voyelle). Un article de Janet Werker et collègues mentionne également que des locuteurs francophones, québécois cette fois, confondent parfois /ð/ avec /d/ (Werker et al., 1992). Dans la première expérience décrite, les sujets entendent des syllabes de forme consonne + /a/ et doivent noter la consonne qu'ils entendent. Les locuteurs anglophones identifient la bonne consonne dans plus de 90% des cas, mais les francophones dans 70% des cas seulement. Dans la plupart des cas erronés (le pourcentage exact n'est pas donné dans l'article), le /ð/ est reconnu comme un /d/ ou un /t/. Cependant, il semble que les locuteurs francophones européens aient moins de mal que les québécois à distinguer la fricative alvéolaire de la plosive alvéolaire (Brannen, 2011, p. 226).

Les réalisations du phonème /r/ de l'anglais et du /ʁ/ du français sont très différentes. Hallé et ses collaborateurs montrent que, probablement parce que le /r/ (du moins en anglais américain) est produit avec des lèvres arrondies, comme /w/, les francophones ont du mal à distinguer ces deux phonèmes (Hallé et al., 1999). Ils font l'hypothèse que /r/ est entendu comme un « mauvais exemplaire » de /w/, et donc qu'il s'agit d'un cas de *Category Goodness*

[9] Des études sur la perception du /h/ japonais (dont les caractéristiques sont très proches du /h/ anglais dans la plupart des contextes) montre que même des locuteurs francophones naïfs n'ayant jamais été en contact avec le japonais font très peu de fautes de discrimination entre des mots commençant ou non par un /h/ (Kamiyama & Nakamura-Delloye, 2015). Ils trouvent là aussi un effet lexical perturbateur (les apprenants débutants ont tendance à préférer les mots qu'ils connaissent et non ceux qui ont la bonne forme sonore).

Difference du modèle PAM (cf. 2.1.1). La discrimination entre les phonèmes est possible mais n'est pas excellente de ce fait.

Les phonèmes /ŋ/ et /ɲ/, qui correspondent tous les deux à des nasales (vélaire pour l'anglais, palatalisée pour le français), ne semblent pas poser de problème en perception. Il est vrai que le /ŋ/ anglais est maintenant également utilisé en français (par exemple, le mot « parking » est en général prononcé /paʁ kiŋ(g)/).

Les affriquées /tʃ/ et /dʒ/ ne sont pas phonémiques en français (les emprunts en français de mots anglais tels que *chips* ou *joker* tendent à utiliser les fricatives /ʃ/ et /ʒ/ plutôt que les sons originels, bien que les deux prononciations soient acceptées). On peut imaginer que les francophones analysent ces phonèmes comme des suites de deux phonèmes, /t/ suivi de /ʃ/ ou /d/ suivi de /ʒ/, ce qui ne devrait pas poser de problème en compréhension. Il est aussi possible qu'ils assimilent /tʃ/ et /ʃ/ d'une part, /dʒ/ et /ʒ/ d'autre part (cas de *Category Goodness Difference*), ce qui peut rendre plus difficile la discrimination de paires minimales telles que *ship* et *chip* ou *leisure* et *ledger*. Cependant, nous n'avons pas trouvé d'études répondant à ces questions.

2.1.2.2. Les voyelles

Les systèmes vocaliques du français et de l'anglais sont beaucoup plus différents que leurs équivalents consonantiques, comme nous le constatons dans le **Erreur ! Source du renvoi introuvable.**.

	Voyelles du français	Voyelles anglais *RP*	Voyelles anglais *GA*
antérieures	i **y** e **ø** ɛ **œ ɛ̃** a	i: **ɪ** ɛ **æ**	id RP
centrales	ə	**ɜ: ʌ** ə	**ɝ: ʌ** ə/ **ɚ**
postérieures	u o ɔ **ɔ̃** **ɑ̃**	u: **ʊ** ɔ: **ɒ ɑ**:	id RP (moins **ɒ**)
diphtongues		**eɪ aɪ ɔɪ əʊ aʊ ɪə eə ʊə**	**eɪ aɪ ɔɪ oʊ aʊ**

Tableau 2.2 - tableau comparatif des voyelles du français et de l'anglais britannique et américain, classées par position de la langue et complexité (en gras, voyelles françaises qui n'existent pas en français, et inversement pour les voyelles soulignées)

Les cellules du Tableau 2.2 sont organisées de la même façon que celles du tableau des consonnes, par position de la langue allant de l'avant vers l'arrière. Les voyelles du français présentent des contrastes qui n'existent pas en anglais : non arrondie/ arrondie, comme /i/ (voyelle antérieure haute non arrondie, comme dans « mis ») vs. /y/ (idem mais arrondie, comme dans « mû »), ainsi qu'un contraste orale/ nasale qui est également non phonémique

en anglais (ɔ/ ɔ̃, « pote » vs. « ponte »). Quand plusieurs voyelles utilisent la même position de la langue en français, elles sont indiquées dans le tableau de la façon suivante : la première est orale non arrondie, la deuxième arrondie et la dernière, le cas échéant, nasale.

Les voyelles de l'anglais comportent également des distinctions inconnues des francophones : d'une part, il existe un contraste voyelles simples / voyelles diphtonguées (*met* vs. *mate*), et d'autre part, un contraste voyelles longues/ courtes (*meet* vs. *mitt*), qui correspond aussi à des différences de qualité (parfois analysées comme des contrastes tendu/relâché[10]). On peut visualiser plus facilement les différences de qualité sur les trapèzes vocaliques suivants (Figure 2.2) représentant les voyelles non diphtonguées, tirés de Capliez (2016), avec le trapèze des voyelles du français à gauche et celles de l'anglais (britannique) à droite :

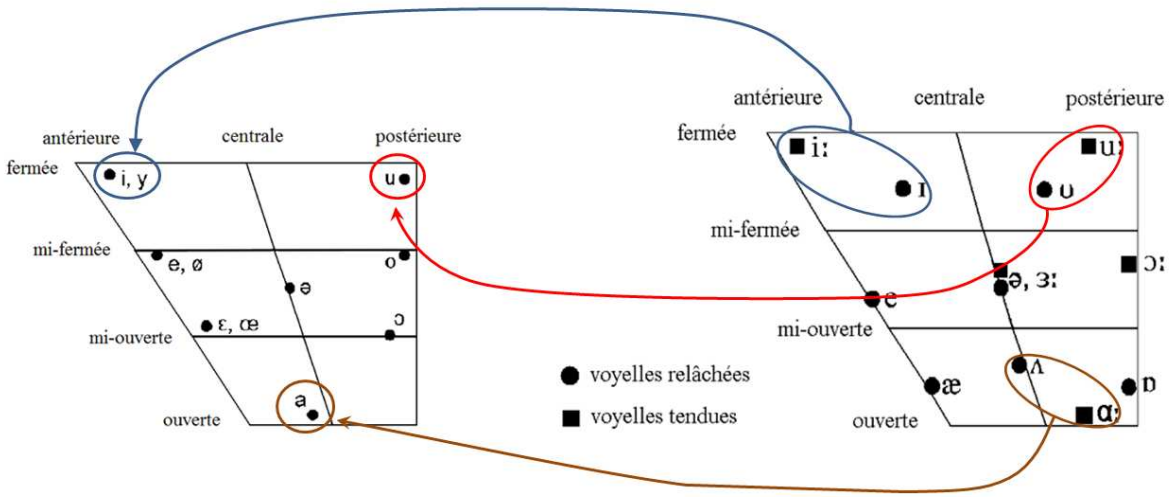


Figure 2.2 - trapèzes des voyelles simples du français (à gauche) et de l'anglais britannique (à droite), avec les contrastes problématiques entourés en bleu (voyelles antérieures fermées), en rouge (voyelles postérieures fermées), et en marron (voyelles ouvertes centrales/postérieures)

A partir de ces deux représentations, on peut imaginer que les contrastes iː/ɪ et uː/ʊ correspondraient à des cas de *Category Goodness Difference* du PAM, avec un phonème (la voyelle longue ou tendue, respectivement iː et uː) correspondant à un « bon » exemplaire du phonème français correspondant (i et u), et la voyelle courte ou relâchée, (ɪ ou ʊ), correspondant à un « mauvais » exemplaire. La distinction devrait donc être possible mais difficile. Pour le contraste /ɑː/ vs /ʌ/, il risque d'être plus difficile à percevoir si les deux phonèmes sont analysés comme des mauvais exemplaires du phonème français /a/. Enfin, on peut également s'attendre à une mauvaise discrimination entre les deux voyelles postérieures

[10] même si ces catégories sont controversées cf. Durand (2005)

/ɑ:/ (*palm*) et /ɒ/ (*pot*) étant donné leur proximité en anglais britannique (cependant, le trapèze n'indique pas si les voyelles sont arrondies ou non, ce qui peut aider à différencier ces deux voyelles dans la mesure où /ɒ/ est arrondie mais pas /ɑ:/). En anglais américain, la voyelle /ɒ/ n'est pas utilisée dans la plupart des dialectes sauf en Nouvelle Angleterre (Labov et al., 2005, p. 13), et le « o » de *pot* n'est pas arrondi pour la plupart des locuteurs (Wells, 1982). La question de la distinction entre /ɑ:/ et /ɒ/ ne se pose donc pas.

Les études sur la perception des francophones indiquent effectivement que la perception des phonèmes vocaliques de l'anglais est plus difficile que celle des consonnes. Iverson et ses collègues étudient l'effet d'un entraînement à haute variabilité sur la perception des voyelles britanniques par des sujets francophones (Iverson et al., 2012). L'entraînement à haute variabilité consiste à faire entendre aux sujets le même phonème dans différents contextes (différents mots), et prononcé par différents locuteurs anglophones. Ils trouvent que les contrastes les moins maîtrisés sont dans l'ordre croissant de réussite avant entraînement (nous avons gardé ceux réussis à moins de 70%) /ɒ/ vs. /ɔ:/ (*pot* et *port*), /i:/ vs. /ɪ/ (*heel* et *hill*), /ɑ:/ vs /ɒ/ vs. /æ/ (*part*, *pot* et *pat*), /ɒ/ vs. /ʌ/ (*cot* et *cut*), /ʌ/ vs /ɜ:/ (*bud* et *bird*), /əʊ/ vs /ɔ:/ (*hole* et *haul*). L'exercice proposé consiste à trouver l'intrus, comme dans les études sur les consonnes présentées plus haut. La paire /i:/ - /ɪ/, qui est souvent mise en avant, peut-être du fait de sa « charge fonctionnelle » (A. Brown, 1988), c'est-à-dire parce qu'elle permet de distinguer un grand nombre de paires minimales, n'est donc pas la seule paire de phonèmes vocaliques difficiles à différencier pour les locuteurs francophones. Cependant, dans l'étude d'Iverson et collègues, c'est la seule qui soit (beaucoup) moins bien réussie après entraînement qu'avant, attestant de sa difficulté résistante à l'expérience. Dans ce même article, un exercice d'identification est proposé aux auditeurs francophones, qui, entendant un mot de la forme /b V t/ (c'est-à-dire une voyelle précédée du phonème /b/ et suivie de /t/), doivent choisir à l'écran parmi un certain nombre de mots orthographiés, celui qui correspond à ce qu'ils ont entendu. Par exemple, ils entendent /bi:t/ et doivent choisir entre *bit*, *beat*, *bot*, *boot*, *bat*, etc… (chacun de ces mots est accompagné d'un autre mot courant contenant la même voyelle (*bit* est accompagné de *hit*, *beat* de *meet*, *bot* de *hot*, …), afin d'aider les sujets à se représenter le mot). Les voyelles les moins bien identifiées dans ce cas de figure sont /i:/, /ɑ:/, /ɜ:/ (20% de réussite), /əʊ/, /ɒ/ et /ɔ:/ (30%). Cependant, l'amélioration après entraînement à haute variabilité est nette pour tous ces phonèmes, avec un taux de réussite allant de 50% à plus de 60%. Les auteurs supposent qu'une partie au moins de cette

amélioration est due à une meilleure acquisition de la correspondance graphie-phonie chez les sujets francophones.

Une autre étude récente (Krzonowski et al., 2016), avec des sujets correspondant mieux à notre public (étudiants francophones en première année de licence de langue étrangère LEA ou LLCER), confirme la difficulté de discrimination /i:/ vs /ɪ/ (moins de 70% de réussite), mais rajoute une paire encore moins bien réussie (60%) : /ɑ:/ vs /ʌ/, tandis que la paire /æ/ vs. /ʌ/ est bien réussie (80% de réussite environ), et ne sera probablement pas candidate à être incluse dans les tests de discrimination phonémique que nous allons construire.

2.1.3. Conséquences pour la reconnaissance lexicale

2.1.3.1. difficulté accrue ?

Weber et Broersma (2012) (résumant les résultats d'expériences faites avec Anne Cutler) montrent que la mauvaise discrimination de phonèmes par les apprenants de L2 conduit à une activation lexicale trop importante. Les apprenants néerlandais qui ne font pas la distinction entre les phonèmes anglais /æ/ et /ɛ/ activent à la fois *flash* et *flesh* quand ils entendent *flash* (Broersma, 2012) et le même phénomène est constaté pour les apprenants castillanophones du catalan, qui ne font pas la distinction entre /e/ (fermé, comme dans « parlé ») et /ɛ/ (ouvert, comme dans « parlait » ou « lève ») (C. Pallier et al., 2001). Comme le nombre de paires minimales en anglais n'est pas énorme, comparé au nombre d'homophones comme *rain-reign* (Cutler, 2005), on pourrait penser que le problème de non-discrimination ne contribue que peu à l'accroissement de la difficulté de reconnaissance lexicale[11]. Cependant, nous avons vu que l'activation lexicale n'attendait pas la fin du mot et se déclenchait dès la fin de la première syllabe. Cela veut dire, pour l'exemple des phonèmes /i:/ et /ɪ/, que non seulement l'apprenant sera susceptible de confondre les paires minimales *ship* et *sheep*, *bit* et *beat*, etc, mais surtout que, en entendant le début /bɪ/, l'apprenant activera non seulement *bin*, *bit*, *bicker*, mais aussi *bean*, *beat*, *beach*, etc…, ce qui conduit à un accroissement beaucoup plus important du nombre de mots activés.

Or, on sait depuis Bradlow et Pisoni (1999) que plus le nombre de mots concurrents est important, moins la reconnaissance lexicale est rapide. Les mots qui ont beaucoup de voisins,

[11] par exemple, le site http://www.minpairs.talktalk.net/minimal.html mentionne 500 paires minimales pour /i:/ /ɪ/, dont la moitié sont des mots différents, soit environ 250 paires, dont certaines comprennent des mots très peu courants (*Aries eyries*)

c'est-à-dire qui ont beaucoup de mots qui leur ressemblent, mettent en effet plus de temps à être reconnus que les autres. L'effet de la densité lexicale du voisinage est encore plus sensible chez les auditeurs non-natifs, peut-être justement parce que le voisinage est encore plus dense chez eux faute de discrimination phonémique suffisante. On pourrait donc s'attendre à ce que la discrimination phonémique soit un facteur important qui explique une part non négligeable de la variance en compréhension de l'oral chez les apprenants de L2, mais cette hypothèse ne semble pas validée par les études existantes.

2.1.3.2. Absence de conséquence au niveau supérieur

Nous avons montré que la discrimination phonémique chez les apprenants étrangers est moins bonne que chez les natifs, et que cela devrait avoir des conséquences sur les performances en compréhension de l'oral des apprenants. Cependant, on constate souvent une absence de corrélation entre la conscience phonémique et la compréhension de l'oral. C'est ce que Ann Bradlow appelle en anglais « *the 'scaling up' problem* » (le problème du passage à l'échelle), qu'elle définit ainsi: « *the problem that patterns of phoneme discrimination and identification are very often not directly reflected in patterns of word or sentence recognition* » (Bradlow, 2007, p. 55). De nombreuses études peinent effectivement à trouver une relation entre la capacité à identifier les phonèmes et celle à comprendre les mots ou les phrases.

2.1.3.2.1. bonne discrimination phonémique sans identification lexicale

Dans certaines études, les locuteurs L2 ont une bonne discrimination phonémique, mais celle-ci ne semble pas forcément aider dans la reconnaissance lexicale. Díaz et ses collaborateurs ont par exemple étudié la discrimination æ/e chez des néerlandais parlant couramment l'anglais L2 (Díaz et al., 2012). Sur leurs 55 sujets, 24, soit près de la moitié, ont obtenu des résultats similaires à ceux des natifs à la tâche de catégorisation phonémique (où ils entendaient l'une ou l'autre voyelle prononcée isolément par un natif britannique). Dans la tâche de décision lexicale, par contre, où les sujets devaient décider si les mots qu'ils entendaient existaient en anglais (par exemple *lemp* plutôt que *lamp*), seules 7 personnes obtiennent cette fois des résultats similaires aux natifs. Les sujets néerlandais reconnaissent bien les mots existants, mais la différence se fait sur les non-mots, qui ne sont pas rejetés par les néerlandais (ils acceptent 70% de non-mots comme *lemp*). On peut d'ailleurs remarquer que la tâche est difficile puisque les résultats des natifs n'atteignent pas le plafond non plus pour les mots en /e/ (40% de non-mots en /e/ acceptés à tort, les résultats étant bien meilleurs pour le phonème /æ/). Enfin, une tâche d'identification lexicale testait la capacité des sujets à

identifier quel membre d'une paire minimale ils entendaient (*cattle* ou *kettle*), en sélectionnant l'image correspondante (accompagnée de la forme orthographique du mot). Les sujets néerlandais ont des résultats similaires aux natifs quand le mot présenté contient un /æ/, mais seuls 5 d'entre eux jouent dans la cour des natifs pour le /e/.

On peut tirer deux conclusions de cette expérience. Premièrement, une bonne discrimination des phonèmes hors contexte ne se traduit pas forcément par une bonne discrimination quand ils sont utilisés dans un contexte lexical (le même résultat a d'ailleurs été démontré en acquisition L1 : « *in first language acquisition research, it has been shown that infants' ability to discriminate a contrast does not necessarily mean that they can use the contrast to discriminate words* » (Hayes-Harb, 2007, p. 66). Deuxièmement, cette mauvaise discrimination en contexte lexical n'empêche pas les apprenants d'atteindre un haut niveau en L2. En effet, dans cette étude, les sujets néerlandais, dont certains ont séjourné dans un pays anglophone, suivent tous des cours en anglais à l'université et se décrivent comme ayant un très haut niveau d'anglais (*highly proficient*).

2.1.3.2.2. mauvaise discrimination phonémique, bonne compréhension

Deux études confirment qu'il est tout à fait possible d'avoir une bonne compréhension dans une langue étrangère dont on ne discrimine pas tous les phonèmes. Baese-Berk et Samuel (2016), par exemple, ont trouvé des problèmes de perception phonologique de base chez 30 hispanophones apprenants de basque, pourtant évalués à un niveau de compétence intermédiaire dans cette langue (y compris en compréhension) : aucun d'entre eux n'entend la différence entre les trois fricatives du basque (lamino-dentale, apico-alvéolaire et post-alvéolaire) après plusieurs années d'expérience avec cette langue. Escudero et Wanrooij (2010) étudient l'influence des connaissances orthographiques sur la perception des phonèmes, mais incluent dans leur étude une tâche purement auditive, où les 204 sujets hispanophones apprenant le néerlandais doivent discriminer deux phonèmes proches avec une tâche XAB (ils entendent un stimulus naturel d'un phonème X, et doivent choisir entre 2 stimuli artificiels A et B celui qui se rapproche le plus de X). La moitié de ces hispanophones ont un niveau A1 sur l'échelle du CECR selon le test de compréhension de l'oral de DIALANG, l'autre moitié se classant au niveau C2, et les chercheurs n'ont trouvé aucune différence entre la perception des voyelles du néerlandais entre ces deux groupes (le néerlandais ayant une quinzaine de voyelles contre cinq pour le castillan).

Nous ne nous attendons pas à ce que les francophones aient des problèmes particuliers de discrimination entre les phonèmes /ɛ/ et /æ/, mais on peut s'attendre au vu de cette étude à ce que la non-discrimination entre /i:/ et /ɪ/, par exemple, ne soit pas rédhibitoire pour la compréhension de l'oral. D'autres études corrélatoires plus générale, que nous présentons ensuite, ne trouve pas de corrélation importante.

2.1.3.2.3. étude corrélatoire

Dans sa thèse, Naouel Zoghlami (Zoghlami, 2015, p. 174) trouve que la variable « discrimination auditive », même si elle est faiblement corrélée au résultat en compréhension de l'oral (0,39), ne rajoute pas de pouvoir explicatif par rapport aux autres variables dans une analyse de régression (étude menée auprès d'apprenants francophones et arabophones de l'anglais L2/ L3). La tâche est tirée de l'*Oxford Placement Test* (Allan, 2004), et comprend des items tels que *The team need new shirts/ shorts*. Une corrélation très proche (0,36) entre CO et discrimination phonémique a été trouvée chez des lycéens japonais, avec une tâche d'identification de phonème entendu (I. Wilson et al., 2011). Une autre étude avec des élèves de primaire français débutants en anglais (Hilton et al., 2016) montre une corrélation encore plus faible entre discrimination auditive et compréhension de l'anglais oral (0,28).

2.1.3.3. Tentatives d'explication

Comment expliquer cette corrélation faible ou absente entre discrimination phonémique et compréhension de l'oral? Deux pistes d'explication peuvent être trouvées dans la littérature, qui sont intéressantes à la fois pour le développement d'un test de sensibilité phonémique, et pour la prise en compte d'autres facteurs. D'une part, le phonème n'est peut-être pas l'unité de base la plus pertinente en perception de la parole, et c'est peut-être notre système d'écriture alphabétique qui nous influence à tort en nous poussant à privilégier l'unité phonémique aux dépens d'autres unités possibles (Wauquier-Gravelines, 1999). D'autre part, comme nous l'avons vu dans la partie présentant les modèles actuels en compréhension de l'oral, le cerveau fait feu de tout bois pour traiter le signal entrant, et les processus descendants peuvent en partie compenser un éventuel déficit d'activation aux niveaux inférieurs de traitement, en particulier pour les apprenants L2. Nous analyserons ces deux hypothèses tour à tour.

2.1.3.3.1. mise en cause du statut du phonème

Dès le début des analyses spectrales dans les années 1950, le statut du phonème s'est posé (Hawkins, 2004). Malgré de nombreuses recherches, il n'a jamais été possible de trouver un

invariant qui caractérise toutes les réalisations d'un même phonème (quels que soient sa position dans un mot, le contexte droit et gauche, la vitesse d'élocution, le sexe du locuteur, etc..). En effet, l'information « phonémique » n'est pas présente uniquement dans un segment, mais au contraire est distribuée sur les segments adjacents, au point qu'une distinction en finale peut avoir une conséquence dès le début du mot. Coleman (2003) rappelle par exemple qu'une distinction de voisement en finale (*lent* vs. *lend*) peut s'entendre dès le /l/ initial qui est légèrement plus long et sombre (vélarisé) pour *lend* que pour *lent*.

Certains chercheurs ont donc émis l'idée qu'une unité plus grande, la syllabe en l'occurrence, serait plus adéquate. Dès 1952, Cooper et collaborateurs déclaraient que « *the perception of these stimuli, and also, perhaps, of their spoken counterparts, requires the consonant-vowel combination as a minimal acoustic unit* » (F. S. Cooper et al., 1952, cité par Hawkins 2004, p. 12). Greenberg (1999), qui étudie un corpus de productions spontanées, plaide lui aussi pour la syllabe : d'une part les mots (anglais) sont pour la plupart monosyllabiques à l'oral (80% des instances dans son corpus), et d'autre part, il y a moins de variation au niveau de la syllabe, ou plus précisément, la variation peut être systématisée au niveau de la syllabe. L'attaque est en effet en général préservée (dans près de 85% des cas, et même 90% en cas d'attaque complexe avec plusieurs consonnes), tandis que la voyelle noyau est souvent modifiée par rapport à la forme de citation (35% des cas), et que le coda est souvent non réalisé (28% du temps). Greenberg souligne que la préservation de l'attaque est probablement due à son importance en peception, en particulier pour la segmentation. De plus, l'accent opère au niveau de la syllabe. Harley (2007) rappelle par ailleurs que la syllabe est une unité plus accessible à la conscience que les phonèmes. C'est le cas chez les enfants avant l'apprentissage de la lecture (Bosse & Zagar, 2016, p. 573, appellent la syllabe « l'élément sublexical le plus facilement accessible aux enfants non lecteurs »), mais aussi chez les adultes n'ayant pas appris à lire avec un système alphabétique (comme C. Read et al., 1986, l'ont montré avec des adultes lecteurs de chinois mandarin). Avant ou sans l'apprentissage d'une langue alphabétique, le cerveau humain a beaucoup de mal à décomposer le langage en éléments infra-syllabiques.

D'un autre côté, le phonème est peut-être au contraire une unité trop grande pour la perception de la parole. Nous avons vu en effet que les locuteurs utilisent des informations allophoniques pour traiter le signal sonore. Les études décrites dans le premier chapitre montrent que les locuteurs natifs savent prendre en compte ces informations, qui leur permettent de segmenter

le signal même quand ils ne connaissent pas certains mots. D'autres études montrent que les apprenants L2 sont également capables de repérer les indices infra-phonémiques. Altenberg (2005), Ito et Strange (2009) et Shoemaker (2014) arrivent à des résultats convergents avec des apprenants d'anglais L2 de langue maternelle respectivement espagnole, japonaise et française. A chaque fois, les sujets réussissent assez bien à segmenter les expressions où la frontière lexicale est marquée par un coup de glotte (devant une voyelle), et arrivent à distinguer *a nice man* (sans coup de glotte) de *an ice man* (avec un coup de glotte devant la voyelle de *ice*). Ils ont plus de mal à utiliser l'information apportée par l'aspiration mais ont tout de même des résultats meilleurs que le hasard pour distinguer *keeps ticking* (avec un [t^h] aspiré en début de mot devant une voyelle) de *keep sticking* (sans aspiration après un /s/). Cependant, les stimuli utilisés dans ces études sont assez artificiels car ils ont été enregistrés dans un contexte peu naturel (*Say* ____ *again*) et bien articulés pour les besoins des expériences (en particulier, prononcés un peu plus lentement que la vitesse habituelle : Altenberg, 2005, p. 336), et il n'est pas sûr que les participants auraient eu d'aussi bons résultats avec des stimuli authentiques. Fox Tree et Meijer (2000) montrent par exemple que des stimuli produits par des locuteurs non entraînés donnent des résultats très différents de ceux enregistrés par des locuteurs professionnels (dans leur cas, pour l'interprétation d'informations prosodiques).

Ces expériences démontrent tout de même que les allophones et, plus généralement, les variations infra-phonémiques ne sont pas simplement des variantes en contexte que le locuteur utilise pour se simplifier la vie en production et qu'il faut « re-normaliser », c'est-à-dire traduire en phonèmes, afin d'arriver à la forme « pure » (de citation). Ces variantes sont certes utilisées par le locuteur pour plus de facilité, parce qu'il est impossible d'articuler les sons indépendamment les uns des autres. En effet, les articulateurs se préparent naturellement à prononcer un nouveau son avant que le précédent soit terminé (coarticulation qui conduit à une assimilation régressive, c'est-à-dire une influence du son suivant sur le son précédent), et ces mêmes articulateurs restent encore un peu dans la position du son précédent après le début du suivant (coarticulation qui conduit à une assimilation progressive, c'est-à-dire une influence du son précédent sur le son suivant). Mais ces variantes aident également l'auditeur, de L2 comme de L1, qui lui aussi se « prépare » à entendre le son suivant dès le son précédent, et qui est surpris si ce n'est pas ce qu'il attend (Dahan & Magnuson, 2006). Cela lui permet aussi de « récupérer » une information qui a pu lui échapper et qui reste disponible sur le segment suivant, rendant ainsi plus robuste la perception de la parole. Loin d'être

assimilable à un alourdissement de la tâche de l'auditeur qui devrait se débarrasser des variations allophoniques avant d'avoir accès à une représentation phonémique et au sens, ces indices infra-phonémiques sont au contraire exploités par l'auditeur pour traiter plus efficacement le signal. Dans cette optique, il n'est donc pas forcément gênant que les apprenants L2 n'aient pas (encore) fait le travail d'abstraction nécessaire à l'analyse des mots en phonèmes (et donc en assimilant toutes les variantes allophoniques à un même phonème), puisque l'utilisation des informations infra-phonémiques peut au contraire les aider (mais rien ne dit non plus qu'ils soient capables d'utiliser efficacement ces informations pour anticiper les sons qui vont suivre).

En conclusion, on peut citer Sara Hawkins selon laquelle « *it is useful to regard phonemes as primarily units of maximal phonological contrast for identification of lexical items in citation form, rather than as an obligatory first stage in understanding connected speech* » (Hawkins, 2004, p. 12). Certains chercheurs comme Goldinger et Azuma (2003) vont plus loin et nient toute « réalité » au phonème comme unité a priori de la perception de l'oral. Pour eux, c'est le contexte, et la tâche qu'on demande aux locuteurs d'effectuer qui rend telle ou telle unité plus ou moins saillante. Goldinger (1998) propose d'ailleurs une théorie exemplariste où les mots sont reconnus sans passer par une étape de reconnaissance d'unités phonologiques intermédiaires, mais simplement par similarité avec des exemplaires de mots stockés en mémoire (chaque exemplaire lexical étant mémorisé avec sa voix et son contexte). Nous reparlerons plus en détail de la théorie des exemplaires dans la troisième partie de ce travail.

2.1.3.3.2. compensation par processus descendants

Nous avons examiné une première explication possible au rôle apparemment peu important de la compétence de discrimination phonémique dans la compréhension de l'oral. Une autre raison réside probablement dans l'utilisation de stratégies compensatoires, et en particulier dans l'exploitation des connaissances lexicales.

Field (2004) a par exemple montré que les auditeurs L2 peuvent passer outre les informations phonémiques quand leurs connaissances lexicales les poussent dans une autre direction : en cas d'informations contradictoires, certains d'entre eux font plus confiance à leurs connaissances lexicales qu'à leur traitement du signal acoustique. Il a fait entendre à 48 lycéens/étudiants internationaux faisant un stage d'anglais dans une école de langue en Angleterre, de niveau élémentaire ou intermédiaire, une liste de phrases se terminant par un

mot rare, qu'ils ne connaissaient probablement pas, mais assez proche dans sa prononciation d'un mot beaucoup plus courant. Les phrases étaient construites de telle façon que le contexte soit favorable au mot rare et non à son voisin plus courant (*They're lazy in that office; they like to shirk*. [*not WORK*]), et les apprenants devaient noter sur une feuille le mot entendu en fin de phrase. Bien qu'un certain nombre d'entre eux se soient clairement basés sur le signal pour proposer une réponse (juste ou parfois fausse, en proposant un non-mot), ou n'aient pas répondu, un tiers de réponses étaient des substitutions de mots plus courants, y compris, dans la moitié des cas, d'une catégorie syntaxique ne convenant pas dans le contexte. Il semble donc que ces apprenants (de niveau assez faible) ne se fient pas à l'information phonémique, mais qu'ils ont au contraire une stratégie lexicale, qui a pu être encouragée par la tâche qui leur demandait de noter un seul mot. On peut aussi remarquer que, s'il s'agit d'une stratégie descendante (le niveau lexical prime sur l'information acoustique), la stratégie reste de niveau assez peu élevé, puisqu'il n'y a pas de prise en compte du contexte sémantique et syntaxique. Une deuxième expérience visait à étudier l'interaction entre une éventuelle stratégie contextuelle et les indices acoustiques, en faisant écouter aux mêmes apprenants des phrases dont le dernier mot, courant et contraint par le contexte, était remplacé par un autre, aussi fréquent mais moins contraint par le contexte, tout en restant plausible (*I couldn't listen to the radio because of the boys*. [NOISE] ou *The people at the party were Germans, Italians, Spanish and some friends* [FRENCH]). Les substitutions dans ce cas sont moins nombreuses que dans l'expérience décrite précédemment, mais ont lieu dans 7 des 20 phrases, avec des pourcentages variant de 15 à 60% de substitutions pour ces 7 items. Il n'est donc pas rare que les informations contextuelles priment sur celles qui viennent du signal.

D'autres études avaient déjà montré que les connaissances antérieures pouvaient pallier les manques linguistiques. Long (1990) a fait écouter à 188 étudiants américains inscrits en troisième trimestre d'espagnol (niveau intermédiaire) deux textes pour lesquels ils avaient des connaissances préalables soit sommaires (la ruée vers l'or en Californie au 19ème siècle), soit étoffées (groupe de rock U2). Les résultats au premier test étaient corrélés à leur note d'espagnol du trimestre précédent, ce qui n'était pas le cas pour le deuxième. Donna Long en conclut que les connaissances préalables dans le deuxième cas peuvent permettre de suppléer aux connaissances linguistiques défaillantes, alors que pour le premier texte, les étudiants étaient obligés de se reposer presque uniquement sur leurs connaissances linguistiques, d'où la corrélation avec leur note de langue.

Enfin, l'étude à grande échelle de Tsui et Fullilove (1998) a aussi montré que l'utilisation de schémas déduits du contexte au début de l'écoute (ou même avant, en utilisant la question et les réponses possibles lisibles avant le début de l'écoute d'un test de compréhension sous format QCM) aide les apprenants L2 (dans ce cas des lycéens de Hong Kong apprenant l'anglais) à condition que ce schéma soit utilisable jusqu'à la fin du texte. Les questions à schéma cohérent entre le début et la fin du texte sont en effet mieux réussies que celles à schéma discordant, en particulier pour les candidats plus faibles. Cela montre que l'utilisation des informations contextuelles, et donc la compensation par processus descendant, facilite la compréhension.

2.1.4. Conclusion

Nous avons vu que l'utilisation de connaissances lexicales, ou plus généralement du contexte, pouvait primer sur les informations tirées du traitement du signal acoustique, et pouvait ainsi compenser les insuffisances du traitement phonémique pour les L2. On peut d'ailleurs rappeler ici que même chez natifs, la discrimination phonémique hors contexte ne se fait pas sans problèmes, comme en témoignent les matrices de confusion dont les premières ont été calculées par George Miller et Patricia Nicely (1955) pour les consonnes (les fricatives sont par exemple moins bien reconnues que les autres consonnes, en particulier /θ/ et /ð/ qui sont souvent prises pour /s/ et /z/) ou Peterson et Barney (1952) pour les voyelles (une des voyelles posant le plus de problèmes est /ɛ/, souvent confondue avec /ɪ/ ou /æ/).

A la lumière de ce qui précède, on ne s'attend pas à ce que la discrimination phonémique joue un rôle crucial dans la compréhension de l'oral en L2. Nous avons en effet vu plusieurs fois que le problème éventuel ne semblait pas tant se situer au niveau de la perception elle-même, même si les effets de l'aimant perceptuel sont indéniables, qu'au niveau lexical. Les apprenants L2 peuvent apprendre à entendre la différence entre certains phonèmes de la L2, mais cela n'est pas forcément suffisant pour qu'ils utilisent ensuite ces informations lors de la reconnaissance lexicale, pour écarter les mots qui ne correspondent pas au schéma qu'ils entendent. Ce phénomène peut être relié au modèle *Chunk-and-Pass* que nous avons décrit dans le premier chapitre. Lorsque des unités d'un niveau inférieur (ici celui des unités phonologiques) sont regroupées (*chunked*) pour former une unité de niveau supérieur (le mot, au niveau lexical), une partie de l'information est perdue. C'est ce que ses concepteurs appellent le *lossy chunking*, ou « regroupement avec perte de détails » (Christiansen & Chater, 2016, p. 8). On peut imaginer que, pour les apprenants L2, les différences acoustiques qu'ils

sont capables de percevoir n'ont pas encore atteint le statut de phonème de la langue étudiée et ne remontent pas au niveau lexical. Le /h/ qu'ils sont capables d'entendre, par exemple, ne remonte pas dans leur représentation du mot *happy*, et son absence ne sera pas remarquée.

Nous allons tout de même essayer dans nos expérimentations d'isoler l'importance de la discrimination phonémique pour nos étudiants en construisant un test où les phonèmes sont traités au niveau de la syllabe, vu son statut privilégié dans le traitement du signal acoustique, suivi d'une partie où ils seront utilisés dans des mots, pour essayer d'isoler les effets éventuels du biais lexical. Nous tirerons parti des études résumées plus haut sur la difficulté de perception de certains phonèmes par les francophones pour construire des items qui soient discriminants (ni trop faciles, ni trop difficiles) pour notre public. La construction de ces tests sera décrite dans la deuxième partie de cette thèse.

2.2. Le rôle du suprasegmental

Du fait de notre biais alphabétique (Wauquier-Gravelines, 1999), il est assez facile de se représenter le signal sonore comme une suite de phonèmes qu'il faut reconnaître progressivement. Cependant, parce que la prosodie n'est pas marquée dans notre système d'écriture, il est plus difficile de ne pas négliger son importance potentielle en compréhension de l'oral. D'après Di Cristo (2013, p. 2), la prosodie est le « champ d'étude d'un ensemble de phénomènes, tels que l'accent, le rythme, les tons, l'intonation, la quantité, les pauses et le tempo qui constituent ce qu'il est convenu d'appeler les éléments prosodiques ou éléments suprasegmentaux du langage ». C'est donc un phénomène supra-segmental, qui a un champ d'application supérieur aux segments (les sons individuels), pouvant concerner les syllabes ou des groupes de syllabes (groupes de souffle par exemple). Ces syllabes peuvent être plus ou moins accentuées, c'est-à-dire varier (Frost, 2011; Huart, 2010) :

- en intensité : les syllabes accentuées sont prononcées avec plus de force et sont entendues comme plus « fortes » (*loud* en anglais) ;
- en longueur : les syllabes accentuées sont plus longues que les non accentuées (toutes choses égales par ailleurs) et durent donc plus longtemps ;
- en hauteur mélodique : les syllabes accentuées sont produites avec une fréquence fondamentale plus haute et sont donc perçues comme plus aiguës ;
- en degré d'articulation : les formants de transition d'un son à l'autre dans les syllabes accentuées sont plus évidents, les syllabes accentuées sont mieux articulées (Cho, 2005;

Lindblom, 1990), et les voyelles par exemple y sont plus facilement identifiées (Warner & Cutler, 2017).

En anglais, l'accentuation de certaines syllabes implique la désaccentuation (et souvent la réduction) des autres syllabes, ce qui est tout aussi important à percevoir pour l'auditeur (et peut-être plus difficile pour l'apprenant). Au niveau supérieur aux syllabes, la perception du rythme (succession de syllabes de longueur plus ou moins importante) et de l'intonation (variation de la courbe mélodique, c'est-à-dire de la fréquence fondamentale, sur différentes parties de l'énoncé) joue également un rôle dans la compréhension des énoncés.

Cependant, il doit être clair qu'il n'est pas facile de séparer niveaux segmental et suprasegmental, puisque le suprasegmental a une influence sur les segments et se réalise linéairement dans la suite de segments, même s'il se laisse mieux analyser comme quelque chose qui se « superpose » aux segments. Une syllabe plus longue, par exemple, implique que la voyelle et peut-être la ou les consonnes qui la composent soient elles-mêmes allongées. Ainsi que le remarque Roach (2009, p. 69) cité par Capliez (2016, p. 1), « [*the use of the term suprasegmentals*] *sometimes give[s] the misleading impression that prosody is something optional, added like a coat of paint, when in reality at least some aspects of prosody are inextricably bound up with the rest of speech* ».

2.2.1. Difficultés pour les apprenants francophones

2.2.1.1. accent fixe et accent libre

Nous avons vu que les phonèmes ou les contrastes phonémiques de L2 qui n'existent pas dans la langue maternelle peuvent être difficiles à percevoir du fait que notre système perceptif a été peu à peu sculpté par notre L1 et a du mal à s'adapter à d'autres langues : selon l'expression de Patricia Kuhl (1993), nous sommes à la naissance des « citoyens du monde » (*citizens of the world*), mais nous devenons ensuite « prisonniers de notre langue maternelle » (*language-bound*). L'accent lexical a lui aussi un statut très différent en anglais et en français : il est contrastif en anglais, mais pas en français. En français, l'accent est porté par la dernière syllabe du mot ou groupe de mots (Christophe et al., 2004), qui est allongée par rapport aux autres syllabes, mais n'est jamais contrastif, c'est-à-dire qu'il n'est pas utilisé pour différencier deux mots de sens différent. En anglais, en revanche, il existe des paires minimales qui ne diffèrent que par l'accent (ˈ*perfect* accentué sur première syllabe, « parfait », et *per*ˈ*fect* accentué sur la deuxième syllabe, « parfaire »), et il est donc important

de le percevoir (et de le placer correctement en production). L'accent est libre en anglais, ce qui veut dire que la place de l'accent lexical y est imprévisible (H. Altmann, 2006). Cela peut être illustré avec les tableaux d'accentuation lexicale (Tableau 2.3) qui aident à visualiser le contraste entre le français et l'anglais, tirés de Delattre (1965, p. 29), cité par Frost (2011)[12]. On y voit que toutes les positions sont exploitées par l'anglais, tandis qu'en français seule la dernière syllabe peut recevoir un relief particulier:

Mots de				
1 syllabe	100%			
2 syllabes	0%	100%		
3 syllabes	0%	0%	100%	
4 syllabes	0%	0%	0%	100%
Syllabes :	1ère	2ème	3ème	4ème

Français

Mots de				
1 syllabe	100%			
2 syllabes	74%	26%		
3 syllabes	55%	39%	6%	
4 syllabes	33%	36%	29%	2%
Syllabes :	1ère	2ème	3ème	4ème

Anglais

Tableau 2.3 - comparaison des syllabes accentuées en français et en anglais (d'après Delattre 1965, p.29)

Nous allons voir dans le prochain paragraphe les difficultés que cela peut susciter chez les francophones.

2.2.1.2. la surdité accentuelle

Un concept qui a eu beaucoup de retentissement est celui de surdité accentuelle (*stress deafness*), qui a été développé par Emmanuel Dupoux et ses collaborateurs (Dupoux et al., 1997, 2008, 2010). Ils montrent que les auditeurs francophones, qui n'ont pas besoin de tenir compte de l'accent lexical dans leur langue maternelle, ont des difficultés à percevoir l'accentuation. Cependant, ces difficultés ne se situent pas à un niveau acoustique pur : dans des tâches de discrimination simple (AX), où il s'agit de distinguer deux exemplaires (*tokens*) de mots inventés (prononcés par un néerlandophone) qui ne contrastent que par l'accentuation, les locuteurs français réussissent aussi bien que les espagnols, pour qui l'accent est contrastif en L1. Les difficultés commencent dès qu'il y a plus de variabilité phonétique, et donc que les francophones doivent s'abstraire des exemplaires utilisés et construire une représentation plus schématique, par exemple dans une tâche de type AXB avec des locuteurs différents. Dans ce protocole, les sujets francophones entendent deux mots de trois syllabes (A et B) qui ne

[12] Une étude plus récente (Clopper, 2002, p. 5) arrive à des résultats similaires. Si l'on convertit en pourcentages les chiffres bruts donnés dans cette étude nous arrivons aux résultats suivants : pour les mots de 2 syllabes, accentuation à 78% et 22% respectivement sur les première et deuxième syllabe, pour les mots de 3 syllabes, 58%, 34% et 8%, et pour les quatre syllabes, 17%, 46%, 35% et 1%.

diffèrent que par l'accentuation, prononcés par une locutrice, et un mot (X) prononcé par un locuteur, et doivent indiquer si celui-ci est semblable à A ou à B. Ils commettent alors environ 20 % d'erreurs (contre moins de 5 % pour les hispanophones). Ils ont donc des difficultés à s'abstraire de la différence de voix et à construire une représentation qui prenne en compte, au-delà de la différence de timbre et de hauteur, le fait que deux exemplaires différents soient accentués sur la même syllabe. L'équipe de Dupoux montre ensuite (Dupoux et al., 2008) que le problème principal se situe au niveau lexical : quand ils sont testés avec de vrais mots espagnols bien ou mal accentués dans une tâche de décision lexicale, les francophones apprenant l'espagnol L2 acceptent autant de mots mal accentués (environ 60%) quel que soit leur niveau, alors que le nombre de mots bien accentués qu'ils reconnaissent correctement augmente avec le niveau. Il semble tout simplement que la syllabe accentuée ne soit pas encodée dans l'entrée lexicale des mots qu'ils connaissent en espagnol, et que ces entrées soient donc sous-spécifiées sur ce plan. Encore une fois, il peut s'agir d'une conséquence du *lossy chunking*, à savoir que quand on passe du niveau phonologique (inférieur) au niveau lexical (supérieur), certaines informations comme l'accentuation sont perdues. Tous les mots qui ont la bonne forme phonémique sont acceptés, quelle que soit leur accentuation (correcte ou incorrecte). Nous retrouverons à plusieurs reprises cette difficulté à rejeter des formes incorrectes chez les apprenants L2.

Dans sa thèse, Heidi Altmann (2006) étudie également la perception (et la production) de l'accent lexical par des locuteurs de différentes langues, y compris l'anglais et le français, en utilisant des mots inventés (*nonce words*), à l'aide d'une tâche où les auditeurs doivent entourer la syllabe qu'ils pensent accentuée sur une représentation orthographique des mots qu'ils entendent. Elle montre encore une fois que les locuteurs de langues à accent libre, comme l'anglais ou l'espagnol, ont de bien meilleurs résultats que ceux de langues à accent fixe, comme le français ou l'arabe.

Ces résultats, qui ont été mis en lumière soit sur des mots inventés, soit sur l'espagnol, se retrouvent-ils en anglais L2 ? Annie Tremblay (2008) étudie spécifiquement la perception de l'accent en anglais L2 par des locuteurs de français (canadien) L1. Dans son expérience d'amorçage intermodal (*cross-modal priming*), des locuteurs de français et d'anglais L1 écoutent le début d'une phrase anglaise qui finit par la première syllabe accentuée ou non d'un mot polysyllabique (*Very few still remember the MYS-*, pour *MYStery,* ou *Very few still remember the mis-*, pour *misTAKE*). Deux mots qui sont compatibles d'un point de vue

segmental avec ce début apparaissent ensuite à l'écran, dont l'un correspond aussi au schéma accentuel (par exemple *MYStery*, c'est la cible), et l'autre non (*misTAKE*, le distracteur). Les apprenants francophones, quel que soit leur niveau, plafonnent à 60% de réponses correctes quand la première syllabe est accentuée, tandis que les Canadiens anglophones natifs ont 73% de bonnes réponses[13]. Quand la première syllabe est inaccentuée, la performance des sujets travaillant en L2 augmente avec le niveau. Celle des apprenants intermédiaires et intermédiaires forts (qui habitent depuis un an en moyenne au Canada) est au niveau du hasard (50%), tandis que celle des apprenants avancés (qui habitent depuis quatre ans en moyenne au Canada) n'est pas significativement différente de celle des natifs, même si elle est un peu plus faible (58 contre 65%).

Nous pouvons tirer deux conclusions de cette expérience. D'une part, la perception de l'accent n'apparaît pas très corrélée avec la compétence générale en langue puisqu'elle n'augmente pas forcément avec le niveau (estimé dans cette étude par un test de closure et une tâche de lecture à haute voix, et corrélé à la durée de l'immersion en contexte L2). D'autre part, cette perception n'est pas très bonne de façon générale chez les francophones, mais les natifs trouvent également la tâche difficile, ce qui renforce le manque de corrélation probable avec la compétence générale.

On peut penser cependant que la corrélation avec la compréhension de l'oral sera plus élevée, dans la mesure où cette expérience ne portait que sur l'utilisation de l'accent pour l'accès lexical (pour différencier deux mots). Or, nous avons vu que l'accent est également utilisé en anglais pour la segmentation lexicale (pour repérer où commencent et finissent les mots). Il est possible (comme on l'a vu avec les résultats des natifs) que l'utilisation de l'accent pour différencier des mots soit marginale, ou du moins non nécessaire, mais que son utilisation pour la segmentation soit essentielle. Dans ce cas, la corrélation avec la compréhension de l'oral peut être non négligeable.

2.2.2. Corrélation avec la compréhension de l'oral

A notre connaissance, il existe très peu d'études sur une éventuelle corrélation entre capacité à percevoir l'accent lexical et compréhension de l'oral. Il en existe beaucoup plus,

[13] Ce résultat n'est pas très bon non plus, et peut être expliqué par le fait que l'information accentuelle est souvent redondante en anglais, puisqu'elle est doublée d'une information segmentale : les syllabes accentuées ont toujours une voyelle pleine, tandis que les syllabes inaccentuées ont souvent une voyelle réduite ; de plus, le nombre de paires minimales contrastant par l'accent est assez réduit (Cutler, 2005).

paradoxalement, sur la corrélation avec les capacités de lecture en L1, depuis que Kitzen (2001) a enrichi la théorie de l'origine phonologique de la dyslexie chez les anglophones (*phonological processing deficit theory of dyslexia*) en montrant que ce n'était pas seulement un problème de traitement des phonèmes, mais aussi de la prosodie et du rythme. On ne peut qu'être frappé, d'ailleurs, par le parallélisme entre les troubles décrits chez les dyslexiques et le traitement « normal » de la langue chez les apprenants de L2 dont nous avons vu quelques exemples (voir aussi Hilton, 2009, p. 60).

Premièrement, on remarque l'imprécision et la sous-spécification des représentations phonologiques : « *Phonological representations may be imprecise, degraded, or incomplete in dyslexic individuals* » (Kitzen, 2001, *abstract*), ce qui encore une fois renforce l'idée que la corrélation entre sensibilité à l'accent lexical et compréhension n'est pas forcément très importante si les natifs dyslexiques peuvent avoir l'une (la compréhension de l'oral) sans l'autre (la sensibilité accentuelle). Deuxièmement, la lenteur de traitement (dont nous avons peu parlé, mais qui se retrouve dans toutes les études psycholinguistiques comparant les temps de réaction des natifs et des locuteurs L2, entre autre l'étude de Tremblay (2008) décrite au paragraphe précédent) : « *Deficient speed in retrieving phonological codes may also contribute to impaired reading* » (ibid., *abstract*). Nous reparlerons des études sur les dyslexiques anglophones dans la deuxième partie quand nous décrirons la réutilisation dans le cadre de notre étude des tests utilisés par certaines d'entre elles.

Pour le lien avec la compréhension de l'oral chez les apprenants L2, nous n'avons trouvé qu'un article et une thèse. L'article de Meerman et al. (2014), très court, confirme que les apprenants L2 (de langue maternelle japonaise dans leur cas) réussissent mieux à accepter les mots bien accentués qu'à rejeter les mal accentués. Dans leur analyse utilisant la modélisation par équations structurelles pour traiter les données (*structural equation modeling*, ou *SEM*), ils trouvent que seules les connaissances lexicales prédisent la variance en compréhension de l'oral, et que la sensibilité à l'accent lexical n'est pas corrélée aux connaissances lexicales ni à la compréhension de l'oral. Il semble donc qu'ici aussi, le schéma accentuel ne fasse pas partie de la connaissance lexicale des apprenants. Cependant, il faut prendre ces résultats avec précaution dans la mesure où le test de compréhension de l'oral comporte peu de questions (18 seulement, séparées en 6 variables pour le *SEM*) et que son coefficient de fiabilité (alpha de Cronbach) est assez bas (0,56).

L'autre référence pertinente est la thèse de Megumi Tabata (2016), qui étudie les relations entre l'éducation musicale au rythme, la sensibilité prosodique et la compréhension de l'oral en anglais chez des enfants et adolescents japonais. Elle trouve (p.138) que la compétence prosodique prédit 28% de la variance en compréhension de l'oral (la corrélation entre les deux variables est de 0,50). L'élément le plus important à l'intérieur de la compétence prosodique est la sensibilité à l'accent contrastif. Dans ce test, les sujets entendent 16 phrases de type « *I wanted blue and BLACK socks* » (avec un accent contrastif sur *black*), et voient 2 images à l'écran, l'une avec des chaussettes noires, l'autre avec des bleues. On demande à l'enfant de cliquer sur ce que l'autre personne a oublié (ici les chaussettes noires). On peut se demander s'il ne choisit pas simplement le mot qu'il entend le mieux (parce que plus long et mieux articulé puisqu'accentué), et si ce test ne mesure pas, de ce fait, la compréhension du mot accentué, d'où la corrélation avec le score de compréhension global. Cependant, les autres parties du test de sensibilité prosodique étant également corrélées à la compréhension de l'oral, il y a probablement un effet prosodique. Ces autres parties concernent la capacité à segmenter les mots ou groupes de mots (différencier « *chocolate, biscuits and milk* » de « *chocolate biscuits and milk* »).

Que pouvons-nous retenir de ces études ? Dans l'une, le repérage de l'accent lexical n'est pas corrélé à la compréhension de l'oral chez les L1 japonais, mais les résultats sont basés sur un test de CO un peu court. Dans l'autre, également avec des sujets japonais (mais plus jeunes), l'utilisation de l'accent contrastif et des indices prosodiques de segmentation semble prédire une partie de la compréhension de l'oral. Notre étude avec des sujets francophones pourra peut-être éclairer un peu la question.

2.3. Les connaissances lexicales

Nous avons vu lors de la présentation des modèles de la compréhension de l'oral que le lexique y tenait une place centrale, à tel point qu'une grande partie des recherches en psycholinguistique sur la compréhension de l'oral se focalise sur la reconnaissance des mots. Selon Broersma et Cutler (2008, p. 23), « *Word recognition is the central component of language processing; there is no sentence without the words comprising it, and phonemic contrast only occurs because it distinguishes words* ». Les apprenants L2 n'ont pas besoin de découvrir le concept de mot (contrairement à celui d'accent lexical dont certains ignorent l'existence), d'autant moins quand ce sont, comme dans notre cas, des locuteurs d'une langue indo-européenne (le français L1) qui en apprennent une autre (l'anglais L2), appartenant au

même groupe typologique (langues analytiques avec flexion, dérivation et composition), et dont les échanges lexicaux avec la première ont été très intenses au cours des siècles, depuis la conquête anglo-normande au 11ème siècle jusqu'à l'hégémonie de la langue anglaise dans le monde technologique contemporain (Walter, 2003).

La difficulté pour nos apprenants se situe plutôt, d'une part, dans le nombre de mots à acquérir, et d'autre part, dans la richesse du lexique et son degré de structuration (regroupés en anglais sous le terme de *vocabulary depth*). Nous examinerons ces deux points tour à tour après une brève mise au point terminologique, et une présentation des liens démontrés entre connaissances lexicales et compréhension de l'oral.

2.3.1. Terminologie : lexique, vocabulaire, lemme et famille de mots

Le terme « vocabulaire » se place au niveau du mot : un vocabulaire est simplement un ensemble de mots qui peuvent être liés du point de vue thématique, ou faire l'objet d'un enseignement (Picoche, 2011). C'est le terme que nous utiliserons de préférence dans un contexte pédagogique. Un lexique correspond à une acception plus technique, et suppose un degré de structuration de l'ensemble des mots qui le composent (Szudarski, 2018). C'est pourquoi l'on parle de « lexique mental », puisque l'on suppose que les mots connus sont liés par des relations de synonymie, d'antonymie, de similarité de forme, de simultanéité d'acquisition, etc. (Aitchison, 1987). Cependant, la distinction n'étant pas suivie systématiquement par les auteurs (essentiellement anglophones) que nous citerons, qui utilisent le terme anglais *vocabulary size* pour parler de l'étendue des connaissances lexicales, nous utiliserons la plupart du temps les deux termes indifféremment.

Un lexique et un vocabulaire sont des ensembles de mots, mais il reste à définir ce qu'est un mot. La tâche n'est pas simple, mais un article récent (Brysbaert et al., 2016) portant sur l'estimation de la taille du lexique connu des natifs anglophones, commence par une clarification de certains termes, annoncée dès son titre : *How Many Words Do We Know? Practical Estimates of Vocabulary Size Dependent on Word Definition, the Degree of Language Input and the Participant's Age*. L'unité la plus utile pour l'étude du vocabulaire est le lemme[14], défini comme « *an uninflected word from which all inflected words are derived* » (ibid., p.2). Ainsi, dans les listes lemmatisées, *be, am, is, are, was, were, being* et

[14] nous ne distinguerons pas ici lemme et lexème, même si cette distinction a de l'importance dans certaines théories de l'accès lexical en production (Roelofs et al., 1998)

been seront tous classés sous le même lemme *be*. L'autre unité largement utilisée est la famille de mots, dont la définition est présentée comme suit : « *a group of lemmas that are morphologically related form a word family. The various members are nearly always derivations of a base lemma or compounds made with base lemmas* » (ibid., p.2). Il ne s'agit plus uniquement de différentes formes fléchies comme pour le lemme mais également des dérivés morphologiques : ainsi la famille *do* inclura non seulement *does, did, done* et *doing*, mais aussi *undo, redo, doable* et *doer*. Nous verrons dans ce qui suit que l'utilisation de l'une ou l'autre de ces unités n'est pas anodine et conduit à des différences importantes dans les estimations de taille de lexique. En effet, d'après les chiffres de Brysbaert et al. (2016), une famille compte en moyenne 3 à 4 lemmes : ils identifient pour l'anglais 61800 lemmes et 18269 familles de mots, soit 3,38 lemmes par famille.

Pour Bauer et Nation (1993), qui se placent explicitement dans une perspective d'acquisition L2, une famille de mots est simplement un groupe de mots de même racine dont l'apprentissage ne requiert pas d'effort supplémentaire une fois qu'on en connaît un des membres. Les familles ne sont donc pas identiques pour tous les apprenants, en fonction des affixes qu'ils connaissent :

> *[A] word family consists of a base word and all its derived and inflected forms that can be understood by a learner without having to learn each form separately. So,* watch*,* watches*,* watched*, and* watching *may all be members of the same word family for a learner with a command of the inflectional suffixes of English. As a learner's knowledge of affixation develops, the size of the word family increases. The important principle behind the idea of a word family is that once the base word or even a derived word is known, the recognition of other members of the family requires little or no extra effort.* (Bauer & Nation 1993, p.254)

Ils élaborent une échelle de niveaux de connaissance morphologique pour l'anglais qui commence au niveau 1, où les apprenants débutants n'ont aucune connaissance flexionnelle ni dérivationnelle, et où chaque mot rencontré est une famille à lui seul. L'échelle va jusqu'au niveau 7, où tous les affixes sont connus y compris les moins transparents (par exemple les préfixes *ab-* ou *ad-*), et où le sens des racines grecques ou latines est utilisé pour comprendre d'autres mots savants de même origine (par exemple *calligraphy* et *calisthenics*). L'apprenant pris en exemple dans la citation a seulement atteint le niveau 2, où seules les flexions sont connues, mais pas les dérivations possibles de la racine. Cette conception intéressante centrée sur l'apprenant rend cependant difficile l'utilisation du concept « famille de mots », dont la signification change d'une personne à l'autre ou pour la même personne au cours du temps.

Dans la pratique, les listes de fréquence créées par Paul Nation (2006) par la suite sont calibrées au niveau 6, c'est-à-dire que les familles de mots utilisent tous les affixes sauf les plus rares, les plus irréguliers et les moins transparents, et que les racines classiques ne sont pas considérées comme connues.

2.3.2. Lien avec les compétences langagières

La corrélation entre les connaissances lexicales et le niveau de langue étrangère est solidement établi par de nombreuses études, même si la taille de cette corrélation varie d'une étude à l'autre, en fonction de la L1 des apprenants, de leur niveau global, des compétences testées, et des outils utilisés pour les mesurer. Le lien est d'ailleurs si important que des tests de vocabulaire sont parfois utilisés comme tests de positionnement (Meara & Jones, 1988). Mecartty (2000), par exemple, montre, avec un public d'étudiants anglophones apprenant l'espagnol, que les connaissances lexicales jouent un rôle non négligeable en compréhension de l'oral (elles expliquent 14% de la variance du score de CO - la corrélation est de 0,38), bien que leur rôle soit moins important qu'en compréhension de l'écrit (où elles en expliquent 25%). Miralpeix et Muñoz (2018) trouvent des chiffres comparables, avec une corrélation de 0,42 entre taille de vocabulaire et compréhension aurale, et 16% de variance expliquée. D'autres études observent des corrélations plus importantes. Staehr a publié deux études sur le lien entre taille du vocabulaire et compréhension de l'oral, une en 2008 avec des apprenants de niveau intermédiaire faible, et une autre en 2009 avec des apprenants avancés (dans les deux cas, il s'agissait d'élèves ou étudiants danois apprenant l'anglais). Dans la première étude (Stæhr, 2008), la taille du vocabulaire explique 39% de la variance en compréhension de l'oral (et 72% en CE). Dans l'étude qui suit (Stæhr, 2009), avec des apprenants avancés ayant fait des séjours à l'étranger, la taille du vocabulaire explique 50% de la variance de CO (avec une corrélation de 0,70). La profondeur, très corrélée à la taille, ajoute 2% d'explication. Dans une réplication récente (Noreillie et al., 2018) avec des lycéens flamands de niveau B1 en anglais, le pourcentage de la variance en CO expliqué par le test de taille de vocabulaire est de 40%. L'étude de Hilton (2006) dans un contexte universitaire français confirme aussi le rôle essentiel du vocabulaire, non seulement pour le niveau global en anglais, mais plus précisément pour la compréhension aurale. Dans son étude, les connaissances lexicales (mesurées en modalité écrite) expliquent 33% de la variance en CO. Enfin, Milton et al. (2010) sont les seuls à avoir utilisé un test de vocabulaire aural pour

évaluer l'importance du rôle des connaissances lexicales en compréhension de l'oral, et ont trouvé qu'elles expliquaient 40% de la variance de CO.

Dans toutes ces études, le pourcentage de variance expliqué par les connaissances lexicales varie de 14 à 50%, avec une majorité d'études se situant entre 30 et 40% (la corrélation, elle, varie entre 0,38 et 0,70). Nous verrons dans la partie expérimentale si notre étude réplique ces résultats. Cependant, nous exposerons auparavant les difficultés de l'acquisition lexicale pour des apprenants L2, en nous focalisant sur l'étendue et la profondeur du lexique à acquérir.

2.3.3. Caractéristiques du lexique à acquérir

2.3.3.1. étendue des connaissances lexicales (vocabulary size)

Les premières estimations de l'étendue du lexique anglais datent du 19ème siècle, et font état de 10 000 mots pour une personne « ordinaire ». Kirkpatrick (1891), par exemple, obtient ce chiffre à partir d'un comptage des mots du roman de Daniel Defoe, *Robinson Crusoe*, qu'il estime accessible à tout anglophone ordinaire. En le multipliant par deux, il obtient 20 000 mots pour une personne ayant fait des études supérieures. Enfin, à partir d'un échantillonnage de mots d'un dictionnaire visant l'exhaustivité contemporaine mais non historique, *Webster's Unabridged Dictionary*, il estime son propre vocabulaire (celui d'un professeur d'université) à plus de 35 000 mots.

Après être passés par des ordres de grandeur beaucoup plus importants (150 000 chez Seashore et Eckerson en 1940), les chiffres des études contemporaines arrivent à des valeurs tout à fait similaires aux premières estimations. Goulden et al. (1990) proposent ainsi une estimation de 17 000 bases lexicales (ce qu'on appellerait maintenant des familles de mots) pour un natif anglophone. Une base lexicale inclut un lexème et ses formes fléchies (*govern*, *governs*, *governing*, *governed*), c'est-à-dire un lemme dans la définition que nous avons donnée plus haut, ainsi que ses dérivés (*government*), mais pas les mots composés auxquels elle peut contribuer, ni les dérivés non transparents (les mots composés ne sont pas pris en compte dans l'estimation, sauf s'ils sont opaques et orthographiés sans espace, comme *cupboard*, auquel cas ils comptent comme des bases). Plus récemment, Brysbaert et ses collègues, en utilisant des familles de mots élargies, incluant plus de dérivés, trouvent une estimation d'un peu plus de 11 000 familles connues pour un anglophone (américain) moyen à 20 ans, et 13 000 à 60 ans (Brysbaert et al., 2016). Il est intéressant d'inclure ci-dessous (Tableau 2.4) une version adaptée et mise à jour du tableau présenté dans cet article sur les

variations des estimations du nombre de mots connus par des anglophones natifs depuis le siècle dernier. En plus de la définition opérationnelle du mot, la tâche utilisée est également incluse parce qu'elle peut influer sur l'estimation finale. On peut remarquer par exemple que Milton et Treffers-Daller (2013), en utilisant la liste de mots de Goulden et al. (1990), mais avec une tâche de production plutôt que de reconnaissance, arrivent à un total beaucoup moins important (moins de 10 000 familles connues contre plus de 17 000 dans l'étude originale).

référence	estimation (mots connus)	définition du « mot »	tâche
Seashore et Eckerson (1940)	150 000	formes fléchies et dérivées + noms composés+ noms propres + affixes	QCM
Nusbaum et al. (1984)	14 400	lemmes	auto-évaluation (degré de familiarité)
Goulden et al. (1990)	17 200	familles de mots	autoévaluation (connu/ inconnu)
D'Anna et al. (1991)	17 000	lemmes	estimations subjectives
Anderson et Nagy (1993)	40 000	lemmes	tests variés
Zechmeister et al. (1995)	12 000	lemmes (cf. D'Anna)	QCM
Milton et Treffers-Daller (2013)	9 800	familles de mots (cf. Goulden)	production de synonymes
Brysbaert et al. (2016)	11 000 (20 ans) 13 000 (60 ans)	familles de mots	test oui/non (avec non-mots)

Tableau 2.4 - variation des estimations du nombre de mots connus par les anglophones natifs (typiquement étudiants en début d'études universitaires), adapté de Brysbaert et al. (2016)

On constate que les estimations varient de moins de 10 000 à plus de 17 000 quand l'unité est la famille de mots, de 12 000 à 40 000 quand c'est le lemme, et que ce n'est que quand la définition du mot inclut séparément les formes fléchies et dérivées, mais aussi les noms composés, les noms propres et les affixes, que l'on arrive à de très gros chiffres (150 000 mots). Nous retiendrons ici le chiffre de 17 000 familles de mots proposé par Goulden, Nation et Read (1990), obtenu à partir d'un test de reconnaissance et non de production, et donc plus compatible avec notre contexte d'habileté réceptive.

Faut-il que les apprenants L2 maîtrisent ces 17 000 familles ? La tâche paraît insurmontable si l'on considère la taille du lexique acquis pendant la scolarité primaire et secondaire des élèves français. Nous avons deux sources pour connaître cette taille. D'une part, nous pouvons estimer un ordre de grandeur en multipliant le nombre d'heures de cours d'anglais suivis en moyenne avant l'entrée à l'université par l'estimation du nombre de mots appris par heure de cours. D'autre part, nous pouvons chercher des études empiriques sur le sujet. Pour la

première méthode, un élève français qui s'inscrit en première année d'université a suivi en moyenne 50 heures de cours d'anglais par an en école primaire (Ministère de l'Education Nationale, 2015), et environ 100 par an dans le secondaire (Ministère de l'Education Nationale, 2010, 2017b, respectivement pour le collège et le lycée), ce qui donne un total de 900 heures environ. Les études sur l'acquisition du vocabulaire en contexte scolaire estiment que le rythme est de 2 à 3 mots par heure en moyenne (Cobb & Horst, 2011, p. 658), ce qui donnerait une fourchette de 1 800 à 2 700 mots à l'entrée à l'université. Milton et Meara (1995), quant à eux, calculent que pendant leur carrière scolaire/ universitaire (en dehors des périodes à l'étranger, et sans compter le primaire), les apprenants acquièrent entre 500 et 600 mots par an (le nombre d'heures d'enseignement auxquelles ce chiffre correspondrait n'est pas précisé). Si l'on ne compte que les années dans le système secondaire (7), cela donnerait une fourchette de 3 500 à 4 200 mots. Ce chiffre correspond mieux à la seule étude que nous ayons trouvée estimant la taille du lexique des étudiants entrant dans le système supérieur français : Hilton (2006) trouve que les étudiants en première année de licence LEA (Langues Etrangères Appliquées) connaissent 3 900 mots en moyenne, mesurés avec le test de Hever[15], avec des résultats variant de 1 000 à plus de 7 000 mots. Les chiffres de référence dans cet article étant ceux d'Anderson et Nagy (1993), l'unité est donc le lemme et non la famille (cf. Tableau 2.4), et le nombre de familles de mots connues est certainement inférieur. En tout état de cause, nous pouvons garder une estimation intermédiaire de moins de 4 000 mots, auquel cas il resterait à l'étudiant de licence à acquérir plus de 13 000 (familles de) mots en 3 ans.

Devant ces chiffres décourageants (déjà constatés par Arnaud et al., 1985 pour le secondaire), les chercheurs en acquisition et en didactique se sont demandé quelle était la taille de vocabulaire nécessaire pour fonctionner de façon satisfaisante dans la langue étrangère : non plus le vocabulaire total à acquérir, mais le vocabulaire minimum.

2.3.3.2. le minimum lexical

Pour essayer de répondre à la question du vocabulaire minimum à acquérir pour des apprenants L2, il faut commencer par décider pour quel usage. Les recherches se sont d'abord concentrées sur la compréhension de l'écrit. En 1989, Batia Laufer, dans son article, *What percentage of Text-Lexis is essential for comprehension ?* (Laufer, 1989), concluait qu'il fallait, en L2 comme en L1, une couverture lexicale de 95% pour assurer une compréhension « raisonnable » (*reasonable*) des textes écrits, définie par un score de compréhension de 55%

[15] http://test-your-english-now.net/

minimum à une suite de questions ouvertes ou à choix multiple. Cette couverture correspond à 5 000 familles de mots, d'après les études antérieures d'Ostyn et Godin (1985). Quelques années plus tard, Hu et Nation (2000) révisèrent les chiffres de Batia Laufer, en montrant que ce n'est qu'avec une couverture de 98% qu'une majorité d'apprenants atteint une bonne compréhension d'un passage de fiction. Laufer et Ravenhorst-Kalovski (2010) ont également révisé la première estimation de Batia Laufer et montré que si la compréhension visée est « optimale » et non plus seulement minimale (acceptable), une couverture de 98% est nécessaire. Plusieurs études montrent ensuite que cette couverture lexicale n'est accessible qu'avec un vocabulaire de 8 000 familles de mots environ (Laufer & Ravenhorst-Kalovski, 2010; Nation, 2006).

Toutes les études précédentes se fondent sur la compréhension des textes écrits. Les premières études sur la couverture lexicale nécessaire en compréhension de l'oral ont repris les chiffres de couverture lexicale identifiés pour l'écrit (98%), et ont cherché à identifier le nombre de familles de mots correspondant à une telle couverture lexicale pour l'oral : étant donné que la densité lexicale du discours oral est moindre, le chiffre ne peut pas être le même que pour l'écrit. Nation (2006), par exemple, reprend le chiffre de 98% et trouve qu'il correspond à 6 000 ou 7 000 familles de mots à l'oral. De même, Webb et Rogers (2009a, 2009b), à partir de l'analyse de programmes télévisés et de scripts de films, concluent que 3 000 familles de mots suffisent pour une couverture de 95% à l'oral, mais qu'il en faut 6 à 7 000 pour une couverture de 98% (6 000 familles de mots en moyenne pour les films, et 7 000 pour les émissions télévisées).

En 2013, Van Zeeland et Schmitt (2013), soulignant que toutes les études sur la couverture lexicale à l'oral se basent sur des chiffres (95% ou 98%) obtenus lors d'expérimentations sur l'écrit, reprennent le protocole de Hu et Nation (2001) et l'adaptent pour l'oral. Dans quatre narrations informelles (textes authentiques trouvés sur la Toile), ils remplacent 0%, 2%, 5% ou 10% des mots les plus rares par des non-mots, et s'assurent que tous les autres mots appartiennent aux 2 000 mots les plus fréquents. Après avoir testé la compréhension de natifs et non-natifs sur ces quatre versions grâce à des QCM de compréhension, ils montrent qu'avec une couverture lexicale de 90%, le score moyen de compréhension est de 73,5% pour les L2, et de 85% pour les L1. La majorité des apprenants sont donc capables de comprendre une narration informelle dont 1 mot sur 10 leur est inconnu. Cependant, la variabilité dans ce cas est très importante : certains apprenants n'ont que 30% de compréhension, ce qui est au niveau du hasard pour un QCM avec 3 ou 4 propositions de réponse par question. Les auteurs

concluent donc qu'une couverture de 95% est beaucoup plus sûre. En effet, même si la compréhension moyenne dans ce cas n'est que légèrement supérieure au cas de figure précédent (76,5% de compréhension avec 95% de couverture, et 73,5% avec 90%), aucun apprenant n'obtient alors moins de 50% en compréhension. Pour assurer une compréhension adéquate d'une majorité d'apprenants, il est préférable (logiquement) que seul un mot sur 20 soit inconnu, sans que cela exige pour autant des connaissances lexicales extrêmement étendues. En effet, comme mentionné plus haut, cette couverture de 95% correspond à une maîtrise de 3 000 familles de mots environ à l'oral. Ce minimum lexical pour la compréhension de l'oral est beaucoup moins important que pour la compréhension de l'écrit, où, comme nous l'avons vu, le nombre de familles de mots nécessaire à une bonne compréhension est estimé à plus de 8 000 (Nation 2006).

On pourrait penser que nos étudiants, qui connaissent (probablement) 4 000 mots en moyenne à l'entrée à l'université, ont donc déjà un bagage suffisant pour une compréhension adéquate. Cependant, il faut nuancer cette conclusion pour quatre raisons. D'une part, le chiffre de 4 000 mots est une moyenne qui cache beaucoup de variation : dans l'étude de Hilton (2006), certains étudiants ne maîtrisaient que 1 000 mots. Or ce sont précisément ces étudiants qui ont besoin de remédiation, et que nous espérons identifier avec les tests diagnostiques que nous allons développer. Une autre raison de penser que nos étudiants n'ont pas forcément atteint le seuil nécessaire est que ce dernier est exprimé en familles de mots, alors que le test utilisé dans l'étude de Hilton (2006) utilise le lemme et non la famille. Or, l'une des caractéristiques des niveaux faibles est que la connaissance d'un mot n'implique pas celle de tous les membres de sa famille (McLean, 2018; Ward & Chuenjundaeng, 2009). Un étudiant peut connaître l'adjectif *wild*, par exemple, sans forcément connaître le nom *wilderness*, ou le nom *coward*, sans reconnaître le nom dérivé *cowardice*. Troisièmement, toutes les études mentionnées utilisent la forme écrite (et non orale) des mots pour tester le vocabulaire (problème sur lequel nous reviendrons dans la section suivante). Enfin, l'étude de Van Zeeland et Schmitt (2013) qui a abouti au chiffre de 3 000 familles nécessaires pour la compréhension de l'oral se basait sur la compréhension de narrations informelles. Il est probable que la compréhension de cours universitaires ou d'émissions politiques (et de beaucoup d'autres discours oraux), qui ont moins recours à la situation de communication immédiate ou aux indices visuels, nécessitent une couverture lexicale plus importante (par ailleurs, ces genres textuels font appel à des connaissances encyclopédiques que nos étudiants ne possèdent pas nécessairement et sur lesquels ils ne peuvent donc pas toujours s'appuyer

pour pallier des déficiences lexicales éventuelles). D'autres chiffres existent d'ailleurs : l'étude de Staehr (2009), avec un test de compréhension de l'oral utilisant des textes formels, semi-formels et formels, arrive à la conclusion qu'une couverture de 98% est nécessaire, ce qui correspond pour les textes utilisés à 5 000 familles de mots (ou, dans les études de Webb et Rogers (2009a, 2009b) citées plus haut, à 6 000 ou 7 000 familles). Pour toutes ces raisons, il nous semble probable qu'un grand nombre de nos étudiants n'aient pas atteint le minimum lexical requis, ce que nous pourrons vérifier lors des études expérimentales décrites dans la deuxième partie.

2.3.3.3. le rôle de la fréquence

Jusqu'à présent, nous avons examiné l'étendue du lexique minimal à acquérir (nombre de familles de mots nécessaires à la compréhension), sans nous poser la question du choix de ces mots. Si l'on considère que la langue anglaise possède environ 20 000 familles de mots (Brysbaert et al., 2016; Nusbaum et al., 1984), et que nos étudiants ont besoin de connaître entre 3 000 et 6 000 de ces familles (selon qu'on désire une couverture de 95 ou de 98% du discours courant), il reste à décider comment les choisir parmi les 20 000 familles existantes. La réponse qui s'impose, du fait de la structuration du lexique, est celle de la fréquence : les mots ne sont pas tous utilisés avec la même fréquence, et il paraît logique de penser qu'il est plus « rentable » pour les apprenants de connaître les mots plus fréquents, qu'ils ont par définition plus de probabilité de rencontrer que les mots moins fréquents. Cette caractéristique du lexique peut être saisie graphiquement par la courbe représentant le lien entre étendue du vocabulaire et couverture textuelle, comme on peut le constater sur la Figure 2.3 (tirée de Chujo & Utiyama, 2005) : plus la taille du vocabulaire augmente (et moins les mots deviennent fréquents), plus l'augmentation de la couverture textuelle ralentit. En passant de 0 à 2 000 lemmes (l'unité utilisée dans cette étude), on passe de 0 à près de 85% de couverture textuelle, mais avec les 2 000 lemmes suivants, on ne rajoute « que » un peu plus de 5% de couverture, pour passer à 90% (les 2 000 lemmes suivants en rajoutent moins de 3%). Il apparaît clairement que la fréquence lexicale joue un rôle essentiel dans le « retour sur investissement » (pour filer la métaphore économique) de l'apprentissage de nouveaux mots.

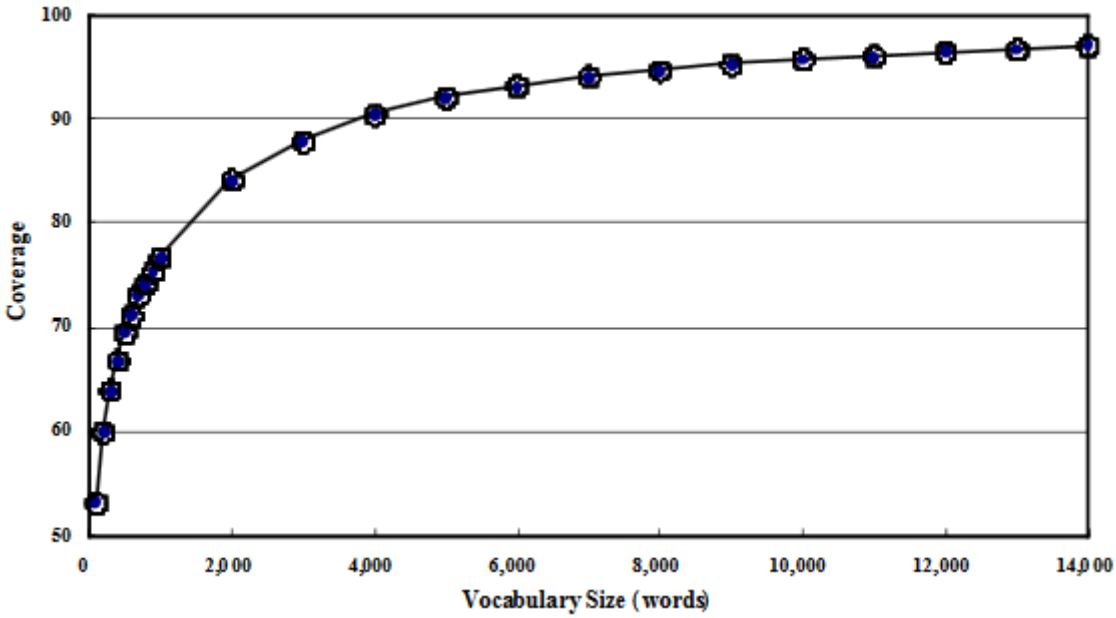


Figure 2.3: pourcentage de couverture textuelle en fonction de la taille du vocabulaire (Chujo et Utiyama 2005)

Dès le début du 20ème siècle, Edward Thorndike (1921, 1931) a produit les premières listes de fréquence de vocabulaire destinées aux enseignants de lecture en anglais L1 (*The Teacher's Word Book* en 1921, et *A Teacher's Word Book of the Twenty Thousand Words Found Most Frequently and Widely in General Reading for Children and Young People* en 1931). Pour la L2, Brezina et Gablasova (2015, p. 2) mentionnent l'existence d'une liste de mots de base dès 1936[16], à laquelle a également collaboré Thorndike : *Interim Report on Vocabulary Selection for the Teaching of English as a Foreign Language* (Faucett et al., 1936). Michael West l'a ensuite révisée pour produire l'une des listes de vocabulaire L2 les plus influentes : la *General Service List* ou *GSL* (West, 1953), qui contient environ 2 000 familles de mots (3 600 lemmes d'après Browne, 2013), complétée ensuite par l'*Academic Word List*, une liste de 560 familles qui rassemble des mots qui ne sont pas dans la *GSL* mais qui ajoutent une couverture importante aux textes universitaires (Coxhead, 2000). Cependant, la liste *GSL*, créée à partir d'un corpus du début du 20ème siècle, contenait des mots obsolètes (*shilling*) mais pas d'autres mots à présent courants tels que *television* ou *computer*. Plusieurs révisions en ont donc été proposées, dont celle de Brezina et Gablasova (2015). Leur *New-GSL* contient 2 500 lemmes (et non familles), qui couvrent plus de 80% de leurs corpus, composé de textes essentiellement écrits et de variété britannique, à savoir le *LOB* (*London-Oslo-Bergen Corpus*), le *BNC* (*British National Corpus*), et deux corpus plus récents, le *BE06 Corpus of British English*, et le *EnTenTen* (un très gros corpus de textes de la Toile). Cependant, quelles

[16] En 1891, W. R. Baird estimait à 300 le nombre de mots nécessaires pour se débrouiller dans un pays étranger : « *Total words needed in a foreign country to get along* » (cité par Seashore & Eckerson, 1940, p. 26)

que soient les qualités de cette nouvelle liste, elle n'atteint pas le chiffre de 3 000, voire 6 000 familles identifié plus haut comme le minimum lexical en compréhension de l'oral. Il est donc nécessaire d'avoir recours à d'autres listes, associées à d'autres corpus, pour atteindre ces chiffres. Ces listes devront compter au moins 6 000 familles de mots, ou, si elles sont lemmatisées, probablement plus de 10 000 lemmes ou 20 000 mots. Il est difficile de donner une estimation plus précise dans la mesure où, même si l'on sait qu'une famille compte en moyenne 3 lemmes et qu'un lemme recouvre en moyenne un peu moins de 2 formes différentes, certains des lemmes appartenant à une famille fréquente peuvent être très rares et peu intéressants à connaître. Par exemple, dans la famille de la base *worth*, le nom pluriel *worthies* est rare et n'a probablement pas vocation à être inclus dans une liste de mots représentant le minimum lexical.

Quel corpus choisir ? Il existe actuellement beaucoup de corpus pour l'anglais consultables en ligne ou dont les informations sur la fréquence lexicale sont téléchargeables librement sous forme de feuille de calcul. Nous présentons dans le tableau ci-dessous (Tableau 2.5) une liste non exhaustive de ces corpus, en précisant lesquels sont accompagnés d'une liste de mots classables par fréquence descendante, ce qui correspond à notre besoin d'identifier les « x » mots les plus fréquents.

corpus	variété	oral (pourcentage)	date	liste ordonnée	taille en mots
BNC	anglais britannique	oui (10%)	années 1990	oui (6 200 lemmes)	100 millions
COBUILD/ Celex	plusieurs variétés	oui (25%)	1991	non	18 millions
MICASE	anglais américain	oui (100%)	1997-2002	non	1,8 millions
COCA	anglais américain	oui (20%)	1990-2017	oui (5 000 lemmes)	560 millions
COCA-Academic	anglais américain	non (0%)	1990-2015	oui (20 000 lemmes)	120 millions
$SUBTLEX_{US}$	anglais américain	oui (100%)	1900-2007	oui (60 000 mots non lemmatisés)	51 millions
Academic spoken corpus	plusieurs variétés	oui (100%)	1997-2015	oui (1 740 familles de mots)	13 millions
BNC + COCA (P. Nation)	plusieurs variétés	oui (non précisé)	1990-2017	oui (10 000 familles)	non précisée

Tableau 2.5 – liste partielle de corpus de l'anglais disponibles en ligne avec listes de mots associées

Etant donné le contexte dans lequel doivent fonctionner nos étudiants (anglais académique), il serait intéressant d'utiliser le corpus MICASE (*Michigan Corpus of Academic Spoken*

English), qui utilise des transcriptions d'anglais académique américain. Cependant, ce corpus consultable en ligne s'utilise uniquement comme un concordancier qui, pour un mot (ou expression) donné, permet de voir et d'entendre des exemples de contexte dans lequel celui-ci est utilisé, accompagné d'informations précises sur la nature de ce contexte. Il ne fournit pas d'informations générales sur la fréquence de ce mot dans le corpus tout entier. Le corpus d'anglais oral académique (*Academic spoken corpus*) développé par Dang, Coxhead et Webb (2017) pour créer une nouvelle liste d'anglais académique (*Academic Spoken Word List*) est également intéressant mais accompagné d'une liste très courte (1 700 familles seulement). Il nous faut donc nous rabattre sur les corpus plus classiques de l'anglais britannique (*British National Corpus* ou BNC) ou de l'anglais américain (*Corpus of Contemporary American English* ou COCA, M. Davies, 2009). Cependant, ces deux corpus sont eux aussi accompagnés de listes de 5 ou 6 000 lemmes seulement, et ne suffisent donc pas à nos besoins. La partie académique du corpus COCA est accompagnée d'une liste lemmatisée de 20 000 mots, mais qui est basée sur un corpus exclusivement écrit (Gardner & Davies, 2014). C'est pourquoi nous nous sommes tournée vers la liste de familles de mots de Paul Nation (2017), compilée à partir d'un corpus hybride mêlant des textes du BNC et de COCA. Cette liste composée de 10 000 familles peut être consultée en ligne via l'outil VocabProfiler (Cobb, s. d.), qui permet d'entrer une liste de mots et de recevoir en sortie la bande de fréquence de chacun de ces mots. Cependant, comme la famille de mots n'est pas toujours l'unité la plus appropriée pour des apprenants qui n'ont pas nécessairement une connaissance étendue de la morphologie dérivationnelle de la L2 (McLean, 2018), nous ferons également usage de la liste accompagnant le corpus $SUBTLEX_{US}$, crée par Marc Brysbaert et Boris New (2009). Ce corpus, constitué à partir de sous-titres de films et de séries télévisées américaines, ne contient pas d'anglais académique, mais correspond probablement mieux à l'input authentique qu'ont pu recevoir nos apprenants avant l'entrée à l'université. Il est accompagné d'une liste de 60 000 mots non lemmatisés, et classés non pas par leur nombre d'apparitions dans le corpus (fréquence brute), mais par le nombre de textes dans lesquels ils sont présents (mesure de diversité contextuelle). Son principe d'organisation est donc totalement différent de celui des listes de Paul Nation, et nous permettra de faire des comparaisons intéressantes. Enfin, nous utiliserons aussi la base de données multilingue Celex (Baayen et al., 1995), qui utilise pour l'anglais le corpus COBUILD (J. M. Sinclair, 1987), et qui permet comme VocabProfiler une consultation facile en ligne, mais avec des résultats basés sur le lemme et non la famille de mots.

Notons pour terminer que la fréquence ne devrait pas être un critère exclusif du choix des mots à apprendre (Gougenheim et al., 1964). Ward et Chuenjundaeng (2009) remarquent par exemple que les listes de fréquence sont établies à partir de corpus généralistes qui correspondent peut-être à ce à quoi les natifs sont en général exposés au cours de leur vie, mais reflètent probablement assez peu l'input reçu par des apprenants L2 (ce qui peut conduire à des comportements assez différents en termes de sensibilité à la fréquence lexicale, cf. Diependaele et al., 2013). Cependant, on pourrait objecter que l'important n'est pas ce que nos apprenants ont entendu jusque-là, mais plutôt ce qu'ils sont censés comprendre à partir de maintenant. Dans cette perspective, un corpus généraliste peut tout à fait convenir. Une autre objection plus gênante est que, comme le rappelle Dee Gardner (2007), les mots les plus fréquents sont en général polysémiques, et il n'est pas clair que connaître un des sens du mot permette de le comprendre dans tous les contextes où il est utilisé. Par ailleurs, beaucoup de ces mots sont utilisés dans des collocations dont le sens n'est pas toujours transparent ni connu des apprenants (comme nous le verrons un peu plus loin). Tous ces facteurs font qu'un test de connaissance lexicale basé uniquement sur des mots isolés comme celui que nous allons construire surestime probablement les connaissances des apprenants. Afin de remédier à cette faiblesse, nous tenterons de le conjuguer à une évaluation des connaissances phraséologiques (collocationnelles) des étudiants. Cependant, la fréquence reste un critère important à prendre en compte, dans la mesure où les mots très fréquents se retrouvent dans tous les genres et à travers les époques (Brezina & Gablasova, 2015).

2.3.4. Difficultés : « profondeur » du lexique

Dans les études précédemment citées visant à estimer l'étendue du lexique des anglophones natifs, les tests s'effectuent soit en demandant aux sujets s'ils connaissent les mots présentés, soit en leur demandant de produire un synonyme, une définition, ou de trouver la bonne réponse parmi plusieurs propositions (QCM). Dans tous les cas, les mots sont proposés à l'écrit et uniquement en reconnaissance. Ces études correspondent ainsi à des connaissances assez superficielles.

2.3.4.1. aspects de la connaissance d'un mot chez Nation (1990, 2001)

Nation (1990), conscient de ce problème, a proposé une des définitions les plus utilisées de ce que « connaître un mot » veut dire. Il distingue plusieurs aspects : la forme (orale et écrite), le sens (le concept lui-même et ses associations), l'usage (qu'il nomme *function*, et qui correspond à la fréquence et au registre/genre), et la syntaxe (appelée *position*, qui comprend

le patron grammatical et les collocations du mot). Nation a ensuite légèrement remanié cette classification dans son livre *Learning vocabulary in another language* (2001, p. 49), et y a notamment ajouté la connaissance de la morphologie du mot. Il a également rassemblé les caractéristiques sous trois aspects seulement : forme (orale, écrite, morphologie), sens (lien forme-sens, concept et associations) et usage (fonction grammaticale, collocations et fréquence/registre). Une autre dimension traverse ces caractéristiques, celle qui distingue la compétence réceptive de la compétence productive. Nous adaptons ci-dessous (Tableau 2.6) le tableau que Nation propose, y soulignant les questions correspondant à la réception (R) pour plus de lisibilité.

Forme		
Forme orale	R	A quoi ressemble ce mot à l'oral ?
	P	Comme ce mot est-il prononcé ?
Forme écrite	R	A quoi ressemble ce mot à l'écrit ?
	P	Comment ce mot s'écrit-il ?
Morphologie	R	Quelles parties peut-on reconnaître dans ce mot ?
	P	De quelles parties ai-je besoin pour exprimer ce sens ?
Sens		
Forme et sens	R	Que veut-dire ce mot ?
	P	Quel mot utiliser pour exprimer ce sens ?
Concept et référents	R	Qu'est-ce qui est inclus dans ce concept ?
	P	A quoi peut se référer ce concept ?
Associations	R	A quels autres mots nous fait penser ce mot ?
	P	Quels autres mots utiliser à la place de celui-ci ?
Usage		
Structure grammaticale	R	Dans quelle structure ce mot est-il utilisé ?
	P	Quelle structure faut-il utiliser avec ce mot ?
Collocations	R	A quels mots s'attendre avant et après celui-ci ?
	P	Quels mots faut-il utiliser avec celui-ci ?
Fréquence, registre, …	R	Où, quand et à quelle fréquence s'attendre à rencontrer ce mot ?
	P	Où, quand et à quelle fréquence faut-il utiliser ce mot ?

Tableau 2.6 - Aspects de la compétence lexicale selon Nation (2001, p.49), en réception (R) et en production (P)

Même si ce n'est pas précisé dans le texte, il nous semble que Paul Nation a organisé ces aspects par ordre d'importance décroissante, ou de difficulté croissante : la forme et le sens sont les caractéristiques les plus importantes (le « signifiant » et le « signifié » de Saussure), et correspondent aux connaissances en général testées pour estimer la taille ou étendue du vocabulaire (*vocabulary size*, ou *vocabulary breadth*), par opposition à la richesse du vocabulaire (*vocabulary depth*), qui suppose pour chaque mot une connaissance plus approfondie, celle au moins de la structure grammaticale associée et des collocations dans

lequel il est utilisé, voire du registre et de sa fréquence (il est possible également que l'organisation du tableau découle des modèles de traitement du langage, allant des sons au discours, en passant par le mot, unité de sens).

2.3.4.2. reconnaissance de la forme orale

Pour la question qui nous préoccupe, celle de la compréhension de l'oral, les aspects pertinents sont ceux qui sont accompagnés d'un « R » et soulignés dans le Tableau 2.6. Si nous nous intéressons d'abord à la forme, c'est la question « A quoi ressemble ce mot à l'oral ? » qui importe. Pourtant, l'étendue du lexique est calculée à partir de l'écrit dans presque toutes les études décrites précédemment. Ceci peut poser un problème, comme le remarque Christine Goh (2000, p. 61) : « *It is likely that for some [students], sound-to-script relationships have not been fully automatised. Therefore, although they knew certain words by sight, they could not recognise them by sound. Put another way, their listening vocabulary was underdeveloped.* » Nous savons également que de nombreux mots anglais reconnus à l'écrit par les francophones ne le sont pas forcément à l'oral. Hilton (2003), par exemple, à partir d'une tâche de décision lexicale, identifie un certain nombre de mots transparents (également appelés mots congénères, Bogaards, 1994) que les francophones reconnaissent presque aussi vite que les anglophones à l'écrit, mais qu'ils rejettent comme des non-mots à l'oral. Nous reproduisons ici le tableau des 16 mots pour lesquels la différence est la plus grande entre écrit et oral (Tableau 2.7).

mot	**%age de réponses correctes, ECOUTE**	**%age de réponses correctes, LECTURE**
creature	**06**	**100**
rebel	**18**	**94**
muscle	**19**	**65**
fraction	**24**	**88**
theory	**31**	**100**
freedom	**35**	**100**
jury	**35**	**94**
signal	**35**	**100**
rival	**38**	**76**
agent	**41**	**82**
angle	**44**	**82**
issue	**44**	**82**
client	**44**	**65**
finance	**44**	**65**
volume	**44**	**65**
fabric	**47**	**81**

Tableau 2.7 - reconnaissance des mots transparents (cognates) à l'écrit et à l'oral, d'après Hilton (2003)

Tous ces mots, à part *freedom*, sont transparents à l'écrit, même si deux sont des faux-amis (*issue* et *fabric*, qui n'ont pas le même sens en anglais et en français). Ils ont tous (à part *freedom*) une étymologie latine et ont généralement été adoptés en anglais par l'entremise du français, excepté *rival* et *theory* qui ont été empruntés directement au latin, d'après l'*Oxford English Dictionary* (Simpson, 1989). Ils sont faciles à reconnaître à l'écrit car l'orthographe est la même en anglais et en français, hormis *rebel* (« rebelle » en français), *theory* (« théorie ») et *fabric* (« fabrique »), dont l'orthographe reste cependant très proche (« créature » en français se distingue également par son accent aigu). Par contre, ils sont difficiles à reconnaître à l'oral du fait des changements de prononciation qui sont intervenus en anglais depuis leur emprunt au français (ou au latin, avec un emprunt parallèle en français).

Nous pouvons prendre comme illustration le premier mot du tableau, *creature*, prononcé /ˈkri:tʃ ə(r)/, qui cumule deux difficultés. D'une part, la prononciation de la voyelle initiale correspondant au digraphe <ea>, qui est prononcé comme une seule voyelle longue /i:/[17]. D'autre part, la consonne /t/ a subi un phénomène de palatalisation du fait du /j/ qui suivait, ce qui fait que le /tj/ s'est transformé en /tʃ/ (*yod-coalescence* Dauer, 1993). Ces deux phénomènes (sans compter la réduction de la voyelle finale, et en anglais britannique la disparition du /r/ final) font qu'il est très difficile pour un francophone non averti de reconnaître dans /ˈkri:tʃ ə(r)/ le /kʁe a tyʁ/ français.

La difficulté en reconnaissance aurale du lexique peut donc venir de la distance entre la prononciation attendue au vu de la forme écrite et la prononciation effective. Contrairement à ce qu'on pourrait penser, la connaissance de la forme écrite du mot joue ainsi un rôle en compréhension de l'oral pour des apprenants lettrés. Alors que ce rôle est clairement positif en compréhension de l'écrit, puisqu'elle permet à peu de frais aux apprenants francophones d'acquérir rapidement un vocabulaire conséquent en reconnaissance à l'écrit en anglais, son rôle est plus ambigu en compréhension aurale. La connaissance de l'orthographe peut être utile si elle aide à la reconnaissance aurale du mot, mais l'effet contraire peut également se produire. Paola Escudero et ses collaborateurs montrent par exemple que l'effet de la connaissance orthographique est positif si les correspondances graphème-phonème sont régulières dans la L2 et si elles correspondant à celles de la L1, mais peuvent être néfastes dans le cas contraire (Escudero et al., 2014). Etant donné que l'anglais est irrégulier dans ses

[17] A l'époque du Grand changement vocalique (*Great Vowel Shift*), qui a eu lieu entre le 14ème et le 16ème siècles en Grande Bretagne, le /e:/ correspondant au digraphes <ea> s'est transformé en /i:/ - la suite orthographique <éa> n'était d'ailleurs pas un digraphe au départ, puisqu'en français elle correspond à 2 voyelles différentes. (Chevillet, 1994)

correspondances graphèmes-phonèmes, et que ces correspondances, au moins pour les voyelles, sont différentes de celles du français, on peut s'attendre à ce que cela pose des difficultés pour nos apprenants.

2.3.4.3. corrélation entre étendue et richesse (profondeur) du lexique

Nous avons vu qu'il était essentiel qu'en plus du sens des mots (le signifié), nos apprenants connaissent leur forme orale, et pas seulement leur forme écrite. Cependant, ces trois caractéristiques ne sont que les premières des neuf identifiées par Nation (2001), que nous avons présentées dans le Tableau 2.6.

Dans son article *Dimensions of lexical competence*, Meara (1996) reconnaît que toutes les caractéristiques énumérées par Nation sont importantes pour caractériser la connaissance des mots individuels, mais souligne qu'il est difficile en pratique de tester ces connaissances : avec un échantillon restreint à 50 mots, par exemple, cela reviendrait à administrer aux apprenants 50 mots x 8 caractéristiques, soit 400 questions. D'autre part, il fait l'hypothèse que l'étendue et la profondeur du lexique sont fortement corrélées chez la plupart des individus : « *It would be unusual, for example, to find someone with a vocabulary of 10,000 words who did NOT know that 'child' is a common word, used in slightly formal situations, that it is a noun, makes its plural with 'ren', and is associated with 'boy', 'girl', 'parent' and so on.* » (ibid., p.44). Autrement dit, l'étendue et la profondeur du vocabulaire évoluent de concert, du fait des conditions d'acquisition du lexique: « *Most people acquire L2 words from exposure to the language, not from learning lists of words in the abstract, and it is inevitable that while they are doing this, they also acquire a broader knowledge about the words they already know* » (ibid., p.44). Il considère qu'en deçà de 5 000 mots, une mesure de la taille du vocabulaire suffit à caractériser le lexique d'un apprenant. Ce n'est qu'au-delà qu'il peut être important de caractériser la « profondeur » du vocabulaire par l'intensité des liens que chaque mot entretient avec les autres éléments du lexique. Il conclut en faisant l'hypothèse que cela peut être fait au niveau du lexique entier et non au niveau de chaque mot : au lieu de tester pour chaque mot les liens qu'il entretient avec d'autres mots, sa fréquence, ses collocations, Meara propose de caractériser l'organisation du lexique lui-même en montrant que tous les mots sont reliés entre eux dans le lexique mental des locuteurs (par exemple avec une tâche d'association libre).

Dans l'article de Meara (1996), la corrélation supposée entre étendue et profondeur du lexique reste spéculative. Cependant, d'autres linguistes ont cherché à aborder le problème de façon

empirique. Schmitt fait une revue de la littérature qui a étudié la question entre-temps et parvient à la même conclusion que Meara : « *For higher frequency words, and for learners with smaller vocabulary sizes, there is often little difference between size and a variety of depth measures* » (Schmitt, 2014, p. 913)[18]. Il remarque par ailleurs qu'en réception, la profondeur a moins d'importance qu'en production. En effet, en réception, la construction grammaticale, les collocations, le genre de texte sont donnés par le contexte et ne sont pas à construire comme en production. D'autres chercheurs sont encore plus réservés sur le concept de « profondeur » des connaissances lexicales, du moins pour une utilisation dans un contexte d'évaluation : « *vocabulary depth has been valuable in furthering the thinking in the field, but its ill-defined, cover-all nature makes it inappropriate as a construct to be used in assessment procedures* » (Gyllstad, 2013).

Au vu des études décrites ci-dessus, de la compétence langagière que nous étudions (réception et non production), et du public auquel nous nous intéressons (étudiants ayant besoin de remédiation, donc probablement avec un lexique peu étendu), il nous semble possible de conclure que dans le contexte de cette étude, une mesure de l'étendue du vocabulaire (nombre de mots reconnus) sera pertinente et suffisante. Dans la deuxième partie de cette étude sur l'opérationnalisation de nos construits, nous étudierons plus en détail les différentes formes de tests qui ont été proposés pour estimer cette taille du lexique. Nous suivrons aussi les conseils de Staehr en construisant un test de reconnaissance aurale : « *a study of the relationship between vocabulary size and listening should ideally be based on a vocabulary test that involves hearing the target words rather than reading them*» (Stæhr, 2009, p. 597).

2.4. Intégration : connaissances phraséologiques et morphosyntaxiques

Le rôle des connaissances lexicales est essentiel en compréhension, mais un texte n'est pas une simple collection d'items lexicaux qu'il suffirait de décoder un à un pour en comprendre le sens. Le sens de tous ces mots doit être intégré au fur et à mesure de leur reconnaissance dans une structure syntaxique. Le tout n'est pas la simple somme des parties, et la structure syntaxique et les combinaisons de mots apportent elles-mêmes du sens (D. Lee, 2001). Marslen-Wilson et Tyler (1980) montrent par exemple à quel point il est plus facile de traiter des phrases syntaxiquement bien formées que des listes de mots (phrases dont les mots sont dans le désordre). C'est une illustration du processus de *chunking*, processus fondamental de

[18] Vermeer (2001) arrive à la même conclusion avec des enfants de maternelle en L1.

la cognition humaine dont nous avons décrit l'importance, et qui consiste à regrouper des éléments au niveau supérieur.

Nous analyserons le processus de regroupement des mots dans des ensembles plus grands en L2 d'abord sous l'angle phraséologique (expressions à plusieurs mots telles que les collocations, idiomes, etc…), puis sous l'angle morphosyntaxique. D'après Panisse (2009, p. 7), la morphosyntaxe « porte aussi bien sur les formes des mots, flexions régulières et irrégulières, variantes irrégulières de certains noms et verbes, l'agencement des marques syntaxiques autour du nom (déterminants, etc.), du verbe (pronoms, etc.), de l'adjectif, de l'adverbe, et enfin de l'organisation des mots et groupes de mots dans un énoncé ou une phrase. ». Nous parlerons d'abord du rôle des mots fonctionnels (les « marques syntaxiques » de Panisse), puis de celui des structures syntaxiques proprement dites.

2.4.1. Connaissances phraséologiques et non compositionnalité du sens

Comme nous l'avons suggéré, il est possible de comprendre tous les mots d'une phrase sans comprendre le sens de la phrase. Cela peut être dû à un manque de connaissances, non pas lexicales proprement dites mais plutôt collocationnelles, qui entrave l'activation du sens d'un ensemble plurilexical. De fait, le sens d'une combinaison de mots n'est pas toujours déductible de la combinaison du sens de chaque mot pris isolément : en d'autres termes, le sens n'est pas toujours compositionnel. Martinez et Murphy (2011) créent deux textes écrits à partir des mêmes mots (dont plus de 98% appartiennent aux 2 000 mots les plus fréquents en anglais) pour illustrer cette difficulté. L'un des deux textes contient beaucoup d'expressions idiomatiques (*living large, that's neither here nor there, to come across as*), tandis que l'autre utilise les mêmes mots de façon plus compositionnelle, c'est-à-dire que le sens du tout est déductible de celui des parties (*I don't live by any large cities, I like it there, …*). Le deuxième texte est beaucoup mieux compris que le premier par des apprenants brésiliens alors qu'ils contiennent tous les deux les mêmes mots. De plus, si les apprenants sont capables d'estimer assez précisément leur compréhension du deuxième texte, ils ont tendance à surestimer leur compréhension du premier (dont ils se rendent tout de même compte qu'il est plus difficile). C'est une démonstration expérimentale de la difficulté liée à l'utilisation d'expressions figées et autres expressions idiomatiques.

Cette difficulté est d'autant plus grande que ces expressions sont difficiles à repérer quand elles sont formées à partir de mots courants dont le sens est connu. Bishop (2004) montre que

les apprenants L2 (anglais) repèrent plus facilement les mots inconnus que les expressions inconnues (comme *come in for* dans le sens de « être confronté à », formée uniquement de mots connus car très fréquents). Dans son expérience, les sujets lisaient un texte contenant des mots et des expressions inconnus d'eux, avant de répondre à des questions de compréhension. Ils avaient la possibilité pendant la lecture de cliquer sur les éléments dont ils ignoraient le sens afin d'avoir accès à un glossaire. Le nombre de clics sur les mots inconnus est significativement supérieur à celui sur les expressions inconnues, sauf quand ces dernières sont mises en valeur typographiquement. Les apprenants ont donc effectivement du mal à détecter la présence d'expressions inconnues formées de mots eux-mêmes connus.

D'autres linguistes ont montré que ces expressions ne sont pas rares : Erman et Warren (2000) appellent *prefabs*, ou combinaisons de mots préfabriquées, toute séquence de deux mots au moins dont l'un détermine (ou du moins restreint) le choix de l'autre, même si le sens reste compositionnel : « *a combination of at least two words favored by native speakers in preference to an alternative combination which could have been equivalent had there been no conventionalization* ». Travaillant à partir d'un corpus de textes écrits et oraux du *London Lund Corpus of Spoken English* et du corpus *Lancaster-Oslo-Bergen*, elles calculent (avec la définition certes très large des *prefabs* citée plus haut) que plus de la moitié des textes (et même près de 60% à l'oral) est formée de combinaisons préfabriquées telles que *good friends, not bad, I guess,* ou *go to seminars*.

Martinez et Schmitt (2012, p. 304) utilisent une définition plus restrictive de ce qu'ils appellent *phrasal expressions*[19] , puisqu'elle suppose une certaine non compositionnalité: « *a fixed or semi-fixed sequence of two or more co-occurring but not necessarily contiguous words with a cohesive meaning or function that is not easily discernible by decoding the individual words alone* ». Ils trouvent également que ces expressions sont très courantes. A partir du *British National Corpus*, ils en créent une liste (la liste PHRASE, pour *PHRASal Expressions*) sur le modèle des listes de vocabulaire fréquent, et identifient 505 expressions qui sont aussi courantes que les 5 000 mots les plus fréquents en anglais (avec une définition un peu moins restrictive des collocations, Shin et Nation (2008) en trouvaient 567 parmi les 3 000 mots les plus fréquents). Ces expressions sont formées à 95% des 1 000 mots les plus courants en anglais (d'où, comme nous l'avons mentionné, un problème pour les repérer), et

[19] Une des constantes de la recherche sur ces expressions est la variété des termes utilisés pour les désigner, comme souligné par Wray (2000), qui identifie une cinquantaine d'étiquettes différentes en anglais, dont *chunks, formulas, multiword units,* ou *ready-made expressions.*

on trouve parmi les premières *have to*, *there is/are*, *of course*, *I mean*, ou *go on*. Il nous semble que cette liste est particulièrement intéressante pour notre contexte, dans la mesure où le critère de non compositionnalité est particulièrement pertinent pour l'activité de réception. En compréhension, de fait, seules les expressions à sens non compositionnel seront difficiles : une expression qui n'est pas repérée comme telle mais qui a un sens qui se déduit de celui de ses parties (*make a decision*) ne devrait pas poser de problèmes. En production, par contre, l'apprenant devra choisir entre plusieurs possibilités (*make/ take /do/ form/ … a decision*), et devra donc connaître la collocation la plus courante, alors que cette même collocation est donnée à l'auditeur en compréhension aurale. La liste PHRASE de Martinez et Schmitt nous paraît donc intéressante, à la fois parce que son contenu est adapté à la compréhension, et parce que les expressions qui y figurent sont classées par ordre de fréquence, ce qui est utile pour fabriquer un test. Martinez (2011) a d'ailleurs développé un test associé, que nous pourrons utiliser pour évaluer les connaissances phraséologiques (terme que nous utiliserons pour désigner ces expressions) de nos étudiants.

2.4.2. Reconnaissance des mots fonctionnels

En parlant des connaissances lexicales, nous avons traité tous les mots de la même façon, mais il est possible d'isoler une classe de mots qui présentent une utilisation grammaticale spécifique : il s'agit des mots grammaticaux ou « fonctionnels » (*function words*, Fries, 1952). Ce sont typiquement des déterminants (*the, a*, …), des auxiliaires (*is, will,…*), des pronoms (*you, them, …*) ou des conjonctions (*and*), ainsi que des prépositions (*of*). Les auteurs de la grammaire *Longman Grammar of Spoken and Written English* (Biber et al., 1999) présentent les spécificités de ces mots dans un tableau que nous reproduisons ici (Tableau 2.8). Par rapport aux autres mots du lexique (qu'ils appellent « mots lexicaux », mais qui sont parfois désignés sous le terme de « mots à contenu » (*content words*, Field 2008) ou, chez Quirk et al. (1989), « mots à classe ouverte », *open-class items*, ou encore « mots référentiels » en psycholinguistique), les mots fonctionnels sont plus fréquents, plus courts, beaucoup moins nombreux (D. Brown, 2017, en compte 446 en anglais)[20] et moins accentués, ils sont souvent invariables (même si ce n'est pas le cas des auxiliaires anglais, qui ont parfois, dans le cas de *be,* plus de formes que les verbes lexicaux) et ont un sens moins facilement définissable. De plus (comme l'indique leur étiquette dans la grammaire de Quirk et al. (1989), *closed-class items*), ils ne sont pas productifs, c'est-à-dire qu'il est rare que de nouveaux éléments soient

[20] Dang et Webb (2016) en comptent 176 dans leur *Essential Word List* destinée à des débutants.

intégrés à leur catégorie, qui est assez « fermée » (parce que l'apparition de nouveaux éléments fait suite à un processus de grammaticalisation qui a lieu sur le long terme). Il est ainsi beaucoup plus facile d'imaginer la création d'un nouveau nom (catégorie ouverte) que d'un nouveau pronom (catégorie fermée).

caractéristiques	**mots lexicaux**	**mots fonctionnels**
fréquence	basse	haute
longueur	longs	courts
sens lexical	oui	non
morphologie	variables	invariables
productivité	oui (classe ouverte)	non (classe fermée)
nombre	nombreux	peu nombreux
accentuation	accentués	désaccentués

Tableau 2.8 - différences entre mots lexicaux et mots fonctionnels en anglais, d'après Biber et al. (1999, p.55)

Deux caractéristiques qui nous intéressent particulièrement pour la compréhension de l'oral sont la brièveté des mots fonctionnels et leur prononciation généralement désaccentuée (et très fortement neutralisée en anglais oral). Cela peut les rendre particulièrement difficiles à reconnaître dans le flux de la parole malgré leur haute fréquence (qui devrait normalement aider à leur reconnaissance) :

> *As closed-class words are often pronounced as weak syllables, they are likely to have short, centralized vowels or no vowel at all as well as reduced, imperfectly articulated consonants. The set of lexical competitors activated by such poor input could be quite large and yet, because the acoustic evidence is so poor, have no clear front runners*. (Shillcock & Bard, 1993, p. 182, cités par Field 2008, p. 416)

Le *Manuel de phonologie anglaise* de Michel Viel (2003, p. 86-88), par exemple, présente sous forme de tableau une liste des mots à forme réduite à l'oral, qui sont tous des mots fonctionnels : 20 auxiliaires (dont 3 formes de *have*, et 8 de *be*), 6 prépositions, 16 pronoms, 6 déterminants et 7 conjonctions. Nous reproduisons ci-dessous (Tableau 2.9) la première ligne de chaque sous-catégorie proposée dans le tableau de ce manuel (nous avons conservé la colonne « ne réduisent pas » parce que paradoxalement, elle contient la forme réduite de l'adverbe négatif *not*, non mentionnée par ailleurs dans le tableau). Nous constatons que certaines de ces formes réduites perdent leur voyelle (/v/ pour *have*, /n/ pour *and*) et ne forment plus alors une syllabe indépendante. Ils deviennent des clitiques qui s'accrochent au mot qui les suit ou qui les précède et sont donc d'autant plus difficiles à détecter.

		Formes pleines	Formes réduites	Ne réduisent pas	
Auxiliaires	have	**hæv**	**həv > əv, v**	haven't	ˈ**hævnt**
Prépositions	at	**æt**	**ət**		
Pronoms	me	**mi:**	**mi, mə**		
Déterminants	a	**eɪ**	**ə**		
Conjonctions	and	**ænd**	**ənd, ən > nd, n**		

Tableau 2.9 - exemples de formes réduites des mots fonctionnels anglais par catégorie (d'après Viel 2003)

Trois autres caractéristiques jouent cependant en faveur des mots fonctionnels en reconnaissance : le fait qu'ils soient fréquents (et donc rencontrés souvent par les apprenants), que la plupart d'entre eux aient une morphologie invariable (il n'y a donc pas besoin d'apprendre à identifier les différents membres de leur famille puisqu'ils n'ont pas de formes dérivées, sauf les auxiliaires qui ont des formes fléchies), et qu'ils soient peu nombreux à l'intérieur de leur catégorie (ils ont donc moins de concurrents potentiels). Pour savoir si les caractéristiques favorables ou défavorables priment lors de l'écoute, Field (2008b) compare la reconnaissance des mots lexicaux et celle des mots fonctionnels en anglais chez des natifs et non natifs de différents niveaux. Il utilise une tâche de transcription où les sujets écoutent un texte et où, à chaque fois que l'enregistrement s'arrête, ils doivent noter les 4 ou 5 derniers mots entendus. Les natifs sont au plafond (même s'ils reconnaissent légèrement mieux les mots à contenu), mais les apprenants L2 ont un écart de 20% de réussite entre la transcription des mots fonctionnels et celle des mots à contenu, à l'avantage de ces derniers. Les mots fonctionnels sont donc effectivement problématiques en contexte.

Sosa et MacFarlane (2002) tentent d'expliquer pourquoi la fréquence n'aide pas forcément à la reconnaissance de certains éléments. Ils étudient la reconnaissance de la préposition *of* à partir de phrases tirées d'un corpus de conversations authentiques. Ils constatent tout d'abord (avec des auditeurs natifs) que moins de la moitié (45%) des occurrences de *of* dans leur corpus sont repérées. D'autre part, les temps de réaction sont plus courts et le nombre d'erreurs moins élevé quand *of* fait partie d'expressions moins courantes (*care of*, *much of*), alors que les collocations très courantes (*sort of*, *lot of*) entraînent plus d'erreurs et des temps de réaction plus longs (ils remarquent d'ailleurs que les temps de réaction toutes fréquences confondues sont deux fois plus longs que dans les expériences classiques avec des stimuli artificiels). L'explication avancée est que les collocations très fréquentes sont « groupées » (*chunked*) et que le mot fonctionnel *of* perd alors son individualité. Sa reconnaissance devient plus difficile et plus longue parce qu'il faut d'abord le « désemballer » (*unpack*) de la collocation dans lequel il est inséré et qui est reconnue comme un tout (Erman & Warren, 2000).

Tout ceci explique en partie pourquoi ces mots sont peu utilisés dans les tests de vocabulaire (Kremmel, 2016, p. 981), bien qu'ils soient très présents au tout début des listes de fréquence (par exemple, les 20 premiers mots de la liste non lemmatisée de Brysbaert et New (2009) sont *the, to, a, you, and, it, 's, of, for, I, in, on, is, that, what, be, have, are* et *this*, et les 20 premiers lemmes du *British National Corpus* sont *the, be, of, and, a, in, to, have, it, to, for, I, that, you, he, on, with, do, at, by*). En effet, les tests de vocabulaire présentent souvent les mots à tester de façon isolée. Or, la prononciation isolée des mots fonctionnels (leur « forme du dictionnaire » ou forme de citation) est très différente de leur prononciation en contexte, et ne nous informe pas sur la capacité des auditeurs à les reconnaître en conditions réelles.

L'autre difficulté qui peut expliquer l'absence de mots fonctionnels dans les tests de vocabulaire est qu'ils ont souvent un sens abstrait et difficile à définir (absence de « sens lexical » selon le bilan présenté dans le Tableau 2.8). Pour *the*, par exemple, on trouve dans le dictionnaire (Longman 1991) non une définition, mais une explicitation de l'usage : « *used before nouns when the referent has been previously specified by context or by circumstance* ». Pour ce même mot, les grammaires de l'énonciation francophones, qui cherchent à dégager une valeur centrale invariante qui couvre tous les usages, proposent par exemple « l'indication d'un travail perceptif et interprétatif antérieur » (Lapaire & Rotgé, 1991, p. 110), ou « une sorte de décrochage par rapport au réel » (Adamczewski & Delmas, 1982, p. 208). Ces propositions sont très abstraites, et il n'est pas possible de les inclure dans un test de vocabulaire au milieu de définitions de mots à sens plein. Il est ainsi difficile d'intégrer ces mots dans un test de reconnaissance (QCM) ou de production du sens. C'est pourquoi nous avons choisi d'inclure la reconnaissance des mots fonctionnels dans un test grammatical où ils seront contextualisés à l'intérieur de phrases complètes.

2.4.3. Le rôle des connaissances syntaxiques

Les structures syntaxiques, comme les collocations et les mots lexicaux ou fonctionnels que nous avons étudiés jusqu'à présent, sont l'appariement d'une forme et d'un sens (ou d'une fonction). La forme des structures syntaxiques, comme celle des mots fonctionnels, a tendance à être courte, voire à ne pas être réalisée phonétiquement : la structure transitive SVO, par exemple, correspond à un ordre des mots et ne rajoute pas de segments sonores aux mots dont elle est composée (mais la prosodie, et en particulier l'intonation, apporte des informations syntaxiques essentielles : Cutler et al., 1997). Par ailleurs, la forme des structures syntaxiques peut être discontinue, c'est-à-dire qu'elle intervient à plusieurs endroits de la

phrase (par exemple, pour une structure pseudo-clivée, *what S V is O*). Ces deux propriétés font qu'elles peuvent être difficiles à repérer, et ce d'autant plus que les mots fonctionnels dont elles sont en partie formées ont souvent une forme réduite. Elles diffèrent ainsi des collocations, qui sont également des combinaisons de mots, mais qui sont beaucoup plus locales, puisque les mots qu'elles combinent sont adjacents. Du point de vue sémantique, le sens des structures syntaxiques (comme celui des mots fonctionnels) est assez abstrait. Pour la structure ditransitive en anglais (*Mary gave him the book*, *Sue taught me all I know*), par exemple, le sens exprimé peut être vu comme le transfert d'une entité (objet ou information) d'un agent à un récipiendaire, avec une focalisation sur le résultat du transfert (D. Lee, 2001, p. 75). Encore une fois, cette définition ne pourra pas être utilisée telle quelle dans un test de connaissances grammaticales.

Les structures syntaxiques sont donc, comme les mots fonctionnels, difficiles à repérer et leur sens est complexe à définir. Est-ce à dire qu'elles posent des problèmes importants aux apprenants ? Van Patten (2002) montre que cela peut être le cas en liant leurs difficultés éventuelles au phénomène de redondance grammaticale. De fait, le sens exprimé par les structures morphosyntaxiques est souvent présent à d'autres endroits de la phrase à travers des éléments lexicaux. Par exemple, les formes grammaticales passées sont souvent accompagnées d'autres repères temporels passés (*last week, in 2001*, etc.), et le *s* de la troisième personne du singulier du présent en anglais est toujours accompagné d'un sujet lui-même singulier et troisième personne. L'apprenant n'a pas vraiment besoin de faire attention à ces formes grammaticales peu saillantes puisque leur sens est présent ailleurs de façon plus saillante, ce qui explique la difficulté d'acquisition de certaines formes grammaticales pourtant extrêmement fréquentes : « *A language learner might never get round to noticing low salience cues, particularly when the interpretation accuracy afforded by the other more obvious cues does well enough for everyday communicative survival.* » (N. C. Ellis, 2008, p. 379). Ce phénomène d'apprentissage associatif s'appelle le blocage (*blocking*, N. C. Ellis, 2006), et fait qu'une fois qu'un événement – pour nous, un sens– est associé de façon fiable à un premier indice – par exemple un adverbe temporel– il est difficile de rajouter ensuite à l'association un autre indice – par exemple un suffixe verbal.

D'autres problèmes de compréhension peuvent découler de l'influence de la langue maternelle ou de stratégies de compréhension inadaptées à la langue étudiée. Par exemple, Van Patten et Cadierno (1993) montrent que la tendance des anglophones apprenant l'espagnol à interpréter des structures avec objet préverbal (de type OV ou OVS) comme des

phrases SV ou SVO peut avoir son origine soit dans l'influence de l'anglais L1, où le sujet est toujours préverbal, soit dans l'utilisation d'une stratégie universelle qui pousse les apprenants à interpréter le premier groupe nominal de la phrase comme un agent.

Van Patten et ses collaborateurs ont essentiellement travaillé sur la compréhension de l'espagnol par des anglophones. On peut se demander si des études comparables existent sur les difficultés spécifiques aux francophones pour l'acquisition de l'anglais. C'est le cas de la thèse de Maud Pélissier (2018), qui utilise le *Competition Model* de MacWhinney (2005) pour étudier l'acquisition de la syntaxe de l'anglais en réception par un public similaire au nôtre (étudiants de licence qui se spécialisent en anglais). Le modèle de compétition suppose que l'apprentissage des correspondances entre forme et fonction fasse appel à la compétition entre plusieurs indices (*cues*) dont l'apprenant se sert pour interpréter l'input. Au début de l'apprentissage, les indices qui priment sont ceux venus de la L1 (par exemple, l'ordre des mots) ou ceux qui sont particulièrement saillants (par exemple, des adverbes temporels plutôt que des suffixes grammaticaux). Maud Pélissier montre, en utilisant les potentiels évoqués, que les francophones ont effectivement plus de mal à rejeter des questions grammaticalement incorrectes si elles sont plus proches de la structure française (**Did Mary finished our dinner ?*, dont le prétérit/ participe passé *finished* ne les choque pas dans la mesure où le passé en français utilise aussi auxilaire + participe passé : « a fini ») que si elles en sont éloignées. **Had Mary finish our dinner ?* est ainsi plus facile à rejeter parce que cette structure (incorrecte) ne correspond pas au plus-que-parfait français, composé d'un auxiliaire au passé suivi d'un participe passé.

Une autre linguiste qui s'est intéressée à l'interaction entre les connaissances morphosyntaxiques et la compréhension chez les francophones est Ruth Huart. Sa *Grammaire orale de l'anglais* (Huart, 2002) mentionne par exemple que la confusion observée chez des étudiants français qui croient entendre *her lady's ship* au lieu de *her ladyship* peut venir d'une méconnaissance de la règle d'accentuation de la tête du groupe nominal (*ladyship*, en un mot, est accentué sur la première syllabe, tandis que dans *her lady's ship*, la tête, qui reçoit l'accent principal, est *ship*). Il nous semble cependant que la confusion peut dans ce cas tout autant provenir d'une ignorance lexicale, *ladyship* (rang 22 231 dans Brysbaert et al., 2016) étant beaucoup plus rare que *lady* (498) ou *ship* (1602). Un autre exemple proposé par Huart (ibid., p.42) est la confusion entre *only essential personnel will be kept* et **only a central personnel will be kept* qui aurait pu être évitée grâce à une meilleure connaissance du fonctionnement grammatical du nom *personnel* (indénombrable donc incompatible avec le déterminant *a*). Il

s'agit là d'exemples intéressants mais anecdotiques, et nous n'avons pas trouvé d'autres travaux sur le même sujet, malgré la pléthore d'études sur les problèmes grammaticaux en production (par exemple en anglais L2, Payre-Ficout, 2007; Sournin-Dufossé, 2007; Vraciu, 2012).

Nous terminerons avec des expériences qui montrent que les natifs eux-mêmes font parfois des erreurs d'interprétation liées à un défaut d'analyse syntaxique. Certains natifs interprètent par exemple *The dog was bitten by the man* comme si le chien était l'agent, et non la victime de la morsure (Ferreira & Patson, 2007). Ferreira et ses collaborateurs appellent cela *Good enough comprehension* : quand il y a un conflit entre les informations syntaxiques (ici la structure passive) et les informations sémantiques (en général, ce sont les chiens qui mordent, et non les humains), ces dernières gagnent parfois parce que les auditeurs se contentent d'une analyse syntaxique superficielle et s'appuient sur leurs connaissances du monde pour comprendre les phrases proposées[21]. Lim et Christianson (2013) mettent en évidence un effet similaire (toujours avec le passif) avec des apprenants L1 coréen L2 anglais.

Nous nous tournerons à présent vers les études générales qui étudient la corrélation entre connaissances morphosyntaxiques et compréhension aurale chez les apprenants L2.

2.4.4. Corrélation avec la compréhension de l'oral

Bien que le rôle des connaissances syntaxiques ne soit probablement pas aussi important en compréhension qu'en production (parce qu'en compréhension les structures morphosyntaxiques sont fournies et il « suffit » de les reconnaître, tandis qu'en production il faut les recréer ex nihilo, et que d'autre part les locuteurs se contentent parfois d'une compréhension superficielle en s'appuyant sur les indices sémantiques), on peut tout de même s'attendre au vu des paragraphes précédents à ce qu'il y ait une corrélation entre les deux. Par exemple, les mots fonctionnels étant souvent réduits, ce sont les connaissances grammaticales qui aident à leur reconnaissance en contexte : le /z/ de /ʃɪz ˈkʌmɪŋ/ sera interprété comme une variante de *is* et le /z/ de /ʃɪz bɪn hɪə/ comme un *has* grâce à la connaissance des formes auxiliées des aspects respectivement progressif (*She's coming*) et parfait (*She's been here*). Brunfaut et Révész (2015), qui s'intéressent à un grand nombre de variables qui peuvent

[21] Un effet similaire avait déjà été montré pour l'écrit par Wason et Reich (1979) avec une structure complexe contenant plusieurs négations : la phrase *No head injury is too trivial to be ignored* était généralement interprétée comme voulant dire « aucune blessure à la tête ne doit être négligée (ignorée) » alors qu'elle veut dire le contraire.

rendre l'écoute difficile (caractéristiques du texte, de la tâche ou de l'auditeur), trouvent pourtant que la complexité syntaxique (présence de subordination, de négation, etc…) ne semble pas corrélée à la difficulté de la tâche.

Les études sur la question montrent qu'il existe tout de même un lien entre connaissances grammaticales et compréhension de l'oral, même si ce lien n'est pas forcément très étroit. Mecartty (2000), que nous avons déjà citée à propos des connaissances lexicales, trouve une corrélation de 0,26 entre ces deux variables. Cependant, le modèle statistique de régression hiérarchique montre que la variable « connaissances grammaticales » n'apporte rien une fois que le lexique est pris en compte. Zoghlami (2015) trouve des résultats tout à fait similaires : les connaissances grammaticales (mesurées avec des phrases à trous du *OPT Grammar*) présentent une corrélation de 0,30 avec la compréhension de l'oral, et n'expliquent pas de variance supplémentaire dans une analyse de régression, une fois la reconnaissance lexicale et l'étendue du vocabulaire prises en compte. La seule étude que nous ayons trouvée où les connaissances grammaticales étaient plus fortement corrélées à la compréhension de l'oral que les connaissances lexicales est celle d'Andringa et al. (2012), où la corrélation est de 0,68 avec le vocabulaire, mais de 0,77 avec la grammaire. La compétence grammaticale était testée avec un test aural de grammaticalité alors que les instruments des autres études utilisaient l'écrit ; ceci explique peut-être en partie les résultats différents, et nous encourage à utiliser également un test aural visant à mesurer les connaissances grammaticales en L2. Nous pouvons citer une étude comparable en compréhension de l'écrit : Shiotsu et Weir (2007) trouvent également que les connaissances syntaxiques jouent un plus grand rôle que le vocabulaire (alors que la plupart des études sur la CE montrent le contraire), mais la corrélation entre connaissances syntaxiques et lexicales est telle (85%) que les auteurs se demandent s'il ne s'agit pas en fait du même construit.

Pour ce qui est du rôle des connaissances phraséologiques en compréhension, nous n'avons trouvé que deux études qui jettent une lumière indirecte sur la question. Nous avons déjà mentionné la première, qui est celle de Brunfaut et Révész (2015) sur (entre autres) les caractéristiques du texte oral et la difficulté de la tâche de compréhension aurale. La présence d'expressions phraséologiques très courantes (*as well as, deal with*) rend la tâche plus facile, tandis que les expressions rares la rendent plus difficile. La présence de ces expressions semble donc influencer les réponses des candidats, mais leurs connaissances phraséologiques n'ont pas été évaluées directement. C'est ce que fait la deuxième étude, qui porte cependant sur la compréhension de l'écrit et non de l'oral. Kremmel et ses collègues étudient les liens

entre connaissances lexicales, grammaticales, phraséologiques d'une part et la compréhension de l'écrit de l'autre dans une étude intéressante à plusieurs titres (Kremmel et al., 2017). Tout d'abord, la corrélation constatée entre syntaxe et compréhension de l'écrit est de 0,39, similaire aux études décrites ci-dessus, et cette corrélation est inférieure à celle observée avec le vocabulaire, qui est de 0,83. De plus, la variable la plus prédictive dans leur modèle statistique est celle qui mesure la compétence phraséologique (les connaissances lexicales expliquent une partie supplémentaire de la variance, mais ici non plus les connaissances grammaticales n'expliquent pas de variance supplémentaire une fois les deux autres variables prises en compte). Nous inclurons également un test de connaissances phraséologiques dans notre étude ; il sera intéressant de comparer nos résultats avec les leurs. En conclusion, ils soulignent à quel point les connaissances lexicales, phraséologiques et syntaxiques sont liées, et qu'il serait souhaitable, soit de faire des tests intégrés, soit au minimum de ne négliger aucune de ces trois sous-composantes :

> *[W]e may wish to consider developing tests of lexicogrammar rather than 'pure' syntax or vocabulary tests, or integrating aspects of syntactic or phraseological properties of vocabulary into vocabulary tests. Even if opting for a 'distinct-components' approach, it is important to cast our nets wider and incorporate measurements of phraseological knowledge alongside traditional measures of vocabulary and syntax in the development of diagnostic test batteries*. (Kremmel et al., 2017, p.19)

2.5. Stratégies et automatisation

2.5.1. Manque d'automatisation

Nous avons à présent fait le tour des connaissances linguistiques qui peuvent faire défaut à des apprenants francophones en anglais L2. Nous avons également souligné en fin de premier chapitre que l'automatisation des opérations de bas niveau était un facteur déterminant en compréhension de l'oral, et que ces automatismes pouvaient manquer en L2, compliquant le décodage, la reconnaissance lexicale ou l'intégration syntaxique. Ce manque d'automatisation peut saturer les ressources attentionnelles, entravant la construction du sens au niveau discursif. Ce problème est reconnu par les chercheurs en didactique qui travaillent sur la compréhension de l'oral L2 :

> *[L]istening comprehension processes do not occur automatically [in] L2 learners […] If low-level processes take up much attention to treat small units of meaning, then activating high-level processes (i.e., activation of prior knowledge and context stored in long term memory, construction of relationships between different understood elements) is no longer possible, and this may considerably impair comprehension.* (Roussel et al., 2017, p. 41)

Ce manque d'automatisation est manifeste dans toutes les études qui utilisent un groupe de contrôle natif ou qui comparent les résultats des locuteurs L1 et L2. Ces derniers sont en effet toujours plus lents que les L1 (en réception comme en production), même quand leurs performances par ailleurs sont au même niveau de précision (MacWhinney, 2005, p. 54). Comme l'expliquent les théories de l'automatisation d'Anderson (J. R. Anderson & Schunn, 2000) ou de Logan (1988) que nous avons exposées dans le premier chapitre, ce déficit d'automatisation provient essentiellement de l'expérience moindre des apprenants avec la L2. De ce fait, ces derniers manquent de connaissances sur les fréquences lexicales et les fréquences de cooccurrence qui sont nécessaires à la création de *chunks* et à l'utilisation efficace des prédictions dont nous avons montré qu'elles accéléraient les opérations de compréhension à tous les niveaux (N. C. Ellis, 2003). C'est pourquoi il est en pratique difficile de séparer les connaissances de leur utilisation :

> *Although conceptually skill acquisition can be distinguished from knowledge accumulation, in reality, knowledge accumulation forms part of skill acquisition because, in real L2 learning, exposure to new words goes hand in hand with exposure to words encountered previously*. (Hulstijn et al., 2009, p. 555)

Nous ne testerons pas directement la connaissance des fréquences des mots chez nos étudiants, mais les connaissances de cooccurrence leur seront utiles dans le test de connaissances phraséologiques ainsi que dans celui sur les connaissances grammaticales, dont nous décrirons la conception dans la deuxième partie de cette thèse. Le degré d'automaticité des traitements formels ne sera pas non plus évalué : il nous faudrait pour cela mesurer chez nos sujets les temps de réaction dans des tâches de décision lexicale, par exemple, ce qui est exclu par les conditions matérielles de passation des tests dans l'étude actuelle.

2.5.2. Stratégies compensatoires

Un manque d'automatismes peut conduire à l'utilisation de stratégies compensatoires, supposées remédier au décodage et à l'intégration sémantique et syntaxique défaillants : « *much second language listening is dependent upon the learner's ability to compensate for gaps in understanding* » (Ridgway & Field, 2000, p. 190). Comme nous l'avons souligné dans le premier chapitre, les stratégies sont des actions conscientes mises en œuvre pour résoudre un problème. De nombreuses recherches ont porté sur l'utilisation de stratégies en compréhension de l'oral L2 et les chercheurs semblent unanimes sur ces deux caractéristiques :

> *Strategies are considered as a subclass of plans and are defined by means of two criteria: problem orientedness and consciousness* (Færch & Kasper, 1980, p. 57);
>
> *The defining features of learning strategies are that they are conscious and that they are intended to enhance comprehension, learning or retention* (O'Malley et al., 1989, p. 422);
>
> *conscious mental activity, employed to meet a specific learning goal, and […] transferrable to other situations or tasks* (Graham et al., 2010, p. 3);
>
> *L2 listeners commonly require strategies to help them understand an utterance. These strategies are deliberate, conscious procedures that compensate for actual or anticipated breakdowns in comprehension* (Yeldham, 2017, p. 4).

Nous n'essaierons pas dans cette étude de tester l'utilisation des stratégies par nos étudiants, et ce pour trois raisons. Tout d'abord, presque toutes les recherches actuelles sur l'utilisation des stratégies en compréhension de l'oral utilisent des protocoles de réflexion à haute voix (par ex. Vandergrift, 2003; Zoghlami, 2015), ou des carnets de bord écrits (Goh, 2000), qui permettent aux apprenants de verbaliser les opérations cognitives qu'ils utilisent pendant l'écoute d'un document. Ce ne sont pas des techniques qui seraient faciles à utiliser en ligne, du fait de la difficulté à analyser les réponses automatiquement. L'utilisation de l'instrument *Metacognitive Awareness Listening Questionnaire* développé par Vandergrift et al. (2006) serait éventuellement possible, mais il n'a pas donné de résultats très probants dans une étude récente menée avec un public très similaire au nôtre. Zoghlami (2015) en a créé une version révisée et a montré que ce questionnaire ne discriminait pas entre les sujets de niveau de compréhension élevé et faible. En effet, tous les apprenants de cette étude (des étudiants de première année à l'université se destinant à être spécialistes d'anglais) étaient très conscients des stratégies qu'ils utilisaient et avaient une haute conscience métacognitive, quel que soit leur niveau de langue. La différence se faisait uniquement sur les variables personnelles (en particulier l'anxiété, assez naturelle, des étudiants plus faibles devant une tâche d'écoute).

D'autre part, les résultats de la littérature sur les stratégies sont encore assez flous. Il semblerait que les apprenants L2, quel que soit leur niveau, utilisent les mêmes stratégies, et souvent avec une fréquence semblable, mais ne les utilisent pas de la même façon, ni à partir des mêmes données (Macaro, 2006). En compréhension aurale, Graham et al. (2008) décrivent par exemple le cas d'une élève de niveau faible qui utilise beaucoup la stratégie d'anticipation, supposée être caractéristique des bons auditeurs (Vandergrift, 2003). A partir des questions auxquelles elle sait qu'elle va devoir répondre, elle prévoit les mots qu'elle est susceptible d'entendre (elle applique donc cette stratégie uniquement au niveau lexical), mais

ses problèmes de segmentation et la difficulté à reconnaître les mots connus font que cette stratégie reste inopérante. De même, Zoghlami (2016), contrairement à Vandergrift (2003), constate que les auditeurs inexperts utilisent autant que les experts les stratégies de méta-compréhension comme le *monitoring*. Comme les experts, ils font régulièrement le point sur ce qu'ils ont compris ou pas (mais constatent probablement plus souvent qu'ils n'ont pas compris).

Enfin, nous avons l'intention de lier les résultats de nos tests diagnostiques à des activités de remédiation. Or il n'est pas clair que ces stratégies puissent être enseignées en dehors d'un exercice de compréhension où les apprenants font effectivement l'expérience d'un problème concret (J. Field, 2000). L'enseignement de stratégies hors contexte risque de donner des apprenants qui savent les appliquer quand on leur dit de le faire, mais sont démunis en situation réelle : « *even where learners have become better at using the target strategy, it seems that they may not be capable of employing it appropriately in relation to a particular listening text or of combining it successfully with other strategies that they have encountered* » (ibid., p.192). Malgré le scepticisme de John Field, plusieurs chercheurs l'ont tenté depuis qu'il a écrit ces lignes. Le travail de Stéphanie Roussel nous paraît à cet égard très intéressant. Roussel et al. (2017) comparent par exemple deux types d'entraînement à la compréhension de l'oral, l'un basé sur l'amélioration des processus de bas niveau (perception et segmentation), et l'autre sur l'utilisation de stratégies métacognitives (planification, élaboration d'hypothèses, vérification). Les résultats montrent que la focalisation sur les processus de bas niveau aide particulièrement les apprenants de niveau faible, mais que ceux qui ont un niveau avancé ne bénéficient pas plus d'un type d'apprentissage que de l'autre.

Au vu de tous ces résultats, nous ne sommes pas sûre que l'enseignement des stratégies de compréhension soit assez mûr pour être traduit en activités en autonomie en ligne. Au final, il nous semble plus intéressant d'essayer d'entraîner nos apprenants à automatiser les processus de bas niveau que de promouvoir l'utilisation de stratégies essentiellement compensatoires dont le rôle et l'utilité sont indéniables, mais encore mal compris. Comme le souligne Wilson (2003, p. 338), le but de nos apprenants est bien de comprendre ce qu'ils entendent sans effort, et non d'être obligé d'utiliser ces stratégies : « *the learners' ultimate aim is to rely less on contextual guesswork, and more on hearing what was actually said* ». Le médium de l'ordinateur, qui ne se lasse jamais de la répétition, paraît de plus particulièrement adapté à l'objectif d'automatisation des opérations de bas niveau, que nous avons essayé de représenter graphiquement dans la Figure 2.4. On y voit que chez les auditeurs inexperts, les processus

non automatisés (donc coûteux en attention, en temps et en énergie) correspondent à l'essentiel des moyens mis en œuvre lors de la compréhension de l'oral, et que les processus automatisés sont très minoritaires. Le but des activités de remédiation qui seront éventuellement développées serait d'automatiser le plus de processus possibles afin d'arriver à une situation plus conforme à celle des auditeurs experts, chez qui l'attention est essentiellement centrée sur les processus de très haut niveau (construction du modèle de situation).

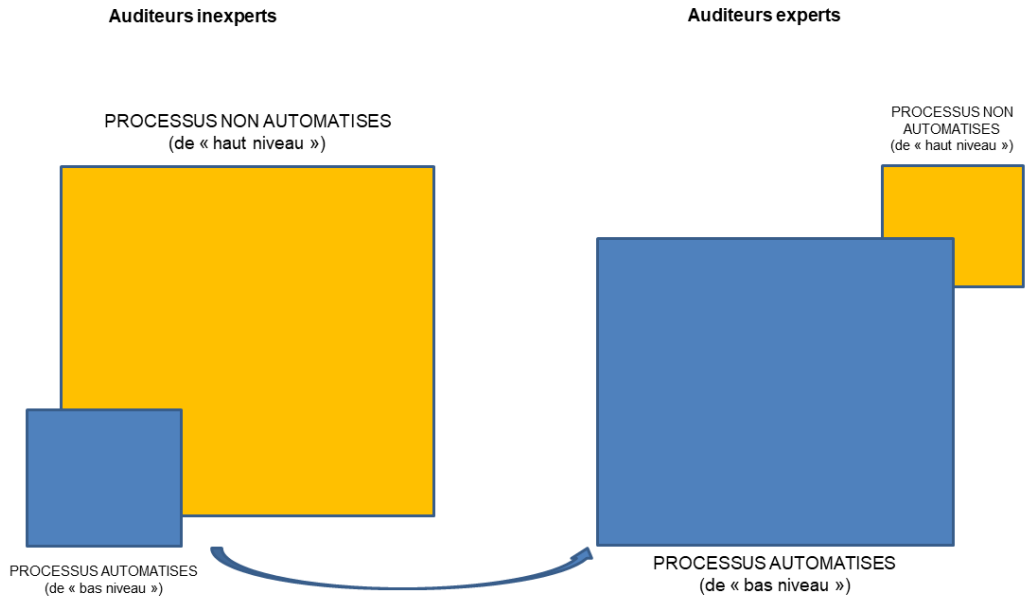


Figure 2.4 - automatisation des processus de bas niveau en compréhension de l'oral

Mentionnons pour terminer que de nombreux auteurs étudiant la compréhension de l'oral L2 intègrent à leurs expériences une mesure de la mémoire de travail de leurs sujets, qui est presque toujours corrélée au résultat en compréhension de l'oral (par ex. Hilton, 2006; Kormos & Sáfár, 2008). Cependant, il s'agit d'une caractéristique individuelle qui est en général vue comme constante chez un individu (Miller, 1956), alors que nous nous intéressons aux caractéristiques pour lesquelles une remédiation est possible[22]. On pourrait penser que le sujet S. F. mentionné dans le premier chapitre, qui était capable à la fin de l'expérience sur la mémorisation de se rappeler de 80 chiffres (Ericsson et al., 1980), avait augmenté sa mémoire de travail. Il n'en était pourtant rien, puisque sa capacité à mémoriser des lettres, par exemple, restait tout à fait ordinaire (6 lettres). Ce qui avait augmenté (comme chez l'apprenti lecteur)

[22] Klingberg (2010) montre cependant qu'un entraînement long, régulier et supervisé peut avoir un effet sur la capacité de mémoire de travail, au moins pour des populations ayant des déficits cognitifs ou attentionnels.

était la taille des éléments qui sont manipulés en mémoire de travail : S. F. avait appris à regrouper les chiffres en ensembles plus grands[23].

Le même genre de processus doit opérer en compréhension de l'oral (où la plupart des membres d'une communauté linguistique atteignent le statut d'expert, contrairement au jeu d'échecs) : grâce à l'exposition à la langue et à l'apprentissage statistique (et implicite) des fréquences de co-occurrence (cf. chapitre 1), les apprenants augmentent la taille des unités de traitement (de plusieurs phonèmes à plusieurs mots). Cette plus grande facilité à regrouper des éléments au niveau supérieur permet d'éviter l'interférence avec les nouveaux éléments qui arrivent, et donc au final de traiter plus d'éléments à la fois (Christiansen & Chater, 2016). Dans cette analyse, les différences individuelles découlent de l'expérience des sujets plus que de la capacité de l'entité séparée que serait la mémoire de travail (Ericsson & Kintsch, 1995).

2.6. Conclusion

En conclusion, nous pouvons confirmer ce que nous avions avancé à la fin de notre premier chapitre. D'après les études empiriques que nous avons résumées dans ce deuxième chapitre sur la compréhension L2, en particulier chez les francophones apprenant l'anglais, la corrélation entre les connaissances linguistiques et la compréhension de l'oral varie en fonction du niveau de traitement considéré, comme on le voit dans le Tableau 2.10.

On peut également remarquer que cette place centrale du lexique va dans le sens des théories linguistiques actuelles qui supposent que la syntaxe est « dans » le lexique. Le programme chomskien est par exemple passé de la grammaire générative (Chomsky, 1957), où les règles grammaticales sont clairement séparées du lexique sur lequel elles s'appliquent, au programme minimaliste, où les règles syntaxiques sont très peu nombreuses et l'essentiel des informations sur la possibilité de combinaison des mots entre eux se trouve au niveau du lexique, dans chaque entrée lexicale (Levin, 1993).

[23] Des études sur les joueurs d'échec dans les années 1960 et 1970 (Chase & Simon, 1973) avaient déjà montré que ce qui différenciait les joueurs exceptionnels (grands maîtres) de joueurs experts plus ordinaires était la capacité des grands maîtres à percevoir les configurations de l'échiquier dans des ensembles (*chunks*) d'ordre supérieur ; par contre, ils ne retenaient pas mieux les positions de pièces placées au hasard sur l'échiquier.

niveaux de traitement et connaissances	niveau de difficulté prévisible (chap. 1)	corrélations constatées/ variance expliquée	références	remarques
Connaissances phonémiques	++	Corrélation faible (0,39) Pas de variance expliquée	Zoghlami 2015	Problème du passage à l'échelle (*scaling up*)/ *lossy chunking*
Connaissances prosodiques	++	Corrélation inexistante à moyenne (0,50), sans ou avec variance expliquée (28%)	Meerman et al. 2014 Tabata 2016	Peu d'études à ce jour, résultats peu clairs
Connaissances lexicales	+++	Corrélation faible (0,39) à importante (0,7). Variance expliquée de 14 à 50%	Cf. 2.3.2 (p. 78)	En général, connaissances les plus corrélées à CO (résultat robuste, beaucoup d'études)
Connaissances phraséologiques		(corrélation importante en CE)	Kremmel et al. 2015	Pas d'étude sur CO
Connaissances grammaticales	+	Corrélation faible (0,30) à importante (0,77). Pas de variance expliquée une fois les connaissances lexicales prises en compte.	Mecartty 2000 Zoghlami 2015 Andringa et al. 2012	Forte corrélation dans la seule étude utilisant un test grammatical aural (Andringa et al.)

Tableau 2.10 - corrélations constatées entre connaissances linguistiques et compréhension de l'oral (résumé des études présentées dans le chapitre 2)

En tout état de cause, ce chapitre nous a également permis de recueillir des informations précieuses sur les connaissances qu'il faudra inclure dans les tests diagnostiques que nous allons construire, parce qu'elles risquent particulièrement de faire défaut aux francophones en compréhension de l'oral : connaissances phonémiques, prosodiques, lexicales, phraséologiques et morphosyntaxiques. Dans le chapitre qui suit, le dernier de cette partie théorique, nous étudierons ce qu'implique la construction et l'utilisation de tests diagnostiques et l'analyse de leurs résultats.

Chapitre 3

Diagnostic et théorie des tests

Comme nous l'avons exposé dans l'introduction générale, le but de cette étude est de faire des propositions en vue d'élaborer une suite de tests diagnostiques qui permette d'identifier les lacunes et les acquis des apprenants en compréhension de l'oral, et de leur proposer des pistes de remédiation. Ceci correspond à la définition succincte d'un « test diagnostique » dans le *Multinlingual Glossary of Language Testing Terms* (Association of Language Testers in Europe, 1998, p. 242) : « Utilisé pour découvrir les points forts et les lacunes d'un apprenant. Les résultats peuvent servir à des prises de décision concernant la formation future, l'apprentissage ou l'enseignement. » Dans ce chapitre, nous allons présenter plus en détail les caractéristiques des tests diagnostiques tels que présentés dans la littérature, puis les analyses quantitatives et qualitatives qu'il est recommandé de suivre pour les valider scientifiquement, en mettant l'accent sur les concepts de validité et de fiabilité.

3.1. Caractéristiques des tests diagnostiques

3.1.1. Classification

Plusieurs classifications ont été proposées pour les tests de langue. Dans un des premiers articles théoriques sur le sujet, *Fundamental Considerations in Testing English Proficiency of Foreign Students* (1961), John B. Carroll propose de caractériser les tests par leur finalité (*purpose*). Il distingue les tests d'aptitude (*aptitude*), qui visent à prédire les chances de réussite des apprenants dans le cours de langue étrangère qu'ils s'apprêtent à suivre, et les tests sommatifs (*achievement*), pour s'assurer des progrès accomplis. Les tests diagnostiques ne sont pas mentionnés, même s'il est noté qu'une des caractéristiques importantes des tests doit être de permettre de diagnostiquer les faiblesses des apprenants: « *making useful diagnoses of learning difficulties or areas of ignorance on the part of the examinees* » (p.320).

Bachman (1990) reprend en partie la classification de Carroll dans un ouvrage dont le titre fait écho à son texte fondateur: *Fundamental Considerations in Language Testing*. Les finalités sont redéfinies comme les types de décision qui sont prises à partir des résultats aux tests. Ces résultats peuvent ainsi servir à sélectionner les candidats à une formation (*entrance tests*), à grouper les apprenants de niveau équivalent (*placement tests*), à choisir les activités ou le contenu le mieux adapté à l'apprenant (*diagnostic tests*, à propos desquels Bachman remarque que « *virtually any language test has some potential for providing diagnostic information* », p.60) ou à décider de passer à la partie suivante du programme quand les élèves maîtrisent la partie précédente (*achievement tests*, qui peuvent chez Bachman contrairement à Carroll être formatifs aussi bien que sommatifs). Bachman remarque que les tests de langue peuvent également être utilisés à des fins de recherche, en linguistique appliquée essentiellement (acquisition, enseignement, ou nature des compétences).

Brown (1996, p. 8), à partir d'une distinction entre tests normés (*norm-referenced*, dont les résultats sont interprétés par rapport à ceux de la population de référence) et tests critériés (*criterion-referenced*, dont les résultats sont interprétés par rapport au domaine de référence, c'est-à-dire aux contenus qu'on espère voir acquis), distingue quatre grandes catégories de tests de langue selon l'objectif visé. Parmi les tests normés, il place les tests de compétence générale (*proficiency*), type TOEFL, et les tests de positionnement (*placement*), qui servent à répartir les étudiants dans des groupes de niveau; parmi les tests critériés, les tests sommatifs (*achievement*), qui ont lieu en fin d'instruction pour vérifier que les buts ont été atteints et les contenus acquis, et enfin les tests diagnostiques (*diagnostic*), qui servent à indiquer les points forts et faibles des apprenants, en général avant le début de l'enseignement.

Cette dernière distinction entre quatre grands types de tests nous paraît la plus répandue actuellement (cf. aussi Hughes, 2002), et ménage aux tests diagnostiques une place importante. Cependant, comme le remarque Alderson (2005), la diffusion des tests diagnostiques reste finalement assez modeste. Même le test DIALANG, un test diagnostique dont il décrit le développement dans son ouvrage (*Diagnosing Foreign Language Proficiency: The Interface between Learning and Assessment*) est essentiellement utilisé comme test de positionnement (Alderson & Huhta, 2011, p. 32). Cela peut être dû au fait qu'il faut du temps, non seulement pour administrer un test diagnostique (le test DIALANG est composé de 5 parties indépendantes prenant chacune environ 30 minutes, sans compter le test de vocabulaire et l'auto-évaluation initiaux qui peuvent rajouter une demi-heure supplémentaire), mais aussi pour interpréter ensuite ses résultats et décider de la marche à suivre. Alderson et

al. (2015) montrent par exemple que les enseignants manquent souvent de temps pour établir des diagnostics fins de leurs élèves et leur proposer des parcours d'apprentissage individualisés correspondants. Une autre hypothèse pour expliquer le faible développement des tests diagnostiques est qu'ils ne présentent pas d'enjeux économiques forts, contrairement aux tests de compétence générale autour desquels s'est développée une industrie florissante: « *Within this examination and testing industry, there is currently very little or no use of diagnostic tests, which in any case, rarely exist*. » (Alderson et al., 2015, p. 2)

3.1.2. Avantages des tests à faible enjeu

3.1.2.1. définition

Les tests diagnostiques sont des tests à faible enjeu (*low-stakes*, Alderson et al., 2015, p. 2), ce qui présente plusieurs avantages du point de vue de leur développement et de leur administration. Le terme « à faible enjeu » renvoie au fait que les décisions prises suite à un test diagnostique n'ont pas de conséquences importantes sur la vie des apprenants, et peuvent être amendées facilement, comme on le voit dans le Tableau 3.1, traduit et adapté de Bachman (2004). Les décisions (prises par un enseignant ou une institution éducative) suite à un test diagnostique concernent essentiellement les activités de remédiation à donner à faire à un apprenant afin de l'aider à atteindre le niveau initialement attendu (dans notre cas, il s'agira d'activités supplémentaires à effectuer en ligne et en autonomie). Il est donc relativement simple de corriger une erreur de diagnostic si elle survient : l'apprenant peut simplement arrêter de suivre le programme de remédiation si son niveau s'avère suffisant. Le coût d'une décision erronée sera essentiellement une perte de temps pour l'apprenant qui aura fait des activités au final non nécessaires. Une erreur dans l'autre sens (étudiant en difficulté non repéré par le test diagnostique) peut certes être plus difficile à corriger et demande de la vigilance de la part des enseignants qui devront alors repérer de leur propre chef les difficultés de l'apprenant afin de l'orienter vers les activités de remédiation. Dans les deux cas, l'institution devra faire preuve de souplesse et de réactivité afin d'autoriser (s'il s'agit d'un programme insitutionnalisé) les changements d'affectation des participants. Cependant, il est possible que ce que les enseignants considèrent comme un enjeu faible ne soit pas envisagé de la même façon par certains apprenants : le fait d'être affecté à un programme de remédiation ou de soutien peut être mal vécu si on pense avoir un bon niveau, et peut ainsi donner l'impression de perdre la face. Un effort de présentation et un entretien personnalisé seront souhaitables dans ce cas-là.

Test à fort enjeu	Test à faible enjeu
Décision *majeure* qui affecte la vie du candidat	Décision *mineure*
Erreurs de décision *difficiles* à corriger	Erreurs *faciles* à corriger
Coût *élevé* des mauvaises décisions	*Faible* coût des mauvaises décisions

Tableau 3.1 - importance relative des décisions (d'après Bachman 2004, p.12)

Les tests de compétence générale qui servent de certification (TOEFL, CLES, etc.), sont, eux, des tests à fort enjeu, du fait des conséquences du résultat sur la vie future du candidat: ils peuvent être utilisés pour autoriser ou non une poursuite d'études (en particulier à l'étranger), pour l'obtention d'un diplôme, d'un visa de travail, d'un permis de séjour ou d'un poste en entreprise. Il est donc essentiel pour les candidats d'obtenir un bon résultat (souvent défini par un score minimal à atteindre). Les examens terminaux de fin d'année à l'université sont également des tests à fort enjeu, dans la mesure où ils déterminent le passage dans l'année supérieure ou l'obtention d'un diplôme.

3.1.2.2. problèmes de sécurité

Du fait de l'importance des résultats des tests à fort enjeu, les tentatives de fraude ne sont pas rares : « *As long as there have been tests for which important, high-stakes decisions are made, there have been people endeavoring to find a means for artificially inflating their scores* » (Wollack & Fremer, 2013, p. 1). Par ricochet, d'importantes mesures de sécurité sont nécessaires, et ce à plusieurs niveaux. Premièrement, les items ne peuvent pas être rendus publics, et doivent être renouvelés régulièrement afin de s'assurer que les candidats qui ont déjà passé le test ne puissent pas mémoriser les items ou les décrire à d'autres personnes qui s'apprêtent à le passer. Les items ont ainsi une durée de vie limitée, et un financement pérenne est nécessaire pour développer continuellement de nouveaux items afin de constituer une banque d'items conséquente.

Deuxièmement, les candidats peuvent parfois frauder pendant l'épreuve (copier les réponses sur les feuilles des candidats voisins, obtenir les réponses de l'extérieur, via des oreillettes, téléphones portables ou autres technologies, préparer à l'avance des réponses possibles, stockées sur support papier, numérique ou tout autre support plus créatif, ou envoyer à sa place un autre candidat plus compétent). En réaction, il est nécessaire de prévoir d'importantes mesures de sécurité. A titre d'exemple, le site du test IELTS géré par le British Council, annonce :

> *A strict set of protocols is in place to safeguard every aspect of the IELTS test, including: tight regulations surrounding test papers; biometric test taker registration and verification systems; training of test centre staff to help them identify imposters, detect fraudulent behaviour and prevent cheating; strict test conditions; routine scrutiny of test results before release.*[24]

Enfin, la fraude n'est pas nécessairement cantonnée aux candidats. Les centres d'examen sont également en concurrence entre eux pour attirer des candidats, et peuvent être le lieu de trafics lucratifs, par exemple en garantissant aux candidats un bon résultat en échange d'une somme d'argent[25]. Aux Etats-Unis, où un programme de tests systématiques à grande échelle a été introduit en 2001 (*No Child Left Behind*), visant à évaluer (et améliorer) l'efficacité de l'enseignement primaire et secondaire, le fait que les résultats aux tests déterminent les financements fédéraux obtenus et parfois le salaire des enseignants a conduit certains de ces derniers à modifier les réponses de leurs élèves (Jacob & Levitt, 2003), ou a poussé certains districts à décourager les élèves faibles de se présenter le jour du test. Ces cas de fraude ont parfois été découverts grâce à l'analyse automatique des données des tests, avec des algorithmes conçus pour repérer les anomalies (Levitt & Lin, 2015). Ces anomalies sont ensuite vérifiées en suivant les procédures prévues en cas de suspicion de fraude. Tout ceci nécessite bien sûr du temps, et un financement supplémentaire.

Les procédures sont plus simples pour un test diagnostique. En effet, ses résultats ne sont qu'indicatifs, ne sont utilisés que par l'apprenant et en général par l'enseignant responsable, et ne sont pas rendus publics ni utilisés par d'autres individus ou organisations. Les enjeux sont donc assez minces, et les erreurs sont finalement assez peu coûteuses, selon Bachman :

> *If a classroom teacher, for example, errs and decides to move to the next lesson before the class is ready, the cost is the amount of time and effort wasted before the error is discovered. In this case, the cost may be minimal if the teacher is able to make the necessary adjustment quickly.* (Bachman, 1990, p. 57)

De ce fait, il n'est pas nécessaire d'investir d'importantes sommes d'argent dans le développement de ce type de test :

> *If the costs associated with decision errors are minimal, then it would be wasteful to expend a great deal of time and effort to assure high levels of reliability and validity. On the other hand, if the potential costs of errors are great, it would be unethical not to make every effort to achieve the highest levels of reliability and validity possible.* (Bachman, 1990, p. 57)

---

[24] https://www.ielts.org/about-the-test/test-security

[25] Pour un exemple de scandale récent (2014) au Royaume-Uni, voir http://www.bbc.com/news/uk-26024375; pour les Etats-Unis (2019), voir https://en.wikipedia.org/wiki/2019_college_admissions_bribery_scandal

Cela ne revient pas à dire qu'il est inutile de faire des efforts pour qu'un test diagnostique soit de la meilleure qualité possible, bien sûr (nous reviendrons plus loin sur les questions de validité et de fiabilité), mais simplement que les erreurs éventuelles seront mieux pardonnées par le contexte de leur utilisation. D'autre part, un test diagnostique se révèle moins coûteux à concevoir, développer et administrer que d'autres types de tests, dans la mesure où aucune mesure de sécurité particulière n'est nécessaire à son administration et à la transmission des scores. Les questions qui le composent pourront également avoir une durée de vie plus longue, puisqu'il ne sera pas nécessaire de les changer régulièrement, les candidats n'ayant pas d'intérêt réel à les transmettre à d'autres futurs candidats (même si la culture de la « triche » pourrait pousser certains à appliquer cette stratégie en toutes circonstances, y compris quand les enjeux ne le justifient pas).

3.1.3. Intérêt de l'administration par ordinateur

3.1.3.1. capacités de stockage

La plupart des auteurs soulignent la capacité des tests diagnostiques à fournir une image assez détaillée des acquis des apprenants (nous avons mis en relief les passages intéressants dans les citations suivantes):

> *a test that has been designed and developed specifically to provide **detailed information** about the specific content domains that are covered in a given program or that are part of a general theory of language proficiency* (Bachman, 1990, p. 60);
>
> *for assessment information to be used effectively it needs to be **detailed**, innovative, relevant and diagnostic, and to address a variety of dimensions rather than being collapsed into one general score.* (Shohamy, 1992, p. 515);
>
> *It is difficult and time-consuming to construct a test which provides **detailed diagnostic information*** (A. Davies et al., 1999, p. 43);
>
> *Diagnostic tests are used to identify learners' strengths and weaknesses. […] But it is not so easy to obtain a **detailed analysis** of a student's command of grammatical structures* (Hughes, 2002, p. 15);
>
> *Diagnostic tests should enable a **detailed analysis** and report of responses to tasks, and must give **detailed feedback** which can be acted upon*. (Alderson, 2005, p. 257).

Cette exigence de résultats détaillés permet de mieux identifier les besoins de chaque apprenant, mais il en découle une durée importante. Comme nous l'avons déjà noté, par exemple, le test DIALANG peut durer deux heures et demie (ou plus, selon la vitesse de l'étudiant) si toutes ses sous-parties sont utilisées. Par ailleurs, il faut trouver un moyen de

conserver ces informations détaillées de telle façon qu'elles soient exploitables par la suite, ce qui peut se révéler difficile, comme le remarque Alderson (2015, p. 38): « *the average SFL classroom teacher is faced with many learners, and diagnosing, or even remembering which learner has which problem is an enormous task.* »

Plusieurs auteurs soulignent que ces problèmes peuvent être en partie résolus grâce à l'outil informatique. Ainsi, la citation d'Alderson reproduite ci-dessus continue de la manière suivante: « *Hence, computer-based diagnosis, and the keeping of electronic records tracking diagnosis, treatment and progress, should be tools that every SFL teacher has available and knows how to use* » (ibid., p. 38). Le besoin de conserver des résultats détaillés pour chaque apprenant fait donc de l'ordinateur un outil privilégié de passation de ces tests. Par ailleurs, l'ordinateur, couplé à la passation en ligne sur internet (et le fait que ce sont des tests à faible enjeu, donc moins sujets à des problèmes de sécurité), permet également de rendre plus agréable le temps important consacré à ces tests: « *delivering tests over the Internet means that individuals can take a test at the time and place of their choosing* » (Alderson, 2005, p. 254).

3.1.3.2. rétroaction

Un autre avantage des ordinateurs et de leur capacité de stockage et de traitement est le fait qu'ils peuvent produire un retour formatif (*feedback*) facilement et immédiatement. Le *feedback* (parfois aussi traduit en français par « rétroaction », ALTE 1998, p.238) est défini comme « *actions taken by (an) external agent(s) to provide information regarding some aspect(s) of one's task performance* » (Kluger & DeNisi, 1996). Son importance est connue depuis longtemps, puisque dès le début du 20ème siècle, Edward Thorndike (1927) décrivait une « loi de l'effet » (*law of effect*). L'expérience consistait à demander à des sujets d'estimer la longueur d'une languette de papier, d'abord sans rétroaction (c'est la partie qui serait appelée « prétest » dans un schéma d'expérimentation moderne), puis avec rétroaction si le sujet était dans le groupe expérimental, ou toujours sans s'il était dans le groupe contrôle, puis de nouveau sans rétroaction pour tout le monde (la partie « post-test »). Les groupes expérimentaux, à qui on a dit dans la partie médiane si leur réponse était juste ou fausse (sans d'ailleurs leur donner la bonne réponse, ni leur dire dans quel sens ils s'étaient trompés), se sont beaucoup plus améliorés au post-test que le groupe contrôle (une deuxième expérience décrite dans le même article, où les sujets apprennent à dessiner des lignes d'une longueur donnée avec un bandeau sur les yeux, donne des résultats allant dans le même sens, mais

moins fiables). C'est une première démonstration de l'importance du feedback, dont l'effet est supposé ici renforcer la connexion entre la question et la réponse (dans la cadre de la théorie behavioriste de l'époque où une rétroaction positive était vue comme une récompense, une négative comme une punition). Plus récemment, le livre *How People Learn* (Bransford et al., 2000), dont l'ambition était de vulgariser les résultats de la recherche en acquisition et en éducation pour qu'elle puisse être utilisée dans les classes par les enseignants, affirmait également: « *In order for learners to gain insight into their learning and their understanding, frequent feedback is critical: students need to monitor their learning and actively evaluate their strategies and their current levels of understanding* » (p.78). Enfin, la « méta-méta-analyse » faite par John Hattie[26] dans son livre *Visible Learning: A Synthesis of Over 800 Meta-analyses Relating to Achievement* (2009) identifie le feedback comme un des éléments majeurs de la situation pédagogique. Cependant, assurer un retour formatif pour chaque apprenant prend du temps, comme le savent bien les enseignants qui n'apprécient pas toujours l'activité répétitive et prenante de correction de copies. Les ordinateurs, eux, peuvent fournir ce retour automatiquement pour les tâches simples que sont généralement les questions incluses dans les tests diagnostiques (QCM et apparentées).

D'autres chercheurs se sont posé la question du moment à privilégier pour cette rétroaction, en particulier s'il est préférable de la proposer tout de suite après la réponse de l'étudiant (*feedback* immédiat) ou à un autre moment, par exemple à la fin de l'activité (*feedback* différé, ou *delayed feedback* en anglais). La méta-analyse d'Azevedo et Bernard (1995) montre que, dans un contexte informatique (*computer-based instruction*), la rétroaction immédiate est plus efficace que la rétroaction différée. La première est associée à une taille d'effet de 0,8, c'est-à-dire un effet important d'après l'échelle de Cohen (1992), tandis que la deuxième a un effet petit à moyen (0,35, un effet tout de même non négligeable par rapport à l'absence de rétroaction).

Une autre méta-analyse – plus complexe – de la même époque, celle de Kluger et DeNisi (1996), montre que la rétroaction est efficace surtout quand elle est dirigée vers le niveau le plus « bas », celui de la tâche, un peu moins efficace si elle est dirigée vers le niveau intermédiaire, celui de l'organisation de la tâche (par exemple, choix des stratégies). Son effet peut même être négatif s'il porte l'attention sur le niveau « haut », c'est-à-dire le sujet (*self*) lui-même (par exemple, un compliment peut avoir un effet négatif sur les performances

[26] certes controversée dans sa méthode, cf. Higgins & Simpson (2011)

futures). Ces conclusions rejoignent celles d'Alderson (2005) dans son livre sur DIALANG. Alderson décrit les analyses qualitatives basées sur des entretiens ou des observations d'étudiants en train d'utiliser le système DIALANG au moment de son pilotage, et avant son déploiement définitif. Les formes de rétroaction préférées des sujets interviewés étaient les résultats détaillés par item donnés tout de suite après la fin du test, et montrant quels items avaient été réussis, et lesquels non, en les classant selon leurs objectifs. L'étudiant peut ainsi voir combien d'items de compréhension aurale testant l'inférence il a réussi, et lesquels précisément (il peut avoir accès au texte de l'item en recliquant dessus), ou bien combien d'items de structure grammaticale sur les verbes, etc. Il s'agit d'un feedback centré sur la tâche, avec des informations très précises. Par contre, un type de rétroaction beaucoup moins apprécié était ce qui s'appelle dans DIALANG « *explanatory feedback* », qui essaie d'identifier les causes de l'écart éventuel entre le résultat de l'autoévaluation par l'étudiant et son résultat dans DIALANG. Un exemple de réaction négative montre que c'est le fait que cette rétroaction se rapproche du niveau du sujet et s'éloigne de la tâche qui a pu susciter ce rejet chez certains étudiants: « *One student could not understand why a test should provide learners with such information, especially if […] some of the feedback related to personality issues.* » (Alderson, 2005, p. 216)

Nous pouvons retenir pour les tests que nous allons concevoir qu'un *feedback* clair et détaillé sera à prévoir, à destination des étudiants aussi bien que des enseignants. Ces informations devront être disponibles tout de suite après l'achèvement du test.

3.1.3.3. individualisation

L'administration par ordinateur permet également une plus grande individualisation : le temps passé sur chaque tâche par les étudiants peut varier, le nombre d'écoutes peut être adapté, et le retour formatif (*feedback*) peut être personnalisé également. Cela permet à la fois d'aller plus dans le détail des acquis, et d'individualiser l'enseignement, qui pourra être adapté aux besoins de chaque apprenant, besoins qui seront mieux identifiés grâce au test. Ainsi, on s'approche des éléments nécessaires à ce que Terrier et Maury (2015) ont appelé une « autoformation individualisée de masse » (§55). Nous reviendrons dans la dernière partie, au moment de la description du dispositif pédagogique envisagé, sur les caractéristiques nécessaires à un dispositif d'autoformation fructueux.

3.1.4. Autres caractéristiques

Dans son ouvrage écrit avec des collègues finlandais ayant collaboré aux projets DIALANG et DIALUKI, *The Diagnosis of Reading in a Second or Foreign Language* (Alderson et al., 2015), Alderson fait le parallèle avec le diagnostic dans d'autres domaines: automobile, médical, et orthophonie (chapitre 2, pp.18-38). Premièrement, le diagnostic doit permettre d'identifier les faiblesses plutôt que les points forts des étudiants, puisque le but du diagnostic est de proposer ensuite une remédiation ou d'adapter le cours pour répondre aux besoins ainsi identifiés. Quand un patient va voir son médecin, par exemple, il s'y rend généralement parce qu'il a un problème et a besoin du médecin pour identifier plus précisément les causes de ce problème. Il ne s'attend pas à ce que le médecin lui énumère toutes les fonctions physiologiques qui fonctionnent bien chez lui. De même, un mécanicien automobile essaye d'identifier la source d'un problème mécanique. Alors que les descripteurs du CECR se focalisent sur ce que les apprenants savent faire (« *can-do statements* »), et qu'ils doivent être « formulé[s] de manière positive » (Conseil de l'Europe, 2001, p. 30), il est donc important qu'un test diagnostique puisse identifier ce que les apprenants ne savent pas (encore) faire (en quelque sorte, il s'agirait de « *can't-do statements* »). Alderson et ses collègues remarquent également que si un diagnostic peut souvent être posé en orthophonie L1 suite à un constat de déviation par rapport à la norme (c'est-à-dire ce que savent faire en général les apprenants L1 à tel âge), il n'existe pas de norme en L2 à laquelle on puisse comparer les performances des apprenants. Il faut alors se contenter de constater une déviation par rapport à un niveau attendu (à telle étape d'une formation, par exemple), qu'il corresponde ou non au développement normal d'un apprenant donné.

Enfin, il faut si possible que les tests diagnostiques soient basés sur une théorie du développement de la compétence langagière (Alderson, 2005, p. 3, *« informed by SLA research, or more broadly by applied linguistic theory as well as research »*) afin de s'assurer que les points testés soient vraiment cruciaux pour l'amélioration du niveau de langue. Bachman souligne lui aussi qu'un test diagnostique est « *a test that has been designed and developed specifically to provide detailed information about the specific content domains that are covered in a given program or that are part of a general theory of language proficiency. Thus, diagnostic tests may be either theory or syllabus-based* » (Bachman, 1990, p. 60). Etant donné que nos tests ne sont pas liés à des cours ni à un programme particuliers, mais sont au contraire destinés à être utilisés en autonomie en parallèle de la formation universitaire choisie

par l'étudiant, ils doivent être basés sur une théorie psycholinguistique. C'est pour cette raison que nous avons consacré les chapitres précédents à une description des modèles possibles de la compréhension de l'oral et de ses composantes identifiées dans la littérature.

3.1.5. Analyse de tests existants

3.1.5.1. DIALANG

Si de nombreux tests de compétence générale ou de positionnement sont disponibles pour l'anglais, il existe pour l'instant peu de vrais tests diagnostiques. L'un des plus utilisés, parce qu'il est gratuit et ouvert à tous, est le test en ligne DIALANG[27], hébergé à l'université de Lancaster et développé dans le cadre d'un projet européen en parallèle avec l'introduction du CECRL (Alderson 2005). Le retour diagnostique fournit à l'utilisateur plusieurs niveaux d'information. Tout d'abord, l'utilisateur reçoit son niveau par habileté (compréhension de l'oral, de l'écrit, production écrite, vocabulaire, grammaire), indexé sur la grille de niveaux du CECRL, avec une description de ce que recouvre ce niveau, et éventuellement une description des niveaux supérieurs et inférieurs pour comparer. Ensuite, si ce niveau est différent du résultat de l'autoévaluation (facultative) de l'étudiant pour chaque compétence, une explication des raisons possibles de cette différence est proposée (comme nous l'avons vu plus haut, cette information n'est pas particulièrement appréciée des étudiants). Enfin, l'utilisateur a accès au résultat commenté pour chaque item du test, avec une identification de la bonne réponse et de la sous-habileté que l'item est censé tester (par exemple, compréhension globale, ou inférence).

DIALANG propose donc à ses utilisateurs un *feedback* immédiat et très détaillé. Par contre, contrairement au vœu de son concepteur principal (Alderson 2005), les détails du *feedback* diagnostique ne sont pas liés à une théorie du développement des compétences langagières. Les résultats donnés portent toujours sur des micro-habiletés qui couvrent le construit de la compétence étudiée, et qui correspondent donc à tout ce que l'étudiant a su ou n'a pas su faire. Le problème est que ces micro-habiletés ne sont pas mises en relation avec une théorie du développement de la compétence, ce qui fait qu'on retrouve les mêmes sous-habiletés à tous les niveaux. Alderson (2005) montre par exemple qu'il n'y a pas de corrélation entre le niveau des candidats et les sous-habiletés qu'ils ont acquises ou non: ceux de niveau A (utilisateurs élémentaires) sont aussi susceptibles que les niveaux B (utilisateurs indépendants) de savoir

[27] https://dialangweb.lancaster.ac.uk/

inférer, avoir une compréhension générale ou répondre à une question de compréhension détaillée (les trois sous-habiletés de la compréhension dans DIALANG, et qu'on retrouve sous une forme ou une autre dans les autres tests) : « *one's abilities in listening subskills did not relate to a learner's CEFR level or to their performance on the Listening test as a whole. Even low-level learners are able to answer some questions that test inferencing abilities, as well as items testing the ability to understand main ideas.* » (p.146). Il n'est donc pas forcément justifié de diriger ensuite les étudiants vers des activités travaillant telle ou telle micro-compétence, puisque ce n'est pas ce qui déterminera leur passage au niveau supérieur.

3.1.5.2. DELNA et DELTA

Il existe d'autres tests diagnostiques, comme le DELNA (*Diagnostic English Language Needs Assessment*) en Australie (J. Read, 2008), qui permet d'identifier les étudiants nouvellement admis à l'université qui ont besoin d'aide en anglais. Les résultats au test sont présentés sous la forme de six profils-types assortis de recommandations de cours à suivre. A Hong Kong, le DELTA (*Diagnostic English Language Tracking Assessment*) est un test diagnostique que les étudiants peuvent passer à plusieurs reprises pendant leur scolarité pour se rendre compte de leur progrès (Lockwood, 2013). Le résultat indique les sous-habiletés réussies ou non (identifier une information spécifique, inférer le raisonnement du locuteur), ainsi que les types de textes (dialogues et conversations, interviews radio/TV), et les thèmes (commerce et marketing, médias et communication, voyages) qui sont associés à ces textes (Urmston et al., 2013). Cependant, il est assez peu utilisé, probablement parce que l'enjeu n'est pas assez important et qu'il n'est pas intégré aux cours de langue (Urmston & Raquel, 2015). Il souffre également du même problème que DIALANG, en ce que l'information donnée en sortie, c'est-à-dire les sous-habiletés, les thèmes et les types de textes bien ou mal réussis, ne sont pas non plus liés à une théorie du développement des compétences. Il s'agit simplement d'une description des types d'items bien ou mal réussis, une information sur le produit de l'activité de compréhension, et non sur les étapes du processus où se produisent les blocages.

D'après nos recherches, DIALANG, DELTA et DELNA sont les seuls tests diagnostiques utilisés à grande échelle, même si se développent aussi des interprétations diagnostiques des tests de compétence générale, qui fournissent aux étudiants un retour sur les types d'items réussis ou non (Liu, 2015). Etant donné leurs faiblesses théoriques, ces tests se révèlent insatisfaisants, et il semble nécessaire de créer un nouveau test diagnostique des difficultés de compréhension de l'oral basé sur le modèle de compréhension aurale que nous avons

présenté, et dont nous avons identifié les étapes problématiques pour les apprenants L2. C'est ce que nous ferons dans la deuxième partie.

3.2. Théorie Classique des Tests

3.2.1. Validité

3.2.1.1. validité - aspects qualitatifs

Tout test doit passer par un processus de validation, défini par le glossaire d'ALTE comme « processus par lequel on rassemble des preuves pour étayer les conclusions données par les notes des tests » (Association of Language Testers in Europe, 1998, p. 244). Plus simplement, on peut dire que « un test est valide s'il mesure ce qu'il a l'intention de mesurer » (Tagliante et al., 2004, p. 116). Cela suppose tout d'abord que l'on définisse ce que l'on essaye de mesurer. Dans notre cas, il s'agit de montrer que chacun des tests diagnostiques que nous allons concevoir teste la sous-habileté visée (sensibilité phonémique ou accentuelle, reconnaissance du vocabulaire oral, …), ou « construit » du test. Nous suivrons pour ce travail les recommandations de Hughes (2002)[28].

Le premier critère de validation est la validité de contenu (*content validity*). Il faut montrer que le contenu du test, d'après Hughes (2002, p.26), « *constitutes a representative sample of the language skills, structures, etc. with which [the test] is meant to be concerned.* ». Pour juger de la validité de contenu d'un test, il faut donc disposer de spécifications précises qui décrivent la conception du test, et permettent de décider si les items dont il est constitué sont effectivement représentatifs du construit testé. Nous expliquerons en détail dans la deuxième partie de cette thèse comment nous avons choisi les items de nos tests diagnostiques, en étroit lien avec les conclusions théoriques présentées aux chapitres 1 et 2.

Le deuxième élément permettant d'établir la validité d'une démarche évaluative est la validité externe ou « de critère externe » (*criterion-related validity*), démontrée par une comparaison entre l'instrument et une autre mesure fiable de la même compétence ou sous-habileté : « *the degree to which results on the test agree with those provided by some independent and highly dependable assessment of the candidate's ability* » (Hughes, 2002, p. 27). Alors que la validité de contenu doit être prouvée, ou du moins étudiée, avant même que le test soit

[28] Il existe des techniques plus récentes inspirées par le travail de Messick (1990), pour qui la validité est « *an integrated judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment* » (p. 1487), mais plus compliquées à mettre en œuvre, et inutilement sophistiquées pour des tests à faible enjeu.

assemblé et soumis à l'épreuve du terrain, la validité de critère ne peut l'être qu'une fois les résultats de la première expérimentation connus. Ces résultats doivent être comparés à ceux d'un autre test reconnu (*concurrent validity*), ou bien peuvent être corrélés à un critère de réussite (ou non) à la formation, pour montrer que le test avait une validité prédictive (*predictive validity*) et qu'il peut donc servir par la suite à prendre des décisions qui concernent la poursuite d'études, par exemple. Dans les deux cas, une corrélation sera calculée (avec le coefficient de Pearson si les résultats sont numériques, la distribution est normale ou les effectifs sont suffisamment importants). Une valeur acceptable pour ce coefficient sera de 0,7 dans le premier cas (pour un test dont les enjeux ne soient pas trop élevés), et de 0,4 dans le deuxième cas (ibid., p. 26-27). En effet, il est beaucoup plus difficile pour un test passé à un instant *t* de prédire la réussite du candidat plusieurs mois plus tard, de nombreux autres facteurs pouvant entrer en ligne de compte dans l'intervalle.

Toutes ces études aident à définir le « construit » (*construct*) du test, c'est-à-dire ce que les concepteurs cherchent à tester. Cependant, dès qu'il s'agit d'un test indirect (comme un test de compréhension), les spécifications et l'analyse des items composant le test ne sont pas forcément suffisantes. Les tests visant les compétences réceptives sont qualifiés d'indirects car il n'est pas possible d'observer la compréhension directement (les tests de compétence productive peuvent par contre être directs puisqu'on peut tester directement l'activité de production orale ou écrite). Afin de vérifier que le test met bien en jeu les compétences visées, il est possible d'utiliser la technique du « *think aloud* », c'est-à-dire de demander aux candidats de verbaliser leurs pensées alors qu'ils sont en train de répondre aux items du test. C'est également ce que Bachman (1990) appelle l'authenticité interactionnelle, qui montre que les apprenants interagissent avec les items du test comme ils le feraient dans une situation authentique de la vie réelle. Pour un test de compréhension de l'oral, il n'est pas vraiment possible aux candidats de verbaliser leurs processus mentaux pendant qu'ils écoutent, puisque le texte oral et la verbalisation utilisent le même canal. Il est cependant possible d'arrêter régulièrement l'enregistrement pour leur demander ce dont ils viennent de faire l'expérience (cf. par exemple Zoghlami, 2016). On s'approche alors d'un protocole rétrospectif.

3.2.1.2. validité - aspects quantitatifs

Aucun chiffre ne peut résumer à lui seul la validité d'un test, mais il existe un certain nombre d'analyses quantitatives possibles. Tout d'abord, comme dans toute analyse, il est important d'avoir une idée générale de la structure des données. Cette vue d'ensemble est obtenue par

des mesures de tendance centrale (la moyenne et la médiane des résultats de l'échantillon) et des mesures de dispersion des résultats : l'étendue (*range*), c'est-à-dire la différence entre le résultat le plus haut et le plus faible, et l'écart-type (*standard deviation*), une mesure des écarts à la moyenne. Nous inclurons également une mesure d'asymétrie (*skew*), qui est positive quand les scores au-dessus de la moyenne sont plus variés que les scores en dessous de la moyenne, et négative dans la situation inverse. Cela peut être visualisé avec un histogramme de fréquence (Figure 3.1) : une mesure d'asymétrie positive correspond à une « queue » plus longue à droite (figure de droite) et une mesure d'asymétrie négative à une « queue » plus longue à gauche (figure de gauche). Le but dans l'étude étant de diagnostiquer les étudiants ayant besoin de remédiation, c'est dans les scores faibles que nous aurons besoin de plus de détails. Une asymétrie négative serait donc préférable à une asymétrie positive.

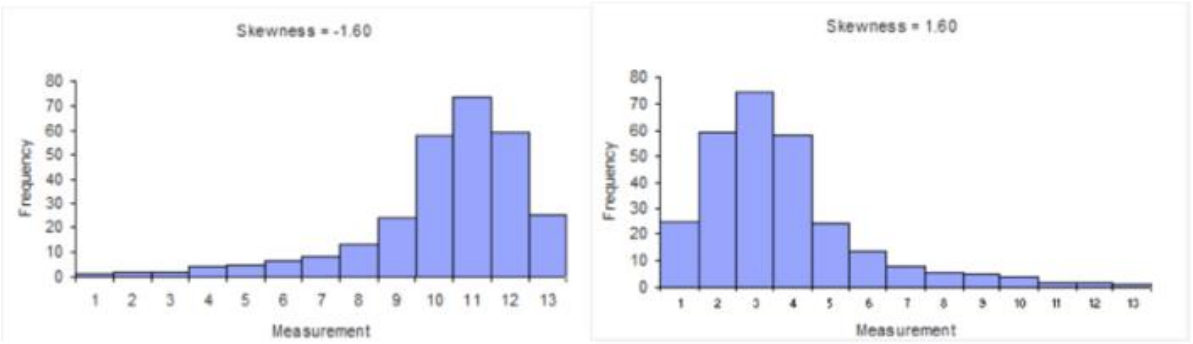


Figure 3.1 - asymétries négative et positive (d'après le site Good Data[29])

Après une première description générale des résultats, il faut vérifier de façon détaillée la qualité de chaque item. Le « coefficient de difficulté » de l'item correspond au pourcentage de candidats qui ont trouvé la réponse juste à cet item (on parle aussi de « facilité » de l'item, du fait que plus il est élevé, plus l'item est facile). Les items ne doivent être ni trop faciles ni trop difficiles. Un item réussi par tout le monde, par exemple, ne peut pas contribuer à diagnostiquer des lacunes. A l'inverse, un item qui n'est réussi par aucun candidat ne nous aidera pas à différencier entre les étudiants n'ayant pas besoin de remédiation et les autres. Le coefficient de difficulté sera donc de préférence compris entre 20 et 80 pourcent (Bachman 2004, p.138). Pour des tests diagnostiques, il vaut mieux que les items soient trop faciles que trop difficiles. En effet, certains items très faciles qui ne sont ratés que par des étudiants très faibles pourront nous aider à identifier ces derniers. Nous accepterons donc des indices allant

[29] https://help.gooddata.com/doc/en/reporting-and-dashboards/maql-analytical-query-language/maql-expression-reference/aggregation-functions/statistical-functions/predictive-statistical-use-cases/normality-testing-skewness-and-kurtosis

jusqu'à 90-95%, à condition que le coefficient de discrimination soit suffisamment élevé (cf. paragraphe qui suit).

Ensuite, il faut vérifier que les questions sont mieux réussies par les candidats dont le niveau est plus élevé que par ceux dont le niveau est plus faible. Ceci est fait grâce à un « coefficient de discrimination », le coefficient de corrélation bisérial de point, qui calcule la corrélation entre la réussite à l'item et la réussite au test. Ce coefficient doit être égal à 0,2 au moins pour que l'item soit de discrimination acceptable, et supérieur à 0,3 pour un item de bonne discrimination (Laveault & Grégoire, 2014, p. 211).

Une dernière vérification importante pour déterminer la validité d'un instrument évaluatif est celle de l'unidimensionnalité du test. Il s'agit de vérifier mathématiquement que tous les items sont corrélés entre eux et vont dans le même sens (varient ensemble), ce qui nous permet d'apporter un argument supplémentaire pour montrer empiriquement que tous les items testent le même construit. Les items d'un test diagnostique donné devraient également être plus corrélés entre eux qu'avec les items de tests visant à mesurer d'autres facteurs. Par exemple, les items du test de sensibilité accentuelle devraient être plus corrélés entre eux qu'avec les items du test de discrimination phonémique, même si on s'attend à ce que les deux tests partagent une variance non négligeable du fait qu'ils font tous les deux appel au traitement du signal sonore et sont dépendant du même input (l'anglais oral). Pour faire ces analyses, nous utiliserons une technique similaire à l'analyse factorielle (l'alpha de Cronbach nous donne une première indication, mais n'est pas suffisant selon Laveault & Grégoire, 2014), à savoir la technique d'analyse en composantes principales. Cette technique vise à vérifier qu'il existe une variable (la composante principale) qui permet d'expliquer l'essentiel de la variance observée chez tous les items d'un test (A. Field et al., 2012, p. 760).

Pour la plupart de ces analyses, nous utiliserons le logiciel libre *R*, qui peut être téléchargé librement en ligne (R Development Core Team, 2005), avec l'interface *R Studio*[30]. Cette suite de programmes possède une base importante de fonctionnalités qui permettent de réaliser les actions les plus importantes : importation de fichier sous forme de tableaux (.csv), traitement sur ces tableaux (modifications, ajout ou suppression de lignes ou de colonnes, et opérations sur ces données), production de graphiques. Par ailleurs, des « bibliothèques » de fonctions (couramment appelées « *packages* » y compris par les utilisateurs francophones) automatisant des opérations supplémentaires ou proposant des graphiques plus élaborés peuvent être

[30] https://www.rstudio.com/

téléchargés facilement pour enrichir les fonctions de base. Nous signalerons systématiquement quels *packages* ont été utilisés pour les analyses que nous proposerons. Ces analyses sont effectuées à l'aide d'appels successifs de fonctions qui peuvent être enregistrées dans un « script » *R* (une suite de fonctions commentées) qui peut ensuite être exécuté de nouveau avec un autre fichier, sans qu'il soit besoin de repasser par chacune des étapes. Des exemples de scripts utilisés sont disponibles dans l'Annexe 7.

3.2.2. Fidélité

3.2.2.1. définition et calcul

La fidélité (ou fiabilité, *reliability*) d'un instrument d'évaluation est sa capacité à produire toujours les mêmes résultats dans les mêmes conditions d'administration, ou sa capacité à démontrer « uniformité, constance ou stabilité des mesures » (Association of Language Testers in Europe, 1998, p. 224). Cette stabilité est essentielle pour garantir la qualité d'un instrument de mesure (par exemple, on s'attend à ce qu'un thermomètre donne toujours le même résultat à température égale). En effet, un test qui n'est pas fidèle est forcément injuste envers certains candidats. Si deux personnes de même habileté sous-jacente (ou la même personne passant le test deux fois) reçoivent deux résultats différents, l'une des deux a forcément été sous-évaluée ou sur-évaluée, et la personne recevant le score inférieur est désavantagée par rapport à l'autre. Ce problème est particulièrement aigu quand le test n'est pas corrigé de manière automatique : il faut alors s'assurer que les différents correcteurs donnent la même note à la même prestation (dans la mesure où nos tests seront corrigés de manière automatique, ce problème ne nous concerne pas directement). Dans le cas d'un test de langue, il est normal que les résultats d'un même étudiant varient légèrement d'une administration à l'autre (ou d'un correcteur à l'autre le cas échéant). Cependant, la variation doit être limitée le plus possible afin que le test soit fiable et les résultats interprétables.

Comment mesurer la fidélité d'un outil d'évaluation donné ? Il n'est pas possible de demander aux étudiants de passer le même test deux fois de suite pour comparer les résultats. Les conditions de passation ne seraient dans ce cas pas les mêmes : la deuxième fois, les étudiants peuvent se souvenir de certaines questions, et ils peuvent aussi être moins motivés puisqu'ils ont déjà passé le test une fois. Les résultats ne seraient donc pas comparables. On pourrait décider de laisser passer suffisamment de temps entre les deux administrations pour s'assurer que personne ne se souvienne des questions. Même si cela était possible, il est probable que le niveau des candidats aurait évolué dans l'intervalle, et encore une fois les

résultats ne seraient pas comparables. L'approche suivie en pratique est donc d'utiliser les résultats d'une seule passation pour estimer la fidélité d'un test. Cela peut se faire de plusieurs manières, décrites tour à tour dans le paragraphe qui suit.

La première méthode est celle des formes équivalentes (*alternate forms method*). Au lieu de faire passer le même test deux fois, on administre deux formes équivalentes du test afin de comparer les résultats. Le problème est bien sûr qu'il est très rare de posséder à l'avance deux formes équivalentes d'un test. La deuxième est la méthode de bipartition ou bissection (*split-half method*), qui simplifie la méthode précédente en supposant que les étudiants passent les deux formes équivalentes en même temps. Le test est divisé en deux parties (qu'on espère) équivalentes, et on calcule la corrélation entre le résultat aux deux parties. Toute la difficulté réside dans la division en deux parties équivalentes, en appariant les items deux à deux. C'est pour cela qu'une troisième méthode est généralement choisie, qui pousse la logique de la méthode de la bipartition à l'extrême, le coefficient alpha (Cronbach, 1951). Ce coefficient (noté α) correspond à la moyenne des coefficients de toutes les bissections possibles, donc de toutes les façons de diviser le test en deux parties[31] (Fulcher, 2010, p. 51). Il est considéré comme une mesure de la cohérence interne (*internal consistency*) du test, parce que plus il est élevé, plus les items vont dans le même sens. On suppose alors que ce parallélisme de comportement signifie que les items mesurent la même chose (cela devra cependant être confirmé par une analyse de l'unidimensionnalité du test, comme nous l'avons vu au paragraphe précédent). Les items doivent en effet varier dans le même sens parce que la variance des items doit refléter une compétence linguistique sous-jacente (le « trait ») et non d'autres caractéristiques individuelles ou l'effet du hasard. Le coefficient alpha doit être supérieur à .7 pour que la fidélité soit considérée comme acceptable (Laveault & Grégoire, 2014, p. 119).

3.2.2.2. améliorer la fidélité

De façon générale, afin d'assurer une bonne fidélité, il est important d'obtenir suffisamment d'informations sur chacune des personnes testées (un test avec une seule question aura moins de chance d'être fiable qu'un test avec beaucoup de questions, où une erreur ponctuelle ne sera pas rédhibitoire). Il faut donc utiliser suffisamment d'items, et il faut que ces items soient indépendants : la réponse à une question ne doit pas conditionner les suivantes. En effet, si un

[31] du moins si une condition est respectée, celle de l'égalité des écarts-types des items, en pratique rarement vérifiée, cf. Laveault et Grégoire (2014, p. 120)

candidat ne réussit pas cette question, il a peu de chance de réussir les autres et les items suivants ne serviront pas à grand-chose. D'après Hughes (2002, p.44), « *Each additional item should as far as possible represent a 'fresh start' for the candidate. By doing this we are able to gain additional information on all the candidates – information that will make the test results more reliable.* »

Il est important également que les modalités de passation soient clairement définies afin que les conditions d'administration soient les plus similaires possibles pour tous les candidats. Elles doivent aussi être les plus agréables possibles, afin que les candidats réussissent à montrer ce dont ils sont capables. Pour un test de compréhension de l'oral ou de discrimination auditive, une qualité du son identique pour tous les candidats, et la meilleure possible, sont des éléments essentiels. Dans notre cas, tous les tests ont été passés dans des salles équipées d'ordinateur et de casques à l'Université Grenoble Alpes en présence d'un(e) enseignant(e), même si à terme ils seront disponibles à distance et sans supervision.

3.2.3. Utilité

Depuis la fin des années 1980, les chercheurs en évaluation ont recentré le concept de validité sur les conséquences de l'usage des tests (Bachman, 1990; Messick, 1990). En effet, si un test mesure bien ce que l'on désire tester mais qu'il est détourné pour d'autres usages, ou s'il amène d'autres conséquences négatives non prévues au départ (*negative washback*), alors le test, ou du moins ses utilisations, ne peut pas être considéré comme valide(s) : « *the purpose of a test is central to test validity: the validity of a test derives directly from the inferences and actions that are based on the interpretations of a test score, and cannot be separated from the purpose on the basis of which a test was constructed* » (Alderson & Huhta, 2011, p. 31). Il est donc essentiel pour les concepteurs d'exposer le plus clairement possible le construit de leurs tests et les contextes dans lesquels ils imaginent qu'ils peuvent être utilisés. C'est ce que nous ferons dans les chapitres qui suivent.

Mentionnons pour finir les concepts de praticité et de validité faciale. Si l'on désire qu'un test soit effectivement utilisé, il faut qu'il soit pratique, c'est-à-dire que sa mise en œuvre n'excède pas les ressources disponibles (Bachman & Palmer, 1996, p. 35-37). Ces ressources comprennent le temps dont on dispose, à la fois pour la conception du test, mais aussi pour son administration. Enfin, la validité faciale est importante pour que le test soit accepté comme valide par ses utilisateurs (Hughes 2002, p.33). Il faut en effet que les candidats (mais

également l'administration qui s'occupe d'organiser les passations) aient confiance dans la validité et l'utilité du test afin d'être motivés pour donner le meilleur d'eux-mêmes. C'est pourquoi les formats de questions non traditionnels ou trop indirects, qui peuvent décontenancer les candidats qui n'ont pas l'habitude d'être testés de cette façon, doivent être soigneusement expliqués et justifiés. Comme nous envisageons à terme une passation des tests en ligne et en autonomie, sans la présence d'enseignants pour expliquer et justifier, nous essaierons d'éviter les formats d'items trop originaux ou trop compliqués.

3.3. Modèles de Réponse à l'Item

Les analyses statistiques exposées dans la section 3.2 (p. 123) utilisent des indices dont la valeur dépend des sujets à qui on a administré les tests analysés. Par exemple, un même item aura un indice de difficulté différent s'il est passé par un groupe de niveau débutant ou de niveau avancé. Pour comparer le niveau de plusieurs étudiants, il faut donc leur faire passer les mêmes questions et comparer leurs scores. Cependant, il existe d'autres techniques psychométriques qui permettent de comparer le niveau de candidats qui ne répondent pas exactement aux mêmes questions. Les candidats débutants, par exemple, ne sont alors pas obligés de passer les questions difficiles, et les candidats avancés peuvent « sauter » les questions faciles. Ces tests s'appellent des tests adaptatifs, dans lesquels le choix de questions à soumettre dépend du niveau des candidats, estimé à partir des réponses aux questions précédentes. Nous n'allons pas dans cette étude concevoir de tests diagnostiques adaptatifs, mais nous allons expliquer rapidement leur fonctionnement parce que le test de compréhension de l'oral que nous utiliserons comme variable à expliquer est un test partiellement adaptatif.

Afin de comparer les résultats de candidats qui n'ont pas tous répondu aux mêmes questions, il faut que ces résultats soient placés sur une échelle de difficulté qui ne dépende pas des items effectivement passés par les candidats. C'est ce qui est fait dans les « modèles de réponse à l'item » (Laveault & Grégoire, 2014). Ces modèles supposent que la probabilité de répondre juste à un item dépend de la compétence du candidat (appelée « trait latent » parce que non observable directement) et augmente de façon non linéaire, suivant une courbe en « S » représentée ci-dessous (Figure 3.2). On peut remarquer que cette courbe se démarque d'une conception déterministe où un candidat n'aurait aucune chance de répondre correctement à un item en deçà d'un certain niveau, et 100% de chances au-delà. En effet, un candidat faible aura toujours une chance de trouver la bonne réponse en utilisant sa compétence même très

partielle, et un candidat fort aura toujours une chance de se tromper (mais de moins en moins à mesure que sa compétence augmente).

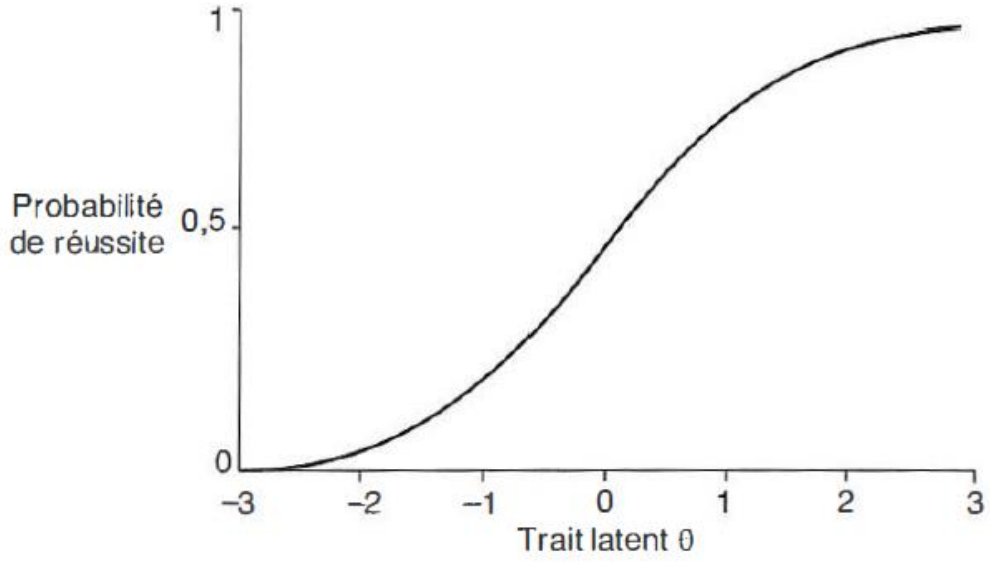


Figure 3.2 - Courbe caractéristique de l'item (Laveault & Grégoire 2014, p.281) montrant la probabilité de réussite à un item d'un test en fonction de la compétence (trait latent) d'un candidat

Par convention, la difficulté d'un item est la valeur du trait latent qui correspond à une probabilité de réussite de 0,5. Sur le graphique de la Figure 3.2, par exemple, un candidat avec un trait latent de 0 aura 50% de chances de réussir cet item, et « 0 » est également la mesure de la difficulté de l'item (par convention encore, la valeur de difficulté moyenne d'un groupe d'items est fixée à 0, ce qui donne un point de départ à l'échelle). Si la courbe était déplacée vers la droite, l'item serait plus difficile (le trait latent devrait avoir une valeur supérieure à 0), et le contraire si elle était déplacée vers la gauche. De cette façon, un lien est établi entre difficulté de l'item et compétence du candidat (trait latent), et chaque item passé donnera des renseignements supplémentaires sur cette compétence, sous forme d'estimation plus précise du trait latent. Dans un test adaptatif, on peut, après chaque item ou après passage d'un groupe d'items, utiliser cette évaluation du trait latent pour proposer de nouveaux items ou groupes d'items différenciés. C'est la méthode utilisée dans le test de positionnement SELF dont nous utiliserons le résultat en compréhension de l'oral, et qui sera décrit plus en détail dans le chapitre 9 de notre thèse.

3.4. Conclusion de la première partie et questions de recherche

Dans cette première partie, nous avons présenté notre cadre théorique, issu de trois courants habituellement séparés de la linguistique. Le premier chapitre, consacré à la modélisation de la compréhension de l'oral en L1, utilise essentiellement des travaux de psycholinguistique. Le deuxième chapitre sur les difficultés spécifiques aux auditeurs apprenants (d'anglais) L2 présente également des résultats issus de la psycholinguistique, mais s'inspire en outre de la recherche en didactique et acquisition des langues étrangères (une branche de la linguistique appliquée, Hilton, 2009). Enfin, c'est le champ théorique de l'évaluation en langues (qui fait également partie de la linguistique appliquée) qui faisait l'objet de ce troisième chapitre, qui présentait les outils que nous allons utiliser dans la partie expérimentale qui suit. Cette partie expérimentale (partie 2) sera consacrée à la construction et à la validation initiale de tests diagnostiques utilisés dans la suite de l'étude. Dans la troisième partie, nous présenterons et analyserons les résultats de l'étude corrélatoire qui étudiera le lien entre les tests diagnostiques élaborés dans la deuxième partie et la compréhension de l'oral. Nous ferons à partir de ces résultats des propositions de conception d'activités de remédiation.

La question centrale qui sert de fil conducteur à cette thèse est la suivante : comment aider les étudiants francophones à améliorer leur compréhension de l'oral en anglais langue étrangère ? Dans la première partie, nous avons montré que les connaissances phonologiques, prosodiques, lexicales, grammaticales et phraséologiques étaient toutes parties prenantes de l'ensemble complexe de processus qu'est la compréhension de l'oral. A partir de là, on peut se demander quel est l'apport relatif de ces connaissances au processus dans son ensemble. Nous avons vu que de nombreuses études existent sur la corrélation entre chacun de ces facteurs pris séparément, et la compréhension de l'oral, mais qu'aucune n'a encore étudié un ensemble aussi complet de facteurs contributeurs potentiels. D'autre part, aucune étude n'a à notre connaissance interrogé le lien entre les connaissances phraséologiques et la compréhension aurale.

Avant de pouvoir répondre à cette question centrale, il nous faut disposer d'outils diagnostiques permettant d'évaluer ces connaissances et d'identifier les lacunes éventuelles des sujets testés. Nous voudrions de plus que ces outils puissent être accessibles aux étudiants dans un cadre de formation en autonomie, et reliés à des activités de remédiation également accessibles à distance et en autonomie. Les questions qui instruiront la construction de nos tests diagnostiques seront donc les suivantes :

- Des tests de discrimination phonémique, de sensibilité prosodique centrée sur la perception de l'accent, de reconnaissance aurale du vocabulaire, de jugement de grammaticalité aurale et de connaissances phraséologiques produisent-il des scores fiables (du point de vue de la cohérence interne), et permettent-il de discriminer les apprenants de niveau faible de ceux de niveau élevé ?
- Ces tests sont-ils par ailleurs utilisables en pratique dans les conditions qui président à cette étude, c'est-à-dire en ligne, en autonomie, dans un temps total voisin d'une heure ?
- Les scores sont-ils interprétables au vu des résultats présentés dans la littérature sur l'acquisition ; en particulier, les mots fréquents sont-ils mieux reconnus ou analysés, les contrastes ou unités qui existent également dans la L1 sont-ils mieux réussis que les autres ?

Les réponses à ces questions détermineront la validation de nos instruments, et nous obligeront sans doute à modifier légèrement les versions initiales en enlevant les items qui dysfonctionnent (parce qu'ils sont trop difficiles, trop faciles, ou pas assez discriminants). Une fois les tests validés, nous les utiliserons dans la dernière partie pour une étude corrélatoire générale. Nous nous poserons alors la question de la relation entre les résultats à tous ces tests :

- Tous ces tests sont-ils corrélés entre eux ?
- Les tests sont-ils corrélés avec les résultats en compréhension de l'oral, ce qui est un premier pas vers la confirmation expérimentale de leur rôle dans cette compétence ?
- Quel est leur rôle respectif ? Les études que nous avons résumées trouvent en général que les connaissances lexicales sont les plus corrélées avec la compréhension de l'oral ; retrouvons-nous ce résultat ? Qu'en est-il des connaissances phraséologiques, qui n'ont pour l'instant pas été testées dans ce cadre ?

Chapitre 4

Plan de l'expérimentation

Dans ce chapitre, nous présenterons les démarches d'élaboration, de pilotage et de validation des différents outils que nous avons ensuite utilisés pour notre étude corrélatoire. Le Tableau 4.1 permet de visualiser le déroulement chronologique de nos expérimentations dont les groupes de sujets et les objectifs seront détaillés dans les paragraphes qui le suivent.

4.1. Déroulement chronologique

date	groupe	échantillon	tâches et objectifs/résultats
février-mars 2017	pilote	étudiants LLCER L1 N = 60	Pilotage de 2 instruments Phrasal Vocabulary Size Test (PVST) Test de sensibilité prosodique Résultats : Validation d'une version courte du PVST
automne 2017	1	étudiants LLCER L1 N = 150	Expérimentation à grande échelle Test de sensibilité prosodique Test de reconnaissance écrite du vocabulaire (LexTALE) Résultats : Validation d'une version du test de sensibilité prosodique Abandon de LexTALE
février-mars 2018	2	étudiants LLCER L1 N = 79	Pilotage de 2 instruments Test de reconnaissance aurale du vocabulaire Test de discrimination phonémique Résultats : Non utilisés (problème de récupération des données)
automne 2018	3	étudiants LLCER L1 N = 118	Expérimentation à grande échelle Test de reconnaissance aurale du vocabulaire Test de discrimination phonémique

		Etudiants L LANSAD N = 54	Test de jugement de grammaticalité aurale Phrasal Vocabulary Size Test (PVST) Test de sensibilité prosodique Résultats : Validation des tests de reconnaissance aurale du vocabulaire, de discrimination phonémique et de jugement de grammaticalité aurale Utilisation des résultats pour l'étude corrélatoire avec le test de positionnement SELF (compréhension de l'oral)

Tableau 4.1 - tableau récapitulatif des expérimentations mises en place

4.2. Les groupes de sujets

4.2.1. Groupe pilote (expérimentation février-mars 2017)

Nous avons réalisé une première étude exploratoire en février 2017 (au deuxième semestre de l'année universitaire 2016-2017) auprès d'étudiants inscrits en première année de licence LLCER (Langues, Littératures et Civilisations Etrangères et Régionales) parcours anglais à l'Université Grenoble Alpes (N= 61). Tous les étudiants inscrits en première année d'anglais (N=183) étaient au départ visés, mais seuls 61 d'entre eux ont effectué les deux tests pilotés, à savoir le test de connaissances phraséologiques (*Phrasal Vocabulary Size Test,* ou PVST) et le test de sensibilité prosodique. Ceci s'explique par l'abandon d'un certain nombre d'étudiants à l'issue du premier semestre d'étude, par le fait que d'autres rejoignent la filière en cours de semestre, par l'absence ponctuelle de certains lors de l'une ou plusieurs des séances pendant lesquelles se déroulaient les tests, et enfin par quelques problèmes techniques dans les salles équipées d'ordinateurs utilisées pour l'administration. Aucun questionnaire n'ayant été distribué à ces étudiants, nous n'avons pas d'informations biographiques les concernant. Les résultats assez élevés obtenus lors de cette étude pilote nous ont conduite à essayer d'avancer les tests au premier semestre, afin de toucher plus d'étudiants, avant qu'un nombre important d'entre eux (ceux qui auraient le plus besoin d'aide) ne décrochent. Nous ferons allusion aux résultats de ce pilotage quand ils nous ont été utiles pour améliorer les versions subséquentes de nos tests, en particulier pour le test de connaissances phraséologiques (*Phrasal Vocabulary Size Test*), qui sera décrit au chapitre 9. Ce test avait en effet déjà fait l'objet d'une étude de validation par son concepteur (Martinez, 2011), et nous nous sommes contentée de vérifier qu'il fonctionnait pour notre public. Pour le test de sensibilité prosodique que nous avons conçu, le pilotage nous a essentiellement servi pour des

raisons techniques et ergonomiques, afin d'améliorer la version qui a fait l'objet de l'expérimentation suivante.

4.2.2. Groupe expérimentation 1 (automne 2017)

En début de premier semestre universitaire 2017-2018, une autre expérimentation a donc été lancée, dans l'espoir d'obtenir un échantillon plus représentatif de la population qui s'inscrit habituellement en première année de licence LLCER anglais. Cette fois-ci, nous avons réussi à obtenir les données de 150 étudiants (dont 140 avaient également passé le test de positionnement en début d'année). L'expérimentation a eu lieu en octobre et novembre 2017, au bout de 3 à 6 semaines de cours universitaires. Les participants ont rempli un questionnaire leur demandant quelques renseignements personnels (âge, sexe) ainsi que des éléments de biographie langagière (langue maternelle, langues étrangères étudiées, séjours dans des pays anglophones). Ces informations sont présentées dans le tableau qui suit (Tableau 4.2).

âge		*âge LE*		*séjours anglo.*		*sexe*	
moyenne	18.8	moyenne	9.2	moyenne	67	F	75,3%
médiane	18	médiane	9	médiane	7	M	24,7%
mode	18	mode	11	mode	0		
écart-type	1.89	écart-type	2.56	écart-type	236.34		
kurtose	11.0	kurtose	2.98	kurtose	34.79		
coeff. d'asymétrie	2.96	coeff. d'asymétrie	0.60	coeff. d'asymétrie	5.72		
plage	12	plage	18	plage	1800		
minimum	17	minimum	2	minimum	0		
maximum	29	maximum	20	maximum	1800		
nb échantillons	148	nb d'échantillons	146	nb d'éch.	148		

Tableau 4.2 - statistiques descriptives des données biographiques des participants à l'expérience 1 (test de conscience accentuelle) : âge au moment de l'expérience, âge de début d'apprentissage de la première langue étrangère, nombre de jours en passés en pays anglophones, et sexe

Comme on peut le constater dans les première et dernière colonnes, les étudiants ont presque 19 ans en moyenne, et 75% sont des étudiantes. Cette surreprésentation féminine est traditionnelle dans les filières littéraires, et en particulier en langues (Ministère de l'Education Nationale, 2017a, p. 158). Tous ont le français comme langue maternelle sauf 8 : 3 sont de langue maternelle arabe (nous avons classé un locuteur de darja, l'arabe dialectal algérien, dans cette catégorie), 2 portugais, 1 lingala, 1 mongol, 1 roumain, 1 turc et 1 vietnamien. Parmi les francophones, 10 sont bilingues de naissance, avec comme autre langue maternelle le lingala (2), l'allemand (1), l'anglais (1), l'arabe (1), le créole (1), le russe (1), le shimaoré (1), le turc (1), le vietnamien (1). Pratiquement tous (95%) ont commencé par étudier l'anglais

comme première langue vivante étrangère (LVE), ce qui est un peu plus élevé que la moyenne nationale (92%, Ministère de l'Education Nationale, 2017a, p. 73), mais logique pour des étudiants qui ont choisi de se spécialiser en anglais. En moyenne, ils ont commencé à apprendre l'anglais à 9 ans, c'est-à-dire en avant-dernière année d'école primaire (CM1). Etant donné qu'ils ont 18-19 ans en 2017, ils étaient en CM1 en 2007 ou 2008. L'enseignement d'une langue vivante étrangère en primaire, qui existait déjà auparavant sous forme d'initiation dans des écoles volontaires, a été généralisé en France en 1998 sous le ministre Claude Allègre (Ministère de l'Education Nationale, 1998), et rendu obligatoire en 2001 sous le ministre Jack Lang (Ministère de l'Education Nationale, 2002), en commençant par le CM2 et en avançant ensuite progressivement la classe de début: CM1, puis CE2 en 2002 et CE1 en 2007 (Ministère de l'Education Nationale, 2007)[32]. Nos étudiants auraient donc dû commencer l'anglais en CE2, en 2006. Cette différence entre les données biographiques et les instructions ministérielles peut s'expliquer de deux façons. L'introduction des langues vivantes en primaire ne s'est pas faite sans mal, du fait du manque de formation et de confiance en soi des maîtres et maîtresses qui ne se sentaient pas toujours compétents pour enseigner cette nouvelle matière (Magnat, 2013, p. 326). Il est donc possible que de nombreuses écoles n'aient pas réussi à mettre en place cet enseignement cinq ans après la réforme de 2002. L'autre explication est que nos étudiants ne se souviennent pas précisément de l'âge auquel ils ont commencé l'anglais : ils savent qu'ils en faisaient en CM1, mais ne gardent pas de souvenir précis des classes précédentes (certains nous ont confié leurs doutes à ce sujet à la sortie de l'expérimentation).

On peut constater de grandes inégalités dans les séjours à l'étranger des étudiants (troisième colonne du Tableau 4.2) : la médiane du nombre de jours passés en pays anglophone est de 7, ce qui veut dire que la moitié de notre cohorte a passé moins d'une semaine en pays anglophone (de fait, un tiers n'y est jamais allé et le mode est à 0). D'un autre côté, un nombre non négligeable y a passé plusieurs mois, voire plusieurs années, ce qui est reflété par une moyenne beaucoup plus élevée que la médiane (67 jours).

Suite à ce constat, nous avons décidé de diviser nos étudiants en trois groupes de niveau à partir des résultats au test de positionnement de début d'année, obligatoire pour tous les étudiants qui s'inscrivent en première année pour la première fois. Le niveau exigé à l'entrée en licence LLCER, ou du moins considéré comme nécessaire afin de pouvoir suivre les cours

[32] L'enseignement d'une langue vivante en CP est obligatoire depuis 2016 (une initiation existait déjà auparavant)..

avec profit, est le niveau B2 (qui est également le niveau attendu en fin de lycée). Nous avons donc choisi d'identifier un groupe « faible », de niveau a priori insuffisant, B1 ou inférieur, un groupe « moyen », de niveau satisfaisant (B2), et un niveau « avancé », de niveau supérieur au minimum attendu (C1 ou plus). Le groupe d'avancés en compréhension de l'oral étant très restreint (15 étudiants seulement ont obtenu un résultat C1 en CO), nous y avons ajouté les étudiants ayant obtenu un résultat B2 en CO mais avec une proposition de positionnement global « en route vers C1/C2 », ce qui donne des groupes plus équilibrés (32 « avancés », 61 « moyens » et 58 « faibles »). Le tableau qui suit (Tableau 4.3) présente les caractéristiques biographiques de chacun de ces groupes. Les groupes « moyens » et « avancés » ne semblent pas présenter de caractéristiques biographiques distinctes, mais les étudiants du groupe « faible » sont un peu plus âgés en moyenne, ont commencé l'anglais un peu plus tard, et ont moins séjourné en pays anglophone. On pourrait dire que les groupes B2 et C1 se rapprochent plus de l'étudiant idéal qui suit le parcours prévu dans les programmes et les diagrammes (Figure 4.1) du ministère de l'éducation nationale français: début de l'anglais en CE2 (d'après les programmes de l'époque) et passage direct (sans redoublement et sans réorientation) du lycée au supérieur, à 18 ans.

	âge		*début LE*		*Séjours (j)*	
	M (sd)	étendue	M (sd)	étendue	M (sd)	étendue
"faibles" (n=51)	19.1 (2.4)	17-29	9.8 (2.7)	5-18	50 (202)	0-1400
"moyens" (n=61)	18.4 (1.4)	17-26	9.1 (2.5)	4-20	79 (298)	0-1800
"avancés" (n=31)	18.5 (1.2)	17-23	8.8 (2.6)	2-12	78 (160)	0-750

Tableau 4.3 - statistiques descriptives de trois variables biographiques (âge, début d'apprentissage de l'anglais, et nombre de jours passés en pays anglophone) en fonction du groupe de compétence des participants à l'expérience 1

Ceci est particulièrement vrai des étudiants avancés dont les résultats présentent moins de variation que ceux des autres groupes (les écarts-types y sont plus réduits). On peut remarquer également que, si les étudiants du groupe faible ont passé environ deux fois moins de temps en moyenne en pays anglophone que ceux des autres groupes, on trouve cependant dans tous les groupes des étudiants qui ne sont jamais allés à l'étranger ainsi que d'autres qui y ont passé plusieurs années. Dans le groupe avancé, la durée maximale passée à l'étranger (2 ans) est plus faible que dans les autres groupes (où elle est de 4 ou 5 ans), ce qui peut être dû au fait que pour réussir dans le système scolaire et universitaire français, il vaut mieux avoir une bonne connaissance de ce système et ne pas avoir passé trop de temps à l'étranger (ce qui peut

accréditer l'idée que le test de positionnement mesure la compétence scolaire en même temps que le niveau de langue).

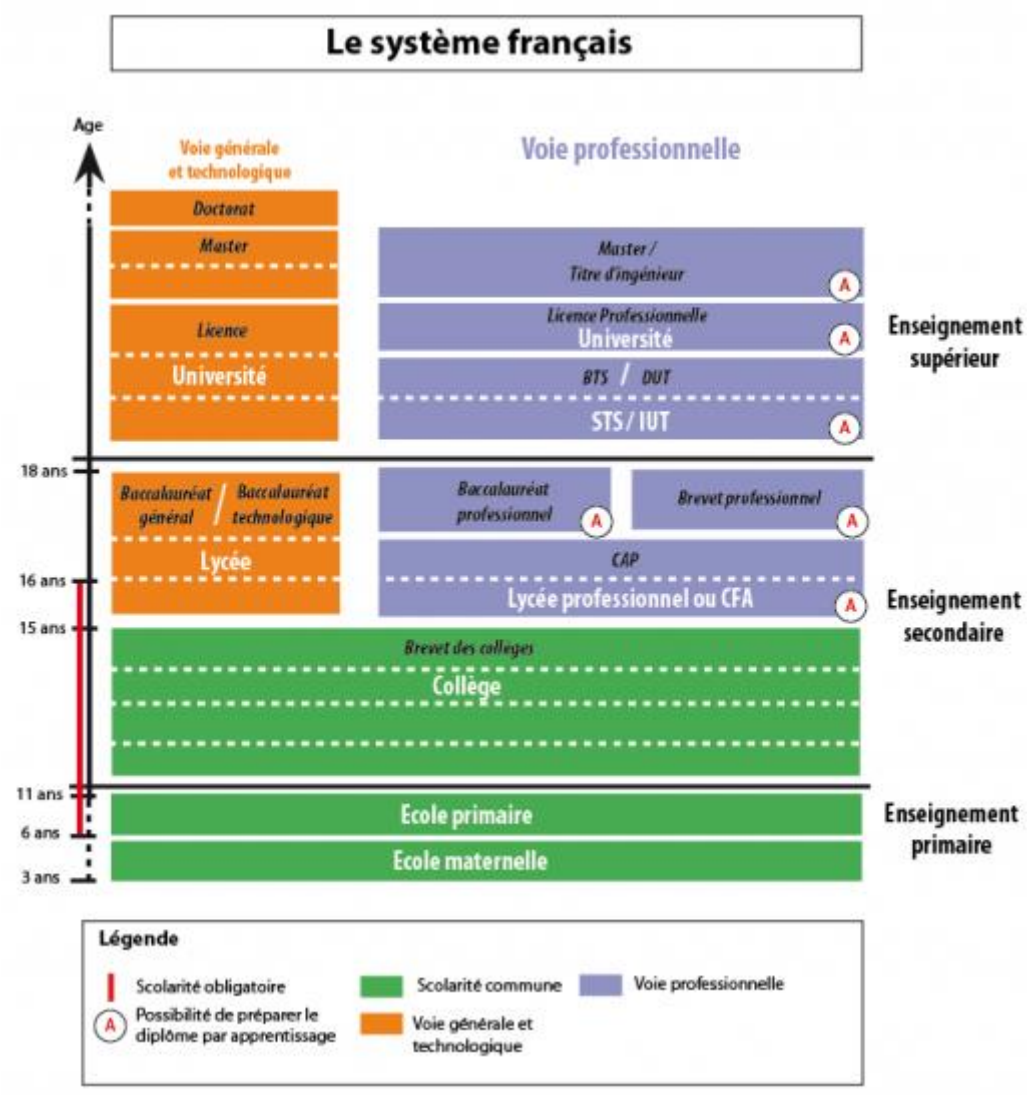


Figure 4.1 - système éducatif français: parcours 'classique' (site https://peda.net/)

Ces étudiants ont passé le test de sensibilité prosodique, le PVST et le test LexTALE (Lemhöfer & Broersma, 2012). Cette étude nous a permis de proposer une version fiable du test de sensibilité prosodique (dont la validation sera décrite dans le chapitre qui suit), et de constater que LexTALE ne convenait pas à nos besoins : il s'agit d'un test de reconnaissance écrite du vocabulaire, qui discriminait très peu entre les étudiants de différents niveaux. C'est pourquoi nous avons décidé d'en concevoir une version aurale et de la piloter lors de l'expérimentation suivante.

4.2.3. Groupe 2 (expérimentation pilote février-mars 2018)

79 étudiants ont participé à l'étude pilote du test de reconnaissance aurale du vocabulaire et de celui de discrimination phonémique au deuxième semestre 2017-2018. Un problème technique de serveur ne nous a permis de récupérer que des résultats partiels, encourageants mais incomplets, que nous ne présenterons donc pas. Ces deux tests, ainsi que celui de jugement de grammaticalité aurale, ont fait l'objet d'une dernière étude de validation à la rentrée 2018.

4.2.4. Groupe 3 (expérimentation octobre 2018)

En début de premier semestre universitaire 2018-2019, une dernière expérimentation a été lancée, en utilisant la version des tests diagnostiques validée précédemment par les autres groupes (pour le PVST et le test de sensibilité prosodique), ou en la validant concurremment (pour le test de reconnaissance du vocabulaire, de discrimination phonémique et de grammaticalité), afin de pouvoir faire une étude corrélatoire complète entre tous les tests diagnostiques et le résultat en compréhension de l'oral au test de positionnement de début d'année. Nous espérions toucher environ 150 étudiants inscrits en première année LLCER anglais comme à l'automne 2017, mais seuls 118 étudiants ont passé les tests. En effet, une réorganisation partielle de la maquette de première année a eu lieu entretemps et les étudiants inscrits en licence bilangue, c'est-à-dire suivants les cours de LLCER anglais et en même temps ceux de LLCER d'une autre langue (allemand, espagnol, italien ou russe), n'étaient plus censés suivre le cours dont les créneaux étaient prévus pour le passage des tests.

Nous avons donc cherché à élargir notre échantillon en faisant appel aux étudiants inscrits en cours d'anglais LANSAD (LANgues pour les Spécialistes d'Autres Disciplines), ce qui nous a permis également de toucher plus d'étudiants de niveau A1 ou A2 en compréhension de l'oral, afin que les niveaux faibles soient mieux représentés. Nous avons lancé un appel par l'intermédiaire des responsables d'anglais LANSAD au Service des Langues, et cinq collègues ont proposé de faire passer les tests à leurs étudiants, totalisant six groupes et 76 étudiants. Parmi ces 76 étudiants, 54 ont répondu au questionnaire langagier, et 41 ont passé tous les tests (y compris le positionnement de début d'année). Ils sont inscrits aussi bien en sciences (chimie, informatique, mathématiques, physique, sciences de l'ingénieur) qu'en sciences humaines (arts du spectacle, histoire de l'art, langues étrangères, lettres, linguistique, philosophie) ou sciences sociales (économie, gestion, psychologie). Ils peuvent difficilement, vu leur faible nombre, être représentatifs des milliers d'étudiants qui prennent des cours

d'anglais au Service des Langues chaque année, mais ils couvrent tout de même une grande partie des champs disciplinaires de licence proposés sur le campus grenoblois.

Nous présenterons rapidement les caractéristiques des étudiants LLCER, qui sont similaires à celles du groupe d'automne 2017, et les comparerons brièvement à celles des étudiants LANSAD (Tableau 4.4). Nous constatons que la cohorte LLCER 2018 a le même âge (19 ans en moyenne) que celle de la rentrée 2017. Il s'agit encore essentiellement d'étudiants qui viennent de passer le baccalauréat (la médiane et le mode sont ici aussi à 18 ans), avec quelques étudiants plus âgés, en redoublement, réorientation ou reprise d'études. Ils disent eux aussi avoir commencé à étudier l'anglais vers 9 ans. Nous pensons encore une fois que cet âge ne correspond pas à la réalité, dans la mesure où beaucoup considèrent (communication personnelle) que les « vrais » cours d'anglais n'ont commencé qu'au collège, même quand ils ont bénéficié d'une initiation à l'école primaire dont ils ne se souviennent pas forcément. Enfin, la moyenne du nombre de jours passés à l'étranger est un peu plus haute qu'en 2017 (88 jours au lieu de 67, soit 3 mois environ au lieu de 2), mais le mode reste à 0, et la médiane n'est pas beaucoup plus élevée (9 au lieu de 7 jours). La moyenne plus élevée provient donc essentiellement de quelques individus avec des valeurs extrêmes, ayant passé beaucoup plus de temps en pays anglophone (le maximum est de 8 ans au lieu de 4 en 2017).

	âge		*début LE*		*séjours (j)*		*sexe*
	M (sd)	étendue	M (sd)	étendue	M (sd)	étendue	
LLCER (n=118)	19.03 (2.04)	17-33	9.47 (1.84) (n=116)	6-15	88 (340) (n=116)	0-2800	75% F
LANSAD (n=54)	19.9 (2.78)	17-31	9.93 (2.18) (n=54)	6-16	10 (30) (n=54)	0-210	76% F

Tableau 4.4 - statistiques descriptives de 3 variables biographiques (âge, début d'apprentissage de l'anglais, et nombre de jours passés en pays anglophone) en fonction de l'origine académique des participants à l'expérience 2

Les étudiants LANSAD, qui suivent des cours d'anglais pour non-spécialistes de première ou deuxième année de licence, n'ont pas un profil très différent des étudiants LLCER de première année, sauf en ce qui concerne les séjours en pays anglophones. Ils ont un an de plus (ce qui est normal dans la mesure où certains sont en deuxième année) et ils ont commencé l'anglais à peu près au même âge (à 10 ans au lieu de 9 ans et demi), mais aucun d'entre eux n'a passé plus de 7 mois en pays anglophone (la moyenne est de 10 jours au lieu de 88). La proportion de filles est presque identique à celle en LLCER, ce qui peut paraître étonnant dans la mesure où la filière langues étrangères est l'une des plus fortement féminisées en France. Un échantillon plus varié d'étudiants, dont certains inscrits en filière scientifique où les filles

sont minoritaires, devrait faire baisser cette proportion (57% des inscrits en licence toutes filières confondues sont des filles, Ministère de l'Education Nationale, 2017a, p. 177). Nous n'avons pas d'explication à ce phénomène, qui peut être dû à la petite taille de notre échantillon, mais aussi au fait que la langue étrangère peut être optionnelle dans certains cursus, et choisie moins fréquemment par des garçons.

4.3. Développement des tests diagnostiques et analyses statistiques

Dans les chapitres 5 à 8, nous allons décrire la conception des tests diagnostiques de sensibilité prosodique, de discrimination phonémique, de reconnaissance du lexique aural et de jugement de grammaticalité aurale que nous avons développés pour ce travail. Nous nous inspirerons de tests et d'items typiquement utilisés dans la littérature psycholinguistique, dans la mesure où (à part pour les connaissances lexicales) nous n'avons pas trouvé dans les études de didactique ou d'évaluation en langues de tests sur les sous-habiletés qui nous intéressent. Cependant, nous aimerions souligner ici qu'il s'agit bien de tests diagnostiques et non de tests utilisés pour une étude véritablement psycholinguistique. Nous visons des tests courts (et donc pas nécessairement exhaustifs) qui serviront à étudier tous les aspects du traitement de l'oral. Nous aurons donc des nombres d'items bien inférieurs à ceux typiquement observés dans les études psycholinguistiques, et les différentes variables en jeu ne seront pas nécessairement réparties de façon équilibrée entre les différents items.

Les tests proposés doivent pouvoir se faire en ligne, et surtout sans supervision. Nous excluons donc tout test où le temps de réponse est crucial : les étudiants ne disposent pas toujours d'une connexion optimale, ce qui peut ralentir l'enregistrement de leurs réponses, et peuvent être distraits brièvement par l'environnement dans lequel ils passent les tests, même s'il leur est demandé de prévoir une plage horaire où ils sont sûrs de pouvoir travailler en ligne sans être dérangés. Enfin, le fait que les tests ne soient pas associés à des enjeux importants (*low-stakes*), s'il diminue le risque de fraude, peu rationnelle dans ce contexte, augmente par contre le risque que les étudiants ne donnent pas le maximum d'eux-mêmes et ne fassent pas preuve d'une réactivité maximale[33].

Les tests seront présentés dans l'ordre chronologique de leur développement : d'abord le test de sensibilité prosodique, puis le test de discrimination phonémique, et enfin les tests de

[33] Cependant, le temps de réponse est tout de même enregistré automatiquement et nous verrons que l'étude de sa corrélation avec les résultats aux tests peut fournir des informations intéressantes.

reconnaissance aurale du vocabulaire et de jugement de grammaticalité aurale. Nous terminerons par une brève description des autres tests utilisés dans cette étude, que nous n'avons pas développés personnellement (PVST) ou qui n'ont pas été développés dans le cadre de cette thèse (SELF). Nous analyserons les tests diagnostiques développés spécifiquement pour cette thèse en suivant les lignes directrices présentées dans le chapitre 3, et résumées dans le tableau ci-dessous (Tableau 4.5).

type d'analyse	but de l'analyse	critère choisi	valeurs visées
analyses qualitatives	validité de construit	couverture du construit (choix items)	(conception du test avant pilotage)
	validité externe	(concurrente ou prédicitve)	non analysée
	praticité	temps moyen temps maximum	10 minutes en moyenne
analyses quantitatives	statistiques descriptives du score	- tendance centrale : score moyen - dispersion : écart-type et étendue - asymétrie	 asymétrie négative
	fidélité	alpha de Cronbach	$\alpha > 0,7$
	fonctionnement des items	- coeff. difficulté - discimination	entre 0,2 et 0,95 > 0,2 (acceptable) > 0,3 (bon)
	unidimensionnalité	ACP	tous items corrélés positivement à composante principale

Tableau 4.5 - Résumé des analyses qualitatives et quantitatives mises en œuvre pour la validation des tests diagnostiques

Lors de la description de la construction de nos tests diagnostiques, nous justifierons soigneusement le choix des items pour montrer qu'ils couvrent bien le construit de la sous-habileté en question. Après le pilotage, nous vérifierons que le temps moyen de passation est raisonnable (autour de dix minutes par test, pour un temps total moyen d'une heure environ) - cet élément fait partie de la praticité (Bachman & Palmer, 1996). Ensuite, nous analyserons quantitativement les scores, pour vérifier qu'ils sont suffisamment dispersés (écart-type et étendue), et, s'ils sont asymétriques, que cette asymétrie est plutôt négative (avec plus de détails dans les scores en dessous de la moyenne de l'échantillon). Nous vérifierons également que le test est fiable, avec un coefficient α est supérieur à 0,7 (Laveault, 2012), et que les items présentent une difficulté et une discrimination acceptables. La discrimination est estimée à l'aide du coefficient de corrélation bisérial de point, qui calcule la corrélation entre la réussite à l'item et la réussite au test. Nous utiliserons pour cela, dans *R*, la fonction

score.multiple.choice[34] de la « bibliothèque » (« *package* ») *psych* (Revelle, 2017). Enfin, une fois les items assemblés pour la version finale du test, nous vérifierons avec une ACP (analyse en composantes principales) qu'il existe une dimension principale à laquelle tous les items sont corrélés positivement (Husson et al., 2016), l'unidimensionnalité étant une des composantes de la validité d'un test.

Nous n'avons pas analysé la validité de critère externe de nos tests, qui nous semblent trop pointus pour être validés d'un point de vue prédictif. Même si nous avions trouvé une corrélation entre le score de discrimination phonémique, par exemple, et la moyenne obtenue par les étudiants à un module d'anglais ou à leur première année d'étude, il est probable que cette corrélation aurait émergé par le truchement d'une compétence plus globale et plus directement liée à la réussite, comme la compréhension de l'oral, ou le niveau d'anglais général. Une comparaison avec les résultats à des tests déjà validés (validité externe concurrente) n'aurait été possible que pour le test de connaissances lexicales, où il existe plusieurs tests reconnus internationalement. Cependant, étant donné le temps de contact réduit dont nous disposions avec les groupes d'étudiants qui ont participé à nos expérimentations, nous n'avons pas pu leur faire passer d'autres tests que ceux développés pour cette étude. Au final, c'est l'étude corrélatoire que nous présenterons dans la dernière partie qui s'apparentera le plus à une recherche de validité de critère externe. Nous y analyserons en effet les corrélations de nos tests diagnostiques avec un test de compréhension de l'oral développé indépendamment.

Pour la plupart de nos tests, nous analyserons plus en détail les résultats pour vérifier qu'ils correspondent aux attentes issues des études psycholinguistiques (on s'attend par exemple à ce que les items plus fréquents soient mieux connus). Pour ces analyses, nous serons amenée à utiliser des tests statistiques. Nous résumons ici les tests utilisés ainsi que leurs conditions d'utilisation, afin de pouvoir nous y référer facilement dans la suite de cette thèse (Tableau 4.6). Beaucoup de ces tests statistiques (dits « paramétriques ») supposent que les données sont issues d'une distribution normale, ce que nous vérifierons à la fois par une inspection visuelle du graphique des résultats (histogramme de fréquence des scores, par exemple), et par le test de Shapiro-Wilk (A. Field et al., 2012, p. 182). Quand ces conditions ne sont pas respectées, nous utiliserons des tests non paramétriques, moins sensibles mais plus robustes, parce qu'utilisant les rangs des valeurs des variables et non les valeurs elles-mêmes.

[34] La seule exception sera le premier test développé (sensibilité prosodique), pour lequel c'est la fonction sjt.itemanalysis de la bibliothèque « sjPlot » qui a été utilisée.

Nous réaliserons en particulier des comparaisons de moyennes (par exemple, comparaison de la réussite à plusieurs groupes d'items différents, ou comparaison de la fréquence moyenne de deux groupes de mots). Pour comparer deux moyennes (c'est-à-dire pour vérifier que les deux échantillons auxquels elles sont associées sont probablement issus de deux populations différentes, ou au contraire de la même population), nous utiliserons le test t (souvent appelé « test de Student » en France, « *Student* » étant le pseudonyme utilisé par son auteur lors de sa première publication). Quand ses conditions d'utilisation ne seront pas respectées, nous nous tournerons vers le test de Wilcoxon (un équivalent du test Mann-Whitney, et qui est utilisé dans *R* à la place de ce dernier, A. Field et al., 2012, p.373). Quand nous aurons à comparer plus de deux moyennes, nous utiliserons une analyse de variance (ANOVA). Les résultats d'une ANOVA indiquent simplement si les moyennes sont significativement différentes ou non, mais pas si la différence entre chaque paire de moyennes est également significative. C'est pourquoi nous compléterons cette analyse par un test post-hoc des étendues de Tukey avec correction pour comparaisons multiples. Quand les conditions d'utilisation d'une ANOVA ne seront pas respectées, nous utiliserons le test de Kruskal-Wallis, complété ensuite par un test de Wilcoxon deux à deux (avec une correction Bonferroni pour le fait qu'on opère de multiples tests sur les mêmes données). Comme il est d'usage en sciences humaines (A. Field et al., 2012, p. 51-52), nous considérerons que les résultats des tests statistiques inférentiels sont significatifs dès que $p<0,05$.

test	objectif	nom	conditions
distribution normale	vérifier conditions de validité des tests paramétriques	Shapiro-Wilk	
égalité de moyennes	comparer 2 moyennes	t-test (« Student »)	- nombre d'observations > 30 ou distribution normale - variances égales (Welch si variances non égales)
	comparer 2 moyennes	Wilcoxon (Mann-Whitney)	aucune (sur rangs)
	comparer plus de 2 moyennes	ANOVA, suivi de test de Tukey 2 à 2	- normalité des résidus - (variances égales)
	comparer plus de 2 moyennes	Kruskal-Wallis, suivi de Wilcoxon 2 à 2	aucune (sur rangs)

Tableau 4.6 - résumé des principaux tests statistiques utilisés dans la thèse (d'après informations de Field et al., 2012)

Chapitre 5

Test de sensibilité prosodique

5.1. Rappels du cadre théorique

Nous avons vu dans notre première partie que la prosodie, que nous avons circonscrite à l'accentuation pour les besoins de cette étude, pouvait être utilisée à deux étapes différentes du processus de compréhension de l'oral en anglais : la segmentation d'une part, et l'accès lexical de l'autre. Nous avons trouvé une seule étude sur chaque sujet pour des apprenants L2 : une thèse de 2016 (Tabata, 2016), qui semble trouver un lien entre compréhension de l'oral d'une part, et sensibilité à l'accent contrastif et aux indices de segmentation d'autre part chez les adolescents japonais ; et un article (Meerman et al., 2014), qui ne trouve pas de corrélation entre la sensibilité à l'accent lexical et la compréhension de l'oral. Ces deux études portent sur des apprenants japonophones et demandent donc à être confirmées par d'autres études avec des francophones, ce que nous allons tenter de faire ici.

Par ailleurs, nous avons rappelé dans la première partie les principes fondateurs d'une approche diagnostique en enseignement : à un problème diagnostiqué doit pouvoir être mis en regard une proposition de remédiation, faute de quoi l'apprenant risque d'être découragé ou bloqué dans son parcours d'apprentissage. Nous avons donc décidé de concentrer notre démarche sur l'accent lexical, pour lequel il existe plus de matériel de remédiation et beaucoup d'activités en autonomie.

5.2. Inventaire d'instruments d'évaluation

5.2.1. Test de « parole réitérée » (*reiterant speech*)

Parmi les tests utilisés dans les études existantes pour mesurer la sensibilité à la prosodie, le test de parole réitérée est peut-être l'un des plus anciens. Il consiste à utiliser des items dans lesquels toutes les syllabes ont été remplacées par une seule et même syllabe, et à ne garder

ainsi que les indices prosodiques puisque les indices segmentaux normaux ont disparu (Nakatani & Schaffer, 1978). Ces items sont enregistrés par un locuteur natif entraîné, qui réalise un énoncé cible tout en n'articulant qu'une seule syllabe (*ma*, par exemple), ou même un seul son (*mm*, ce qui donne de la « parole chantonnée », *hummed speech*). Au lieu de parole naturelle, on peut aussi utiliser des syllabes synthétisées, ou une combinaison des deux techniques. Ce type de test a été utilisé pour étudier deux phénomènes différents. D'une part, la segmentation : Nakatani et ses collègues par exemple demandent à leurs sujets de situer les frontières de mots dans ce flux chantonné, mais cette tâche apparaît difficile pour des sujets natifs, qui ont un taux de réussite variant (selon les items) entre 60% et 85% (avec seulement deux items au dessus de 75%). D'autre part, le test de parole réitérée a été utilisé pour étudier l'accès lexical, via la reconnaissance du patron accentuel : Kitzen (2001), puis Whalley et Hansen (2006) et Goswami et al. (2013) demandent à leurs sujets (enfants et adultes, travaillant en langue maternelle) à quels noms de personnages ou de titres de films les suites de syllabes qu'ils entendent correspondent (par exemple, « dee DEE dee DEE » correspond à *The Lion King* et non pas à *Casablanca*). Les taux de réussite sont au plafond pour les adultes non dyslexiques, mais le test exige que les sujets connaissent les expressions correspondant aux patrons accentuels. Les mots à reconnaître sont présentés au préalable à l'écrit chez Kitzen, et représentés par des images dans l'expérience de Goswami et ses collègues qui ont vérifié avant le test que les enfants connaissaient les représentations iconiques utilisées (écartant les items correspondants si ce n'est pas le cas).

Nous n'utiliserons pas la technique de la parole réitérée ou chantonnée pour étudier la segmentation, puisque même les natifs trouvent la tâche difficile. Pour ce qui est de l'accès lexical via la reconnaissance du patron accentuel, nous ne pouvons pas sélectionner au préalable le lexique déjà connu des étudiants et devrons donc renoncer à utiliser cette tâche également (ce serait cependant une tâche séduisante à utiliser pour tester les acquisitions des apprenants après un enseignement de prosodie).

5.2.2. Test avec phrases traitées avec un filtre « passe-bas » (*low-pass filter*)

Une autre technique utilisée dans la recherche psycholinguistique pour mesurer les traitements prosodiques sans l'interférence des informations segmentales est la filtration du signal acoustique écouté par les sujets. On élimine tout ce qui se trouve au-delà d'une certaine fréquence, pour que les formants des segments, leurs transitions, etc. disparaissent et qu'il ne reste que la fréquence fondamentale (qui indique la hauteur de la voix) et les informations

rythmiques. Wood et Terrell (1998) ont utilisé cette tâche avec des enfants dyslexiques ou non en L1, qui devaient reconnaître entre deux énoncés normaux (présentés à l'oral) celui qui correspondait à une phrase filtrée donnée.

Comme le précédent, ce paradigme expérimental paraît intéressant pour un test diagnostique, de par sa focalisation exclusive sur le rythme de la phrase et la simplicité de la tâche. Cependant, aucun exemple d'item n'est donné dans l'article, ce qui n'aide pas à la conception d'un test similaire. Enfin, les enfants ont des résultats assez faibles (pratiquement au niveau du hasard pour les dyslexiques, et avec un taux de réussite de l'ordre de 65% pour les lecteurs normaux), ce qui indique que la tâche risque d'être trop difficile pour des apprenants L2.

5.2.3. Tâche psycholinguistique d'amorçage (ou d'identification) intermodal

Le paradigme d'amorçage joue un rôle très important dans la recherche psycholinguistique depuis sa mise en lumière par Meyer et Schvaneveldt (1971) il y a plusieurs décennies[35]. Le phénomène est exploité lors des tâches de décision lexicale (LDT ou *Lexical Decision Task*), pour sonder les effets de différents types d'information (par exemple sémantique, ou ici prosodique) sur la vitesse de reconnaissance des mots. Lors d'une tâche de décision lexicale, on présente aux sujets une suite de lettres ou de sons, et on leur demande d'indiquer le plus rapidement possible si l'ensemble constitue un mot (dans une langue donnée). Dans le paradigme d'amorçage, on présente avant le mot à reconnaître une autre information (sous forme de mot, d'image, de phrase) et on en observe l'effet sur le temps de réaction mesuré très précisément. L'étude initiale de Meyer et Schvaneveldt avait montré que la présentation d'un premier mot relié sémantiquement à un deuxième facilite (accélère) la reconnaissance de ce deuxième mot.

Cooper et ses collègues utilisent un paradigme expérimental d'amorçage intermodal qui implique deux modes de présentation : oral et écrit (N. Cooper et al., 2002). Dans ce paradigme, les sujets entendent une suite de phonèmes correspondant à un début de mot, et doivent ensuite effectuer le plus rapidement possible une tâche de décision lexicale sur un mot écrit qui apparait à l'écran (« ce mot existe-t-il ou pas ? »). Les auteurs montrent que si la portion de mot utilisée pour l'amorçage (à l'oral) a les mêmes segments et le même patron accentuel que le mot à reconnaitre (à l'écrit), le temps de réponse est plus rapide qu'avec les

[35] Nous y avons d'ailleurs déjà fait allusion dans le premier chapitre, à propos de la réanalyse de Fine et al. (2013) du phénomène d'amorçage en termes de prédiction – voir section 1.4.4.4.

mêmes segments mais un patron accentuel différent (MUSic est reconnu plus vite après avoir entendu la syllabe accentuée MUS- que ne l'est muSEum, et vice-versa après la syllabe non accentuée mus-). Le contour prosodique est donc utilisé pour l'accès lexical.

La tâche utilisée par Tremblay (2008) est également intermodale, mais il s'agit d'une tâche d'identification et non de décision lexicale. Les sujets entendent une phrase qui se termine par la première syllabe accentuée ou non d'un mot plurisyllabique. A l'écran apparaissent ensuite deux mots qui sont tous les deux compatibles avec l'amorce sur le plan segmental, mais dont un seul correspond au schéma accentuel, et les sujets doivent décider le plus vite possible lequel des deux mots correspond à l'amorce entendue.

Nous avons décidé d'écarter ces deux tâches de nos tests diagnostiques pour deux raisons. Premièrement, dans le contexte d'utilisation de nos tests, nous ne pourrons pas prendre en compte le temps de réaction, une information essentielle pour l'interprétation des résultats. Deuxièmement, dans les deux expériences, les anglophones n'ont pas de très bons résultats à cette tâche en langue maternelle (70% de bonnes réponses chez Tremblay par exemple) : il semble que les syllabes semblables sur le plan segmental activent les mots correspondants même si le patron accentuel est différent. Cela est certainement dû au fait que l'information accentuelle est en général confondue avec l'information segmentale en anglais, puisque les syllabes accentuées sont pleines et les non accentuées sont le plus souvent réduites (Cutler et al., 1997).

5.2.4. Tests de jugement de syllabe accentuée de (mots ou) pseudomots

Altmann (2006) utilise une technique plus simple pour étudier l'acquisition de l'accent lexical chez des apprenants étrangers : afin de neutraliser les connaissances lexicales, elle a créé des pseudomots, que les sujets entendent et voient écrits en même temps, et dont ils doivent entourer la syllabe accentuée. Ce test peut présenter l'inconvénient de confondre les indices segmentaux et accentuels, mais Altmann fait attention d'inclure des syllabes réduites ou non dans les syllabes non accentuées (et effectivement, les résultats sont meilleurs quand les syllabes inaccentuées sont réduites, surtout en dernière position). D'autre part, cette collusion de types d'indices, segmentaux et accentuels, est conforme à ce qui se passe en conditions réelles. Davis et Kelly (1997) utilisent une variante de ce test où les sujets, au lieu d'entourer la syllabe accentuée du mot sur sa représentation orthographique, doivent simplement

indiquer le numéro de la syllabe accentuée (avec des mots ou pseudomots bisyllabiques, donc 1 ou 2).

Ce paradigme est bien adapté à notre situation, de par sa simplicité (les consignes sont facilement compréhensibles), et le degré de difficulté adéquat. Les natifs anglophones ont en effet des résultats presque au plafond dans les deux expériences (plus de 90% de réussite), alors que les non-natifs francophones ont beaucoup de difficultés dans l'expérience d'Altmann. Nous pouvons donc supposer que ces items seront discriminants pour notre population.

5.2.5. Tâche de discrimination AX

Les tâches de type AX, où les sujets entendent un stimulus « X » et doivent dire s'il est identique (ou similaire) à un autre stimulus « A », sont souvent utilisées en psycholinguistique (en général en conjonction avec le temps de réponse, qui est une donnée importante dans la plupart des études). C'est le type de tâche utilisé par Dupoux et son équipe pour montrer que les francophones perçoivent l'accent au niveau acoustique : ils sont capables de différencier deux pseudomots ne se distinguant que par l'accent quand ils sont prononcés par le même locuteur (Dupoux et al., 1997).

Leong et ses collègues utilisent le même genre de tâche avec des adultes dyslexiques (Leong et al., 2011). Au lieu d'utiliser des pseudomots, ils tentent de neutraliser en partie les connaissances lexicales en utilisant des vrais mots, mais qui peuvent être bien ou mal accentués (créant par là des non-mots). Leurs sujets entendent soit le même mot accentué deux fois de façon identique, soit le mot accentué de deux façons différentes. Même s'ils trouvent une différence significative entre le taux de bonnes réponses chez les dyslexiques et les lecteurs « normaux », les réponses sont presque au plafond (respectivement 95% et 99%). C'est pourquoi ils imaginent une adaptation de cette tâche où les exemplaires à discriminer sont des mots différents (toujours bien ou mal accentués), et il faut décider s'ils ont le même patron accentuel ou pas. Cette variante est beaucoup plus difficile que l'autre, et produit des résultats beaucoup plus discriminants (respectivement moins de 60% et plus de 85% de bonnes réponses).

Nous pouvons imaginer réutiliser cette tâche, bien qu'elle soit légèrement difficile pour les natifs également. Comme la tâche n'est pas habituelle (comparer le patron accentuel de deux

mots différents, qu'ils soient bien ou mal accentués), il faudra porter une attention particulière à la consigne pour s'assurer de sa bonne compréhension par nos apprenants.

5.2.6. Tests de jugement de mots bien ou mal accentués et test de compréhension à choix forcé (QCM)

Meerman et al. (2014) utilisent un test assez simple, qui consiste à faire entendre aux apprenants des mots bien ou mal accentués, parmi lesquels ils doivent identifier les mots mal accentués. Même si les résultats sont croisés ensuite avec ceux d'un test de vocabulaire sur des mots différents, ce test a l'inconvénient de ne pas vraiment contrôler les connaissances lexicales. Si un apprenant ne connaît pas bien le mot (s'il ne l'a jamais ou pas souvent entendu à l'oral, par exemple), il ne pourra pas répondre à la question posée alors qu'il est peut-être sensible au schéma accentuel. S'il a souvent entendu le mot prononcé de façon incorrecte par ses pairs, il n'identifiera pas non plus le mot comme mal prononcé. Nous ne pourrons donc pas utiliser ce test en l'état (en tout état de cause, les auteurs ne trouvent pas de corrélation avec les connaissances lexicales et en concluent que les formes lexicales ne sont pas mémorisées avec l'accent chez les nipponophones apprenant l'anglais).

Pour évaluer la compréhension de l'accentuation en contexte, soit pour segmenter des groupes de mots, soit pour comprendre les implications d'un accent contrastif, Tabata (2016) utilise des QCM où le stimulus oral est associé à un choix entre deux images. Par exemple, pour le syntagme *chocolate ice cream and cookies* (qu'il ne faut pas confondre avec *chocolate, ice cream, and cookies*) il faut choisir entre une image représentant deux éléments (de la glace au chocolat et des biscuits) et une autre avec trois éléments (du chocolat, de la glace et des biscuits). Ces tests ont été développés à Queen Margaret University College à Edimbourg par Sue Peppé (Peppé & McCann, 2003) et fait actuellement l'objet d'une commercialisation comme outil de diagnostic des troubles langagiers associés à l'autisme. Les items visant la segmentation à l'intérieur d'un syntagme ne sont pas particulièrement bien réussis (61%) par les adolescents nopponophones de l'étude de Tabata (2016, p. 120) et donc assez discriminants. On peut donc imaginer les réutiliser pour notre test.

5.2.7. Tableau récapitulatif

Dans les six sous-sections qui précèdent, nous avons fait une revue des techniques utilisées pour étudier la perception et l'acquisition de l'accentuation (au niveau des mots et des phrases). Cette revue n'est pas exhaustive car il y manque en particulier des tâches nécessitant

du matériel sophistiqué dont nous ne disposons pas (oculométrie ou potentiels évoqués par exemple). Le Tableau 5.1 résume les caractéristiques des paradigmes présentés ci-dessus, avec, en dernière colonne, un rappel de notre jugement concernant la pertinence de la démarche dans notre contexte évaluatif. Les trois tâches qui nous semblent le mieux adaptées à nos besoins sont la tâche de jugement de la syllabe accentuée (avec de vrais mots et des pseudomots), la tâche de discrimination entre deux patrons accentuels et le QCM de compréhension. La tâche de jugement de syllabe accentuée, sélectionnée pour sa simplicité et la facilité de sa mise en œuvre, peut être discriminante avec des francophones. Cependant, il nous a semblé que certains de nos étudiants, qui n'auraient jamais entendu parler d'accentuation lexicale, pourraient avoir du mal avec le côté métalinguistique de cette tâche. Nous avons donc également décidé d'inclure la tâche de discrimination AX de Leong et al. (2011), où il suffit de donner une réponse « oui » ou « non » à la question « est-ce que ces 2 mots ont le même rythme ? », question qui nous paraît moins abstraite que « quelle syllabe de ce mot est accentuée ? ». Nous comparerons d'ailleurs les résultats à ces deux tâches pour confirmer ou infirmer cette hypothèse. Enfin, nous avons ajouté quelques items de QCM de segmentation, tirés de Tabata (2016), qui ne demandent aucune connaissance métalinguistique (« choisissez l'image qui correspond à ce que vous entendez »).

test	source(s)	remarques	pertinence
parole réitérée	Nakatani et al. 1978	difficile pour natifs	non
parole réitérée	Goswami et al. 2003	interférence des connaissances lexicales	non
filtre passe-bas (AXB)	Wood & Terrell 1998	pas d'exemples d'items difficile pour natifs	non
amorçage intermodal	Cooper et al. 2002 Tremblay 2008	difficile pour natifs	non
jugement syllabe accentuée	Davis & Kelly 1997 Altmann 2006	avec mots ou pseudomots	oui
discrimination AX	Dupoux et al. 1997 Leong et al. 2011	avec même mot ou même patron accentuel	oui
jugement accent correct	Meerman et al. 2014	interférence connaissances lexicales	non
QCM compréhension	Tabata 2016	aucune connaissance métalinguistique requise	oui

Tableau 5.1 - récapitulatif des tests permettant d'évaluer la sensibilité à l'accentuation et évaluation de leur pertinence dans notre contexte

5.3. Construction du test de sensibilité accentuelle

5.3.1. Matériel expérimental (stimuli)

Suite à cette analyse des paradigmes utilisés en psycholinguistique, nous avons donc choisi, pour le test de sensibilité prosodique, quatre groupes d'items, comportant chacun une vingtaine de questions (sauf pour le dernier qui est plus restreint) et correspondant à trois tâches différentes :

- tâche 1 : 20 items tirés de Leong et al. (2011) dont deux items d'entraînement au début du test (tâche de comparaison de deux patrons accentuels) ;
- tâche 2a : 21 mots difficiles à prononcer correctement pour les francophones (tâche d'identification de la syllabe accentuée avec des mots existants) ;
- tâche 2b : 20 pseudomots tirés de Altmann (2006), pour lesquels il faut également identifier la syllabe accentuée ;
- tâche 3 : 6 items de QCM inspirés de Tabata (2016), pour lesquels il faut choisir l'image correspondant à l'expression entendue.

5.3.1.1. tâche 1 : comparaison de patrons accentuels

Les 20 premiers items, pour la première tâche, ont été fabriqués au départ (Leong et al., 2011) à partir de 20 mots de quatre syllabes accentués sur la première syllabe (comme *CAterpillar*), et de 20 autres mots de quatre syllabes accentués sur la deuxième (comme *maTERnity*), présentés dans le Tableau 5.2.

mots accentués sur syllabe 1	**mots accentués sur syllabe 2**
auditory	botanical
categorize	capacity
caterpillar	curriculum
cauliflower	debatable
citizenship	delivery
comfortable	democracy
dandelion	discovery [mal enregistré]
delicacy	facility
difficulty	harmonica
educator	historical
fertilizer	magnificent
lavatory	manipulate
mercenary	maternity
military	miraculous
monastery	necessity
organizer	participant
pacifier	pistachio
punishable	remarkable
secondary	ridiculous
voluntary	velocity

Tableau 5.2 - liste des mots utilisés pour créer les items du test de sensibilité accentuelle (tirés de Leong et al., 2011)

Les mots choisis pour les deux listes sont de fréquence similaire : les 20 mots accentués sur la première syllabe ont une fréquence par million dans le corpus COBUILD (base de données CELEX disponible en ligne, Baayen et al., 1995) de 19,35 (écart-type 30,88) et ceux accentués sur la deuxième syllabe de 19,75 (écart-type 18,23). Pour vérifier que ces moyennes sont effectivement comparables (ce qui ne paraît pas évident étant donné que les écarts-types sont très différents), nous avons effectué un test de Wilcoxon (parce que la distribution des fréquences de l'échantillon n'est pas normale) : W=1,746, p=0,08. Nous considérerons donc, comme dans l'article original, que la fréquence des deux listes est équivalente (même si on s'approche d'une différence significative). Au final, nous n'avons utilisé que 39 mots : le mot *discovery* a été enlevé car il y a eu une erreur lors de l'enregistrement et c'est *discover* qui a été prononcé.

Les mots du Tableau 5.2 ont été enregistrés une fois avec la première syllabe accentuée, et une autre fois avec la deuxième syllabe accentuée, ce qui a donné au final 80 mots (20 bien accentués sur la première syllabe, 20 mal accentués sur la première syllabe, 20 bien accentués sur la deuxième, et 20 mal sur la deuxième). Comme les items originaux n'étaient plus disponibles, ils ont été réenregistrés par une locutrice native d'anglais américain, enseignante d'anglais, mais sans formation particulière en phonétique. Pour lui faire prononcer les mots accentués de façon inhabituelle, des exemples lui ont été donnés (« *Prononce 'MAternity' comme si tu disais 'MAtter dada'* »), et plusieurs répétitions ont parfois été nécessaires. Pour vérifier que la prononciation du même mot accentué soit sur la première (*MAternity*), soit sur la deuxième syllabe (*maTERnity*) était effectivement différente, nous les avons analysées avec le logiciel Praat (Boersma & Heuven, 2001). Comme nous l'avons vu à la section 2.2 (chapitre 2), les syllabes accentuées diffèrent en anglais des non accentuées de par leur durée, leur hauteur, et leur intensité. Nous avons donc calculé de façon semi-automatique dans Praat ces valeurs pour les deux versions de chaque mot, et les résultats de ces analyses sont présentés dans le Tableau 5.3. La moitié haute du tableau correspond à l'analyse de la première syllabe, et la moitié basse à la deuxième syllabe. La partie gauche correspond aux syllabes inaccentuées, et la droite aux accentuées. Pour faciliter la lecture de l'exemple, nous avons surligné en gris la prononciation inhabituelle de *maternity* (première syllabe accentuée, deuxième syllabe non accentuée).

On observe que les syllabes accentuées sont effectivement plus longues, plus aiguës et plus fortes que les non accentuées (en moyenne), quelle que soit leur position. Pour savoir si ces différences sont significatives, nous avons utilisé le test *t* sur échantillons appariés, puisque

les deux listes appariées de 39 mots fournissent un échantillon suffisamment grand pour que le test soit pertinent. La quatrième colonne du Tableau 5.3 montre que toutes les différences entre les deux versions de chaque mot sont hautement significatives. Les syllabes accentuées se distinguent donc bien des inaccentuées (et vice-versa), ce qui peut permettre aux auditeurs de les différencier en utilisant une ou plusieurs de ces caractéristiques, seules ou en combinaison puisqu'elles se renforcent les unes les autres.

	non accentuée	*accentuée*	*t(38)*
première syllable	**ma**TERnity	***MA****ternity*	
durée en ms *(sd)*	125.3 (*40.7*)	174.3 (*46.6*)	-7.05***
F0 moyenne en Hz *(sd)*	191.5 (*10*)	219.5 (*14.4*)	-13.09***
intensité médiane (*sd*)	77.63 (*3*)	80.76 (*1.8*)	-6.77***
deuxième syllable	MA**ter**nity	ma**TER**nity	
durée en ms *(sd)*	153 (*44.8*)	175.3 (*37.6*)	-3.51**
F0 moyenne en Hz *(sd)*	198.6 (*10.4*)	208.6 (*15.3*)	-4.17***
intensité médiane (sd)	77.6 (*3.7*)	80 (*1.9*)	-4.15***

Tableau 5.3 - analyse acoustique des 1ère et 2ème syllabes, accentuées ou non, des 39 mots utilisés dans le test de sensibilité prosodique (la prononciation inhabituelle du mot utilisé comme exemple est surlignée en gris), * : p< .05 ** : p< .01 *** : p< .001

Des paires de mots ont ensuite été créées en mélangeant les deux patrons d'accentuation (habituel et non habituel), en essayant d'équilibrer toutes les combinaisons possibles. Cependant, comme nous créons un test diagnostique destiné à l'enseignement et non un test psycholinguistique, l'équilibrage n'a pas besoin d'être strict. Il s'agit simplement d'éviter des tendances observables afin que les bonnes réponses ne soient pas prévisibles pour les candidats. Le Tableau 5.4 présente quelques exemples d'items (la liste complète se trouve en Annexe 2).

numéro	**mots**	**patrons accentuels**	**accentuation habituelle (H) / inhabituelle (I)**	**même accentuation**
3	maternity botanical	/1000/ /1000/	I - I	oui
4	facility-necessity	/0100/-/0100/	H - H	oui
6	democracy-velocity	/1000/- /0100/	I - H	non
8	dandelion-mercenary	/1000/- /1000/	H - H	oui
10	historical-curriculum	/0100/ -/1000/	H - I	non

Tableau 5.4 - exemples d'items du test de conscience accentuelle

5.3.1.2. tâche 2a : identification de la syllabe accentuée (mots)

Les 21 items suivants demandent aux candidats d'identifier la syllabe accentuée dans des mots, cette fois toujours prononcés correctement. Comme nous sommes à la recherche d'items discriminants, donc pas trop faciles, notre choix s'est porté sur des mots de trois

syllabes, où la répartition entre mots accentués sur la première ou deuxième syllabe (55 vs. 39%) est plus équilibrée que chez les mots de deux syllabes (75 vs 25%, cf. deuxième chapitre, section 2.2.1.1). Pour que la tâche soit suffisamment difficile pour notre public, nous avons choisi des mots qui figurent parmi les mots mal accentués à l'oral dans les rapports du Capes (MEN, 2010, p.39) et d'agrégation (MEN, 2011, p. 93, p.113) publiés chaque année par le ministère de l'éducation nationale français. En effet, nous faisons l'hypothèse que s'ils sont mal accentués, c'est peut-être en partie qu'ils ne sont pas bien analysés lors du processus de compréhension (selon Tremblay, 2008, qui trouve un lien entre accentuation en production et accentuation en compréhension). Tous ces mots sont connus des francophones à l'écrit puisque ce sont des mots transparents (*cognates*) : *consider, etiquette, artifice,…*, mais sous leur forme orale, certains peuvent être connus de certains candidats et d'autres non. Douze d'entre eux sont accentués sur la première syllabe, et neuf sur la deuxième, ce qui correspond à peu près à la répartition mentionnée plus haut dans le lexique anglais en général pour les mots de trois syllabes (57% vs. 43% dans notre test, et 55% vs. 39% dans le lexique). Ces mots ont été téléchargés du site de prononciation en ligne collaboratif FORVO (Forvo, 2008).

num	mot	syllabe accentuée	fréquence *mpm* (Spoken Celex)	rang (COCA)	mot très fréquent
1	adjective	1	8	5000+	non
2	argument	1	108	1189	oui
3	artifice	1	0	5000+	non
4	character	1	164	2049	oui
5	consider	2	137	2509	oui
6	contribute	2	39	1319	oui
7	develop	2	210	485	oui
8	encourage	2	81	1188	oui
9	etiquette	1	2	5000+	non
10	harvested	1	0 (V)	4985	non
11	implicit	2	2	5000+	non
12	interpret	2	24	2783	oui
13	interview	1	42 (N) 33 (V)	947	oui
14	motivate	1	16	3570	non
15	negative	1	42	1478	oui
16	occurrence	2	5	5000	non
17	processes	1	147	392(n)/2771(v)	oui
18	rhetoric	1	2	3748	non
19	satire	1	5	5000+	non
20	specific	2	60	982	oui
21	syntactic	2	1	5000+	non

Tableau 5.5 – liste et caractéristiques des mots utilisés pour la deuxième tâche du test de sensibilité prosodique (la fréquence de *Spoken Celex* est donnée en mots par million)

Onze de ces mots sont très fréquents. Ils font partie des 3 000 mots les plus courants selon le *Corpus of Contemporary American English* (COCA, M. Davies & Gardner, 2013) ce qui

correspond environ à une fréquence supérieure à 20 par million dans la partie orale du corpus Celex (Baayen et al., 1995), basé sur COBUILD (J. Sinclair, 1991). Dix autres mots sont de fréquence moyenne ou basse : ils ne font pas partie des 5 000 mots les plus courants de COCA (à l'exception de *motivate* et *rhetoric*, qui appartiennent à la bande de fréquence allant des rangs 3 000 à 4 000), et ont une fréquence inférieure à 20 par million.

5.3.1.3. Tâche 2b : identification de la syllabe accentuée (pseudomots)

La troisième partie du test tente de remédier au biais introduit dans la tâche précédente par l'utilisation de mots qui peuvent être connus à l'oral de certains étudiants mais pas d'autres, en utilisant des pseudomots trisyllabiques tirés de la thèse d'Altmann (2006). Ceci nous permet également d'introduire quelques mots accentués sur la dernière syllabe (ce qui implique d'ailleurs la présence d'un accent secondaire sur la première syllabe). Altmann demande à ses sujets d'entourer sur une feuille la syllabe qu'ils pensent accentuée, mais nous avons préféré garder la même consigne qu'à la partie précédente (identifier la syllabe accentuée), dans la mesure où la représentation orthographique des pseudomots utilisés dans l'étude originelle (*leytauwma*, *nafeepa*) pourrait déconcerter les candidats. Les enregistrements originaux n'étant pas disponibles, ces mots ont été réenregistrés par un locuteur d'anglais britannique (RP) ayant des connaissances en phonétique. Comme on peut le voir dans le Tableau 5.6, sept d'entre eux sont accentués sur la première syllabe, huit sur la deuxième, et cinq sur la dernière, pour un total de 20 items.

num	mot	syllabe accentuée (accent primaire)
1	menoysa	2
2	soibena	1
3	soidetta	2
4	zedoola	2
5	fellazee	3
6	linnesoo	3
7	taremma	2
8	faysaboo	3
9	pecoitay	2
10	koyvalee	1
11	pagenoo	1
12	leytauwma	2
13	nafeepa	1
14	dafeanoo	2
15	roodela	1
16	joamaray	3
17	savossauw	2
18	delloyma	1
19	savvaney	3
20	lisseda	1

Tableau 5.6 - liste des pseudomots utilisés pour la deuxième tâche du test de sensibilité prosodique

5.3.1.4. tâche 3: QCM de compréhension

Contrairement aux précédents, les derniers items, inspirés de Tabata (2016) et de Peppé et McCann (2003), ne font pas explicitement référence aux concepts d'accent et de rythme dans la consigne, mais demandent simplement aux candidats de sélectionner l'image correspondant à l'expression entendue. Les expressions sont à chaque fois composées des mêmes items lexicaux (Figure 5.1), mais leurs structures syntaxiques sous-jacentes diffèrent : soit les deux premiers mots forment un nom composé (*chocolate biscuits*), qui est coordonné avec un deuxième nom (*and jam*), soit les trois noms sont coordonnés (*chocolate, biscuits and jam*).

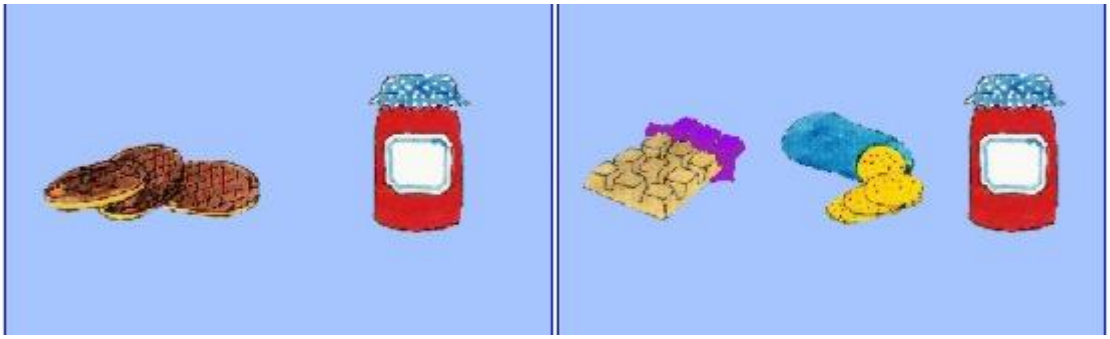


Figure 5.1 - exemple d'images associées à un item de compréhension du test de sensibilité prosodique : *chocolate biscuits and jam* vs. *chocolate, biscuits and jam*

Nous avons inséré dans des phrases les quatre items retenus de Tabata (2016) et légèrement modifiés (nous n'avons pas gardé les items utilisant plusieurs adjectifs de couleur, qui nous semblaient peu naturels : *black and pink & red* vs. *black & pink, and red*). Nous y avons ajouté deux items tirés de Harley et al. (1995), qui utilisent un test similaire pour comparer l'utilisation des indices prosodiques et syntaxiques en compréhension de l'oral chez les enfants anglophones L1 ou L2. Les six items finaux sont présentés dans le Tableau 5.7.

n	phrase avec mots composés	phrase sans mots composés
1	*I need chocolate biscuits, jam and juice.*	*I need chocolate, biscuits, jam, and juice.*
2	*I need chocolate milk, tea, and coffee.*	*I need chocolate, milk, tea, and coffee.*
3	*I need ice cream, broccoli, and cheese*	*I need ice, cream, broccoli, and cheese*
4	*I need plastic shoes, glasses, and shorts*	*I need plastic, shoes, glasses, and shorts*
5	*Have you ever seen a dragonfly?*	*Have you ever seen a dragon fly?*
6	*Where's Mikey?*	*Where's my key?*

Tableau 5.7 – items de la troisième tâche du test de sensibilité prosodique, contrastant des phrases ayant les mêmes items lexicaux ou les mêmes segments, mais des structures syntaxiques différentes

Lors du pilotage du test à l'automne 2017, un problème technique a empêché les images associées à ces six items de s'afficher à l'écran. L'analyse qui suit porte donc uniquement sur les 61 premiers items (tâches 1, 2a et 2b). Nous avons cependant rajouté une nouvelle fois les items de la tâche 3 (moins le cinquième, certains étudiants nous ayant dit ne pas connaître le

mot *dragonfly*) lors de l'expérimentation finale à l'automne 2018, mais en proposant un choix entre deux phrases écrites plutôt qu'entre deux images. Nous mentionnerons brièvement les résultats de ce rajout en fin de section 5.4.2.

5.3.2. Administration du test

Les items décrits en section 5.3.1 (tâche 1 de comparaison de patrons accentuels et tâche 2a/b d'identification de la syllabe accentuée de mots/pseudomots) ont été assemblés dans l'interface auteur du système d'administration de tests « SELF » (Cervini et al., 2013), développé à l'Université Stendhal-Grenoble3 (avant qu'elle ne soit fusionnée dans l'Université Grenoble Alpes) dans le cadre du projet Innovalangues (Masperi, 2012). Le test a ensuite été passé par les étudiants, en présence d'un enseignant, dans un des laboratoires de langues équipés en ordinateurs des sites concernés (Grenoble et Valence). A chaque fois que l'utilisateur valide sa réponse à un item, l'item suivant apparaît. Il n'est pas possible de ne pas répondre ni de répondre « je ne sais pas ». Un écran apparaît entre chaque partie du test pour faire la transition, et demande à l'utilisateur s'il est prêt à continuer.

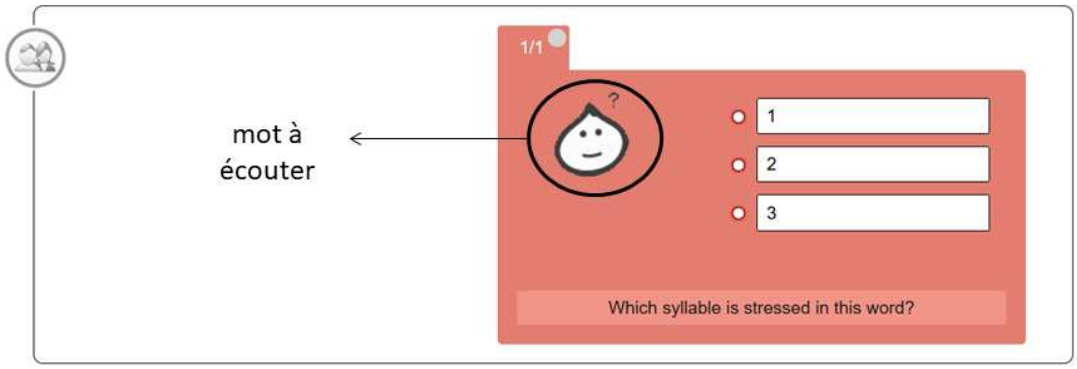


Figure 5.2 - interface de la plateforme d'administration de test SELF: exemple d'un item du test de sensibilité prosodique (tâche d'identification de la syllabe accentuée)

La Figure 5.2 présente un exemple d'écran de l'interface du test. L'icône en forme de goutte (ou d'oignon) accompagnée d'un point d'interrogation indique l'endroit où il faut cliquer pour entendre le mot à analyser. Les fichiers exports de résultats, comprenant la réponse choisie par chaque étudiant à chaque item, le temps passé sur chaque item, et le score total, ont ensuite été récupérés sous format .csv pour être traités et analysés.

5.4. Résultats

5.4.1. Statistiques descriptives globales

Nous mettrons ici en oeuvre les analyses quantitatives résumées dans le chapitre 4 (Tableau 4.5 - Résumé des analyses qualitatives et quantitatives mises en œuvre pour la validation des tests diagnostiquesTableau 4.5). Le test a été passé par un échantillon de 143 candidats, avec un temps moyen d'un peu plus de 11 minutes (11,2), ce qui excède légèrement l'objectif fixé de 10 minutes, mais reste dans des limites raisonnables. Cependant, il faut noter que les candidats les plus lents ont mis un peu plus de 20 minutes (21,2).

Le Tableau 5.8 présente en deuxième colonne le score moyen (45,5/61, soit 75% de réussite). Ce score moyen est assez élevé : contrairement à ce que nous aurions pu penser à la lecture de la littérature, et en particulier de la théorie de la surdité accentuelle (Dupoux et al., 2008), nos étudiants semblent tout à fait capables de distinguer les syllabes accentuées en contexte lexical, qu'ils connaissent le mot ou non. D'autres études plus récentes (Michelas et al., 2016) montrent d'ailleurs que les francophones arrivent en fait assez bien à discriminer l'accentuation en contexte lexical, même s'ils sont plus lents que pour la discrimination phonémique. La dispersion des scores peut être appréhendée par l'écart-type et l'étendue. On constate qu'il existe un écart très important de 44 points entre le résultat le plus haut (61) et le plus bas (17) : si certains étudiants sont au plafond, d'autres ont beaucoup de mal avec ces tâches de sensibilité accentuelle. Les coefficients d'asymétrie et de kurtose ont une valeur acceptable (entre -2 et 2), indiquant une répartition des scores raisonnablement symétrique et étalée. La valeur négative de l'asymétrie montre que les scores sont plus étalés dans les valeurs faibles. Comme nous l'avons expliqué dans la première partie, cela peut être un atout pour notre test qui serait ainsi plus à même de distinguer les apprenants ayant besoin de remédiation.

n	moy.	é.t.	médiane	min	max	étendue	asymétrie	kurtose	se
143	45.52	9.87	47	17	61	44	-0.44	-0.4	0.83

Tableau 5.8 - statistiques descriptives duscore total du test de sensibilité prosodique

L'histogramme des scores (Figure 5.3) confirme l'étalement des résultats en dessous de la moyenne. Il n'est pas clair au vu du graphique si la distribution se rapproche d'une distribution normale ou pas. Nous le vérifions avec le test de Shapiro-Wilk ($W = 0,966$, $p < 0.01$). Ces scores ne sont donc pas normalement distribués, et nous ne pourrons pas les soumettre à des tests statistiques de type paramétrique.

Par ailleurs, l'alpha de Cronbach du test est de 0,9, bien au-dessus du seuil acceptable de 0,7 (Laveault, 2012). Cela signifie que les items vont globalement dans le même sens et que la cohérence interne du test est excellente.

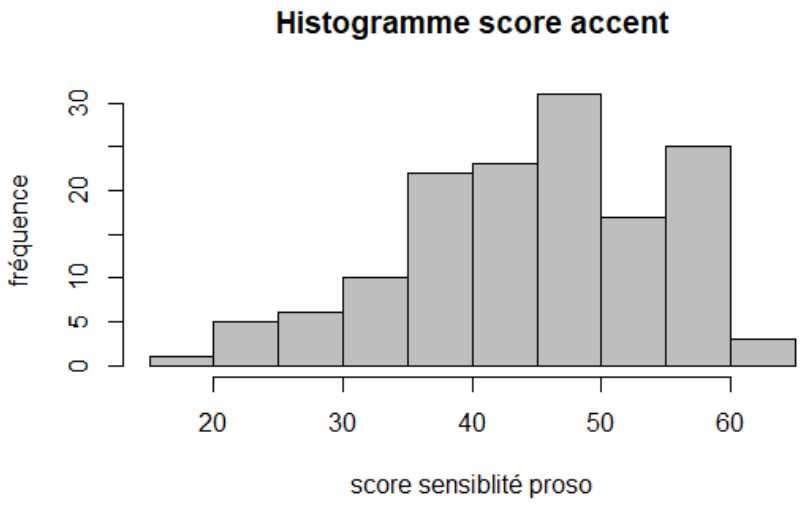


Figure 5.3 - histogramme des scores de sensibilité prosodique

5.4.2. Analyse des items

Nous allons à présent analyser pour chaque item de notre test de sensibilité prosodique l'indice de difficulté (aussi appelé *p-value*), c'est-à-dire le taux de réussite à l'item (qui doit être compris entre 0,2 et 0,9 environ), et l'indice de discrimination, mesuré par le coefficient de corrélation bisérial de point, qui mesure la corrélation (Pearson) entre la réussite à l'item et la réussite au test. Cet indice doit être supérieur ou égal à 0,2 pour qu'un item fonctionne de façon acceptable (0,3 pour qu'il fonctionne vraiment bien). Le Tableau 5.9 présente les résultats de ces analyses (bibliothèque sjPlot de *R*). Les résultats des items des tâches 1, 2a et 2b sont présentés en parallèle, chaque (sous-)tâche occupant quatre colonnes du tableau.

La première colonne indique le nom de l'item, dans lequel on peut reconnaître le contenu de l'item (mots utilisés). Vient ensuite le pourcentage de candidats ayant choisi la bonne réponse (la difficulté de l'item), l'écart-type et le coefficient de discrimination. Pour le premier item de la première tâche, *comfortable.cauliflower*, par exemple, qui est un item d'entraînement, les candidats devaient indiquer si les deux mots avaient le même schéma accentuel (ce qui est le cas ici, l'item contenant les deux mots avec leur schéma accentuel canonique, accentués sur la première syllabe). Soixante-six pour cent des candidats ont choisi la bonne réponse (oui), ce qui signifie que l'item est assez difficile (la mauvaise réponse a donc été choisie par un tiers des participants). Comme il s'agit du premier item du test et d'un item d'entraînement, il

est normal que les étudiants l'aient trouvé difficile. Le coefficient de discrimination est mauvais (inférieur à 0,2), ce qui signifie que la réussite à cet item est très peu corrélée à la réussite au test dans son entier. Autrement dit, les « bons » et les « mauvais » candidats ont presque autant de chance de trouver la bonne réponse (ou de se tromper) à cet item. Encore une fois, ce n'est pas inattendu pour le premier item du test.

tâche 1				**tâche 2a**				**tâche 2b**			
nom item	**diff.**	**é.t.**	**item discr.**	**nom item**	**diff.**	**é.t.**	**item discr.**	**nom item**	**diff.**	**é.t.**	**item discr.**
comfortable.cauliflower	0.66	0.5	**0.138**	adjective	0.85	0.36	0.416	menoysa2	0.88	0.32	0.378
military.magnificent	0.66	0.5	0.298	argument	0.78	0.42	0.443	soibena1	0.86	0.35	0.398
maternity.botanical	0.74	0.4	0.331	artifice	0.8	0.4	0.45	soidetta2	0.84	0.37	0.358
facility.necessity	**0.92**	0.3	**0.034**	character	0.8	0.4	0.469	zedoola2	0.79	0.41	0.338
maternity.botanical	0.71	0.5	0.305	consider	0.77	0.42	0.535	fellazee3	0.57	0.5	0.262
dandelion.mercenary	0.62	0.5	0.349	contribute	0.73	0.44	0.55	linnesoo3	0.8	0.4	0.307
lavatory.fertilizer	0.76	0.4	0.375	develop	0.69	0.46	0.504	taremma2	0.67	0.47	0.395
historical.curriculum	0.71	0.5	0.282	encourage	0.78	0.42	0.468	faysaboo3	0.78	0.41	0.334
delicacy.monastery	0.66	0.5	0.337	etiquette	0.71	0.46	0.421	pecoitay2	0.8	0.4	0.312
capacity.ridiculous	0.69	0.5	0.379	harvested	0.78	0.42	0.402	**koyvalee1**	**0.91**	0.29	0.262
cauliflower.caterpillar	0.52	0.5	0.292	implicit	0.59	0.49	0.35	paggenoo1	0.88	0.32	0.308
difficulty.voluntary	0.78	0.4	0.336	interpret	0.74	0.44	0.459	leytauwma2	0.9	0.3	0.357
comfortable.organizer	0.69	0.5	0.313	interview	0.59	0.49	0.337	naffeepa1	0.71	0.45	0.355
educator.categorize	0.64	0.5	0.37	motivate	0.77	0.42	0.452	dafeanoo2	0.82	0.39	0.405
secondary.military	0.82	0.4	**0.151**	negative	0.82	0.39	0.497	roodela1	0.83	0.38	0.482
auditory.citizenship	0.77	0.4	0.202	occurrence	0.8	0.4	0.366	joamaray3	0.64	0.48	0.463
punishable.pacifier	0.76	0.4	**0.122**	processes	0.84	0.37	0.315	savossauw2	0.47	0.5	0.212
magnificent.delivery	0.78	0.4	0.224	rhetoric	0.83	0.38	0.458	**delloyma1**	0.74	0.44	**0.111**
participant.manipulate	0.76	0.4	**0.127**	satire	0.73	0.45	0.405	savvaney3	0.42	0.5	0.318
miraculous.pistachio	0.8	0.4	0.31	specific	0.68	0.47	0.394	lisseda1	0.83	0.38	0.462
				syntactic	0.85	0.36	0.34				

Tableau 5.9 - analyse des items de sensibilité prosodique par tâche et sous-tâche ; les items dont le nom est en gras ont de mauvais indices de difficulté (en gras, items très faciles, *p-value* > .9) ou de discrimination (en gras et surlignés en gris, items peu discriminants, indice <.2)

Parmi les autres items, deux sont très faciles (92% de bonnes réponses pour *facility.necessity* et 91% pour *koyvalee1,* en gras). Comme le coefficient de discrimination de *facility.necessity* n'est pas particulièrement bon, il pourra être éliminé de la version finale du test (nous garderons tout de même *koyvalee1*, dont la discrimination est bonne). L'indice de discrimination (coefficient de discrimination bisérial de point) est mauvais (inférieur à 0,2) pour cinq items (en gras, et coefficients surlignés en gris), en plus du premier item d'entraînement. On peut remarquer que quatre de ces items se trouvent dans la première

tâche, dont le format peu classique a peut-être déconcerté certains candidats. Tous les items de la deuxième tâche (2a), plus classique (identification de la syllabe accentuée de mots existants), ont bien fonctionné.

La version finale du test de sensibilité prosodique piloté à l'automne 2017 contiendra donc 13 items de la première tâche (après enlèvement des 2 items d'entraînement initiaux, des 4 items restants qui ne discriminent pas bien, et d'un item dupliqué par erreur, *maternity.botanical*), l'ensemble des 21 items du groupe 2, et 19 items du groupe 3, pour un total de 53 items. A l'automne 2018, lors de la dernière expérimentation à grande échelle, nous avons rajouté les cinq items de la tâche 3 (QCM de compréhension) qui avaient subi un dysfonctionnement technique lors du pilotage, et trois de ces items se sont révélé de bonne qualité : *chocolate milk* vs. *chocolate, milk* (indice de difficulté 0,79, indice de discrimination 0,27), *ice cream* vs. *ice, cream* (diff. 0,71, discr. 0,43), et *Mikey* vs. *my key* (diff 0,73, discr 0,37). Nous avons décidé de les conserver dans la version finale du test, parce que la tâche, qui ne demande aucune connaissance métalinguistique, nous paraît particulièrement intéressante (même si elle est représentée par peu d'items au final). La version finale utilisée compte donc en fait 56 items.

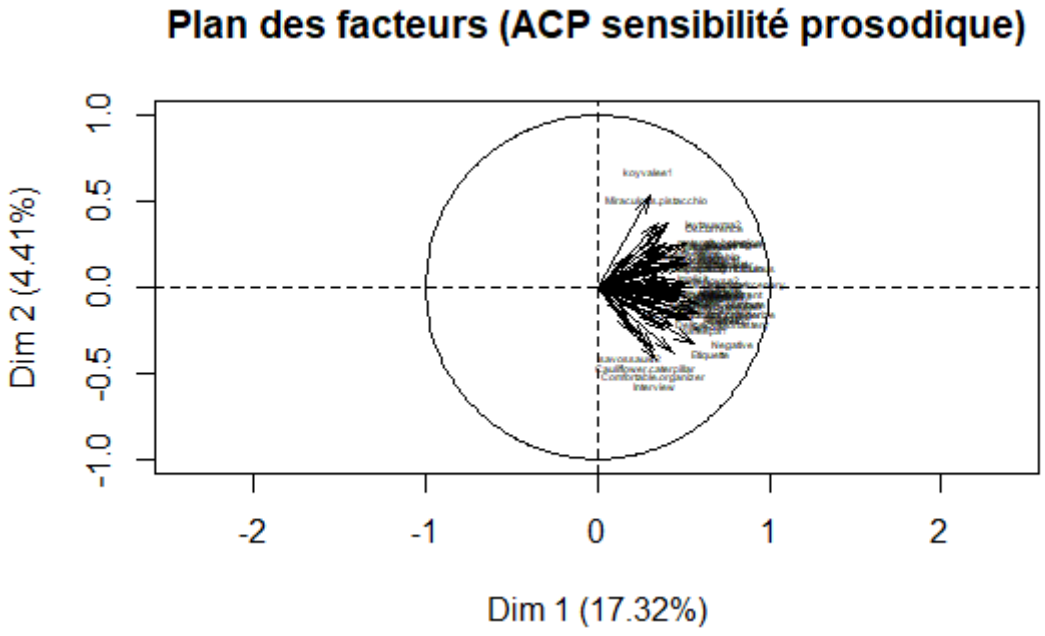


Figure 5.4 - résultat graphique de l'ACP du test de sensibilité prosodique : plan des facteurs (deux premières dimensions)

Pour évaluer l'unidimensionnalité du test final, nous avons utilisé une analyse en composants principaux, dont le but est d'identifier un facteur principal qui synthétise le maximum d'informations apportées par les différents items du test. Nous ferons ce constat par inspection

visuelle du graphe du cercle des corrélations de toutes les variables avec l'axe principal de l'ACP (Figure 5.4). Nous constatons sur le graphique que tous les items du test corrèlent de façon positive avec le facteur principal identifié (parce que tous les vecteurs les représentant sont orientés vers la droite). De plus, ce facteur explique plus de 17% de la variance du test (comme indiqué dans la parenthèse à droite du nom de l'abscisse de la Figure 5.4), ce qui est un résultat acceptable (Husson et al., 2016).

5.4.3. Analyse approfondie

5.4.3.1. comparaison des tâches 1, 2a et 2b

Nous avons émis l'hypothèse que la première tâche, qui consiste à comparer le « rythme » de deux mots, demandait moins de connaissances métalinguistiques que la tâche d'identification de la syllabe accentuée (la notion de rythme étant plus accessible que celle d'accent lexical). Nous avons cependant remarqué au paragraphe précédent que la tâche 2a semblait avoir mieux fonctionné que les autres, peut-être parce qu'elle correspond à une tâche assez classique (identifier la syllabe accentuée de mots existants). Les deux autres tâches sont moins conventionnelles : dans la troisième (tâche 2b), la consigne était semblable à celle de la tâche 2a, mais les mots étaient inventés, et dans la première, les candidats devaient comparer le schéma accentuel de mots existants qui pouvaient être bien ou mal accentués. Pour comparer le fonctionnement des trois tâches, nous avons calculé pour chaque groupe la moyenne des indices de difficulté et de discrimination (Tableau 5.10).

tâche	difficulté moyenne (é.t.)	discrimination moyenne (é.t.)
1	0.72 (0.09)	0.26 (0.10)
2a	0.76 (0.08)	0.43 (0.07)
2b	0.76 (0.14)	0.34 (0.09)

Tableau 5.10 - Comparaison des items des tâches 1, 2a et 2b du test de sensibilité prosodique

On constate que la difficulté moyenne des trois groupes d'items est similaire (0,72, 0,76 et 0,76). Par contre, la discrimination moyenne est assez contrastée. Pour le premier groupe, elle est de 0,26, c'est-à-dire acceptable mais sans plus (rappelons que l'indice de discrimination est mauvais en dessous de 0,2, acceptable entre 0,2 et 0,3, et bon au-delà). Pour le deuxième groupe, elle est au contraire très bonne (0,43), et bonne pour le dernier (0,34). Le test de Kruskal-Wallis montre que ces différences sont significatives (χ^2 = 26,9, df = 2, p<0,001), et le test post-hoc de Wilcoxon avec correction Bonferroni montre que cela est vrai aussi bien pour la différence entre les tâches 1 et 2a ($p < 0,001$) et 2a et 2b ($p < 0,01$), que pour celle entre les tâches 1 et 2b ($p < 0,05$). Il semble donc que plus les items utilisent une tâche

classique, plus ils discriminent entre les candidats, et que contrairement à notre hypothèse, la première tâche, moins métalinguistique en apparence, n'est pas du tout plus facile pour nos étudiants. Nous garderons cependant les 13 items du premier groupe qui ont résisté à l'analyse précédente, et qui nous permettent d'avoir un test plus long, donc plus robuste et plus fiable.

Pour ce qui est de la première tâche, on peut aussi supposer que les jugements « identiques » sont plus difficiles à effectuer que les jugements « différents », dans la mesure où il faut faire abstraction du fait que les phonèmes et le sens de mots diffèrent pour remarquer que les patrons accentuels sont identiques. Cependant, les jugements identiques sont en fait réussis à 0,74 (écart-type 0,085), donc mieux que les différents à 0.70 (é.-t. 0,086), et la différence n'est pas significative d'après le test de Wilcoxon (W = 59, p = 0,49).

5.4.3.2. influence de la fréquence lexicale

Les items de la deuxième tâche du test de sensibilité prosodique se répartissent en mots très fréquents et moins fréquents. Il est donc intéressant de faire l'analyse de ces items : d'une part, on pourrait s'attendre à ce que les mots moins fréquents à l'oral aient un patron accentuel plus difficile à reconnaître, mais d'autre part, nous avons vu dans la première partie qu'un contexte lexical connu pouvait défavoriser les apprenants. Nous avons déjà noté que les mots de fréquence nulle (les pseudomots du troisième groupe d'items) ne sont pas plus difficiles à analyser prosodiquement pour nos étudiants que les mots de mots existants, ce qui semble aller à l'encontre de la première hypothèse. Nous reproduisons dans le Tableau 5.11 les items de la tâche 2a avec leur indice de difficulté. Les quatre premières colonnes correspondant aux mots fréquents, et les quatre dernières aux mots plus rares.

mots fréquents	**fréquence (Sp. Celex)**	**rang (COCA)**	**diff.**	**mots plus rares**	**fréquence (Sp. Celex)**	**rang (COCA)**	**diff.**
interview	42 (N) 33 (V)	947	.59	implicit	2	5000+	.59
specific	60	982	.68	etiquette	2	5000+	.71
develop	210	485	.69	satire	5	5000+	.73
contribute	39	1319	.73	motivate	16	3570	.77
interpret	24	2783	.74	harvested	0	4985	.78
consider	137	2509	.77	occurrence	5	5000	.8
encourage	81	1188	.78	artifice	0	5000+	.8
argument	108	1189	.78	rhetoric	2	3748	.83
character	164	2049	.8	syntactic	1	5000+	.85
negative	42	1478	.82	adjective	8	5000+	.85
processes	147	392 (N) 2771 (V)	.84				
moy (é.t.)			.75 (.07)				.77 (.08)

Tableau 5.11 - répartition des items de sensibilité prosodique (tâche 2a) en 2 groupes en fonction de leur fréquence (nombre d'occurrences par million de mots dans le corpus oral Spoken Celex, et rang dans le corpus COCA (« 5000+ » = n'appartient pas aux 5000 premiers mots), et présentés par difficulté décroissante

La difficulté moyenne des items avec des mots fréquents ou plus rares est pratiquement identique : 0,75 pour les mots fréquents (écart-type 0,07), et 0,77 pour les moins fréquents (écart-type 0,08). On peut vérifier avec un test *t* (étant donné que les variances sont presque égales, et malgré le petit nombre d'observations) que la différence n'est pas significative (t=-0,73, df = 19, p = 0,48). On peut souligner d'ailleurs qu'on trouve des mots difficiles (0,59 de réussite) à la fois dans les mots fréquents (*interview*) et dans les mots plus rares (*implicit*), et de même pour les mots bien réussis, qui peuvent aussi bien être fréquents (*processes*) que rares (*rhetoric*).

Les étudiants n'ont donc pas plus de difficultés à identifier la syllabe accentuée de mots existants rares ou fréquents, résultat peu surprenant étant donné que les pseudomots sont également assez bien réussis. Il est donc possible que nos apprenants aient une compétence prosodique décontextualisée qui leur permet d'analyser un patron prosodique indépendamment de son instanciation et de faire abstraction du contexte lexical dans lequel celui-ci est utilisé. On peut également se demander si la syllabe accentuée joue un rôle : puisque l'accentuation sur la première syllabe est la plus fréquente pour les mots de trois syllabes en anglais, il est possible que les mots qui suivent cette accentuation par défaut soient plus faciles à reconnaître. Les mots accentués sur la première syllabe sont effectivement mieux réussis : 0,77 (0,07) de réussite contre 0,74 (0,08) pour les mots accentués en deuxième syllabe, mais la différence n'est pas significative non plus (t = 1,09, df = 19, p = 0,29).

Nous n'avons donc aucune piste pour expliquer la difficulté pour nos sujets des deux mots nettement moins bien réussis que les autres, *interview* et *implicit*. L'un est très fréquent (*interview*), l'autre est beaucoup plus rare (*implicit*), l'un est accentué sur la première syllabe (*interview*), l'autre sur la deuxième (*implicit*). Il semble que ces différences soient idiosyncratiques et dépendent de chaque mot, même si dans l'ensemble, encore une fois, les étudiants sont tout à fait capables d'identifier la syllabe accentuée dans les trois-quarts des cas.

5.4.4. Conclusion

Nous avons montré dans ce chapitre que le test de conscience prosodique que nous avons élaboré était fiable, avec une très bonne cohérence interne (α = 0,9), et une distribution qui

illustre une bonne discrimination entre les candidats faibles et les forts. Le test initial comptait 6 items présentant une discrimination trop faible, mais les 53 autres items ont fonctionné correctement, avec difficulté et discrimination acceptables. Nous avons donc utilisé ce test de 53 items (une fois écartés les six items de qualité insatisfaisante) pour la suite de l'expérimentation (nous l'avons cependant par la suite augmenté de trois items qui n'avaient pas pu être pilotés lors du pilotage décrit dans ce chapitre).

Nous avons de plus constaté que, contrairement à ce à quoi on pourrait s'attendre, des étudiants francophones, sans formation particulière en prosodie (car elle est peu enseignée dans l'éducation secondaire, Voise, 2010), sont capables de comparer deux patrons accentuels ou d'identifier la syllabe accentuée de mots connus ou inconnus, fréquents ou non. Nous n'avons pas observé d'effet lexical, ce qui montre que, dans ce type de tâche, le fait de connaître ou non le mot n'influence pas la réponse des apprenants. Cette constatation peut être vue de façon positive : les apprenants francophones ont acquis une sensibilité à l'accentuation qui fait qu'ils sont capables d'identifier la syllabe accentuée d'une suite de syllabes, contextualisée ou non. D'un autre côté, on peut voir ce résultat de façon plus négative : les étudiants français ne sont pas meilleurs pour repérer la syllabe accentuée des mots qu'ils connaissent, par rapport à des mots inconnus. Il est donc tout à fait plausible, comme plusieurs chercheurs en ont fait l'hypothèse (Dupoux et al., 1997; Tremblay, 2008), que les mots (connus) ne soient pas mémorisés avec leur schéma accentuel.

Chapitre 6

Test de discrimination phonémique

6.1. Rappels du cadre théorique

Les difficultés des apprenants L2 à utiliser des contrastes phonémiques qui n'existent pas dans leur L1, et en particulier des francophones apprenant l'anglais, sont avérées. Pour les consonnes, il s'agit essentiellement de la confusion entre les fricatives /θ/ (*thin*, inconnue en français) et /s/ et /f/, d'une part, et /ð/ (*this*, également inconnue) et /z/ et /v/ d'autre part, ainsi que de la présence ou absence de /h/. Pour les voyelles, nous avons vu à la section 2.1.2.2 que les contrastes les plus difficiles à acquérir sont /iː/ vs. /ɪ/ (*heel* et *hill*) et /ɑː/ vs /ʌ/ (*heart* vs. *hut*), mais aussi /ɒ/ vs. /ɔː/ (*pot* vs. *port*), /ɑː/ vs /ɒ/ vs. /æ/ (*part*, *pot* et *pat*), /ɒ/ vs. /ʌ/ (*cot* et *cut*), /ʌ/ vs /ɜː/ (*bud* et *bird*), /əʊ/ vs /ɔː/ (*hole* et *haul*). D'autres paires comme /æ/ vs. /ʌ/ sont en général mieux réussies.

Parmi ces contrastes potentiellement intéressants à tester, nous n'avons pas gardé ceux qui avaient des réalisations ou des distributions trop différentes en anglais britannique et anglais américain, comme le /ɔː/, le /əʊ/ ou le /ɒ/ (Larreya & Schottman, 2013). Nous avons par contre gardé le contraste /æ/ vs. /ʌ/, réputé plus facile, mais qui peut justement nous aider à repérer les étudiants en grande difficulté. Nous avons par ailleurs rajouté un contraste non mentionné dans la littérature sur la perception, mais identifié comme posant des problèmes en production : /e/ vs /eɪ/ (*met* vs. *mate*, Terrier, 2011, p. 45).

6.2. Inventaire d'instruments d'évaluation

Beaucoup de tests de conscience phonémique (en particulier en L1) comportent un élément de production orale. Cependant, comme nous étudions la compréhension, et que nous avons besoin de tests autocorrectifs qui puissent être passés en autonomie, nous ne pourrons pas faire usage de ces tests. Nous nous concentrerons donc sur les tests qui font uniquement

intervenir la réception, et ne présenterons pas non plus les paradigmes expérimentaux utilisant les potentiels évoqués (par ex. Mah et al., 2016).

6.2.1. Test de discrimination AX sur phonèmes, allophones ou paires minimales

Le test le plus simple est probablement de faire écouter deux sons à des sujets et de leur demander s'il s'agit d'un même son ou de deux sons différents. C'est le test utilisé par Patricia Kuhl (1991) dans l'expérience décrite dans le chapitre 2 (section 2.1.1) pour étudier la discrimination phonétique chez les adultes (chez les enfants, c'est la procédure « *head turn* » qui est utilisée), et en particulier pour étudier le phénomène d'aimant perceptuel des voyelles prototypiques, difficiles à distinguer des voyelles phonétiquement proches. Il s'agissait là d'une expérience avec des natifs (et sur la discrimination phonétique et non phonémique de voyelles L1), mais le même test a été utilisé pour étudier la discrimination phonémique en L2. Hayes-Harb (2007), par exemple, l'exploite pour estimer l'efficacité de divers entraînements à la discrimination entre [k] (non aspiré) et [g] suivis d'une voyelle (par exemple [gæ] vs. [kæ]) chez des anglophones, avec des stimuli artificiels (la distinction est non phonémique en anglais en position initiale, les deux étant des allophones du phonème /g/, c'est donc une distinction nouvelle pour les sujets).

Cependant, à part quand les distinctions sont très ténues (sons très proches chez Kuhl, 1991, ou sur un continuum artificiel entre deux phonèmes comme chez Christophe Pallier et al., 1997), ou quand la distinction est entièrement nouvelle pour les sujets (Hayes-Harb, 2007), ce type de test n'est pas assez discriminant car il donne rapidement des résultats au plafond : par exemple, les sujets de Hayes-Harb obtiennent plus de 93% de bonnes réponses à toutes les questions demandant une réponse « même ». Nous ne pourrons donc pas utiliser ce test.

6.2.2. Test de discrimination « cherchez l'intrus » (*oddity*) ou test AXB sur phonèmes, syllabes ou mots

Le test « cherchez l'intrus », qui consiste à présenter aux sujets une suite de phonèmes ou de syllabes (en général trois) et à leur demander lequel est différent des autres, est très utilisé. C'est notamment un des tests présents dans l'étude de Strange et Dittman (1984) sur l'identification de la distinction /r/ - /l/ en anglais américain par des japonophones (avec des stimuli artificiels), et surtout dans l'étude classique de Flege et McKay (2004) sur la perception des voyelles de l'anglais canadien par des italophones installés au Canada anglophone depuis plus ou moins longtemps. Les stimuli sont naturels, prononcés par trois

voix différentes, et les résultats sont intéressants dans la mesure où ils sont au plafond pour les natifs (ce qui montre que la tâche n'est pas trop difficile), mais très contrastés pour les apprenants (entre 55 et 95% de réussite selon les voyelles considérées), ce qui permet de créer des items discriminants.

Ce test peut être rapproché des tests de discrimination AXB (par exemple Bundgaard-Nielsen et al., 2011, sur la discrimination des voyelles d'anglais australien par des japonophones), où l'on présente également trois exemplaires aux sujets (dans ce cas des bisyllabes où les voyelles étudiées sont insérées dans une structure /hVba/). Dans cette tâche, le premier et le dernier exemplaire sont toujours différents, et la tâche consiste à décider si l'exemplaire du milieu ressemble plus au premier ou au dernier exemplaire[36]. Nous avons choisi d'utiliser la version « classique » de cette tâche, plus simple à expliquer aux étudiants qui doivent pouvoir faire le test en autonomie (c'est également une des tâches choisies par Krzonowski et al., 2016, qui nous ont permis d'utiliser leurs enregistrements).

6.2.3. Test d'identification de phonèmes (classification) ou d'identification lexicale

Dans le test d'identification de phonèmes, également très répandu, les sujets entendent un phonème ou un mot et doivent identifier ce qu'ils ont entendu. Il s'agit en général d'un choix forcé, c'est-à-dire que les réponses possibles sont déjà indiquées, et les participants doivent entourer ou cocher la réponse choisie. Parmi les études que nous avons déjà citées, beaucoup utilisent plusieurs tests, dont un test d'identification de phonèmes. C'est le cas par exemple de Strange et Dittman (1984) qui, en plus du test de discrimination présenté plus haut, utilisent deux tests d'identification : un avec des stimuli artificiels, et un appelé « test de paires minimales » avec des stimuli naturels, où les sujets doivent entourer sur une feuille lequel des deux éléments de la paire minimale ils pensent avoir entendu. Les participants ont environ 70% de réussite, ce qui paraît assez discriminant pour envisager d'utiliser cette tâche dans notre test. Krznowski et al. (2016) utilisent également un test d'identification à choix forcé où les sujets doivent cliquer sur un bouton de clavier d'ordinateur pour indiquer à l'écran quel phonème contient le mot (monosyllabique) qu'ils viennent d'entendre. Les résultats là aussi

[36] On peut rapprocher cette tâche de celles utilisées en L1 où il faut appuyer sur un bouton dès que le son qu'on entend change (un intrus s'immisce dans la suite entendue), mais qui n'est pas utilisable dans notre contexte du fait de sa complexité (il faut alors mesurer le temps de réponse des sujets)

sont loin d'être au plafond (pour des étudiants francophones), avec entre 60 et 80% de bonnes réponses selon la voyelle.

Cependant, ce test requiert soit d'utiliser des symboles phonétiques pour représenter les phonèmes, soit de proposer aux sujets une représentation orthographique des sons, syllabes ou mots proposés. Cela pose un problème quand les tests sont conçus pour être utilisés en autonomie, comme c'est notre cas. En effet, la plupart des lycéens n'ont jamais appris l'alphabet phonétique au lycée (Frost & Henderson, 2013, §25, mentionnent que de nombreux enseignants français ne se sentent pas à l'aise avec l'API) et ne sont donc pas capables d'utiliser les symboles correspondants pour choisir la bonne réponse parmi les choix proposés (même si certains connaissent cet alphabet, d'ailleurs, le fait que certains le connaissent et d'autres non introduit un biais dans l'expérience qui fait que les résultats ne sont pas interprétables). L'autre solution, qui consiste à proposer une représentation orthographique, n'est pas satisfaisante non plus dans la mesure où tous les lycéens n'ont pas forcément acquis les correspondances graphèmes-phonèmes de l'anglais. Comme le notent Iverson et al. (2012, p. 157), « *The problem with identification tasks is that a listener must know the language well enough to use the response labels correctly*. » En effet, ils constatent dans leur expérience sur la perception et la production des voyelles de l'anglais par des francophones que leurs sujets ne maîtrisent pas la prononciation de certains digraphes vocaliques comme <ou> (ibid., p. 155). Ces difficultés nous ont finalement conduite à écarter cette tâche de notre test.

6.3. Construction du test

6.3.1. Matériel expérimental

Nous avons donc choisi d'administrer un test de discrimination de type « cherchez l'intrus », avec des suites de trois mots prononcés par trois locuteurs différents, afin de proposer aux candidats des tâches suffisamment difficiles pour qu'elles soient discriminantes y compris pour les niveaux élevés, le tout sans nécessiter de connaissances explicites sur l'orthographe ou la transcription phonétique de l'anglais. Dans ces trois mots, l'un est un « intrus », formant une paire minimale avec les 2 autres. Par exemple, pour le contraste /ɑ:/ vs. /ʌ/, nous avons une suite *bard – bard – bud* (prononcés par trois locuteurs différents), où le troisième mot est l'intrus. Nous nous sommes concentrée sur les phonèmes ou les contrastes identifiés dans la littérature scientifique comme étant particulièrement difficiles pour les francophones. La liste des 10 contrastes et des 35 items retenus au final est présentée dans le Tableau 6.1.

contrastes voyelles	paire 1	paire 2	paire 3	paire 4	paire 5
/ɑ:/ vs. /ʌ/	bard bud	carp cup	heart hut	park puck	
/ɑ:/ vs. /æ/	bark back	ban barn	card cad	hard had	
/æ/ vs. /ʌ/	had Hudd	lack luck	mat mutt	tat tut	
/i:/ vs. /ɪ/	beak bick	beat bit	cheap chip	deed did	
/u:/ vs. /ʊ/	fool full	Luke look	pool pull	shooed should	
/ɛ/ vs. /eɪ/	fell fail	pet pate	sess sace	tet tate	
contrastes consonnes					
/f/ vs. /θ/	cliff clith				
/s/ vs. /θ/	face faith	force fourth	truce truth	mouse mouth	seam theme
/z/ vs. /ð/	loathe loze	uzzer other			
/z/ vs. /θ/	zin thin	ms. myth			

Tableau 6.1 - contrastes phonémiques et paires minimales retenus pour le test de discrimination phonémique

Certains des items font intervenir des mots existants (par exemple *force* ou *fourth*), d'autres non (*loze* ou *bik*), afin de couvrir l'éventail des difficultés possibles de discrimination, en ou hors contexte lexical (nous avons vu au chapitre deux que la discrimination était a priori plus facile hors contexte, au moins pour les niveaux faibles, Mah et al., 2016).

Les mots enregistrés utilisés comme stimuli proviennent de trois sources: les paires minimales illustrant les cinq premiers contrastes vocaliques ont été enregistrées pour l'étude de Krzonowski et al. (2016), par 18 locuteurs (neuf hommes et neuf femmes), de langue maternelle anglaise, originaires du Sud-Est de l'Angleterre[37]. Nous avons essayé de répartir ces voix sur l'ensemble des items, sans mélanger les sexes en général. La plupart des items proposent trois mots prononcés par trois hommes différents, ou trois mots prononcés par trois femmes différentes. La deuxième source d'enregistrements est le site collaboratif FORVO (Forvo, 2008), auquel nous avons également eu recours pour certains des items utilisés pour le test de sensibilité prosodique, et utilisé ici pour le dernier contraste vocalique (/ɛ/ vs. /eɪ/) et pour les contrastes consonantiques. Cependant, le site FORVO ne propose que des enregistrements de mots existants. Pour les mots inventés de ces cinq derniers contrastes, nous avons fait appel à deux locuteurs d'anglais britannique, un homme et une femme. De ce fait, les suites de trois mots illustrant ces contrastes sont parfois prononcées par des locuteurs de sexe différent. La liste complète des items exacts se trouve en Annexe 3.

6.3.2. Administration du test

[37] nous remercions Jennifer Krzonowski de nous avoir généreusement donné accès à ses enregistrements

De même que pour les autres tests diagnostiques développés pour ce travail, nous avons utilisé la plateforme d'administration SELF. Une copie du premier écran avec les instructions proposées est disponible en Figure 6.1. La Figure 6.2 représente une copie d'écran d'un item quelconque du test. L'interface est la même que pour le test de conscience prosodique; seule la consigne située en bas de l'écran change. L'expérimentation a eu lieu à l'automne 2018.

Figure 6.1 - écran d'accueil du test de discrimination phonémique dans l'interface d'administration SELF

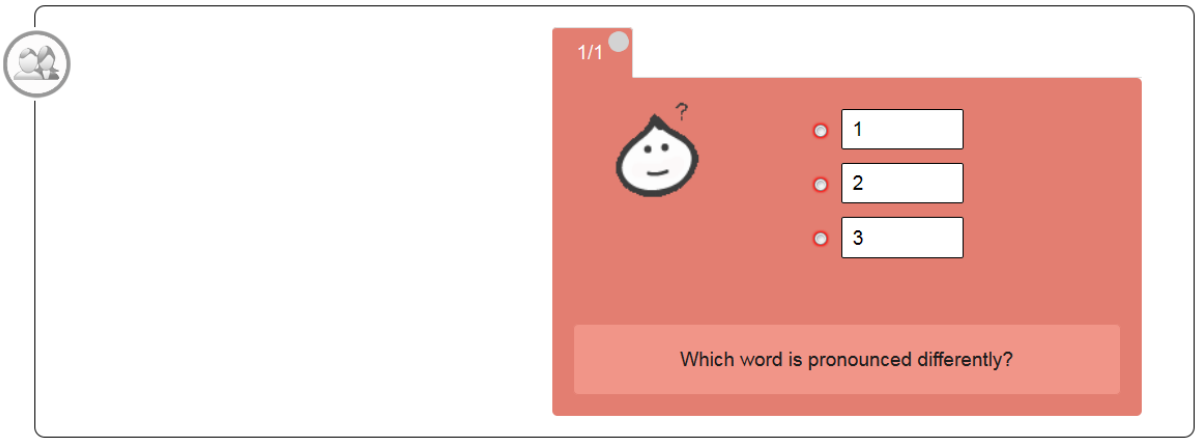


Figure 6.2 - écran d'un item du test de discrimination phonémique (interface SELF)

6.4. Résultats

6.4.1. Statistiques descriptives globales

En moyenne, les 183 étudiants ont eu besoin de 6,4 minutes (écart-type 1,8) pour passer ce test, ce qui est clairement en deçà de notre limite de 10 minutes. Le candidat le plus lent a mis 16 minutes.

Les résultats du test, passé par 183 étudiants, sont présentés dans le Tableau 6.2 ci-dessous. La moyenne obtenue aux 35 items du test (23,74/35, soit 68%), est bien au-dessus du hasard, qui est à 33% puisque les sujets avaient une chance sur trois de trouver la bonne réponse à chaque item. La tâche a donc été bien comprise, mais il existe un écart important de 24 points (colonne « étendue ») entre le plus faible résultat (7, en deçà du hasard qui serait entre 11 et 12 points) et le meilleur (31). Les coefficients d'asymétrie et de kurtose ont une valeur acceptable (entre -2 et 2), indiquant une répartition des scores raisonnablement symétrique et étalée. La valeur négative de l'asymétrie montre que les scores sont plus étalés dans les valeurs faibles, ce qui peut nous permettre de mieux distinguer les apprenants ayant besoin de remédiation.

n	moyenne	écart-type	médiane	min	max	étendue	asymétrie	kurtose
183	23.74	4.42	25	7	31	24	-1.17	1.74

Tableau 6.2 - Statistiques descriptives du score total du test de discrimination phonémique

L'histogramme des scores (Figure 6.3) confirme l'étalement des résultats en dessous de la moyenne. Il ne semble pas que les scores suivent une distribution normale (du fait de la longue queue à gauche, produisant une asymétrie), ce que nous pouvons vérifier avec le test de normalité de Shapiro-Wilk : W = 0,92 ($p < 0,001$). Nous devons rejeter l'hypothèse nulle selon laquelle la distribution de nos données suit une loi normale, et nous ne pourrons donc pas par la suite leur appliquer de tests paramétriques.

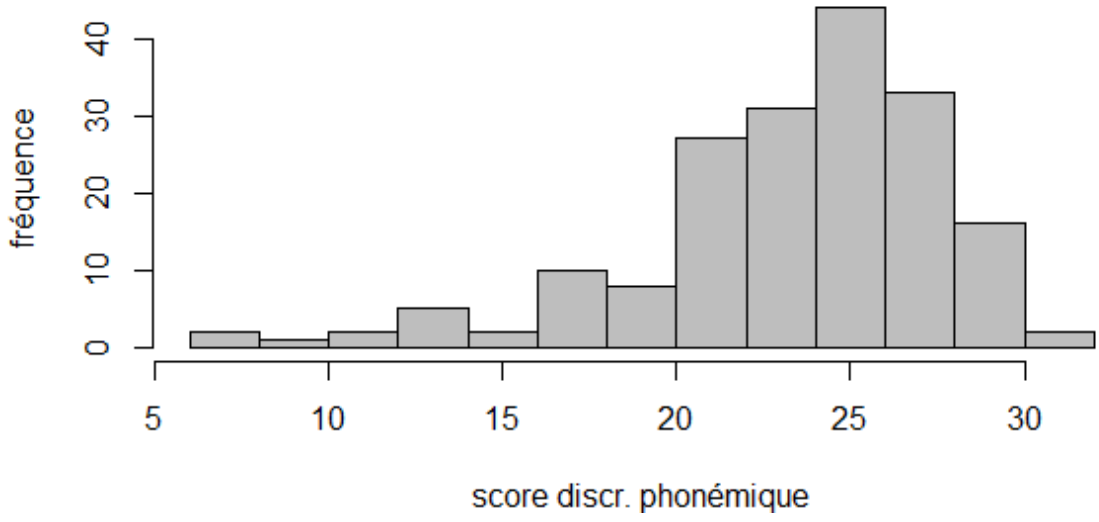


Figure 6.3 - histogramme des scores du test de discrimination phonémique

La corrélation entre le temps total passé sur le test et le score total, que nous avons calculée avec un coefficient de Spearman (car aucune des deux variables ne suit une distribution normale), est très faible, et n'est pas significative (*rho* = -0.105, p = 0.16). Le fait que *rho* soit

négatif nous apprend cependant que les meilleurs apprenants ont tendance à être un peu plus rapides que les autres.

L'alpha de Cronbach pour cette passation est de 0,72, ce qui est au-dessus de la valeur acceptable de 0,7. Nous pouvons donc considérer que la fiabilité du test est suffisante.

6.4.2. Analyse des items

Pour chaque item du test, nous calculons l'indice de facilité (ou difficulté), qui est le taux de réussite à l'item et doit être compris entre 0,2 et 0,9 environ afin que les items ne soient ni trop difficiles ni trop faciles, et l'indice de discrimination, mesuré par le coefficient de corrélation bisérial de point, qui mesure la corrélation (Pearson) entre la réussite à l'item et la réussite au test. Cet indice doit être au moins égal à 0,2 pour qu'un item fonctionne de façon acceptable (0,3 pour qu'il fonctionne vraiment bien).

Le Tableau 6.3 présente les résultats de ces analyses (effectuées avec le logiciel *R*, bibliothèque *psych*, Revelle, 2017). La première colonne indique le numéro de l'item, la deuxième le nom de l'item, dans lequel on peut reconnaître la paire minimale correspondante. Vient ensuite le numéro de la clé, c'est-à-dire de la proposition qui est la bonne réponse, suivi du pourcentage de candidats ayant choisi chacune des propositions de réponse. Pour le premier item, *bardbud*, par exemple, la bonne réponse était la troisième proposition (l'intrus était le dernier mot de la suite de trois mots). 22% des candidats ont choisi le premier mot, 10% le deuxième, et 67% le troisième (la bonne réponse). La septième colonne contient le coefficient de discrimination bisérial de point. Pour *barbud*, il est de 0,44, ce qui montre que cet item discrimine très bien entre les « bons » et les « mauvais » candidats. La colonne suivante indique le nombre de candidats qui ont passé l'item. On constate que tous les items ont été passés par tous les candidats, il n'y a donc pas de données manquantes. Enfin, l'avant-dernière colonne indique la difficulté de l'item, qui a la même valeur que le pourcentage de candidats qui ont choisi la bonne proposition (ici 67%), et la dernière, l'écart-type.

On constate que quatre items sont extrêmement faciles (entre 92 et 94% de bonnes réponses, en gras), mais comme leur coefficient de discrimination est bon, cela signifie que le peu de candidats qui ne trouvent pas la bonne réponse à ces items sont des candidats très faibles. Ces items servent probablement à différencier les « très faibles » candidats des simplement « faibles ». Nous avons donc décidé de conserver ces items qui peuvent nous aider à identifier

les candidats ayant le plus besoin de remédiation. Un item est trop difficile (*cliffclith*, 14% de réussite), mais nous l'étudierons plus en détail dans le paragraphe qui suit.

num	nom	clé	choix prop 1	choix 2	choix 3	discrim.	n	diff.	é.t.
1	bardbud	3	0.22	0.10	0.67	0.44	183	0.67	0.47
2	carpcup	1	0.58	0.32	0.10	0.37	183	0.58	0.49
3	barkback	1	0.84	0.10	0.05	0.39	183	0.84	0.37
4	bitbeat	3	0.10	0.05	0.85	0.55	183	0.85	0.36
5	barnban	3	0.05	0.14	0.81	0.39	183	0.81	0.39
6	hadhud	3	0.07	0.06	0.87	0.33	183	0.87	0.34
7	lookluke	3	0.05	0.06	0.89	0.47	183	0.89	0.32
8	chickcheek	1	0.51	0.16	0.33	0.39	183	0.51	0.50
9	hearthut	2	0.32	0.60	0.08	0.38	183	0.60	0.49
10	cardcad	1	0.90	0.06	0.04	0.42	183	0.90	0.31
11	**bickbeak**	2	0.58	0.27	0.14	**0.04**	183	0.27	0.45
12	lackluck	2	0.14	0.80	0.06	0.27	183	0.80	0.40
13	**cesssace**	3	0.03	0.05	0.92	0.42	183	**0.92**	0.28
14	parkpuck	2	0.40	0.44	0.15	0.21	183	0.44	0.50
15	pullpool	2	0.14	0.71	0.15	0.24	183	0.71	0.45
16	muttmat	3	0.19	0.17	0.64	0.25	183	0.64	0.48
17	tatetet	1	0.90	0.04	0.07	0.44	183	0.90	0.31
18	diddeed	2	0.21	0.46	0.33	0.30	183	0.46	0.50
19	shouldshooed	2	0.05	0.31	0.64	0.22	183	0.31	0.46
20	tattarttot	2	0.10	0.84	0.07	0.49	183	0.84	0.37
21	**fullfool**	2	0.72	0.21	0.07	**-0.16**	183	0.21	0.41
22	**fellfail**	3	0.04	0.04	0.92	0.40	183	**0.92**	0.28
23	hardhad	2	0.31	0.61	0.08	0.38	183	0.61	0.49
24	**petpate**	3	0.05	0.03	0.92	0.41	183	**0.92**	0.28
25	**thinzin**	3	0.03	0.03	0.94	0.39	183	**0.94**	0.24
26	**cliffclith**	1	0.14	0.83	0.04	**0.05**	183	**0.14**	0.34
27	loatheloze	3	0.02	0.09	0.89	0.42	183	0.89	0.31
28	otheruzzer	1	0.67	0.30	0.04	0.44	183	0.67	0.47
29	mythmis	1	0.68	0.03	0.28	0.28	183	0.68	0.47
30	mouthmouse	2	0.10	0.81	0.09	0.36	183	0.81	0.39
31	themeseam	2	0.10	0.68	0.22	0.38	183	0.68	0.47
32	trucetruth	2	0.56	0.39	0.04	0.23	183	0.39	0.49
33	facefaith	1	0.85	0.06	0.09	0.35	183	0.85	0.36
34	**force4th**	1	0.36	0.21	0.44	**0.05**	183	0.36	0.48
35	sighthigh	3	0.04	0.08	0.87	0.37	183	0.87	0.33
									Cronbach's α : 0.72

Tableau 6.3 - résultats de l'analyse des items du test de discrimination phonémique : les items dont le nom est grisé ont de mauvais indices de difficulté (en gras, items très faciles, *p-value* > .9) ou de discrimination (en gras et surlignés en gris, items peu discriminants, indice <.2)

Le coefficient de discrimination bisérial de point est proche de zéro ou même négatif pour 4 autres items (*bickbeak*, *fullfool*, *cliffclith* et *force4th*), ce qui signifie que la réussite à ces items n'est pas du tout corrélée à la réussite au test dans son entier. Ces items sont d'ailleurs parmi les plus difficiles du test, avec respectivement 27%, 21%, 14% et 36% de réussite. Dans chaque cas, c'est un distracteur qui a été choisi en majorité comme réponse pour l'intrus. L'item *fullfool* illustre ce problème: alors que la clé (l'intrus) est le deuxième mot, une majorité de candidats a choisi le premier mot. Après réécoute de l'item, il semble que le problème vienne du dernier mot. Les deux premiers mots sont effectivement clairement différents, mais la prononciation du troisième ne permet pas de distinguer s'il s'agit d'un /u:/ ou d'un /ʊ/ (on peut remarquer d'ailleurs que la charge fonctionnelle (A. Brown, 1988) du contraste /u:/ - /ʊ/ est assez faible et qu'il ne permet pas de différencier beaucoup de mots en anglais). Parmi les trois autres items défectueux, *cliffclith* souffre d'un problème similaire : les étudiants ont choisi le deuxième mot comme intrus parce qu'il est prononcé par une voix masculine très basse, alors que les deux autres sont prononcés par une femme. Ils ont visiblement été distraits par la différence de hauteur de voix et n'ont pas réussi à se concentrer sur le contraste consonantique en finale. Les deux autres items qui ont dysfonctionné ne nous semblent pas problématiques d'un point de vue acoustique. Dans *bickbeak*, qui teste le contraste /ɪ/ - /i:/, le premier mot contient un /ɪ/ assez long qui a visiblement déconcerté les candidats qui l'ont choisi comme intrus (le contraste est souvent présenté comme une différence de longueur plutôt que de qualité dans certains manuels et les explications de certains enseignants), alors que les deux autres mots de l'item étaient clairement différents l'un de l'autre, le deuxième avec /bi:k/ et le troisième avec /bɪk/. C'était donc un item difficile, mais qui aurait pu bien fonctionner. Après réécoute, il est difficile de comprendre pourquoi le dernier item, *force4th,* n'a pas départagé les candidats, l'intrus (*fourth*) est clairement le premier mot de la suite. Il a d'ailleurs un taux de réussite (36%) légèrement meilleur que celui des trois autres items que nous venons d'examiner.

Pour résumer, parmi nos 35 items, 27 ont bien ou très bien fonctionné, 4 sont très faciles mais permettent d'identifier des apprenants très faibles et sont donc à garder, 3 sont clairement trop difficiles et sont donc à écarter, et le dernier est tangent (parce qu'un peu difficile). Comme nous avons décidé de conserver ce dernier item, il nous reste un test où nous gardons 32 items sur 35, et que nous pouvons continuer à analyser. Dans la version initiale, l'alpha de Cronbach (indice de fiabilité) du test était de 0,72, juste au-dessus du seuil acceptable de 0,7 (Laveault

2012). L'élimination des trois items dysfonctionnels permet de faire remonter cette valeur à 0,76, ce qui est tout à fait satisfaisant.

Pour évaluer l'unidimensionnalité du test final, nous avons utilisé une analyse en composants principaux, dont le but est d'identifier un facteur principal qui synthétise le maximum d'informations apportées par les différents items du test. Nous ferons ce constat par inspection visuelle du graphe du cercle des corrélations de toutes les variables avec l'axe principal de l'ACP (Figure 6.4). Nous constatons sur le graphique que tous les items du test corrèlent de façon positive avec le facteur principal identifié (parce que tous les vecteurs les représentant sont orientés vers la droite). De plus, ce facteur explique plus de 15% de la variance du test (comme indiqué dans la parenthèse en bas), ce qui est un résultat acceptable (Husson et al., 2016).

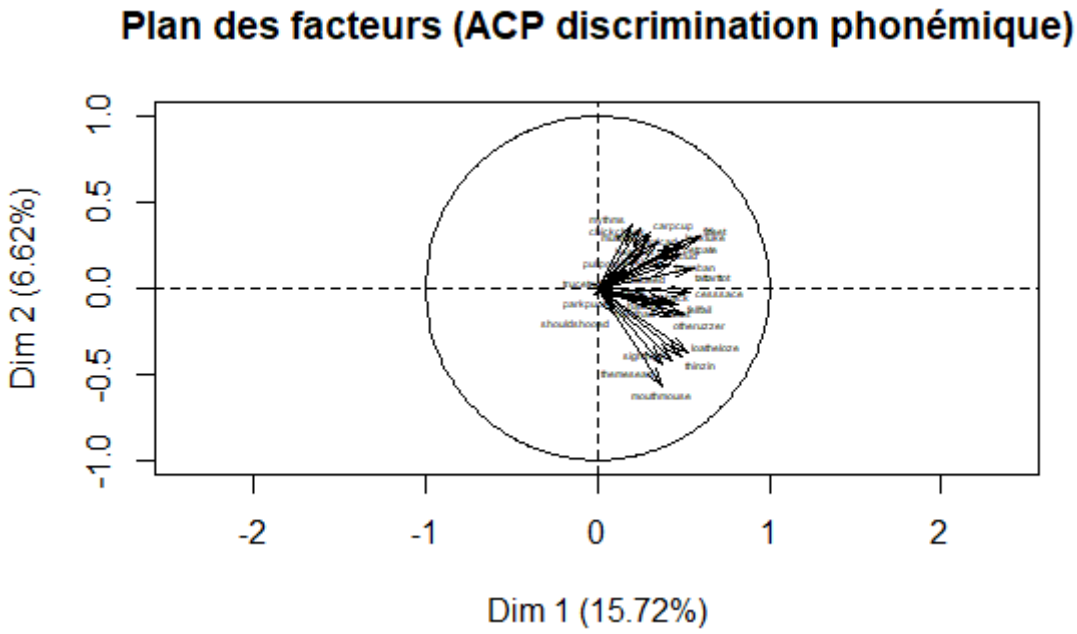


Figure 6.4 - ACP discrimination phonémique

6.4.3. Analyse approfondie

6.4.3.1. difficulté constatée des contrastes

Comme nous sommes dans le cadre d'un test diagnostique et non psycholinguistique, nous avons un nombre d'items par condition (ici, par contraste phonémique) assez réduit, entre un et cinq dans notre test. Nous ne pouvons donc pas tirer de conclusions sur les différences éventuellement observées entre ces contrastes, mais nous pouvons vérifier que nos données vont bien dans le sens de la littérature scientifique, en observant l'ordre de difficulté des

contrastes pour nos étudiants. Le Tableau 6.4 présente les résultats pour les paires minimales représentées par plus de quatre items (par ordre croissant de réussite de gauche à droite).

contraste	/i:/ - /ɪ/	/u:/ - /ʊ/	/a:/ - /ʌ/	/s/ - /θ/	/æ/ - /ʌ/	/æ/ - /a:/	/ɛ/ - /eɪ/
diff moy	.52	.53	.57	.66	.79	.79	.92
exemple	cheek - chick	pool - pull	bard - bud	face - faith	lack - luck	ban - barn	fell - fail
Krzonowski et al. (2016)	mal réussi		mal réussi		bien réussi		
Iverson et al. (2012)	mal réussi .5						bien réussi .8

Tableau 6.4 - difficulté moyenne observée en fonction du contraste phonémique, classée par ordre croissant, accompagnée des résultats correspondants des études de Krzonowski et al. (2016) et Iverson et al. (2012)

On peut observer deux groupes de contrastes vocaliques : d'une part, les contrastes /i:/ - /ɪ/, /u:/ - /ʊ/ et /a:/ - /ʌ/, qui se révèlent difficiles (moins de 60% de réussite), et d'autre part les contrastes /æ/ - /ʌ/, /æ/ - /a:/ et /ɛ/ - /eɪ/, beaucoup plus faciles (autour de 80% de réussite et plus). Ces résultats correspondent assez bien à ceux trouvés dans la littérature, que nous avons ajoutés sous les nôtres dans le tableau à titre de comparaison : parmi nos trois contrastes difficiles, deux ont également donné beaucoup de mal aux sujets de Krzonowski et al. (2016) et Iverson et al. (2012), dont nous avons décrit les résultats plus haut (le troisième, /u:/ - /ʊ/, n'était pas un objet d'étude, peut-être à cause de sa faible charge fonctionnelle en anglais). Parmi les trois contrastes faciles, deux sont également bien réussis par les sujets de ces deux études (le troisième, /æ/ - /a:/, n'est pas étudié). Le contraste consonantique /s/ - /θ/, dont la difficulté est médiane entre celle du groupe de contrastes vocaliques faciles et du groupe difficile, n'a pas pu être comparé à d'autres études faute d'avoir pu trouver de résultats chiffrés publiés sur la question. Cependant, les résultats sur les voyelles nous permettent d'ajouter un élément à la validation de notre test, dont les résultats correspondent à ceux d'autres études publiées.

6.4.3.2. corrélation entre temps et réussite au niveau des items

Nous avions constaté un peu plus haut qu'il n'y avait (pratiquement) pas de corrélation au niveau du test entre le temps total passé par les candidats et leur score. Au niveau des items, au contraire, la corrélation (toujours avec Spearman) entre le temps passé et la difficulté existe de façon plus prononcée : *rho* = -0.51 (p = 0.003). Comme le coefficient est négatif, cela signifie que plus les items sont faciles (leur taux de réussite augmente), plus le temps de réponse diminue. Nous pouvons visualiser cette relation sur le graphique suivant (tableau 19) :

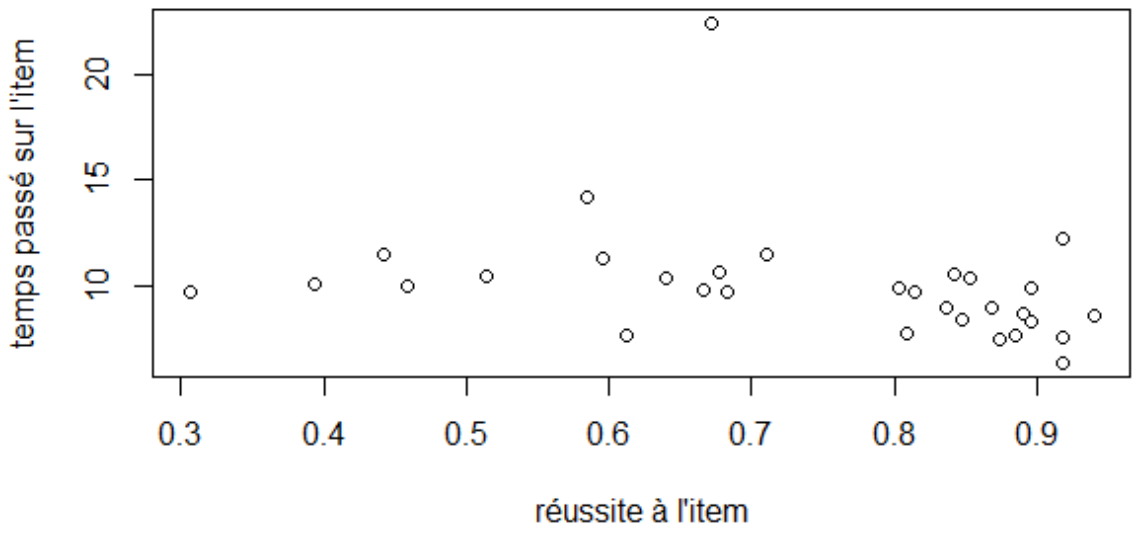


Figure 6.5 - représentation graphique de la relation entre temps passé et réussite à l'item (test de discrimination phonémique)

On constate que cette relation, quoique non négligeable (r = -0,5), est difficile à distinguer sur le graphique, peut-être parce que, pour les items très difficiles sur la gauche du graphique (entre 0,3 et 0,55 de réussite), la corrélation apparaît nulle, et qu'elle a l'air de se vérifier surtout pour les items à partir de 0,6 de difficulté environ. Il semble donc que, avant que les items ne deviennent vraiment trop difficiles, les apprenants sont capables de se rendre compte que certains items sont plus difficiles, et de prendre alors plus de temps pour tenter d'y répondre (par exemple en réécoutant les trois mots de l'item). C'est peut-être pour cela que l'on ne trouve pas de corrélation au niveau du test entier : certains apprenants entendent tout de suite le contraste et répondent très vite, tandis que d'autres ont besoin de plusieurs écoutes avant de repérer l'intrus, mais tous sont capables, du moment qu'ils disposent du temps qui leur convient, d'arriver au même résultat. On verra qu'il n'en est pas de même pour toutes les compétences testées.

6.4.4. Conclusion

Nous pouvons ainsi considérer qu'en l'état, le test de discrimination phonémique fonctionne assez bien: il contient 32 items de difficulté variée, de très facile à très difficile, et qui discriminent de façon satisfaisante entre les candidats. Il possède également une assez bonne fiabilité (alpha de Cronbach > 0,7). Nous avons de plus un premier élément de validité, dans la mesure où les résultats en termes d'échelle de difficulté des contrastes sont conformes à ceux trouvés dans la littérature. Nous poursuivrons cette étude de validité dans la troisième partie avec des analyses corrélationnelles avec les résultats en compréhension de l'oral.

Chapitre 7

Test de reconnaissance du lexique aural

7.1. Rappels du cadre théorique

Nous avons vu dans la première partie de cette étude que de nombreuses analyses vont dans le sens d'une place centrale du lexique dans la compétence langagière, et en particulier en compréhension de l'oral. Nous avons conclu que dans notre contexte, nous aurions besoin d'un test de reconnaissance (plutôt que de production), et qu'il serait intéressant d'utiliser un test de reconnaissance aurale, qui peut donner des indications plus fiables que les tests écrits sur les problèmes des apprenants en compréhension de l'oral (les francophones reconnaissant plus de mots à l'écrit grâce aux mots transparents communs au français et à l'anglais). Dans ce qui suit, nous mentionnerons les quelques tests auraux existants, mais, étant donné que la grande majorité des tests de vocabulaire utilisent la modalité écrite, nous décrirons également les tests écrits les plus courants.

7.2. Inventaire d'instruments d'évaluation

L'intérêt pour les tests de vocabulaire date de la fin du dix-neuvième siècle, et dès 1929, Verner Sims écrivait un article intitulé *The Reliability and Validity of Four Types of Vocabulary Tests*, où il comparait les formats suivants toujours utilisés aujourd'hui : les tests où les sujets doivent produire le sens du mot dont la forme leur est donnée, qu'il appelle *identification tests* (cela correspond à ce qu'on appellerait aujourd'hui *meaning recall*) ; les test à choix multiples (*meaning recognition*) ; les tests à appariement, qu'il appelle *matching tests*; et les tests « oui/non », qu'il nomme *checking tests* (appelés aujourd'hui en anglais soit *checklist tests*, soit *yes/no tests*). Nous allons examiner ces formats tour à tour, avant de considérer d'autres modalités non mentionnées dans cet article fondateur.

7.2.1. Test à production du sens (*meaning recall*)

Les tests où les sujets doivent montrer qu'ils connaissent le sens des mots proposés en produisant ce sens sont assez peu utilisés aujourd'hui, parce qu'ils sont difficiles à corriger et donc à standardiser. On leur préfère en général des formats plus contraints comme ceux qui sont décrits ensuite (QCM, appariement, vrai/faux, …), qui ne demandent en général que la reconnaissance du sens (*meaning recognition*). Cependant, cela ne signifie pas que les tests de production du sens ne sont d'aucune utilité. D'une part, la technique est très couramment employée dans la phase de validation des tests, souvent lors d'un entretien qui suit leur passation, pour montrer que les formats contraints ne surestiment pas (trop) les connaissances des candidats. Pellicer-Sanchez et Schmitt (2012) trouvent par exemple qu'un test oui/non sous-estime très légèrement les connaissances des candidats comparé au résultat d'un entretien où c'est la reconnaissance du sens qui est demandée, mais les surestime assez nettement si on exige d'eux la production du sens.

D'autre part, il existe un test qui a eu un certain succès et qui inclut un élément de production du sens. En 1996, Marjorie Wesche et T. Sima Paribakht ont proposé un format de test qui rende compte de la « profondeur » des connaissances lexicales des candidats en leur laissant la possibilité de se placer à plusieurs niveaux sur une échelle de connaissance du vocabulaire (*VKS* ou *Vocabulary Knowledge Scale*). Un exemple est proposé dans le Tableau 7.1.

I. Je n'ai jamais vu ce mot.
II. J'ai déjà vu ce mot, mais je ne connais pas son sens.
III. J'ai déjà vu ce mot, et je pense qu'il signifie _____ (synonyme ou traduction)
IV. Je connais ce mot. Il signifie _____ (synonyme ou traduction)
V. Je peux utiliser ce mot dans une phrase : _____

Tableau 7.1 - Echelle de connaissance du vocabulaire (Vocabulary Knowledge Scale) d'après Wesche et Paribakht (1996)

Nous ne pourrons utiliser ce format du fait qu'il est difficilement compatible avec la correction automatique, mais c'est un format très intéressant en auto-évaluation.

7.2.2. Questions à choix multiple (QCM)

7.2.2.1. QCM écrit

Contrairement aux tests de production du sens, le format à choix multiple est facile à corriger et sa correction peut être automatisée. Il a donc été très utilisé, depuis le *Thorndike Test of Word Knowledge* de Thorndike, 1921 (cité par Sims, 1929), jusqu'aux tests plus récents décrits ici.

Le VST (*Vocabulary Size Test*) développé par Paul Nation et David Beglar (2007) comporte 140 questions à choix multiple, chaque bande de fréquence de 1 000 familles de mots étant représentée par dix mots. Il couvre donc toutes les bandes de fréquence jusqu'à celle de 14 000, et permet d'estimer la taille du vocabulaire réceptif des participants. Le Tableau 7.2 présente un exemple d'item du test, tiré de la cinquième bande de fréquence. On peut remarquer qu'il n'est pas suffisant pour choisir la bonne réponse de savoir que *miniature* se réfère à quelque chose de petit ; une connaissance plus approfondie est nécessaire puisque les distracteurs se réfèrent également à des choses « petites » (le mot *small* est présent dans toutes les options de réponse). Ce test reste cependant un test de reconnaissance (*meaning recognition*, et non *meaning recall*) qui encourage les sujets à utiliser leurs connaissances, même partielles (Nation et Beglar précisent explicitement qu'il n'y a pas d'option « je ne sais pas » pour encourager les étudiants à répondre même s'ils ne sont pas entièrement sûrs). Enfin, les mots sont présentés dans une courte phrase de contexte non définitoire, compatible avec toutes les propositions de réponse, qui donne la catégorie syntaxique du mot et aide à rendre la tâche plus naturelle.

1. miniature : It is a **miniature**. a. a very small thing of its kind b. an instrument for looking at very small objects c. a very small living creature d. a small line to join letters in handwriting

Tableau 7.2 - exemple d'item du Vocabulary Size Test (Nation & Beglar, 2007)

7.2.2.2. QCM aural

Il existe depuis quelques années un test de vocabulaire aural, le LVLT ou *Listening Vocabulary Levels Test* (McLean et al., 2015). Ce test bilingue est calqué sur le VST à l'écrit, mais n'est disponible qu'en anglais-japonais pour l'instant. Il couvre les 5 000 mots anglais les plus fréquents, et estime la taille du vocabulaire aural avec 24 questions à choix multiple en langue maternelle par bande de fréquence de 1000 mots, pour un total de 120 items. Le tableau qui suit (Tableau 7.3) présente un exemple d'item du LVLT, avec les options de réponses traduites du japonais. On peut constater que le mot, comme dans le VST, est donné seul puis en contexte (entendre tout d'abord le mot isolé est important à l'oral pour aider les sujets à segmenter l'énoncé qui suit malgré les phénomènes éventuels d'allophonie en fonction du contexte de début et de fin de mot).

Les candidats n'ont droit qu'à une seule écoute par item (ce qui donne deux instances du mot), et le test dure plus de 30 minutes (McLean et al., 2015, p. 3). Malgré l'intérêt indéniable

de ce test, la longueur nous paraît rédhibitoire et nous avons décidé de ne pas l'adapter pour des participants francophones.

1. [Les participants entendent: « *School. This is a big school* »] a. banque b. poisson c. école d. maison

Tableau 7.3 - exemple d'item du Listening Vocabulary Levels Test, adapté de McLean et al. (2015)

Enfin, nous avons trouvé trace d'une première tentative de test de vocabulaire aural dans le contexte nippon, avec un format un peu différent (Mizumoto & Shimamoto, 2008). Dans ce test bilingue composé de 160 items, les candidats voient un mot japonais écrit sur leur feuille de réponse, et entendent quatre mots anglais parmi lesquels ils doivent reconnaître celui qui correspond à la traduction proposée en japonais. Au lieu d'un test de reconnaissance du sens, il s'agit alors d'un test de reconnaissance de la forme (*form recognition*), ce qui est moins pertinent pour les activités de réception où la forme est donnée. Ce n'est donc pas un format que nous retiendrons.

7.2.3. Test d'appariement multiple

Le *Vocabulary Levels Test* est décrit en 1996 par Paul Meara comme *« the nearest thing we have to a standard test in vocabulary »* (Meara, 1996, p. 38). Développé par Paul Nation (1990) comme le *VST*, ce test a ensuite été modifié par Beglar et Hunt (1999) puis Schmitt et al. (2001). Les versions révisées suivent le format original, celui d'un test d'appariement multiple, avec six items lexicaux à gauche, et trois définitions correspondant à trois de ces mots à droite, comme on peut le voir dans le Tableau 7.4.

(a)	1. blame	
	2. hide	___ keep out of sight
	3. hit	___ have a bad effect
	4. invite	___ ask
	5. pour	
	6. spoil	

Tableau 7.4 - exemple d'item du Vocabulary Levels Test de Nation (1990)

Ce format inhabituel est un peu compliqué à implémenter sur ordinateur, et ne nous paraît pas avoir d'avantages particuliers par rapport à un QCM simple (en particulier, le nombre de distracteurs change au cours de la tâche, puisque le candidat élimine au fur et à mesure les distracteurs ; les trois items ne sont donc pas totalement indépendants). D'ailleurs, une

nouvelle version du VLT sous format QCM a été proposée récemment (McLean & Kramer, 2015) pour essayer de pallier ces problèmes.

7.2.4. Test « oui/non » ou « liste à cocher » (*checklist test*)

7.2.4.1. version écrite

Le format le plus simple est probablement le test « oui/non », qui est un test de reconnaissance du vocabulaire (*form recognition*): les candidats doivent indiquer quels mots ils (re)connaissent, sans avoir besoin de montrer qu'ils en connaissent le sens. Dans la terminologie de Read (1993), il s'agit d'un test « non vérifiable », parce que basé sur les déclarations personnelles des candidats (*self-report*). Ce type de test a été utilisé depuis les premières études sur la taille du vocabulaire en L1, dès le début du 20ème siècle (par exemple, le *English Vocabulary Test* de Starch cité par Sims, 1929, p.92). L'inconvénient de ce test, reconnu rapidement (Sims le compare défavorablement au QCM, à l'appariement ou à la production du sens), est justement le côté invérifiable, qui peut poser un problème quand les apprenants ont tendance, consciemment ou non, à cocher des mots qu'ils ne connaissent en fait pas. A partir de ce constat, la pratique a été de rajouter des pseudomots pour essayer d'évaluer la tendance du candidat à surévaluer ses connaissances, et d'ajuster le score en conséquence. C'est Paul Meara (Meara, 1996; Meara & Jones, 1988) qui a popularisé ce format de test, dont une version a aussi été utilisée pour le test de niveau initial du test en ligne *Dialang* décrit en première partie (chapitre 3). Meara souligne que les avantages de ce format sont, entre autres, sa rapidité (ce qui est important pour un contexte diagnostique tel que le nôtre), et le fait qu'il n'est pas parasité par d'autres compétences. Dans un QCM, au contraire, il faut en général lire les définitions, ce qui met en jeu la compétence de compréhension de l'écrit, et pose la question du rôle des termes utilisés dans les propositions de réponse.

Cependant, une fois qu'on ajoute des pseudomots à ce type de test, la question est de savoir comment prendre en compte les « fausses alarmes » (*false alarms*, terme venu de la théorie de détection du signal), c'est-à-dire les pseudomots qui sont reconnus à tort comme de vrais mots par les participants. Différentes formules ont été proposées (Beeckmans et al., 2001), mais celle qui fonctionne le mieux semble être la plus simple, c'est-à-dire la moyenne du pourcentage de bonnes réponses aux mots et aux pseudomots (Lemhöfer & Broersma, 2012).

Ce format, du fait de sa rapidité et de sa simplicité, nous paraît tout à fait indiqué pour notre étude, même s'il faut rester conscient des inconvénients potentiels en termes de différence de stratégies de réponse entre étudiants qui ont tendance à surévaluer leurs connaissances et étudiants qui en ont une vision plus lucide.

7.2.4.2. version aurale

Nous avons trouvé un seul test de type oui/non où les mots sont présentés sous forme orale et non écrite. Il s'agit de *Aural Lex* (Milton & Hopkins, 2006), décrit comme durant de 10 à 15 minutes en moyenne et étant composé de 120 mots (dont 20 pseudomots). Les 100 « vrais » mots forment 5 groupes de 20 mots, correspondant chacun à une bande de fréquence lexicale (de la bande 0-1 000 à 4 000 -5 000). Cela permet ainsi d'inférer la proportion des 5 000 mots les plus courants connue de chaque sujet. Nous avons réussi à obtenir l'accès à ce test non publié[38], mais trop tard pour l'utiliser pour cette étude. Nous avions entre temps créé un autre test (moins long, mais sur des principes similaires), dont nous décrirons la conception un peu plus loin.

Nous mentionnerons ici encore une fois (voir la section 5.2.3 pour une présentation plus approfondie) le format de décision lexicale avec mesure du temps de réaction, parce que certains chercheurs incluent dans le construit de compétence lexicale la rapidité de traitement ou *fluency* (Laufer & Nation, 2001). Partant de ce constat, d'autres chercheurs ont ajouté à un test de vocabulaire oui/non la prise en compte de la vitesse. Pellicer-Sanchez et Schmitt (2012), par exemple, montrent que les mots acceptés comme tels mais dont le sens n'est en fait pas connu par les sujets (ce dernier point étant vérifié à l'aide d'un entretien suite au passage du test) correspondent à un temps de réponse plus long de la part des sujets (de l'ordre de 50% en moyenne). Cela pourrait donc constituer une autre façon de vérifier l'exactitude des réponses des candidats, sans avoir besoin d'ajouter de pseudomots. Cependant, comme nous l'avons remarqué en introduction à la deuxième partie (section 4.3), nous ne sommes pas en mesure de prendre en compte de façon fine le temps de réaction dans nos expérimentations.

7.2.5. Autres formats : dictée

Fountain et Nation (2000, cité aussi dans Nation, 1990, p.86, avec la date 1974) proposent d'utiliser comme test de positionnement une dictée qu'ils considèrent comme un test de

[38] Nous remercions James Milton d'avoir généreusement accepté de le partager avec nous.

vocabulaire en anglais. Le texte dicté est composé de mots progressivement moins fréquents et de groupes de souffle progressivement plus longs. Même si la présentation orale est intéressante, et le temps d'administration assez court (dix minutes environ), la validité de ce format comme test de vocabulaire ne nous paraît pas établie : en effet, la dictée suppose, en plus des connaissances lexicales (même conçues de façon étendue comme chez Nation), la capacité de segmenter l'énoncé, et de le maintenir en mémoire en écrivant. De plus, elle comporte des connaissances qui sont peu pertinentes dans notre contexte, à savoir celle de l'orthographe des mots (pertinente, bien sûr, en production de l'écrit, mais non pas en compréhension de l'oral). En bref, c'est un type de test « intégratif » (*integrative*, qui s'oppose à *discrete point*, J. W. Oller & Conrad, 1971), c'est-à-dire qu'il requiert l'intégration de plusieurs sous-habiletés, contrairement à des tests qui sont focalisés sur une sous-habileté particulière (comme la compétence lexicale ou grammaticale). C'est un format qui est donc peu adapté à un test diagnostique, où l'on a au contraire besoin de tests ciblés afin d'identifier le plus précisément possible la source des difficultés des apprenants.

7.3. Construction du test

7.3.1. Choix des stimuli et des items

Ayant étudié ces différents formats, nous avons décidé d'utiliser un test oui/non qui nous permet de tester en un temps restreint un nombre de mots plus important qu'un QCM. Nous avons choisi d'adapter le test Lextale (Lemhöfer & Broersma, 2012) à l'oral. Lextale est un test de reconnaissance lexicale écrite créé pour servir de test de positionnement rapide (de l'ordre de cinq minutes) pour des expériences de psycholinguistique en langue étrangère ou seconde, plus fiable que l'autoévaluation souvent demandée aux participants dans le même but, et beaucoup moins lourd à organiser qu'un test de compétence générale. C'est au départ un test oui/non de reconnaissance écrite du vocabulaire, comprenant 40 mots et 20 pseudo-mots (plus trois mots d'entraînement pour commencer : deux « vrais » mots et un pseudo-mot). Ces 60 mots sont tirés d'un test précédent de Paul Meara, non publié mais mentionné dans certains de ses articles (Meara & Jones, 1988), appelé *10k*, qui évalue la taille du vocabulaire des apprenants sur les 10 000 familles de mots les plus fréquentes de l'anglais. Le test original comprenait 240 items, dont le quart qui fonctionnait le mieux a été conservé pour Lextale (en essayant de garder des mots de différentes fréquences). Les items du test sont présentés dans le Tableau 7.5, accompagnés de la bande de fréquence de leur famille de mots, calculée avec le site *Lextutor* (Cobb, s. d.).

Etant donné que le nombre d'items dans les premières bandes de fréquence (mots très courants) nous a paru insuffisant, nous avons enlevé cinq mots moins fréquents, que nous avons représentés par des mots barrés dans le tableau. Parmi eux, deux ont une prononciation qui peut les rendre ambigus pour des apprenants : *flaw* peut être confondu avec *floor*, dont la prononciation est identique en anglais britannique (/flɔ:/), et *wrought*, dont la prononciation est différente en anglais britannique (/rɔ:t/) et en anglais américain (/rɑ:t/), peut facilement être confondu avec *wrote*, (/rəʊt/ en anglais britannique, /roʊt/ en américain) et fausser ainsi les résultats du test. Si les étudiants déclarent connaître ce mot, il serait tout à fait possible qu'ils aient cru reconnaître *wrote* et non *wrought*. Nous avons également écarté *recipient*, qui est un faux-ami (un mot transparent qui a un sens différent en anglais et en français, ce qui fait que les étudiants qui en reconnaissent la forme n'en connaissent pas forcément le sens), *cylinder*, et *savoury*, qui a une fréquence très différente en anglais britannique (où il a une fréquence moyenne, puisqu'il est dans la septième bande de notre tableau 28) et en anglais américain (où il est très peu utilisé : dans la liste tirée de $SUBTLEX_{US}$, par exemple, il est au 45 000ème rang). Nous avons remplacé ces cinq mots par des mots plus courants, appartenant aux trois premières bandes de fréquence (qui nous intéressent pour repérer les étudiants en difficulté), et identifiés dans l'étude de Hilton (2003) comme étant des mots transparents pour des francophones à l'écrit mais non reconnus à l'oral : *angle, creature, issue, muscle* et *theory*. Nous avons gardé quelques mots peu fréquents (bandes de fréquence 10 et 11) afin de pouvoir éventuellement distinguer les étudiants de niveau B2 de ceux de niveau C1.

bande de fréquence	**mots**
1	cleanliness **issue**
2	breeding **creature** denial lengthy muddy **muscle** moonlit
3	allied **angle** scholar **theory**
4	censorship dispatch fluid *generic* hasty hurricane ~~flaw~~ ~~recipient~~
5	carbohydrate festivity turtle ~~cylinder~~
6	eloquence fray nourishment scornful screech slain stoutly turmoil
7	ingenious lofty majestic ~~savoury~~ ~~wrought~~
8	celestial shin
9	
10	ablaze awry bewitch listless plaintively upkeep
11	rascal unkempt
pseudomots	~~abergy~~ alberation crumper destription exprate fellick interfate kermshaw kilp magrity mensible *platery* plaudate proom pudour pulsh purrage quirty rebondicate skave spaunch **tailent**

Tableau 7.5 - items du test de reconnaissance du lexique aural, adapté du test LexTALE (Lemhöfer & Broersma, 2012), par bande de fréquence (les mots barrés du test LexTALE ont été remplacés par les mots en gras, et les mots en italique correspondent aux items d'entraînement)

7.3.2. Administration du test

Comme les autres tests diagnostiques du projet, le test de reconnaissance aurale du vocabulaire a été administré via la plateforme SELF. Une copie de l'écran d'accueil est reproduite en Figure 7.1 - écran d'accueil du test de reconnaissance aurale du vocabulaire (SELF)Figure 7.1.

Figure 7.1 - écran d'accueil du test de reconnaissance aurale du vocabulaire (SELF)

Tous les items ont le même écran d'interface, reproduit en Figure 7.2. La consigne choisie pour le test oui/ non est « *Do you know this word ?* ». En effet, il nous a semblé qu'une question sur la connaissance personnelle du mot serait moins à même d'entraîner de fausses alarmes qu'une question sur l'existence du mot en anglais (« *Does this word exist in English ?* »), qui pourrait être interprétée comme une question sur la conformité des pseudomots aux règles orthographiques, phonologiques ou morphologiques de l'anglais.

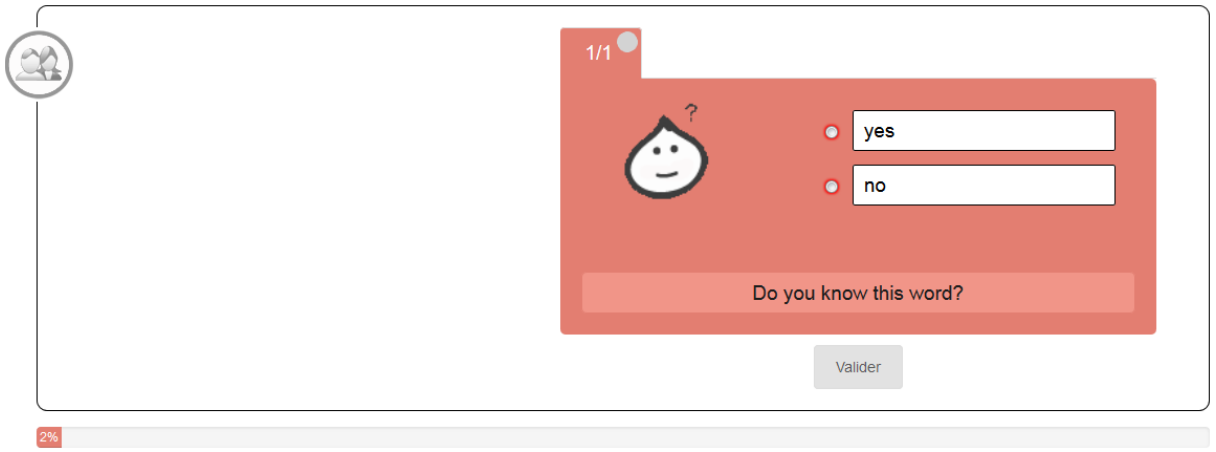


Figure 7.2 – écran d'administration du test de reconnaissance du vocabulaire aural

7.4. Résultats

7.4.1. Statistiques descriptives globales

En moyenne, les étudiants ont eu besoin de 7,8 minutes (écart-type 3) pour passer ce test, ce qui est satisfaisant puisque en dessous de notre limite de 10 minutes (même si le plus lent a mis 25 minutes).

Les résultats du test, passé par 183 étudiants, sont présentés dans le Tableau 7.6 ci-dessous. La moyenne obtenue aux 64 items du test (notre test a quatre items de plus que le test original parce que nous n'avons gardé que deux items d'entraînement au lieu de trois et utilisé le troisième comme item du test, que nous avons rajouté un pseudomot, et que deux des items que nous voulions enlever ont été gardés par erreur), n'est pas très élevée : 38,67/64, soit 60%.

n	moyenne	écart-type	médiane	min	max	étendue	asymétrie	kurtose
183	38.67	6.61	38	25	63	38	0.59	0.5

Tableau 7.6 - statistiques descriptives du score total au test de reconnaissance lexicale

Nous observons par ailleurs que le coefficient d'asymétrie est positif, ce qui confirme l'impression de difficulté du test : il distingue bien les candidats assez forts (étalement des scores vers la droite, comme on peut le constater sur l'histogramme de fréquence en Figure 7.3), mais est peut-être insuffisant pour bien distinguer les candidats plus faibles (peu d'étalement dans les scores faibles, contrairement à ce dont nous aurions besoin). Un test de Shapiro-Wilk nous confirme que la distribution n'est pas normale ($W = 0,98$, $p < 0,01$), ce qui nous interdit d'utiliser des tests paramétriques pour la suite de l'analyse de ce test dans son état actuel.

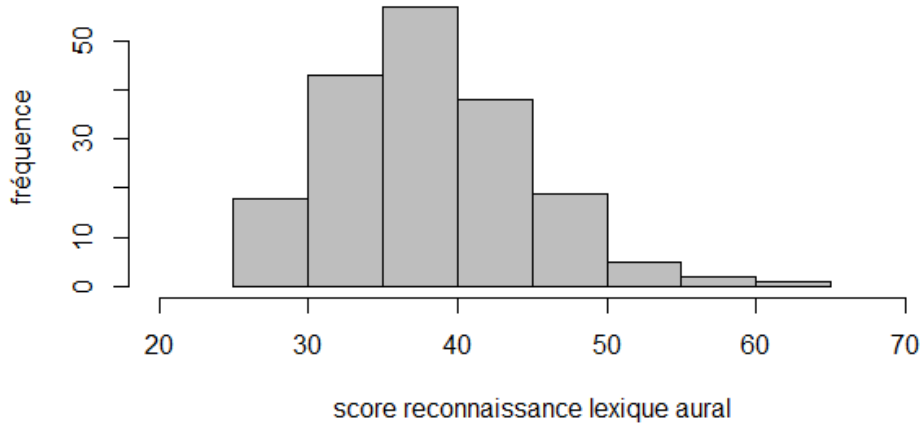


Figure 7.3 - histogramme du score global au test de reconnaissance du lexique aural

La fiabilité, estimée avec le coefficient alpha de Cronbach (0,72), est suffisante, même avant suppression des items insatisfaisants.

7.4.2. Analyse des items

Nous allons examiner pour ce test la difficulté de chaque item, qui doit être comprise entre 0,2 et 0,9 environ, et son indice de discrimination, qui doit être supérieur à 0,2. Afin de faciliter la lecture du Tableau 7.7, nous présentons les résultats des analyses sur deux ensembles de colonnes, à gauche celles correspondant aux 43 mots et à droite aux 21 pseudomots. Les items sont classés par ordre alphabétique dans les deux cas. On peut tout de suite constater que les vrais mots fonctionnent beaucoup mieux que les pseudomots. Seuls 7 mots sur 43 ont un indice de discrimination inacceptable (*ablaze*, *ingenious* et *awry*, ainsi que *savoury* qui a été inclus par erreur, ont en particulier un indice très mauvais, inférieur à 0,1). Inversement, seuls 5 pseudomots sur 21 ont un indice totalement acceptable (*destription*, *exprate*, *kermshaw*, *mensible* et *rebondicate*).

Si l'on regarde les vrais mots de façon plus approfondie, trois sont très faciles, avec un indice de facilité supérieur à 0,9 (*creature, ingenious* et *turtle*), dont deux qui ont un indice de discrimination acceptable et peuvent donc être gardés tout de même afin de nous aider à repérer les apprenants en grande difficulté. Le troisième (*ingenious*), avec un indice de discrimination de 0,06, ne semble pas être mieux reconnu par les étudiants forts que faibles et ne nous aide pas à identifier nos sujets à faible vocabulaire aural. Trois autres mots ont un indice de discrimination compris entre 0,1 et 0,2 : *listless, scornful* et *stoutly*. Etant donné qu'il s'agit dans les trois cas de mots difficiles (bande de fréquence 10 pour *listless*, comme on l'a vu dans le tableau 28, et bande 6 pour *scornful* et *stoutly*, qui sont reconnus seulement par 22 et 37% de sujets), il est inutile de les garder pour notre test. Il nous resterait ainsi 36 items pour les « vrais » mots (43 moins 7).

mots				**pseudomots**		
nom item	**discrim.**	**difficulté**		**nom item**	**discrim.**	**difficulté**
ablaze	**0.09**	0.55		affordish	**-0.08**	0.68
allied	0.54	0.40		alberation	**0.18**	0.34
angle	0.25	0.83		crumper	**0.07**	0.70
awry	**-0.09**	0.49		destription	0.35	0.30
bewitch	0.29	0.59		exprate	0.28	0.56
breeding	0.34	0.82		fellick	**-0.16**	0.61
carbohydrate	0.29	0.38		interfate	**-0.17**	0.59
celestial	0.47	0.39		kermshaw	0.20	0.72
censorship	0.48	0.69		kilp	**0.13**	0.81
cleanliness	0.36	0.32		magrity	**0.03**	0.81
creature	0.26	**0.92**		mensible	0.20	0.76

denial	0.54	0.58	plaudate	**0.17**	0.83
dispatch	0.27	0.87	proom	**-0.03**	0.75
eloquence	0.33	0.68	pudour	**-0.10**	0.44
festivity	0.28	0.85	pulsh	**0.09**	0.81
fluid	0.24	0.88	purrage	**-0.10**	0.23
fray	0.21	0.67	quirty	**0.05**	0.43
hasty	0.30	0.51	rebondicate	0.22	0.68
hurricane	0.38	0.86	skave	**0.05**	0.49
ingenious	**0.06**	**0.93**	spaunch	**0.08**	0.63
issue	0.45	0.83	tailent	**0.10**	0.63
lengthy	0.50	0.37			
listless	**0.12**	0.51			
lofty	0.31	0.37			
majestic	0.44	0.83			
moonlit	0.43	0.31			
muddy	0.45	0.62			
muscle	0.45	0.72			
nourishment	0.46	0.60			
plaintively	0.34	0.46			
rascal	0.42	0.39			
savoury	**0.03**	0.83			
scholar	0.21	0.72			
scornful	**0.14**	0.22			
screech	0.45	0.37			
shin	0.29	0.49			
slain	0.32	0.61			
stoutly	**0.19**	0.37			
theory	0.26	0.87			
turmoil	0.30	0.38			
turtle	0.20	**0.95**			
unkempt	0.21	0.48			
upkeep	0.39	0.35			
					Cronbach's α = 0.72

Tableau 7.7 - résultats de l'analyse des items du test de reconnaissance aurale du lexique, mots à gauche et pseudomots à droite : les items dont le nom est grisé ont de mauvais indices de difficulté (en gras, items très faciles, *p-value* > .9) ou de discrimination (en gras et surlignés en gris, items peu discriminants, indice <.2))

Comme nous l'avons dit, les pseudomots sont très peu discriminants, et six d'entre eux ont même un coefficient de discrimination négatif, ce qui signifie que les candidats forts les réussissent moins bien que les faibles, c'est-à-dire qu'ils les reconnaissent comme des mots alors qu'ils n'existent pas (il semble donc que, malgré nos précautions sur la formulation de la consigne, qui demandait aux étudiants s'ils connaissaient les mots présentés, certains ont pu l'interpréter comme : « Est-ce que ce mot ressemble à un mot anglais ? »). Nous retrouverons

ce phénomène au moment de l'analyse du test de jugement de grammaticalité aurale, et l'analyserons plus en détail à ce moment-là.

Nous pouvons conclure provisoirement que le test final après analyse des items comprendra 36 items pour les « vrais » mots et 5 seulement pour les pseudomots, soit 41 items en tout, ce qui fait 23 de moins qu'au départ. Le test ainsi écourté, malgré un nombre d'item moindre, a une meilleure fiabilité, puisque l'alpha de Cronbach passe à 0,84, ce qui est un très bon résultat Nous recalculons le score total de notre cohorte dans la nouvelle version du test, et représentons sa distribution avec un histogramme reproduit en Figure 7.4 :

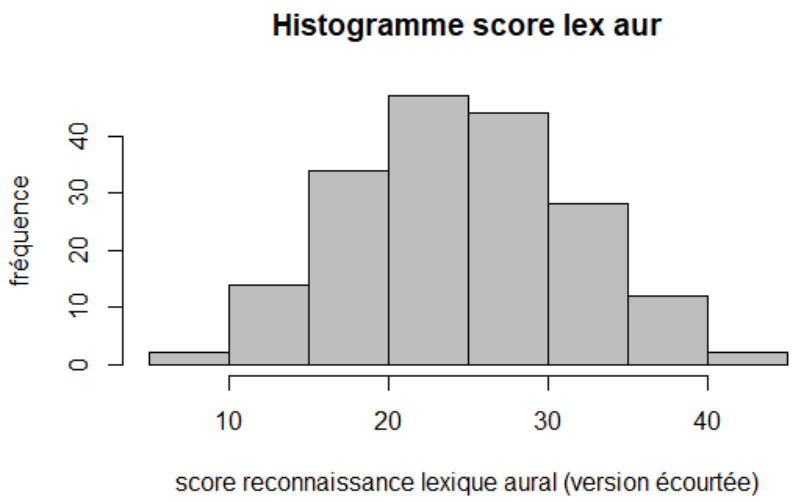


Figure 7.4 - histogramme de fréquence du score total à la version finale du test de reconnaissance du lexique aural

Nous présentons également les statistiques descriptives de la version écourtée dans le Tableau 7.8. Nous constatons que le score moyen a peu bougé (il est maintenant à 61% au lieu de 60, soit 24,96/41), mais l'étendue s'est proportionnellement accrue, passant de 36 pour 64 items à 32 pour 41 items ; surtout, l'asymétrie en faveur des scores élevés a presque disparu.

n	moyenne	écart-type	médiane	min	max	étendue	asymétrie	kurtose
183	24.96	6.77	25	9	41	32	0.07	-0.49

Tableau 7.8 - statistiques descriptives du score total au test de reconnaissance lexicale après suppression des items qui dysfonctionnent

Nous vérifions d'ailleurs que la distribution des scores ne diffère plus de la normale avec un test de Shapiro-Wilk (W = 0,99, p-value = 0,35). Nous pourrons donc après tout utiliser des tests paramétriques si besoin est.

Pour évaluer l'unidimensionnalité du test final, nous utilisons une analyse en composants principaux, dont le but est d'identifier un facteur principal qui synthétise le maximum

d'informations apportées par les différents items du test. L'inspection visuelle du cercle des corrélations de toutes les variables avec l'axe principal de l'ACP (Figure 7.5) nous permet de constater que tous les items du test corrèlent de façon positive avec le facteur principal identifié (même si certains sont beaucoup plus corrélés avec l'axe vertical). De plus, ce facteur explique environ 15% de la variance du test (comme indiqué dans la parenthèse en bas), ce qui est un résultat acceptable.

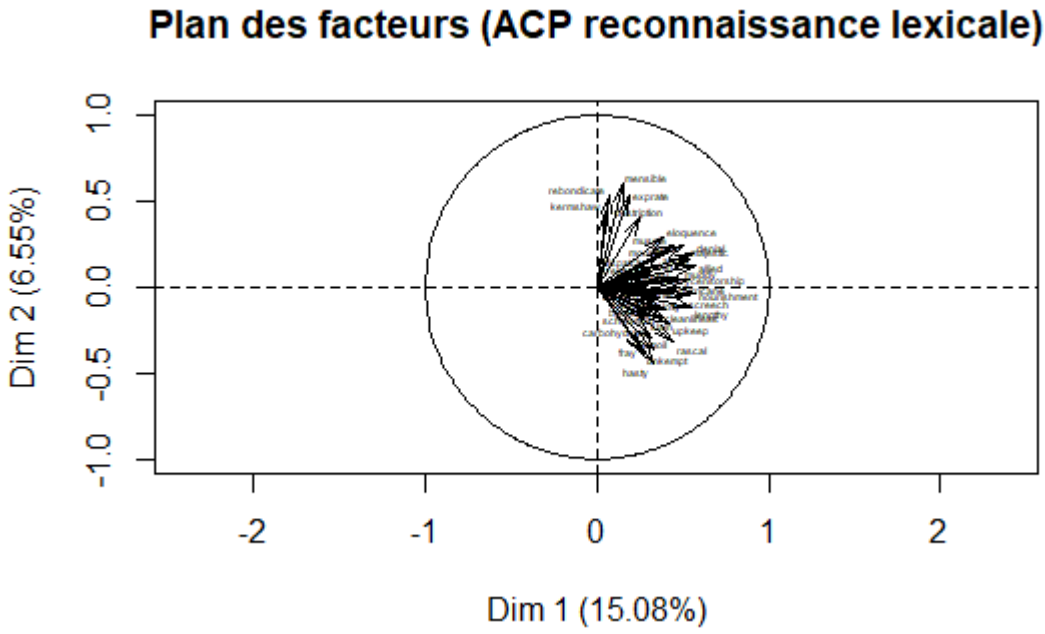


Figure 7.5 - ACP reconnaissance aurale du lexique

7.4.3. Analyse approfondie

7.4.3.1. influence de la fréquence

Nous n'avons pas réussi à détecter une influence de la fréquence lexicale dans le test de sensibilité prosodique, absence que nous avons interprétée comme une incapacité de nos apprenants à mémoriser le patron accentuel des mots qu'ils connaissent. Cependant, nous nous attendons naturellement à ce que la fréquence lexicale joue plus directement dans un test de reconnaissance du vocabulaire.

Nous avons construit le test par rapport aux bandes de fréquence des familles de mots de Nation (2017). Il est donc logique de tester si l'appartenance à une bande plus fréquente est corrélée à une meilleure reconnaissance (indice de difficulté/ facilité plus élevé). Nous avons choisi de diviser les bandes de fréquence en trois groupes. Dans le premier groupe se trouvent les mots appartenant à une famille des trois premières bandes. Ces mots très fréquents font

partie des 3 000 familles de mots dont nous avons montré dans le deuxième chapitre qu'elles correspondent à une couverture de 95% à l'oral, et sont indispensables pour une compréhension acceptable. Le deuxième groupe contient les mots de moyenne fréquence des bandes 4 à 6. Une étendue de 6 000 familles de mots correspond en effet à l'oral à une couverture de 98%, associée à une bonne compréhension. Enfin, les mots de basse fréquence appartiennent aux bandes 7 à 11. Cette division en trois groupes peut être visualisée dans le Tableau 7.9 ci-dessous.

bande de fréquence (famille de mots)	mots	nb mots	difficulté moyenne
1, 2, 3	allied angle breeding cleanliness creature denial issue lengthy moonlit muddy muscle scholar theory	13	0.64
4, 5, 6	carbohydrate censorship dispatch eloquence festivity fluid fray hasty hurricane nourishment scornful screech slain stoutly turmoil turtle	16	0.62
7, 8, 9, 10, 11	ablaze awry bewitch celestial ingenious listless lofty majestic plaintively rascal savoury shin unkempt upkeep	14	0.55

Tableau 7.9 - division des items du test de reconnaissance aurale en 3 groupes selon les bandes de fréquence de Nation (2017), avec nombre de mots par groupe et difficulté moyenne pour chaque groupe

Nous avons rajouté à droite du tableau une colonne avec la réussite moyenne aux items correspondant aux mots de chaque groupe. On constate que la difficulté augmente à mesure que la fréquence diminue. Pour savoir si cette tendance est significative, nous avons besoin de tester l'hypothèse de non-égalité des trois moyennes. Comme nous avons plus de deux moyennes à comparer, que n'avons pas beaucoup d'items dans chaque groupe et que la distribution des scores à l'intérieur n'est pas normale, nous utilisons un test de Kruskal-Wallis. Ce test montre que la différence entre les moyennes des trois groupes (mots très fréquents, moyennement fréquents et rares) n'est pas significative : Kruskal-Wallis χ^2= 1,25, df =2, p = 0,54.

On peut tenter d'expliquer cette absence surprenante de rôle de la fréquence (ou du moins un rôle trop subtil pour être détecté avec le peu d'items dont nous disposons) de deux façons différentes : l'une est que la présence de mots transparents (*cognates*) dans chacun des groupes brouille l'influence de la fréquence : les étudiants peuvent reconnaître les mots entendus non parce qu'ils sont fréquents, mais parce qu'ils ont réussi à les identifier grâce à leur connaissance du lexique français (malgré les différences phonologiques). Nous examinerons cette hypothèse au paragraphe suivant. L'autre possibilité est que les bandes de fréquence des familles de mots ne correspondent pas bien à la réalité de l'input anglais reçu par nos étudiants, et que, comme l'ont souligné des chercheurs que nous avons cités dans la

première partie (par ex., Ward & Chuenjundaeng, 2009), la famille de mots n'est pas une unité adéquate pour rendre compte des connaissances des apprenants de niveau faible. On peut imaginer, par exemple, que la plupart de nos étudiants auraient reconnu le mot *clean* sous la forme /kli:n/, alors que peu d'entre eux (32%) ont déclaré connaître /ˈklɛn li nəs/, *cleanliness*. Les deux mots appartiennent pourtant à la même famille (et qui plus est, à la bande 1, la plus fréquente), mais la présence de suffixes et l'alternance vocalique dans la racine[39] empêchent les francophones de reconnaître *clean* dans *cleanliness*.

Partant de ce constat, on peut essayer de tester le rôle de la fréquence avec un instrument plus précis que la famille de mots. Comme nous l'avons expliqué dans la première partie, nous disposons d'une liste (non lemmatisée) de 60 000 mots tirés du corpus $SUBTLEX_{US}$ (Brysbaert & New, 2009), classés selon une mesure de diversité textuelle (nombre de textes de leur corpus dans lesquels ils sont présents). Nous présentons dans le Tableau 7.10 une nouvelle division de nos items en trois groupes selon cette nouvelle mesure. Comme nous ne disposons pas de lignes directrices pour les rangs où s'arrêtent les mots très fréquents, moyennement fréquents ou rares quand ils ne sont pas classés par familles de mots, nous avons choisi des bornes arbitraires qui scindent notre petit corpus en trois parties grossièrement égales : nous avons considéré que les mots très fréquents appartiennent aux 10 000 premiers mots du corpus, que les mots moyennement fréquents sont entre le 10 001ème et le 20 000ème rang, et que les mots rares se trouvent au-delà.. On peut considérer que 10 000 mots correspondent à un peu moins de 3 000 familles : le corpus de Nation composé de 14 000 familles de mots contient 62 000 formes (Martinez, 2011, p.75), c'est-à-dire un peu plus de 4 formes par famille de mots. 3 000 familles de mots équivalent dans ce cas-là à 13 000 formes environ, et 10 000 formes à un peu plus de 2 300 familles.

groupe de fréquence (mot non lemmatisé)	mots	nb mots	difficulté moyenne
0 - 10 000 premiers mots	angle breeding creature denial dispatch fluid hasty hurricane ingenious issue muddy muscle rascal scholar theory turtle	16	0,77
10 001 – 20 000	allied celestial censorship fray lengthy lofty majestic nourishment shin slain turmoil	12	0,51
20 001 et +	ablaze awry carbohydrate cleanliness bewitch eloquence festivity listless moonlit plaintively savoury scornful stoutly unkempt upkeep	15	0,49

Tableau 7.10 - items du test de reconnaissance aurale groupés par fréquence selon le corpus SUBTLEXUS (Brysbaert & New, 2009), avec nombre de mots et difficulté moyenne par groupe

[39] /i:/ dans l'adjectif et /ɛ/ dans le nom dérivé, parce que le nom dérivé n'a pas subi le grand changement vocalique qui a transformé la plupart des voyelles longues aux 14ème et 15ème siècles

On s'aperçoit que certains mots ont changé complètement de groupe de fréquence, passant de très fréquent à rare (*cleanliness, moonlit*), alors que d'autres passent au contraire de rare à très fréquent (*rascal*). La difficulté moyenne des items de chaque groupe est indiquée dans la dernière colonne du tableau. On observe, comme tout à l'heure, une augmentation de la difficulté (moyenne de réussite plus basse) à mesure que la fréquence diminue. La différence de moyenne est cette fois significative entre les mots des trois groupes (Kruskal-Wallis $\chi^2 = 16.5$, df =2, $p < 0.001$). Le test de Kruskal-Wallis ne nous dit pas exactement quels groupes sont différents. Au vu des moyennes, on peut supposer que le groupe d'items très fréquents sera différent des deux autres moins fréquents, mais que la différence entre ces deux derniers ne sera pas significative. C'est ce qu'on vérifie avec un test de Wilcoxon 2 à 2 (avec une correction Bonferroni pour le fait qu'on opère de multiples tests sur les mêmes données) : le groupe très fréquent est significativement différent du groupe moyennement fréquent ($p<0.01$) et du groupe rare ($p < 0.01$), mais ces derniers ne sont pas différents l'un de l'autre.

Nous avons donc réussi à observer l'effet de la fréquence lexicale en utilisant une liste de fréquence dont l'unité est le mot, et non le lemme ou la famille. Cela va dans le sens de la mise en cause actuelle de l'utilité (et de la réalité) de la famille de mot comme unité lexicale, à la fois pour des apprenants comme les nôtres qui ne disposent pas de connaissances morphologiques étendues (McLean, 2018), mais également pour les natifs (Brysbaert & New, 2009). Les compétences en morphologie dérivationnelle (avec toutes les complications phonologiques qui y sont associées en anglais) ne sont pas forcément répandues chez tous les locuteurs, et surtout ces connaissances n'excluent pas l'utilisation d'unités moins abstraites en parallèle, à savoir le lemme, ou même le mot non lemmatisé. Par ailleurs, le corpus dont est tiré cette deuxième liste de fréquence est un corpus oral de sous-titres de films et séries télévisées américaines, qui reflète probablement mieux l'input authentique reçu par les étudiants en dehors des cours, ce qui explique peut-être la meilleure corrélation qu'avec les données de Nation (2017) tirées de corpus plus équilibrés entre écrit et oral, et entre textes formels et informels.

7.4.3.2. influence de la transparence

Poursuivons maintenant l'analyse du test en nous penchant sur les réponses aux mots transparents. Nombre de chercheurs (Cobb, 2000; Elgort, 2012; Lemhöfer et al., 2008; Meara, 1996) ont montré le rôle facilitateur des mots transparents (*cognates*) dans la reconnaissance du vocabulaire écrit d'une langue étrangère. Nous avons trouvé une seule étude qui mette en

lumière ce même effet dans un test de vocabulaire aural anglais (le *Peabody Picture Vocabulary Test*), avec des enfants néerlandophones (De Wilde & Eyckmans, 2017). Il est donc intéressant d'expliciter cet effet pour notre test de reconnaissance aurale avec une population de jeunes adultes, et entre l'anglais et le français.

Nous avons choisi une définition assez large des mots transparents. Comme Elgort, nous considérons que sont transparents tous mots similaires dans les deux langues du point de vue de la forme et du sens, que ce soit parce qu'ils ont une étymologie commune ou parce qu'ils sont le résultat d'un emprunt dans un sens ou dans l'autre : « *loan words or lexical borrowings are classified as cognates, alongside words that have a common ancestor, as long as they match the criterion of having similar form and meaning across the L1 and L2* » (Elgort, 2012, p. 255). Inversement, certains mots de racine commune ne sont plus transparents parce que leurs sens ont trop divergé (*fray* / « frayer »). La liste des mots anglais classés par statut (transparent ou non) est disponible dans le Tableau 7.11. La majorité de ces mots a été emprunté par l'anglais au français, mais certains ont fait l'objet d'un emprunt dans l'autre sens (*dispatch* -> « dispatcher »), et d'autres forment des paires anglais / français dont les deux membres viennent du latin (*scholar* / scolaire).

statut	**mots**	**nb mots**	**difficulté moyenne**
mots transparents	allied (.619) angle (1/ .794) celestial censorship creature (.875/ .812) denial dispatch eloquence festivity fluid ingenious issue (1/ .833) majestic muscle (1/ .771) nourishment plaintively rascal scholar theory (.571/ .718) turtle	20	0.72
mots non transparents	ablaze awry bewitch breeding carbohydrate cleanliness fray hasty hurricane lengthy listless lofty moonlit muddy savoury scornful screech shin slain stoutly turmoil unkempt upkeep	23	0.50

Tableau 7.11 - classement des mots du test de reconnaissance aurale selon leur statut de mot transparent ou non, avec entre parenthèses le coefficient de distance (orthographique et phonétique) de Schepens et al. (2013) pour les mots transparents les plus fréquents

Cependant, il n'est pas toujours facile de décider si deux mots sont transparents d'une langue à l'autre, puisque, au moins à l'oral, la transparence n'est jamais totale, dans la mesure où les systèmes phonologiques des deux langues sont différents. Pour nous aider, nous avons utilisé la liste de Schepens et collègues (Schepens et al., 2013), qui ont testé une méthode automatique pour déterminer la distance orthographique et phonétique entre des équivalents sémantiques de diverses langues européennes, dont l'anglais et le français. Comme ils n'ont testé que des mots très fréquents, il nous manque certains chiffres, mais leur étude nous

permet par exemple de justifier l'inclusion du mot *issue* parmi les mots transparents, du fait du recouvrement partiel des sens dans les deux langues, même si son sens principal est différent en anglais et en français. Nous avons ajouté dans le Tableau 7.11, aux mots pour lesquels nous disposons de ces informations, le coefficient de distance (orthographique/ phonétique). Plus le chiffre est élevé, et plus les mots sont proches de leur équivalent français.

Nous voyons dans la dernière colonne du tableau que les mots transparents sont effectivement mieux reconnus en moyenne que les mots non transparents. Un test de Wilcoxon (comparaison de deux moyennes) nous confirme que cette différence est significative ($W = 377$, $p<0{,}001$).

7.4.3.3. mots vs. pseudomots

Penchons-nous à présent sur quelques pseudomots qui se sont révélés particulièrement difficiles pour nos étudiants. Les trois stimuli les plus difficiles à rejeter parmi les pseudo-mots, à savoir *purrage* (23% de réussite), *destription* (30%), et *alberation* (34%), ont été reconnus (à tort) comme des mots par une large majorité de nos apprenants. Nous pensons qu'il s'agit là d'un exemple de ce que Perfetti et Hart (2002) appelle la *lexical quality hypothesis*, une hypothèse qui a été développée pour l'écrit, mais qui nous semble pouvoir être étendue à la compréhension de l'oral. L'idée est que, pour une bonne compréhension (à l'écrit), la rapidité d'accès ne suffit pas, et qu'il faut également que l'information récupérée soit de bonne qualité, définie par Perfetti et son équipe comme la précision de l'orthographe, de la prononciation et du sens du mot reconnu. Diependale et ses collaborateurs ajoutent : « *L2 lexical representations will, on average, be of a lower precision than those in L1. For example, to an L2 speaker, the English words "squirrel" and "quarrel" may be more similar and thus more confusable than to a native speaker, who can quickly decide whether he is presented with the one or the other* » (Diependaele et al., 2013, p. 847).

Il nous semble que nous sommes ici devant le même phénomène: pour une majorité de nos étudiants, les pseudomots *purrage*, *destription* et *alberation* sont suffisamment proches de *porridge, description* et *aberration* pour être confondus avec eux. Il ne s'agit probablement pas d'une difficulté à distinguer les phonèmes (à part peut-être pour *purrage* et *porridge* qui ne diffèrent que par leurs voyelles) : *destription* et *description* se distinguent par le contraste /t/ - /k/ qui ne pose pas de problème en principe pour les francophones (même s'il peut être difficile à entendre sur un enregistrement). Quant à *alberation* et *aberration*, outre la différence de voyelle initiale qui peut être difficile à distinguer, ils diffèrent par la présence de

la consonne /l/ qui encore une fois ne devrait pas poser de problème. Il nous semble qu'un manque de précision de la représentation lexicale (ici au niveau aural) pourrait être à l'origine de cette difficulté à rejeter ces pseudomots.

Diependale et ses collaborateurs poursuivent en montrant que l'étendue du vocabulaire joue un rôle important dans la précision des représentations : « *basic individual differences in lexical processing […] can be attributed to a single causing factor—namely, vocabulary size (or lexical proficiency) in the target language* » (Diependaele et al., 2013, p. 858). L'étendue du vocabulaire devrait donc être corrélée à la capacité à rejeter les pseudomots proches des mots connus. Or nous avons vu que très peu de pseudomots ont un coefficient de discrimination acceptable, ce qui signifie que les étudiants avec un vocabulaire aural plus étendu ne savent pas forcément mieux rejeter les pseudomots. Même si elle est séduisante (en particulier pour *destription* qui est, lui, discriminant, avec un coefficient de 0,35), l'hypothèse de la qualité des représentations lexicales ne peut donc pas tout expliquer. Nous reviendrons d'ailleurs sur la difficulté chez nos apprenants à rejeter les formes non attestées dans le chapitre suivant sur l'analyse du test de jugement de grammaticalité.

7.4.4. Conclusion

Nous pouvons conclure provisoirement que le test de reconnaissance lexicale aurale fonctionne bien. Dans sa version écourtée, longue de 41 items (où il ne reste que six pseudo-mots), il a une bonne fiabilité ($\alpha = 0,84$), et discrimine bien entre les étudiants faibles et forts (mais il pourrait donner plus de détails sur les étudiants faibles s'il était un peu plus facile).

Nous avons par ailleurs obtenu deux éléments de validité supplémentaires en faveur de notre test. Premièrement, comme on pouvait s'y attendre à la lecture de la littérature, les mots transparents sont mieux reconnus que les mots opaques. Deuxièmement, les mots fréquents sont mieux reconnus que les mots rares, du moins quand la fréquence est opérationnalisée au niveau du mot et non de la famille de mots, qui est probablement une unité peu pertinente pour nos apprenants.

Enfin, nous avons observé que nos candidats, qu'ils soient forts ou faibles, ont tous beaucoup plus de mal à rejeter les pseudomots qu'à reconnaître les vrais mots. Ce phénomène pourrait s'expliquer en partie par l'hypothèse de la qualité lexicale (la qualité des représentations lexicales –en particulier aurales- de nos étudiants serait médiocre, sous-spécifiée, ce qui les empêcherait de bien distinguer deux mots proches par la forme), mais pas entièrement. Il

pourrait également s'agir d'un phénomène partiellement culturel : Eyckmans (2004) trouve aussi une proportion très importante de fausses alarmes dans un test oui/non de reconnaissance du vocabulaire (écrit) avec des apprenants francophones du néerlandais. Enfin, il peut paraître normal que des apprenants d'anglais langue étrangère soient sensibles à la possibilité que des mots avec une consonance anglaise (parce que conformes aux règles phonologiques de cette langue), et prononcés par un(e) anglophone, puissent effectivement faire partie de la langue, d'autant plus qu'ils savent qu'ils n'ont qu'une connaissance partielle du lexique anglais. Nous avions essayé d'éviter ce type de réaction avec une formulation portant sur leurs connaissances personnelles (« Connaissez-vous ce mot ? ») plutôt que sur le lexique anglais (« Ce mot existe-t-il ? »), mais cela n'a visiblement pas suffi.

Chapitre 8

Test de jugement aural de grammaticalité

8.1. Rappels du cadre théorique

Nous avons mentionné dans le chapitre 2 de la première partie (section 2.4) plusieurs caractéristiques souhaitables pour un test diagnostique des connaissances morphosyntaxiques qui reflète leur rôle dans la compréhension aurale. Nous voudrions tout d'abord que les structures étudiées soient contextualisées dans des phrases. Même s'il est possible d'isoler un sens (très abstrait) pour des éléments morphosyntaxiques isolés (*the*, *-s*, *might*), leur domaine d'application naturel est la phrase, d'où le besoin de contextualisation. Ce besoin est d'autant plus criant quand il s'agit d'éléments discontinus (*have -en*) ou de structures syntaxiques qui opèrent sur les constituants principaux de la phrase (par exemple la structure passive).

Il faut également que les stimuli soient oraux. En effet, une des difficultés du traitement syntaxique en anglais réside dans le fait que les mots fonctionnels et les suffixes grammaticaux sont désaccentuées et peuvent ainsi être difficiles à percevoir pour les apprenants étrangers. Il est donc important de tester leur capacité à repérer ces éléments grammaticaux de faible saillance dans le flux de la langue orale.

8.2. Inventaire d'instruments d'évaluation

Rod Ellis (1991) remarque que les premières études sur l'acquisition des langues étrangères ou secondes, dans les années 1970, utilisaient essentiellement des tests de production, mais que, à partir des années 1980, à la fois pour des raisons pratiques (facilité d'administration et de correction), et théoriques (tests plus contrôlés pour tester des hypothèses précises), le jugement de grammaticalité a pris une place prépondérante. Nous décrirons ces deux paradigmes tour à tour.

Etant donné nos contraintes techniques de passation, nous n'envisagerons pas les mesures en ligne telles que la lecture de phrases, grammaticales ou non, contrôlée par le sujet (*self-paced reading*), ou l'oculométrie pendant l'écoute avec un paradigme de monde visuel, utilisées par exemple par Suzuki et al. (2017). Ces tâches sont intéressantes dans la mesure où elles n'attirent pas l'attention du sujet sur la forme des phrases mais uniquement sur leur sens, tout en donnant des informations sur le traitement des structures étudiées grâce au temps de réaction. Cependant, elles ne sont pas compatibles avec notre cahier des charges (tests autocorrectifs sans suivi fin des temps de réaction).

8.2.1. Tests de production

Dans les années 1970, suite à l'intérêt soulevé par l'étude de Roger Brown (1973) sur le développement du langage enfantin et en particulier sur les séquences d'acquisition des morphèmes grammaticaux en L1, Heidi Dulay et Marina Burt (1974) ont développé un instrument de production orale suscitée par des images pour les apprenants L2, le *Bilingual Syntax Measure*. Cet instrument n'est pas autocorrectif et nécessite un intervenant qui pose des questions à propos des images et enregistre les réponses (qui doivent ensuite être analysées et évaluées). Nous ne pourrons donc pas l'utiliser, mais il nous semble important de le mentionner étant donné le retentissement de l'étude pour laquelle il a été créé (il a été cité plus de 2 000 fois depuis sa parution, d'après *Google Scholar*).

Les tests d'imitation (*elicited imitation*) sont également utilisés depuis les années 1960-1970 (Naiman, 1974). Il s'agit de demander aux sujets de répéter des phrases contenant les structures grammaticales étudiées. Du fait de la longueur ou de la complexité des phrases, il n'est pas possible de les répéter de mémoire sans les avoir d'abord comprises, puis reconstruites au moment de la répétition. C'est une tâche intéressante, toujours utilisée (par exemple, R. Ellis, 2005; Tracy-Ventura et al., 2014), mais qui n'est pas non plus envisageable dans notre contexte du fait que son analyse n'est pas automatisable pour l'instant.

Il existe également des formats plus contraints et qui peuvent être autocorrectifs, comme les textes à trous. Pour évaluer l'importance des connaissances grammaticales (et les comparer aux connaissances lexicales) dans la compréhension de l'oral et de l'écrit d'étudiants anglophones apprenant l'espagnol, Mecartty (2000) utilise des phrases à trous accompagnées de quatre propositions de réponse possibles par trou (il s'agit donc d'une sorte de QCM, qui s'éloigne d'un vrai test de production). Ce genre de format permet de tester la connaissance

de mots grammaticaux (déterminants, conjonctions, etc…), mais il est plus difficile à utiliser pour tester la connaissance de structures qui impliquent la phrase entière, comme le passif. Par ailleurs, c'est un format peu adapté à la modalité aurale, que nous aimerions utiliser pour notre test, et c'est pour cette raison que nous l'avons écarté.

8.2.2. Tests de jugement de grammaticalité

D'après Susan Gass (1983, p. 274), « *Judgments of grammaticality refer to a speaker's intuition concerning the nature of a particular utterance. The basic question is whether or not a given utterance (usually a sentence) is well-formed*. ». Le jugement de grammaticalité peut être un simple jugement oui/non, en temps limité ou pas, avec ou sans correction demandée, et avec des stimuli oraux ou écrits. Pélissier (2018), par exemple, utilise un test aural avec un public universitaire similaire au nôtre, pour tester l'acquisition de la morphologie du passé après entraînement implicite ou explicite. C'est un format assez simple qui est tout à fait adapté à notre contexte.

Un autre format intéressant est celui d'Andringa et al. (2012), qui font écouter à leurs apprenants néerlandophones les quatre premiers mots d'une phrase et leur demandent d'indiquer si c'est le début d'une phrase possible ou pas (ils trouvent une corrélation élevée entre ce test et une mesure de compréhension aurale). Cependant, nous n'avons pas retenu ce format parce qu'il nous semble qu'il risque de demander un effort cognitif supplémentaire aux apprenants, qui peuvent être amenés à imaginer la suite de la phrase eux-mêmes, et donc à fournir la contextualisation qui manque dans chaque phrase. Par ailleurs, du fait de nos conditions de passation à terme (en autonomie), nous préférons conserver un format plus classique qui risque moins de décontenancer les apprenants quand ils seront seuls face aux questions.

8.2.3. Test de connaissance explicite de règles grammaticales

Dans certains tests de jugement de grammaticalité, les apprenants L2 doivent montrer qu'ils sont capables d'identifier la faute en entourant le segment problématique (Bialystok, 1979; Mecartty, 2000), en corrigeant la phrase (S. Gass, 1983; Gutiérrez, 2013; Mecartty, 2000; Vafaee et al., 2017), en expliquant pourquoi la phrase est erronée (Gutiérrez, 2013; Vafaee et al., 2017), ou en sélectionnant la règle violée par la phrase présentée (Bialystok, 1979). Ces tâches font clairement appel à des connaissances métalinguistiques explicites, et ne nous paraissent pas adaptées à notre public : nous voulons en effet nous assurer que les candidats

qui ne connaissent pas les règles correspondant aux erreurs, qui manquent de vocabulaire métalinguistique mais qui sentent que certaines phrases sont problématiques (en utilisant leur intuition, leur connaissance implicite des probabilités de cooccurrence, etc.) soient capables aussi d'exécuter la tâche. Certains de nos étudiants sont en effet partiellement autodidactes : tout en ayant suivi des cours dans le système scolaire (français ou étranger), ils ont également travaillé en pays anglophone (année de césure avant d'entrer dans l'enseignement supérieur), ou passent beaucoup de temps à écouter de la musique ou regarder des vidéos en anglais. Il est possible que ces étudiants, grâce à une quantité d'input importante, aient intériorisé les régularités syntaxiques, sans être pour autant capables de les expliciter. Les études qui utilisent des tâches métalinguistiques montrent d'ailleurs à quel point les apprenants peuvent parfois être démunis face à elles (R. Ellis, 1991).

Certaines études utilisent un test de connaissances métalinguistiques encore plus poussées : dans le *Metalinguistic Knowledge Test* (Elder, 2009), les apprenants doivent montrer qu'ils sont capables non seulement d'identifier les règles violées par des phrases agrammaticales, mais aussi d'utiliser un certain nombre de termes métalinguistiques (par exemple ceux qui nomment les différentes catégories syntaxiques). Nous n'envisageons pas plus d'utiliser ce type de test, n'ayant pas d'objectifs portant sur le niveau d'expertise métalinguistique de nos sujets.

8.3. Construction du test

Nous avons donc décidé d'utiliser un jugement de grammaticalité simple, qui convient à notre contexte de compétence réceptive, et utilisant des stimuli oraux. Comme dans les tests diagnostiques décrits précédemment, nous avons cherché à varier la difficulté et le contenu des items proposés aux apprenants, à la fois pour que les niveaux de nos étudiants soient représentés, mais aussi pour que la validité de contenu soit assurée. Il faut en effet s'assurer que le construit de compétence grammaticale réceptive soit le mieux couvert possible, tout en proposant un nombre d'items tel qu'ils puissent être administrés dans un laps de temps assez court. Nous avons besoin d'un test qui discrimine particulièrement bien dans les niveaux faibles et chercherons donc à avoir un nombre suffisant d'items reflétant ces niveaux (B1 et inférieur).

8.3.1. Choix des stimuli et des items

Les manuels traditionnels de grammaire anglaise sont souvent organisés en quatre grands domaines : le groupe nominal, le groupe verbal, la phrase simple et la phrase complexe (par exemple, la *Grammaire explicative de l'anglais* de Larreya & Rivière, 2010). La couverture de ces grands pans de la grammaire est le premier critère qui a présidé à notre choix d'items, que nous présentons ci-dessous dans le Tableau 8.1. Les quatre grands domaines de la syntaxe sont bien représentés, avec un accent particulier sur le groupe verbal, qui présente des problèmes récurrents pour nos étudiants dans le choix des temps, des aspects et des modaux. Certains items sont comptés deux fois dans le tableau, dans deux catégories différentes : la phrase incorrecte **When was written this book ?*, par exemple, pour laquelle la difficulté réside dans la formation de l'interrogative avec un groupe verbal passif, est comptée une fois pour les questions et une autre pour le passif dans la catégorie « syntaxe de la phrase simple ».

domaines	structures	nb d'items
groupe nominal	géntitif /of (2), quantifieurs (3), démonstratifs (2), adjectifs invariables (1), singulier/ pluriel (1), nom (in)dénombrable (1), post-modifieurs du N (2)	12
groupe verbal	verbes à particule (1), verbes irréguliers (3), modalité (2), temps et aspect (4), présent simple (2), sous-catégorisation verbale (2)	14
phrase simple	ordre des mots (2), négation (2), passif (2), questions (3), phrase à sujet locatif (1)	10
phrase complexe	complémentation verbale :V-ing/ (to) V/ that (4), interrogative indirecte (1), subordination nominale (2), subordonnées adverbiales (1), extraposition (1)	9

Tableau 8.1 - items du test de jugement de grammaticalité aurale classés par domaine syntaxique

Le deuxième critère influençant la construction des items de notre test est celui du niveau de compétence grammaticale qu'ils représentent. Afin de nous aider à établir une correspondance entre les niveaux du CECRL et les structures grammaticales qu'on peut supposer acquises à chaque niveau, nous avons utilisé les documents produits par l'association EAQUALS (*European Association for Quality Language Services*) et le projet *English Profile*. En effet, le CECRL lui-même est un document neutre par rapport aux langues européennes dont il tente de décrire les étapes d'acquisition. Il décrit ce que les apprenants sont censés savoir faire à chaque niveau (les descripteurs performatifs dits « *can do* »), en terme d'activités langagières, fonctions de communication, genres de textes, contextes d'utilisation, degré d'aisance et de complexité, etc., mais ne donne d'exemples illustratifs dans aucune langue. Des documents d'accompagnement, créés comme le CECRL sous l'égide du Conseil de l'Europe, existent (ils préexistent parfois au CECRL, comme le *Threshold Level* dont la première version est parue en 1975 et a été réinterprété ensuite comme une description du niveau B1) et servent de

référentiels pour chaque niveau. Pour l'anglais, il s'agit de *Breakthrough* (Trim, 2001, niveau A1), *Waystage* (Van Ek & Trim, 1990b, niveau A2), *Threshold* (Van Ek, 1975; Van Ek & Trim, 1990a, B1), et *Vantage* (Van Ek & Trim, 2001, B2). Ces référentiels sont essentiellement des répertoires de notions langagières et de fonctions de communication, accompagnés d'exemples, mais leur objectif n'est pas d'identifier les items lexicaux ni les structures grammaticales qui pourraient être caractéristiques de tel ou tel niveau.

C'est pourquoi nous nous sommes tournée vers les ressources de l'association EAQUALS, en particulier le *Core Inventory for General English* (North et al., 2011). A partir de l'analyse des descripteurs du CECRL, de manuels pour l'enseignement de l'anglais utilisés typiquement dans le système universitaire ou la formation continue en Europe (mais pas dans le système scolaire secondaire), des programmes de diverses écoles de langues, et d'enquêtes auprès d'enseignants d'anglais, ce document propose (entre autres) une liste des structures grammaticales correspondant à un niveau donné dans plus de 80% des ressources analysées. C'est ainsi qu'on peut constater que le *present simple* est introduit au niveau A1, mais que le *future perfect continuous* ne l'est qu'au niveau B2.

Savoir à quels niveaux sont enseignées quelles structures est important, mais cela ne nous dit pas vraiment à quel niveau elles sont utilisées par les apprenants (en compréhension ou en production). Comme l'a souligné Corder (1967), *input* ne veut pas dire *intake*, et les structures auxquelles l'apprenant est exposé pendant l'enseignement ne sont pas forcément intégrées. Nous avons donc également fait appel aux ressources produites par le projet *English Profile*, dont l'un des objectifs est d'essayer de caractériser de façon empirique, à partir de productions d'apprenants (bien que les auteurs affirment ne pas séparer la compréhension et la production), les propriétés lexicales et grammaticales essentielles de chaque niveau du CECRL. Ces propriétés, appelées *criterial features*, sont définies ainsi :

> *'Criterial features' are linguistic properties from all aspects of language (phonology, morphology, syntax, semantics, discourse, etc.) which can distinguish the different proficiency (here CEFR) levels from one another and thus can serve as a basis for the estimation of a learner's proficiency level. […][T]hey capture essential distinguishing properties of the CEFR proficiency levels*. (Salamoura & Saville, 2010, p. 102)

Pour identifier ces propriétés, les chercheurs du projet se basent sur l'analyse d'un corpus d'apprenants, le *Cambridge Learner Corpus*, et ont produit un site Internet qui permet de

naviguer dans leur base de données à partir d'items lexicaux[40] ou de structures grammaticales[41]. La base lexicale permet de voir à quel niveau apparaissent les différents sens des items lexicaux dans les productions d'apprenants. Ces différents sens sont souvent associés à des structures différentes, et peuvent donc nous donner des renseignements utiles (un exemple est fourni en Figure 8.1 pour le verbe *depend*, où l'on constate qu'il apparaît de façon suivie dans les productions d'apprenants à partir du niveau B1, mais que ce n'est qu'au niveau B2 qu'apparaît le sens « se reposer sur (quelqu'un) »). Un exemple de l'interface grammaticale est fourni en Figure 8.2.

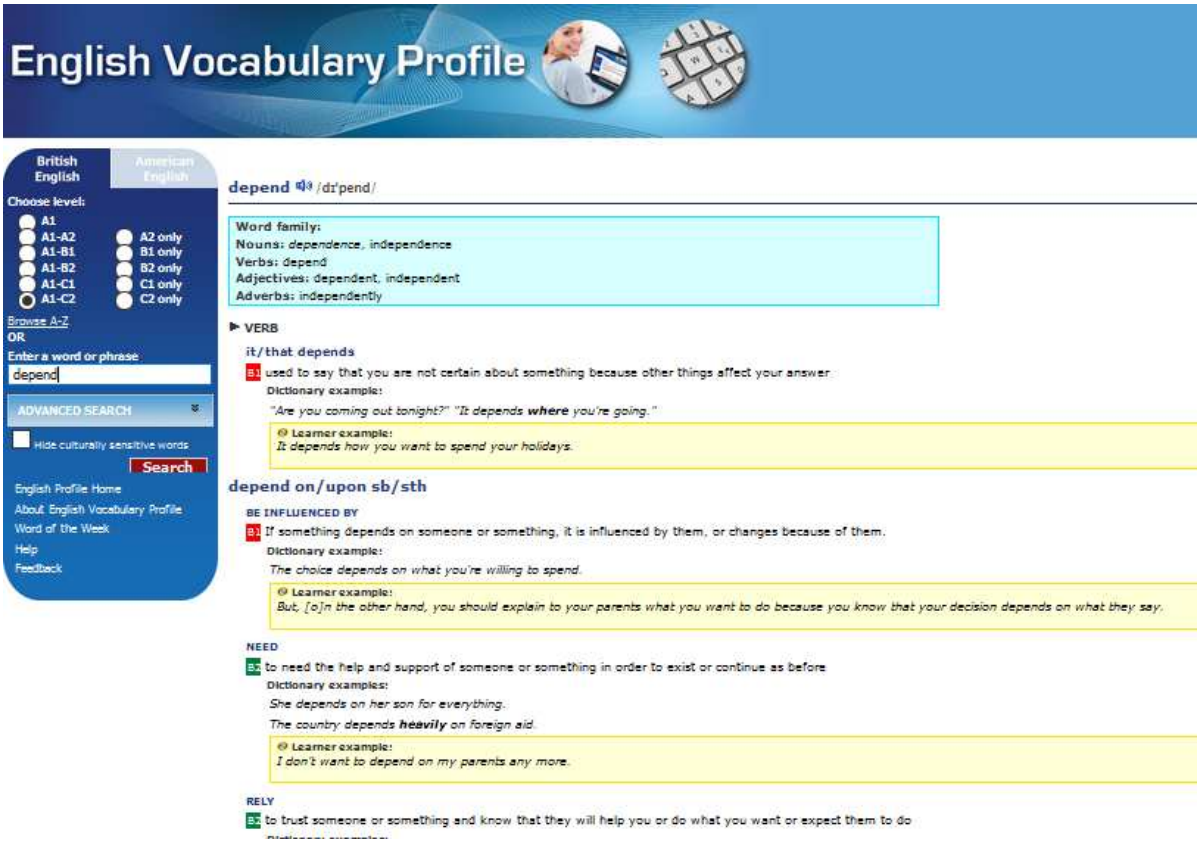


Figure 8.1 - exemple d'écran de résultat (verbe *depend*) du site *English Vocabulary Profile*

Ces ressources ne sont pas encore tout à fait suffisantes car, comme le CECRL, elles décrivent ce que les apprenants savent faire à chaque niveau, mais pas ce qu'ils ne savent pas (encore) faire. Or, notre test de grammaticalité comporte nécessairement des phrases erronées que nous aimerions également lier à des niveaux donnés. C'est ce qu'ont également essayé de faire les chercheurs du projet *English Profile*, en cherchant à caractériser les niveaux du CECR par des structures en cours d'acquisition (*developing language features*), définies comme « *features*

[40] http://vocabulary.englishprofile.org (English Profile, 2012)
[41] http://www.englishprofile.org/english-grammar-profile/egp-online (English Profile, 2015)

that appear at a certain level but […] are unstable, i.e. they are not used correctly in a consistent way » (Salamoura & Saville, 2010, p. 110). Cependant, ils constatent en pratique que, même si les erreurs diminuent de façon significative d'un niveau à l'autre pour un certain nombre de structures, ce n'est qu'aux niveaux C, et en particulier au niveau C2, que l'amélioration est la plus nette. A ces niveaux, l'utilisation des structures déjà acquises devient plus cohérente, et les erreurs disparaissent peu à peu :

> *[A]s learners progress from level A1 through to B2, they gradually acquire new structures which can be identified as characteristic of each level. Once they reach C levels, learners' progress is characterised by increased structural accuracy and by greater lexical accuracy and range rather than by the addition of new structures to their repertoire.* (English Profile, 2011, p. 15)

Il nous parait difficile dans ces conditions d'attribuer un niveau à des phrases erronées, dans la mesure où on trouve encore au niveau C1 des productions telles que « *the tour was a completely disaster* » (ibid., p. 30) et « *many advices are hidden inside* » au niveau C2 (ibid., p. 35).

SuperCategory	SubCategory	Level	Can-do statement
ADJECTIVES	combining	A1	FORM: COMBINING TWO ADJECTIVES WITH 'AND' Can use 'and' to join a limited range of common adjectives.
ADJECTIVES	combining	A2	FORM: COMBINING TWO ADJECTIVES WITH 'BUT' Can use 'but' to join a limited range of common adjectives, after 'be'.
ADJECTIVES	combining	B1	FORM: BEFORE THE NOUN Can use a comma to combine two adjectives used before the noun, following the usual order of adjective types.
ADJECTIVES	combining	B1	FORM: COMBINING COMPARATIVE ADJECTIVES WITH 'AND' Can use 'and' to join a limited range of comparative adjectives.► adjectives: comparatives
ADJECTIVES	combining	B1	FORM: COMBINING MORE THAN TWO ADJECTIVES Can use commas and 'and' to join more than two adjectives, after 'be'.
ADJECTIVES	combining	B1	FORM: COMBINING THE SAME COMPARATIVE ADJECTIVE WITH 'AND' Can use 'and' to repeat a comparative adjective to indicate change over time, usually after 'become' or get. ► adjectives: comparatives
ADJECTIVES	combining	B1	FORM: COMPOUND ADJECTIVES Can use a limited range of compound adjectives ('good-looking', 'well-known')

Figure 8.2 - exemple d'écran de résultat du site *English Grammar Profile* (syntaxe des adjectifs en fonction du niveau CECRL)

Nous avons donc décidé de caractériser les phrases erronées par le niveau où la forme correcte est utilisée de façon caractéristique. Par exemple, la phrase incorrecte *She's name is Anna* correspond à la phrase correcte *Her name is Anna*. La deuxième ligne du Tableau 8.2 ci-dessous, qui présente la liste des items et les raisons pour lesquelles ils ont été placés dans tel ou tel niveau, explique pourquoi cette phrase est étiquetée A1. D'une part, le document EAQUALS indique que les déterminants possessifs (appelés adjectifs possessifs dans la

terminologie EAQUALS) sont enseignés au niveau A1, et d'autre part, le projet *English Grammar Profile* (EGP) constate que *her* fait partie des possessifs dont l'utilisation (correcte) est caractéristique du niveau A1 (dans le tableau, nous citons les règles telles qu'elles apparaissent sur le site du EGP, ici « *Can use possessive determiners 'my' 'your' 'his' 'her' 'our' before nouns* »). Quand une phrase incorrecte a deux formes correctes possibles correspondant à des niveaux différents, l'item a été mis entre les deux niveaux. Par exemple, **I haven't the keys* correspond en anglais standard soit à *I don't have the keys* (négation avec *don't*, niveau A1), soit à *I haven't got the keys* (négation du *present perfect*, niveau A2), et a donc été placé en A1/ A2.

Quand la structure est attachée à un item lexical particulier, nous donnons la référence du site *English Vocabulary Profile* (EVP), accompagnée d'un exemple tiré de leur corpus d'apprenants. Par exemple, la phrase (incorrecte) **This stroke me as important* est étiquetée B2 (cinquième ligne avant la fin du Tableau 8.2), parce que la forme correcte du verbe au prétérit, *struck*, est utilisée de façon caractéristique au niveau B2 d'après EVP.

Nos 45 items couvrent ainsi les niveaux A1 à C1, comme on peut le constater dans le Tableau 8.2. Comme notre objectif principal est d'identifier les étudiants en difficulté, nous avons essayé de nous concentrer sur la frontière entre les niveaux faibles (A2 et B1) et le niveau juste au-dessus (B2), c'est pourquoi 35 des 45 items se trouvent à ces niveaux. Un item (**I don't have any car*) est sans niveau identifié car nous n'avons pas réussi à trouver mention de la règle grammaticale correspondante (négation des dénombrables singuliers avec *not a* plutôt que *not any*) associée à un niveau particulier. Nous signalons cependant que le projet *English Profile* note que ce genre d'erreur diminue significativement entre les niveaux C1 et C2.

Par ailleurs, nos différentes sources donnent parfois des informations contradictoires. A la cinquième ligne du tableau, par exemple, on peut constater que l'item incorrect **I have seen it last year* a été placé au niveau A1 parce que EGP (*English Grammar Profile*) indique que le prétérit (qui correspond à la forme correcte *I saw it last year*) est utilisé de façon caractéristique au niveau A1 pour certains verbes très courants (dont on suppose que *see* fait partie) et que EVP (*English Vocabulary Profile*) confirme que l'expression *last year* est utilisée en A1. Nous n'avons donc pas tenu compte de l'information du projet EAQUALS selon laquelle la différence entre le prétérit et le *present perfect* est enseignée au niveau B1.

niveaux (nb)	items	justification
A1 (6)	Every person is important	EGP A1 Can use a limited range of quantifying determiners with singular nouns ('a', 'every')
	*She's name was Anna	EAQ A1 possessive adjectives EGP A1 Can use possessive determiners 'my' 'your' 'his' 'her' 'our' before nouns
	*There are differents possibilities	EGP A1 Can form simple noun phrases by pre-modifying plural nouns with an adjective and no determiner EGP A1 Can use 'there are' + plural noun phrase as complement
	*I have seen it last year	[EAQ B1 present perfect/ past simple] EVP A1 last year EGP A1 Can use the affirmative form [of the past simple] with a limited range of regular and irregular verbs
	*I like very much sweets	EAQ A1 very basic intensifiers EGP A1 Can use 'very much' with verbs expressing preference
	*I am boring in class	EVP A1 'bored': *I can play when I'm bored*
A1/A2 (1)	*I haven't the keys	EGP A1 Can form negative statements of main verbs in the present simple with 'don't' EGP A2 Can form negative statements of main verbs in the present perfect
A2 (12)	A friend of mine came	EGP A2 Can use 'of mine' after 'friend'
	Those dogs are trained	[EAQ A1 demonstrative adjectives] EGP A2 Can use 'those' with plural nouns
	I was born early	EVP A2 'born': *She was born 2 months ago*
	I have been to Rome	EAQ A2 present perfect EGP A2 Can use the present perfect simple to talk about experiences up to now
	What happens next?	EGP A2 Can use the present simple for timetabled events in the future
	*This children are lost	[EAQ A1 demonstrative adjectives] EGP A2 Can use 'these' with plural nouns
	*She's one of my friend	EGP A2 Can use a range of quantifying determiners + 'of' + determiner (+ N)
	*Have a good travel	EVP A2 'trip': *Have a nice time on your trip*
	*I spended last week end in London	EVP A2 'spend' *I spent 500 dol[la]rs*
	*The crash arrived because it was dark	EVP A2 'happen' *The accident happened*
	*I'm agree with you	EVP A2 'agree': *Do you agree with me?*
	*I explain you the situation	EVP A2 'explain' *explain the information to our friends* EGP A2 transitive frames with a following PP, NP-V-NP-PP, P =*to* EGP A2 Can use 'will' to talk about willingness and offers
A2/B1 (2)	We may go sightseeing later	EAQ B1 Modals – might, may, will, probably EGP A2 Can use 'may' to talk about weak possibility referring to the present and the future
	*Sorry, I forget to send it	EVP B1 'forget' *I almost forgot to…* EGP A2 Can use the past simple with an increasing range of verbs
B1 (11)	Most people think so	EGP B1 Can use an increasing range of quantifying determiners with both plural nouns and uncountable nouns ('most',…)
	What were you thinking?	EAQ B1 past continuous EGP B1 Can use the question form [of the past continuous]

	That city seems very large	EVP B1 'seems' *She seems very friendly*
	The idea that they might win is absurd	EAQ B1 Modals – might, may, will, probably EVP B2 'absurd' EGP A2 Can use 'might' to talk about weak possibility
	Guess where it is	EGP B1 indirect questions
	Before doing anything you need to think	EGP B1 Can use a non-finite subordinate clause with 'before' and 'after' + '-ing'
	*the house who's on the other side	EAQ B2 relative clauses EGP A2 Can use a defining relative clause with 'which' as the subject
	*She is here since 2 years now	EGP B1 Can use the present perfect simple with 'since' to talk about duration
	*It's depend on the situation	EVP B1 'depend' *It depends on what they say*
	*I want that you stay	EGP B1 Can use some verbs of requesting and commanding followed by a direct object and a 'to-'infinitive
	*He helped her making the cake	EGP B1 Can use 'help' + object +infinitive
B2 (9)	I took my coat off	[EAQ A2 Common phrasal verbs] EGP B2 Can use phrasal verbs + nouns as object + particle
	The professor I gave the book to has left	EGP B2 Can use defining relative clauses ending in a preposition
	This game is sure to be a winner	EGP B2 Can use the full range of expressions with 'be' + infinitive ('be likely to', 'be bound to', 'be sure to', …)
	Smoking is known to cause cancer	EGP B2 raising verb (passive)
	They worried about him drinking	EGP B2 NP-V-P-NP-V(+ing)
	It's amazing to think that it's true	EGP B2 *it* extraposition with infinitival phrases
	They wanted the children found	EGP B2 NP-V-NP-AdjP (object control)
	*From what says Joe, it's not true	EGP B2 Can report speech directly inverting the subject and verb in the reporting clause where the subject is a proper noun or noun phrase
	*This stroke me as important	EVP B2 'strike': *what struck me…*
C1 (3)	I saw them drive away	EGP C1 Can use some verbs connected with the senses + direct object + infinitive without 'to'
	*When was written this book	EAQ C1 all passive forms
	This tent sleeps four.	EGP (no level) locative subject
indéfini (1)	*I don't have any car	EGP diminue significativement entre C1 et C2

Tableau 8.2 - items par niveau du CECR (les phrases incorrectes sont précédées d'une astérisque), avec justification du niveau par référence aux projets EAQUALS (EAQ), English Grammar Profile (EGP) ou English Vocabulary Profile (EVP). Les crochets indiquent les informations contradictoires non prises en compte.

8.3.2. Administration du test

Comme les précédents, le test de grammaticalité aurale a été administré par l'intermédaire de la plateforme SELF. La Figure 8.3 reproduit l'écran d'accueil du test, et la Figure 8.4 l'interface utilisée pour tous les items du test.

Figure 8.3 – écran d'accueil du test de jugement de grammaticalité aurale

Figure 8.4 - écran d'administration d'un item du test de jugement de grammaticalité aurale

8.4. Résultats

8.4.1. Statistiques descriptives globales

En moyenne, les étudiants ont eu besoin de 6,5 minutes (écart-type 2) pour passer ce test, ce qui est satisfaisant puisque inférieur à notre limite de 10 minutes. La personne la plus rapide a mis 3,6 minutes et la plus lente 13 minutes. L'écart-type de 2 minutes nous indique que les deux tiers de notre échantillon mettent entre 4 et 8 minutes environ (moyenne +/- un écart-type), et que la quasi-totalité mettent entre 2 et 10 minutes environ (moyenne +/- deux écarts-types).

Les résultats du test, passé par 184 étudiants, sont présentés dans le Tableau 8.3 ci-dessous. La moyenne obtenue aux 45 items du test est encore plus faible qu'au test de reconnaissance lexicale (56% de réussite ici : 25,3/ 45, avec un écart-type de 6,25). Sachant que le hasard est à 50% (puisque c'est un test oui/non), cela montre la grande difficulté qu'éprouvent nos étudiants à accomplir ce type de tâche, avec des items dont la majorité sont faciles ou très faciles (32 items sur 45 correspondent aux niveaux A1 à B1). L'écart entre le score le plus faible (12/45, soit 27% de bonnes réponses) et le plus élevé (42/45, soit 93% de bonnes réponses) est par ailleurs très important, avec 30 points.

n	moyenne	écart-type	médiane	min	max	étendue	asymétrie	kurtose
184	25.31	6.25	24	12	42	30	0.49	-0.32

Tableau 8.3 - statistiques descrptives du score total (sur 45 points possibles) au test de jugement de grammaticalité aurale

Le coefficient d'asymétrie est positif, ce qui confirme l'impression de difficulté du test : il distingue bien les candidats assez forts (étalement des scores vers la droite, comme on peut le constater sur l'histogramme de fréquence en Figure 8.5), mais est peut-être insuffisant pour bien distinguer les candidats plus faibles (peu d'étalement dans les scores faibles à gauche).

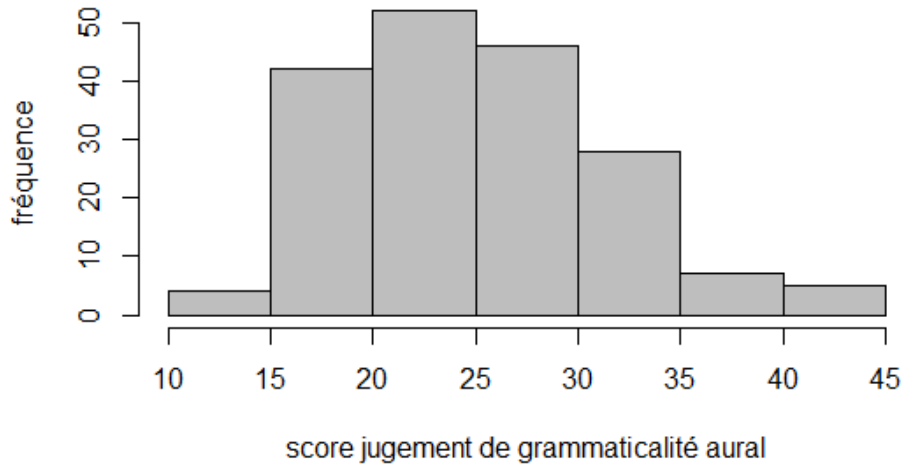


Figure 8.5 - histogramme du score global au test de reconnaissance du lexique aural

Un test de Shapiro-Wilk confirme que la distribution n'est pas normale ($W = 0,97$, $p<0,001$), ce qui nous interdit d'utiliser des tests paramétriques pour la suite de l'analyse des scores du jugement de grammaticalité.

La corrélation entre le temps total passé sur le test et le score total, que nous avons calculée avec un coefficient de Spearman (car aucune des deux variables ne suit une distribution

normale), est faible (*rho* = -0.22). Elle est cependant significative (p < 0,01). Le fait que *rho* soit négatif nous apprend que les meilleurs apprenants ont tendance à prendre un tout petit peu moins de temps. Cependant, il n'est pas certain que cette relation se retrouve en conditions réelles de passation, où les étudiants pourront gérer leur temps comme ils le désirent. Dans le cas présent, même s'il n'y avait aucun enjeu, les étudiants ont tous passé le test dans une salle sous la surveillance (bienveillante) d'un enseignant.

Le coefficient alpha de Cronbach est de 0,78, ce qui est au-dessus de la valeur acceptable de 0,70. Nous pouvons donc considérer que la fiabilité du test est bonne, même avant suppression des items insatisfaisants.

Pour évaluer l'unidimensionnalité du test, nous avons utilisé une analyse en composants principaux, dont le but est d'identifier un facteur principal qui synthétise le maximum d'informations apportées par les différents items du test. Nous ferons ce constat par inspection visuelle du graphe du cercle des corrélations de toutes les variables avec l'axe principal de l'ACP (Figure 8.6). Nous constatons sur le graphique que tous les items du test corrèlent de façon positive avec le facteur principal identifié (parce que tous les vecteurs les représentant sont orientés vers la droite). De plus, ce facteur explique presque 20% de la variance du test (comme indiqué dans la parenthèse à droite du nom de l'abscisse de la figure), ce qui est un résultat acceptable (Husson et al., 2016).

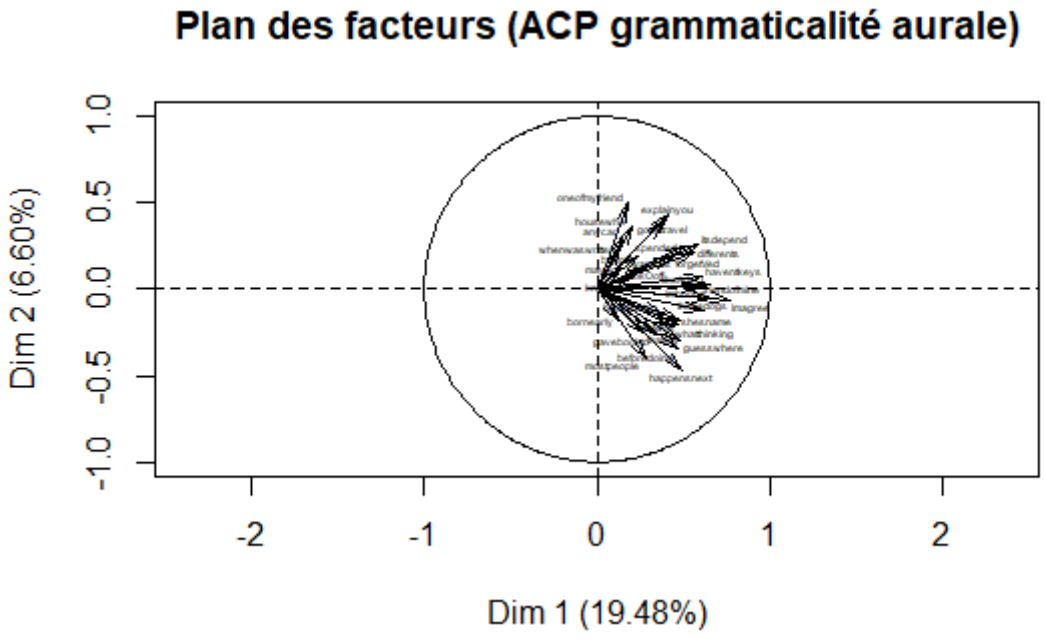


Figure 8.6 - plan des facteurs ACP des items du test de jugement de grammaticalité aurale

8.4.2. Analyse des items

Pour plus de facilité de lecture, nous présentons dans le Tableau 8.4 les résultats par item en deux groupes de colonnes, celles correspondant aux 22 phrases correctes à gauche, et celles pour les 23 phrases incorrectes à droite, toutes présentées par ordre de difficulté croissante. Dans chaque groupe, la première colonne correspond au nom de l'item (une version abrégée de la phrase à laquelle il correspond), la deuxième colonne contient la difficulté (le pourcentage de bonnes réponses à l'item), et la troisième le coefficient de discrimination (qui indique dans quelle mesure les candidats forts ont mieux répondu à l'item que les faibles).

phrases correctes				**phrases incorrectes**		
nom item	**difficulté**	**discrim.**		**nom item**	**difficulté**	**discrim.**
sleeps4	0.29	**0.15**		anycar	0.13	0.24
maygo	0.32	0.26		goodtravel	0.16	0.42
childrenfound	0.35	**0.03**		oneofmyfriend	0.25	0.29
sureto	0.52	**0.08**		**perflastyear**	0.26	**-0.07**
thosedogs	0.54	0.53		explainyou	0.29	0.45
himdrinking	0.56	0.40		**helpOVing**	0.36	**-0.06**
ideathat	0.60	0.46		forgetVed	0.37	0.47
friendofmine	0.66	0.54		whenwaswritten	0.40	0.22
seeObjV	0.68	**0.13**		differents	0.41	0.55
beento	0.70	0.29		Imboring	0.42	0.54
gavebookto	0.70	0.29		**strokeme**	0.43	**0.09**
bornearly	0.72	0.20		**crasharrived**	0.45	**0.00**
whatthinking	0.72	0.39		spended	0.48	0.48
guesswhere	0.75	0.40		haventkeys	0.51	0.53
thatcity	0.78	0.29		housewho	0.56	0.26
beforedoing	0.79	0.41		likeAdvO	0.58	0.57
happensnext	0.80	0.41		itsdepend	0.58	0.56
knownto	0.80	0.21		**whatsaysJoe**	0.59	**0.00**
mostpeople	0.82	0.28		Imagree	0.59	0.68
tookOoff	0.83	0.30		**thischildren**	0.61	**0.12**
amazingto	**0.91**	**0.16**		**isheresince**	0.65	**0.14**
everyperson	**0.92**	**0.12**		wantthat	0.66	0.44
				shesname	0.82	0.43
moyenne	0.67				0.46	
						alpha de Cronbach 0.78

Tableau 8.4 - indices de difficulté et de discrimination des items du test de jugement de grammaticalité (phrases correctes à gauche, incorrectes à droite) ; les indices de discrimination inférieurs à 0,2 sont surlignés en gris et mis en gras (ainsi que le nom des items correspondants) ; les indices de difficulté (facilité) supeérieurs à 0,9 sont en gras

On constate que la difficulté varie pour les phrases correctes de 29% de bonnes réponses (*This tent sleeps four*, rejetée par une grande majorité de candidats), à 92% (*Every person is*

important, qui avait d'ailleurs été identifiée comme un item A1). Pour les phrases incorrectes, la difficulté varie de 13% (**I don't have any car*, qui peut éventuellement être considérée comme correcte dans certains contextes) et 16% (**Have a good travel !,* un item A2) à 82% (*She's name was Anna*, item A1). L'éventail des difficultés est donc très large.

Six phrases correctes ont un coefficient de discrimination trop faible pour pouvoir en principe être conservées comme items du test (leur coefficient, inférieur à 0,2, est en gras et surligné en gris dans la troisième colonne du tableau). Deux de ces items ont un coefficient de difficulté (facilité) un peu trop élevé (supérieur à 0,9). Ils sont très bien réussis de façon générale, par plus de 90% des candidats, et c'est pour cette raison qu'ils ont du mal à discriminer les étudiants forts des faibles (les quelques candidats qui se sont trompés n'ont pas beaucoup plus de chances d'être faibles que forts). Cependant, étant donné qu'il est utile d'avoir quelques items faciles dans le test (pour encourager les candidats très faibles), nous avons décidé de garder celui avec le meilleur coefficient de discrimination (*amazingto*, avec 0,16, assez proche de notre limite de 0,2). Nous n'enlèverons donc que cinq items corrects.

Parmi les phrases incorrectes, sept doivent être rejetées parce qu'insuffisamment discriminantes. Deux d'entre elles ont même un coefficient négatif (*I can help you making the cake* et *I have seen it last year*, respectivement B1 et A1), ce qui signifie que les étudiants forts ont eu plus de mal à les rejeter que les faibles. Nous reviendrons sur ce constat dans les prochains paragraphes.

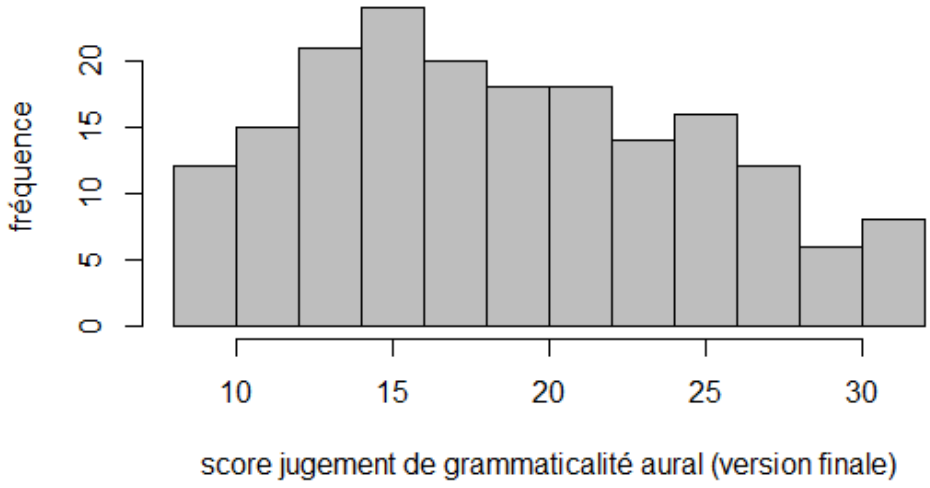


Figure 8.7 - histogramme du score global au test de reconnaissance du lexique aural dans sa version écourtée

Le test final aura donc 12 items en moins, et sera composé de 17 phrases correctes et 16 phrases incorrectes. Le coefficient de fiabilité (alpha de Cronbach) de cette version écourtée

est de 0,85. La distribution reste non normale (coefficient de Shapiro-Wilk W = 0,97, p<0,001), mais le coefficient d'asymétrie diminue un peu (il passe de 0,49 dans la version initiale à 0,25). La distribution demeure cependant légèrement asymétrique comme on peut le constater dans l'histogramme ci-dessus (Figure 8.7). Malgré l'attention portée à la conception des items, le test reste ainsi un peu trop difficile pour notre population. Nous allons essayer d'éclaircir pourquoi dans les paragraphes qui suivent.

8.4.3. Analyse approfondie

8.4.3.1. corrélation avec le niveau estimé des items

Etant donné que nous avons conçu le test en essayant de proposer des phrases à différents niveaux de difficulté, nous pouvons tenter de vérifier que ces niveaux correspondent à la difficulté objective pour nos étudiants. Pour obtenir des groupes qui ne soient pas trop déséquilibrés, nous avons analysé l'item A1/A2 (**I haven't the keys*) avec les 6 items A1, les deux items A2/B1 avec les items B1, et les trois items C1 avec les items B2, arrivant ainsi à 4 groupes de 7 items (A1 et A1/A2), 12 items (A2), 13 items (A2/B1 et B1), et 12 items (B2 et C1). Nous n'avons pas inclus l'item de niveau indéfini. Nous voyons sur la Figure 8.8 que, contrairement à notre hypothèse, la difficulté n'augmente pas avec le niveau : les phrases ont sensiblement la même médiane de difficulté quel que soit le niveau, et, même si aucune phrase de niveau A1 (boîte de gauche) n'a l'air extrêmement difficile (en dessous de 40% de bonnes réponses), l'éventail de difficulté des phrases B2/C1 (boîte de droite) a l'air relativement similaire à celui des phrases A1.

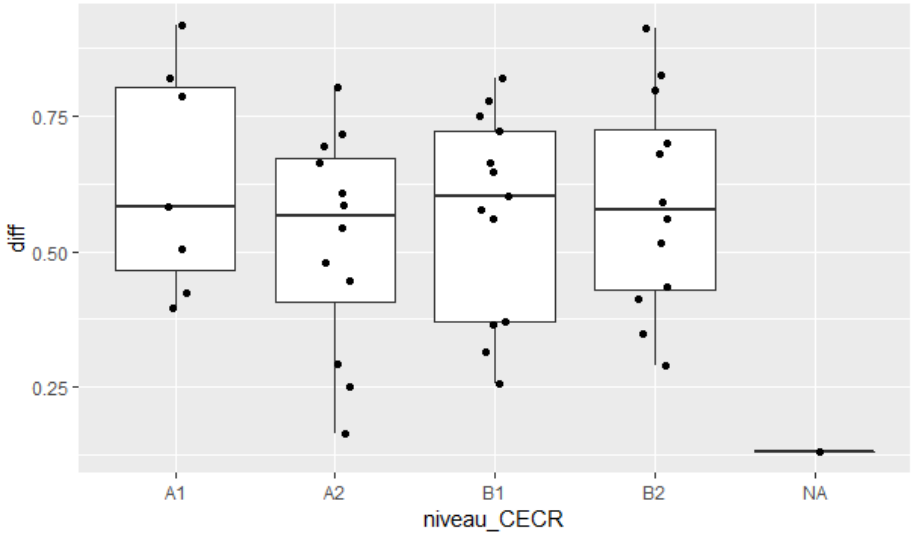


Figure 8.8 - difficulté des phrases du test de jugement de grammaticalité aurale par niveau de conception des phrases

On peut essayer de comprendre ce résultat de deux façons différentes, qui ne sont pas mutuellement exclusives. Premièrement, on pourrait objecter que les niveaux des items ont été estimés à partir de productions d'apprenants, alors que nous travaillons en réception. Cependant, il paraît peu probable qu'il n'y ait aucune correspondance de niveau entre les deux modalités : même s'il peut exister une asymétrie réception et production, surtout au début de l'acquisition, du fait que la tâche de production demande plus de ressources que la tâche de compréhension (Chater et al., 2016), cette asymétrie ne devrait pas modifier l'ordre de difficulté des items. Il serait plus logique de penser que la compréhension des structures qui sont caractéristiques de la production à un niveau donné pourrait être caractéristique d'un niveau inférieur ou égal en compréhension (par exemple, Fraser et al. (1963) trouvent, avec des enfants anglophones, une corrélation de 0,72 entre le rang de difficulté en compréhension et celui en production de 10 structures de l'anglais, la compréhension étant toujours mieux réussie que la production). Cela ne devrait donc rien changer à l'ordre de difficulté de nos items.

C'est pourquoi il faut se tourner vers d'autres tentatives d'explication. Comme pour le test de reconnaissance lexicale, nous constatons qu'il est beaucoup plus facile pour nos apprenants d'accepter les phrases correctes que de rejeter les phrases incorrectes. En effet, la difficulté moyenne des items correspondant aux phrases correctes est de 0,67, tandis que celle des items incorrects est de 0,46 (voir dernière ligne du Tableau 8.4). Les phrases incorrectes sont donc particulièrement mal réussies, et acceptées plus souvent qu'elles ne sont rejetées. Ce résultat n'est pas nouveau, et d'autres études ont également trouvé que les phrases grammaticales sont plus faciles à juger que les agrammaticales pour les apprenants L2 (Bialystok, 1979; R. Ellis, 1991; Gutiérrez, 2013; Loewen, 2009) : Gutiérrez par exemple trouve comme nous une différence de plus de 20% entre le résultat de ses sujets (anglophones apprenant l'espagnol) aux phrases grammaticales et agrammaticales d'un test de jugement de grammaticalité (écrite) avec ou sans limite de temps. Certaines études récentes font ainsi l'hypothèse que le jugement des phrases grammaticales et agrammaticales ne fait pas appel aux mêmes connaissances (R. Ellis, 2005; Gutiérrez, 2013) : les unes feraient appel à des connaissances implicites, et les autres des connaissances explicites. Bien que cette hypothèse soit controversée (Vafaee et al., 2017), nous allons examiner ces deux types d'items séparément, et relancer les analyses d'abord pour les phrases correctes, puis pour les phrases incorrectes.

8.4.3.2. difficulté des phrases correctes

Comme le nombre de phrases correctes n'est pas le même dans chaque niveau, et qu'il y en a proportionnellement moins dans les niveaux plus bas, cela nous oblige à réorganiser nos groupes : un groupe de 6 items A1/A2, un groupe de 7 items B1, et un groupe de 9 items B2/C1. Les résultats sont visibles dans le Tableau 8.5 ci-dessous. Même si la difficulté moyenne semble croître (légèrement) avec le niveau dans cette nouvelle configuration des items, la variabilité entre items est clairement au moins aussi importante à l'intérieur de chaque groupe qu'entre les groupes (nous pouvons vérifier avec un test de Kruskal-Wallis que la différence des moyennes n'est pas significative : χ^2=0,41, df =2, p=0,82). Nous pouvons toutefois observer que le nombre des items très difficiles, surlignés en gris dans le Tableau 8.5, augmente avec le niveau : on ne trouve aucun item réussi par moins de 50% de l'échantillon en A1/A2, mais on en trouve 1/7, soit 14%, en B1, et 2/9, soit 22% en B2/C1.

niveau	items	difficulté moyenne
A1/A2	Every person is important (0,92) What happens next (0,80) I have been to Rome (0,72) I was born early (0,70) A friend of mine came (0,66) Those dogs are trained (0,54)	0,72
B1	Most people think so (0,82) Before doing anything you need to think (0,79) That city seems very large (0,78) Guess where it is (0,75) What were you thinking (0,72) The idea that they might win is absurd (0,60) We may go sightseeing later (0,32)	0,68
B2/C1	It's amazing to think that it's true (0,91) I took my coat off (0,83) Smoking is known to cause cancer (0,80) The professor I gave the book to has left (0,72) I saw them drive away (0,68) They worried about him drinking (0,56) This game is sure to be a winner (0,52) They want the children found (0,35) This tent sleeps four (0,29)	0,63

Tableau 8.5 - difficulté observée des items du test de jugement de grammaticalité aurale, par groupe de niveau, avec difficulté croissante à l'intérieur de chaque groupe (items réussis par moins de la moitié de l'échantillon surlignés en gris)

Essayons d'analyser les résultats à l'intérieur de chaque groupe. Dans le groupe A1/A2, certaines phrases sont effectivement très bien (*Every person is important*, 0,92) ou bien (*What happens next ?*, 0,80) réussies, mais d'autres ont visiblement posé des problèmes à nos étudiants : *Those dogs are trained* n'est accepté que par 54% des étudiants, et, plus étonnant,

A friend of mine came par seulement 66%. Pour la première phrase, deux explications sont possibles : soit le démonstratif *those* est beaucoup moins bien connu de nos étudiants que de ceux des projets EAQUALS et *English Profile*, soit c'est un problème lexical (méconnaissance du participe passé *trained*). Pour tester cette deuxième hypothèse, nous avons vérifié à quel niveau apparaît le verbe *train*, et c'est effectivement au niveau B1 seulement qu'il apparaît. Cependant, il nous semble que ce verbe peut tout de même être connu des étudiants faibles, parce que le mot *training* est utilisé en français ; d'autre part, 54% de bonnes réponses paraît faible même pour un item B1. C'est pourquoi nous pensons que le problème vient probablement du démonstratif *those*, dont la prononciation est peut-être mal connue de nos étudiants (nous avons remarqué de façon anecdotique dans nos cours que certains étudiants le prononcent /ðu:z/). Le fait que l'indice de discrimination soit supérieur à 0,5 montre également que cet item a été bien mieux réussi par les étudiants forts que par les faibles, ce qui peut signifier qu'il était simplement mal placé en A2 et que son niveau est supérieur à ce que nous pensions (B1/B2 plutôt que A1/A2). On peut remarquer que l'item incorrect **This children are lost* a été également mal réussi (correctement rejeté par 60% d'étudiants seulement), ce qui peut pointer un problème spécifique aux démonstratifs, peut-être lié à leur forme orale. L'inadéquation entre nos résultats et les niveaux proposés par les projets EAQUALS et English Profile pourrait donc être dû en partie, non à la différence entre compréhension et production comme initialement envisagé, mais à celle entre écrit et oral.

Nous avons beaucoup plus de mal à rendre compte du fait que *A friend of mine came* ne soit accepté que par les deux tiers de notre échantillon. Cette phrase nous paraît bien placée en A2, d'une part parce que tous les mots qui la composent, ainsi que leur combinaison, sont extrêmement courants, et d'autre part parce que les thèmes qu'elle évoque (description de soi, domaine personnel) sont typiques des niveaux A1 et A2 du CECR. De plus, il n'y a pas de raison de penser qu'elle puisse être plus difficile à l'oral qu'à l'écrit. La seule explication qui pourrait être avancée est que les étudiants ont analysé la phrase consciemment et ont (à juste titre) trouvé étrange l'utilisation simultanée de *of* et du possessif *mine* (cette structure s'appelle d'ailleurs le double génitif). Le fait que cette phrase soit très discriminante (coefficient de 0,54) montrerait alors que les étudiants faibles se reposent plus sur les règles explicites et les forts plus sur leur intuition née d'une plus grande exposition à la langue, ce qui correspond à certains résultats antérieurs sur les jugements de grammaticalité (Goss et al., 1994, cité par Loewen, 2009).

Examinons maintenant les items B1, dont deux sont particulièrement difficiles. Le premier, *The idea that they might win is absurd* (accepté par 60% d'étudiants seulement), nous paraît objectivement difficile, avec une subordonnée nominale dans laquelle apparaissent un modal épistémique (*might*), et l'adjectif *absurd* caractéristique plutôt du niveau B2 (voir Tableau 8.2), même s'il est transparent pour des francophones. Le deuxième item, par contre, *We may go sightseeing later*, a été encore beaucoup moins bien réussi, puisqu'accepté par un tiers seulement (32%) de notre échantillon, alors que sa structure est beaucoup plus simple, et que, d'après English Profile, *sightseeing* apparaît bien au niveau B1. Après vérification dans le corpus $SUBTLEX_{US}$ (déjà utilisé pour le test de reconnaissance lexicale), le mot *sightseeing* est en fait beaucoup moins courant (environ 5 fois moins) que l'adjectif *absurd*. Rappelons que $SUBTLEX_{US}$ est un corpus de sous-titres de films et séries télévisées américaines. Il est possible qu'à l'oral, la première phrase soit plus courante que la deuxième -d'ailleurs, le modal *might* ($200^{ème}$ mot le plus courant) y est également un peu plus utilisé que *may* ($237^{ème}$). Cependant, il n'est pas sûr que cela suffise à expliquer la très grande différence d'acceptation entre les deux phrases. Il est possible que, là encore, les étudiants aient tenté une analyse grammaticale et se soient dit que l'adverbe *later* devait être utilisé avec le modal du futur, *will*. Cet item a un coefficient de discrimination acceptable (0,26), ce qui montrerait encore une fois que les étudiants de niveau plus élevé se reposent sans doute moins sur ce genre de règles et plus sur leur intuition que les étudiants de niveau plus faible.

Le troisième groupe de 9 items a été conçu pour être au niveau B2/C1. Parmi eux, 5 ont été réussis par moins de 70% de l'échantillon, ce qui correspond à la difficulté attendue pour ces niveaux. Ce sont à chaque fois des structures complexes et/ou peu courantes : une phrase avec un sujet non agentif (locatif) dans *This tent sleeps four* (un exemple de ce qu'on appelle parfois la voie moyenne, entre l'actif et le passif), un verbe suivi d'une complétive réduite dans *They want the children found* (plus couramment, *They want the children to be found*), une structure avec montée du sujet dans *This game is sure to be a winner* (qui correspond à *It is sure that this game will be a winner*), un verbe prépositionnel suivi d'une proposition en V-ing avec sujet à l'accusatif (*They worried about him drinking*), et un verbe de perception suivi d'une complétive réduite avec base verbale dans *I saw them drive away* (alors que la forme en V-ing est probablement plus courante).

Inversement, trois des phrases B2/C1 ont été réussies à plus de 80%. Il s'agit de *It's amazing to think that it's true* et *Smoking is known to cause cancer*, qui sont toutes deux des structures complexes (une infinitive extraposée dans le premier cas et un passif avec infinitive

complément dans le deuxième), mais dont la proximité avec le français (*C'est incroyable de penser que c'est vrai* et *(le fait de) Fumer est connu pour causer le cancer*) a pu aider à leur reconnaissance. La troisième structure, *I took my coat off* (83% de réussite) n'existe pas en français, mais nous avons signalé dans le Tableau 8.2 ci-dessus que les verbes à particule les plus courants, dont fait partie *take off* (c'est le 28ème verbe à particule le plus courant d'après Garnier & Schmitt, 2015), sont enseignés en général au niveau A2. D'autre part, la forme *take it off* apparaît également au niveau A2 d'après *English Vocabulary Profile*. Nous avions choisi de placer cette structure en B2 parce que ce n'est qu'à ce niveau que la structure verbe + GN + particule (*take my coat off*) est utilisée de façon productive d'après *English Grammar Profile*, même si la structure verbe + pronom + particule (*take it off*) apparaît dès le niveau A2. On peut constater que, au moins en compréhension et pour ce verbe très courant, la forme avec groupe nominal complet ne pose pas de problème.

Notre tentative de trouver un parallèle entre niveau CECR et difficulté des items s'est donc soldée par un demi-échec. Nous avons certes réussi, en ne gardant que les phrases correctes, à observer une légère tendance à la difficulté croissante (et à l'augmentation du nombre d'items très difficiles) quand le niveau CECR augmente. Cependant, cette tendance s'accompagne d'une énorme variabilité à l'intérieur de chaque niveau : aux niveaux A1/A2 comme B1 et B2/C1, on trouve à la fois des items très faciles et des items difficiles voire très difficiles. Nous avons avancé plusieurs hypothèses pour expliquer la difficulté plus élevée que prévue des items A1 à B1 : le fait que l'estimation des niveaux CECR soit basée sur des données de production écrite alors que notre test utilise la compréhension de l'oral, et le fait que les étudiants faibles fassent appel à des connaissances grammaticales explicites qui les induisent en erreur. La difficulté moindre qu'attendue pour certains items B2/C1 a été expliquée, à l'inverse, par une transparence structurelle avec le français. Cependant, ce sont des hypothèses qui demandent à être confirmées.

8.4.3.3. difficulté des phrases incorrectes : influence du français, faible saillance perceptuelle et rôle des connaissances lexicales

Dans l'analyse qui précède, nous avons choisi d'écarter les phrases incorrectes qui semblaient fausser les résultats par niveau ; nous nous penchons maintenant sur ces phrases. Nous les reproduisons dans le Tableau 8.6 ci-dessous, où nous constatons qu'elles se sont révélées objectivement très difficiles : une seule phrase, **She's name was Anna*, est rejetée correctement par plus des deux tiers de l'échantillon (elle est tout de même acceptée par 18%

de nos étudiants). Toutes les autres ont entre 13 et 66% de réussite. Nous constatons encore une fois une énorme variabilité à l'intérieur de chaque niveau, et aucune corrélation entre niveau estimé et difficulté. Le niveau où l'on constate la plus grande difficulté moyenne, et le plus grand nombre de phrases très difficiles à rejeter, est A2 (les phrases incorrectes acceptées comme correctes par plus des deux tiers de l'échantillon sont surlignées en gris). Comme il paraît impossible de dégager une tendance au vu de ces résultats (sauf à dire que l'item le plus facile, souligné, se trouve bien en A1), nous nous contenterons d'émettre des hypothèses quant à l'origine des difficultés pour nos étudiants : l'influence du français, la faible saillance perceptuelle de certains marqueurs, et le manque de connaissances lexicales.

niveau	items	difficulté moyenne
A1 et A1/A2	*She's name was Anna (0.82) *I like very much sweets (0.58) *I haven't the keys (0.51) *I am boring in class (0.42) *There are differents possibilities (0.41) *I have seen it last year (0.26)	0.50
A2	*This children are lost (0.61) *I'm agree with you (0.59) *I spended last week-end in London (0.48) *The crash arrived because it was dark (0.45) *I explain you the situation (0.29) *She's one of my friend (0.25) *Have a good travel (0.16)	0.40
B1 et A2/B1	*I want that you stay (0.66) *She is here since 2 years now (0.65) *It's depend on the situation (0.58) *The house who's on the other side (0.56) *Sorry, I forget to send it (0.37) *He helped her making the cake (0.36)	0.53
B2/C1	*From what says Joe, it's not true (0.59) *This stroke me as important (0.43) *When was written this book (0.40)	0.47

Tableau 8.6 - phrases incorrectes classées par groupe de niveau et par difficulté croissante (les items réussis par moins du tiers de l'échantillon sont surlignés en gris, et par plus des deux tiers, soulignés)

hypothèse 1:influence du français

Si l'on regarde les quatre items les moins bien réussis (en gris), les phrases correspondantes couvrent le groupe nominal (**Have a good travel*, **She's one of my friend*), les temps/aspects (**I have seen it last year*), et la sous-catégorisation verbale (**I explain you the situation*). Elles ne sont donc pas cantonnées à un domaine spécifique qui pourrait être particulièrement responsable de la difficulté du test, et il faut se tourner vers d'autres explications, au premier rang desquelles se trouve l'influence du français. Nous nous plaçons ici dans une longue tradition d'analyse contrastive (S. M. Gass & Selinker, 1983). Les phrases incorrectes que

nous avons choisies correspondent en effet à des fautes fréquemment commises par les francophones du fait de l'influence du français : par exemple **I'm agree with you*, souvent entendu dans les interactions en classe, ressemble à « Je suis d'accord avec vous/toi ». Il nous semble que l'influence du français peut expliquer les très mauvais résultats d'au moins deux des quatre phrases les moins bien réussies.

Commençons par l'item le plus difficile. Dans **Have a good travel* (16% de réussite), le nom indénombrable *travel* est utilisé de façon incorrecte comme dénombrable ; or, le nom français correspondant, « voyage », est dénombrable. Le deuxième, **I have seen it last year* (26%), utilise le *present perfect* avec une référence temporelle passée, ce qui est incorrect mais correspond à l'utilisation parfaitement acceptable du passé composé français dans le même contexte : « Je l'ai vu(e) l'an dernier ». On peut s'étonner de ce dernier résultat étant donné le temps passé en classe et dans les manuels sur la différence entre le prétérit et le *present perfect* en anglais (Payre-Ficout, 2007). On peut cependant remarquer que nous l'avions placé en A1 contrairement aux indications de EAQUALS, qui indique que c'est au niveau B1 que la différence entre le prétérit (la forme correcte) et le *present perfect* (la forme qu'ont tendance à utiliser les apprenants francophones) est généralement étudiée (cf. Tableau 8.2). D'autre part, cet item a un coefficient de discrimination très légèrement négatif, ce qui signifie que les candidats forts ne l'ont pas mieux réussi que les faibles. On peut supposer qu'ici encore, les faibles ont fait appel plus que les forts à leur connaissance explicite des règles grammaticales, mais avec succès cette fois.

Ces phrases font ainsi partie de l'input régulièrement entendu en classe par les apprenants, et, s'ils jugent les phrases par rapport à leur familiarité, il est normal qu'elles leur paraissent acceptables. Nous rejoignons ainsi les conclusions de Hopp (2016), qui montre que l'input (correct et incorrect) reçu en cours de langue étrangère influence le traitement de la langue par les apprenants, qui utilisent les mêmes mécanismes que les natifs, mais travaillent à partir de données différentes. C'est peut-être ainsi que l'on peut expliquer la légère tendance des phrases les moins bien réussies à se trouver dans les niveaux les plus faciles (A1 et A2) : ce sont des phrases entendues par nos apprenants depuis le tout début de leur apprentissage, et qui leur paraissent d'autant plus naturelles qu'elles ont été entendues souvent et dans de nombreux contextes (en d'autres termes, elles sont fossilisées, Selinker, 1972).

hypothèse 2 : faible saillance perceptuelle

Une autre explication possible à la difficulté de nos items tient au caractère oral des stimuli. Le troisième item très mal réussi (**She's one of my friend*) est incorrect parce qu'il manque au nom *friend* la marque du pluriel. Il est tout à fait possible que cette absence soit plus difficile à remarquer à l'oral qu'à l'écrit. Ceci pourrait venir d'une part, encore une fois, de l'influence du français (où le marqueur de pluriel est indiqué à l'écrit mais non prononcé à l'oral), mais aussi de sa faible saillance perceptuelle (*perceptual salience*, Goldschneider & DeKeyser, 2001). Il s'agit effectivement d'un seul phonème rattaché à la syllabe qui le précède et qui peut facilement se perdre dans le flux auditif, même dans un contexte « facile » où c'est le dernier phonème de la phrase. Cela rejoint les résultats de Bell et al. (2015), qui montrent la difficulté de perception d'un morphème similaire (*-ed* du prétérit) par des apprenants L2 de niveau intermédiaire faible.

Il existe d'autres items du test où un problème de perception peut expliquer les faibles taux de réussite. **There are differents possibilities* (40% de réussite), par exemple, présente un /s/ supplémentaire (incorrect) à la fin de l'adjectif *different*. Il est possible que nos apprenants ne l'aient pas entendu (d'autant plus qu'il est suivi d'une consonne, un contexte difficile d'après Bell et al., 2015). Cependant, ce résultat peut aussi s'expliquer par l'influence du français, où le pluriel est marqué sur les adjectifs[42]. Nous avons par ailleurs déjà remarqué plus haut que la prononciation du démonstratif dans **This children are lost* pouvait poser un problème si la distinction entre *this* /ðɪs/ et these /ði:z/ n'est pas reconnue à l'oral. Le /s/ intempestif après le *it* dans **It's depend on the situation* n'est peut-être pas remarqué non plus (manque de saillance perceptive), pas plus que son absence parallèle après *depend*. Cependant, le problème peut aussi être dû à l'influence de l'input entendu en classe de langue, dans la mesure où cette erreur est assez courante dans les productions des apprenants (peut-être à cause de la surgénéralisation de la forme *It's* en début de phrase).

hypothèse 3 : connaissances morphosyntaxiques attachées à un item lexical

D'autres sources de difficulté nous semblent attachées à des items lexicaux particuliers. Il suffit alors que les apprenants connaissent moins bien cet item lexical qu'un autre qui a des caractéristiques similaires mais un comportement différent pour que l'erreur survienne. C'est typiquement le cas pour la structure dative, qui alterne entre deux formes en anglais (on appelle ce phénomène *dative alternation*). Certains verbes acceptent les deux structures (*I gave the present to her*/ *I gave her the present*), et d'autres une seule des deux (*I explained the*

[42]nous trouvons d'ailleurs régulièrement cette forme dans les productions écrites de nos étudiants

*problem to her/ *I explained her the problem*). Cela peut expliquer la difficulté de la phrase incorrecte **I explain you the situation*, correctement rejetée par 29% seulement de notre échantillon (alors que l'absence d'aspect aurait également dû pousser les étudiants à rejeter cette phrase). D'autres chercheurs ont déjà montré la difficulté du phénomène d'alternance syntaxique avec le datif (*dative alternation*) pour les L2, même de niveau avancé, avec certains verbes (R. Ellis, 1991).

Les formes irrégulières du prétérit testées par certaines de nos phrases sont également attachées à des items lexicaux spécifiques : *forget/forgot* (**Sorry, I forget to send it*, 37% de réussite), *spend/spent* (**I spended last week-end in London*, 48% de réussite) et *strike/struck* (**This stroke me as important*, 43% de réussite). La littérature sur l'acquisition des formes du prétérit anglais nous apprend que plusieurs variables influent sur leur utilisation correcte en L2 (McDonald & Roussel, 2010). Ces variables comprennent la fréquence (plus le verbe est fréquent, mieux sa forme au prétérit est réussie), la présence d'une plosive alvéolaire /t/ ou /d/ à la fin de la base verbale (ces finales entraînent plus d'erreurs, que ce soit de régularisation incorrecte de verbes irréguliers ou d'absence de suffixe alors que le verbe est régulier), ainsi que l'existence de verbes amis (qui ont les mêmes sonorités et le même comportement) ou ennemis (qui ont les mêmes sonorités mais un comportement différent). La difficulté de nos items ne peut pas s'expliquer par la première de ces variables, dans la mesure où aucun des trois verbes n'est de basse fréquence : *forget, spend* et *strike* font tous les trois partie de la bande de fréquence 1 rassemblant les 1 000 familles de mots les plus fréquentes (Nation, 2017). Par contre, *forget* et *spend* finissent respectivement par /t/ et /d/, un contexte difficile même pour des natifs. Marchman (1997) trouve par exemple que des enfants anglophones de 3 à 13 ans utilisent beaucoup plus souvent la base verbale pour le prétérit (*zero-marking*) quand les verbes se terminent par /t/ ou /d/, que ces verbes soient réguliers (*melt*) ou à changement vocalique (*feed*). C'est une erreur de surgénéralisation normale dans la mesure où tous les verbes pour lesquels le prétérit a la même forme que la base verbale se terminent par un /t/ (*hit*) ou un /d/ (*shed*). Cela pourrait expliquer pourquoi nos apprenants ne sont pas gênés par l'absence de marque à *forget* (cependant, il est possible également qu'ils n'aient pas remarqué que le contexte requérait un prétérit). Le contexte alvéolaire a donc certainement contribué à la difficulté de ces deux premiers items.

Pour expliquer la difficulté de *strike*/**stroke*, il ne reste que la troisième variable, le voisinage. Le verbe *strike* possède certes un certain nombre de voisins ennemis (*like*, *bike*, *hike, spike*), mais ce sont des verbes réguliers, qui ne devraient pas influer sur l'acceptabilité de **stroke*

comme prétérit. D'autres verbes qui présentent la même alternance vocalique, comme *ride/rode*, sont considérés comme trop éloignés pour être de véritables voisins, dans la mesure où seule la voyelle est commune (mais ils peuvent tout de même avoir une influence). Il reste ce que McDonald et Roussel (2010) appellent les « voisins de voyelle » (*vowel neighborhood*). L'idée est que l'existence de formes verbales semblables en tout point à la base d'un verbe (*strike*) à l'exception de la voyelle centrale (*streak, stroke, struck*) induit en erreur les apprenants, et peut leur faire penser que ce sont des formes passées possibles pour le verbe considéré. Même s'il est douteux que nos étudiants connaissent vraiment le verbe *stroke*, moins courant que *strike*, ils connaissent probablement le nom *stroke* (*to have a stroke, to suffer a stroke…*). Cette hypothèse reste donc assez hasardeuse, mais nous ne pouvons aller plus loin dans l'analyse faute d'items similaires sur lesquels la tester. Par ailleurs, cet item est très peu discriminant (0,09) contrairement aux deux autres sur le prétérit, ce qui montre que le problème persiste aux niveaux avancés (ce qui peut éventuellement aller dans le sens d'une influence du mot *stroke*, moins courant et donc mieux connu des étudiants plus avancés).

8.4.4. Conclusion

L'écart important de réussite entre les phrases grammaticales et agrammaticales (un résultat courant dans la littérature) nous a amenée à les examiner séparément. Nous avons trouvé une très légère tendance (non significative) à l'augmentation de la difficulté avec le niveau CECR pour les phrases grammaticales, mais pas pour les phrases agrammaticales. Ceci rejoint les observations des chercheurs du projet *English Profile* qui remarquent que ce sont plutôt les structures correctement utilisées (phrases grammaticales) qui caractérisent le développement langagier des apprenants L2. Les erreurs grammaticales persistent en effet tout au long de l'apprentissage, et ce n'est qu'aux niveaux C (auxquels peu de nos étudiants sont parvenus, et qui ne sont pas l'objet principal de cette étude orientée vers les étudiants ayant besoin de remédiation) que leur fréquence baisse de façon vraiment importante (English Profile, 2011).

Le niveau CECR (tel qu'opérationnalisé dans cette étude) semble peu à même d'expliquer les variations de difficulté de nos items (cf. aussi Prodeau et al., 2012, pour la même constatation en français langue étrangère). Nous avons eu recours à d'autres explications, dont la principale est l'influence du français, qui peut expliquer à la fois la facilité inattendue de certaines structures complexes pour lesquelles il existe une structure parallèle en français, et la difficulté à rejeter des phrases agrammaticales influencées par le français et entendues pendant toute la scolarité dans des classes d'élèves francophones. Nous avons également

constaté le rôle probable de l'oralité des stimuli, qui rend plus difficile la perception de morphèmes grammaticaux du fait de leur faible saillance perceptive, et celui des règles apprises, qui peuvent aider les apprenants aussi bien que les induire en erreur quand elles sont mal assimilées.

Nous arrivons ainsi à une conclusion paradoxale : notre test a de bonnes propriétés psychométriques (bonne fiabilité, items bien discriminants et de difficulté variée), mais notre construit n'a été que partiellement validé dans la mesure où il est difficile d'affirmer ce qui influence, précisément, les choix des sujets : analyses grammaticales (réussies ou non), phonologiques, lexicales, ou simple comparaison intuitive de la phrase entendue avec des ensembles stockés en mémoire (*exemplars*). Cependant, ces résultats ambigus sont, paradoxalement, tout à fait conformes à ceux trouvés dans la littérature : après avoir porté sur la question de savoir s'ils étaient des tests de compétence ou de performance (R. Ellis, 1991), la discussion scientifique cherche actuellement à comprendre si les tests de jugement de grammaticalité mobilisent des connaissances implicites ou explicites (Pélissier, 2018), ou implicites et « explicites automatisées » (Vafaee et al., 2017), si les phrases grammaticales et agrammaticales testent la même chose ou pas (Gutiérrez, 2013), etc. – « *It is not clear, in fact, precisely what subjects base their judgments on* », (R. Ellis, 1991, p. 164). Nous conservons pour l'instant ce test (dans sa version améliorée après élimination de quelques items non performants) et continuerons sa validation dans la partie suivante lors des analyses de corrélation avec les autres tests développés dans ce travail.

Chapitre 9

Autres tests et conclusion

9.1. PVST

Nous avons mentionné dans le deuxième chapitre de la première partie (section 2.4.1) le rôle important que pouvaient prendre les connaissances phraséologiques au moment de l'étape d'intégration, c'est-à-dire de la combinaison du sens des différents (groupes de) mots reconnus. Nous avons rappelé que certaines combinaisons de mots n'ont pas un sens compositionnel, et que le sens du tout n'est pas immédiatement déductible de celui des parties. C'est pourquoi il est important de connaître ces expressions, qui contribuent également au processus de *chunking* dont dépend l'efficacité des processus de compréhension.

Gyllstadt (2009) a développé deux tests de collocation, COLLEX et COLLMATCH, dont le premier est composé de questions à choix multiple où le candidat doit choisir la collocation la plus courante parmi trois propositions (Figure 9.1) et le deuxième (COLLMATCH) est un test oui/non où il faut décider si le groupe présenté est une combinaison de mots fréquente (une collocation) ou pas (Figure 9.2).

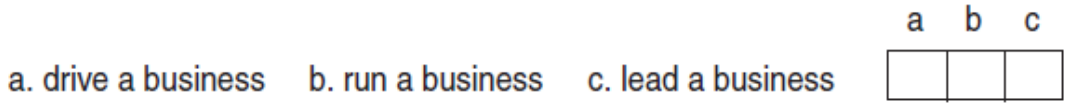


Figure 9.1 - exemple d'item du test de connaissances collocationnelles COLLEX (Gyllstad, 2009, p.157)

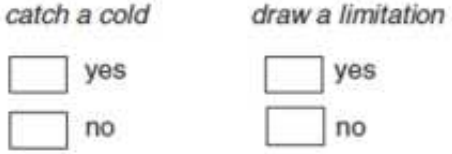
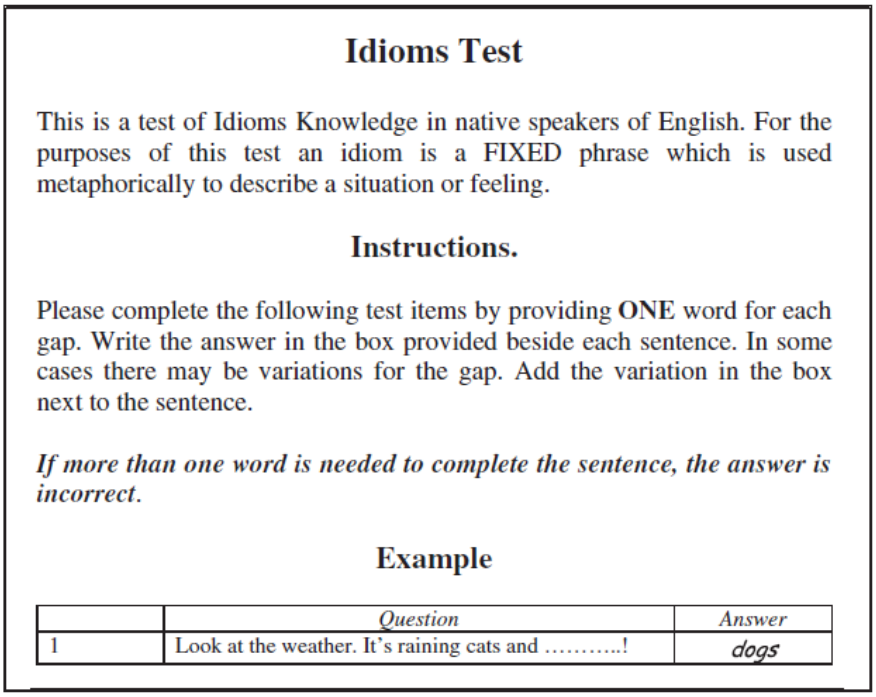


Figure 9.2 - exemple d'item du test de connaissances collocationnelles COLLMATCH (Gyllstad, 2009, p.158)

Ces tests concernent uniquement les suites verbe + nom (*say a prayer, run a business*), et seraient plus appropriés dans un contexte de production que de compréhension, dans la mesure où beaucoup de ces collocations sont compositionnelles et la difficulté n'est pas de les comprendre mais de sélectionner le verbe qui complète le mieux le nom. Citons ici un autre test tourné vers la production, celui de McGavigan (2009, cité par Milton, 2009), qui porte sur les expressions idiomatiques. Un exemple d'item de ce test est présenté en Figure 9.3. Les expressions dans ce cas ne sont pas compositionnelles, mais sont très peu fréquentes et ne semblent donc pas essentielles pour notre public en compréhension.



Idioms Test

This is a test of Idioms Knowledge in native speakers of English. For the purposes of this test an idiom is a FIXED phrase which is used metaphorically to describe a situation or feeling.

Instructions.

Please complete the following test items by providing **ONE** word for each gap. Write the answer in the box provided beside each sentence. In some cases there may be variations for the gap. Add the variation in the box next to the sentence.

If more than one word is needed to complete the sentence, the answer is incorrect.

Example

	Question	*Answer*
1	Look at the weather. It's raining cats and ……..!	dogs

Figure 9.3 - exemple d'item du test d'expressions idiomatiques de McGavigan, (2009), cité par Miton (2009, p. 152)

C'est pourquoi nous avons préféré utiliser le seul test existant (à notre connaissance) de connaissances phraséologiques générales, le *Phrasal Vocabulary Size Test* (PVST) de Martinez (2011). Le PVST est composé de 50 questions à choix multiple qui évaluent la reconnaissance du sens des expressions proposées (*meaning recognition*). Il s'appuie sur la liste PHRASE (Martinez & Schmitt, 2012) que nous avons déjà décrite à la section 2.4.1 (chapitre 2), et qui identifie 505 expressions dont la fréquence les place parmi les 5 000 mots anglais les plus courants. Il est divisé en 5 parties correspondant chacune à une bande de mille mots, et il comporte 10 questions par partie, soit 50 questions dans sa version initiale. Deux exemples sont présentés dans le Tableau 9.1, l'un de la première bande de fréquence (les mille mots les plus fréquents), et l'autre de la quatrième. Le test complet se trouve en Annexe 8.

Cette structure en bandes de fréquence (semblable à celle du VST ou du LVLT décrits plus hauts) permet d'estimer un nombre total d'expressions connues. En effet, chaque bande de fréquence, représentée par dix items, comprend un nombre d'expressions identifié (par exemple, 32 parmi la première bande de fréquence des 1 000 familles de mots les plus courantes, ou 105 parmi la cinquième bande allant de 4 000 à 5 000). Cela permet, grâce à une règle de trois, d'inférer le nombre d'expressions connues par bande de fréquence (par exemple, en multipliant le résultat à la première bande par 3,2, ou celui à la cinquième bande par 10,5). La structure par bande de fréquence permet également de vérifier le rôle de la fréquence dans la réussite aux items. Cette vérification ne semble pas avoir été prévue dans l'étude initiale de Martinez (2011), mais nous l'effectuerons sur nos propres résultats.

at all: I don't like it **at all**.	
a.	all the time
b.	in any way
c.	at first
d.	sometimes
come across: They **came across** a hotel.	
a.	stayed in
b.	opened
c.	were near
d.	found

Tableau 9.1 - deux exemples d'item du Phrasal Vocabulary Size Test (Martinez 2011), l'un plus facile, dans la première bande de fréquence, le second plus difficile, dans la quatrième bande de fréquence

Le PVST est un test écrit (le seul de notre ensemble de tests), que nous avons décidé de garder sous cette forme pour conserver un temps de passation raisonnable. Il nous paraît compléter de façon intéressante le test de reconnaissance aurale de mots isolés décrit plus haut, dans la mesure où cet autre test est basé sur les déclarations des candidats (non vérifiable dans la nomenclature de Read, 1993) alors que celui-ci, certes plus long, permet une certaine vérification de la reconnaissance du sens grâce à l'utilisation de questions à choix multiple.

Nous avons piloté ce test dans sa version papier avec un groupe de 61 étudiants en février 2017 (voir section 4.2.1 sur le plan de l'expérimentation). La passation étant assez longue (15 minutes en moyenne), nous avons décidé de faire une analyse d'items et de ne pas garder les items les moins discriminants. Etant donné que ces analyses ne concernent pas un test que nous avons nous-même développé, elles sont disponibles en Annexe 8 uniquement. Nous avons ainsi conservé un test de 40 items, que nous avons ensuite intégré à la plateforme SELF également utilisée pour les autres tests. La liste des expressions retenues est présentée dans le Tableau 9.2 par bande de fréquence. La dernière colonne du tableau contient le nombre total

d'expressions identifiées par l'étude de Martinez (2011) dans chaque bande de fréquence, et indique le nombre d'expressions représentées par chaque point du PVST (calculé en divisant le nombre total d'expressions par bande de fréquence par le nombre d'items dans cette bande).

bande de fréquence	expressions du test	n	nb total expressions dans bande
1	go on lead to so that at all I mean at least be likely to deal with used to	9	32 (chaque point "vaut" 3,6 expressions)
2	so far to do with take over in particular for instance as soon as be about to be expected to	8	84 (1 pt = 10,5)
3	give up feel like turn out other than all over in touch at once in time	8	129 (1 pt = 16,5)
4	prove to next door run out in light of by no means come across happen to even so	8	157 (1 pt = 16,1)
5	by far straight away turn down (to be) to blame take for granted as of can tell	7	103 (1 pt = 14,7)

Tableau 9.2 – liste et nombre d'expressions utilisées, valeur de chaque point par bande de fréquence dans le Phrasal Vocabulary Size Test (PVST, Martinez 2011)

Dans cette version modifiée et en ligne, le PVST dure environ dix minutes. Un inconvénient du raccourcissement du test est que chaque expression représente un nombre plus important d'expressions de sa bande de fréquence, mais cela reste dans des proportions raisonnables : comme on le voit dans le Tableau 9.2, chaque item du test représente au plus 16,5 expressions (dans un test comme le VST de Nation et Beglar, 2007, qui comporte 10 items par bande de fréquence de 1 000 familles, chaque item représente 100 familles de mots). Cependant, il ne nous semble pas particulièrement intéressant d'estimer le nombre d'expressions connues par nos étudiants (contrairement à la taille du vocabulaire, par exemple) : comme ce chiffre n'est pas couramment utilisé, il ne représente rien pour les étudiants, et un score en pourcentage de mots connus par bande de fréquence paraît aussi parlant.

Nous terminerons cette présentation du PVST avec les résultats par bande de fréquence visibles dans le Tableau 9.3. On constate effectivement une difficulté croissante à mesure que la fréquence des expressions diminue, sauf pour la bande 4 qui est un peu mieux réussie que la bande 3. Un test de Kruskal-Wallis confirme que ces différences sont significatives (Kruskal-Wallis χ^2 = 20,5, df=4, p <0,001), et un test de Wilcoxon apparié (avec ajustement Bonferroni) nous apprend que seules la différence entre les résultats aux items de la première

bande et ceux des autres bandes est significative (p <0,05 pour la différence entre les bandes 1 et 2, p<0,01 pour les bandes un et quatre, et p<0,001 pour les autres différences). Il y a donc bien un effet de la fréquence, même si ce dernier ne se révèle que quand on compare les expressions extrêmement fréquentes et les autres.

	bande 1	bande 2	bande 3	bande 4	bande 5
moyenne (é.-t.)	7,77 (2,35)	6,82 (2,17)	6,30 (2,35)	6,46 (2,29)	6,21 (2,33)

Tableau 9.3 - difficulté moyenne par bande de fréquence du PVST (expérimentation de février 2017), chaque bande contenant dix items

9.2. Compréhension de l'oral: test de positionnement SELF

La compétence de compréhension de l'oral qui est la pierre de touche de notre dispositif (et la variable à expliquer dans nos analyses statistiques) est évaluée par le test SELF, qui a été développé au sein du projet Innovalangues (Masperi 2012), auquel nous avons participé. Il est déployé en anglais depuis la rentrée universitaire 2015. C'est un test de positionnement en ligne qui utilise un algorithme semi-adaptatif : tous les étudiants passent le même minitest initial, composé de 36 questions, et sont ensuite, en fonction de leur résultat, envoyés vers un autre groupe de 24 items qui peuvent être de niveau faible (niveau CECR A1-A2), moyen (B1) ou avancé (B2-C1). Même si chaque candidat répond à 60 questions (36 dans le minitest initial et 24 dans la deuxième étape), tous les étudiants ne sont pas exposés aux mêmes questions, ce qui permet de limiter le temps de passation du test (60 minutes en moyenne) en proposant aux candidats des questions de niveau adapté. Cette architecture est explicitée dans la Figure 9.4.

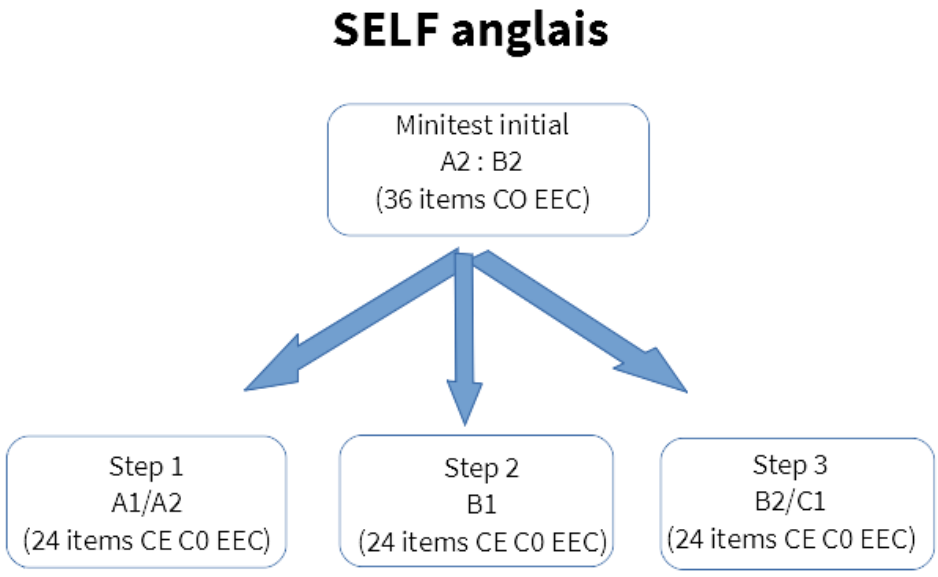


Figure 9.4 - structure du test de positionnement SELF anglais

Le test SELF évalue non seulement la compréhension de l'oral, mais également celle de l'écrit (CE), ainsi que l'expression écrite courte (EEC). L'écran de sortie du test de positionnement fournit à l'étudiant une proposition de groupe de niveau dans lequel il peut s'inscrire, ainsi qu'un score par sous-compétence (un exemple est proposé dans la Figure 9.5). Nous n'utiliserons que le score de la sous-compétence de compréhension de l'oral, qui est calculé avec les 18 questions de CO du minitest initial auxquelles s'ajoutent 9 items de CO pour ceux qui sont envoyés vers l'étape A1/A2, 8 pour l'étape B1 et 7 pour l'étape B2/C1 (soit un total de 27, 26 ou 25 questions selon le niveau). Nous ne pourrons donc pas utiliser pour les analyses qui suivront le score total de CO de l'étudiant : un score de 20 n'aura pas du tout le même sens pour un étudiant passé par l'étape A1/A2 que pour un autre passé par l'étape B2/C1. C'est pourquoi nous utiliserons, pour certaines analyses, le niveau de la sous-compétence de CO proposé en sortie (A1, A2, B1, B2 ou C1), et pour d'autres, la valeur du logit (estimation du trait latent) calculée avec un modèle de réponse à l'item dont nous avons expliqué le principe dans le troisième chapitre de la première partie.

Figure 9.5 - écran de sortie du test de positionnement SELF: proposition de groupe et scores par sous-compétence

Le test SELF aspire à tester la compétence communicative, malgré les contraintes qu'impose le format en ligne autocorrectif avec questions à choix multiple (Cervini & Jouannaud 2015). Cela se traduit par une attention particulière vis-à-vis de la contextualisation et de l'authenticité. Les documents utilisés sont dans la mesure du possible des textes authentiques (c'est-à-dire non créés pour des apprenants de langue étrangère), surtout aux niveaux élevés, et les questions requièrent aussi bien la compréhension fine du lexique (exemple de niveau B2, Tableau 9.5), que la compréhension de fonctions de communication, par exemple par inférence (exemple de niveau A2, Tableau 9.4). Comme de courts extraits sont utilisés, une contextualisation est proposée par un champ « contexte ». Les deux exemples qui suivent illustrent la structure des items du test, qui comprennent donc le contexte, le texte proprement

dit, la question à laquelle il faut répondre, et des propositions de réponse dont une seule est correcte (la clé), et les autres sont incorrectes (les distracteurs). Lors de l'administration du test, l'ordre de la clé et des distracteurs est aléatoire.

champs	
contexte	*Jack is talking to Sarah about another housemate, Frank*
texte	*Jack: Do you know where Frank is, Sarah?* *Sarah: Well, I heard music from his room earlier.*
question	*What does Sarah mean?*
proposition de réponse 1 (clé)	*Frank is probably in his room.*
proposition 2 (distracteur)	*Frank forgot to turn the music off.*
proposition 3 (distracteur)	*Frank's loud music bothers Sarah.*

Tableau 9.4 - exemple d'item de compréhension de l'oral du test SELF de niveau A2, avec un focus pragmatique

champs	
contexte	*A documentary*
texte	*On the West shore of Hudson's Bay, the low rolling terrain offers little protection against the bitter cold of the Northwest wind. Tyupak and Agiyutak, having left their families in the igloos of the winter camp, are coming in to the post to trade. They look forward to a mug up of tea and pilot biscuits at the trader's house.*
question	*Why did the two Eskimos leave their families?*
proposition de réponse 1 (clé)	*They are going to buy and sell things*
proposition 2 (distracteur)	*They are looking for protection*
proposition 3 (distracteur)	*They have nothing left to eat*

Tableau 9.5 - exemple d'item de compréhension de l'oral du test SELF de niveau B2, avec un focus lexical (*trade*)

Les concepteurs ont fait le choix du « tout à l'oral », ce qui signifie que la question et les propositions de réponse pour chaque item sont proposées à l'oral et non à l'écrit. On peut voir ci-dessous un exemple de l'interface du test pour les tâches de compréhension de l'oral (Figure 9.6). Le seul écrit qui apparaît est en français (« Reste 1 écoute »), sous l'icône du dialogue à écouter. Les autres icônes ne sont pas commentées (les étiquettes rectangulaires ont été rajoutées ici à l'image pour expliciter leur signification), mais une vidéo de tutoriel est disponible avant de commencer le test ou pendant la passation pour les participants qui souhaitent des éclaircissements ou ne trouvent pas l'interface intuitive. Quand une seule question porte sur le texte, ce dernier ne peut être écouté qu'une seule fois, mais la question et les propositions de réponse peuvent être écoutées un nombre illimité de fois, y compris avant l'écoute du texte proprement dit. Cela constitue une entorse à l'authenticité interactionnelle,

dans la mesure où un auditeur ne sait pas toujours exactement ce qu'il cherche à entendre dans une situation de communication authentique, mais cela paraît un compromis acceptable dans un test dont la durée est très limitée.

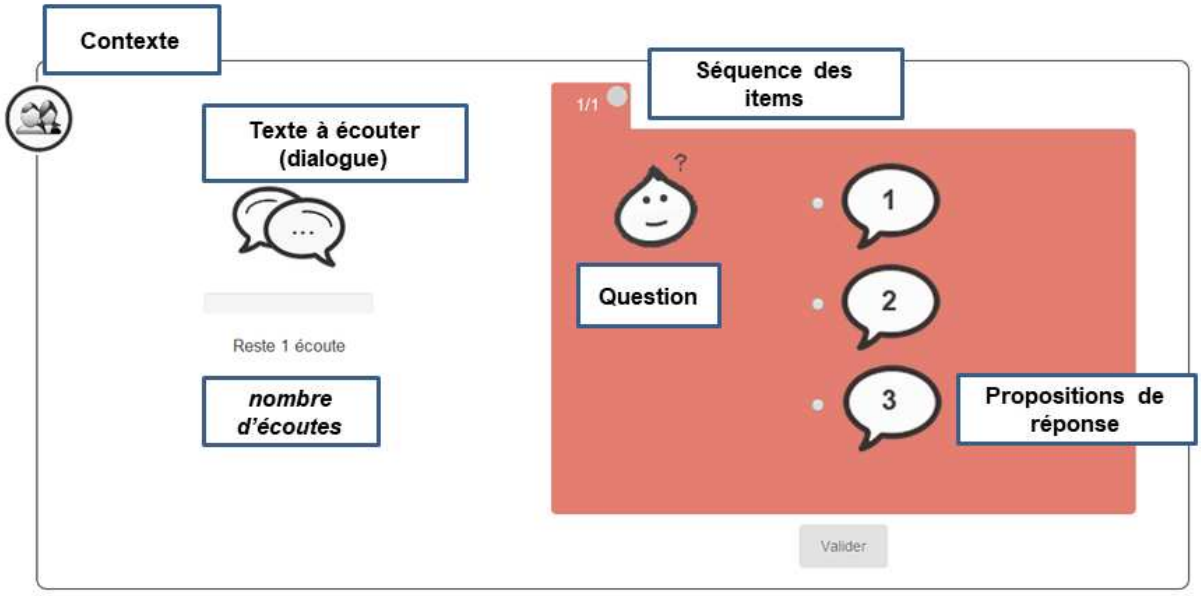


Figure 9.6 - interface du test SELF (compréhension de l'oral)

L'un des avantages du test SELF de notre point de vue est que tous les étudiants qui s'inscrivent en première année d'anglais doivent le passer au moment de leur inscription administrative, ce que tous font, à l'exception de quelques retardataires. Nous disposons donc de données presque complètes pour cette population, sans avoir besoin d'organiser de nouvelles passations. Ce test est également utilisé pour les étudiants du public LANSAD qui choisissent de suivre des cours d'anglais au Service des Langues de l'université, même si un nombre non négligeable d'entre eux échappe à la passation suite à des questions d'organisation : inscription tardive, changements d'option, défaut d'information, etc.

9.3. Conclusion de la deuxième partie

Nous rappelons ici les trois premières questions de recherche que nous avons soulevées à la fin de la première partie et qui concernent la validation de nos instruments de mesure diagnostiques :

- Des tests de discrimination phonémique, de conscience prosodique centrée sur la perception de l'accent, de reconnaissance aurale du vocabulaire, de jugement de grammaticalité aurale et de connaissances phraséologiques produisent-ils des scores

fiables (du point de vue de la cohérence interne), et permettent-ils de discriminer les apprenants de niveau faible de ceux de niveau élevé ?

- Ces tests sont-ils par ailleurs utilisables en pratique dans les conditions qui président à cette étude, c'est-à-dire en ligne, en autonomie, dans un temps total voisin d'une heure ?
- Les scores sont-ils interprétables au vu des résultats présentés dans la littérature sur l'acquisition ; en particulier, les mots fréquents sont-ils mieux reconnus ou analysés, les contrastes ou unités qui existent également dans la L1 sont-ils mieux réussis que les autres ?

Nous avons pu pour chacun de ces tests répondre par l'affirmative à la première question. Tous nos tests sont fiables, avec un alpha de Cronbach supérieur à 0,7, et avec des taux de discrimination pour les items dont ils sont composés allant d'acceptable (supérieur à 0,2) à excellent (supérieur à 0,4 ou 0,5). Nous avons calculé la valeur moyenne du coefficient de discrimination des items pour chacun des tests, qui est supérieur à 0,3 pour tous ; ces valeurs se trouvent dans la quatrième colonne du Tableau 9.6 ci-dessous, qui résume les propriétés de tous les tests diagnostiques utilisés. L'étendue des scores est également satisfaisante, même si certains tests (notamment celui de jugement de grammaticalité, qui présente une petite asymétrie positive) sont peut-être un peu trop difficiles pour notre population et manquent un peu de sensibilité dans les scores faibles.

Pour ce qui est de la deuxième question, tous ces tests ont été administrés en ligne, et, à part un problème ponctuel de serveur pour une des passations (qui a affecté l'enregistrement des résultats mais apparemment pas la passation elle-même), aucun incident n'a été constaté. Cependant, il est important de souligner que presque tous les tests supposent le traitement de matériau sonore et qu'ils nécessitent l'utilisation d'écouteurs et de casques de bonne qualité. C'était en général le cas dans nos expérimentations étant donné qu'elles ont eu lieu dans les laboratoires de langues de notre université, mais c'est un point sur lequel il faudra être vigilant si les tests sont passés à terme, comme prévu, à distance, depuis le domicile des étudiants ou n'importe quel autre lieu de leur choix.

La faisabilité du passage en autonomie des tests n'a pas été étudiée en détail ni expérimentée, puisque tous les tests ont été passés en notre présence. Pour chaque groupe, les étudiants recevaient d'abord une explication à l'oral des buts de l'expérimentation, ainsi que deux feuilles : d'une part le questionnaire biographique langagier qu'ils devaient remplir, et d'autre

part une feuille de route (disponible en Annexe 1) détaillant les tests à passer, ainsi que le rappel des instructions de chacun de ces tests (instructions accessibles également sur la page d'accueil de chacun d'entre eux). Le guidage était donc plus important que lors d'une passation en autonomie. Ainsi, la quasi-totalité des étudiants ont choisi de passer les tests dans l'ordre dans lequel ils étaient présentés sur la feuille de route, même si aucune consigne n'avait été donnée à ce propos. Il est donc certain que les étudiants se sont servis du guidage proposé. Cependant, nous avons l'expérience du passage de tests en autonomie avec le test de positionnement SELF, qui est passé tous les ans à distance par les étudiants LLCER au moment de leur inscription administrative au mois de juillet ou d'août (voire début septembre), sans qu'ils aient besoin de se déplacer sur le campus. Pour la plupart de ces étudiants, la passation se passe très bien et ne donne lieu à aucun guidage particulier (hormis l'envoi par messagerie électronique des codes de connexion et du calendrier de passation, et du lien vers un mode d'emploi). Cependant, un petit nombre d'étudiants chaque année rencontre des difficultés d'ordre technique ou organisationnel et contactent l'UFR pour essayer de les résoudre, ou choisissent de venir passer les tests sur place. Il est probable que la passation de nos tests diagnostiques serait vécue à peu près de la même façon, et il faudra prévoir un guidage plus poussé pour certains étudiants qui en font la demande. Nous reviendrons sur ce point dans la troisième partie, dans le chapitre sur les conséquences pédagogiques.

Enfin, le temps de passation total de nos tests nous paraît acceptable, puisque, si on ajoute le temps de passation moyen de chaque test, nous arrivons à un total d'un peu plus de 40 minutes en moyenne, comme on peut le constater dans la deuxième colonne du Tableau 9.6. Certains étudiants prennent beaucoup plus de temps, mais nous avons vu qu'il n'y avait pas forcément de corrélation entre le temps passé et la réussite aux tests, ce qui nous encourage à penser que notre public cible (les étudiants faibles ayant besoin de remédiation) ne sera pas nécessairement susceptible d'être celui ayant besoin de plus de temps. Nous avons par ailleurs calculé le temps total par étudiant ayant passé les cinq tests, et l'étudiant le plus lent a mis 95 minutes, soit un peu plus d'une heure et demie. Au final, ces chiffres nous paraissent satisfaisants et pourront être communiqués aux étudiants sous la forme suivante : en moyenne, le passage des cinq tests diagnostiques prend 45 minutes (un peu moins de 10 minutes par test), et une heure trente au maximum.

instrument	tps moyen (min)	alpha (fiabilité)	discrim. moy	unidim.	influence fréq. lex.	influence prox. fr.
PHON	6,4	0,76	0,37	oui		oui
PROSO	11,2	0,9	0,35	oui	non	
AURLEX	7,8	0,84	0,34	oui	oui	oui
AURGRAM	6,5	0,85	0,40	oui		oui
PVST	10	0,93	0,45	oui	oui	oui
tot	42 min					

Tableau 9.6 - résumé des propriétés des tests diagnostiques développés : temps de passation moyen, fiabilité (α de Cronbach), coefficient de discrimination moyen des items du test, unidimensionnalité, influence constatée de la fréquence lexicale et de la proximité du français

La dernière question portait sur la validité de construit de nos tests. L'idée était de montrer que ces derniers, bâtis à partir de spécifications inspirées des résultats de la recherche en acquisition des L2, se sont effectivement comportés comme on s'y attendait. Nous nous sommes intéressée en particulier à deux variables : la fréquence lexicale et la proximité avec le français, qui est la langue maternelle de la quasi-totalité de nos sujets (colonnes six et sept du Tableau 9.6). Pour cette deuxième variable, nous avons montré qu'elle jouait un rôle facilitateur dans trois des tests considérés : dans le test de discrimination phonémique, les contrastes présentés comme difficiles pour les francophones dans la littérature (parce qu'ils n'existent pas en français ou sont assimilés à un seul phonème en français) ont effectivement été les moins réussis. Pour le test de reconnaissance aurale du vocabulaire, les mots transparents pour des francophones ont été mieux reconnus que les autres. Enfin, pour le test de jugement de grammaticalité, nous avons remarqué que les structures syntaxiques parallèles à celles du français pouvaient expliquer le taux de réussite important à certains items, et inversement, le phénomène de transfert pouvait sous-tendre l'acceptation erronée de structures incorrectes en anglais.

L'influence de la fréquence lexicale a été plus difficile à mettre en évidence. Dans le test AURLEX, elle n'est apparue qu'en utilisant une liste de fréquence non lemmatisée (Brysbaert & New, 2009), plutôt que basée sur les familles de mots (Nation, 2017), ce qui va dans le sens de la recherche en acquisition des L2 (McLean, 2018). Dans le test de sensibilité prosodique, il n'a pas été plus facile pour nos étudiants d'identifier le schéma accentuel des mots courants que des mots rares (cependant, ce résultat correspond en partie à la théorie de « surdité accentuelle » chez les francophones, Dupoux et al., 1997). Enfin, nous n'avons pas cherché à valider nous-même l'utilisation du PVST que nous n'avons pas développé personnellement et qui a déjà fait l'objet d'une étude de validation (Martinez, 2011), mais nous avons pu y mettre en évidence un certain effet de la fréquence (des collocations cette fois). Dans l'ensemble,

l'influence de la fréquence est donc visible, sauf quand un traitement lexical s'opère au détriment du traitement phonologique.

Chapitre 10

Etude corrélatoire

La dernière partie de cette thèse est composée de deux chapitres avec une focalisation assez différente. Le chapitre 10 présente les résultats de l'étude corrélatoire qui met en rapport les résultats aux différents instruments diagnostiques présentés dans la deuxième partie. Ces analyses statistiques sont au cœur de notre travail, puisqu'elles permettront ensuite de tirer les conclusions qui seront exploitées dans le chapitre 11. Ce dernier portera sur les conséquences pédagogiques de nos résultats pour l'enseignement et l'évaluation de la compréhension de l'oral dans le contexte universitaire français. Nous ne proposerons pas une nouvelle didactique de la compréhension aurale, mais simplement des actions complémentaires de remédiation en autonomie, avec l'idée que le travail sur les composantes d'un processus complexe peut conduire à des améliorations non négligeables.

10.1. Rappel des questions de recherche

Dans la deuxième partie de cette étude, nous avons décrit la conception de cinq tests diagnostiques, dont nous avons montré qu'ils étaient fiables, qu'ils avaient une étendue de difficulté acceptable et qu'ils discriminaient bien entre étudiants en difficulté ou non. Nous les avons analysés séparément, mais ce qui nous intéresse à présent, c'est de les utiliser pour étudier la contribution relative de chacune des compétences qu'ils représentent à la compétence plus globale de compréhension de l'oral, dans une perspective diagnostique. Dans ce chapitre, nous allons ainsi mettre en rapport les résultats de nos étudiants aux différents instruments décrits et analysés dans la partie précédente. Rappelons que nous souhaitons étudier le rôle de la discrimination phonémique, de la sensibilité prosodique, de la reconnaissance aurale du vocabulaire, du jugement aural de grammaticalité, et des connaissances phraséologiques dans la compréhension de l'oral en anglais langue étrangère. Toutes ces variables sont des variables explicatives (ou indépendantes), que nous voulons

utiliser pour expliquer la variable de compréhension de l'oral (variable à expliquer, ou dépendante). En fin de première partie, nous avions ainsi identifié les questions de recherche suivantes (nous avons ici inversé la première et la deuxième question) :

- Dans quelle mesure les tests diagnostiques sont-ils corrélés avec les résultats en compréhension de l'oral, ce qui est un premier pas vers la confirmation expérimentale de leur rôle dans cette compétence ?
- Tous les tests diagnostiques sont-ils corrélés entre eux ?
- Quel est leur rôle respectif ? Les études que nous avons résumées trouvent en général que les connaissances lexicales sont les plus corrélées avec la compréhension de l'oral ; retrouvons-nous ce résultat ? Qu'en est-il des connaissances phraséologiques, qui n'ont pour l'instant pas été testées dans ce cadre ?

Pour répondre à la première question de recherche, qui porte sur la corrélation entre les résultats aux différents tests diagnostiques pris séparément et le niveau de compréhension de l'oral (CO) estimé par le test de positionnement SELF, nous avons besoin d'une mesure d'association entre deux variables. Les tests diagnostiques génèrent des résultats numériques, mais le test mesurant le niveau en CO utilise l'échelle du CECRL, avec pour valeurs possibles A2, B1, B2, ou C1 (aucun étudiant n'ayant reçu le niveau A1, et le test SELF ne faisant pas la différence entre les niveaux C1 et C2). Le « niveau SELF CO » n'est donc pas une variable numérique, mais une variable ordinale, avec des valeurs non numériques qui peuvent être rangées de la plus petite à la plus grande. Pour étudier le lien entre chaque test diagnostique (variable numérique) et la compréhension de l'oral (variable ordinale), nous ne pourrons donc utiliser de corrélation classique de type Pearson, qui suppose que les deux variables soient numériques (même si l'une des deux peut aussi être dichotomique). Nous utiliserons le *tau* de Kendall, noté τ. Ce coefficient permet de calculer la force de l'association entre une variable numérique et une variable ordonnée (Howell, 2009, p. 306). Comme le *rho* de Spearman, qui est plus connu, le *tau* de Kendall transforme les valeurs des deux variables en rangs, mais elle est plus robuste que *rho* quand il y a beaucoup d'ex-aequo (rangs identiques), ce qui est notre cas puisque nous n'avons que quatre valeurs possibles pour les rangs en CO (chaque niveau CECR correspondant à un rang). Le *tau* de Kendall offre en outre l'avantage de pouvoir déterminer une valeur pour *p* (la probabilité que les résultats constatés puissent être compatibles avec l'hypothèse nulle).

Pour l'estimation globale de la taille de l'effet, les lignes directrices de Cohen (1992), résumées dans Larson-Hall (2009, p. 119) sont souvent utilisées en sciences humaines : corrélation de petite taille à partir de 0,1, de taille moyenne à partir de 0,3, et de grande taille à partir de 0,5. Cependant, Plonsky et Oswald (2014) ont montré que dans le champ de la recherche en acquisition des L2, ces estimations devaient être revues à la hausse : une corrélation de petite taille commencera à 0,25, de taille moyenne à 0,4 et de taille importante à 0,6. Pour ce qui est des pourcentages de variance expliquée, les chiffres correspondants (obtenus par quadrature des chiffres de corrélation) sont de 3% (variance commune faible), 16% (variance commune moyenne), et 36% (importante variance commune).

Pour aller plus loin dans l'analyse de nos données, nous utiliserons des tests d'analyse de variance ANOVA. Ces tests statistiques nous permettront de comparer les moyennes obtenues à chaque test diagnostique par les étudiants de différents niveaux CECR en CO et de savoir si les différences de moyenne éventuellement constatées sont statistiquement significatives ou si elles peuvent être dues au hasard. Ensuite, comme un test ANOVA ne nous dit pas si la différence entre chaque paire de moyennes prises deux par deux est également significative, nous effectuerons un test post-hoc des étendues de Tukey avec correction pour comparaisons multiples (voir section 4.3 pour un résumé des tests statistiques principaux utilisés et leurs conditions de validité). Enfin, pour évaluer le pourcentage de variation expliquée associé à l'analyse de variance, nous utiliserons la statistique êta-carré (Howell, 2009, p. 344). Les conditions d'utilisation d'une ANOVA (Howell, 2009, p. 325-326) sont la normalité des résidus du modèle linéaire et l'homogénéité des variances (qui doivent varier dans un facteur de moins de 4, c'est-à-dire que l'écart-type doit varier dans un facteur de 2 au maximum). S'il s'avère que ces conditions d'utilisation d'une ANOVA ne sont pas respectées, nous nous tournerons vers le test non-paramétrique de Kruskal-Wallis, qui travaille sur les rangs et non sur les valeurs numériques, et résiste donc bien à la violation de normalité (Howell, 2009, p. 683). Pour savoir si les moyennes prises deux par deux sont également différentes, nous nous tournerons vers un test post-hoc de Wilcoxon apparié avec correction Bonferroni pour comparaisons multiples.

Notre deuxième question de recherche porte sur les corrélations existant entre tous les tests proposés. Nous nous attendons à ce qu'ils soient tous corrélés entre eux, dans la mesure où les habiletés qu'ils recouvrent ont des caractéristiques communes, notamment le traitement de l'anglais oral. Cependant, nous voudrions également vérifier qu'ils constituent des construits séparés, c'est-à-dire que les tests diagnostiques mesurent vraiment des capacités ou

connaissances différentes. Cette vérification sera importante en vue de notre troisième question de recherche. Cette dernière a trait à l'exploration du rôle respectif des différentes variables explicatives mesurées par nos tests dans la compréhension de l'oral. Pour tenter de répondre à cette question, nous utiliserons une technique de régression, la régression logistique, que nous présenterons plus loin.

10.2. Méthode : sujets et instruments (rappels)

Nous avons organisé une expérimentation auprès d'étudiants en première année de licence LLCER anglais (n = 124) et d'étudiants de licence inscrits dans des cours d'anglais LANSAD (n = 66) au premier semestre de l'année universitaire 2018-2019. Nous avons supposé que ces publics ne sont pas encore foncièrement différents à ce stade de leurs études. En effet, même s'ils ont certainement un meilleur niveau qui les a conduits à choisir de poursuivre des études supérieures en licence d'anglais, les étudiants LLCER ne peuvent pas encore être qualifiés de « spécialistes » de la langue anglaise après quelques semaines d'étude seulement. Nous avons montré lors de la description de nos groupes de sujets (chapitre 4 sur le plan de l'expérimentation, section 4.2.4) qu'ils avaient effectivement des caractéristiques biographiques similaires en ce qui concerne l'âge au moment de l'expérimentation (respectivement 19 et 20 ans en moyenne), le sexe (74 et 75% de filles) et l'âge auquel ils disent avoir commencé les cours d'anglais à l'école (9,5 et 10 ans). Cependant, les étudiants LLCER ont passé en moyenne beaucoup plus de temps à l'étranger (près de 3 mois, contre 10 jours pour les LANSAD, même si dans les deux cas le mode est à 0).

Ces sujets ont passé les tests de discrimination phonémique (ci-après PHON), de conscience prosodique (PROSO), de reconnaissance aurale du vocabulaire (AURLEX), et de jugement de grammaticalité aurale (AURGRAM), ainsi que le *Phrasal Vocabulary Size Test* (PVST), mesurant les connaissances phraséologiques. Ils ont également passé le test de positionnement SELF en début d'année, dont nous n'avons gardé que le résultat en compréhension de l'oral (SELF_CO).

10.3. Résultats

10.3.1. Résultats globaux

Les résultats globaux sont présentés dans le tableau qui suit (Tableau 10.1). Nous constatons encore une fois que les tests ont bien fonctionné : ils ont un coefficient alpha (fiabilité par

cohérence interne) qui va d'acceptable (0,76 pour le test de discrimination phonémique) à excellent (0,93 pour le PVST). Ils présentent une large étendue de scores, avec pour trois d'entre eux (AURLEX, PROSO et PVST) des étudiants qui atteignent le plafond, et des scores qui ne sont qu'à un ou deux points du maximum pour les deux autres (PHON et AURGRAM). Les deux tests les mieux réussis sont ceux qui ont trait aux traitements de bas niveau (PHON et PROSO), avec des taux de réussite autour de 70% (72 et 69%) alors que les tests portant sur la reconnaissance ou la compréhension du vocabulaire et de la grammaire (AURLEX, AURGRAM et PVST) ont posé plus de difficulté, avec un taux de réussite autour de 60% (61, 58 et 59%).

		échantillon complet						**LLCER**			**LANSAD**		
mesure	k	n	moy (é.t.)	moy en %	étend.	se	α	n	moy	éc.-type	n	moy	écart-type
AURLEX	41	183	24.96 (6.77)	61	9 - 41	0.50	.84	117	27.71	5.80	66	20.09	5.54
PHON	32	183	23.12 (4.51)	72	5 - 30	0.33	.76	117	23.72	4.44	66	22.06	4.46
AURGRAM	33	184	19.21 (6.13)	58	8 - 32	0.45	.85	118	22.03	5.47	66	14.17	3.41
PROSO	56	180	38.38 (9.92)	69	13 -56	0.74	.9	115	41.57	8.97	65	32.74	9.02
PVST	40	174	23.77 (9.4)	59	7 - 40	0.71	.93	116	28.05	7.73	58	15.21	5.98

Tableau 10.1 - résultats des groupes de sujets aux différents tests diagnostiques développés pour cette étude, avec pour chaque test le nombre d'items (k), d'étudiants (n), moyennes et écarts-types, étendue des scores et alpha de Cronbach

Tous les étudiants n'ayant pas passé le test de positionnement, il ne nous reste que 158 sujets (122 LLCER et 36 LANSAD), répartis dans les niveaux CECR en compréhension de l'oral indiqués dans le Tableau 10.2. Nous remarquons que les étudiants LLCER ont de bien meilleurs résultats que ceux mentionnés dans l'introduction de cette thèse : dans l'étude de Payre-Ficout (2011, portant sur une cohorte d'étudiants de l'année universitaire 2009-2010), moins de 25% des primo-entrants en L1 LLCER atteignaient le niveau B2 en CO, alors qu'ils sont ici 62% (76 sur 122). Il est possible que la population s'inscrivant en première année d'anglais ait changé entre temps (9 ans séparent les deux cohortes), ou que le niveau de CO ait augmenté (du fait de l'accès facilité aux films ou séries anglophones en version originale, par exemple). Cependant, un tel écart s'explique probablement aussi par les instruments différents utilisés dans les deux études, DIALANG dans la première et SELF dans la présente étude. Quant aux étudiants LANSAD, ceux dont nous disposons des résultats de positionnement se trouvent tous dans les groupes A2 ou B1. Les enseignants LANSAD qui

ont accepté de participer à notre étude avaient donc essentiellement des groupes de niveaux globalement faibles.

self_CO	**tous**							**LLCER**					**LANSAD**			
niveau	A2	B1	B2	C1	tot	A2/B1	B2/C1	A2	B1	B2	C1	tot	A2	B1	B2	C1
nb.	38	44	63	13	158	82	76	5	41	63	13	122	33	3	0	0

Tableau 10.2 - résultats du test SELF en compréhension de l'oral pour l'échantillon total et chacun des sous-échantillons (LLCER et LANSAD) : nombre d'étudiants par niveau CECR

Le nombre d'étudiants est relativement équilibré dans chacun des niveaux (38, 44 et 63 en A2, B1 et B2), à l'exception du groupe de niveau C1 qui compte peu de sujets (13). Cet équilibre est plus flagrant quand on compare l'échantillon d'étudiants de niveau insuffisant, c'est-à-dire A2 ou B1 (82 sujets), et celui de niveau suffisant (B2 ou C1, 76 sujets).

Nous allons maintenant nous pencher sur les relations entre chacune de nos variables diagnostiques et le niveau de CO. Présentons tout d'abord les résultats pour tous les tests en fonction de ce niveau. Nous avons réparti les étudiants en quatre groupes selon leur niveau de compréhension de l'oral (variable à expliquer), correspondant aux niveaux A2, B1, B2 et C1/C2 (nommé ici C1) du CECR. Le Tableau 10.3 présente les résultats des quatre groupes aux cinq tests diagnostiques.

	AURLEX (k=41)		*AURGRAM (k=33)*		*PHON (k=32)*	
	M (sd)	étendue	M (sd)	étendue	M (sd)	étendue
"A2" (n=38)	18.5 (4.2)	9-27	12.9 (2.7)	8-18	21.7 (4.4)	8-27
"B1" (n=43)	25.9 (4.6)	16-40	19.3 (4.6)	10-30	22.7 (4.3)	10-29
"B2" (n=63)	28.7 (5.2)	20-41	23.5 (4.9)	11-32	24 (4.6)	8-30
"C1" (n=13)	32.6 (3.7)	27-41	26.4 (5.3)	14-32	26.3 (2.1)	22-29

	PROSO (k=56)		*PVST (k=40)*	
	M (sd)	étendue	M (sd)	étendue
"A2" (n=37)	33.1 (8.7)	13-52	13.7 (4.9) (n=34)	8-26
"B1" (n=43)	38.1 (9.2)	16-54	23.6 (7.4)	9-37
"B2" (n=63)	42.1 (8.5)	17-56	30.5 (6.1)	12-39
"C1" (n=11)	49.4 (4.5)	44-55	34.9 (5)	25-40

Tableau 10.3 - résultats aux tests de discrimination phonémique, de sensibilité prosodique, de reconnaissance du vocabulaire aural, de jugement de grammaticalité aural et de connaissances phraséologiques en fonction du niveau de compréhension de l'oral (SELF CO)

On observe dans chaque colonne M que les moyennes aux cinq tests augmentent avec le niveau, ce qui correspond à nos prédictions. On constate également, dans les colonnes « étendue », que pour presque toutes les variables, les scores minimum et maximum augmentent avec le niveau (les deux exceptions étant d'une part le test PHON où pratiquement tous les groupes semblent avoir la même étendue de résultats, et d'autre part le

test PROSO où des étudiants de tous les groupes s'approchent du maximum). Pour toutes les variables sauf PHON, l'écart-type des résultats de chaque groupe de niveau ne varie pas dans un facteur supérieur à 2 (même si AURGRAM s'en approche avec l'écart-type le plus bas à 2,7, et le plus élevé à 5,3), ce qui veut dire qu'une des conditions liées à l'utilisation d'une ANOVA est respectée. Pour PHON, par contre, l'écart-type le plus haut (4,6 pour le groupe B2) est plus du double de celui de C1 (2,1), ce qui nous interdira probablement d'utiliser un test ANOVA classique (ce que nous vérifierons plus loin).

Nous allons à présent observer ces résultats plus en détail pour chaque test. Nous commencerons par visualiser les données graphiquement pour chaque variable avec des diagrammes en boîtes à moustache qui permettent de comparer visuellement les tendances centrales et la dispersion de la variable numérique pour chaque valeur de la variable catégorielle (pour nous, A2, B1, B2, C1 en compréhension aurale). Suite à ces visualisations, nous effectuerons les tests de corrélation et de significativité mentionnés plus haut.

Rappelons ici nos hypothèses et les résultats trouvés dans d'autres études de ce type (résumés précédemment en 2.6). Nous nous attendons à trouver une corrélation faible avec les connaissances phonémiques, conformément à Zoghlami (2015), Hilton et al. (2016) et Wilson et al. (2011), les trois études que nous ayons trouvées qui testent cette corrélation. Certaines études que nous avons résumées dans le deuxième chapitre parlent également du problème de passage à l'échelle (*scaling up*) lorsqu'on passe du niveau de traitement phonémique au niveau de traitement lexical. Cependant, ces études ne testent pas en général empiriquement la corrélation entre les deux.

Pour ce qui est des connaissances prosodiques, nous nous attendons également à une corrélation faible avec la compréhension de l'oral. Nous avons dans le deuxième chapitre résumé les deux seules études trouvées sur la question, l'une ne trouvant pas de corrélation (Meerman et al., 2014), et l'autre trouvant une corrélation moyenne (Tabata, 2016), avec des sujets japonophones.

La corrélation entre les connaissances lexicales et la compréhension de l'oral est très bien établie et nous nous attendons à trouver une corrélation moyenne à importante. Par contre, nous n'avons pas trouvé d'études sur la corrélation avec les connaissances phraséologiques. Une étude sur la compréhension de l'écrit trouvant une corrélation importante, nous nous attendons à ce que cette corrélation existe pour la compréhension de l'oral également. Pour ce

qui est des connaissances grammaticales, les corrélations trouvées variant de faibles à importantes, nous nous attendons à reproduire ces résultats, avec un niveau moyen.

10.3.2. Discrimination phonémique et compréhension de l'oral

Commençons par le test de discrimination phonémique (PHON). L'association avec la CO mesurée par le tau de Kendall est de $r_\tau = 0{,}28$, $p<0{,}001$. La corrélation est donc significative, mais faible.

Analysons maintenant les résultats plus en détail. Conformément à ce que nous avions noté dans le Tableau 10.3, la Figure 10.1 montre que la dispersion de PHON est grande à l'intérieur de chaque niveau (sauf en C1 où les résultats sont clairement concentrés en haut du graphique). Une tendance à la hausse se dessine cependant en passant d'un niveau à l'autre. On constate par ailleurs que les étudiants pour lesquels nous n'avons pas d'informations de niveau pour la compréhension de l'oral (dernière colonne à droite, notés « NA ») ont des résultats qui paraissent assez similaires à ceux des étudiants A2 ou B1, avec une dispersion semblable (même si leur médiane est légèrement inférieure).

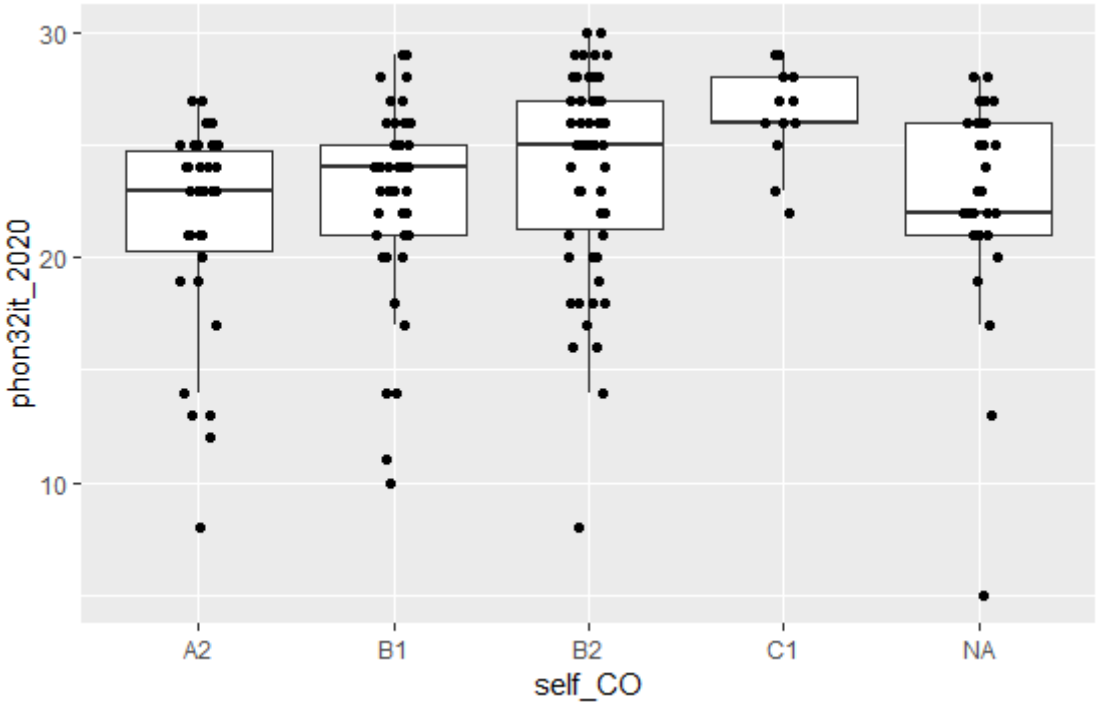


Figure 10.1 - résultats au test de discrimination phonémique en fonction du niveau CECR de compréhension de l'oral

Afin de savoir si les différences entre les moyennes des quatre groupes sont statistiquement significatives, nous avons effectué un test de Kruskal-Wallis (nous n'avons pas pu réaliser de

test ANOVA car les résidus ne sont pas normalement distribués). Les résultats montrent que la différence de moyenne entre les quatre groupes est statistiquement significative, $\chi^2(3) = 19{,}25$, $p<0{,}001$. Pour savoir si la différence entre chaque paire de niveaux pris deux par deux est également significative, nous avons effectué un test post-hoc de Wilcoxon apparié avec correction Bonferroni pour comparaisons multiples. Nous trouvons que la différence entre A2 et B1 n'est pas significative, ni celle entre B1 et B2, ni celle entre B2 et C1. Pour constater une différence entre sous-groupes, il faut monter de deux niveaux : la différence entre A2 et B2 est statistiquement significative, $p<0{,}01$ (et a fortiori entre A2 et C1), ainsi que celle entre B1 et C1 ($p<0{,}01$). On peut donc affirmer qu'au fur et à mesure que le niveau de compréhension de l'oral augmente, la capacité de discrimination phonémique a tendance à augmenter elle aussi. Cependant, cette tendance, quoique statistiquement significative quand on augmente de deux niveaux CECR, n'est pas très marquée.

A l'occasion de ces analyses des résultats de tests de comparaisons de moyennes, nous aimerions réfléchir à la proposition d'un score de césure pour séparer les étudiants ayant besoin de remédiation de ceux qui n'en ont probablement pas besoin. Etant donné que notre niveau de référence est le niveau B2, nous voudrions idéalement que tous les sujets A2 et B1 soient séparés des sujets B2 et C1 par ce score. Cependant, la dispersion des résultats de PHON dans tous les niveaux est si importante qu'un tel score de césure qui effectuerait une partition, même très imparfaite, n'existe pas. C'est pourquoi nous ne proposerons pas de score de césure pour PHON pour l'instant.

10.3.3. Sensibilité prosodique et compréhension de l'oral

La sensibilité prosodique est elle aussi corrélée avec la compréhension de l'oral : le tau de Kendall est de $r_\tau = 0{,}38$, $p<0{,}001$. La corrélation est donc significative, et un peu plus importante que pour la discrimination phonémique (même si elle reste assez faible).

Nous constatons de nouveau dans la Figure 10.2, représentant les scores au test de sensibilité prosodique (PROSO) en fonction du niveau CECR en CO, une grande variabilité à l'intérieur de chaque niveau (seuls les quelques étudiants de niveau C1 ont des résultats clairement regroupés en haut du tableau), mais des médianes qui augmentent tout de même régulièrement à l'occasion du passage d'un niveau à l'autre. Afin de savoir si les différences entre les moyennes des quatre groupes sont statistiquement significatives, nous avons effectué un test ANOVA (dont les résidus sont cette fois normalement distribués, test de Shapiro-Wilk :

w=0,99, p=0,19). Les résultats montrent que la différence de moyenne entre les quatre groupes est statistiquement significative, F(3,145) = 13,74, p<0,001. Pour savoir si la différence entre chaque paire de niveaux pris deux par deux est également significative, nous avons effectué un test post-hoc des étendues de Tukey avec correction pour comparaisons multiples. Nous trouvons, comme pour le test de connaissances phonémiques, que la différence de moyenne entre un niveau CECR et le niveau immédiatement supérieur n'est pas statistiquement différente de 0 (sauf pour le passage de B2 à C1, p<0,05). La différence de moyenne quand on augmente de deux niveaux ou plus (A2 à B2, A2 à C1 ou B1 à C1), est, elle, statistiquement significative (p<0,001 dans les trois cas). Le calcul de la statistique êta-carré nous permet d'évaluer le pourcentage de variance expliquée : nous trouvons ici η^2 = 0,22, soit une proportion assez faible, mais non négligeable, de variance expliquée.

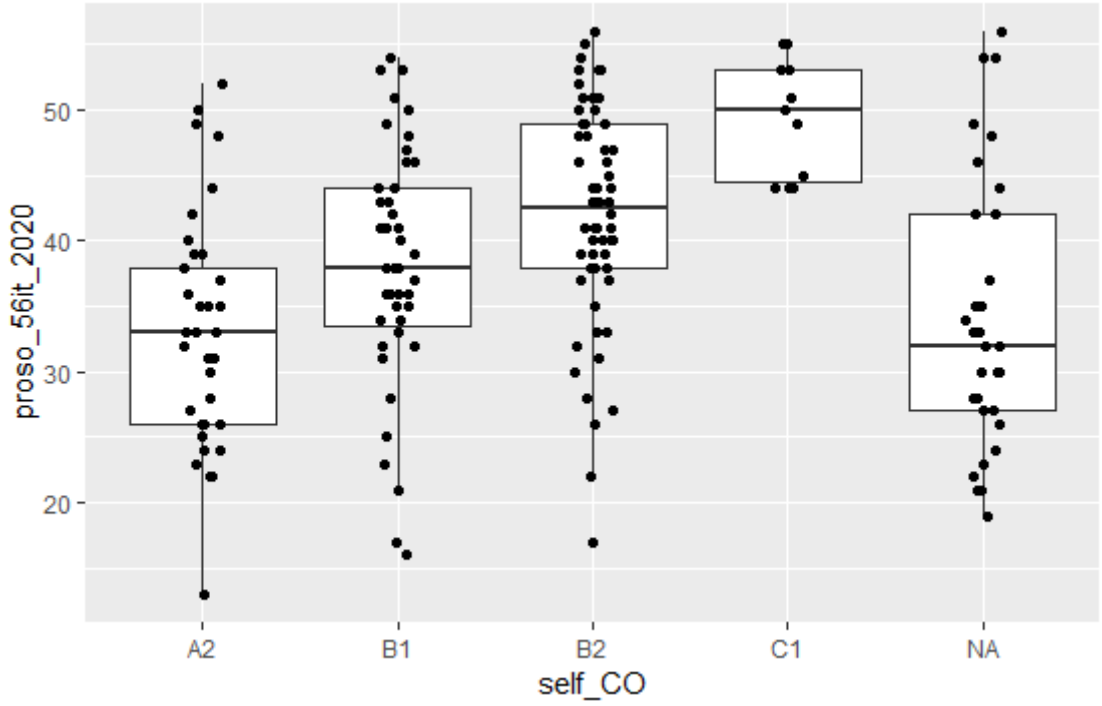


Figure 10.2 - résultats au test de sensibilité prosodique en fonction du niveau CECR de compréhension de l'oral

La proposition d'un score de césure pour le test PROSO, qui aiderait à distinguer entre les étudiants ayant besoin de remédiation et les autres, est peut-être légèrement plus facile que pour le test de discrimination phonémique. On observe par exemple sur la Figure 10.2 que, si l'on choisit le score médian des B1 (38), les trois quarts de l'échantillon A2 sont compris en dessous de ce score (puisqu'il correspond au troisième quartile de la boîte à moustache A2), et un quart de l'échantillon B2 seulement (et aucun C1). Cependant, cette première approche n'est que provisoire. Nous reviendrons sur cette question après l'étude corrélatoire générale.

10.3.4. Reconnaissance aurale du vocabulaire et compréhension de l'oral

AURLEX est elle aussi significativement corrélée avec la compréhension de l'oral : le tau de Kendall est de $r_\tau = 0,55$, $p<0,001$. La corrélation est donc plus importante que pour les deux tests précédents : c'est une corrélation moyenne à importante d'après Plonsky et Oswald (2014).

Nous constatons dans la Figure 10.3, représentant les scores en fonction du niveau CECR en CO, une variabilité à l'intérieur de chaque niveau moins grande que pour les deux tests précédents, et qui semble augurer d'une meilleure séparation entre les niveaux. Les médianes augmentent clairement d'un niveau à l'autre, sauf pour les deux niveaux centraux (B1 et B2) où l'augmentation est très minime.

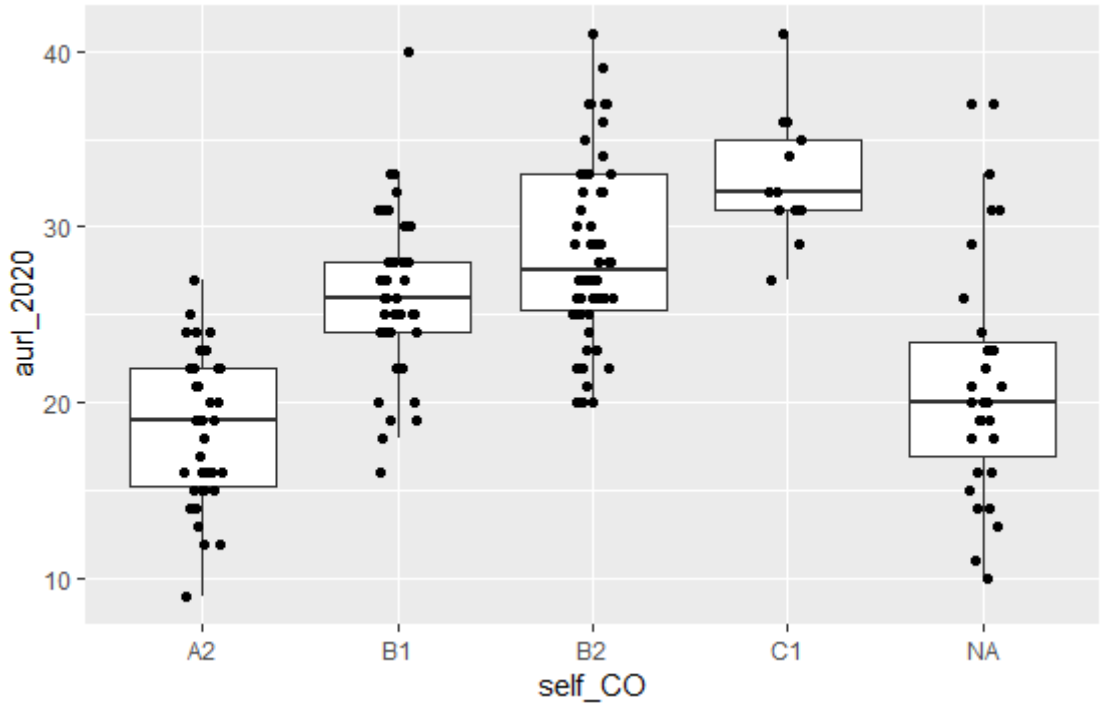


Figure 10.3 - résultats au test de reconnaissance aurale du vocabulaire en fonction du niveau CECR de compréhension de l'oral

Afin de savoir si les différences entre les moyennes des quatre groupes sont statistiquement différentes de 0, nous avons effectué un test ANOVA (dont les résidus sont normalement distribués, test de Shapiro-Wilk : w=0,99, p=0,23). Les résultats montrent que la différence de moyenne entre les quatre groupes est statistiquement significative, $F(3,148) = 47,01$, $p<0,001$. Le test post-hoc des étendues de Tukey avec correction pour comparaisons multiples montre que toutes les différences sont statistiquement significatives à $p<0,001$, sauf pour le passage de B1 à B2 et de B2 à C1, où elles sont également significatives, mais avec $p<0,05$. A mesure

que s'élève le niveau de CO, le résultat moyen en reconnaissance aurale du vocabulaire augmente donc également significativement. Le calcul de la statistique êta-carré nous permet d'évaluer le pourcentage de variance expliquée : nous trouvons ici $\eta^2 = 0,49$, soit une proportion beaucoup plus importante que pour la sensibilité prosodique.

Quatre valeurs très proches peuvent être proposées pour un score de césure entre étudiants ayant besoin de remédiation et les autres : le score maximal en A2 (qui est de 27), ce qui permettrait d'inclure tous les étudiants très faibles ; le troisième quartile de B1 (28), qui permet d'inclure la majeure partie de ces étudiants qui restent d'un niveau insuffisant ; le premier quartile de B2 (25), qui permet de ne pas inclure trop d'étudiants de niveau suffisant ; et la valeur minimale en C1, qui est la même que le maximum en A2 et permet donc d'exclure tous les étudiants de bon niveau. Cette dernière valeur (27) paraît la plus adaptée pour l'instant. Elle conduit à l'inclusion de près de la moitié de l'effectif B2 dans les étudiants ayant besoin de remédiation, ce qui peut paraître excessif. Cependant, il est tout à fait possible que cela corresponde à un besoin réel d'approfondissement des connaissances lexicales chez ces étudiants.

10.3.5. Connaissances phraséologiques et compréhension de l'oral

Les connaissances phraséologiques mesurées par le PVST sont elles aussi significativement corrélées avec la compréhension de l'oral : le tau de Kendall est de $r_\tau = 0,62$, $p<0,001$. La corrélation est donc assez importante, un peu plus que celle observée avec les connaissances lexicales (mots isolés).

La Figure 10.4 permet de constater encore une fois une augmentation claire de la médiane des résultats au passage d'un niveau à l'autre, une dispersion assez importante, en particulier aux niveaux B1 et B2, et quelques résultats au plafond en C1. Le test ANOVA sur les résultats du PVST en fonction du niveau de CO montre que la différence de moyenne entre les trois groupes est statistiquement significative, $F(3,142) = 63,84$, $p<0,001$ (les résidus du modèle sont normalement distribués, test de Shapiro-Wilk $W = 0,99$, $p = 0,31$). Le test post-hoc des étendues de Tukey avec correction pour comparaisons multiples montre que toutes les moyennes de groupes sont significativement différentes à $p<0,001$, sauf la différence entre B2 et C1 qui n'est pas significative ($p =0,14$). Le calcul de la statistique êta-carré nous permet d'évaluer le pourcentage de variance expliquée : nous trouvons ici $\eta^2=0,58$, soit une proportion importante de la variance.

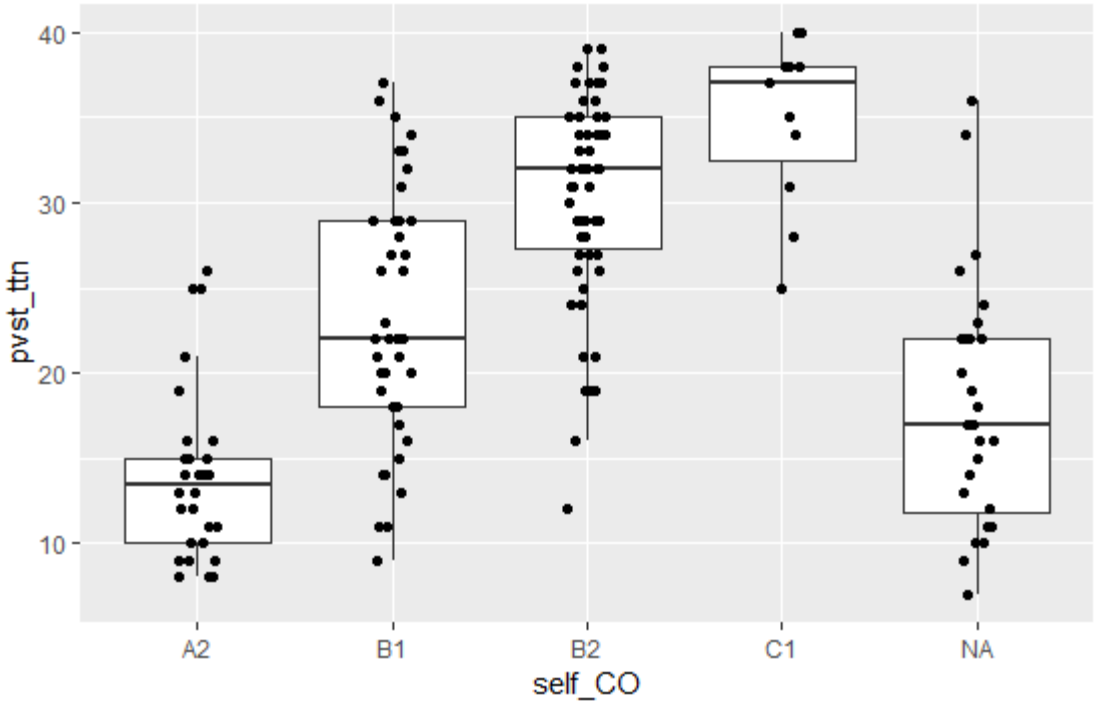


Figure 10.4 - résultats au test de connaissances phraséologiques en fonction du niveau CECR de compréhension de l'oral

Pour le score de césure, nous pourrions proposer encore une fois le minimum du groupe C1 (25), qui est presque égal au maximum du groupe A2 (26), englobe plus de la moitié des résultats B1, et moins du quart des B2.

10.3.6. Jugement aural de grammaticalité et compréhension de l'oral

Les connaissances grammaticales sont elles aussi significativement corrélées avec la compréhension de l'oral, le tau de Kendall est de $r_\tau = 0{,}61$, $p<0{,}001$. Ce résultat est très proche de celui obtenu avec les connaissances phraséologiques.

La Figure 10.5 présente les résultats au test de jugement de grammaticalité aurale (AURGRAM) en fonction du niveau de compréhension de l'oral. On constate encore une fois une grande dispersion des scores de B1 et B2, et une valeur aberrante en C1 qui serait plus à sa place en A2. Nous avons d'ailleurs remarqué que le/la même étudiant(e) avait également eu le résultat le plus bas du groupe C1 au PVST. Rappelons que le test de positionnement SELF est passé en autonomie par la majorité des étudiants, dont certains peuvent tricher malgré l'absence d'enjeu apparent. Il est donc possible que le niveau de CO attribué dans ce cas ne corresponde pas à la compétence effective. C'est pourquoi nous avons décidé de ne pas prendre en compte les résultats de ce/tte candidat/e dans la suite de nos analyses (nous aurons

donc un sujet de moins). Cependant, la médiane augmente clairement d'un niveau à l'autre (un peu moins de B2 à C1), et plusieurs étudiants obtiennent le score maximal en B2 et C1. Les résultats test ANOVA (dont les résidus sont normalement distribués, test de Shapiro-Wilk : w=0,99, p=0,23) montrent que la différence de moyenne entre les quatre groupes est statistiquement significative, F(3,149) = 54,59, p<0,001. Le test post-hoc des étendues de Tukey avec correction pour comparaisons multiples montre comme pour le PVST que toutes les différences sont statistiquement significatives à p<0,001, sauf pour le passage de B2 à C1, où la différence de moyenne n'est pas statistiquement significative (p=0,13). Le résultat moyen en jugement de grammaticalité aurale augmente donc également significativement avec le niveau de CO. Le calcul de la statistique êta-carré nous permet d'évaluer le pourcentage de variance expliquée : nous trouvons ici η^2 = 0,55, soit une proportion importante de la variance.

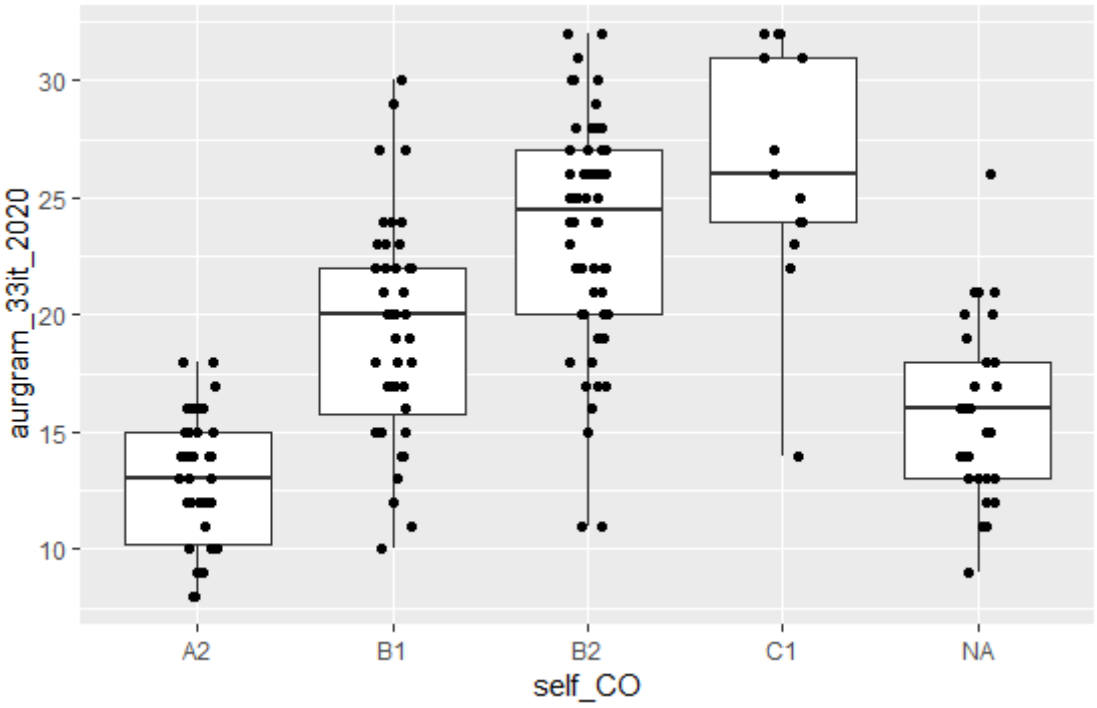


Figure 10.5 - - résultats au test de jugement de grammaticalité en fonction du niveau CECR de compréhension de l'oral (y compris le résultat aberrant d'un candidat C1)

Un score de césure possible pour AURGRAM serait la médiane de B1 (20), qui permet d' « attraper » tous les étudiants A2, la moitié donc des B1, pas plus du quart des B2 et aucun C1 (une fois exclue la valeur aberrante mentionnée plus haut).

10.3.7. Conclusion intermédiaire

Nous savons donc qu'il existe pour les cinq tests considérés une différence entre les résultats des différents groupes de niveau, en fonction de la compétence de compréhension de l'oral. Nous avons utilisé deux méthodes pour mettre ce résultat en évidence. D'une part, nous avons utilisé un simple coefficient de corrélation, le *tau* de Kendall (approprié parce qu'une de nos deux variables est une variable ordinale). Signalons que ces corrélations ne peuvent pas être comparées directement avec des corrélations de type Pearson (utilisées dans les articles que nous avons résumés dans la deuxième partie) dans la mesure où le *tau* de Kendall donne typiquement des valeurs légèrement inférieures à celles de Pearson (Howell, 2009, p. 306). Cependant, c'est surtout l'ordre de grandeur des corrélations, ainsi que leur taille respective pour nos différentes variables, qui nous intéressent ici.

Pour la discrimination phonémique, nous confirmons le résultat de Wilson et al. (2011), Zoghlami (2015) et Hilton (2016) d'une corrélation faible ($r_\tau = 0,28$), qui explique très peu de variance commune entre les connaissances phonémiques et la compréhension de l'oral. Pour la prosodie, nous confirmons les résultats de Tabata (2016) d'une corrélation faible à moyenne ($r_\tau = 0,38$). Pour les trois derniers tests, la corrélation avec la compréhension de l'oral est moyenne à importante (0,55, 0,62 et 0,61). Ce résultat était déjà connu pour les connaissances lexicales et grammaticales, comme nous l'avons rappelé plus haut. Cependant, c'est pour le test de connaissances phraséologiques (PVST) que la corrélation est la plus haute ($r_\tau = 0,62$). Il s'agit ici d'un nouveau résultat, qui fait le pendant de celui trouvé par Kremmel et al. (2017) pour la compréhension de l'écrit, où la corrélation observée entre compréhension de l'écrit et connaissances phraséologiques était légèrement plus importante qu'avec les connaissances lexicales proprement dites. Il est possible également (dans notre cas) que le test PVST apporte un peu plus d'informations parce qu'il est un peu plus long (dix minutes au lieu de sept environ pour AURLEX et AURGRAM), et qu'il suppose la reconnaissance du sens des expressions et pas uniquement de leur forme (même si l'accès à la forme peut entraîner celui au sens).

Ces résultats ont été confirmés en tous points (et résumés dans le tableau Tableau 10.4) par le calcul du coefficient êta carré (η^2) qui permet d'estimer le pourcentage de variance expliquée associé à une analyse de variance. Nous avons également trouvé que les variables PHON et PROSO expliquent peu de variance (0,22 au maximum, pour PROSO), et que AURLEX, PVST et AURGRAM en expliquent une proportion importante, environ la moitié (entre 0,45 et 0,58), avec encore une fois le PVST qui explique plus de variance que AURGRAM, qui à son tour en explique plus qu'AURLEX (donc le même ordre que pour les corrélations). Nous

avions déjà observé cette partition en deux groupes, d'une part PHON et PROSO, et d'autre part AURLEX, AURGRAM et PVST, au moment de notre description des résultats bruts. Les tests PHON et PROSO ont effectivement été mieux réussis (10% de plus en moyenne) que les tests AURLEX, AURGRAM et PVST. On peut donc imaginer que les tests portant sur les processus de bas niveau (traitement du signal sonore), mieux réussis par les étudiants faibles, discriminent moins entre les faibles et les forts.

test	PHON	PROSO	AURLEX	PVST	AURGRAM
tau (r_τ)	0,28	0,38	0,55	0,62	0,61
êta-carré (η^2)		0,22	0,45	0,58	0,55

Tableau 10.4 - valeurs de corrélation entre chacune des variables explicatives et la variable à expliquer (CO), mesurées par le tau de Kendall, et proportion de variance associée à chaque analyse de variance (êta-carré)

Nous avons par ailleurs pu observer que les tests diagnostiques peinaient parfois à distinguer les niveaux CECR adjacents. Cette difficulté n'est pas vraiment gênante pour la distinction B2/C1 (le fait que l'échantillon C1 soit très réduit n'arrange probablement pas les choses), où la différence de moyenne n'est pas statistiquement significative dans deux tests sur cinq (PHON et AURGRAM), et où p est compris entre 0,05 et 0,01 dans deux autres cas (bien que le niveau alpha 0,05 soit généralement accepté en sciences humaines et sociales, c'est une probabilité d'écarter à tort l'hypothèse nulle parfois jugée excessive, V. E. Johnson, 2013). Les étudiants C1 comme B2 ont un niveau considéré comme suffisant et peuvent donc être rangés dans le même groupe puisque ni les uns ni les autres n'ont besoin de remédiation. La différence entre A2 et B1, qui n'est pas non plus significative dans deux tests sur cinq (les deux tests les mieux réussis par les étudiants faibles, à savoir PHON et PROSO), nous intéresse plus. Même si ces deux groupes ont besoin de remédiation, il peut être pertinent d'un point de vue pédagogique de faire une distinction entre étudiants faibles et très faibles (nous pouvons d'ailleurs remarquer que d'autres études n'arrivent pas non plus à différencier étudiants A2 et B1, par exemple Kremmel, 2017, pour les connaissances lexicales).

La différence qui est la plus cruciale pour notre étude est la différence entre B1 et B2, qui marque le basculement du groupe « niveau insuffisant » au groupe « niveau suffisant ». Pourtant, deux de nos tests (PHON et PROSO toujours) n'arrivent pas à les distinguer. Cet état de fait peut avoir plusieurs explications. Une première possibilité est que le test de positionnement utilisé (SELF CO) ne donne pas de résultats assez précis et ne distingue pas bien entre les niveaux B1 et B2 pour la compréhension de l'oral. Une deuxième est que ce sont nos tests diagnostiques qui ne sont pas assez sensibles (bien que, dans les deux cas, les tests aient fait l'objet d'études de validation qui ont montré leur fiabilité et leur utilité). Il est

également possible qu'il y ait un effet seuil et qu'une fois acquise une capacité minimale à distinguer les phonèmes ou les schémas prosodiques, ce sont d'autres connaissances qui deviennent plus importantes pour la CO. Cependant, le fait que PHON et PROSO ne distinguent pas non plus entre étudiants A2 et B1 ne plaide pas en faveur de cette hypothèse. Une dernière possibilité est que la distinction entre B1 et B2 ne recouvre effectivement pas de différences de connaissances phonémiques ou prosodiques suffisamment importantes pour être visibles avec notre taille d'échantillon et que la différence se fasse, soit sur d'autres connaissances, soit par une utilisation différente (par exemple, plus rapide ou automatique) de ces connaissances.

C'est pourquoi nous pensons qu'il pourrait être intéressant de prendre en compte, au final, le temps de réponse aux items de certains tests, qui permettrait d'évaluer indirectement l'automaticité de certains processus. Par ailleurs, nous avons peut-être trop restreint le construit de notre test de sensibilité prosodique. Dans sa version actuelle, il teste essentiellement la sensibilité à l'accentuation lexicale. Il aurait pu être intéressant d'y inclure plus d'items de sensibilité prosodique au sens large, en particulier pour tester la segmentation lexicale, à l'image des tout derniers items (qui demandent par exemple à l'utilisateur de choisir entre *my key* et *Mikey*).

10.4. Exploration des relations entre toutes les variables

Comme annoncé, nous commencerons par observer les relations entre toutes nos variables (les cinq variables explicatives PHON, PROSO, AURLEX, PVST et AURGRAM et la variable à expliquer CO) en utilisant une matrice de corrélation (Figure 10.6). Bien que SELF_CO soit une variable ordinale, et que certaines des autres variables, quoique numériques, ne suivent pas une loi normale, nous utiliserons le coefficient de Pearson. En effet, d'après Falissard (2011, p. 36), une matrice de corrélation Pearson peut être utilisée avec profit même avec des variables non numériques ou suivant une loi non normale lors de l'observation de leurs relations, du moment qu'aucun test statistique n'est prévu sur ces mêmes variables.

Cette matrice (bibliothèque *GGally* de *R*) présente à la fois les valeurs de corrélation entre les variables prises deux à deux (au-dessus de la diagonale), la forme générale de la courbe correspondant à l'histogramme des valeurs prises par chaque variable (sur la diagonale), et les nuages de points qui représentent graphiquement les relations entre deux variables (sous la diagonale). Elle nous permet de confirmer certaines observations faites précédemment. Les

courbes de nos tests nous rappellent que seule AURLEX a une distribution quasi-normale, qu'AURGRAM présente une asymétrie positive (peu de détails dans les scores faibles), et qu'au contraire PHON et PROSO présentent une asymétrie négative (plus de détails dans les scores faibles). Le PVST semble avoir une tendance bimodale, avec un pic dans les scores faibles et un autre dans les scores élevés. Par ailleurs, les nuages de points permettent bien de visualiser la différence entre les variables numériques de nos tests diagnostiques, et la variable ordinale qu'est SELF_CO, répartie sur quatre verticales correspondant à ses quatre valeurs possibles (A2, B1, B2 et C1).

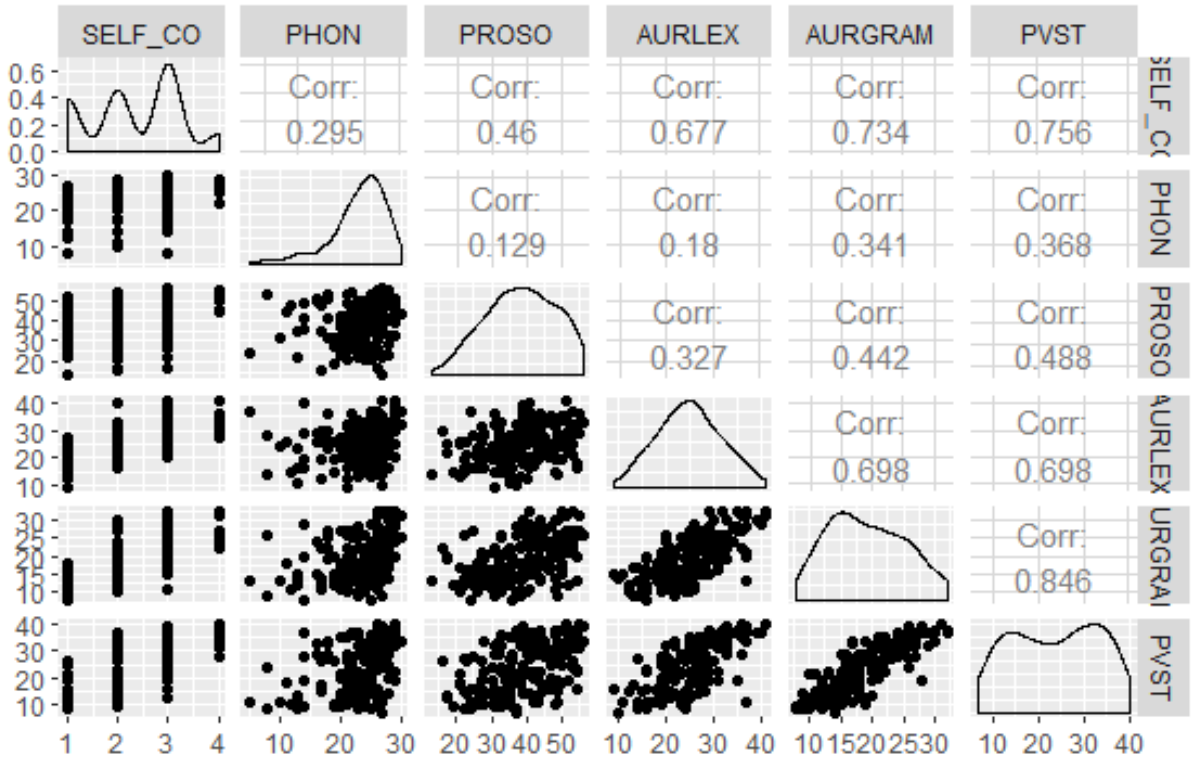


Figure 10.6 - matrice de corrélation entre les variables de l'étude (PHON, PROSO, AURLEX, PVST, AURGRAM et SELF_CO)

En ce qui concerne les corrélations, nous retrouvons sur la première ligne la hiérarchie de corrélations avec SELF_CO observée précédemment en utilisant le *tau* de Kendall (les valeurs sont ici légèrement supérieures puisqu'elles utilisent le *r* de Pearson) : corrélation faible avec PHON (0,3), moyenne avec PROSO (0,46), et importante avec AURLEX, AURGRAM et PVST (entre 0,68 et 0,76), la corrélation la plus importante étant observée avec le PVST. De manière schématique, on peut dire que plus l'on « monte » dans les niveaux de traitement (et donc plus les unités traitées sont grandes), plus la corrélation avec la CO est élevée. Cela peut paraître logique dans la mesure où l'on s'approche de plus en plus du produit fini de la compréhension (la prise en compte du sens et la création d'un modèle de

situation). Comme les premiers niveaux de traitement du signal sont plus éloignés du produit fini, d'autre processus peuvent intervenir entretemps (comme le *lossy chunking*, ou la compensation par processus descendants) et affaiblir la corrélation avec la compréhension finale.

Dans la deuxième ligne du tableau, nous constatons que PHON est très peu corrélé aux autres tests et que la corrélation avec PROSO, en particulier, est quasiment inexistante (0,13, donc en dessous du seuil de Plonsky et Oswald, 2014, pour une corrélation « faible »). Il semble que les deux compétences mesurées par PHON et PROSO soient en partie orthogonales, et qu'il n'y ait pas de lien entre la capacité à discriminer les phonèmes et celle d'identifier l'accent lexical. Cela peut paraître surprenant dans la mesure où ces deux tests évaluent le traitement du signal sonore aux niveaux inférieurs. Certaines études sur la dyslexie en L1 (Goswami et al., 2010) peinent d'ailleurs également à trouver une corrélation entre discrimination phonémique et sensibilité prosodique chez les enfants dyslexiques ou non, malgré les théories de l'acquisition L1 qui supposent par exemple que les acquisitions prosodiques sont un préalable aux acquisitions phonémiques (Pierrehumbert, 2003). La sensibilité prosodique, elle (troisième ligne du tableau), est plus clairement liée à AURLEX, AURGRAM et PVST. La corrélation (faible, 0,33) avec AURLEX est attendue dans la mesure où PROSO mesure essentiellement la sensibilité à l'accent lexical. La corrélation (moyenne, 0,44 et 0,49) avec AURGRAM et PVST est peut-être plus surprenante parce qu'elle est plus importante qu'avec AURLEX. Cependant, trouver une corrélation entre sensibilité prosodique et connaissances phraséologiques et grammaticales n'est pas aberrant, dans la mesure où la prosodie d'une phrase est partiellement liée à sa structure grammaticale (Cutler et al., 1997).

Enfin, cette matrice nous montre surtout que trois de nos variables explicatives sont très corrélées entre elles (lignes quatre et cinq). AURGRAM et PVST, en particulier, ont une corrélation de 0,85. AURLEX est un peu moins corrélée aux deux autres, avec une corrélation proche de 0,70 (une corrélation qui reste très importante). Ce résultat n'est pas totalement surprenant, puisque de nombreuses études ont constaté la forte corrélation entre connaissances lexicales et grammaticales (par exemple, Guo & Roehrig, 2011; Purpura, 1999; Shiotsu & Weir, 2007), ce qui a poussé certains de leurs auteurs à proposer l'hypothèse qu'il s'agit en fait d'un seul et même construit, ou du moins d'un gradient qui n'a pas de frontière nette. Une corrélation de 0,7 est justement la limite à partir de laquelle la multicolinéarité peut devenir gênante si l'on veut ensuite faire une étude de régression (Crossley et al., 2012). D'après Field

et ses collaborateurs : « *Multicollinearity between predictors makes it difficult to assess the individual importance of a predictor. If the predictors are highly correlated, and each accounts for similar variance in the outcome, then how can we know which of the two variables is important? Quite simply, we can't tell which variable is important – the model could include either one, interchangeably* » (A. Field et al., 2012, p. 276). Cela n'empêche pas de faire une analyse de régression, mais cette multicolinéarité ne nous permettra pas vraiment, une fois l'analyse de régression effectuée, de déterminer l'importance relative des variables fortement corrélées entre elles. Ainsi, si une variable ne s'avère pas expliquer de variance supplémentaire par rapport à une variable principale, parce qu'elle est très corrélée avec cette dernière, il paraît difficile de dire que la première variable, non retenue dans les prédicteurs de l'analyse de régression, ne joue pas de rôle (Howell, 2009, p. 551-552).

10.5. Régression logistique binaire

10.5.1. Présentation

Ce qui nous intéresse maintenant, c'est d'étudier la contribution respective de ces cinq tests à l'explication de la variance de compréhension de l'oral. Nous ne pouvons pas utiliser directement les valeurs trouvées lors de l'analyse par variable, car les pourcentages de variance expliquée par chacune d'entre elles ne s'additionnent pas (d'ailleurs, leur somme totale est supérieure à 100, puisque les trois variables les plus corrélées à la CO, AURLEX, PVST et AURGRAM, expliquent chacune environ 50% de variance). En effet, toutes nos variables étant corrélées entre elles, une grande partie de la variance expliquée l'est de façon commune.

Les modèles statistiques de régression linéaire sont souvent utilisés pour montrer l'apport de plusieurs variables (numériques ou non) à l'explication d'une variable numérique continue. Nous ne pourrons pas dans notre cas appliquer ce genre de modèle, car la variable que nous voulons expliquer (la compréhension de l'oral) est une variable ordinale, qui prend quatre valeurs ordonnées : A2, B1, B2 et C1. Rappelons que, comme le test SELF utilisé pour obtenir ce niveau est un test semi-adaptatif, tous les étudiants ne passent pas les mêmes items, et le score total obtenu au test ne peut donc être comparé d'un étudiant à l'autre. C'est pourquoi nous nous sommes tournée vers un autre modèle que nous examinerons à présent[43].

[43] Le modèle de réponse à l'item utilisé pour le calcul du score SELF exploite en fait des valeurs numériques, le logit, mais ces valeurs ne sont pas accessibles directement car elles n'apparaissent dans les données exportées.

Le modèle de régression logistique binaire permet d'expliquer une variable binaire (schématiquement, 0 ou 1) à l'aide d'une ou plusieurs variables indépendantes (explicatives). Dans notre cas, nous souhaitons utiliser les cinq variables PHON, PROSO, AURLEX, AURGRAM et PVST pour prédire un niveau suffisant, B2 ou C1 (codé comme 1) ou insuffisant, A2 ou B1 (codé comme 0) en compréhension de l'oral. Nous basculons ainsi en quelque sorte dans une tâche de classification, en transformant la variable ordinale SELF_CO en variable binaire (que nous appelerons CO_b). Au lieu de prédire une valeur numérique, une régression logistique permet de calculer la probabilité qu'un individu soit classé dans une catégorie ou dans une autre à partir des données disponibles.

Pour visualiser ce que cela signifie, prenons l'exemple simplifié de la relation entre la variable PVST et la variable binaire CO_b (Figure 10.7). Nous voyons que les deux variables n'entretiennent pas une relation linéaire et qu'une droite (de régression linéaire) serait donc une représentation très peu satisfaisante de cette relation. La régression logistique suppose que la relation entre ces deux variables est mieux représentée par une courbe en S (logistique), visible en gris sur la figure[44].

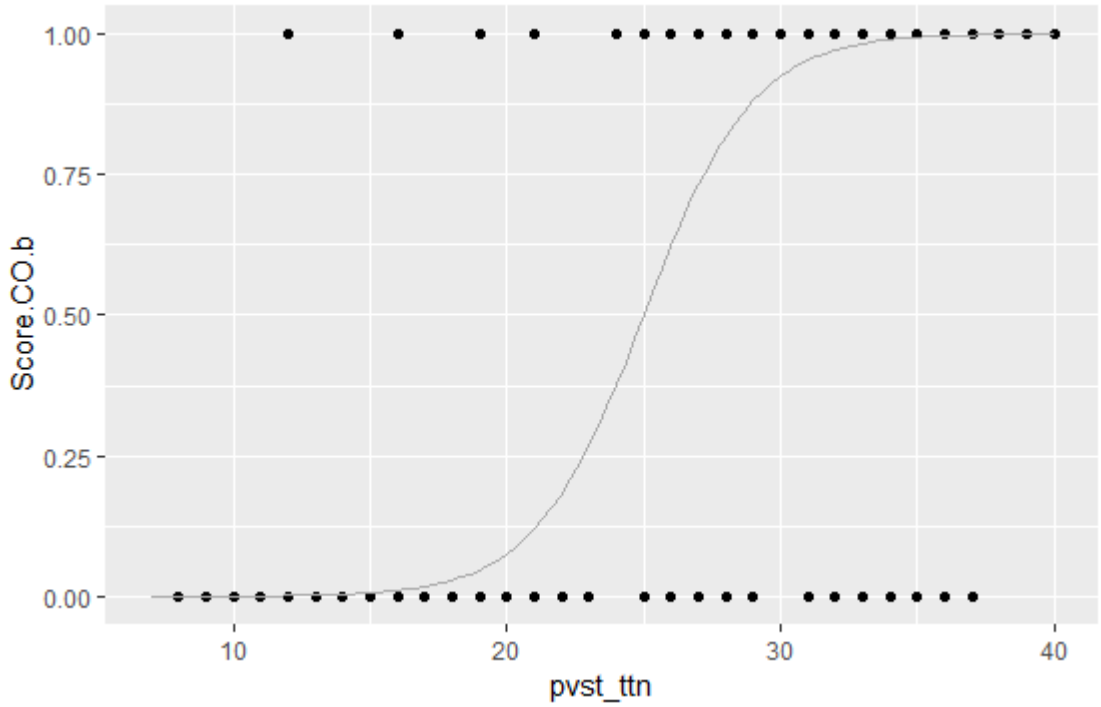


Figure 10.7 – nuage de points représentant la relation entre la variable continue PVST et la variable binaire CO.b, avec courbe de régression logistique superposée pour illustration

[44] Cette courbe sert uniquement d'illustration et n'a pas été générée à partir d'un modèle calculé sur les données du PVST.

Les points qui se situent en bas à gauche correspondent à des scores faibles au PVST (autour de 10), et sont associés à une probabilité quasi nulle d'avoir un niveau satisfaisant (B2 ou C1) en CO (notons cependant qu'un étudiant avec un score de 12 au PVST en haut à gauche du graphique est classé dans la catégorie « CO satisfaisante » - le modèle ne rendra pas bien compte de ce résultat). Inversement, les points en haut à droite du graphique correspondent à des scores élevés au PVST, associés à une probabilité élevée d'obtenir un niveau satisfaisant en CO. La courbe passe de la probabilité nulle à la probabilité maximale en formant un S au milieu du graphique, le point d'inflexion se situant autour du score 25 au PVST. Cela signifie que, selon le modèle logistique (imaginaire ici), un étudiant avec un score de 25 au PVST a 50% de chances d'avoir un score satisfaisant en CO.

Ce principe de régression logistique, présenté ici avec une seule variable explicative (modèle simple), peut être généralisé à plusieurs variables explicatives, ce que nous ferons dans les paragraphes qui suivent. Au préalable, nous utiliserons la régression logistique binaire simple pour confirmer le score de césure pour chaque test.

10.5.2. Confirmation des scores de césure (régression logistique simple)

Comme nous venons de le voir, la régression logistique permet d'identifier un score qui corresponde à une probabilité de 50% d'avoir un niveau suffisant en compréhension de l'oral. Nous voudrions utiliser ce point de bascule entre niveau insuffisant et niveau suffisant comme score de césure pour chaque test. Ce score de césure servira à repérer les étudiants qui ont besoin de remédiation dans un domaine particulier, et à les informer de manière plus générale sur le sens à donner aux scores qu'ils ont obtenus. Un score seul est en effet peu informatif s'il n'est pas accompagné d'une aide à l'interprétation, qui permet de jauger si l'on a obtenu un résultat faible, acceptable ou très satisfaisant. Nous avons donc calculé le modèle logistique associé à chaque test diagnostique par rapport à la compréhension aurale. Nous ne présenterons pas en détail les résultats des modèles (disponibles en Annexe 9), mais simplement les résultats du calcul du score de césure pour chaque test, c'est-à-dire le score associé à une probabilité de 50% d'avoir un niveau de B2 ou plus en compréhension aurale. Les étudiants obtenant un score inférieur seront dirigés vers des activités de remédiation dans le domaine correspondant au test diagnostique. Le tableau Tableau 10.5 présente ces scores, accompagnés des pourcentages d'étudiants à chaque niveau qui seraient ainsi concernés par une remédiation éventuelle.

	PHON	PROSO	AURLEX	PVST	AURGRAM
score de césure	25	41	27	26	21
A2 sous score	74%	84%	97%	97%	100%
B1 sous score	67%	58%	58%	53%	69%
B2 sous score	40%	40%	41%	17%	29%
C1 sous score	8%	0%	0%	0%	0%
score inspection visuelle		38	27	25	20

Tableau 10.5 - scores de césure (obtenus par régression logistique sur la compréhension de l'oral) proposés pour chaque test diagnostique, proportion d'étudiants concernés par la remédiation pour chaque niveau de CO, et rappel des scores de césure proposés suite à l'inspection visuelle des diagrammes de la section 10.3

On constate tout d'abord que ces scores sont assez proches (voire identique dans le cas d'AURLEX) de ceux envisagés suite à l'inspection visuelle des diagrammes en boîtes à moustaches dans les sections 10.3.3 et suivantes (et rappelés ici dans la dernière ligne du tableau). En ce qui concerne les niveaux extrêmes, tous les scores proposés excluent les étudiants de niveau C1 de la remédiation (sauf pour le test PHON où un étudiant C1 est inclus), et les scores de césure pour AURLEX, PVST et AURGRAM permettent d'inclure pratiquement tous les étudiants A2. Ce n'est pas le cas de PHON et PROSO, mieux réussis par les étudiants faibles que les autres tests : un quart des étudiants A2 échappent ainsi au besoin de remédiation en discrimination phonémique. Les résultats sont moins tranchés pour les étudiants B1, dont 50 à 70% sont concernés par la remédiation selon les tests. Enfin, si les scores de césure pour PVST et AURGRAM permettent d'exclure la grande majorité des étudiants B2 de la remédiation, 40% de ces étudiants sont tout de même concernés pour PHON, PROSO et AURLEX. Ce n'est pas forcément un défaut de ces tests (ni des scores de césure) : les étudiants de niveau B2 peuvent tout à fait avoir besoin de progresser dans certains domaines, en particulier en connaissances lexicales (besoin de plus grande précision).

Ces résultats montrent clairement que le passage d'un niveau à l'autre du CECR ne se fait pas de manière abrupte et soudaine : il s'agit d'un continuum, et l'amélioration est continue d'un niveau à l'autre. Les niveaux CECR ont été superposés artificiellement, pour des raisons de lisibilité et de praticité (parce que cela correspond à la division traditionnelle en débutants/ intermédiaires/ avancés), sur ce continuum sous-jacent (De Jong & Zheng, 2016). Cependant, on s'attend tout de même à ce que tous les étudiants B1 aient besoin de remédiation dans au moins un des domaines testés, et qu'inversement aucun étudiant B2 n'en ait besoin dans tous les domaines. Après analyse de nos données avec les scores de césure proposés, nous avons constaté que quatre étudiants B1 (sur 44) avaient un score supérieur au score de césure dans tous les tests diagnostiques, et qu'à l'inverse, un étudiant B2 (sur 63) était concerné par tous les domaines de remédiation. Nous considérons qu'il s'agit là d'un taux d'erreur acceptable, même si certains étudiants B1 (10%) échappent ainsi à la remédiation.

10.5.3. Analyse principale (régression logistique multiple)

Revenons à présent à la question du rôle relatif de chacune de nos variables explicatives, et calculons les résultats du modèle de régression logistique avec tous nos prédicteurs (Tableau 10.6). La condition de validité d'un modèle logistique est un nombre suffisant d'observations : au moins cinq à dix événements par variable explicative sont nécessaires[45] (Falissard, 2011, p. 90). Comme nous avons cinq variables explicatives, cette condition requiert au moins 25 à 50 observations (par niveau de CO_b). Nous avons vu (Tableau 10.2) que nous avions 82 sujets avec un niveau de CO_b insatisfaisant (B1 ou moins), et 76 avec un niveau satisfaisant (B2 ou plus), ce qui est donc suffisant pour satisfaire à cette condition. La première ligne du Tableau 10.6 ci-dessous (après le rappel de la fonction utilisée dans *R*) présente la distribution des résidus du modèle, c'est-à-dire des distances entre le modèle proposé et un modèle idéal qui rendrait parfaitement compte des données. Les résidus sont symétriques et à peu près centrés sur 0, ce qui est un point positif pour notre modèle.

La suite du Tableau 10.6 présente les coefficients calculés par l'analyse de régression logistique. Nous constatons que le seul coefficient significatif est celui associé au PVST, ce qui n'est pas inattendu dans la mesure où, dans notre analyse précédente, le PVST était la variable la plus corrélée à la CO, et que les autres variables également bien corrélées avec la CO (AURGRAM et AURLEX) étaient elles-mêmes très corrélées avec le PVST et n'expliquent donc pas forcément de variance supplémentaire par rapport à cette première variable. Les deux dernières variables, PHON et PROSO, peu corrélées à la CO, avaient peu de chance d'être prédictives de ce niveau. Nous pouvons à présent calculer le R^2 associé à ce modèle (A. Field et al., 2012, p. 317-318), représentant le pouvoir explicatif du modèle, ou le pourcentage de variance expliquée, qui est de 39%. Ce résultat (tout en représentant une taille d'effet importante, Plonsky & Oswald, 2014) est un peu décevant par rapport au pourcentage de variance expliquée en utilisant une seule variable, présenté lors des analyses de variance, où le PVST expliquait à lui seul 58% de la variance. Cela peut venir du fait que nous avons perdu en richesse d'information, dans ce modèle logistique binaire où la variable ordonnée CO (qui prenait quatre valeurs possibles, de A2 à C1) est transformée en variable binaire correspondant à un niveau suffisant ou insuffisant. D'autre part, nous avons vu que certains de nos tests peinaient à distinguer le niveau B1 du niveau B2. Nous aurions pu espérer que la

[45] Une deuxième condition, la linéarité entre les prédicteurs et le logarithme du rapport des cotes (*log odds ratio*) ne sera pas vérifiée ici.

combinaison des résultats à plusieurs tests puisse améliorer les choses, mais cela n'est apparemment pas le cas.

```
Call:
glm(formula = Score.CO.b ~ aurl_2020 + aurgram_32it_2020 + pvst_ttn +
    phon31it_2020 + proso_56it_2020, family = "binomial", data = foo)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2024  -0.5423  -0.2179   0.6756   2.2929

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)       -7.62622    1.94651  -3.918 8.93e-05 ***
aurl_2020          0.02933    0.05671   0.517   0.6050
aurgram_32it_2020  0.09480    0.07150   1.326   0.1848
pvst_ttn           0.13720    0.04818   2.848   0.0044 **
phon31it_2020      0.02718    0.05561   0.489   0.6251
proso_56it_2020    0.01645    0.02712   0.606   0.5442
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 190.26  on 137  degrees of freedom
Residual deviance: 116.75  on 132  degrees of freedom
  (51 observations deleted due to missingness)
AIC: 128.75

Number of Fisher Scoring iterations: 5
```

Tableau 10.6 - résultats du modèle de régression logistique multiple principal, avec cinq variables explicatives (AURLEX, AURGRAM, PVST, PHON et PROSO) et une variable binaire à expliquer (CO.b), séparant les sujets de niveau suffisant (B2 ou C1) de ceux de niveau insuffisant (A2 ou B1)

Pour visualiser l'adéquation du modèle aux données, nous proposons un diagramme (Figure 10.8) qui permet de comparer d'une part la probabilité prédite pour chaque étudiant, au vu de ses résultats, d'avoir un niveau suffisant en compréhension de l'oral, et d'autre part son niveau effectif. Chaque croix de la courbe logistique représente un sujet (les sujets sont rangés par probabilité croissante d'avoir un niveau suffisant de CO d'après le modèle). Les croix de couleur bleu foncé correspondent aux 82 sujets de niveau B1 ou inférieur d'après SELF, et les croix de couleur bleu clair aux 76 sujets de niveau B2 ou plus. On constate que le modèle fonctionne globalement bien : les sujets sont placés sur un continuum de probabilité qui reflète la variabilité de leurs résultats dans les différents tests diagnostiques. Les premiers sujets en bas à gauche, considérés par le modèle comme ayant moins de 25% de chances d'être de niveau B2, sont effectivement essentiellement des croix foncées. Inversement, les sujets en haut à droite du modèle ont, dans leur quasi-totalité, été effectivement positionnés dans le niveau B2 par le test SELF (croix bleu clair). C'est entre 25% et 75% qu'on observe le plus d'enchevêtrement de croix claires et foncées, mais toujours avec une proportion

croissante de bleu clair quand on monte dans la courbe. Cependant, le modèle n'est pas parfait (ce que nous savions, dans la mesure où il n'explique « que » 38% de la variance des données) puisque quelques étudiants B2 (bleu clair) se retrouvent en bas de la courbe, et quelques étudiants B1 ou moins (bleu foncé) s'insèrent également en haut de la courbe. Ce sont les étudiants qui seront mal repérés par les tests diagnostiques : soit ils seront dirigés vers les activités de remédiation dont ils n'ont en fait pas besoin, soit (ce qui est plus grave à nos yeux), ils échapperont au soutien dont ils auraient besoin.

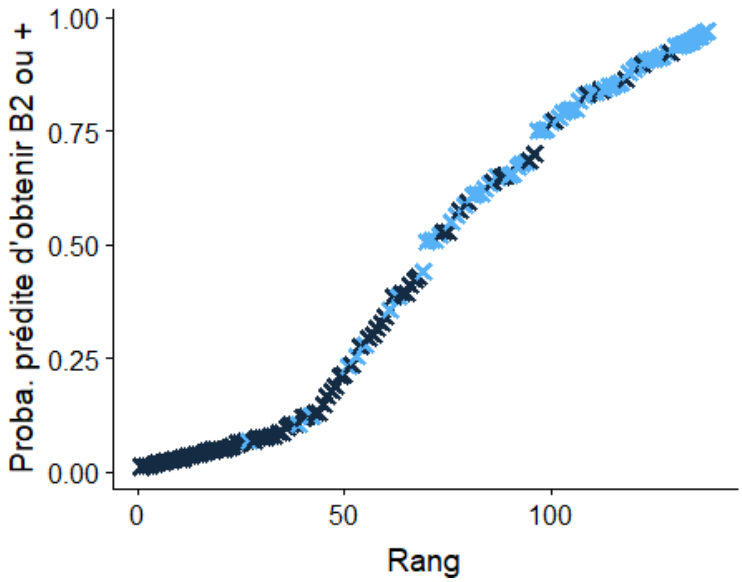


Figure 10.8 - courbe logistique de comparaison entre la probabilité prédite d'avoir un niveau satisfaisant et le niveau effectif (croix bleu foncé = niveau insuffisant B1 ou -, croix bleu clair = niveau suffisant, B2 ou +)

10.5.4. Analyse exploratoire

Dans un esprit exploratoire, pour mieux comprendre nos données, nous avons décidé de lancer une dernière régression logistique binaire, pour voir s'il était possible de distinguer les étudiants très faibles (A2) en CO, des étudiants de niveau moyen et avancé (B1 et plus). Signalons que cette distinction recoupe en grande partie, dans notre échantillon, la variable « filière », puisque la plupart des étudiants de niveau A2 sont des étudiants LANSAD, et que, inversement, la plupart des étudiants de niveau B1 ou plus sont des étudiants LLCER. Nous avons vu que, pour une régression logistique, il faut au moins cinq à dix sujets par variable explicative, et qu'avec nos cinq variables explicatives, nous avons donc besoin d'au moins 25 sujets (limite des conditions de validité), et dans l'idéal 50 sujets de niveau A2. Nous en avons 38 (cf Tableau 10.2), ce qui n'est pas idéal, mais tout de même acceptable (il nous reste alors 120 étudiants de niveau B1 et plus). Notre modèle sera cependant moins robuste de ce

fait, et la prudence sera de mise dans l'interprétation des résultats (en particulier pour ce qui est de leur généralisation).

Nous constatons à la lecture des résultats du modèle (Tableau 10.7) que les trois variables explicatives qui contribuent significativement à ce deuxième modèle de régression logistique sont AURLEX, AURGRAM et PVST, c'est-à-dire les trois variables les plus corrélées à la compréhension aurale. Leur rôle respectif est ici différent de celui qu'elles avaient dans le modèle précédent (où seul PVST était retenu dans les prédicteurs significatifs), dans la mesure où c'est AURLEX (et donc les connaissances lexicales) qui apparaît comme le plus prédictif d'un score très faible en CO. Il est ainsi possible que, pour les étudiants de niveau très faible, un travail préalable sur le lexique soit plus profitable qu'un travail sur les expressions phraséologiques ou les structures grammaticales. Ce modèle explique 64% de la variance dans les données, ce qui est beaucoup plus important que notre premier modèle, mais probablement moins généralisable du fait du petit nombre d'observations. Ces conclusions demanderaient donc à être confirmées par une nouvelle étude avec un nombre plus important d'étudiants de niveau A2.

```
Call:
glm(formula = Score.CO.B1.b ~ aurl_2020 + aurgram_33it_2020 +
    pvst_ttn + phon32it_2020 + proso_56it_2020, family = "binomial",
    data = foo)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.88758   0.00130   0.02514   0.20738   2.20126

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)       -11.996003   3.749967  -3.199  0.00138 **
aurl_2020           0.323743   0.111240   2.910  0.00361 **
aurgram_33it_2020   0.326518   0.132936   2.456  0.01404 *
pvst_ttn            0.151257   0.074195   2.039  0.04149 *
phon32it_2020      -0.099566   0.101606  -0.980  0.32712
proso_56it_2020    -0.003133   0.041459  -0.076  0.93975
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 151.821  on 137  degrees of freedom
Residual deviance:  54.602  on 132  degrees of freedom
  (51 observations deleted due to missingness)
AIC: 66.602

Number of Fisher Scoring iterations: 8
```

Tableau 10.7 - résultats du modèle de régression logistique multiple exploratoire, avec cinq variables explicatives (AURLEX, AURGRAM, PVST, PHON et PROSO) et une variable binaire à expliquer (CO.b), séparant les sujets de niveau faible (A2) de ceux de niveau moyen ou bon (B1 et plus)

10.6. Conclusion

Au terme de cette étude de corrélation et de régression, nous pouvons tirer quelques conclusions pour la suite. Nous avons déjà montré plus haut (lors de la conclusion partielle de la section 10.3.7) que les réponses à notre première question de recherche (Dans quelle mesure les tests diagnostiques sont-ils corrélés avec les résultats en compréhension de l'oral ?) confirment les résultats des travaux scientifiques existants sur la compréhension de l'oral : faible corrélation avec connaissances phonémiques et prosodiques d'une part, et corrélation importante avec les connaissances lexicales et grammaticales de l'autre. Une corrélation importante a par ailleurs été constatée (pour la première fois) avec les connaissances phraséologiques.

Grâce à notre deuxième question de recherche (Tous les tests diagnostiques sont-ils corrélés entre eux ?), nous avons pu observer que les tests se répartissaient en trois groupes. Le premier groupe est formé des trois tests diagnostiques très corrélés avec la compréhension aurale (AURLEX, AURGRAM et PVST), qui sont également très corrélés entre eux. A l'intérieur de ce groupe, les deux tests qui concernent l'étape d'intégration d'Anderson ou d'interprétation de Cutler et Clifton (combinaison du sens des mots, cf. sections 1.2.1 et 1.2.2), à savoir AURGRAM et PVST, sont tellement corrélés (0,85) que la séparabilité de leur construit peut être mise en doute. Les deux tests restants, PHON et PROSO, peu corrélés avec la compréhension de l'oral, sont également très peu corrélés entre eux, contrairement à ce à quoi nous nous attendions, et appartiennent donc à deux « groupes » différents. Nous avons vu que cette absence de corrélation constatée entre connaissances phonémiques et prosodiques, quoique surprenante, a déjà été observée dans des études antérieures en L1 anglais (Goswami et al., 2010). Par ailleurs, PHON n'est que faiblement corrélé avec les tests du premier groupe, tandis qu'on observe pour PROSO une corrélation moyenne avec ces mêmes tests.

Notre dernière question de recherche portait sur le rôle respectif de chacune de nos variables explicatives dans la variance de la variable à expliquer, la CO. Pour répondre à cette question, nous avons transformé notre variable ordinale SELF_CO (qui prend les valeurs A2, B1, B2 et C1) en variable binaire correspondant aux valeurs « niveau insuffisant » en CO (A2 et B1) et « niveau suffisant » (B2 et C1), et utilisé une technique de régression logistique binaire. Nous avons trouvé que les connaissances phraséologiques étaient les seules prédictives du niveau de CO, et que les autres variables n'expliquaient pas de variance supplémentaire. C'est en

quelque sorte la conclusion logique de la réponse à nos deux premières questions de recherche, puisque nous savions déjà qu'AURLEX, AURGRAM et PVST étaient très corrélées à la CO (donc très susceptibles d'être prédictives de ce niveau), et qu'elles étaient très corrélées entre elles (donc susceptibles de ne pas apporter beaucoup d'informations supplémentaires les unes par rapport aux autres). Une analyse exploratoire utilisant une régression logistique séparant notre échantillon en « niveau faible » (A2) et « niveau intermédiaire et avancé » (B1 et plus) a également apporté des résultats préliminaires pointant une importance plus grande des connaissances lexicales pour les niveaux très faibles. En vue du chapitre sur les conséquences pédagogiques de nos résultats, nous allons dans ce qui suit résumer les apports de chaque test diagnostique et envisager des améliorations futures à leur apporter éventuellement.

Nous n'avons trouvé que peu de corrélation entre la capacité de discrimination phonémique opérationnalisée par le test PHON et la compréhension aurale, conformément à l'absence de passage à l'échelle (*scaling up*) notée par Bradlow (2007), qui peut être expliquée en partie par le phénomène postulé de *lossy chunking* (Christiansen & Chater, 2016), qui implique une perte d'informations lors du passage d'un niveau de traitement à un autre. L'utilisation de processus descendants pour compenser les manques éventuels au niveau du traitement phonologique peuvent également contribuer à cette absence de corrélation. Dans notre cas, cependant, il semble que cette faible corrélation puisse s'expliquer en partie, non pas par un mauvais traitement phonologique, mais au contraire par les relativement bons résultats obtenus par nos sujets au test de discrimination phonémique (70% de réussite), y compris par les étudiants de niveau A2 et B1 en CO. Nous avions remarqué lors de l'analyse du test de discrimination phonémique que les étudiants avaient passé plus de temps sur les items les plus difficiles. Il est possible que la prise en compte du temps de réponse eût permis de mieux identifier les étudiants faibles, qui peuvent se différencier des étudiants plus avancés par un traitement moins rapide et moins automatique, qui n'apparaît pas dans notre test diagnostique mais se traduit par une moins bonne compréhension en temps réel.

Il est possible également qu'un test se focalisant uniquement sur le traitement phonologique au niveau lexical soit finalement plus utile pour notre contexte. En effet, le test PHON comportant à la fois des mots et des pseudomots, les étudiants ont pu privilégier une stratégie « pré-lexicale » (Melnik & Peperkamp, 2019a). Nous avions fait attention à ce qu'ils ne puissent pas utiliser une stratégie purement acoustique (« Est-ce que j'entends exactement la même chose ou pas ? »), en incluant systématiquement différents locuteurs (parfois de sexe

différent) dans nos items de « cherchez l'intrus ». En effet, nous savions déjà que des items demandant un traitement uniquement acoustique sont bien réussis par les apprenants francophones, même sur des sons et contrastes n'existant pas en français (Dupoux et al., 1997; Mah et al., 2016). Nos items requièrent donc de faire abstraction des différences de hauteur de voix, de durée, de timbre, etc., pour arriver à une représentation abstraite (probablement phonémique) des mots entendus et d'identifier celui qui est différent des autres. Cependant, notre test n'implique pas l'accès lexical, c'est-à-dire que les formes entendues n'activent pas forcément une forme phonologique stockée en mémoire et associée à un mot particulier. D'après Melnik et Peperkamp (2019), c'est à ce niveau-là que les apprenants francophones font l'expérience des problèmes les plus importants. Leur étude (sur la perception du /h/ anglais par des étudiants francophones de niveau intermédiaire) montre que même quand les francophones ont une bonne perception du /h/, ils acceptent des mots inexistants (*usband*) parce qu'ils n'ont pas une représentation phonologique précise du mot, qui est probablement sous-spécifié quant à la présence d'une consonne initiale. Si cette hypothèse est vraie (et notre analyse des résultats du test AURLEX, section 7.4.3.3, va dans le même sens), ce n'est donc pas la perception des phonèmes qu'il faudrait tester (ou du moins pas seulement), mais plutôt la représentation phonologique du mot chez nos apprenants. Nous pourrions alors proposer deux types d'items. Les premiers sont des items de décision lexicale avec des formes correctes ou incorrectes (auquel cas nous retombons sur les items de notre test AURLEX, mais avec des pseudomots formés à partir de mots véritables dont un phonème difficile a été enlevé ou remplacé par un autre phonème proche). Les seconds sont des items qui lient la forme orale et la forme écrite d'un mot, par exemple en faisant écouter un mot (*hair*), et en leur faisant choisir entre deux orthographes possibles (*air* ou *hair*). Nous avions écarté ce type d'item justement parce qu'il teste également la connaissance des correspondances phono-graphématiques, mais il serait peut-être intéressant de les intégrer tout de même, précisément pour cette raison[46].

Comme pour PHON, nous avons constaté avec PROSO une corrélation assez faible entre les résultats au test diagnostique de sensibilité prosodique et le niveau de compréhension aurale. Ici aussi, il est possible que la prise en compte du temps de réponse puisse aider à caractériser le traitement prosodique des étudiants plus faibles, et en particulier son automatisation. Mais comme pour le test PHON, l'intégration de nouveaux items pourrait également être envisagée.

[46] Cependant, nous devons noter que le test utilisé dans la thèse de Zoghlami (2015) utilisait des items de ce type, et que la corrélation avec la compréhension de l'oral était tout de même faible, comme nous l'avons déjà souligné.

Nous avons résumé dans le premier chapitre (Figure 1.4) les études de Mattys et de son équipe (par ex. Mattys et al., 2005), qui montrent que les informations prosodiques jouent un rôle moins important que les informations lexicales (comme nous l'avons également constaté dans notre étude), mais que la prosodie devient essentielle dans des conditions d'écoute dégradées (en présence de bruit de fond, par exemple). Les items de notre test ayant été enregistrés dans de bonnes conditions acoustiques et sans bruit de fond dans un studio d'enregistrement, ils ne sont pas représentatifs des conditions d'écoute naturelle où il est rare que la voix qui produit le discours que nous essayons de comprendre soit la seule source sonore dans l'environnement. C'est pourquoi nous pensons qu'il pourrait être intéressant d'inclure des items avec bruit de fond dans un tel test. Par ailleurs, nous avons déjà mentionné que nous aurions pu intégrer plus d'items testant la segmentation lexicale (*my key* vs. *Mikey*), alors qu'actuellement, le test PROSO est essentiellement composé d'items d'identification de la syllabe accentuée de mots ou pseudomots. Nous avons vu qu'en ce qui concerne le test PHON, il serait intéressant d'essayer d'accéder au traitement lexical plutôt que pré-lexical (et donc de monter d'un niveau de traitement). Pour le test PROSO, il serait de même intéressant de monter aussi d'un niveau et de tester l'utilisation par nos apprenants de la prosodie au niveau de la phrase (et pas uniquement du mot), en particulier pour la segmentation lexicale.

Parmi nos trois tests restants, penchons-nous sur le cas d'AURGRAM et du PVST. Nous avons vu que les résultats à ces deux tests étaient étroitement corrélés. Nous pensons que cela peut être dû à la façon dont nous avons opérationnalisé le construit des connaissances grammaticales. Les études qui trouvent peu de corrélation entre compréhension aurale et connaissances grammaticales utilisent des tests à trous écrits (par ex. Mecartty, 2000; Zoghlami, 2015), qui correspondent probablement assez peu au traitement nécessaire lors de la compréhension du matériau oral. La seule étude qui trouve une corrélation importante (et, comme dans cette thèse, plus importante qu'avec les connaissances lexicales), utilise comme nous un test aural, de type jugement de grammaticalité (Andringa et al., 2012). La corrélation importante entre ce type de test et un test de connaissances phraséologiques pourrait s'expliquer par la stratégie employée par les apprenants pour effectuer ces jugements. Nous avions souligné lors de notre analyse d'AURGRAM (section 8.4.4) qu'il n'était pas facile de savoir au vu des résultats (ni à la lecture de la littérature scientifique sur la question) quelle était la source principale des réponses de nos étudiants (connaissances grammaticales, phonologiques, lexicales, ou comparaison avec des phrases déjà entendues). Etant donné la corrélation très élevée avec le PVST, pour lequel aucune analyse grammaticale ni traitement

de la langue orale ne sont requis, et dont les expressions sont composées de mots extrêmement fréquents connus des étudiants[47], c'est la dernière possibilité qui paraît la plus plausible. Il est donc probable que pour beaucoup d'étudiants et/ou d'items, le jugement aural de grammaticalité était un jugement de familiarité (« Est-ce que cela ressemble à quelque chose que j'ai déjà entendu ? Est-ce que ça sonne juste ? »). Nous reviendrons sur cette question en présentant la théorie des exemplaires dans le chapitre qui suit.

[47] Les mots composant les items du PVST appartiennent tous à la première bande de fréquence de Nation (2017), c'est-à-dire aux 1 000 familles de mots les plus fréquentes, mis à part 5 mots : *likely, instance, prove, blame,* et *granted*, qui appartiennent à la deuxième bande de fréquence.

Chapitre 11

Conséquences pédagogiques et remédiation

Nous avons pour l'instant travaillé sur les traitements cognitifs et les connaissances linguistiques à l'œuvre lors de la compréhension aurale, dans une perspective diagnostique. Cela nous a permis d'identifier les champs problématiques pour nos étudiants, mais l'objectif d'une évaluation diagnostique est de proposer in fine des pistes de remédiation aux apprenants. C'est ce que nous allons tenter de faire dans ce chapitre, où nous passerons ainsi du versant « traitement » au versant « acquisition » de la langue étrangère. Il faut cependant noter que ces pistes resteront spéculatives : en effet, elles découlent de notre étude corrélatoire, mais ne viennent pas d'une expérimentation sur la remédiation elle-même. Corrélation n'équivaut pas à « causation » : nous avons constaté une forte corrélation entre compréhension de l'oral d'une part et connaissances lexicales, phraséologiques et grammaticales de l'autre. Cela ne signifie pas que ces connaissances « causent » la compréhension (Milton, 2013, p. 75) ; comme nous l'avons d'ailleurs montré dans la première partie, elles n'en sont qu'un élément. Cependant, les modèles de la CO que nous avons présentés (1.2.) expliquent comment ces connaissances, à l'intersection entre le bas niveau (traitement du signal sonore, son et prosodie) et le haut niveau (traitement du sens jusqu'à l'élaboration d'un modèle de situation) peuvent se retrouver au centre du processus. La corrélation n'indique pas non plus dans quel sens intervient la relation : est-ce que les connaissances lexicales, phraséologiques et grammaticales sont un préalable à la CO, ou est-ce qu'on connaît beaucoup de vocabulaire, d'expressions et de structures parce qu'on passe beaucoup de temps à écouter ? La question est un peu artificielle et la relation va probablement dans les deux sens[48] : « *The relationship between comprehension skill and knowledge of language is most likely reciprocal. More text processing leads to more knowledge, which leads to better comprehension skills* » (Andringa et al., 2012, p. 71). Nous

[48] Une relation similaire existe entre la conscience phonémique et l'apprentissage de la lecture en langue maternelle, qui se renforcent l'un l'autre (Bosse & Zagar, 2016).

commencerons par une exploration plus en profondeur de nos résultats et poursuivrons par une présentation des principales caractéristiques du dispositif envisagé, centré sur l'apprentissage des connaissances lexicales et phraséologiques.

11.1. Connaissances lexicales et phraséologiques au centre des processus d'apprentissages langagiers

Dans le premier chapitre de cette thèse, nous avons exposé le système hiérarchique de processus imbriqués dont dépend la compréhension de l'oral, avec les connaissances linguistiques qui leur correspondent. Les connaissances lexicales peuvent être considérées comme centrales, dans la mesure où elles sont à l'intersection des processus de traitement formel (traitement du signal acoustique) et de traitement sémantique. Les connaissances phraséologiques et grammaticales sont également essentielles, en ce qu'elles facilitent le processus de regroupement des mots (*chunking*) qui permet un traitement efficace et rapide des données langagières qui arrivent continuellement. Dans le deuxième chapitre, nous avons vu que les connaissances lexicales sont au cœur des difficultés des apprenants, de par la quantité d'éléments à acquérir (plusieurs milliers de familles de mots), mais aussi du fait des différents aspects de la connaissance d'un mot (formes écrite et orale, sens, collocations, morphologie, etc.). Les connaissances phraséologiques posent également certaines difficultés, en particulier le repérage des collocations et expressions figées, souvent formées de mots très courants, et dont le sens non compositionnel n'est pas forcément identifié. Dans l'étude corrélatoire générale décrite au chapitre précédent (chapitre 10), nous avons constaté que les connaissances phraséologiques (mesurées par un test écrit de reconnaissance du sens) étaient les plus prédictives du niveau de compréhension aurale, et que les connaissances grammaticales (mesurées par un test de jugement de grammaticalité aural, donc d'un format très différent) étaient fortement corrélées aux connaissances phraséologiques, mettant en doute la séparabilité de leur construit. Dans ce qui suit, nous ne nous référerons donc plus qu'aux connaissances phraséologiques, supposant qu'elles ne forment qu'un avec les connaissances grammaticales en réception. Cette même étude corrélatoire a montré que les connaissances lexicales étaient elles aussi très corrélées aux connaissances phraséologiques, quoique dans une moindre mesure. Même si elles n'expliquent pas de variance supplémentaire dans le modèle séparant les apprenants de niveau A2 et B1 en compréhension de l'oral de ceux de niveau B2 et plus, il semble qu'elles jouent un rôle plus important pour expliquer les difficultés des étudiants de niveau A2, comme l'a montré notre modèle de

régression logistique exploratoire. Nous voudrions apporter dans ce qui suit quelques arguments supplémentaires en faveur de la place centrale du lexique, non plus lors du traitement langagier, mais cette fois au cours de l'apprentissage d'une L2 en milieu institutionnel (scolaire ou universitaire).

11.1.1. Lexique et apprentissage phonologique

Travailler le lexique ne veut pas dire s'abstenir de travailler les autres compétences : il s'agit au contraire de travailler les autres compétences à travers le lexique. Pour ce qui est de la prosodie, le mot est bien sûr le lieu le plus naturel pour un travail sur l'accent lexical, mais nous avons vu au chapitre précédent que c'était également l'unité la plus appropriée pour travailler la discrimination phonémique : ce qui pose problème en effet chez les apprenants de langue étrangère est moins le traitement phonétique-phonologique lui-même que la représentation phonologique des mots en mémoire, et c'est donc cela qu'il faut essayer de travailler en priorité. Nous revenons ici à une idée déjà mentionnée lors de l'analyse du test AURLEX, en section 7.4.3.3 : il est important d'améliorer la précision des représentations lexicales (une idée d'abord prônée par Perfetti & Hart, 2002, pour la lecture). Darcy et ses collaborateurs, par exemple, situent l'origine de certaines difficultés de reconnaissance lexicale dans les representations trop imprécises des apprenants (« *imprecise or fuzzy lexical representations* », Darcy et al., 2013, p. 374), et Melnik et Peperkamp font le même constat avec des apprenants francophones à propos de la reconnaissance des mots anglais commençant par /h/ : « *intermediate to advanced French learners—in addition to the difficulty they may have with the perception of /h/—have imprecise lexical representations of /h/-initial words* » (Melnik & Peperkamp, 2019b, p. 17).

Afin d'améliorer la précision des représentations lexicales, une solution semble être d'augmenter la taille du lexique des apprenants. En effet, quand la taille du lexique augmente, il devient difficile de conserver des représentations très imprécises. Pour Walley (2007), les représentations lexicales – chez les enfants en L1 (Hallé & Boysson-Bardies, 1996) comme chez les apprenants de langue étrangère – sont au départ sous-spécifiées, se précisant peu à peu sous la pression d'un nombre de mots connus toujours plus grand. Quand le vocabulaire est très restreint, il est facile de distinguer les mots les uns des autres par une représentation holistique de leur prononciation, et par les traits saillants qui distinguent chaque mot d'un autre. Au fur et à mesure que le vocabulaire se densifie, de plus en plus de mots différents se ressemblent formellement, et leur mémorisation oblige à une spécification plus fine de ces

formes[49]. Bianco (2016, p. 84) remarque par exemple, à propos de l'acquisition L1 : « l'augmentation de la taille du lexique pousse l'enfant à se centrer sur les unités phonologiques qui permettent de distinguer les mots les uns des autres ». Bungaard-Nielsen et al. (2011) constatent ainsi une corrélation entre étendue du vocabulaire et capacités de discrimination phonémique (avec des japonophones apprenant l'anglais en Australie). Nous avons là un cercle vertueux, où le fait d'apprendre du lexique permet d'affiner ses représentations phonologiques, et, en retour, le nouveau lexique qui est appris est mieux représenté : « Le stockage de nouveaux mots en mémoire dépend de la qualité de l'identification des différents phonèmes qui composent la structure phonologique lors de l'audition » (Magnat, 2013, p. 42).

11.1.2. Lexique et phraséologie

Si le mot est l'unité naturelle pour travailler l'accent lexical, et également l'unité appropriée pour travailler les contrastes phonémiques, qu'en est-il des collocations, qui impliquent la combinaison de plusieurs mots, et donc un traitement supérieur au niveau lexical ? On pourrait penser que, comme elles forment un tout du point de vue du sens, elles sont difficilement décomposables. Wray (2000, p. 465), par exemple, donne cette définition des unités phraséologiques : « *a sequence, continuous or discontinuous, of words or other meaning elements, which is, or appears to be, prefabricated: that is, stored and retrieved whole from memory at the time of use, rather than being subject to generation or analysis by the language grammar* ». Si ces collocations sont préfabriquées et traitées comme un tout, il n'est peut-être pas intéressant, d'un point de vue didactique, de les aborder au niveau des mots qui les composent. Dans son livre de 2008, Alison Wray va plus loin, et fait l'hypothèse que les collocations sont traitées comme des morphèmes, c'est-à-dire comme des associations forme-sens élémentaires : « *a word or word string […] that is processed like a morpheme, that is, without recourse to any form-meaning matching of any sub-parts it may have* » (Alison Wray, 2008, p. 12).

Siyanova-Chanturia (2015) explique au contraire que le fait que les formules (*formulaic language*) soient traitées (comprises et prononcées) plus rapidement ne signifie pas forcément qu'elles soient mémorisées comme un tout sans qu'on ait accès à leurs constituants. La

[49] Wedel (2012) étend cette idée à l'échelle du lexique entier : l'existence de paires minimales force à distinguer entre phonèmes proches, et plus il y a de paires minimales distinguant deux phonèmes proches, plus la distinction a des chances de se maintenir.

rapidité de traitement peut être attribuée à deux causes (travaillant probablement de concert) : d'une part, le tout devient plus saillant par rapport aux parties (qui restent cependant accessibles), et d'autre part, comme les mots composant la formule sont souvent entendus ensemble, leur fréquence de co-occurrence est élevée, ce qui conduit les auditeurs à anticiper l'arrivée des mots suivants, et donc à les traiter plus rapidement. Il est donc possible d'avoir accès au tout à partir des parties. D'un point de vue pédagogique, étant donné la taille du lexique à apprendre, il y a besoin d'un principe organisateur qui mette un peu d'ordre dans les milliers d'items lexicaux à acquérir (« *the chaos of the lexicon* » d'après Lewis, 1997, p. 67), et ce sera le ou les mot(s) le(s) plus saillant(s) de la collocation qui pourra tenir lieu d'accroche pour la mémorisation. C'est d'ailleurs aussi de cette façon que les expressions figées sont en général accessibles dans un dictionnaire. Michael Lewis, tout en soulignant qu'il ne peut y avoir un seul principe organisateur qui fonctionne pour toutes les collocations, propose par exemple d'utiliser comme mots-clés certains verbes supports comme *take* (Lewis, 1997, p. 75).

Au final, le mot peut être un point d'entrée pour les structures de niveau plus élevé (grammaticales ou phraséologiques) dans lesquelles il est inséré. On peut ainsi se rapporter aux propos de John Sinclair : « *learners would do well to learn the common words of the language very thoroughly, because they carry the main patterns of the language* » (J. Sinclair, 1991, p. 72). Dans nos propositions de remédiation, c'est donc le mot qui sera au centre du dispositif, le point focal à partir duquel les autres connaissances (qu'elles appartiennent aux niveaux inférieurs, phonologiques ou morphologiques, ou supérieurs, phraséologiques ou grammaticaux) seront abordées.

11.2. Contextualisation et compositionnalité

11.2.1. Approche compositionnelle ou holistique

Une approche diagnostique qui essaie d'identifier les composantes essentielles de la compréhension pour faire porter la remédiation précisément sur ces éléments peut paraître incompatible avec une approche « holistique » de la langue étrangère à enseigner qui a les faveurs de la communauté éducative depuis l'avènement de la méthode communicative dans les années 1970. A l'époque, la focalisation sur la « compétence communicative » (Celce-Murcia et al., 1995; Hymes, 1972; Savignon, 1976) s'était faite en réaction à une vision de la langue centrée sur la forme, et dont le sens (en particulier pragmatique) semblait avoir été évacué. Cette focalisation sur la forme supposait également une approche allant du plus petit

vers le plus grand (du mot à la phrase, par exemple). Roger Shuy (1981), dans un article au titre évocateur, *The Holistic View of Language*, caractérise ainsi l'approche qu'il qualifie de réductionniste (parce qu'elle tente de réduire la complexité du système en se focalisant sur ses parties) : « *The underlying learning theory […] is that learners learn best small things before large things and that by taking natural language apart and by cutting it into little pieces, the learner can best benefit.* » (p.105). Il insiste au contraire, comme Hymes avant lui, sur le contexte (textuel, discursif, culturel, social) comme une caractéristique nécessaire de l'apprentissage des langues (L1 ou L2). Les éléments langagiers ne peuvent être appris isolément et ne prennent tout leur sens que quand ils sont contextualisés, c'est-à-dire insérés dans une situation d'énonciation, un contexte social, et remplissant une fonction langagière que les apprenants ont besoin d'exprimer. L'approche actionnelle prônée par le CECR est également dans cette lignée, puisque le CECR reprend la notion de compétence communicative (en la redéfinissant légèrement), et insiste sur le contexte d'utilisation de la langue comme variable essentielle de son apprentissage. Il distingue par exemple les « domaines » dans lesquelles les activités langagières peuvent être contextualisées, les « situations » possibles à l'intérieur de ces domaines, et les textes écrits ou oraux, à produire ou à comprendre, qui leur sont associés (Conseil de l'Europe, 2001, p. 15; 45-46). L'attention portée au contexte n'est d'ailleurs pas cantonnée à la recherche en didactique, puisqu'elle joue également un rôle crucial dans les théories linguistiques développées depuis les années 1970 : la sociolinguistique (Hymes, 1972; Labov, 1972), l'analyse du discours (Widdowson, 1978), la linguistique de corpus (M. Davies, 2009; J. M. Sinclair, 1987), et plus généralement la linguisique basée sur l'usage (par exemple la grammaire cognitive de Langacker, 1987).

Ce rôle essentiel du contexte n'empêche cependant pas une focalisation simultanée, dans un cadre didactique, sur les « petits morceaux » de la langue, et en particulier sur les mots[50]. Nous retrouvons là la conclusion à laquelle nous sommes arrivée en fin de première partie : pour pouvoir se focaliser pleinement sur la construction du sens (le tout), il faut avoir suffisamment automatisé le traitement de la forme (les parties). Nous pensons qu'il est possible d'adopter dans notre contexte une démarche où, tout en nous attachant particulièrement aux mots (qui sont les unités de base qui nous intéressent, et dont il faut apprendre à reconnaître la forme orale, les collocations, etc.), nous pouvons en même temps

[50] On peut faire le parallèle avec l'apprentissage de la lecture en langue maternelle, où un enseignement explicite du code (correspondances graphèmes- phonèmes), est nécessaire, mais avec un travail de la compréhension en parallèle (Bianco, 2016; Goigoux, 2016), donc une focalisation à la fois sur les unités élémentaires (lettres et sons), et sur le tout que forme un texte.

porter attention au contexte dans lequel ils sont utilisés. Il faut donc essayer de conjuguer, d'une part, une contextualisation qui mette en avant le sens, et d'autre part, un travail décontextualisé, dont on espère qu'il peut conduire à une meilleure généralisation (puisque non lié à un contexte particulier), et qu'il permette de s'affranchir de la charge cognitive inhérente au travail sur tous les niveaux de traitement à la fois.

Une approche compositionnelle, qui décompose le matériau à apprendre en morceaux de taille moins impressionante (que celle d'un énoncé ou d'un texte entiers, par exemple), peut aussi être source de motivation selon Ostyn et Godin (1985). En effet, il est plus facile dans ce genre de cadre de se fixer des objectifs raisonnables et quantifiés, et de constater ses progrès après travail. De même, comme nous l'avions souligné dès notre introduction générale, une approche qui se focalise sur les composants d'un processus peut plus facilement être associée à une remédiation : « *[L]earners need to know what aspects of their performance can be improved and, critically, how they can make that improvement. It is the process of understanding what the goal is, where the learner is now and how they can move towards the goal* » (Fulcher, 2010, p. 69). D'autre part, le retour formatif (*feedback*) peut alors être plus précis, comme nous l'avions remarqué au chapitre trois en résumant l'étude de Kluger et Denisi (1996), qui montrait qu'un *feedback* sur l'activité (niveau inférieur), était mieux perçu et plus efficace que quand il portait sur des niveaux plus élevés (tâche ou sujet).

11.2.2. Contextualisation et authenticité

La contextualisation des éléments linguistiques est souvent liée à la notion « d'authenticité » en didactique des langues : toute contextualisation n'est certes pas authentique (il est parfaitement possible de créer des textes spécifiquement pour un public d'apprenants), mais, inversement, la décontextualisation est toujours artificielle, puisque les sons, les mots, les phrases ne sont que rarement utilisés hors contexte en situation de communication naturelle. Gilmore (2007) passe en revue les avantages liés à l'utilisation de documents authentiques (qu'il définit simplement comme des documents créés par et pour des natifs) dans l'apprentissage des langues étrangères. Le premier avantage mis en avant est que l'authenticité peut jouer un rôle important pour la motivation des apprenants (bien que l'on manque d'études pour évaluer son efficacité réelle, Gilmore, 2007, p. 108). En effet, les apprenants ont l'impression d'être exposés à la « vraie » langue, et donc d'être mieux préparés ensuite à une utilisation en interaction avec des natifs anglophones ou d'autres étrangers avec qui l'anglais sera la langue de communication partagée.

L'authenticité garantit également des contextes plus variés, une richesse lexicale et syntaxique plus représentative de la langue en général – une idée qui a émergé très tôt : « *The great advantage of natural, idiomatic texts over artificial 'methods' or 'series' is that they do justice to every feature of the language* » (Sweet, 1899, cité par Gilmore, 2007, p. 97). On pourra donc faire l'hypothèse que la contextualisation, par l'intermédiaire de documents authentiques, permettra de mieux exposer les apprenants à un vocabulaire, à des structures (et à des voix) plus variées. Elle peut également donner une représentation plus fidèle de la langue à acquérir, et en particulier de la fréquence d'occurrence et de co-occurrence des éléments, qui ne sera pas biaisée par les préférences ou les objectifs pédagogiques des enseignants ou des méthodes utilisées. Or, nous avons vu (section 1.4.4.2) qu'il était important pour le traitement rapide du signal que les apprenants soient sensibilisés à la fréquence des unités linguistiques, et en particulier la fréquence de leur co-occurrence avec d'autres éléments (N. C. Ellis, 2002). C'est cette connaissance qui leur permet de prédire la suite du message, d'anticiper ce qu'ils vont entendre et ainsi d'être plus efficaces dans le traitement de la langue orale.

11.2.3. Variabilité et théorie des exemplaires

Nous avons déjà parlé de la variabilité inhérente à la parole au paragraphe 1.1.3 (due à la vitesse d'élocution et à la hauteur de la voix, mais nous pouvons rajouter ici l'âge, l'appartenance ou l'identification à un groupe social, etc.). Nous avions d'autre part résumé les études de Johnson (2004) et Greenberg (1999) qui montraient qu'une majorité de mots en discours ne sont pas prononcés sous leur forme de citation (c'est-à-dire avec leur prononciation hors contexte, dite « du dictionnaire »). Pierrehumbert (2016) passe en revue les effets du contexte sur la forme orale des mots : les mots plus fréquents ou plus prévisibles en contexte sont articulés moins clairement (comme il est plus facile de les prédire, moins d'informations sont nécessaires pour les reconnaître, comme nous l'avons vu en 1.4.4.4, et les locuteurs en profitent pour faire moins d'efforts d'articulation). La prosodie, qui dépend de la structure grammaticale et informationnelle de la phrase, modifie aussi la prononciation des mots (par exemple, dans les mots non accentués, il est beaucoup plus difficile de faire la différence entre accent primaire et accent secondaire, Plag et al., 2011).

Il est important que les apprenants soient à même de reconnaître les mots dans le flux du discours par-delà ces différences. On pourrait penser que la présentation d'une forme abstraite décontextualisée (par exemple phonémique, ou même orthographique, et en tout cas sous-

spécifiée) permette de couvrir toute cette variabilité et donc de généraliser à tous les contextes dans lesquels le mot peut être rencontré, la sous-spécification rendant cette forme abstraite compatible avec beaucoup d'instances contextualisées et donc spécifiées. C'est l'idée même de la « normalisation » du flux sonore entrant (qui correspond à l'étape « décodage » du modèle de Cutler et Clifton, cf. Figure 1.3). Cependant, nous avons vu que les auditeurs exploitent également des informations sub-phonémiques (en particulier pour prédire le son qui va suivre), qui devraient logiquement être perdues lors de la normalisation phonémique (phénomène de coarticulation, 1.4.4.1). D'autre part, ils sont également sensibles aux voix qui prononcent les mots qu'ils entendent, une information qui devrait également disparaître lors de la transformation en suite de phonèmes : Hintzman et al. (1972) montrent entre autres que les auditeurs se souviennent s'ils ont entendu un nouveau mot de vocabulaire avec une voix de femme ou d'homme.

Confrontés à ces données, certains chercheurs ont rejeté l'idée d'une normalisation nécessaire, en faveur d'une théorie exemplariste (ou épisodique). Pour Goldinger (1998), par exemple, chaque exemplaire lexical entendu et reconnu est mémorisé avec sa voix (caractéristiques phonétiques) et son contexte. Quand on réentend le mot, on envoie une sonde (*probe*) en mémoire à long terme, qui renvoie en écho tous les épisodes similaires stockés en mémoire. Plus il y en a, et plus l'écho est puissant (et plus le temps de réponse est donc rapide, ce qui correspond aux données psycholinguistiques montrant que les mots fréquents sont reconnus plus rapidement que les autres, 1.4.4.2). Quand c'est un mot courant, l'effet d'une voix individuelle est noyé dans la masse, tandis que s'il est rare, le « *token repetition effect* » (le fait de mieux reconnaître un mot s'il est prononcé avec une voix qu'on a déjà entendue) sera beaucoup plus important. L'effet d'abstraction est donc obtenu par superposition de tous les épisodes singuliers (*generic echo*), et le fait d'avoir entendu le même mot prononcé par de nombreuses voix différentes permet ainsi une meilleure abstraction (il n'y a donc pas de contradiction entre mémorisation de détails très spécifiques et généralisation, au contraire). Barcroft et Summers montrent, avec des apprenants L2, que l'exposition à de nombreux exemplaires différents de nouveaux mots, et surtout prononcés par des voix différentes, aide à la mémorisation, justement parce qu'elle permet la création de catégories plus générales et abstraites, et, au final, plus robustes : « *[A]coustically varied instances of each new lexical item in the input combine to form a representation that is more robust than would have been obtained by an equivalent number of acoustically consistent instances of the same item (i.e., multiple presentations of the same speech signal)* » (Barcroft & Sommers, 2005, p. 405).

Cette variété de voix entendues dessert légèrement la mémorisation à court terme (puisque l'on reconnaît mieux un nouveau mot prononcé par une voix familière), mais est préférable à long terme[51]. En effet, si l'apprenant s'habitue aux caractéristiques spécifiques d'une seule voix (par exemple celle de son enseignant), son système perceptif risquera d'être « surajusté » à cette voix, et d'utiliser des indices phonétiques qui lui sont spécifiques et ne se généralisent pas à celle d'autres locuteurs.

L'entraînement perceptuel à haute variabilité (*High Variability Perceptual Training,* ou HVPT) exploite le même phénomène, mais au niveau de la construction des catégories phonémiques cette fois (les sons sont alors contextualisés dans des mots). Lively et al. (1993) l'utilisent par exemple pour aider des locuteurs japonais à mieux distinguer le contraste anglais /r/ - /l/ : on leur propose un grand nombre d'exercices de discrimination avec des stimuli (mots contenant /r/ ou /l/) produits par plusieurs locuteurs, dans plusieurs environnements phonétiques (début ou fin de mot, placé après une voyelle ou une consonne). Les entraînements durent quelques heures, réparties sur plusieurs séances, et peuvent aussi être composés d'exercices d'identification des sons et de « cherchez l'intrus » (par exemple, Krzonowski et al., 2016). Lively et ses collègues ont montré que la généralisation nécessite plusieurs voix et plusieurs environnements phonétiques. Pour être sûr que les apprenants travaillent au niveau lexical et non pré-lexical, Melnik et Peperkamp (2019a) proposent de travailler uniquement avec des mots existants, et mentionnent que les apprenants apprécient de savoir que tous les mots qui leur sont présentés lors des exercices existent vraiment. Elles montrent de plus que cet entraînement a alors des conséquences sur la reconnaissance lexicale et pas seulement sur la construction des catégories phonémiques.

Nous en arrivons donc à la conclusion qu'il est vain d'opposer contextualisation et décontextualisation, et qu'il faut au contraire promouvoir à la fois une conscience des petits éléments constitutifs de la langue (pour nous, les items lexicaux), et en parallèle une exposition à ces mêmes éléments dans de nombreux contextes différents qui reflètent leur usage le plus fidèlement possible, dans toute sa richesse (par exemple de co-occurrence droite et gauche) et sa variabilité.

[51] La variété des contextes utilisés peut également avoir des conséquences sur l'acquisition du sens des mots (et pas uniquement de leur forme), si l'on considère que le sens d'un mot est en grande partie défini par ses contextes gauche et droit, et que, comme le proposent Lund et Burgess (1996), les mots peuvent être représentés sous la forme de vecteurs dans un espace multidimensionnel qui représente les co-occurrences auxquelles un apprenant a été exposé (voir aussi Elman, 2004).

11.3. Apprentissage du lexique

Nous venons de voir qu'il faut essayer de conjuguer dans nos activités une présentation décontextualisée des mots, avec leur prononciation canonique, et l'exposition à ces mêmes mots dans des contextes les plus variés possible. Il existe déjà de nombreuses recherches sur l'apprentissage du lexique, en L1 comme en L2, qu'il s'agisse de l'apprentissage complètement décontextualisé de listes de mots, de l'apprentissage totalement contextualisé (par exemple lors de la lecture d'un livre pour le plaisir, ou, à l'oral, par l'écoute extensive, c'est-à-dire l'exposition simple à la langue dans des contextes authentiques comme des émissions de télé, des films, des chansons, des cours, conférences, conversations, etc.), ou prenant des formes intermédiaires (apprentissage du vocabulaire contenu dans un texte travaillé en classe de langue). La plupart de ces études utilisent la forme écrite des items lexicaux considérés, mais certaines (que nous mettrons en exergue) portent également sur leur forme orale. Nous n'avons pas le loisir ici de faire un état des lieux complet des études, en nombre considérable, sur la question, mais nous présenterons brièvement quelques résultats sur l'acquisition en L2, qui montrent que généralement, le vocabulaire est mieux retenu s'il est présenté hors contexte que contextualisé (par ex. Folse, 2013; Laufer & Shmueli, 1997), avec un équivalent L1 plutôt qu'une définition en L2 (Laufer & Shmueli, 1997; Ramachandran & Rahim, 2004), et à l'écrit plutôt qu'à (ou en plus de) l'oral (Ronan Brown et al., 2008; Vidal, 2011).

11.3.1. Apprentissage décontextualisé (listes de mots)

On sait que l'apprentissage de listes de mots est efficace pour la rétention à long terme, d'autant plus que les mots sont accompagnés d'un équivalent en L1 plutôt que d'un synonyme ou définition en L2 : dans l'expérience de Laufer & Schmueli (1997), des lycéens israéliens apprenant l'anglais ont retenu, cinq semaines après le traitement, 75% des mots présentés seuls ou dans une courte phrase, et 60% des mots présentés dans un texte (le test de rétention étant un QCM avec synonymes en anglais), montrant qu'il est plus facile de retenir une liste de mots écrits décontextualisés (ou peu contextualisés) que contextualisés. De plus, les mots dont une traduction en L1 était donnée ont été mieux retenus, dans toutes les conditions.

Il est donc efficace d'apprendre du vocabulaire ainsi en vue d'un test sur la reconnaissance du sens (QCM), mais qu'en est-il pour la compréhension aurale ? Van Zeeland (2013) montre que la connaissance des mots hors contexte ne se traduit pas nécessairement par leur

reconnaissance en contexte. Les apprenants qui connaissent parfaitement le mot hors contexte (ils doivent en donner une traduction après l'avoir vu à l'écrit et entendu à l'oral) ne le reconnaissent pas dans une narration orale dans 20% des cas (il ne s'agit pas uniquement d'un problème de segmentation de la langue orale, puisque 9% des mots connus hors contexte ne sont pas reconnus non plus dans un texte écrit). Inversement, ceux qui ne connaissent pas le mot sont aidés par le contexte dans 25% des cas.

En situation d'enseignement, l'utilisation d'une liste de mots à préparer en vue de la réalisation d'une tâche de CO semble également peu efficace, du moins quand le travail s'effectue juste avant la tâche, et que la forme orale des mots n'est pas travaillée explicitement. Chang (2007) a donné à trois groupes d'étudiants taïwanais une liste de mots anglais à étudier en vue d'un exercice de compréhension. Le premier et le deuxième groupes disposaient respectivement d'une semaine et d'une journée pour apprendre les mots individuellement, et le troisième groupe de 30 minutes seulement, pour travailler les mots en classe, de façon collective, juste avant d'écouter le texte. Comme on pouvait s'y attendre, plus les étudiants ont eu de temps pour apprendre les mots de la liste, meilleurs sont leurs résultats au test de vocabulaire. Pour la compréhension de l'oral, par contre, les résultats du premier groupe ne sont que marginalement supérieurs à ceux des autres groupes, le problème principal étant qu'ils n'avaient pas suffisamment appris la forme orale des mots à étudier. Une étude similaire (Chang & Read, 2006) a cependant montré que, même avec un travail préalable (pendant l'heure qui précède le test de CO) sur une liste de vocabulaire comprenant la forme orale des mots, les résultats sont moins bons qu'avec d'autres types d'aide (prévisualisation des questions, deuxième écoute ou travail en L1 sur le thème du texte). Il semble que les apprenants, au lieu d'essayer de comprendre le texte, tentent au contraire d'y reconnaître les mots qu'ils viennent d'apprendre et qu'ils n'ont pas encore assimilés, étant donné le délai très court entre le travail hors contexte et la tâche de compréhension.

Au final, la présentation décontextualisée semble utile et efficace pour établir le lien initial forme-sens (premier niveau de la connaissance d'un mot selon Nation, 2001, cf. 2.3.4.1), à condition de ne pas oublier de présenter également la forme orale, mais ce lien initial devra probablement être enrichi ensuite par d'autres rencontres (contextualisées) avec le mot, afin d'accéder à d'autres aspects de la connaissance du mot. Bien qu'il y ait beaucoup de réticences à proposer des listes de mots aux apprenants, surtout dans un cadre communicatif (encore une fois parce que ces mots sont alors complètement décontextualisés), ces listes peuvent donc avoir leur utilité : la rapidité de l'appariement forme-sens, mais aussi la

possibilité de révision systématique pour une mémorisation à long terme. Les listes de mots insérées dans une application informatique permettant de mettre en place un programme de révision espacée (*spaced repetition*) semblent particulièrement efficaces (Nakata, 2008). Mentionnons pour terminer l'utilisation de carnets de vocabulaire par les étudiants de langue étrangère. Ces carnets, similaires à une longue liste de mots ou à un petit dictionnaire, contiennent des mots sortis de leur contexte, mais dont le choix, ainsi que les informations attachées à chaque mot, est laissé libre à l'apprenant. C'est donc un outil hautement personnalisable (« *a fully personalised learning aid* », Lewis, 1997, p. 76), qui paraît tout à fait adapté à une situation de remédiation (et qui peut éventuellement être dématérialisé, et accessible depuis un ordinateur ou un téléphone portable).

11.3.2. Intérêt d'une présentation multimodale

Nous savons que nos étudiants ont besoin de travailler sur la forme orale des mots qu'ils connaissent souvent mal, mais la question se pose de l'utilité de la présentation simultanée de la forme écrite. La théorie du double codage de Paivio (1991), qui suppose que les modalités visuelle, auditive, haptique, etc. sont indépendantes les unes des autres et se renforcent mutuellement pour la mémorisation[52], plaide en faveur de cette double présentation. La théorie de la charge cognitive de Sweller (2014) va également dans le même sens, en supposant un « effet de modalité » qui fait que l'utilisation simultanée des canaux auditif et visuel permet d'augmenter la capacité d'attention et de mémorisation du sujet[53]. Les études sur l'apprentissage lexical en langue étrangère, que ce soit en condition incidente (*incidental acquisition*), c'est-à-dire sans intention délibérée d'apprendre du vocabulaire, ou en condition d'apprentissage délibéré (explicite), semblent confirmer ces deux théories (dont la deuxième s'est d'ailleurs inspirée de la première).

Brown et al. (2008) montrent par exemple que l'écoute seule est beaucoup moins efficace que la présentation simultanée des formes écrite et orale dans une expérience d'acquisition incidente de l'anglais par des japonophones. Ils comparent l'acquisition dans trois conditions : exposition extensive orale, écrite, ou orale et écrite simultanément (dans les trois cas, il s'agit de lire/ écouter une histoire). Le nombre de mots nouveaux appris est relativement faible dans

[52] A l'intérieur de chaque modalité, une division verbal/ non verbal est également postulée (par exemple, à l'intérieur du mode visuel, un mot écrit et une image ne seront pas traités par le même système, et pourront se renforcer l'un l'autre).

[53] La théorie de la charge cognitive met cependant en garde contre un « effet de redondance » qui fait que, quand un seul canal (par exemple visuel) est suffisant, l'utilisation inutile d'un autre canal en parallèle conduit à une moins bonne réalisation de la tâche (de mémorisation ou de compréhension).

les trois conditions : après trois mois (sans révision entre temps), les élèves ne sont capables de donner le sens que d'un seul des 28 mots testés (dans la condition « lecture »), voire d'aucun (condition « écoute »). Cependant, quand la rétention est mesurée avec un test de reconnaissance du sens (comme dans l'expérience de Laufer et Schmueli, 1997, résumée précédemment), les sujets ont retenu entre 25% (écoute seule) et 46% des mots (lecture, ou lecture et écoute simultanées). L'acquisition lexicale lors de l'écoute seule est donc non négligeable, mais nettement plus faible qu'avec l'écrit en sus. De même, Lin et Yu (2017), dans une expérience d'apprentissage décontextualisé cette fois (avec des *flashcards* électroniques dans une application pour téléphone mobile), constatent lors du post test différé, après deux semaines, que leurs apprenants (des collégiens taïwanais) ont mieux retenu les mots étudiés en modalité combinée texte, image et audio que dans les autres conditions.

La présentation de la forme écrite en plus de la forme orale aide donc à la mémorisation des items lexicaux, même si elle n'a pas que des avantages. Bürki et al. (2019) montrent par exemple que l'apprentissage d'un mot anglais avec son orthographe (en plus de sa prononciation), si elle aide à se rappeler le mot lors d'un test de production de la forme orale, conduit également à une plus grande influence phonologique de la L1 (le français dans leur cas) sur la forme prononcée. Comme nous sommes dans une situation de réception, ce problème potentiel nous concerne moins, mais nous renforce dans l'idée qu'il faut porter une attention particulière à la forme orale, pour faire en sorte que la forme phonologique des mots dans le lexique mental des apprenants soit la plus précise possible. En tout état de cause, nous présenterons la forme écrite en plus de la forme orale lors des activités proposées aux étudiants. Le travail sur la forme écrite peut également être l'occasion de renforcer leur connaissance des correspondances graphèmes-phonèmes de l'anglais, et la forme écrite est bien sûr plus pratique pour garder une trace du mot et y accéder dans un carnet de vocabulaire, par exemple.

11.3.3. Apprentissage explicite et implicite

Nous avons vu que, pour établir le lien forme-sens initial, il est plus efficace d'utiliser une présentation décontextualisée et explicite. En effet, une des difficultés de l'apprentissage lexical incident dans des conditions de lecture ou d'écoute extensive est que le sens n'est pas toujours inférable aisément du contexte. Pigada et Schmitt (2006) montrent par exemple qu'un apprenant peut avoir tendance à inférer le sens sur des critères formels et être induit en erreur par des ressemblances avec d'autres langues de son répertoire (par exemple leur

apprenant de français confondait « sable » et l'anglais « *stable* », bien que le sens de ces deux mots soit peu compatible). Il peut également confondre des mots de forme similaire dans la langue cible (« éteindre », « atteindre », « entendre »), ce que Laufer (1988b) appelle des « *synforms* ». Certains apprenants effectuent également une analyse morphologique erronée et comprennent par exemple *outline* comme voulant dire « *out of the line* » (Laufer, 1988a, p. 12).

Il est donc important d'établir le sens explicitement, surtout dans notre cadre d'activités de remédiation en ligne. L'apprentissage moins guidé, où l'apprenant doit lui-même générer le sens, est parfois considéré comme supérieur à l'apprentissage guidé où le sens est donné dès le départ, parce que le travail cognitif de génération de la solution est supposé aider à la mémorisation subséquente. Cependant, comme nous l'avons vu, cette hypothèse ne se vérifie pas dans les recherches sur l'acquisition des L2, où l'apprentissage lexical est meilleur quand le sens est donné aux apprenants, sous sa forme la plus transparente possible (donc avec une traduction en L1, au moins pour les niveaux faibles). Les psychologues cognitivistes soulignent d'ailleurs que l'important est moins la façon dont on apprend ou rencontre le mot initialement, que ce qu'on en fait ensuite. Selon Anderson et Schunn (2000), par exemple, le mode de mémorisation initiale importe peu : « *There are no magical properties conveyed upon a knowledge structure just because it was self-generated* » (p. 5), et l'essentiel est le temps total passé à pratiquer : « *For competences to be displayed over a lifetime, time on task is by far and away the most significant factor* » (p. 15). La distinction entre apprentissage implicite et explicite ne nous occupera pas davantage. En effet, un paradigme d'apprentissage implicite est peu pertinent dans notre cas, où les apprenants sont engagés dans un processus d'amélioration consciente de leurs compétences en compréhension aurale en anglais. De nombreux psycholinguistes (DeKeyser, 2014; Hulstijn, 2012; Spada, 2015) pensent d'ailleurs que la distinction entre apprentissage implicite et explicite n'est pas utile dans un cadre d'apprentissage en milieu institutionnel, entre autre parce que la question de ce qui les distingue n'est pas encore clairement tranchée :

> *Given how difficult it is to determine whether knowledge is implicit or explicit (and even more whether learning was implicit or explicit), even under controlled laboratory conditions, it stands to reason that the implicit/ explicit distinction in this narrow sense should be of little concern to second language learners and teachers*. (DeKeyser, 2014, p. 106)

11.4. Propositions pour un dispositif de remédiation en CO

11.4.1. Evaluation diagnostique et formative

Pour décrire les caractéristiques du dispositif que nous aimerions concevoir, nous nous appuierons tout d'abord sur l'article de Bennett (2011) sur l'évaluation formative. Dans le chapitre sur l'évaluation, nous n'avons pas mentionné le terme d'évaluation formative, définie par Bennet comme « *assessment for learning* ». Nous avons bien sûr défini l'évaluation diagnostique, qui est au cœur de ce travail, et qui sert à identifier les points forts et les points faibles d'un apprenant, et éventuellement à lui proposer ensuite des pistes d'amélioration (Chapitre 3). Ce n'est que quand les résultats de l'évaluation diagnostique sont effectivement insérés dans un dispositif de formation (que Demaizière, 2008, appelle aussi « système de formation ») que nous pouvons utiliser ce terme. Il est vrai que, dans la pratique, la distinction entre évaluation diagnostique et évaluation formative n'est pas facile à faire (Huhta, 2008). Même si l'évaluation formative est en général basée sur un programme de cours (répondant à la question « Les apprenants ont-ils bien assimilé le contenu du cours ? Peut-on passer à la suite ? »), tandis que l'évaluation diagnostique peut (et doit, à notre avis) s'appuyer sur une théorie du développement de la compétence en jeu (répondant à la question « A quel endroit du processus le problème se pose-t-il ? »), les deux types d'évaluation sont censés être accompagnés d'actions de formation qui doivent aider l'apprenant à s'améliorer. Cependant, la focalisation dans le cas de l'évaluation diagnostique est sur le test, tandis qu'elle porte sur l'enseignement dans l'évaluation formative. Cette dernière doit être composée à la fois d'un test (diagnostique, donc, dans notre cas) et d'une action d'enseignement (remédiation pour nous). Le(s) test(s) ou actions d'évaluation doivent être accompagnées d'un argument de validité (ce que nous avons fait dans la deuxième partie pour chaque test), et les actions d'enseignement doivent être justifiées par un « argument d'efficacité », que nous détaillons dans ce chapitre. Bennet (2011) suppose que cette action a lieu dans une salle de classe, mais nous supposerons ici qu'elle s'insère dans un système de formation hybride (avec une partie de cours en présentiel) ou entièrement à distance.

Le Tableau 11.1 présente l'architecture du dispositif proposé d'évaluation formative. Au moment de leur inscription administrative, les étudiants passent le test de positionnement SELF. En fonction de leur résultat en compréhension de l'oral, ils sont orientés ou non vers les tests diagnostiques que nous avons conçus. Un entretien serait souhaitable à cette étape,

pour expliquer à l'apprenant les buts et le fonctionnement du dispositif, ou lui laisser exprimer ses doutes ou son éventuel désaccord avec le résultat du positionnement. Il est en effet essentiel que l'action de remédiation soit acceptée pour qu'elle ait un effet positif. L'entretien doit également servir à se fixer des objectifs et un calendrier réalistes. Selon leur score à chacun de ces tests, les apprenants sont ensuite dirigés vers les activités de remédiation correspondantes. Dans ce qui suit, nous ne décrirons que le module de remédiation portant sur le lexique, proposé en priorité aux étudiants avec un score inférieur au score de césure aux tests AURLEX et PVST, mais que les autres étudiants seront également encouragés à suivre, étant donné la centralité du rôle du lexique. Pour les étudiants qui ont des difficultés de discrimination phonémiques et d'identification de l'accent lexical, nous aimerions proposer des activités d'entraînement perceptuel à haute variabilité (HPVT) comme décrit plus haut (11.2.3). Pour attirer l'attention des apprenants sur les détails grammaticaux présents à l'oral qu'ils ne remarquent pas dans le flux de la parole, nous aimerions proposer des activités de type « *processing instruction* » (*PI*, Van Patten, 2002). Comme nous l'avons vu (2.4.3), les informations grammaticales (par exemple le temps verbal) sont souvent négligées parce qu'elles sont également exprimées par des éléments lexicaux (par exemple, des adverbes temporels) qui bloquent leur apprentissage. L'idée est de proposer aux apprenants des exemples où aucune information lexicale redondante n'est présente, ce qui les oblige à traiter l'information morphosyntaxique proprement dite pour comprendre la phrase.

action d'évaluation	condition	action de remédiation	temporalité
SELF (positionnement)	SELF_CO < B2	passage des tests diagnostiques + entretien	inscription administrative
PHON	score < 25	HVPT	à négocier
PROSO	score < 41	HVPT	à négocier
AURLEX	score < 27	parcours lexical	à négocier
PVST	score < 26	parcours lexical	à négocier
AURGRAM	score < 21	PI	à négocier

Tableau 11.1 - architecture du dispositif formatif dans lequel s'insèrent les tests diagnostiques

Nous ne décrirons pas plus avant certaines caractéristiques d'une autoformation réussie auxquelles nous avons déjà fait allusion au cours de ce travail : l'élaboration d'un retour formatif (feedback) précis pour chaque activité (voire chaque item), l'autoévaluation, la visualisation de ses progrès qui aide à la motivation (Ng, 2015). Nous nous pencherons cependant sur l'individualisation nécessaire à toute action de remédiation, et entre autres sur la possibilité de choisir les activités sur lesqelles faire porter ses efforts. Pour aider les étudiants à s'organiser, chaque parcours est divisé en 10 à 12 unités qui correspondent au

nombre de semaines dans un semestre universitaire classique mais qui peuvent être complétées au rythme de chacun.

11.4.2. Conception des activités

Pour la conception d'activités de remédiation centrées sur le lexique, nous nous appuierons sur Schmitt (2008), qui décrit les caractéristiques auxquelles les concepteurs doivent s'intéresser en priorité. Tout d'abord, il faut utiliser des activités qui permettent un engagement maximal des apprenants avec les mots considérés, c'est-à-dire qu'elles garantissent qu'ils soient obligés d'y porter attention et de passer du temps à les traiter. Ensuite, il faut maximiser le nombre de répétitions de chaque mot auquel sont exposés les apprenants (nous pouvons ajouter ici qu'il faut que ces écoutes répétées se fassent avec des voix différentes). Les études qui se penchent sur l'effet de la répétition sur l'acquisition du vocabulaire aural trouvent en général qu'il est non négligeable (Ronan Brown et al., 2008; Vidal, 2011), même si le nombre de répétitions nécessaires est plus élevé que pour l'écrit (jusqu'à une trentaine, d'après l'étude de Brown et collègues). Enfin, il faut choisir quels aspects de la connaissance d'un mot sont cohérents avec nos objectifs d'apprentissage. Nous avons déjà montré que ce qui importait dans notre contexte de réception aurale était l'établissement du lien forme orale – forme écrite – sens, ainsi que l'exposition à des contextes variés pour sensibiliser les apprenants aux collocations les plus courantes. Nous commencerons par décrire comment nous avons choisi les mots à intégrer à nos activités de remédiation, avant de présenter la conception des activités elles-mêmes, qui prend en compte le besoin de répétition, d'engagement cognitif, et de contextualisation.

11.4.2.1. Choix des mots

Au moment de concevoir des activités portant sur le lexique, un certain nombre de questions se posent. Tout d'abord, quel lexique choisir ? Nous avons déjà répondu à cette question dans la deuxième partie, où nous avons montré que le plus efficace (du point de vue du retour sur investissement en temps d'apprentissage) était de s'appuyer sur la fréquence lexicale : commencer par se focaliser sur les mots les plus fréquents avant de s'intéresser aux moins fréquents. Nous avons vu également que, même si la fréquence est assez corrélée avec les connaissances lexicales, elle est loin de tout expliquer (section 7.4.3.1), et que le type d'exposition reçu par les apprenants est également essentiel (Milton, 2009, p. 26-28, compare par exemple le profil lexical d'un apprenant idéal, imaginé au départ par Paul Meara, dont les connaissances lexicales décroissent avec la fréquence, et les profils constatés de véritables

étudiants, qui s'en démarquent plus ou moins, mais jamais complètement). Par ailleurs, certains mots utiles n'apparaissent pas dans les listes de fréquences parce qu'ils ne sont pas forcément utilisés à la hauteur du rôle de leurs référents dans la vie quotidienne: « *A purely frequency-based pedagogical list is necessarily biased by the nature of the corpus and would ignore low-frequency words which refer to basic concepts that are useful for communicative purposes but rarely spoken or written about by users of the language* » (Benigno & de Jong, 2019, p. 11). L'étude de Brysbaert et al. (2019) montre que ces mots mieux connus (par les natifs) que leur fréquence ne le laisserait penser peuvent en particulier se référer à des objets ou outils (*sanitizer*) ou être utilisés par des enfants (*unicorn*, *nap*). Enfin, les listes de fréquence ne disent pas quels mots sont particulièrement difficiles pour les apprenants L2. Dans leur étude, Benigno et de Jong (2019) intègrent d'ailleurs les intuitions d'enseignants pour modérer les informations données par la fréquence lexicale (les enseignants doivent indiquer où ils placent l'utilité de chaque mot sur une échelle de type Likert).

A défaut de nous fier exclusivement à la fréquence lexicale, nous pourrions nous reposer sur le vocabulaire acquis pendant la scolarité antérieure des étudiants. Cependant, il n'existe pas pour l'instant en France de syllabus lexical dans l'enseignement des langues étrangères dans le secondaire (Hilton, 2019), ce qui conduit nécessairement à un éclatement des connaissances (Horst, 2010, montre que le choix et l'utilisation du lexique par un enseignant peuvent être assez peu systématiques). S'ils ont exclusivement été exposés à l'anglais en cours (ce qui a peu de chances de correspondre à la réalité, étant donné la présence massive de films, chansons, ou séries anglophones accessibles en France), le vocabulaire que connaissent les étudiants entrant à l'université devrait être étroitement déterminé par celui utilisé dans les manuels (pour le collège), et ensuite par les thèmes traités au lycée, qui sont en général larges et abstraits. Pour la classe de seconde, par exemple, la thématique générale pour l'année, « l'art de vivre ensemble », se décline en huit axes dont « le passé et le présent » ou « la création et le rapport aux arts » (Ministère de l'Education Nationale, 2019). Ce guidage très abstrait complique le travail d'identification du lexique probablement enseigné.

Il n'est pas plus facile de déterminer quel vocabulaire a été enseigné pendant la scolarité d'un élève de collège. Les manuels d'anglais de collège contiennent très peu de vocabulaire commun (Peereman, 2019) : sur une année de collège, seuls 10% des mots sont communs à tous les manuels (cette proportion s'élève à 25% quand on considère les quatre années de collège comme un tout, ce qui reste très faible). De plus, on peut y observer une très mauvaise dispersion, c'est-à-dire que beaucoup de mots ne sont utilisés qu'une fois par manuel et

jamais revisités (la constatation est d'ailleurs la même dans d'autres pays européens comme la Suède, du moins pour les manuels d'anglais au primaire, Norberg & Nordlund, 2018). Saragi et al. (1978) avaient déjà fait un constat similaire, en observant que seuls 40% des mots appartenant à la première bande de fréquence – les 1,000 familles les plus courantes – étaient connus de toute une promotion qui obtenait par ailleurs un résultat moyen de 89% à cette même bande. Ces résultats pointent un besoin criant d'individualisation lors de la remédiation, dans la mesure où même si la plupart des mots fréquents sont connus de presque tout le monde, il y a souvent quelques apprenants (différents à chaque fois) qui ignorent tel ou tel item lexical. Dans l'idéal, il faudrait ainsi que les apprenants choisissent eux-mêmes les mots à travailler. La conception d'activités toutes prêtes ne permet pas un tel degré d'individualisation, mais il est possible d'apprendre aux étudiants à utiliser les outils adéquats (dictionnaires, bases de données, applications internet, …), et de leur y donner accès, afin qu'ils puissent poursuivre leur formation seuls au-delà des modules de remédiation mis à leur disposition. Dans cette perspective, les activités proposées jouent le rôle d'exemples d'activités dont les étudiants peuvent constater l'efficacité en les exécutant, ce qui peut les motiver ensuite à continuer seuls (ou en collaboration, si on va plus loin en imaginant faire créer des activités similaires aux apprenants dans le cadre d'un MOOC, par exemple). Nous proposons à cet effet au début de chaque unité une très courte vidéo théorique expliquant les mécanismes d'acquisition du lexique L2 (combinaison d'apprentissage contextualisé et décontextualisé, par exemple) et présentant des conseils pour utiliser des outils qui permettent de personnaliser l'apprentissage (carnet de vocabulaire papier ou électronique, corpus en ligne, …).

Etant donné qu'il n'est pas possible de se baser sur un syllabus lexical, et comme nous avons montré que les listes de fréquence fondées sur les familles de mots ne correspondaient pas aux connaissances de nos étudiants (section 7.4.3.1), nous utiliserons pour cela d'autres corpus. Nous avons décidé de nous appuyer essentiellement sur les listes lemmatisées des corpus COCA, et *Academic Spoken English* (voir Tableau 2.5 pour une liste partielle de corpus disponibles en ligne), ainsi que sur la liste de mots fonctionnels contenue dans la *Essential Word List* (Dang & Webb, 2016), destinée à des apprenants de niveau débutant ou élémentaire. Comme il n'est pas envisageable de créer des activités ciblant des milliers de mots, nous avons utilisé deux autres critères pour restreindre le choix des items à inclure : notre intuition d'enseignante, et un sondage informel auprès de trois étudiants inscrits en première année LLCER anglais suivant un cours de remédiation en présentiel. Comme

Kremmel (2016), nous avons décidé de surreprésenter les mots de la première bande de fréquence, à la fois pour nous assurer que ces mots essentiels soient bien acquis, mais aussi parce que ce sont les plus utilisés dans les expressions non transparentes. Ces mots sont insérés dans des activités réparties en plusieurs unités. Le nombre de mots nouveaux par unité diminue au fur et à mesure du parcours, parce que chaque unité inclut également la révision du vocabulaire des unités précédentes, dont la quantité augmente à chaque fois. Une unité de révision est prévue toutes les quatre unités (Tableau 11.2).

unité	u1	u2	u3	u4	u5	u6	u7	u8
nb. nouv. mots	20	20	15	0	15	10	10	0
bande fréq. nouv. mots	0-500	500-1 000	1 000-2 000		2 000-3 000	3 000-4 000	4 000-5 000	
bande révision		0- 500	500-1 000	0-2 000	1 000-2 000	2 000-3 000	3 000-4 000	0-4 000

Tableau 11.2 - tableau sysoptique du nombre et de la distribution des items lexicaux dans le dispositif de remédiation par unité d'enseignement, ainsi que la répartition par bande de fréquence des mots nouveaux et recyclés

11.4.2.2. Choix des activités

Les activités conçues doivent permettre, dans un premier temps, d'établir le lien sens – forme écrite – forme orale de chaque mot de façon décontextualisée. Nous avons vu qu'il était en général plus efficace d'exprimer le sens en L1, par une traduction française donc, mais nous ne nous interdisons pas d'utiliser une représentation picturale ou un synonyme anglais si besoin, afin de concevoir des activités variées qui vont permettre d'augmenter le nombre de répétitions de chaque mot ou expression. Les activités proposées sont ainsi les suivantes : appariement forme écrite- traduction L1[54], appariement forme orale-traduction L1[55], appariement forme écrite- forme orale, appariement forme orale/ écrite- définition/ synonyme en anglais, appariement forme orale- image représentant le sens, repérage du mot dans une phrase. Pour la forme orale, nous avons utilisé les enregistrements disponibles sur le site *Forvo* (Forvo, 2008), avec des voix différentes selon les activités quand cela était possible (certains mots sur le site n'ayant pas toujours plusieurs enregistrements de qualité suffisante). Des captures d'écran sont disponibles en Annexe 10. Un exemple d'activité intégrée sur la plateforme d'apprentissage en ligne *Moodle* (*Moodle.org*, s. d.) est proposé en Figure 11.1.

[54] Pour des questions de facilité dans des activités en ligne non supervisées, nous avons opté pour une présentation initiale de la forme écrite.

[55] Chung (2003) affirme qu'il faut que le stimulus (pour nous, la forme orale) soit présenté seul quelques secondes avant la réponse (la traduction L1), pour éviter que la forme L1, déjà bien connue, ne bloque l'apprentissage du mot L2 ; cependant, les deux sont présents simultanément dans notre format d'exercices (appariement).

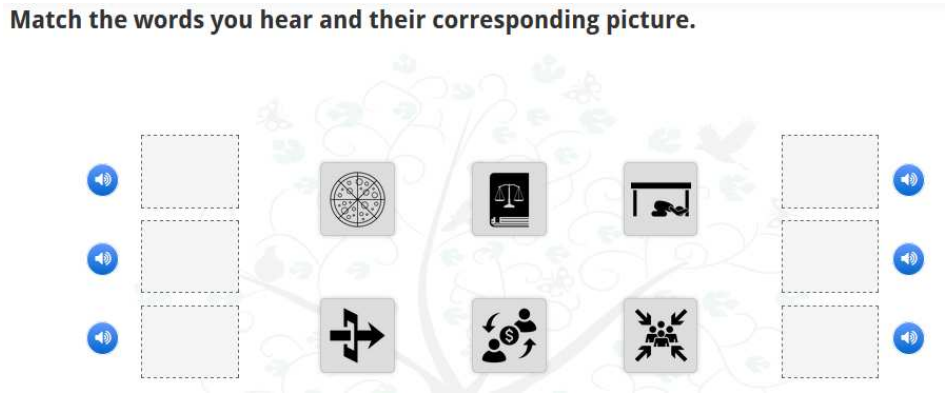


Figure 11.1 - capture d'écran d'une activité d'appariement forme orale – image (sens) pour les mots *whole, law, below, through, trade* et *those*

Dans un deuxième temps, il faut que chaque mot soit entendu dans plusieurs contextes différents, et si possible prononcé par des voix différentes. Pour trouver des extraits audio ou vidéo qui contiennent ces mots, nous avons utilisé le site associé au corpus iWeb[56] de Mark Davies (M. Davies & Kim, 2019), qui donne directement accès, entre autres, à deux autres sites qui permettent d'entendre le mot choisi en contexte. Le premier, *PlayPhrase.me* (Potapenko, s. d.) propose d'entendre le mot (ou expression) dans de très courts extraits de fiction (films ou séries). Le deuxième site, *YouGlish* (*Improve Your English Pronunciation*, s. d.), permet d'accéder à des vidéos de conférences en ligne (un contexte académique, donc) qui contiennent le mot recherché. Il s'agit donc de contextes authentiques, et qui peuvent être motivants pour nos étudiants quand ils reconnaissent une série qu'ils apprécient ou un acteur connu. Pour chaque mot, cinq ou six extraits ont ainsi été choisis, et utilisés dans des activités de reconnaissance du mot, d'identification de la structure ou de l'expression dans lequel il est utilisé, ou de production du sens en contexte. Un exemple d'exercice de reconnaissance aurale (toujours sur *Moodle*) est fourni en Figure 11.2.

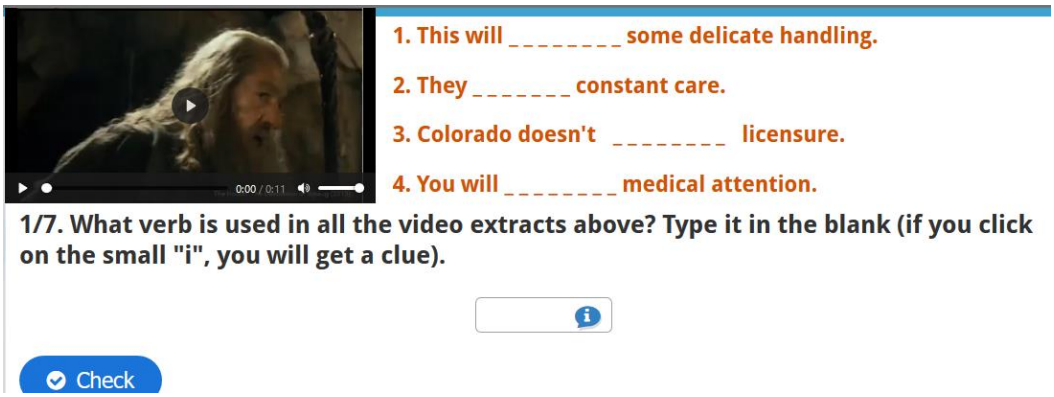


Figure 11.2 - capture d'écran d'un exercice de reconnaissance aurale en contexte (*require*)

[56] https://www.english-corpora.org/iweb/

Enfin, les activités du troisième groupe sont utilisées dans la partie de chaque unité consacrée à la révision des mots introduits précédemment. Il s'agit d'activités de production de la forme, de jeux de mémoire (« memory ») qui vérifient l'association forme orale-sens, ou de jeux de pendu ou de mots croisés pour l'association sens- forme écrite. Les tâches de production de la forme peuvent être associées à des contextes oraux ou écrits. Dans le premier cas, à partir d'extraits vidéo, déjà utilisés ou non dans les activités précédentes, le mot cible qui a été effacé doit être retrouvé (et donné à l'écrit). Dans le deuxième, les contextes sont des phrases ou groupes de phrases tirés du corpus iWeb, et parfois légèrement simplifiés. Dans tous les cas, des indices (en général la première lettre) sont fournis à l'étudiant sur demande, ce qui est important pour des activités non supervisées qui peuvent se révéler assez difficiles à compléter.

Au final, le parcours est conçu pour que chaque mot ciblé soit lu ou entendu une quinzaine de fois (ou plus si l'étudiant a besoin de plusieurs écoutes pour arriver à un score satisfaisant à certains exercices, par exemple). Chacune de ces rencontres avec le mot implique un engagement cognitif puisqu'une opération d'identification ou de production de la forme ou du sens les accompagne : les apprenants sont obligés de remarquer ces mots, et de leur accorder de l'attention (*noticing*, Schmidt, 1990). Cependant, c'est bien sûr insuffisant pour une rétention à long terme, et ces contacts doivent être prolongés par une exposition extensive à la langue orale où le mot sera rencontré de nouveau dans d'autres contextes, prononcé par d'autres voix, etc. Nous espérons que le parcours de remédiation pourra aider les étudiants à comprendre l'importance de cette démarche à long terme.

11.4.3. Pilotage du parcours de remédiation lexicale

Nous avons piloté les trois premières unités du parcours de remédiation avec 11 étudiants de niveau A2 et B1 inscrits en L1 LLCER et LEA. Cela nous a permis de constater que les activités fonctionnaient bien, et que chaque unité correspondait à 30 minutes de travail environ (un temps raisonnable à y consacrer chaque semaine). Un questionnaire a été soumis aux étudiants qui ont noté chaque activité sur une échelle de Likert (de 1 à 5). Les commentaires sur le parcours sont globablement très positifs (« très instructif », « great context »), mais les résultats montrent que plus l'engagement cognitif augmente, moins les activités sont appréciées : les exercices d'établissement du lien forme écrite – forme orale – sens obtiennent en moyenne une note de 4,7/5, ceux de contextualisation 4,3/5, alors que les exercices de rebrassage et de révision n'obtienent que 3,8/5 (ce qui reste bien sûr une bonne

note). Par ailleurs, une forte corrélation émerge entre la satisfaction globale et la réussite globale aux activités ; il est donc important que ces dernières correspondent bien au niveau des apprenants afin qu'ils ne se découragent pas. Nous n'avons pas observé d'effet de plafond, puisque l'étudiant(e) avec un score global approchant de 100% était le (la) plus satisfait(e). Inversement, il est probable que les activités aient été un peu trop difficiles pour les étudiants les plus faibles, dans la mesure où les seuls commentaires négatifs provenaient des deux étudiants avec une réussite moyenne autour de 70% (les autres ayant obtenu plus de 85%). Nous touchons là à une limite de l'utilisation de corpus et de documents authentiques dans l'enseignement des langues étrangères, qui rend plus difficile le contrôle de la difficulté des données langagières que doivent traiter les apprenants.

11.5. Conclusion générale

Notre conclusion générale reprendra les points principaux abordés dans les conclusions partielles précédentes, en particulier la conclusion de la deuxième partie sur la validation de nos tests diagnostiques (9.3), et celle du chapitre 10 sur les relations entre nos différentes variables (10.6). Nous présenterons ensuite les limitations de notre étude et les perspectives pédagogiques sur lesquelles elle débouche.

Dans cette thèse, nous avons adopté une perspective diagnostique sur l'évaluation de la compréhension aurale en anglais langue étrangère, ce qui implique, d'une part, une attention portée aux processus sous-jacents de la compréhension (et pas simplement à son résultat), et d'autre part, un lien explicite avec des activités de remédiation associées : « *it is important to create meaningful linkages between outcomes of diagnosis and subsequent learning and instruction* » (Y.-W. Lee, 2015, p. 295). Nous avons à cet effet conçu cinq tests diagnostiques inspirés des résultats de la recherche en acquisition des langues étrangères, portant sur la discrimination phonémique (chapitre 6), la sensibilité prosodique (chapitre 5), la reconnaissance aurale du vocabulaire (chapitre 7), le jugement aural de grammaticalité (chapitre 8), et les connaissances phraséologiques (PVST, Martinez, 2011), qui sont toutes supposées contribuer à la compréhension de l'oral. Nous avons présenté un argument de validité pour chacun de ces tests, montrant tout d'abord que chacun d'entre eux était fiable, discriminant, et unidimensionnel. Ils sont également pratiques, utilisables en ligne, autocorrectifs, et demandant un temps de passation assez court. Enfin, les items dont ils sont composés reflètent lors de leur administration, comme attendu, une influence de la fréquence lexicale et de la langue maternelle de nos sujets (le français).

Nous avons ensuite poursuivi cette validation (chapitre 10) avec la mise en rapport des résultats à nos cinq instruments diagnostiques, à la fois entre eux, et avec ceux du test SELF en compréhension de l'oral. Cela nous a permis de retrouver encore une fois des résultats bien établis, à savoir d'une part une corrélation importante entre connaissances lexicales et grammaticales, et d'autre part la hiérarchie de corrélations avec la compréhension aurale généralement rapportée dans la littérature scientifique : corrélation faible entre CO et capacité de discrimination phonémique, et moyenne à forte entre CO et connaissances lexicales et grammaticales.

Nous avons également observé une corrélation faible à moyenne entre sensibilité prosodique et compréhension de l'oral en anglais L2, déjà constatée dans une étude précédente avec des sujets japonophones, et que nous confirmons ici avec des sujets francophones. Ce résultat, même s'il n'est pas totalement nouveau, n'est pas non plus encore solidement établi. C'est donc un résultat intéressant, dans la mesure où il est important d'essayer de reproduire les résultats existants en didactique des L2 (et plus généralement en sciences humaines), afin d'asseoir leur fiabilité et leur validité (Morgan-Short et al., 2018). La pratique des réplications est en effet essentielle à l'avancée incrémentale des connaissances scientifiques (Szucs & Ioannidis, 2017). Nous pouvons ajouter ici un résultat plus périphérique, observé auparavant chez des adolescents anglophones, à savoir l'absence apparente de corrélation entre capacités de discrimination phonémique et prosodique, que nous retrouvons également avec notre population d'étudiants en anglais L2. Ces deux résultats demandent bien sûr à être encore confirmés par d'autres études.

Enfin, nous avons trouvé un résultat nouveau, à savoir une forte corrélation entre connaissances phraséologiques et compréhension de l'oral (constatée auparavant avec la compréhension de l'écrit par Kremmel et al., 2015). Ce rôle prépondérant a été confirmé par une étude de régression logistique dans laquelle le seul prédicteur significatif du niveau de compréhension de l'oral était encore une fois les connaissances phraséologiques. Nous avons interprété ce résultat comme montrant l'importance des processus de traitement de niveau supérieur (impliquant le sens) dans la compréhension, et en particulier l'intégration du sens des différents éléments lexicaux reconnus. Il montre également à quel point il est important d'inclure une mesure des connaissances phraséologiques lors de l'évaluation des connaissances linguistiques en général et lexicales en particulier. Une évaluation des connaissances lexicales focalisée uniquement sur les mots isolés aura tendance à surestimer

les connaissances des apprenants, qui ne reconnaissent pas forcément le sens d'expressions non compositionnelles fabriquées à partir de mots très courants.

Perspectives pour l'évaluation et l'enseignement de la compréhension de l'oral

Nous avons conçu cinq tests diagnostiques qui sont à la disposition de la communauté des enseignants d'anglais langue étrangère. Nous avons montré, comme nous venons de le rappeler, que ces tests fonctionnaient bien, mais nous avons également formulé des propositions d'amélioration pour certains d'entre eux, que nous résumerons ici. Une première proposition tient à la prise en compte du temps de réponse à chacun des items, qui peut permettre d'évaluer l'automatisation des connaissances linguistiques des apprenants (elle peut également renseigner sur les stratégies de réponse, par exemple sur les hésitations devant les items potentiellement inconnus dans le test de reconnaissance du lexique). Nous avions délibérément exclu cette prise en compte pour des raisons techniques, le temps de réponse étant difficile à évaluer de façon fiable à distance, et quand les sujets utilisent tous des systèmes informatiques différents pour passer les tests. Cependant, c'est une idée qui reste intéressante et des solutions techniques sont peut-être à creuser.

Une deuxième amélioration possible consisterait à ajouter de nouveaux items à certains de nos tests. Cela permettrait d'affiner le diagnostic en subdivisant certaines habiletés et donc en permettant une identification plus précise des difficultés potentielles de nos étudiants. Nous avons par exemple proposé d'ajouter à notre test de sensibilité prosodique des items avec bruit de fond (qui reproduiraient mieux les conditions authentiques d'écoute), ou à notre test de discrimination phonémique des items testant la connaissance des correspondances graphèmes-phonèmes en anglais. Ce sont des pistes intéressantes, même si cela rallongerait mécaniquement la durée de passation des tests diagnostiques, diminuant ainsi leur praticité.

Pour ce qui est des perspectives pédagogiques, nous avons consacré le chapitre 11 à l'exposition des grandes lignes d'un dispositif de remédiation dont les principes de conception découlent de nos résultats antérieurs : focalisation sur les composantes du processus de compréhension de l'oral (en particulier les connaissances lexicales), importance du *feedback* (dont la forme précise reste à imaginer), personnalisation. A l'intérieur de ce dispositif, nous avons détaillé la conception du parcours de remédiation lexicale, qui tente de conjuguer exposition répétée, décontextualisée et contextualisée, aux items lexicaux, et engagement cognitif. Ce parcours a été réalisé et piloté avec quelques étudiants, montrant ainsi sa

faisabilité et son intérêt, mais il aurait besoin de faire l'objet d'une expérimentation qui pourrait confirmer son efficacité auprès de notre public cible, à savoir les étudiants francophones étudiant l'anglais L2 ayant un niveau insuffisant (B1 ou moins) en compréhension aurale.

Bibliographie

Adamczewski, H., & Delmas, C. (1982). *Grammaire linguistique de l'anglais*. Armand Colin.

Afflerbach, P., Pearson, P. D., & Paris, S. G. (2008). Clarifying Differences between Reading Skills and Reading Strategies. *Reading Teacher*, *61*(5), 364-373. https://doi.org/10.1598/RT.61.5.1

Aitchison, J. (1987). *Words in the Mind : An Introduction to the Mental Lexicon*. John Wiley & Sons.

Alderson, J. C. (2005). *Diagnosing Foreign Language Proficiency : The Interface between Learning and Assessment*. Continuum.

Alderson, J. C., Haapakangas, E.-L., Huhta, A., Nieminen, L., & Ullakonoja, R. (2015). *The Diagnosis of Reading in a Second or Foreign Language*. Routledge.

Alderson, J. C., & Huhta, A. (2011). Can research into the diagnostic testing of reading in a second or foreign language contribute to SLA research? *EUROSLA Yearbook*, *11*(1), 30-52. https://doi.org/10.1075/eurosla.11.04ald

Allan, D. (2004). *Oxford Placement Tests 1 : Test Pack*. OUP Oxford.

Allopenna, P. D., Magnuson, J. S., & Tanenhaus, M. K. (1998). Tracking the Time Course of Spoken Word Recognition Using Eye Movements : Evidence for Continuous Mapping Models. *Journal of Memory and Language*, *38*(4), 419-439. https://doi.org/10.1006/jmla.1997.2558

Altenberg, E. P. (2005). The perception of word boundaries in a second language. *Second Language Research*, *21*(4), 325-358. https://doi.org/10.1191/0267658305sr250oa

Altmann, G. T., & Kamide, Y. (1999). Incremental interpretation at verbs : Restricting the domain of subsequent reference. *Cognition*, *73*(3), 247-264.

Altmann, H. (2006). *The perception and production of second language stress [electronic resource] : A cross-linguistic experimental study /*. University of Delaware.

Anderson, J. R. (1995). *Cognitive psychology and its implications* (Vol. 1-4th Revised edition). Worth Publishers.

Anderson, J. R., & Schunn, C. D. (2000). Implications of the ACT-R Learning Theory : No Magic Bullets. In R. Glaser (Éd.), *Advances in Instructional Psychology, Volume 5 : Educational Design and Cognitive Science* (Vol. 5, p. 1-34). Routledge. http://d-scholarship.pitt.edu/22831/

Anderson, R. C., & Nagy, W. E. (1993). *The Vocabulary Conundrum. Technical Report No. 570* (Technical Report N° 570). University of Illinois at Urbana Champaign. https://eric.ed.gov/?id=ED354489

Andringa, S., Olsthoorn, N., van Beuningen, C., Schoonen, R., & Hulstijn, J. (2012). Determinants of Success in Native and Non-Native Listening Comprehension : An Individual Differences Approach. *Language Learning*, *62*, 49-78. https://doi.org/10.1111/j.1467-9922.2012.00706.x

Arciuli, J. (2018). Reading as Statistical Learning. *Language, Speech, and Hearing Services in Schools*, *49*(3S), 634-643. https://doi.org/10.1044/2018_LSHSS-STLT1-17-0135

Arnaud, P. J. L., Béjoint, H., & Thoiron, P. (1985). A quoi sert le programme lexical? *Les Langues Modernes*, *3/4*, 72-85.

Arnold, J. E., Kam, C. L. H., & Tanenhaus, M. K. (2007). If you say thee uh you are describing something hard : The on-line attribution of disfluency during reference comprehension. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *33*(5), 914-930. https://doi.org/10.1037/0278-7393.33.5.914

Arnold, J. E., Tanenhaus, M. K., Altmann, R. J., & Fagnano, M. (2004). The old and thee, uh, new : Disfluency and reference resolution. *Psychological Science*, *15*(9), 578-582. https://doi.org/10.1111/j.0956-7976.2004.00723.x

Association of Language Testers in Europe (Éd.). (1998). *Multilingual glossary of language testing terms*. Cambridge Univ. Press.

Azevedo, R., & Bernard, R. M. (1995). A Meta-Analysis of the Effects of Feedback in Computer-Based Instruction. *Journal of Educational Computing Research*, *13*(2), 111-127. https://doi.org/10.2190/9LMD-3U28-3A0G-FTQT

Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1995). *CELEX2 LDC96L14 Web Download*. Linguistic Data Consortium.

Bachman, L. F. (1990). *Fundamental considerations in language testing* (Vol. 1-1). Oxford university press.

Bachman, L. F. (2004). *Statistical Analyses for Language Assessment*. Cambridge University Press.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice : Designing and developing useful language tests* (Vol. 1-1).

Baese-Berk, M. M., & Samuel, A. G. (2016). Listeners beware : Speech production may be bad for learning speech sounds. *Journal of Memory and Language*, *89*, 23-36. https://doi.org/10.1016/j.jml.2015.10.008

Barcroft, J., & Sommers, M. S. (2005). Effects of Acoustic Variability on Second Language Vocabulary Learning. *Studies in Second Language Acquisition*, *27*(03). https://doi.org/10.1017/S0272263105050175

Bard, E. G., Shillcock, R. C., & Altmann, G. T. M. (1988). The recognition of words after their acoustic offsets in spontaneous speech : Effects of subsequent context. *Perception & Psychophysics*, *44*(5), 395-408. https://doi.org/10.3758/BF03210424

Barker, L., Gladney, K., Edwards, R., Holley, F., & Gaines, C. (1980). An investigation of proportional time spent in various communication activities by college students. *Journal of Applied Communication Research*, *8*(2), 101-109. https://doi.org/10.1080/00909888009360275

Bauer, L., & Nation, P. (1993). Word Families. *International Journal of Lexicography*, *6*(4), 253-279. https://doi.org/10.1093/ijl/6.4.253

Beeckmans, R., Eyckmans, J., Janssens, V., Dufranne, M., & Velde, H. V. de. (2001). Examining the Yes/No vocabulary test : Some methodological issues in theory and practice. *Language Testing*, *18*(3), 235-274. https://doi.org/10.1177/026553220101800301

Beglar, D., & Hunt, A. (1999). Revising and validating the 2000 Word Level and University Word Level Vocabulary Tests. *Language Testing*, *16*(2), 131-162. https://doi.org/10.1177/026553229901600202

Bell, P., Trofimovich, P., & Collins, L. (2015). Kick the Ball or Kicked the Ball? : Perception of the Past Morpheme –ed by Second Language Learners. *The Canadian Modern Language Review / La revue canadienne des langues vivantes*, *71*(1), 26-51.

Benigno, V., & de Jong, J. (2019). Linking vocabulary to the CEFR and the Global Scale of English : A psychometric model. In A. Huhta, G. Erickson, & N. Figueras, *Developments in language education : A memorial volume in honour of Sauli Takala* (p. 8-29).

Bennett, R. E. (2011). Formative assessment : A critical review. *Assessment in Education: Principles, Policy & Practice*, *18*(1), 5-25. https://doi.org/10.1080/0969594X.2010.513678

Best, C. T. (1995). A direct realist perspective on cross-language speech perception. In Winifred Strange & Workshop on Cross-Language Speech Perception (Éds.), *Speech perception and linguistic experience : Issues in cross-language research* (p. 167-200). York Press.

Bialystok, E. (1979). Explicit and Implicit Judgements of L2 Grammaticality. *Language Learning*, *29*(1), 81-103. https://doi.org/10.1111/j.1467-1770.1979.tb01053.x

Bianco, M. (2016). *Du langage oral à la compréhension de l'écrit*. PUG.

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman Grammar of Spoken and Written English*. Longman.

Bishop, H. (2004). The effect of typographic salience on the look up and comprehension of unknown formulaic sequences. In N. Schmitt (Éd.), *Language Learning & Language Teaching* (Vol. 9, p. 227-248). John Benjamins Publishing Company. https://doi.org/10.1075/lllt.9.12bis

Boersma, P., & Heuven, V. van. (2001). Speak and unSpeak with PRAAT. *Glot International*, *5*(9/10), 341-347.

Bogaards, P. (1994). *Le vocabulaire dans l'apprentissage des langues étrangères*. Hatier.

Bosse, M.-L., & Zagar, D. (2016). La conscience phonémique en maternelle : État des connaissances et proposition d'évolution des pratiques pédagogiques actuelles. *Approche Neuropsychologique des Apprentissages chez l'Enfant A.N.A.E*, *139*(27), 573-582.

Bradlow, A. R. (2007). *Information Flow and Plasticity across Levels of Linguistic Sound Structure : Responses to the Target Papers by Cutler & Weber and by Goldinger*. ICPhS XVI, Saarbrücken.

Bradlow, A. R., & Pisoni, D. B. (1999). Recognition of spoken words by native and non-native listeners : Talker-, listener-, and item-related factors. *The Journal of the Acoustical Society of America*, *106*(4 Pt 1), 2074-2085. https://doi.org/10.1121/1.427952

Brannen, K. (2011). *The Perception and Production of Interdental Fricatives in Second Language Acquisition*. McGill Univeristy.

Bransford, J., National Research Council (U.S.), & National Research Council (U.S.) (Éds.). (2000). *How people learn : Brain, mind, experience, and school* (Expanded ed). National Academy Press.

Brezina, V., & Gablasova, D. (2015). Is There a Core General Vocabulary? Introducing the New General Service List. *Applied Linguistics*, *36*(1), 1-22. https://doi.org/10.1093/applin/amt018

Broersma, M. (2012). Increased lexical activation and reduced competition in second-language listening. *Language and Cognitive Processes*, *27*(7-8). https://doi.org/10.1080/01690965.2012.660170

Broersma, M., & Cutler, A. (2008). Phantom word activation in L2. *System*, *36*(1), 22-34. https://doi.org/10.1016/j.system.2007.11.003

Brown, A. (1988). Functional Load and the Teaching of Pronunciation. *TESOL Quarterly*, *22*(4), 593-606. https://doi.org/10.2307/3587258

Brown, D. (2017). Coverage-based Frequency Bands : A Proposal. *Vocabulary Learning and Instruction*, *6*(2), 52-60.

Brown, J. D. (1996). *Testing in Language Programs*. Prentice Hall Regents.

Brown, Roger. (1973). *A first language : The early stages*. Harvard University Press. http://public.eblib.com/choice/publicfullrecord.aspx?p=3300136

Brown, Ronan, Waring, R., & Donkaewbua, S. (2008). Incidental Vocabulary Acquisition from Reading, Reading-While-Listening, and Listening to Stories. *Reading in a Foreign Language*, *20*(2), 136-163.

Browne, C. (2013). The New General Service List : Celebrating 60 years of vocabulary learning. *The Language Teacher*, *34*(7), 13-15.

Brunfaut, T., & Révész, A. (2015). The Role of Task and Listener Characteristics in Second Language Listening. *TESOL Quarterly*, *49*(1), 141-168. https://doi.org/10.1002/tesq.168

Brysbaert, M., Mandera, P., McCormick, S. F., & Keuleers, E. (2019). Word prevalence norms for 62,000 English lemmas. *Behavior Research Methods*, *51*(2), 467-479. https://doi.org/10.3758/s13428-018-1077-9

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis : A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, *41*(4), 977-990. https://doi.org/10.3758/BRM.41.4.977

Brysbaert, M., Stevens, M., Mandera, P., & Keuleers, E. (2016). How Many Words Do We Know? Practical Estimates of Vocabulary Size Dependent on Word Definition, the Degree of Language Input and the Participant's Age. *Frontiers in Psychology*, *7*. https://doi.org/10.3389/fpsyg.2016.01116

Buck, G. (2001). *Assessing Listening*. Cambridge University Press.

Bundgaard-Nielsen, R. L., Best, C. T., & Tyler, M. D. (2011). Vocabulary Size Is Associated With Second-Language Vowel Perception Performance In Adult Learners. *Studies in Second Language Acquisition*, *33*(3), 433-461. https://doi.org/10.1017/S0272263111000040

Bürki, A., Welby, P., Clément, M., & Spinelli, E. (2019). Orthography and second language word learning : Moving beyond "friend or foe?" *The Journal of the Acoustical Society of America*, *145*(4), EL265-EL271. https://doi.org/10.1121/1.5094923

Capliez, M. (2016). *Acquisition and learning of English phonology by French speakers : On the roles of segments and suprasegments* [Lille 3]. http://www.theses.fr/2016LIL30011

Carroll, J. B. (1961). Fundamental Considerations in Testing English Proficiency of Foreign Students. In *Testing the English Proficiency of Foreign Students*. Center for Applied Linguistics.

Casasanto, L. S. (2008). Does Social Information Influence Sentence Processing? *Proceedings of the Annual Meeting of the Cognitive Science Society*, *30*, 7.

Celce-Murcia, M., Dornyei, Z., & Thurrell, S. (1995). Communicative Competence : A Pedagogically Motivated Model with Content Specifications. *Issues in Applied Linguistics*, *6*(2). https://escholarship.org/uc/item/2928w4zj

Cervini, C., Masperi, M., Jouannaud, M.-P., & Scanu, F. (2013). Defining, modeling and piloting SELF, a new formative assessment test for foreign languages. In J. Colpaert, M. Simons, A. Aerts, & M. Oberhofer (Éds.), *Language Testing in Europe : Time for a new framework?*

Chang, A. C.-S. (2007). The impact of vocabulary preparation on L2 listening comprehension, confidence and strategy use. *System*, *35*(4), 534-550. https://doi.org/10.1016/j.system.2007.06.003

Chang, A. C.-S., & Read, J. (2006). The Effects of Listening Support on the Listening Performance of EFL Learners. *TESOL Quarterly*, *40*(2), 375-397. https://doi.org/10.2307/40264527

Chase, W. G., & Simon, H. A. (1973). Perception in chess. *Cognitive Psychology*, *4*(1), 55-81. https://doi.org/10.1016/0010-0285(73)90004-2

Chater, N., McCauley, S. M., & Christiansen, M. H. (2016). Language as skill : Intertwining comprehension and production. *Journal of Memory and Language*, *89*, 244-254. https://doi.org/10.1016/j.jml.2015.11.004

Chaudron, C. (1988). *Second Language Classrooms : Research on Teaching and Learning*. Cambridge University Press.

Chevillet, F. (1994). *Histoire de la langue anglaise*. Presses universitaires de France.

Cho, T. (2005). Prosodic strengthening and featural enhancement : Evidence from acoustic and articulatory realizations of /ɑ,i/ in English. *The Journal of the Acoustical Society of America*, *117*(6), 3867-3878. https://doi.org/10.1121/1.1861893

Chomsky, N. (1957). *Syntactic Structures*.

Christiansen, M. H., Allen, J., & Seidenberg, M. S. (1998). Learning to Segment Speech Using Multiple Cues : A Connectionist Model. *Language and Cognitive Processes*, *13*(2-3), 221-268. https://doi.org/10.1080/016909698386528

Christiansen, M. H., & Chater, N. (2016). The Now-or-Never bottleneck : A fundamental constraint on language. *Behavioral and Brain Sciences*, *39*. https://doi.org/10.1017/S0140525X1500031X

Christophe, A., Pallier, C., Bertoncini, J., & Mehler, J. (1991). A la recherche d'une unité : Segmentation et traitement de la parole. *L'Année psychologique*, *91*(1), 59-86. https://doi.org/10.3406/psy.1991.29445

Christophe, A., Peperkamp, S., Pallier, C., Block, E., & Mehler, J. (2004). Phonological Phrase Boundaries Constrain Lexical Access I. Adult Data. *Journal of Memory and Language*, *51*(4), 523-547. https://doi.org/10.1016/j.jml.2004.07.001

Chujo, K., & Utiyama, M. (2005). Understanding the Role of Text Length, Sample Size and Vocabulary Size in Determining Text Coverage. *Reading in a Foreign Language*, *17*(1), 1-22.

Chung, K. K. H. (2003). Effects of Pinyin and First Language Words in Learning of Chinese Characters as a Second Language. *Journal of Behavioral Education*, *12*(3), 207-223. https://doi.org/10.1023/A:1025560327860

Church, K. W. (1987). Phonological parsing and lexical retrieval. *Cognition*, *25*(1), 53-69. https://doi.org/10.1016/0010-0277(87)90004-7

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *The Behavioral and Brain Sciences*, *36*(3), 181-204. https://doi.org/10.1017/S0140525X12000477

Clopper, C. G. (2002). Frequency of Stress Patterns in English : A Computational Analysis. *IULC Working Papers*, *2*(1). https://www.indiana.edu/~iulcwp/wp/article/view/02-02

Cobb, T. (s. d.). *Compleat Lexical Tutor v.8.3*. Consulté 9 décembre 2018, à l'adresse https://www.lextutor.ca/

Cobb, T. (2000). One Size Fits All? Francophone Learners and English Vocabulary Tests. *Canadian Modern Language Review*, *57*(2), 295-324. https://doi.org/10.3138/cmlr.57.2.295

Cobb, T., & Horst, M. (2011). Does Word Coach Coach Words? *CALICO Journal*, *28*, 639-661.

Cohen, A. D. (1998). Strategies and processes in test-taking and SLA. In L. F. Bachman & A. D. Cohen (Éds.), *Interfaces between second language acquisition and language testing research* (p. 90-111). Cambridge University Press.

Cohen, J. (1992). Statistical Power Analysis. *Current Directions in Psychological Science*, *1*(3), 98-101. https://doi.org/10.1111/1467-8721.ep10768783

Coleman, J. (2003). Discovering the acoustic correlates of phonological contrasts. *Journal of Phonetics*, *31*(3), 351-372. https://doi.org/10.1016/j.wocn.2003.10.001

Commission Européenne. (2012). *First European Survey on Language Competences : Final report* (p. 244). Luxembourg: Publications Office of the European Union.

Conseil de l'Europe. (2001). *Cadre européen commun de référence pour les langues : Apprendre, enseigner, évaluer* (01 éd.). Editions Didier.

Cooper, F. S., Delattre, P. C., Liberman, A. M., Borst, J. M., & Gerstman, L. J. (1952). Some Experiments on the Perception of Synthetic Speech Sounds. *The Journal of the Acoustical Society of America*, *24*(6), 597-606. https://doi.org/10.1121/1.1906940

Cooper, N., Cutler, A., & Wales, R. (2002). Constraints of lexical stress on lexical access in English : Evidence from native and non-native listeners. *Language and Speech*, *45*(Pt 3), 207-228. https://doi.org/10.1177/00238309020450030101

Corder, S. P. (1967). The Significance of Learner's Errors. *IRAL - International Review of Applied Linguistics in Language Teaching*, *5*(1-4), 161–170. https://doi.org/10.1515/iral.1967.5.1-4.161

Cowan, N. (2010). The Magical Mystery Four : How is Working Memory Capacity Limited, and Why? *Current directions in psychological science*, *19*(1), 51-57. https://doi.org/10.1177/0963721409359277

Coxhead, A. (2000). A New Academic Word List. *TESOL Quarterly*, *34*(2), 213-238. https://doi.org/10.2307/3587951

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297-334. https://doi.org/10.1007/BF02310555

Crossley, S. A., Allen, D., & McNamara, D. S. (2012). Text simplification and comprehensible input : A case for an intuitive approach. *Language Teaching Research*, *16*(1), 89-108. https://doi.org/10.1177/1362168811423456

Cutler, A. (2005, septembre). The lexical statistics of word recognition problems caused by L2 phonetic confusion. *Proceedings of INTERSPEECH 2005*. Interspeech 2005, Lisbon.

Cutler, A., & Carter, D. M. (1987). The predominance of strong initial syllables in the English vocabulary. *Computer Speech & Language*, *2*(3), 133-142. https://doi.org/10.1016/0885-2308(87)90004-0

Cutler, A., & Clifton, C. Jr. (1999). Comprehending spoken language : A blueprint of the listener. In C. M. Brown & P. Hagoort (Éds.), *The neurocognition of language* (p. 123-166). Oxford University Press.

Cutler, A., Dahan, D., & Van Donselaar, W. (1997). Prosody in the Comprehension of Spoken Language : A Literature Review. *Language and Speech*, *40*(2), 141-201.

Cutler, A., & Norris, D. (1988). The role of strong syllables in segmentation for lexical access. *Journal of Experimental Psychology: Human Perception and Performance*, *14*(1), 113-121. https://doi.org/10.1037/0096-1523.14.1.113

Dahan, D., & Magnuson, J. S. (2006). Chapter 8—Spoken Word Recognition. In M. J. T. A. Gernsbacher (Éd.), *Handbook of Psycholinguistics (Second Edition)* (p. 249-283). Academic Press. http://www.sciencedirect.com/science/article/pii/B9780123693747500092

Dahan, D., Magnuson, J. S., Tanenhaus, M. K., & Hogan, E. M. (2001). Subcategorical mismatches and the time course of lexical access : Evidence for lexical competition. *Language and Cognitive Processes*, *16*(5-6), 507-534. https://doi.org/10.1080/01690960143000074

Dang, T. N. Y., Coxhead, A., & Webb, S. (2017). The Academic Spoken Word List. *Language Learning*, *67*(4), 959-997. https://doi.org/10.1111/lang.12253

Dang, T. N. Y., & Webb, S. (2016). Making an essential word list for beginners. In I. S. P. Nation (Éd.), *Making and Using Word Lists for Language Learning and Testing*. John Benjamins Publishing Company.

D'Anna, C. A., Zechmeister, E. B., & Hall, J. W. (1991). Toward a Meaningful Definition of Vocabulary Size: *Journal of Reading Behavior*. https://doi.org/10.1080/10862969109547729

Darcy, I., Daidone, D., & Kojima, C. (2013). Asymmetric lexical access and fuzzy lexical representations in second language learners. *The Mental Lexicon*, *8*(3), 372-420. https://doi.org/10.1075/ml.8.3.06dar

Dauer, R. (1993). *Accurate English* (First Printing edition). Prentice.

Davies, A., Brown, A., Elder, C., Hill, K., Lumley, T., & McNamara, T. (1999). *Dictionary of Language Testing*. Cambridge University Press.

Davies, M. (2009). The 385+ million word Corpus of Contemporary American English (1990–2008+) : Design, architecture, and linguistic insights. *International journal of corpus linguistics*, *14*(2), 159–190.

Davies, M., & Gardner, D. (2013). *A Frequency Dictionary of Contemporary American English : Word Sketches, Collocates and Thematic Lists*. Routledge.

Davies, M., & Kim, J.-B. (2019). The advantages and challenges of "big data" : Insights from the 14 billion word iWeb corpus. *Linguistic Research*, *36*(1), 1-34. https://doi.org/10.17250/khisli.36.1.201903.001

Davis, M. H., Marslen-Wilson, W. D., & Gaskell, M. G. (2002). Leading up the lexical garden path : Segmentation and ambiguity in spoken word recognition. *Journal of Experimental Psychology: Human Perception and Performance*, *28*(1), 218-244. https://doi.org/10.1037/0096-1523.28.1.218

Davis, S. M., & Kelly, M. H. (1997). Knowledge of the English Noun–Verb Stress Difference by Native and Nonnative Speakers. *Journal of Memory and Language*, *36*(3), 445-460. https://doi.org/10.1006/jmla.1996.2503

De Jong, J., & Zheng, Y. (2016). Linking to the CEFR : Validation using a priori and a posteriori evidence. In J. Banerjee & D. Tsagari (Éds.), *Contemporary Second Language Assessment* (Vol. 4, p. 83-100). Bloomsbury Academic. https://eprints.soton.ac.uk/396223/

De Wilde, V., & Eyckmans, J. (2017, août). *Children's incidental knowledge of English before receiving formal instruction*. EUROSLA 2017, Reading, UK.

DeKeyser, R. (2014). Skill Acquisition Theory. In B. VanPatten & J. Williams (Éds.), *Theories in Second Language Acquisition* (p. 106-124). Routledge. https://doi.org/10.4324/9780203628942-11

Delattre, P. (1965). *Comparing the phonetic features of English, French, German and Spanish : An interim report*. Harrap.

DeLong, K. A., Urbach, T. P., & Kutas, M. (2005). Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. *Nature Neuroscience*, *8*(8), 1117-1121. https://doi.org/10.1038/nn1504

Demaizière, F. (2008). Le dispositif, un incontournable du moment. *Alsic. Apprentissage des Langues et Systèmes d'Information et de Communication*, *Vol. 11, n° 2*, Article Vol. 11, n° 2. http://journals.openedition.org/alsic/384

Dennison, H., & Schafer, A. (2010). Online construction of implicature through contrastive prosody. *Proceedings of Speech prosody 2010 conference*. Speech Prosody 2010. http://speechprosody2010.illinois.edu/papers/100338.pdf

DEPP. (2017). *CEDRE 2004-2010-2016 : Compétences en langues des élèves en fin de collège* (N° 17-20). Direction de l'évaluation, de la prospective et de la performance.

Di Cristo, A. (2013). *La prosodie de la parole*. Solal Editeurs.

Díaz, B., Mitterer, H., Broersma, M., & Sebastián-Gallés, N. (2012). Individual differences in late bilinguals' L2 phonological processes : From acoustic-phonetic analysis to lexical

access. *Learning and Individual Differences*, *22*(6), 680-689. https://doi.org/10.1016/j.lindif.2012.05.005

Diependaele, K., Lemhöfer, K., & Brysbaert, M. (2013). The word frequency effect in first- and second-language word recognition : A lexical entrenchment account. *Quarterly Journal of Experimental Psychology*, *66*(5), 843-863. https://doi.org/10.1080/17470218.2012.720994

Dilley, L. C., & McAuley, J. D. (2008). Distal prosodic context affects word segmentation and lexical processing. *Journal of Memory and Language*, *59*(3), 294-311. https://doi.org/10.1016/j.jml.2008.06.006

Dryer, M. S., & Haspelmath, M. (Éds.). (2013). *World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology. https://wals.info/

Duchet, J.-L., & Paillard, M. (1985). Les blocages de compréhension aurale : Phonétique, lexique ou grammaire? *Les Langues Modernes*, *3/4*.

Dulay, H. C., & Burt, M. K. (1974). Natural Sequences in Child Second Language Acquisition. *Language Learning*, *24*(1), 37-53. https://doi.org/10.1111/j.1467-1770.1974.tb00234.x

Dupoux, E., Pallier, C., Sebastian, N., & Mehler, J. (1997). A Destressing "Deafness" in French? *Journal of Memory and Language*, *36*(3), 406-421. https://doi.org/10.1006/jmla.1996.2500

Dupoux, E., Peperkamp, S., & Sebastián-Gallés, N. (2010). Limits on bilingualism revisited : Stress 'deafness' in simultaneous French–Spanish bilinguals. *Cognition*, *114*(2), 266-275. https://doi.org/10.1016/j.cognition.2009.10.001

Dupoux, E., Sebastián-Gallés, N., Navarrete, E., & Peperkamp, S. (2008). Persistent stress 'deafness' : The case of French learners of Spanish. *Cognition*, *106*(2), 682-706. https://doi.org/10.1016/j.cognition.2007.04.001

Durand, J. (2005). La phonétique classique : L'Association Phonétique Internationale et son alphabet. In N. Nguyen, S. Wauquier-Gravelines, & J. Durand, *Phonologie et phonétique : Forme et substance* (p. 25-59).

Elder, C. (2009). Validating a test of metalinguistic knowledge. In R. Ellis (Éd.), *Implicit and Explicit Knowledge in Second Language Learning, Testing and Teaching*. Multilingual Matters.

Elgort, I. (2012). Effects of L1 definitions and cognate status of test items on the Vocabulary Size Test. *Language Testing*, 0265532212459028. https://doi.org/10.1177/0265532212459028

Ellis, N. C. (2002). Reflections on Frequency Effects in Language Processing. *Studies in Second Language Acquisition*, *24*(2), 297-339. https://doi.org/10.1017/S0272263102002140

Ellis, N. C. (2003). Constructions, Chunking, and Connectionism : The Emergence of Second Language Structure. In C. J. Doughty & M. H. Long (Éds.), *The Handbook of Second Language Acquisition*. Blackwell Publishing Ltd. https://doi.org/10.1111/b.9781405132817.2005.x

Ellis, N. C. (2006). Language Acquisition as Rational Contingency Learning. *Applied Linguistics*, *27*(1), 1-24. https://doi.org/10.1093/applin/ami038

Ellis, N. C. (2008). Usage-Based and Form-Focused Language Acquisition : The associative learning of constructions, learned attention, and the limited L2 endstate. In P. Robinson & N. Ellis, *Handbook of Cognitive Linguistics and Second Language Acquisition* (p. 372-405). Routledge.

Ellis, N. C. (2009). Optimizing the Input : Frequency and Sampling in Usage-Based and Form-Focused Learning. In *The Handbook of Language Teaching* (p. 139-158). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781444315783.ch9

Ellis, R. (1991). Grammatically Judgments and Second Language Acquisition. *Studies in Second Language Acquisition*, *13*(2), 161-186. https://doi.org/10.1017/S0272263100009931

Ellis, R. (2005). Measuring Implicit and Explicit Knowledge of a Second Language : A Psychometric Study. *Studies in Second Language Acquisition*, *27*(2), 141-172. https://doi.org/10.1017/S0272263105050096

Elman, J. L. (2004). An alternative view of the mental lexicon. *Trends in Cognitive Sciences*, *8*(7), 301-306. https://doi.org/10.1016/j.tics.2004.05.003

English Profile. (2011). *Introducing the CEFR for Englsih*. https://www.englishprofile.org/resources/information-booklet

English Profile. (2012). *English Vocabulary Profile*. http://vocabulary.englishprofile.org/staticfiles/about.html

English Profile. (2015). *English Grammar Profile*. http://www.englishprofile.org/english-grammar-profile

Ericsson, K. A., Chase, W. G., & Faloon, S. (1980). Acquisition of a memory skill. *Science (New York, N.Y.)*, *208*(4448), 1181-1182. https://doi.org/10.1126/science.7375930

Ericsson, K. A., & Kintsch, W. (1995). Long-term working memory. *Psychological Review*, *102*(2), 211-245. https://doi.org/10.1037/0033-295x.102.2.211

Erman, B., & Warren, B. (2000). The idiom principle and the open choice principle. *Text - Interdisciplinary Journal for the Study of Discourse*, *20*(1), 29–62. https://doi.org/10.1515/text.1.2000.20.1.29

Escudero, P., Simon, E., & Mulak, K. E. (2014). Learning words in a new language : Orthography doesn't always help. *Bilingualism: Language and Cognition*, *17*(02), 384–395. https://doi.org/10.1017/S1366728913000436

Escudero, P., & Wanrooij, K. (2010). The effect of L1 orthography on non-native vowel perception. *Language and Speech*, *53*(3), 343–365.

Eyckmans, J. (2004). *Measuring receptive vocabulary size*. Katholieke Universiteit Nijmegen.

Færch, C., & Kasper, G. (1980). Processes and Strategies in Foreign Language Learning and Communication. *Interlanguage Studies Bulletin*, *5*(1), 47-118. JSTOR.

Færch, C., & Kasper, G. (1986). The Role of Comprehension in Second-language Learning. *Applied Linguistics*, *7*(3), 257-274. https://doi.org/10.1093/applin/7.3.257

Falissard, B. (2011). *Analysis of Questionnaire Data with R*. CRC Press.

Faucett, L., Palmer, H. E., Thorndike, E. L., & West, M. (1936). *Interim report on vocabulary selection for the teaching of English as a foreign language*. P. S. King & Son, Ltd.

Ferreira, F., & Patson, N. D. (2007). The 'Good Enough' Approach to Language Comprehension. *Language and Linguistics Compass*, *1*(1-2), 71-83. https://doi.org/10.1111/j.1749-818X.2007.00007.x

Feyten, C. (1989). *Listening Ability : An Overlooked Dimension of Foreign Language Acquisition.* Eastern Educational Research Association, Savannah, Georgia.

Field, A., Miles, J., & Field, Z. (2012). *Discovering Statistics Using R*. SAGE Publications Ltd.

Field, J. (2000). Finding One's Way in the Fog : Listening Strategies and Second-language Learners. *Modern English Teacher*, *9*(1), 29-34.

Field, J. (2004). An Insight into Listeners' Problems : Too Much Bottom-Up or Too Much Top-Down? *System: An International Journal of Educational Technology and Applied Linguistics*, *32*(3), 363-377. https://doi.org/10.1016/j.system.2004.05.002

Field, J. (2008a). *Listening in the Language Classroom*. Cambridge University Press.

Field, J. (2008b). Bricks or Mortar : Which Parts of the Input Does a Second Language Listener Rely on? *TESOL Quarterly*, *42*(3), 411-432. https://doi.org/10.1002/j.1545-7249.2008.tb00139.x

Fine, A. B., Jaeger, T. F., Farmer, T. A., & Qian, T. (2013). Rapid Expectation Adaptation during Syntactic Comprehension. *PLOS ONE*, *8*(10), e77661. https://doi.org/10.1371/journal.pone.0077661

Flege, J. E., & MacKay, I. R. A. (2004). Perceiving Vowels In A Second Language. *Studies in Second Language Acquisition*, *26*(01), 1–34. https://doi.org/10.1017/S0272263104026117

Folse, K. S. (2013). *Vocabulary Myths : Applying Second Language Research to Classroom Teaching*. University of Michigan Press ELT.

Forster, K. (1976). Accessing the mental lexicon. In R. J. Wales & E. L. Walker, *New Approaches to Language Mechanisms : A Collection of Psycholinguistic Studies*. North-Holland Publishing Company.

Fort, M., Spinelli, E., Savariaux, C., & Kandel, S. (2010). The word superiority effect in audiovisual speech perception. *Speech Communication*, *52*(6), 525-532.

Forvo, L. (2008). *Forvo : Le guide de la prononciation. Tous les mots du monde prononcés par des locuteurs natifs*. Forvo.com. https://fr.forvo.com

Fountain, R. L., & Nation, I. S. P. (2000). A Vocabulary- Based Graded Dictation Test. *RELC Journal*, *31*(2), 29-44. https://doi.org/10.1177/003368820003100202

Fox Tree, J. E., & Meijer, P. J. A. (2000). Untrained speakers' use of prosody in syntactic disambiguation and listeners' interpretations. *Psychological Research*, *63*(1), 1-13. https://doi.org/10.1007/PL00008163

Fraser, C., Bellugi, U., & Brown, R. (1963). Control of grammar in imitation, comprehension, and production. *Journal of Verbal Learning and Verbal Behavior*, *2*(2), 121-135. https://doi.org/10.1016/S0022-5371(63)80076-6

Frenck-Mestre, C. (2002). An on-line look at sentence processing in the second language. In R. R. Heredia & J. Altarriba, *Bilingual Sentence Processing* (Vol. 134, p. 217-236). Elsevier. https://doi.org/10.1016/S0166-4115(02)80012-7

Fries, C. C. (1952). *The structure of English; an introduction to the construction of English sentences.* —. New York : Harcourt, Brace. http://archive.org/details/structureofengli0000frie

Frost, D. (2011). Stress and cues to relative prominence in English and French : A perceptual study. *ournal of the International Phonetic Association*.

Frost, D., & Henderson, A. (2013). Résultats du sondage EPTiES (English Pronunciation Teaching in Europe Survey) : L'enseignement de la prononciation dans plusieurs pays européens vu par les enseignants. *Recherche et pratiques pédagogiques en langues de spécialité. Cahiers de l'Apliut*, *Vol. XXXII N° 1*, 92-113. https://doi.org/10.4000/apliut.3586

Fulcher, G. (2010). *Practical Language Testing* (1 edition). Routledge.

Ganong, W. F. (1980). Phonetic categorization in auditory word perception. *Journal of Experimental Psychology. Human Perception and Performance*, *6*(1), 110-125.

Gaonac'h, D. (2005). La question de l'automatisation en langue étrangère. *Revue parole*, *34/35/36*, 221-242.

Gardner, D. (2007). Validating the Construct of Word in Applied Corpus-based Vocabulary Research : A Critical Survey. *Applied Linguistics*, *28*(2), 241-265. https://doi.org/10.1093/applin/amm010

Gardner, D., & Davies, M. (2014). A New Academic Vocabulary List. *Applied Linguistics*, *35*(3), 305-327. https://doi.org/10.1093/applin/amt015

Garnier, M., & Schmitt, N. (2015). The PHaVE List : A pedagogical list of phrasal verbs and their most frequent meaning senses. *Language Teaching Research*, *19*(6), 645-666. https://doi.org/10.1177/1362168814559798

Gass, S. (1983). The Development of L2 Intuitions. *TESOL Quarterly*, *17*(2), 273-291. https://doi.org/10.2307/3586654

Gass, S. M., & Madden, C. G. (1985). *Input in Second Language Acquisition*. Newbury House Publishers, Inc.

Gass, S. M., & Selinker, L. (1983). *Language Transfer in Language Learning. Issues in Second Language Research*. Newbury House Publishers, Inc.

Gilmore, A. (2007). Authentic materials and authenticity in foreign language learning. *Language Teaching*, *40*(2), 97-118. https://doi.org/10.1017/S0261444807004144

Goh, C. C. M. (2000). A Cognitive Perspective on Language Learners' Listening Comprehension Problems. *System*, *28*(1), 55-75.

Goigoux, R. (2016). *Lire et écrire*.

Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review*, *105*(2), 251-279.

Goldinger, S. D., & Azuma, T. (2003). Puzzle-solving science : The quixotic quest for units in speech perception. *Journal of Phonetics*, *31*(3-4), 305-320. https://doi.org/10.1016/S0095-4470(03)00030-5

Goldschneider, J. M., & DeKeyser, R. M. (2001). Explaining the "Natural Order of L2 Morpheme Acquisition" in English : A Meta-analysis of Multiple Determinants. *Language Learning*, *51*(1), 1-50. https://doi.org/10.1111/1467-9922.00147

Goss, N., Ying-Hua, Z., & Lantolf, J. P. (1994). Two heads may be better than one : Mental activity in second language grammaticality judgments. In E. Tarone, S. M. Gass, & A. D. Cohen (Éds.), *Research methodology in second-language acquisition*. L. Erlbaum.

Goswami, U., Gerson, D., & Astruc, L. (2010). Amplitude envelope perception, phonology and prosodic sensitivity in children with developmental dyslexia. *Reading and Writing*, *23*(8), 995-1019. https://doi.org/10.1007/s11145-009-9186-6

Goswami, U., Mead, N., Fosker, T., Huss, M., Barnes, L., & Leong, V. (2013). Impaired perception of syllable stress in children with dyslexia : A longitudinal study. *Journal of Memory and Language*, *69*(1), 1-17. https://doi.org/10.1016/j.jml.2013.03.001

Gougenheim, G., Michéa, R., Rivenc, P., & Sauvageot, A. (1964). *L'élaboration du français fondamental, 1er degré : Étude sur l'établissement d'un vocabulaire et d'une grammaire de base*. Didier.

Goulden, R., Nation, P., & Read, J. (1990). How Large Can a Receptive Vocabulary Be? *Applied Linguistics*, *11*(4), 341-363.

Graham, S., Santos, D., & Vanderplank, R. (2008). Listening comprehension and strategy use : A longitudinal exploration. *System*, *36*(1), 52-68. https://doi.org/10.1016/j.system.2007.11.001

Graham, S., Santos, D., & Vanderplank, R. (2010). Strategy clusters and sources of knowledge in French L2 listening comprehension. *Innovation in Language Learning and Teaching*, *4*(1), 1-20. https://doi.org/10.1080/17501220802385866

Greenberg, S. (1999). Speaking in Shorthand—A Syllable-centric Perspective for Understanding Pronunciation Variation. *Speech Commun.*, *29*(2), 159–176. https://doi.org/10.1016/S0167-6393(99)00050-3

Grosjean, F. (1980). Spoken word recognition processes and the gating paradigm. *Perception & Psychophysics*, *28*(4), 267-283. https://doi.org/10.3758/BF03204386

Grosjean, F. (1985). The recognition of words after their acoustic offset : Evidence and implications. *Perception & Psychophysics*, *38*(4), 299-310.

Guo, Y., & Roehrig, A. D. (2011). Roles of General versus Second Language (L2) Knowledge in L2 Reading Comprehension. *Reading in a Foreign Language*, *23*(1), 42-64.

Gutiérrez, X. (2013). The Construct Validity of Grammaticality Judgment Tests as Measures of Implicit and Explicit Knowledge. *Studies in Second Language Acquisition*, *35*(3), 423-449. https://doi.org/10.1017/S0272263113000041

Gyllstad, H. (2009). Designing and Evaluating Tests of Receptive Collocation Knowledge : COLLEX and COLLMATCH. In A. Barfield & H. Gyllstad (Éds.), *Researching Collocations in Another Language : Multiple Interpretations* (p. 153-170). Palgrave Macmillan UK. https://doi.org/10.1057/9780230245327_12

Gyllstad, H. (2013). Looking at L2 Vocabulary Knowledge Dimensions from an Assessment Perspective – Challenges and Potential Solutions. In C. Bardel, B. Laufer, & C. Lindqvist, *L2 vocabulary acquisition, knowledge and use : New perspectives on assessment and corpus analysis* (p. 11-28).

Hallé, P. A., Best, C. T., & Levitt, A. (1999). Phonetic vs. Phonological influences on French listeners' perception of American English approximants. *Journal of Phonetics*, *27*(3), 281-306. https://doi.org/10.1006/jpho.1999.0097

Hallé, P. A., & Boysson-Bardies, B. de. (1996). The format of representation of recognized words in infants' early receptive lexicon. *Infant Behavior and Development*, *19*(4), 463-481. https://doi.org/10.1016/S0163-6383(96)90007-7

Harley, B., Howard, J., & Hart, D. (1995). Second Language Processing at Different Ages : Do Younger Learners Pay More Attention to Prosodic Cues to Sentence Structure? *Language Learning*, *45*(1), 43-71. https://doi.org/10.1111/j.1467-1770.1995.tb00962.x

Harley, T. A. (2007). *The Psychology of Language : From Data to Theory* (3 edition). Psychology Press.

Hart, B., & Risley, T. R. (2003). The Early Catastrophe. The 30 Million Word Gap. *American Educator*, *27*(1), 4-9.

Hasher, L., & Chromiak, W. (1977). The processing of frequency information : An automatic mechanism? *Journal of Verbal Learning & Verbal Behavior*, *16*(2), 173-184. https://doi.org/10.1016/S0022-5371(77)80045-5

Hasher, L., & Zacks, R. T. (1984). Automatic processing of fundamental information : The case of frequency of occurrence. *The American Psychologist*, *39*(12), 1372-1388.

Haspelmath, M., & Tadmor, U. (2009). *Loanwords in the World's Languages : A Comparative Handbook*. Walter de Gruyter.

Hattie, J. (2009). *Visible Learning : A Synthesis of Over 800 Meta-analyses Relating to Achievement*. Routledge.

Hauser, M. D., Newport, E. L., & Aslin, R. N. (2001). Segmentation of the speech stream in a non-human primate : Statistical learning in cotton-top tamarins. *Cognition*, *78*(3), B53-64.

Hawkins, S. (2004, juin). *Puzzles and patterns in 50 years of research on speech perception*. From Sound to Sense, MIT.

Hayes-Harb, R. (2007). Lexical and statistical evidence in the acquisition of second language phonemes. *Second Language Research*, *23*(1), 65-94. https://doi.org/10.1177/0267658307071601

Higgins, S., & Simpson, A. (2011). Visible Learning : A Synthesis of over 800 Meta-Analyses Relating to Achievement. By John A.C. Hattie. *British Journal of Educational Studies*, *59*(2), 197-201. https://doi.org/10.1080/00071005.2011.584660

Hilton, H. (2003). L'accès au lexique mental dans une langue étrangère : Le cas des francophones apprenant l'anglais. *Corela. Cognition, représentation, langage*, *1-2*. https://doi.org/10.4000/corela.676

Hilton, H. (2005). Théories d'apprentissage et didactique des langues. *Langues modernes*, *99*(3), 12-21.

Hilton, H. (2006). Quelques aspects de la mémoire verbale en L2. *Recherche et pratiques pédagogiques en langues de spécialité. Cahiers de l'Apliut*, *Vol. XXV N° 2*, 44-60. https://doi.org/10.4000/apliut.2478

Hilton, H. (2009). *Systèmes émergents : Acquisition, traitement et didactique des langues* [Habilitation à diriger des recherches].

Hilton, H. (2019). *Sciences cognitives et didactique des langues. Rapport d'expertise pour le Conseil national de l'évaluation du système scolaire.* CNESCO.

Hilton, H., Lenart, E., & Zoghlami, N. (2016). Compréhension et production en anglais L2 à l'école primaire. *Revue francaise de linguistique appliquee*, *Vol. XXI*(2), 65-80.

Hintzman, D. L., Block, R. A., & Inskeep, N. R. (1972). Memory for mode of input. *Journal of Verbal Learning and Verbal Behavior*, *11*(6), 741-749. https://doi.org/10.1016/S0022-5371(72)80008-2

Hopp, H. (2016). Learning (not) to predict : Grammatical gender processing in second language acquisition. *Second Language Research*, *32*(2), 277-307. https://doi.org/10.1177/0267658315624960

Horst, M. (2010). How well does teacher talk support incidental vocabulary acquisition? *Reading in a Foreign Language*, *22*(1), 161-180.

Howell, D. C. (2009). *Statistical Methods for Psychology*. Cengage Learning.

Howes, D. (1957). On the relation between the intelligibility and frequency of occurrence of English words. *Journal of the Acoustical Society of America*, *29*, 296-305. https://doi.org/10.1121/1.1908862

Hu, M. H.-C., & Nation, P. (2000). Unknown Vocabulary Density and Reading Comprehension. *Reading in a Foreign Language*, *13*(1), 403-430.

Huart, R. (2002). *La grammaire orale de l'anglais*. Editions OPHRYS.

Huart, R. (2010). *Nouvelle grammaire de l'anglais oral*. Editions OPHRYS.

Hughes, A. (2002). *Testing for Language Teachers*. Cambridge University Press.

Huhta, A. (2008). Diagnostic and Formative Assessment. In *The Handbook of Educational Linguistics* (p. 469-482). John Wiley & Sons, Ltd. https://doi.org/10.1002/9780470694138.ch33

Hulstijn, J. H. (2012). Incidental Learning in Second Language Acquisition. In C. A. Chapelle, *The Encyclopedia of Applied Linguistics*. American Cancer Society. https://doi.org/10.1002/9781405198431.wbeal0530

Hulstijn, J. H., Gelderen, A. V., & Schoonen, R. (2009). Automatization in second language acquisition : What does the coefficient of variation tell us? *Applied Psycholinguistics*, *30*(4), 555-582. https://doi.org/10.1017/S0142716409990014

Husson, F., Lê, S., & Pagès, J. (2016). *Analyse de données avec R* (2e édition revue et augmentée). Presses universitaires de Rennes.

Hymes, D. (1972). On Communicative Competence. In J. B. Pride & J. Holmes (Éds.), *Sociolinguistics : Selected readings*. Penguin.

Imhof, M. (2008). What Have You Listened to in School Today? *International Journal of Listening*, *22*(1), 1-12. https://doi.org/10.1080/10904010701802121

Improve Your English Pronunciation. (s. d.). Consulté 6 novembre 2020, à l'adresse https://youglish.com

Ito, K., & Strange, W. (2009). Perception of allophonic cues to English word boundaries by Japanese second language learners of English. *The Journal of the Acoustical Society of America*, *125*(4), 2348-2360. https://doi.org/10.1121/1.3082103

Iverson, P., Pinet, M., & Evans, B. G. (2012). Auditory training for experienced and inexperienced second-language learners : Native French speakers learning English vowels. *Applied Psycholinguistics*, *33*(1), 145-160. https://doi.org/10.1017/S0142716411000300

Jacob, B. A., & Levitt, S. D. (2003). *Rotten Apples : An Investigation of the Prevalence and Predictors of Teacher Cheating* (Working Paper N° 9413). National Bureau of Economic Research. https://doi.org/10.3386/w9413

Johnson, K. (2011). *Acoustic and Auditory Phonetics, 3rd Edition*. Wiley-Blackwell.

Johnson, K. (2004). Massive reduction in conversational American English. In K. Yoneyama & K. Maekawa (Éds.), *Spontaneous Speech : Data and Analysis. Proceedings of the 1st Session of the 10th International Symposium* (p. 29-54).

Johnson, V. E. (2013). Revised standards for statistical evidence. *Proceedings of the National Academy of Sciences*, *110*(48), 19313-19317. https://doi.org/10.1073/pnas.1313476110

Jones, G. (2012). Why Chunking Should be Considered as an Explanation for Developmental Change before Short-Term Memory Capacity and Processing Speed. *Frontiers in Psychology*, *3*. https://doi.org/10.3389/fpsyg.2012.00167

Jusczyk, P. W. (1997). *The Discovery of Spoken Language*. https://mitpress.mit.edu/books/discovery-spoken-language

Kamiyama, T., & Nakamura-Delloye, Y. (2015, août). *Native French speakers' perception of the Japanese /h/ : Ha piece hof cake?* (Paper 0871). The 18th International Congress of Phonetic Sciences, Glasgow, Scotland, United Kingdom. http://www.icphs2015.info/

Kay, P., & Kempton, W. (1984). What Is the Sapir-Whorf Hypothesis? *American Anthropologist*, *86*(1), 65-79.

Kintsch, W., & Greene, E. (1978). The role of culture-specific schemata in the comprehension and recall of stories. *Discourse Processes*, *1*(1), 1-13. https://doi.org/10.1080/01638537809544425

Kintsch, W., & Van Dijk, T. A. (1978). Toward a Model of Text Comprehension and Production. *Psychological Review*, *85*(5), 363-394.

Kintsch, W., Welsch, D., Schmalhofer, F., & Zimny, S. (1990). Sentence memory : A theoretical analysis. *Journal of Memory and Language*, *29*(2), 133-159. https://doi.org/10.1016/0749-596X(90)90069-C

Kirkpatrick, E. A. (1891). Number of Words in an Ordinary Vocabulary. *Science (New York, N.Y.)*, *18*(446), 107-108. https://doi.org/10.1126/science.ns-18.446.107-a

Kitzen, K. R. (2001). Prosodic sensitivity, morphological ability, and reading ability in young adults with and without childhood histories of reading difficulty. *APA PsycNET*.

Klatt, D., & Stevens, K. (1973). On the automatic recognition of continuous speech:Implications from a spectrogram-reading experiment. *IEEE Transactions on Audio and Electroacoustics*, *21*(3), 210-217. https://doi.org/10.1109/TAU.1973.1162453

Klingberg, T. (2010). Training and plasticity of working memory. *Trends in Cognitive Sciences*, *14*(7), 317-324. https://doi.org/10.1016/j.tics.2010.05.002

Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance : A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin*, *119*(2), 254-284. https://doi.org/10.1037/0033-2909.119.2.254

Kormos, J., & Sáfár, A. (2008). Phonological short-term memory, working memory and foreign language performance in intensive language learning. *Bilingualism: Language and Cognition*, *11*(02). https://doi.org/10.1017/S1366728908003416

Krashen, S. D. (1981). *Second language acquisition and second language learning* (Reprinted). Pergamon Press.

Kremmel, B. (2016). Word Families and Frequency Bands in Vocabulary Tests : Challenging Conventions. *TESOL Quarterly*, *50*(4), 976-987. https://doi.org/10.1002/tesq.329

Kremmel, B. (2017). *Development and initial validation of a diagnostic computer-adaptive profiler of vocabulary knowledge* [Thesis, University of Nottingham]. http://eprints.nottingham.ac.uk/49085/

Kremmel, B., Brunfaut, T., & Alderson, J. C. (2017). Exploring the Role of Phraseological Knowledge in Foreign Language Reading. *Applied Linguistics*, *38*(6), 848–870. https://doi.org/10.1093/applin/amv070

Krzonowski, J., Ferragne, E., & Pellegrino, F. (2016). Perception et production de voyelles de l'anglais par des apprenants francophones : Effet d'entraînements en perception et en production. *Journées d'Etudes sur la Parole*. https://hal.archives-ouvertes.fr/hal-01485744

Kuhl, P. K. (1991). Human adults and human infants show a « perceptual magnet effect » for the prototypes of speech categories, monkeys do not. *Perception & Psychophysics*, *50*(2), 93-107.

Kuhl, P. K. (1993). Early linguistic experience and phonetic perception : Implications for theories of developmental speech perception. *Journal of Phonetics*, *21*(1-2), 125-139.

Kuhl, P. K., & Iverson, P. (1995). Linguistic experience and the « perceptual magnet effect. » In W. Strange (Éd.), *Speech perception and linguistic experience : Issues in cross-language research* (p. 121-154). York Press. http://discovery.ucl.ac.uk/23733/

Kuhl, P. K., Williams, K. A., Lacerda, F., Stevens, K. N., & Lindblom, B. (1992). Linguistic experience alters phonetic perception in infants by 6 months of age. *Science (New York, N.Y.)*, *255*(5044), 606-608.

Kuperberg, G. R., & Jaeger, T. F. (2016). What do we mean by prediction in language comprehension? *Language, Cognition and Neuroscience*, *31*(1), 32-59. https://doi.org/10.1080/23273798.2015.1102299

Kurumada, C., Brown, M., Bibyk, S., Pontillo, D. F., & Tanenhaus, M. K. (2014). Is it or isn't it : Listeners make rapid use of prosody to infer speaker meanings. *Cognition*, *133*(2), 335-342. https://doi.org/10.1016/j.cognition.2014.05.017

Labov, W. (1972). *Sociolinguistic Patterns*. University of Pennsylvania Press.

Labov, W., Ash, S., & Boberg, C. (2005). *The Atlas of North American English : Phonetics, Phonology and Sound Change* (Pck Har/CD Edition). Mouton de Gruyter.

Langacker, R. W. (1987). *Foundations of Cognitive Grammar : Theoretical prerequisites*. Stanford University Press.

Lapaire, J.-R., & Rotgé, W. (1991). *Linguistique et grammaire de l'anglais* (3e éd). Presses Universitaires du Mirail.

Larreya, P., & Rivière, C. (2010). *Grammaire explicative de l'anglais*. Pearson.

Larreya, P., & Schottman, W. (2013). *A Pronunciation Guide—Bien prononcer l'anglais*. Nathan.

Larson-Hall, J. (2009). *A Guide to Doing Statistics in Second Language Research Using SPSS* (1 edition). Routledge.

Lass, R. (1984). *Phonology : An Introduction to Basic Concepts*. Cambridge University Press.

Laufer, B. (1988a). A Factor of Difficulty in Vocabulary Learning : Deceptive Transparency. *AILA Review*, *6*, 10-20.

Laufer, B. (1988b). The concept of 'synforms' (similar lexical forms) in vocabulary acquisition. *Language and Education*, *2*(2), 113-132. https://doi.org/10.1080/09500788809541228

Laufer, B. (1989). What percentage of text-lexis is essential for comprehension? In Ch. Lauren & M. Nordman (Éds.), *Special Language : From Humans Thinking To Thinking Machines* (p. 316-323). Multilingual Matters.

Laufer, B., & Nation, P. (2001). Passive vocabulary size and speed of meaning recognition : Are they related? *EUROSLA Yearbook*, *1*, 7-28. https://doi.org/10.1075/eurosla.1.05lau

Laufer, B., & Ravenhorst-Kalovski, G. C. (2010). Lexical Threshold Revisited : Lexical Text Coverage, Learners' Vocabulary Size and Reading Comprehension. *Reading in a Foreign Language*, *22*(1), 15-30.

Laufer, B., & Shmueli, K. (1997). Memorizing New Words : Does Teaching Have Anything To Do With It?: *RELC Journal*, *28*(1), 89-108. https://doi.org/10.1177/003368829702800106

Laveault, D. (2012). Soixante ans de bons et mauvais usages du alpha de Cronbach. *Mesure et évaluation en éducation*, *35*(2), 1. https://doi.org/10.7202/1024716ar

Laveault, D., & Grégoire, J., enseignant en psychologie. (2014). *Introduction aux théories des tests en psychologie et en sciences de l'éducation*. Bruxelles : De Boeck. DL 2014.

Lee, D. (2001). *Cognitive Linguistics : An Introduction*. Oxford University Press Australia & New Zealand.

Lee, Y.-W. (2015). Future of diagnostic language assessment. *Language Testing*, *32*(3), 295-298. https://doi.org/10.1177/0265532214565385

Lemhöfer, K., & Broersma, M. (2012). Introducing LexTALE : A quick and valid Lexical Test for Advanced Learners of English. *Behavior Research Methods*, *44*(2), 325-343. https://doi.org/10.3758/s13428-011-0146-0

Lemhöfer, K., Dijkstra, T., Schriefers, H., Baayen, R. H., Grainger, J., & Zwitserlood, P. (2008). Native language influences on word recognition in a second language : A megastudy. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *34*(1), 12-31. https://doi.org/10.1037/0278-7393.34.1.12

Leong, V., Hämäläinen, J., Soltész, F., & Goswami, U. (2011). Rise time perception and detection of syllable stress in adults with developmental dyslexia. *Journal of Memory and Language*, *64*(1), 59-73. https://doi.org/10.1016/j.jml.2010.09.003

Lerner, Y., Honey, C. J., Silbert, L. J., & Hasson, U. (2011). Topographic Mapping of a Hierarchy of Temporal Receptive Windows Using a Narrated Story. *The Journal of Neuroscience*, *31*(8), 2906-2915. https://doi.org/10.1523/JNEUROSCI.3684-10.2011

Levin, B. (1993). *English Verb Classes and Alternations : A Preliminary Investigation*. University of Chicago Press.

Levinson, S. C. (2000). *Presumptive Meanings : The Theory of Generalized Conversational Implicature*. MIT Press.

Levitt, S. D., & Lin, M.-J. (2015). *Catching Cheating Students* (Working Paper N° 21628). National Bureau of Economic Research. https://doi.org/10.3386/w21628

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*(3), 1126-1177. https://doi.org/10.1016/j.cognition.2007.05.006

Lewis, M. (1997). *Implementing the Lexical Approach : Putting Theory into Practice*. Cengage ELT.

Libben, G., & Jarema, G. (Éds.). (2007). *The Representation and Processing of Compound Words*. Oxford University Press. http://www.oxfordscholarship.com/view/10.1093/acprof:oso/9780199228911.001.0001/acprof-9780199228911

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review*, *74*(6), 431-461.

Liberman, A. M., Delattre, P. C., Gerstman, L. J., & Cooper, F. S. (1956). Tempo of frequency change as a cue for distinguishing classes of speech sounds. *Journal of Experimental Psychology*, *52*(2), 127-137.

Lim, J. H., & Christianson, K. (2013). Integrating meaning and structure in L1–L2 and L2–L1 translations. *Second Language Research*, *29*(3), 233-256. https://doi.org/10.1177/0267658312462019

Lin, C.-C., & Yu, Y.-C. (2017). Effects of presentation modes on mobile-assisted vocabulary learning and cognitive load. *Interactive Learning Environments*, *25*(4), 528-542. https://doi.org/10.1080/10494820.2016.1155160

Lindblom, B. (1990). Explaining Phonetic Variation : A Sketch of the H&H Theory. In W. J. Hardcastle & A. Marchal (Éds.), *Speech Production and Speech Modelling* (p. 403-439). Springer Netherlands. https://doi.org/10.1007/978-94-009-2037-8_16

Liu, H. H.-T. (2015). The Conceptualization and Operationalization of Diagnostic Testing in Second and Foreign Language Assessment. *Working Papers in TESOL and Applied Linguistics*, *14*(1), 1-12. https://doi.org/10.7916/D84F23B7

Lively, S. E., Logan, J. S., & Pisoni, D. B. (1993). Training Japanese listeners to identify English /r/ and /l/. II : The role of phonetic environment and talker variability in learning new perceptual categories. *The Journal of the Acoustical Society of America*, *94*(3 Pt 1), 1242-1255. https://doi.org/10.1121/1.408177

Lockwood, J. (2013). The Diagnostic English Language Tracking Assessment (DELTA) writing project : A case for post-entry assessment policies and practices in Hong Kong universities. *Papers in Language Testing and Assessment*, *2*(1).

Loewen, S. (2009). Grammaticality judgment tests and the measurement of implicit and explicit L2 knowledge. In R. Ellis (Éd.), *Implicit and Explicit Knowledge in Second Language Learning, Testing and Teaching*. Multilingual Matters.

Logan, G. D. (1988). Toward an instance theory of automatization. *Psychological Review*, *95*(4), 492-527. https://doi.org/10.1037/0033-295X.95.4.492

Long, D. R. (1989). Second Language Listening Comprehension : A Schema-Theoretic Perspective. *The Modern Language Journal*, *73*(1), 32-40. JSTOR. https://doi.org/10.2307/327265

Long, D. R. (1990). What You Don't Know Can't Help You : An Exploratory Study of Background Knowledge and Second Language Listening Comprehension. *Studies in Second Language Acquisition*, *12*(1), 65-80.

Luce, P. A. (1986a). A computational analysis of uniqueness points in auditory word recognition. *Perception & Psychophysics*, *39*(3), 155-158. https://doi.org/10.3758/BF03212485

Luce, P. A. (1986b). *Neighborhoods of Words in the Mental Lexicon. Research on Speech Perception.* (Technical Report No. 6). https://eric.ed.gov/?id=ED353610

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers*, *28*(2), 203-208. https://doi.org/10.3758/BF03204766

Lynch, T. (2010). Listening : Sources, Skills, and Strategies. In R. B. Kaplan, *The Oxford Handbook of Applied Linguistics*. https://doi.org/10.1093/oxfordhb/9780195384253.013.0005

Macaro, E. (2006). Strategies for Language Learning and for Language Use : Revising the Theoretical Framework. *The Modern Language Journal*, *90*(3), 320-337. https://doi.org/10.1111/j.1540-4781.2006.00425.x

MacWhinney, B. (2005). A unified model of language acquisition. In J. F. Kroll & A. M. B. D. Groot, *Handbook of Bilingualism : Psycholinguistic Approaches*. Oxford University Press.

Maddieson, & Disner, S. F. (1984). *Patterns of Sounds*. Cambridge University Press.

Magnat, E. (2013). *Le TBI comme instument du développement de la conscience phonémique à l'école : Une approche ergonomique* [These de doctorat, Grenoble]. http://www.theses.fr/2013GRENL005

Magyari, L., & de Ruiter, J. P. (2012). Prediction of Turn-Ends Based on Anticipation of Upcoming Words. *Frontiers in Psychology*, *3*. https://doi.org/10.3389/fpsyg.2012.00376

Mah, J., Goad, H., & Steinhauer, K. (2016). Using Event-Related Brain Potentials to Assess Perceptibility : The Case of French Speakers and English [h]. *Frontiers in Psychology*, *7*. https://doi.org/10.3389/fpsyg.2016.01469

Marchman, V. A. (1997). Children's Productivity in the English Past Tense : The Role of Frequency, Phonology, and Neighborhood Structure. *Cognitive Science*, *21*(3), 283-304. https://doi.org/10.1207/s15516709cog2103_2

Marslen-Wilson, W. D. (1975). Sentence perception as an interactive parallel process. *Science (New York, N.Y.)*, *189*(4198), 226-228. https://doi.org/10.1126/science.189.4198.226

Marslen-Wilson, W. D., & Tyler, L. K. (1980). The temporal structure of spoken language understanding. *Cognition*, *8*(1), 1-71. https://doi.org/10.1016/0010-0277(80)90015-3

Martinez, R. (2011). *The development of a corpus-informed list of formulaic sequences for language pedagogy.* [Unpublished PhD thesis]. University of Nottingham.

Martinez, R., & Murphy, V. (2011). Effect of Frequency and Idiomaticity on Second Language Reading Comprehension. *TESOL Quarterly*, *45*(2), 267-290. https://doi.org/10.5054/tq.2011.247708

Martinez, R., & Schmitt, N. (2012). A Phrasal Expressions List. *Applied Linguistics*, *33*(3), 299-320. https://doi.org/10.1093/applin/ams010

Masperi, M. (2012). *Projet IDEFI Innovalangues : Innovation et transformation des pratiques de l'enseignement-apprentissage des langues dans l'enseignement supérieur*.

Mattys, S. L., White, L., & Melhorn, J. F. (2005). Integration of multiple speech segmentation cues : A hierarchical framework. *Journal of Experimental Psychology. General*, *134*(4), 477-500. https://doi.org/10.1037/0096-3445.134.4.477

McCauley, S., & Christiansen, M. H. (2015). Individual differences in chunking ability predict on-line sentence processing. *Proceedings of the 37th Annual Conference of the Cognitive Science Society*, 1553-1558.

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, *18*(1), 1-86. https://doi.org/10.1016/0010-0285(86)90015-0

McDonald, J. L., & Roussel, C. C. (2010). Past tense grammaticality judgment and production in non-native and stressed native English speakers*. *Bilingualism: Language and Cognition*, *13*(4), 429-448. https://doi.org/10.1017/S1366728909990599

McLean, S. (2018). Evidence for the Adoption of the Flemma as an Appropriate Word Counting Unit. *Applied Linguistics*, *39*(6), 823-845. https://doi.org/10.1093/applin/amw050

McLean, S., & Kramer, B. (2015). The Creation of a New Vocabulary Levels Test. *Shiken*, *19*(2), 1-11.

McLean, S., Kramer, B., & Beglar, D. (2015). The Creation and Validation of a Listening Vocabulary Levels Test. *Language Teaching Research*, *19*(6), 741-760. https://doi.org/10.1177/1362168814567889

McQueen, J. M. (1998). Segmentation of Continuous Speech Using Phonotactics. *Journal of Memory and Language*, *39*(1), 21-46. https://doi.org/10.1006/jmla.1998.2568

Meara, P. (1996). The dimensions of lexical competence. In G. Brown, K. Malmkjaer, & J. Williams, *Performance and Competence in Second Language Acquisition*. Cambridge University Press.

Meara, P., & Jones, G. (1988). Vocabulary size as a placement indicator. In *Applied Linguistics in Society* (P. Grunwell, p. 80-87).

Mecartty, F. H. (2000). Lexical and grammatical knowledge in reading and listening comprehension by foreign language learners of Spanish. *Applied Language Learning*, *11*(2), 323-348.

Meerman, A. D., Kiyama, S., & Tamaoka, K. (2014). To What Extent Does Accent Sensitivity Provide the Foundation for Lexical Knowledge and Listening Comprehension? *Open Journal of Modern Linguistics*, *04*(03), 457-464. https://doi.org/10.4236/ojml.2014.43037

Melnik, G. A., & Peperkamp, S. (2019a). Online phonetic training improves L2 word recognition. *Proceedings of the 41st Annual Conference of the Cognitive Science Society*. https://hal.archives-ouvertes.fr/hal-02397746

Melnik, G. A., & Peperkamp, S. (2019b). Perceptual deletion and asymmetric lexical access in second language learners. *The Journal of the Acoustical Society of America*, *145*(1), EL13-EL18. https://doi.org/10.1121/1.5085648

Messick, S. (1990). Validity of Test Interpretation and Use. *ETS Research Report Series*, *1990*(1), 1487-1495. https://doi.org/10.1002/j.2333-8504.1990.tb01343.x

Meyer, D. E., & Schvaneveldt, R. W. (1971). Facilitation in recognizing pairs of words : Evidence of a dependence between retrieval operations. *Journal of Experimental Psychology*, *90*(2), 227-234. https://doi.org/10.1037/h0031564

Michelas, A., Frauenfelder, U. H., Schön, D., & Dufour, S. (2016). How deaf are French speakers to stress? *The Journal of the Acoustical Society of America*, *139*(3), 1333-1342. https://doi.org/10.1121/1.4944574

Miller, G. A. (1956). The magical number seven plus or minus two : Some limits on our capacity for processing information. *Psychological Review*, *63*(2), 81-97.

Miller, G. A., Heise, G. A., & Lichten, W. (1951). The intelligibility of speech as a function of the context of the test materials. *Journal of Experimental Psychology*, *41*(5), 329-335.

Miller, G. A., & Nicely, P. E. (1955). An Analysis of Perceptual Confusions Among Some English Consonants. *The Journal of the Acoustical Society of America*, *27*(2), 338-352. https://doi.org/10.1121/1.1907526

Millis, K. K., & Just, M. A. (1994). The Influence of Connectives on Sentence Comprehension. *Journal of Memory and Language*, *33*(1), 128-147. https://doi.org/10.1006/jmla.1994.1007

Milton, J. (2009). *Measuring Second Language Vocabulary Acquisition*. Multilingual Matters.

Milton, J. (2013). Measuring the contribution of vocabulary knowledge to proficiency in the four skills. *L2 vocabulary acquisition, knowledge and use New perspectives on assessment and corpus analysis.*, *2*, 57–78.

Milton, J., & Hopkins, N. (2006). Comparing phonological and orthographic vocabulary size : Do vocabulary tests underestimate the knowledge of some learners? *Canadian Modern Language Review*, *63*(1), 127–147.

Milton, J., & Meara, P. (1995). How periods abroad affect vocabulary growth in a foreign language. *ITL - International Journal of Applied Linguistics*, *107*(1), 17-34. https://doi.org/10.1075/itl.107-108.02mil

Milton, J., & Treffers-Daller, J. (2013). Vocabulary size revisited : The link between vocabulary size and academic achievement. *Applied Linguistics Review*, *4*(1), 151–172. https://doi.org/10.1515/applirev-2013-0007

Milton, J., Wade, J., & Hopkins, N. (2010). Aural word recognition and oral competence in English as a foreign language. In R. Chacón-Beltrán, C. Abello-Contesse, M.

Torreblanca-López, & M. D. López-Jiménez (Éds.), *Further insights into non-native vocabulary teaching and learning* (p. 83-97). Multilingual Matters.

Ministère de l'Education Nationale. (1998). *Enseignement des langues vivantes au CM2 à la rentrée 1998- orientations pédagogiques* (Circulaire n°98-135; Bulletin Officiel de l'Education Nationale n°27, p. 1486-1507). MINISTERE DE L'EDUCATION NATIONALE.

Ministère de l'Education Nationale. (2002). *Programme d'enseignement des langues étrangères ou régionales au cycle des approfondissements à l'école primaire », Bulletin OfJiciel de I'Education Nationale, 29.08.2002, (http://www.eduscol.education.fr).*

Ministère de l'Education Nationale. (2007). *Bulletin officiel n°8 du 30 Aout 2007—Programmes de langues étrangères pour l'école primaire (Préambule commun)*.

Ministère de l'Education Nationale. (2010). *Bulletin officiel spécial n°9 du 30 septembre 2010—Programme d'enseignement de langues vivantes du cycle terminal pour les séries générales et technologiques NOR : MENE1019796A arrêté du 21-7-2010—J.O. du 28-8-2010 MEN - DGESCO A1-4*.

Ministère de l'Education Nationale. (2015). *Bulletin officiel n°44 du 26 novembre 2015—Horaires d'enseignement des écoles maternelles et élémentaires arrêté du 9-11-2015—J.O. du 24-11-2015 (NOR MENE1526553A).*

Ministère de l'Education Nationale. (2017a). *Repères & références statistiques sur les enseignements, la formation et la recherche.* (N° 34). Ministère de l'Éducation Nationale.

Ministère de l'Education Nationale. (2017b). *Bulletin officiel n°22 du 22 juin 2017—Enseignements au collège Organisation des enseignements : Modification arrêté du 16-6-2017—J.O. du 18-6-2017 (NOR MENE1717553A).*

Ministère de l'Education Nationale. (2019). *Bulletin officiel n°1 du 22 Janvier 2019—Programme d'enseignement commun et optionnel de langues vivantes de la classe de seconde générale et technologique et des classes de première et terminale des voies générale et technologique*. https://www.education.gouv.fr/bo/19/Special1/MENE1901585A.htm

Miralpeix, I., & Muñoz, C. (2018). Receptive vocabulary size and its relationship to EFL language skills. *International Review of Applied Linguistics in Language Teaching*, *56*(1), 1–24. https://doi.org/10.1515/iral-2017-0016

Mizumoto, A., & Shimamoto, T. (2008). A Comparison of Aural and Written Vocabulary Size of Japanese EFL University Learners. *Language Teaching and Technology*, *45*, 35-52.

Moodle.org. (s. d.). Consulté 7 novembre 2020, à l'adresse https://moodle.org/

Morgan-Short, K., Marsden, E., Heil, J., Ii, B. I. I., Leow, R. P., Mikhaylova, A., Mikołajczak, S., Moreno, N., Slabakova, R., & Szudarski, P. (2018). Multisite Replication in Second Language Acquisition Research : Attention to Form During Listening and Reading Comprehension. *Language Learning*, *68*(2), 392-437. https://doi.org/10.1111/lang.12292

Naiman, N. (1974). *The Use of Elicited Imitation in Second Language Acquisition Research. Working Papers on Bilingualism, No. 2*.

Nakata, T. (2008). English vocabulary learning with word lists, word cards and computers : Implications from cognitive psychology research for optimal spaced learning. *ReCALL*, *20*(1), 3-20. https://doi.org/10.1017/S0958344008000219

Nakatani, L. H., & Dukes, K. D. (1977). Locus of segmental cues for word juncture. *The Journal of the Acoustical Society of America*, *62*(3), 714-719. https://doi.org/10.1121/1.381583

Nakatani, L. H., & Schaffer, J. A. (1978). Hearing « words » without words : Prosodic cues for word perception. *The Journal of the Acoustical Society of America*, *63*(1), 234-245.

Nation, I. S. P. (1990). *Teaching & Learning Vocabulary* (1 edition). Heinle ELT.

Nation, I. S. P. (2001). *Learning Vocabulary in Another Language*. Cambridge University Press. https://doi.org/10.1017/CBO9781139524759

Nation, I. S. P. (2006). How Large a Vocabulary is Needed For Reading and Listening? *The Canadian Modern Language Review*, *63*(1), 59-82. https://doi.org/10.3138/cmlr.63.1.59

Nation, I. S. P. (2017). *The BNC/COCA Level 6 word family lists (Version 1.0.0) [Data file]*. http://www.victoria.ac.nz/lals/staff/paul-nation.aspx

Nation, I. S. P., & Beglar, D. (2007). A Vocabulary Size Test. *The Language Teacher*, *31*(7), 9-13.

Newman, R. S., Clouse, S. A., & Burnham, J. L. (2001). The perceptual consequences of within-talker variability in fricative production. *The Journal of the Acoustical Society of America*, *109*(3), 1181-1196. https://doi.org/10.1121/1.1348009

Ng, F. (2015). *Am I There Yet? : Probing the Effects of Goal Progress Feedback on Cognitive Motivation* [Mémoire de master]. Princeton University.

Norberg, C., & Nordlund, M. (2018). A Corpus-based Study of Lexis in L2 English Textbooks. *Journal of Language Teaching and Research*, *9*(3), 463. https://doi.org/10.17507/jltr.0903.03

Noreillie, A.-S., Kestemont, B., Heylen, K., Desmet, P., & Peters, E. (2018). Vocabulary knowledge and listening comprehension at an intermediate level in English and French as foreign languages : An approximate replication study of Stæhr (2009). *ITL - International Journal of Applied Linguistics*, *169*(1), 212-231. https://doi.org/10.1075/itl.00013.nor

Norris, D., & McQueen, J. M. (2008). Shortlist B : A Bayesian model of continuous speech recognition. *Psychological Review*, *115*(2), 357-395. https://doi.org/10.1037/0033-295X.115.2.357

Norris, D., McQueen, J. M., & Cutler, A. (2016). Prediction, Bayesian inference and feedback in speech recognition. *Language, Cognition and Neuroscience*, *31*(1), 4-18. https://doi.org/10.1080/23273798.2015.1081703

North, B., Ortega, A., & Sheehan, S. (2011). *A Core Inventory for General English*.

Nowrouzi, S., Tam, S. S., Zareian, G., & Nimehchisalem, V. (2015). Iranian EFL Students' Listening Comprehension Problems. *Theory and Practice in Language Studies*, *5*(2), 263. https://doi.org/10.17507/tpls.0502.05

Nusbaum, H., Pisoni, D. B., & Davis, C. K. (1984). *Sizing up the Hoosier Mental Lexicon : Measuring the Familiarity of 20,000 Words* (N° 10; Research on Speech Perception). Indiana University.

Oller, D. K. (1973). The effect of position in utterance on speech segment duration in English. *The Journal of the Acoustical Society of America*, *54*(5), 1235-1247.

Oller, J. W., & Conrad, C. A. (1971). The Cloze Technique and Esl Proficiency. *Language Learning*, *21*(2), 183-194. https://doi.org/10.1111/j.1467-1770.1971.tb00057.x

O'Malley, J. M., Chamot, A. U., & Küpper, L. (1989). Listening Comprehension Strategies in Second Language Acquisition. *Applied Linguistics*, *10*(4), 418-437. https://doi.org/10.1093/applin/10.4.418

Ostyn, P., & Godin, P. (1985). Ralex—An Alternative Approach To Language Teaching. *Modern Language Journal : Devoted to Research and Discussion about the Learning and Teaching of Foreign and Second Languages*, *69*(4), 346.

Paivio, A. (1991). Dual coding theory : Retrospect and current status. *Canadian Journal of Psychology/Revue Canadienne de Psychologie*, *45*(3), 255-287. https://doi.org/10.1037/h0084295

Pallier, C., Colomé, A., & Sebastián-Gallés, N. (2001). The influence of native-language phonology on lexical access : Exemplar-based versus abstract lexical entries. *Psychological Science*, *12*(6), 445-449. https://doi.org/10.1111/1467-9280.00383

Pallier, Christophe, Bosch, L., & Sebastián-Gallés, N. (1997). A limit on behavioral plasticity in speech perception. *Cognition*, *64*(3), B9-B17. https://doi.org/10.1016/S0010-0277(97)00030-9

Parisse, C. (2009). La morphosyntaxe : Qu'est-ce que c'est ? Application au cas de la langue française. *Rééducation Orthophonique*, *47*(238), 7-20.

Payre-Ficout, C. (2007). *L'apprentissage du prétérit et du present perfect dans le cadre scolaire : Étude extensive chez des apprenants francophones du secondaire et des étudiants du supérieur* [Thèse]. Grenoble 3.

Payre-Ficout, C. (2011). Conception et mise en place d'un dispositif hybride pour accompagner les étudiants de première année LLCE dans leur acquisition de l'anglais. *Recherche et pratiques pédagogiques en langues de spécialité. Cahiers de l'Apliut*, *Vol. XXX N° 1*, 102-116. https://doi.org/10.4000/apliut.459

Peereman, R. (2019, novembre 22). *Le corpus APPREL2 : Environnement lexical et listes de fréquence*. Journées d'étude ReAL2 : lexique (mémoire, acquisition, communication), Université Lyon 2. https://real.cnrs.fr/je-lyon2019

Pélissier, M. (2018). *Effets d'entrainements explicites et implicites sur l'acquisition de la syntaxe de l'anglais par des apprenants francophones : Étude en potentiels évoqués* [These de doctorat, Sorbonne Paris Cité]. http://www.theses.fr/2018USPCC091

Pellicer-Sánchez, A., & Schmitt, N. (2012). Scoring Yes–No vocabulary tests : Reaction time vs. nonword approaches. *Language Testing*, *29*(4), 489-509. https://doi.org/10.1177/0265532212438053

Peppé, S., & McCann, J. (2003). Assessing intonation and prosody in children with atypical language development : The PEPS-C test and the revised version. *Clinical Linguistics & Phonetics*, *17*(4-5), 345-354.

Perfetti, C. A., & Hart, L. (2002). The lexical quality hypothesis. In L. Verhoeven, C. Elbro, & P. Reitsma (Éds.), *Studies in Written Language and Literacy* (Vol. 11, p. 189-213). John Benjamins Publishing Company. https://doi.org/10.1075/swll.11.14per

Peterson, G. E., & Barney, H. L. (1952). Control Methods Used in a Study of the Vowels. *The Journal of the Acoustical Society of America*, *24*(2), 175-184. https://doi.org/10.1121/1.1906875

Pickering, M. J., & Garrod, S. (2013). An integrated theory of language production and comprehension. *The Behavioral and Brain Sciences*, *36*(4), 329-347. https://doi.org/10.1017/S0140525X12001495

Picoche, J. (2011). *Le vocabulaire et son enseignement.* (Ressources pour l'école primaire). Ministère de l'Éducation Nationale.

Pierrehumbert, J. B. (2003). Phonetic Diversity, Statistical Learning, and Acquisition of Phonology. *Language and Speech*, *46*(2-3), 115-154. https://doi.org/10.1177/00238309030460020501

Pierrehumbert, J. B. (2016). *Phonological representation : Beyond abstract versus episodic.* 32.

Pigada, M., & Schmitt, N. (2006). Vocabulary Acquisition from Extensive Reading : A Case Study. *Reading in a Foreign Language*, *18*(1), 1-28.

Pinker, S. (1999). *Words and Rules : The Ingredients Of Language* (Reprint edition). Basic Books.

Pisoni, D. B., & Luce, P. A. (1987). Acoustic-phonetic representations in word recognition. *Cognition*, *25*(1), 21-52. https://doi.org/10.1016/0010-0277(87)90003-5

Plag, I., Kunter, G., & Schramm, M. (2011). Acoustic correlates of primary and secondary stress in North American English. *Journal of Phonetics*, *39*(3), 362-374. https://doi.org/10.1016/j.wocn.2011.03.004

Plonsky, L., & Oswald, F. L. (2014). How Big Is "Big"? Interpreting Effect Sizes in L2 Research: Effect Sizes in L2 Research. *Language Learning*, *64*(4), 878-912. https://doi.org/10.1111/lang.12079

Potapenko, E. (s. d.). *PlayPhrase.me : Site for cinema archaeologists.* Consulté 6 novembre 2020, à l'adresse https://www.playphrase.me/

Preston, K. A. (1935). The speed of word perception and its relation to reading ability. *Journal of General Psychology*, *13*, 199-203. https://doi.org/10.1080/00221309.1935.9917878

Prodeau, M., Lopez, S., & Véronique, D. (2012). Acquisition of French as a Second Language : Do developmental stages correlate with CEFR levels? *Apples - Journal of Applied Language Studies*. https://jyx.jyu.fi/dspace/handle/123456789/40865

Purpura, J. E. (1999). *Learner Strategy Use and Performance on Language Tests : A Structural Equation Modeling Approach*. Cambridge University Press.

R Development Core Team. (2005). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. http://www.R-project.org

Ramachandran, S. D., & Rahim, H. A. (2004). Meaning Recall and Retention : The Impact of the Translation Method on Elementary Level Learners' Vocabulary Learning: *RELC Journal*, *35*(2), 161-178. https://doi.org/10.1177/003368820403500205

Rankin, P. T. (1928). The Importance of Listening Ability. *The English Journal*, *17*(8), 623-630. JSTOR. https://doi.org/10.2307/803100

Read, C., Yun-Fei, Z., Hong-Yin, N., & Bao-Qing, D. (1986). The ability to manipulate speech sounds depends on knowing alphabetic writing. *Cognition*, *24*(1), 31-44. https://doi.org/10.1016/0010-0277(86)90003-X

Read, J. (1993). The development of a new measure of L2 vocabulary knowledge. *Language Testing*, *10*(3), 355-371. https://doi.org/10.1177/026553229301000308

Read, J. (2008). Identifying academic language needs through diagnostic assessment. *Journal of English for Academic Purposes*, *7*(3), 180-190. https://doi.org/10.1016/j.jeap.2008.02.001

Regier, T., & Gahl, S. (2004). Learning the unlearnable : The role of missing evidence. *Cognition*, *93*(2), 147-155. https://doi.org/10.1016/j.cognition.2003.12.003

Revelle, W. R. (2017). *psych : Procedures for Personality and Psychological Research*. https://www.scholars.northwestern.edu/en/publications/psych-procedures-for-personality-and-psychological-research

Ridgway, T., & Field, J. (2000). Point and Counterpoint : Listening Strategies--I Beg Your Pardon [and. *ELT Journal*, *54*(2), 179-197.

Roach, P. (2009). *English phonetics and phonology : A practical course. Book: ...* (3. ed., 12. print). Cambridge Univ. Press.

Roelofs, A., Meyer, A. S., & Levelt, W. J. M. (1998). *A case for the lemma/lexeme distinction in models of speaking : Comment on Caramazza and Miozzo (1997)*. 12.

Roussel, S., Gruson, B., & Galan, J.-P. (2017). What Types of Training Improve Learners' Performances in Second Language Listening Comprehension? *International Journal of Listening*, *0*(0), 1-14. https://doi.org/10.1080/10904018.2017.1331133

Rubin, P., Turvey, M. T., & Van Gelder, P. (1976). Initial phonemes are detected faster in spoken words than in spoken nonwords. *Perception & Psychophysics*, *19*(5), 394-398. https://doi.org/10.3758/BF03199398

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science (New York, N.Y.)*, *274*(5294), 1926-1928.

Saffran, J. R., Newport, E. L., & Aslin, R. N. (1996). Word Segmentation : The Role of Distributional Cues. *Journal of Memory and Language*, *35*(4), 606-621. https://doi.org/10.1006/jmla.1996.0032

Salamoura, A., & Saville, N. (2010). Exemplifying the CEFR: criterial features of written learner English from the English Profile Programme. *EUROSLA MONOGRAPHS SERIES*, *1*, 32.

Salverda, A. P., Dahan, D., & McQueen, J. M. (2003). The role of prosodic boundaries in the resolution of lexical embedding in speech comprehension. *Cognition*, *90*(1), 51-89. https://doi.org/10.1016/S0010-0277(03)00139-2

Saragi, T., Nation, I. S. P., & Meister, G. F. (1978). Vocabulary learning and reading. *System*, *6*(2), 72-78. https://doi.org/10.1016/0346-251X(78)90027-1

Savignon, S. J. (1976). *Communicative Competence : Theory and Classroom Practice*. Central States Conference on the Teaching of Foreign Languages, Detroit, Michigan.

Schepens, J., Dijkstra, T., Grootjen, F., & Heuven, W. J. B. van. (2013). Cross-Language Distributions of High Frequency and Phonetically Similar Cognates. *PLOS ONE*, *8*(5), e63006. https://doi.org/10.1371/journal.pone.0063006

Schmidt, R. W. (1990). The Role of Consciousness in Second Language Learning. *Applied Linguistics*, *11*(2), 129-158.

Schmitt, N. (2008). Review article : Instructed second language vocabulary learning. *Language Teaching Research*, *12*(3), 329-363. https://doi.org/10.1177/1362168808089921

Schmitt, N. (2014). Size and Depth of Vocabulary Knowledge : What the Research Shows. *Language Learning*, *64*(4), 913-951. https://doi.org/10.1111/lang.12077

Schmitt, N., Schmitt, D., & Clapham, C. (2001). Developing and exploring the behaviour of two new versions of the Vocabulary Levels Test. *Language Testing*, *18*(1), 55-88. https://doi.org/10.1177/026553220101800103

Seashore, R. H., & Eckerson, L. D. (1940). The measurement of individual differences in general English vocabularies. *Journal of Educational Psychology*, *31*(1), 14-38. https://doi.org/10.1037/h0053494

Selinker, L. (1972). Interlanguage. *IRAL - International Review of Applied Linguistics in Language Teaching*, *10*(1-4). https://doi.org/10.1515/iral.1972.10.1-4.209

Shillcock, R., & Bard, E. (1993). Modularity and the processing of closed-class words. In G. Altmann & R. Shillcock, *Cognitive Models Of Speech Processing : The Second Sperlonga Meeting* (p. 163-183). Psychology Press.

Shin, D., & Nation, P. (2008). Beyond single words : The most frequent collocations in spoken English. *ELT Journal*, *62*(4), 339-348. https://doi.org/10.1093/elt/ccm091

Shiotsu, T., & Weir, C. J. (2007). The relative significance of syntactic knowledge and vocabulary breadth in the prediction of reading comprehension test performance. *Language Testing*, *24*(1), 99-128. https://doi.org/10.1177/0265532207071513

Shoemaker, E. (2014). The exploitation of subphonemic acoustic detail in L2 speech segmentation. *Studies in Second Language Acquisition*, *36*(04), 709–731. https://doi.org/10.1017/S027226311400014X

Shohamy, E. (1992). Beyond Proficiency Testing : A Diagnostic Feedback Testing Model for Assessing Foreign Language Learning. *The Modern Language Journal*, *76*(4), 513-521. https://doi.org/10.2307/330053

Shuy, R. W. (1981). A Holistic View of Language. *Research in the Teaching of English*, *15*(2), 101-111. JSTOR.

Simpson, J. A. (Éd.). (1989). *The Oxford English dictionary* (2. ed). Clarendon Press.

Sims, V. M. (1929). The Reliability and Validity of Four Types of Vocabulary Tests. *The Journal of Educational Research*, *20*(2), 91-96.

Sinclair, J. (1991). *Corpus, Concordance, Collocation*. Oxford University Press.

Sinclair, J. M. (1987). *Looking Up : An Account of the COBUILD Project in Lexical Computing and the Development of the Collins COBUILD English Language Dictionary*. Collins ELT.

Siyanova-Chanturia, A. (2015). On the 'holistic' nature of formulaic language. *Corpus Linguistics and Linguistic Theory*, *11*(2), 285–301. https://doi.org/10.1515/cllt-2014-0016

Slowiaczek, L. M. (1990). Effects of Lexical Stress in Auditory Word Recognition. *Language and Speech*, *33*(1), 47-68. https://doi.org/10.1177/002383099003300104

Sosa, A. V., & MacFarlane, J. (2002). Evidence for frequency-based constituents in the mental lexicon : Collocations involving the word of. *Brain and Language*, *83*(2), 227-236. https://doi.org/10.1016/S0093-934X(02)00032-9

Sournin-Dufossé, S. (2007). *Les théories linguistiques, les pratiques pédagogiques et l'acquisition de la détermination nominale en anglais chez les apprenants francophones* [Thèse, La Réunion]. http://www.theses.fr/2007LARE0013

Spada, N. (2015). SLA research and L2 pedagogy : Misapplications and questions of relevance. *Language Teaching*, *48*(1), 69-81. https://doi.org/10.1017/S026144481200050X

Speer, S. R., Kjelgaard, M. M., & Dobroth, K. M. (1996). The influence of prosodic structure on the resolution of temporary syntactic closure ambiguities. *Journal of psycholinguistic research*, *25*(2), 249-271. https://doi.org/10.1007/BF01708573

Stæhr, L. S. (2008). Vocabulary size and the skills of listening, reading and writing. *The Language Learning Journal*, *36*(2), 139-152. https://doi.org/10.1080/09571730802389975

Stæhr, L. S. (2009). Vocabulary Knowledge and Advanced Listening Comprehension in English as a Foreign Language. *Studies in Second Language Acquisition*, *31*(04), 577–607. https://doi.org/10.1017/S0272263109990039

Strand, E. A., & Johnson, K. (1996). Gradient and Visual Speaker Normalization in the Perception of Fricatives. *Natural Language Processing and Speech Technology, Results of the 3rd KONVENS Conference*, 14–26.

Strange, W., & Dittmann, S. (1984). Effects of discrimination training on the perception of /r-l/ by Japanese adults learning English. *Perception & Psychophysics*, *36*(2), 131-145.

Suzuki, Y. (2017). Validity of new measures of implicit knowledge : Distinguishing implicit knowledge from automatized explicit knowledge. *Applied Psycholinguistics*, *38*(5), 1229-1261.

Sweet, H. (1899). *The Practical Study of Languages : A Guide for Teachers and Learners*. J. M. Dent & Company.

Sweller, J. (2014). Implications of Cognitive Load Theory for Multimedia Learning. In R. Mayer (Éd.), *The Cambridge Handbook of Multimedia Learning*.

Szucs, D., & Ioannidis, J. P. A. (2017). When Null Hypothesis Significance Testing Is Unsuitable for Research : A Reassessment. *Frontiers in Human Neuroscience*, *11*. https://doi.org/10.3389/fnhum.2017.00390

Szudarski, P. (2018). *Corpus Linguistics for Vocabulary : A Guide for Research* (1 edition). Routledge.

Tabata, M. (2016). *The Relationships between Sound Sensitivity, English Prosody Processing, and English Listening Comprehension*. Nagoya.

Tagliante, C., Mègre, B., Breton, G., Dupleix, D., & Houssa, C. (2004). Les tests de langues. *Revue internationale d'éducation de Sèvres*, *37*, 115-121. https://doi.org/10.4000/ries.1474

Terrier, L. (2011). *Méthodologie linguistique pour l'évaluation des restitutions et analyse expérimentale des processus de didactisation du son : Recommandations pour un apprentissage raisonné de la compréhension de l'anglais oral par les étudiants francophones du secteur LANSAD* [Toulouse 3]. http://www.theses.fr/2011TOU30267

Terrier, L., & Maury, C. (2015). De la gestion des masses à une offre de formation individualisée en anglais-LANSAD : Tensions et structuration. *Recherche et pratiques pédagogiques en langues de spécialité. Cahiers de l'Apliut*, *Vol. XXXIV N° 1*, 67-89. https://doi.org/10.4000/apliut.5029

Thorndike, E. L. (1921). *The teacher's word book*. New York Teachers College, Columbia University. http://archive.org/details/teacherswordbook00thoruoft

Thorndike, E. L. (1927). The Law of Effect. *The American Journal of Psychology*, *39*(1/4), 212-222. https://doi.org/10.2307/1415413

Thorndike, E. L. (1931). *A teacher's word book of the twenty thousand words found most frequently and widely in general reading for children and young people,*. http://hdl.handle.net/2027/mdp.39015004965102

Tracy-Ventura, N., McManus, K., Norris, J., & Ortega, L. (2014). "Repeat as much as you can" : Elicited imitation as a measure of oral proficiency in L2 French. In P. Leclercq, A. Edmonds, & H. Hilton, *Measuring L2 Proficiency : Perspectives from SLA* (p. 143-166). Multilingual Matters.

Traxler, M. J., Bybee, M. D., & Pickering, M. J. (1997). Influence of Connectives on Language Comprehension : Eye tracking Evidence for Incremental Interpretation. *The Quarterly Journal of Experimental Psychology Section A*, *50*(3), 481-497. https://doi.org/10.1080/027249897391982

Tremblay, A. (2008). Is second language lexical access prosodically constrained? Processing of word stress by French Canadian second language learners of English. *Applied Psycholinguistics*, *29*(04), 553–584. https://doi.org/10.1017/S0142716408080247

Trim, J. L. M. (2001). *Breakthrough*. Council of Europe. http://www.englishprofile.org/index.php/resources/t-series

Tsui, A. B. M., & Fullilove, J. (1998). Bottom-up or Top-down Processing as a Discriminator of L2 Listening Performance. *Applied Linguistics*, *19*(4), 432-451. https://doi.org/10.1093/applin/19.4.432

Tuinman, A., Mitterer, H., & Cutler, A. (2014). Use of Syntax in Perceptual Compensation for Phonological Reduction. *Language and Speech*, *57*(1), 68-85. https://doi.org/10.1177/0023830913479106

Tulving, E., & Gold, C. (1963). Stimulus information and contextual information as determinants of tachistoscopic recognition of words. *Journal of Experimental Psychology*, *66*(4), 319-327. https://doi.org/10.1037/h0048802

Tyler, M. D., & Cutler, A. (2009). Cross-language differences in cue use for speech segmentation. *The Journal of the Acoustical Society of America*, *126*(1), 367-376. https://doi.org/10.1121/1.3129127

Urmston, A., & Raquel, M. (2015). *Developing a diagnostic, post-entry language assessment amidst changing policy and practice*. 12ème conférence EALTA, Copenhague.

Urmston, A., Raquel, M., & Tsang, C. (2013). Diagnostic testing of Hong Kong tertiary students' English Language proficiency : The development and validation of DELTA. *Hong Kong Journal of Applied Linguistics*, *14*(2), 60-82.

Vafaee, P., Suzuki, Y., & Kachisnke, I. (2017). Validating Grammaticality Judgment Tests : Evidence from Two New Psycholinguistic Measures. *Studies in Second Language Acquisition*, *39*(1), 59-95. https://doi.org/10.1017/S0272263115000455

Vaissière, J. (1983). Language-Independent Prosodic Features. In A. Cutler & D. R. Ladd (Éds.), *Prosody : Models and Measurements* (p. 53-66). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-69103-4_5

Van Ek, J. A. (1975). *The Threshold Level in a European Unit/Credit System for Modern Language Learning by Adults*. Council of Europe.

Van Ek, J. A., & Trim, J. L. M. (1990a). *Threshold 1990*. Cambridge University Press. http://www.englishprofile.org/resources/t-series

Van Ek, J. A., & Trim, J. L. M. (1990b). *Waystage 1990*. Council of Europe/Cambridge University Press. http://www.englishprofile.org/resources/t-series

Van Ek, J. A., & Trim, J. L. M. (2001). *Vantage 2001*. Council of Europe/Cambridge University Press. http://www.englishprofile.org/resources/t-series

Van Patten, B. (2002). Processing Instruction : An Update. *Language Learning*, *52*(4), 755-803. https://doi.org/10.1111/1467-9922.00203

Van Patten, B., & Cadierno, T. (1993). Input Processing and Second Language Acquisition : A Role for Instruction. *The Modern Language Journal*, *77*(1), 45-57. https://doi.org/10.1111/j.1540-4781.1993.tb01944.x

Van Petten, C., & Luka, B. J. (2012). Prediction during language comprehension : Benefits, costs, and ERP components. *International Journal of Psychophysiology*, *83*(2), 176-190. https://doi.org/10.1016/j.ijpsycho.2011.09.015

van Zeeland, H. (2013). L2 vocabulary knowledge in and out of context : Is it the same for reading and listening? *Australian Review of Applied Linguistics*, *36*(1), 52-70. https://doi.org/10.1075/aral.36.1.03van

van Zeeland, H., & Schmitt, N. (2013). Lexical Coverage in L1 and L2 Listening Comprehension : The Same or Different from Reading Comprehension? *Applied Linguistics*, *34*(4), 457-479. https://doi.org/10.1093/applin/ams074

Vandergrift, L. (2003). Orchestrating Strategy Use : Toward a Model of the Skilled Second Language Listener. *Language Learning*, *53*(3), 463-496.

Vandergrift, L., & Goh, C. (2012). *Teaching and Learning Second Language Listening*. Routledge.

Vandergrift, L., Goh, C. C. M., Mareschal, C. J., & Tafaghodtari, M. H. (2006). The Metacognitive Awareness Listening Questionnaire : Development and Validation. *Language Learning*, *56*(3), 431-462. https://doi.org/10.1111/j.1467-9922.2006.00373.x

Vermeer, A. (2001). Breadth and depth of vocabulary in relation to L1/L2 acquisition and frequency of input. *Applied Psycholinguistics*, *22*(2), 217-234.

Vidal, K. (2011). A Comparison of the Effects of Reading and Listening on Incidental Vocabulary Acquisition. *Language Learning*, *61*(1), 219-258. https://doi.org/10.1111/j.1467-9922.2010.00593.x

Viel, M. (2003). *Manuel de phonologie anglaise*. Armand Colin.

Voise, A.-M. (2010). Enseigner la phonologie de l'anglais aux futurs professeurs du primaire. *Recherche et pratiques pédagogiques en langues de spécialité. Cahiers de l'Apliut*, *XXIX*(2), 11-24. https://doi.org/10.4000/apliut.673

Vraciu, E. A. (2012). *La morphologie temporo-aspectuelle chez des apprenants avancés d'anglais langue étrangère : Une étude des facteurs sémantiques, discursifs et inter-linguistiques* [Thesis, Paris 10]. http://www.theses.fr/2012PA100062

Walley, A. C. (2007). Speech learning, lexical reorganization, and the development of word recognition by native and non-native English speakers. In O.-S. Bohn & M. J. Munro

(Éds.), *Language Learning & Language Teaching* (Vol. 17, p. 315-330). John Benjamins Publishing Company. https://doi.org/10.1075/lllt.17.27wal

Walter, H. (2003). *Honni soit qui mal y pense*. Le Livre de Poche.

Ward, J., & Chuenjundaeng, J. (2009). Suffix knowledge : Acquisition and applications. *System*, *37*(3), 461-469. https://doi.org/10.1016/j.system.2009.01.004

Warner, N., & Cutler, A. (2017). Stress Effects in Vowel Perception as a Function of Language-Specific Vocabulary Patterns. *Phonetica*, *74*(2), 81-106. https://doi.org/10.1159/000447428

Warren, R. M., & Warren, R. P. (1970). Auditory illusions and confusions. *Scientific American*, *223*(6), 30-36. https://doi.org/10.1038/scientificamerican1270-30

Wason, P. C., & Reich, S. S. (1979). A verbal illusion. *Quarterly Journal of Experimental Psychology*, *31*(4), 591-597. https://doi.org/10.1080/14640747908400750

Wauquier-Gravelines, S. (1999). Segmentation lexicale de la parole continue : La linéarité en question. *Recherches linguistiques de Vincennes*, *28*, 133-156. https://doi.org/10.4000/rlv.1217

Webb, S., & Rodgers, M. P. H. (2009a). Vocabulary Demands of Television Programs. *Language Learning*, *59*(2), 335-366. https://doi.org/10.1111/j.1467-9922.2009.00509.x

Webb, S., & Rodgers, M. P. H. (2009b). The Lexical Coverage of Movies. *Applied Linguistics*, *30*(3), 407-427. https://doi.org/10.1093/applin/amp010

Weber, A., & Broersma, M. (2012). Spoken Word Recognition in Second Language Acquisition. In C. A. Chapelle (Éd.), *The Encyclopedia of Applied Linguistics*. Blackwell Publishing Ltd. http://doi.wiley.com/10.1002/9781405198431.wbeal1104

Weber, A., & Scharenborg, O. (2012). Models of spoken-word recognition. *Wiley Interdisciplinary Reviews: Cognitive Science*, *3*(3), 387-401. https://doi.org/10.1002/wcs.1178

Wells, J. C. (1982). *Accents of English : Volume 3: Beyond the British Isles*. Cambridge University Press.

Werker, J. F., Frost, P. E., & McGurk, H. (1992). La langue et les Lèvres : Cross-language influences on bimodal speech perception. *Canadian Journal of Psychology*, *46*(4), 551-568.

Wesche, M., & Paribakht, T. S. (1996). Assessing Second Language Vocabulary Knowledge : Depth Versus Breadth. *Canadian Modern Language Review*, *53*(1), 13-40.

West, M. (1953). *A General Service List of English Words*. Longman.

Whalley, K., & Hansen, J. (2006). The role of prosodic sensitivity in children's reading development. *Journal of Research in Reading*, *29*(3), 288-303. https://doi.org/10.1111/j.1467-9817.2006.00309.x

Wheeler, D. D. (1970). Processes in word recognition. *Cognitive Psychology*, *1*(1), 59-85. https://doi.org/10.1016/0010-0285(70)90005-8

White, E. J., Titone, D., Genesee, F., & Steinhauer, K. (2017). Phonological processing in late second language learners : The effects of proficiency and task. *Bilingualism: Language and Cognition*, *20*(1), 162-183. https://doi.org/10.1017/S1366728915000620

White, L., Mattys, S. L., & Wiget, L. (2012). Segmentation Cues in Conversational Speech : Robust Semantics and Fragile Phonotactics. *Frontiers in Psychology*, *3*. https://doi.org/10.3389/fpsyg.2012.00375

White, L., Melhorn, J. F., & Mattys, S. L. (2010). Segmentation by lexical subtraction in Hungarian speakers of second-language English. *The Quarterly Journal of Experimental Psychology*, *63*(3), 544-554. https://doi.org/10.1080/17470210903006971

Widdowson, H. G. (1978). *Teaching Language as Communication*. OUP Oxford.

Wilson, I., Kaneko, E., Lyddon, P., Okamoto, K., & Ginsburg, J. (2011, août 17). *Nonsense-Syllable Sound Discrimination Ability Correlates with Second Language (L2) Proficiency*. ICPhS XVII, Hong Kong.

Wilson, M. (2003). Discovery listening—Improving perceptual processing. *ELT Journal*, *57*(4), 335-343. https://doi.org/10.1093/elt/57.4.335

Wollack, J. A., & Fremer, J. J. (2013). *Handbook of Test Security*. Routledge.

Wood, C., & Terrell, C. (1998). Poor readers' ability to detect speech rhythm and perceive rapid speech. *British Journal of Developmental Psychology*, *16*(3), 397-413. https://doi.org/10.1111/j.2044-835X.1998.tb00760.x

Wray, A. (2000). Formulaic sequences in second language teaching : Principle and practice. *Applied Linguistics*, *21*(4), 463-489. https://doi.org/10.1093/applin/21.4.463

Wray, Alison. (2008). *Formulaic Language : Pushing the Boundaries*. OUP Oxford.

Yeldham, M. (2017). Developing a framework for investigating L2 listeners' longitudinal development. *International Review of Applied Linguistics in Language Teaching*, *57*(2), 235–263. https://doi.org/10.1515/iral-2017-0004

Zechmeister, E. B., Chronis, A. M., Cull, W. L., D'Anna, C. A., & Healy, N. A. (1995). Growth of a Functionally Important Lexicon. *Journal of Reading Behavior*, *27*(2), 201-212.

Zoghlami, N. (2015). *Processus ascendants et descendants en compréhension de l'oral en langue étrangère—Problèmes et retombées didactiques pour la compréhension de l'anglais*. http://www.theses.fr/s113406

Zoghlami, N. (2016). La compréhension de l'anglais oral (L2) : Processus cognitifs et comportements stratégiques. *Recherche et pratiques pédagogiques en langues de spécialité. Cahiers de l'Apliut*, *Vol. 35 N° 1*. https://doi.org/10.4000/apliut.5322

Zwaan, R. A., & Radvansky, G. A. (1998). Situation models in language comprehension and memory. *Psychological Bulletin*, *123*(2), 162-185.

Liste des Figures

Liste des Tableaux

Annexes

Annexe 1 - Feuille de route et Questionnaire de biographie langagière

UGA UFR LE NOM : Prénom :

Vous allez prendre part à une expérimentation dont le but est de déterminer l'importance de la reconnaissance du vocabulaire, de la discrimination phonétique et des connaissances grammaticales dans la compréhension de l'oral, afin de concevoir des activités de remédiation correspondantes.

Vos résultats seront utilisés pour calculer des corrélations entre les différents tests et le test de positionnement de début d'année. Pour plus de renseignements ou si vous voulez connaître les conclusions finales de l'expérimentation, merci de contacter :

marie-pierre.jouannaud@univ-grenoble-alpes.fr

Signez-ici si vous acceptez que vos résultats soient utilisés :

Connectez-vous à : **self.innovalangues.net** , et créez-vous un profil avec une votre adresse mail personnelle ou universitaire.

Etape 1 : test de sensibilité accentuelle Code de session: **proso2018**

Instructions : Le test dure environ 12 minutes. Il est composé de 3 parties : dans la première partie, vous devez indiquer si, à votre avis, les 2 mots que vous entendez ont le même rythme. Certains mots sont accentués correctement, d'autres non, mais vous devez uniquement décider s'ils sont accentués sur la même syllabe ou pas.
Dans la deuxième partie et dans la troisième, vous devez indiquer quelle syllabe est accentuée.
Dans la dernière, vous devez décider quelle expression vous entendez.

Etape 2 : test de vocabulaire oral. Code de session: **voc2018**

Instructions : Le test dure environ 10 minutes. A chaque question, cliquez sur la bulle et écoutez le mot. Indiquez si vous connaissez le mot (« yes ») ou pas (« no »). Si vous êtes sûr(e) que le mot existe, cliquez sur « yes » même si vous ne savez pas exactement ce qu'il veut dire. Si vous n'êtes pas sûr(e), cliquez sur « no ».

Etape 3 : test de sensibilité phonémique (différenciation de sons proches) Code: **phono2018**

Instructions: Le test dure environ 10 minutes. A chaque question, cliquez sur la bulle et écoutez les 3 mots. Le but est de repérer l'intrus, prononcé différemment des autres.

Etape 4 : test de jugement grammatical Code de session : **gram2018**

Instructions : Le test dure environ 10 minutes. A chaque question, cliquez sur la bulle et écoutez la phrase. Indiquez si la phrase vous paraît correcte (« yes ») ou pas (« no »).

Etape 5 : test Phrasal Vocabulary Size Test Code de session : **PVST2018**

Instructions: Le test dure environ 10 minutes. A chaque question, trouvez l'expression synonyme

Merci de votre coopération !

Language Background Questionnaire **NAME/ NOM:**

Merci de compléter ce questionnaire aussi complètement et objectivement que possible. Le contenu de ce questionnaire est strictement confidentiel. Vous n'êtes pas obligé(e)s de répondre à toutes les questions, mais votre coopération nous sera très utile. Merci de votre aide !

1. Age _____
2. Sexe: M F
3. Langue maternelle _____
   Si vous avez plus d'une langue maternelle, merci de les indiquer toutes les deux, et entourez éventuellement celle que vous considérez comme dominante.

4. Merci d'indiquer ici l'anglais ainsi que les autres langues que vous parlez (y compris le français si ce n'est pas votre langue maternelle)

A1 (débutant), A2 (élémentaire) B1 (intermédiaire) B2 (avancé) C1/ C2 (expert)

Langue	Age/ classe de début	Niveau
_____	_____	A1 A2 B1 B2 C1/C2
_____	_____	A1 A2 B1 B2 C1/C2
_____	_____	A1 A2 B1 B2 C1/C2
_____	_____	A1 A2 B1 B2 C1/C2

5. Combien d'années scolaires de cours d'anglais avez-vous eues? _____

6. Si vous avez fait des séjours en pays anglophones, notez où vous êtes allé(e), l'âge que vous aviez et combien de temps vous êtes resté(e) :

Pays	age	durée
_____	_____	_____
_____	_____	_____
_____	_____	_____
_____	_____	_____

Annexe 2 - Items du test de sensibilité prosodique

Les items barrés n'ont pas été conservés dans la version finale du test.

num	stimuli	clé	précisions	
Tâche 1 (comparaison de patrons accentuels – même patron ou pas ?)				
			accentuation habituelle (H) ou inhabituelle (I)	patrons
1	~~comfortable cauliflower~~ (entraînement)	oui	H – H	/1000/ /1000/
2	~~military magnificent~~ (entraînement)	non	H – H	/1000/ /0100/
3	maternity botanical	oui	I – I	/1000/ /1000/
4	~~facility necessity~~	oui	H – H	/0100/ /0100/
5	~~maternity botanical~~	oui	H – H	/0100/ /0100/
6	dandelion mercenary	oui	H – H	/1000/ /1000/
7	lavatory fertilizer	non	I – H	/0100/ /1000/
8	historical curriculum	non	H – I	/0100/ /1000/
9	delicacy monastery	non	H – I	/1000/ /0100/
10	capacity ridiculous	non	H – I	/0100/ /1000/
11	cauliflower caterpillar	non	I – H	/0100/ /1000/
12	difficulty voluntary	oui	H – H	/1000/ /1000/
13	comfortable organizer	oui	I – I	/0100/ /0100/
14	educator categorize	oui	H – H	/1000/ /1000/
15	~~secondary military~~	oui	I – I	/0100/ /0100/
16	auditory citizenship	non	H – I	/1000/ /0100/
17	~~punishable pacifier~~	oui	I – I	/0100/ /0100/
18	magnificent delivery	oui	I – I	/1000/ /1000/
19	~~participant manipulate~~	non	I – H	/1000/ /0100/
20	miraculous pistachio	non	H – I	/0100/ /1000/
Tâche 2a (identification de la syllabe accentuée de mots existants)				
23	**1 item de transition**			
24	adjective	1		
25	argument	1		
26	artifice	1		
27	character	1		
28	consider	2		
29	contribute	2		
30	develop	2		
31	encourage	2		
32	etiquette	1		
33	harvested	1		
34	implicit	2		
35	interpret	2		
36	interview	1		
37	motivate	1		
38	negative	1		
39	occurrence	2		
40	processes	1		
41	rhetoric	1		
42	satire	1		
43	specific	2		
44	syntactic	2		

Tâche 2b (identification de la syllabe accentuée de pseudomots)			
45	**1 item de transition**		
46	menoysa	2	
47	soibena	1	
48	soidetta	2	
49	zedoola	2	
50	fellazee	3	
51	linnesoo	3	
52	taremma	2	
53	faysaboo	3	
54	pecoitay	2	
55	koyvalee	1	
56	pagenoo	1	
57	leytauwma	2	
58	nafeepa	1	
59	dafeanoo	2	
60	roodela	1	
61	joamaray	3	
62	savossauw	2	
63	~~delloyma~~	1	
64	savvaney	3	
65	lisseda	1	
Tâche 3 (identification de la phrase entendue)			
66	**item de transition**		
67	~~I need chocolate, biscuits, jam and juice~~		
68	I need chocolate milk, tea, and coffee		
69	I need ice, cream, broccoli, and cheese		
70	Where's Mikey?		
71	~~Have you ever seen a dragonfly?~~		
72	~~I need plastic, shoes, glasses, and shorts~~		

Annexe 3 - Items du test de discrimination phonémique

Les items barrés n'ont pas été conservés dans la version finale du test.

Num	stimuli	clé	voix
1	bard bard bud	3	F
2	cup carp carp	1	M
3	bark back back	1	F
4	bit bit beat	3	F
5	barn barn ban	3	M
6	had had hud	3	F
7	look look Luke	3	M
8	chip cheap cheap	1	F
9	hut heart hut	2	F
10	card cad cad	1	M
11	~~bick beak bick~~	2	M
12	luck lack luck	2	M
13	cess cess sace	3	F/ M
14	park puck park	2	M
15	pull pool pull	2	F
16	mutt mutt mat	3	F
17	tate tet tet	1	F/ M
18	deed did deed	2	M
19	should shooed should	2	F
20	tat tut tat	2	M
21	~~full fool fool~~	1	M
22	fell fell fail	3	F/M
23	hard had hard	2	F
24	pate pate pet	3	F/M
25	thin thin zin	3	F/M
26	~~elith cliff cliff~~	1	F/M
27	loathe loathe loze	3	F/M
28	uzzer other other	1	M
29	myth ms. ms.	1	F
30	mouth mouse mouth	2	F/M
31	theme seam theme	2	F/M
32	truce truth truce	2	F/M
33	faith face face	1	F/M
34	fourth force force	1	F/M
35	thigh thigh sigh	3	F/M

Annexe 4 - Items du test de reconnaissance aurale du lexique

Les items barrés n'ont pas été conservés dans la version finale du test.

num	stimulus
1	generic (entraînement)
2	platery (entraînement)
3	denial
4	muscle
5	creature
6	mensible
7	~~scornful~~
8	~~stoutly~~
9	~~ablaze~~
10	kermshaw
11	moonlit
12	lofty
13	hurricane
14	flaw
15	~~alberation~~
16	unkempt
17	breeding
18	festivity
19	screech
20	~~savoury~~
21	~~plaudate~~
22	shin
23	fluid
24	spaunch
25	allied
26	slain
27	recipient
28	exprate
29	eloquence
30	cleanliness
31	dispatch
32	rebondicate
33	~~ingenious~~
34	bewitch
35	~~skave~~
36	plaintively
37	~~kilp~~
38	~~interfate~~
39	hasty
40	lengthy

num	stimulus
41	fray
42	~~crumper~~
43	upkeep
44	majestic
45	~~magrity~~
46	nourishment
47	~~affordish~~
48	~~proom~~
49	turmoil
50	carbohydrate
51	scholar
52	turtle
53	~~fellick~~
54	destription
55	~~awry~~
56	censorship
57	celestial
58	rascal
59	~~purrage~~
60	~~pulsh~~
61	muddy
62	quirty
63	~~pudour~~
64	~~listless~~
65	wrought

Annexe 5 - Items du PVST (*Phrasal Vocabulary Size Test*)

Les numéros manquants correspondent aux items enlevés par rapport au test de Martinez (2011), cf Annexe 8.

First 1000

1	go on: It will **go on**.	
	a.	sleep
	b.	repeat
	c.	be fast
	d.	continue

2.	lead to: No one knows what it will **lead to.**	
	a.	want
	b.	have inside
	c.	cause in the future
	d.	find

3.	so that: He sat **so that** they could do it.	
	a.	to make it possible that
	b.	because
	c.	very slowly and then
	d.	before

4	at all: I don't like it **at all**.	
	a.	all the time
	b.	in any way
	c.	at first
	d.	sometimes

5.	I mean: Two, **I mean**, three.	
	a.	I am guessing
	b.	maybe
	c.	then later
	d.	I correct myself

6.	at least: **At least** it is warm.	
	a.	other things may be bad, but
	b.	many days have passed and now
	c.	I cannot believe that
	d.	the least important thing is

7	is likely to: He **is likely to** go.	
	a.	likes to
	b.	can
	c.	wants to
	d.	probably will

9	deal with: I can **deal with** it.	
	a.	fix
	b.	remember
	c.	find
	d.	see

10	used to: I **used to** go.	
	a.	want to travel now
	b.	went there in the past
	c.	usually go there
	d.	always travel there

Second 1000

1	so far: It's good **so far**.	
	a.	until now
	b.	but not really
	c.	sometimes
	d.	from a distance

2	to do with: It is **to do with** money.	
	a.	making
	b.	for
	c.	about
	d.	our

3	take over: They will **take over**.	
	a.	be finished
	b.	have control
	c.	come later
	d.	think about it

4.	in particular: I want that **in particular**.	
	a.	especially
	b.	in private
	c.	because it is different
	d.	maybe

5.	for instance: **For instance**, it is cheaper.	
	a.	maybe
	b.	for a short time
	c.	in my opinion
	d.	as an example

7	as soon as: I'll go **as soon as** I can.	
	a.	from the moment
	b.	only if
	c.	after
	d.	before

9.	be about to: I **am about to** read the newspaper.	
	a.	cannot wait to
	b.	am soon going to
	c.	really like to
	d.	am trying to

10.	be expected to: We **are expected to** do it.	
	a.	are waiting
	b.	hoping to
	c.	must
	d.	are able to

Third 1000

1.	give up: I **give up**.	
	a.	try very hard
	b.	am starting
	c.	will now stop
	d.	exercise

2	feel like: I just did not **feel like** it.	
	a.	love
	b.	want to do
	c.	think about
	d.	try to do

3	turn out: It **turned out** different.	
	a.	started
	b.	seemed
	c.	became
	d.	did not look

4	other than: **Other than** that, it's good.	
	a.	not including
	b.	if you include
	c.	because of
	d.	after

6	all over: It is **all over** the bed.	
	a.	covering
	b.	inside
	c.	on top of
	d.	beside

7.	in touch: Keep **in touch**.	
	a.	feeling it
	b.	communicating
	c.	pushing it
	d.	thinking

9.	at once: I did it **at once**.	
	a.	one time
	b.	many times
	c.	early
	d.	immediately

10	in time: **In time** they bought a house.	
	a.	quickly
	b.	earlier
	c.	eventually
	d.	recently

Fourth 1000

1.	prove to be: It has **proved to be** important.	
	a.	possibly become
	b.	shown itself to be
	c.	continued to be
	d.	never been

2.	next door: It's just **next door**.	
	a.	coming soon
	b.	common
	c.	perfect
	d.	very close

3.	run out: I think we **ran out** of it.	
	a.	had no more
	b.	were bored
	c.	thought
	d.	moved outside

6.	in light of: It was accepted **in light of** the money.	
	a.	despite
	b.	because of
	c.	in addition to
	d.	instead of

7.	by no means: He is **by no means** rich.	
	a.	very
	b.	not at all
	c.	more or less
	d.	considered

8.	come across: They **came across** a hotel.	
	a.	stayed in
	b.	opened
	c.	were near
	d.	found

9.	happen to: She **happened to** call.	
	a.	pretended
	b.	tried hard to
	c.	did not want to
	d.	by chance did

10.	even so: **Even so** it's better.	
	a.	despite that
	b.	that way
	c.	it is the same and
	d.	maybe

Fifth 1000

1.	by far: She is **by far** the most intelligent.	
	a.	trying to be
	b.	not at all
	c.	really
	d.	sometimes

3.	straight away: They did it **straight away**.	
	a.	immediately
	b.	the correct way
	c.	slowly
	d.	because they wanted to

5.	turn down: She **turned down** the money.	
	a.	hid
	b.	lost
	c.	made
	d.	refused

6.	to blame: We are not **to blame**.	
	a.	in total agreement
	b.	interested
	c.	accusing anyone
	d.	the cause of the problem

7.	take for granted: She **took** it **for granted**.	
	a.	kept it
	b.	did not give it importance
	c.	wanted it a lot
	d.	thought about it carefully

8.	as of: It changes **as of** today.	
	a.	starting
	b.	sometime
	c.	perhaps
	d.	because of

9.	can tell: You **can tell**.	
	a.	may speak
	b.	are smart
	c.	can see
	d.	might

Annexe 6 - Items du test de jugement de grammaticalité aurale

Les items barrés n'ont pas été conservés dans la version finale du test.

num	stimulus	clé (1= grammatical, 0= non)
1	I took my coat off	1
2	They worried about him drinking	1
3	I like very much sweets	0
4	A friend of mine came	1
5	~~This stroke me as important~~	0
6	Sorry, I forget to send it	0
7	She's name was Anna	0
8	~~He helped her making the cake~~	0
9	Guess where it is	1
10	I don't have any car	0
11	~~I saw them drive away~~	1
12	~~From what says Joe, it's not true~~	0
13	I haven't the keys	0
14	That city seems very large	1
15	~~This children are lost~~	0
18	There are differents possibilities	0
17	I spended last week end in London	0
16	When was written this book	0
19	~~This tent sleeps 4.~~	1
20	She's one of my friend	0
21	I 'm agree with you	0
22	~~This game is sure to be a winner~~	1
23	What happens next ?	1
24	The idea that they might win is absurd	1
25	I want that you stay	0
26	Smoking is known to cause cancer	1
27	I have been to Rome	1
28	The professor I gave the book to has left	1
29	What were you thinking	1
31	~~I have seen it last year~~	0
30	Before doing anything you need to think	1
32	~~Every person is important~~	1
33	~~The crash arrived because it was dark~~	0
34	~~They wanted the children found~~	1
35	Have a good travel!	0
36	It's depend on the situation	0
37	We may go sightseeing later	1
38	I explain you the situation	0
39	I am boring in class	0
40	Those dogs are trained	1
41	Most people think so	1
42	It's the house who's on the other side	0
43	It's amazing to think that it's true	1
44	~~She is here since 2 years now~~	0
45	I was born early	1

Annexe 7 - Script *R* : exemple du script d'analyse de variance, des corrélations, etc. (Chap.10)

Les lignes commençant par # sont des commentaires ; les autres sont les lignes de commande.

```
# ANALYSE DES RELATIONS ENTRE TESTS DIAGNOSTIQUES ET CO

# vider l'espace de travail
rm(list = ls())

# chargement des bibliothèques de fonctions utilisées
library(tidyverse)
library(psych)
# pour fonctions de stats (eta carré, etc.)
library (sjstats)
# pour comparaisons multiples (Tukey, etc.)
library(multcomp)
#graphiques pour corrélations
library (GGally)

# importer les données du fichier csv dans R ; prise en compte des cases vides
foo <- read.csv2("sessions mar2020.csv", na.strings=c("","NA"))
# donner une allure plus élégante au tableau de données (tidyverse)
foo <- as_tibble(foo)
# examiner la structure du tableau de données et le résumé des données (min, max, etc. par variable)
str(foo)
summary(foo)


##############################STATISTIQUES DESCRIPTIVES#########################
#description de toutes les variables numériques (sans l'asymétrie) de tout l'échantillon, puis par niveau, puis par filière
describe(foo[,3:15], skew=FALSE)
describe(foo[foo$self_CO == "A2",3:15], skew=FALSE)
describe(foo[foo$self_CO == "B1",3:15], skew=FALSE)
describe(foo[foo$self_CO == "B2",3:15], skew=FALSE)
describe(foo[foo$self_CO == "C1",3:15], skew=FALSE)
describe(foo[foo$filiere=="LLCER",3:15], skew=FALSE)
describe(foo[foo$filiere=="SDL",3:15], skew=FALSE)

#effectifs par niveau de CO, tout l'échantillon puis par filière
summary (foo$self_CO)
summary (foo[foo$filiere=="LLCER",]$self_CO)
summary (foo[foo$filiere=="SDL",]$self_CO)
summary(foo$filiere)

############################## BOITES A MOUSTACHE et ANOVAS ######################
#PROSO
# graphique en boîte à moustaches (PROSO par niveau de CO) avec superposition des points
ggplot(foo, aes(x=self_CO, y=proso_56it_2020)) +
  geom_boxplot(outlier.shape=NA) + #avoid plotting outliers twice
  geom_jitter(position=position_jitter(width=.1, height=0))

#corrélation avec Kendall's tau
cor.test(foo$self_CO_num, foo$proso_56it_2020, method="kendall")
```

```
#ANOVA a 1 facteur (variable factorielle), avec fonction lm
modanova_proso<-lm(proso_56it_2020~self_CO, data=foo)
summary(modanova_proso)
eta_sq(modanova_proso)

#vérification des conditions d'utilisation d'anova: assez normal ?
hist(resid(modanova_proso))
shapiro.test(resid(modanova_proso))

#comparaisons 2 à 2 avec Tukey (comp multiples)
mc_tukey_proso <- glht(modanova_proso, linfct=mcp(self_CO="Tukey"))
summary(mc_tukey_proso)

## idem avec PHON, AURLEX, PVST, AURGRAM


##################################CORRELATIONS#####################################

#recoder niveaux CECR en échelle numérique et vérifier le recodage
codes_CO <- c(A1=1,A2=2,B1=3,B2=4,C1=5)
foo$self_CO_num <- sapply(foo$self_CO,function(x)codes_CO[x])
table(foo$self_CO,foo$self_CO_num,deparse.level = 2,useNA = "always")

#matrice de corrélations entre toutes variables (GGally)
var1 <- c("self_CO_num","phon32it_2020","proso_56it_2020","aurl_2020","aurgram_33it_2020",
"pvst_ttn")
ggpairs(foo[,var1], columnLabels = c("SELF_CO","PHON","PROSO", "AURLEX", "AURGRAM",
"PVST"))


###################REGRESSION LOGISTIQUE BINAIRE MULTIPLE###################
#REGRESSION LOGISTIQUE avec niveau insuffisant (A2/B1), suffisant (B2/C1)

#créer une nouvelle variable binaire score.CO.b
foo$Score.CO.b <- ifelse(foo$self_CO_num>2 , 1, 0)
table(foo$self_CO,foo$Score.CO.b, deparse.level = 2)

#calcul du modèle de régression logistique avec toutes variables explicatives
mod2 <- glm(Score.CO.b~aurl_2020+aurgram_33it_2020+pvst_ttn+phon32it_2020+
proso_56it_2020, data=foo, family = "binomial")
summary(mod2)

#calculer un pseudo R2 (MacFadden's pseudo R2): taille d'effet
ll.null2<-mod2$null.deviance/-2
ll.proposed2<-mod2$deviance/-2
(ll.null2-ll.proposed2)/ll.null2

#calcul du p associé
1 - pchisq(2*(ll.proposed2 - ll.null2), df=(length(mod2$coefficients)-1))

# graphique de comparaison entre valeurs prédites et constatées
#avant de faire graphique, enlever lignes avec NA:
foo_comp <- foo [!((is.na(foo$proso_56it_2020)) | (is.na(foo$aurl_2020)) |
(is.na(foo$aurgram_33it_2020)) | (is.na(foo$pvst_ttn)) |(is.na(foo$phon32it_2020))|
(is.na(foo$Score.CO.b))),]
```

```
#mettre d'abord toutes les observations dans l'ordre de probabilité prédite
predicted.data <- data.frame (pba.of.b2=mod2$fitted.values, b2=foo_comp$Score.CO.b)
predicted.data <- predicted.data[
  order(predicted.data$pba.of.b2, decreasing=FALSE),]
predicted.data$rank <- 1:nrow(predicted.data)
# graphique
ggplot(data=predicted.data, aes(x=rank, y=pba.of.b2)) +
  geom_point(aes(color=b2), alpha=1, shape=4, stroke=2) +
  xlab("Rang") +
  ylab("Proba. prédite d'obtenir B2 ou +")

#FIN
```

Annexe 8 – Résultats de l'analyse du PVST (groupe pilote mars 2017)

Le logiciel Winsteps a été utilisé pour ces calculs. Les items en gras sont les 10 items moins performants (soit trop faciles, soit pas suffisamment discriminants) qui ont été enlevés pour créer une version moins longue du PVST (40 items au lieu de 50).

num	nom	n	difficulté	discrim.
1	PVST100001	61	0,93443	0,3477
2	PVST100002	61	0,81967	0,5887
3	PVST100003	61	0,83607	0,4834
4	PVST100004	61	0,77049	0,5179
5	PVST100005	61	0,83607	0,5825
6	PVST100006	61	0,72131	0,5518
7	PVST100007	61	0,78689	0,5226
8	**PVST100008**	**61**	**0,7541**	**0,2375**
9	PVST100009	61	0,70492	0,468
10	PVST100010	61	0,60656	0,5206
11	PVST200001	61	0,70492	0,6346
12	PVST200002	61	0,7541	0,4288
13	PVST200003	61	0,60656	0,6204
14	PVST200004	61	0,95082	0,3566
15	PVST200005	61	0,72131	0,5513
16	**PVST200006**	**61**	**0,70492**	**0,2729**
17	PVST200007	61	0,62295	0,3784
18	**PVST200008**	**61**	**0,55738**	**0,2013**
19	PVST200009	61	0,83607	0,4763
20	PVST200010	61	0,32787	0,5659
21	PVST300001	61	0,85246	0,4759
22	PVST300002	61	0,70492	0,552
23	PVST300003	61	0,72131	0,5391
24	PVST300004	61	0,88525	0,35
25	**PVST300005**	**61**	**0,67213**	**0,2631**
26	PVST300006	61	0,81967	0,5791
27	PVST300007	61	0,72131	0,7154
28	**PVST300008**	**61**	**0,47541**	**0,3415**
29	PVST300009	61	0,21311	0,4329
30	PVST300010	61	0,22951	0,5049
31	PVST400001	61	0,81967	0,5159
32	PVST400002	61	0,81967	0,4947
33	PVST400003	61	0,63934	0,7369
34	**PVST400004**	**61**	**1**	**0**
35	**PVST400005**	**61**	**0,72131**	**0,1469**
36	PVST400006	61	0,44262	0,396
37	PVST400007	61	0,60656	0,4179
38	PVST400008	61	0,2623	0,5417
39	PVST400009	61	0,62295	0,6527
40	PVST400010	61	0,47541	0,3779
41	PVST500001	61	0,90164	0,3837
42	**PVST500002**	**61**	**0,98361**	**0,1459**
43	PVST500003	61	0,7377	0,6219
44	**PVST500004**	**61**	**0,7377**	**0,17**
45	PVST500005	61	0,57377	0,5362
46	PVST500006	61	0,65574	0,4704
47	PVST500007	61	0,40984	0,6037
48	PVST500008	61	0,52459	0,6227
49	PVST500009	61	0,47541	0,4605
50	**PVST500010**	**61**	**0,19672**	**0,3127**

Annexe 9 - Résultats des modèles de régression logistique binaire simple pour la détermination des scores de césure des tests PHON, PROSO, AURLEX, PVST et AURGRAM

Le score de césure correspond à la bascule B1/B2 dans le modèle logistique binaire, et donc une probabilité de 0,5 d'être classé par le modèle en B2 et plus en compréhension de l'oral. L'équation associée à une régression logistique binaire simple est :

ax+b = log (pba(y)/pba(1-y)), où

- « x » est la variable explicative (pour nous, PHON, PROSO, AURLEX, PVST et AURGRAM selon le modèle),
- « y » est la variable à expliquer (niveau suffisant en compréhension de l'oral, SELF_CO_b),
- « a » est le coefficient associé à la variable explicative,
- « b » est l'odonnée à l'origine (*intercept*).

Ici, on cherche la valeur de x qui correspond à une probabilité de 0,5 d'être classé en SELF_CO_b « satisfaisant », donc ax + b = log (0,5/0,5) = log (1) = 0. Le score de césure que nous recherchons aura donc à chaque fois la valeur x = -b/a. Nous arrondissons ensuite ce résultat à l'entier inférieur.

Test PHON

```
glm(formula = Score.CO.b ~ phon32it_2020, family = "binomial",
    data = foo)
Deviance Residuals:
   Min      1Q  Median      3Q     Max
-1.427  -1.146  -0.637   1.101   2.120
Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)      -3.16645    1.03724  -3.053  0.00227 **
phon32it_2020  0.12882    0.04316   2.985  0.00284 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 208.53  on 150  degrees of freedom
Residual deviance: 198.11  on 149  degrees of freedom
  (38 observations deleted due to missingness)
AIC: 202.11
Number of Fisher Scoring iterations: 4
```

score de césure = 3.16645/0.12882 = 24.6

Test PROSO

```
glm(formula = Score.CO.b ~ proso_56it_2020, family = "binomial",
    data = foo)
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1767  -1.0761   0.5639   0.9730   1.6991
Coefficients:
                  Estimate  Std. Error  z value Pr(>|z|)
(Intercept)        3.85360    0.87433   4.408   1.05e-05 ***
proso_56it_2020  -0.09311    0.02132  -4.366   1.26e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 204.20  on 147  degrees of freedom
Residual deviance: 180.82  on 146  degrees of freedom
  (41 observations deleted due to missingness)
AIC: 184.82
Number of Fisher Scoring iterations: 4
```

score de césure = 3.85360/0.09311 = 41.4

Test AURLEX

```
glm(formula = Score.CO.b ~ aurl_2020, family = "binomial", data = foo)
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8539  -0.9466   0.3249   0.8959   2.5177
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)    6.16787    1.09651   5.625 1.85e-08 ***
aurl_2020   -0.23236    0.04142  -5.610 2.02e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 208.53  on 150  degrees of freedom
Residual deviance: 157.85  on 149  degrees of freedom
  (38 observations deleted due to missingness)
AIC: 161.85
Number of Fisher Scoring iterations: 4
```

score de césure = 6.16787/0.23236 = 26.5444

Test PVST

```
glm(formula = Score.CO.b ~ pvst_ttn, family = "binomial", data = foo)
Deviance Residuals:
   Min      1Q  Median      3Q     Max
-2.4399  -0.6862   0.2402   0.5767   2.1202
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   5.35310    0.88367   6.058 1.38e-09 ***
pvst_ttn     -0.20241    0.03184  -6.356 2.07e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Dispersion parameter for binomial family taken to be 1)
   Null deviance: 200.45  on 144  degrees of freedom
Residual deviance: 128.56  on 143  degrees of freedom
  (44 observations deleted due to missingness)
AIC: 132.56
Number of Fisher Scoring iterations: 5
```

score de césure = 5.35310/0.20241 = 26.4468

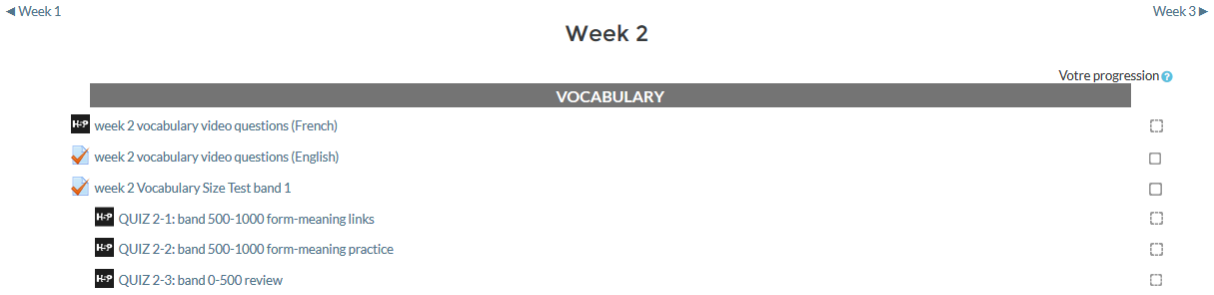
Test AURGRAM

```
glm(formula = Score.CO.b ~ aurgram_33it_2020, family = "binomial",
   data = foo)
Deviance Residuals:
   Min      1Q  Median      3Q     Max
-2.3771  -0.7311   0.3048   0.6776   2.3085
Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)       5.86527    0.93298   6.287 3.25e-10 ***
aurgram_33it_2020  -0.28193    0.04444  -6.344 2.24e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Dispersion parameter for binomial family taken to be 1)
   Null deviance: 209.77  on 151  degrees of freedom
Residual deviance: 141.66  on 150  degrees of freedom
  (37 observations deleted due to missingness)
AIC: 145.66
Number of Fisher Scoring iterations: 5
```

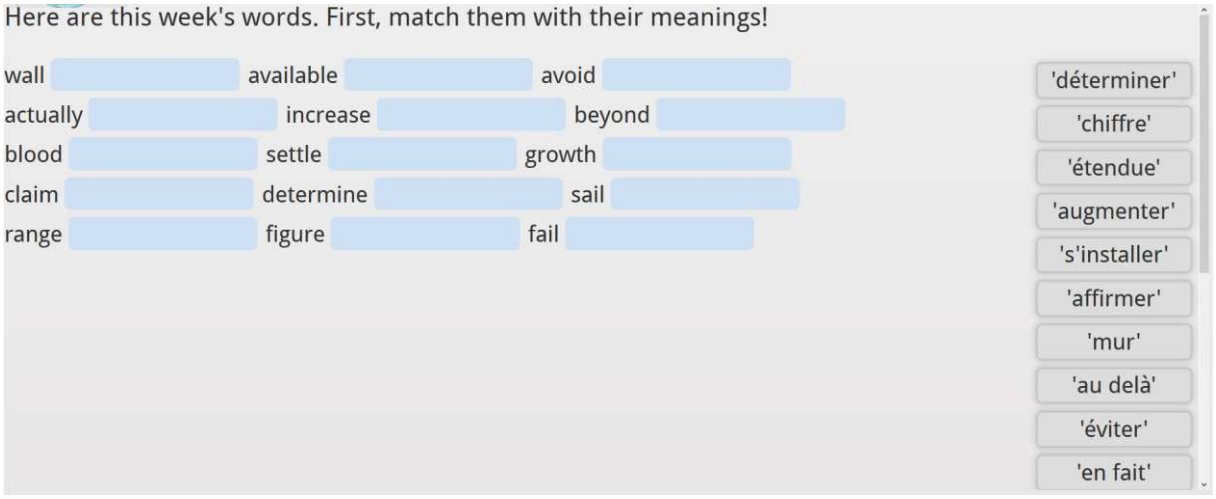
score de césure = 5.86527/0.28193 = 20.8

Annexe 10 - Parcours de remédiation

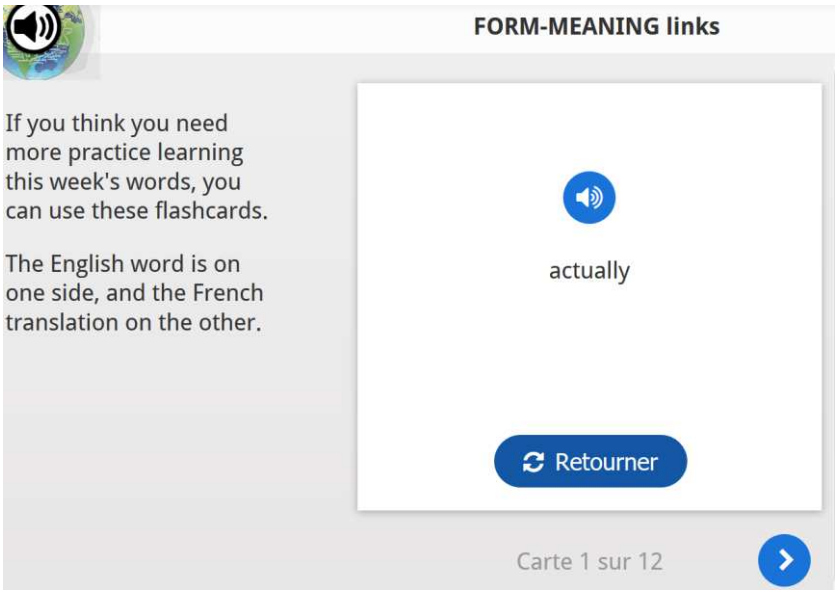
1. Structure d'une semaine du parcours de remédiation :



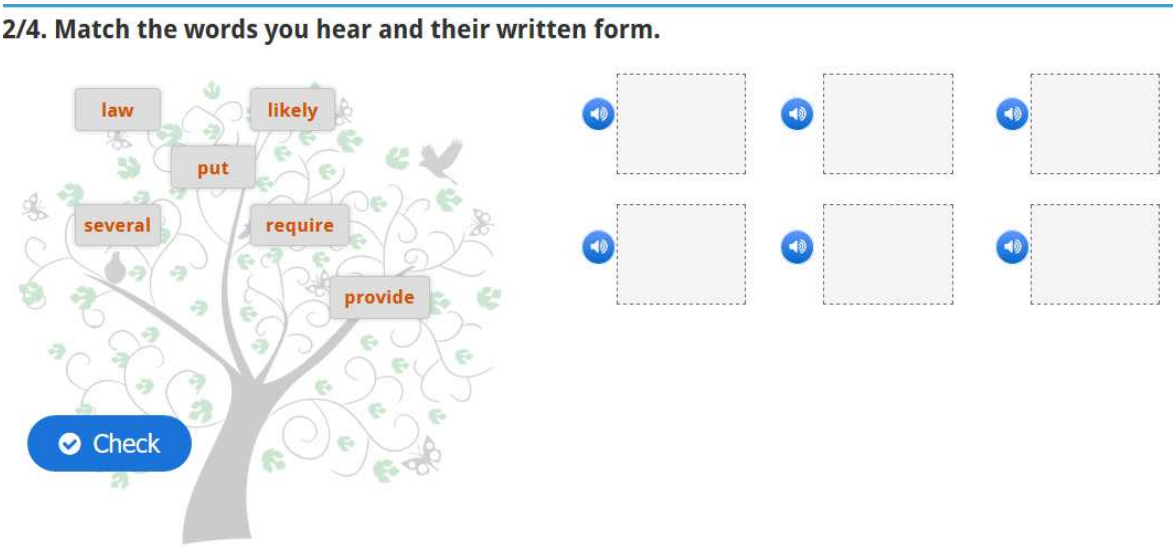
2. Exemple de tâche d'appariement forme écrite – traduction L1 (écran partiel) :



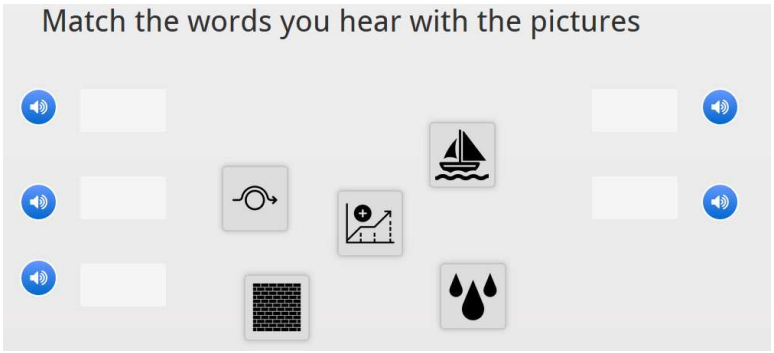
3. Exemple de tâche d'appariement forme orale-traduction L1



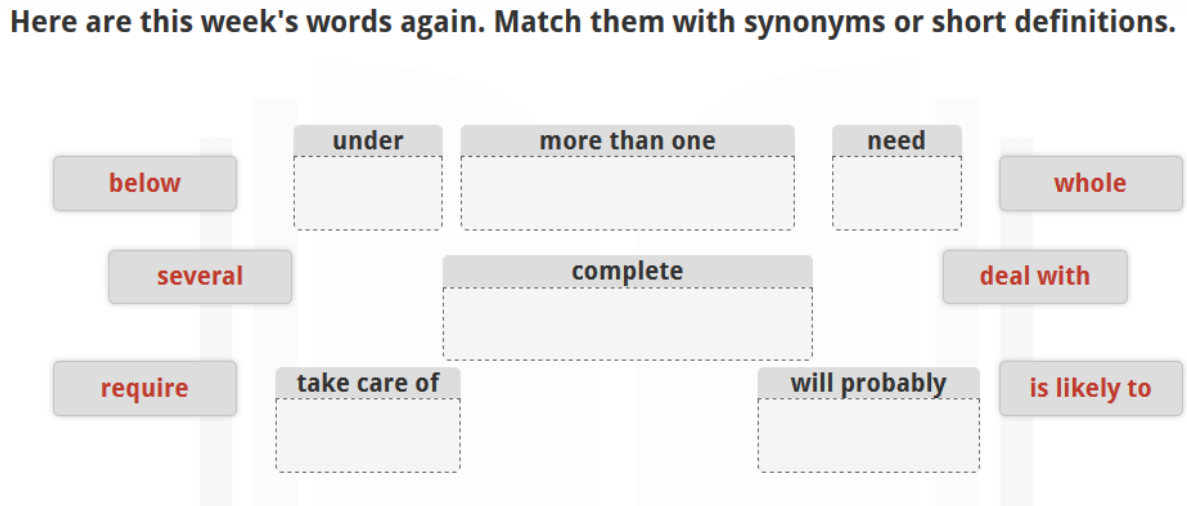
4. Exemple de tâche d'appariement forme écrite- forme orale :



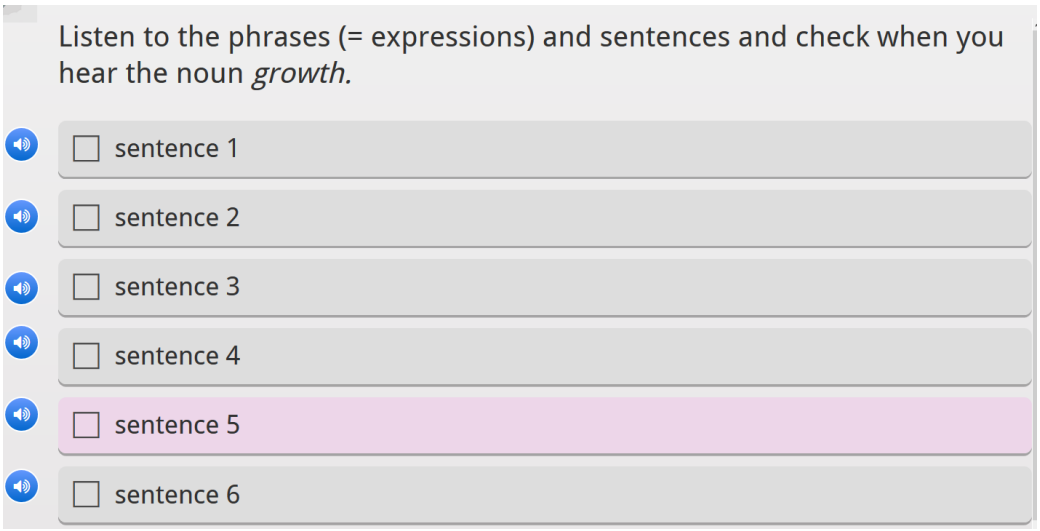
5. Exemple de tâche d'appariement forme orale- image représentant le sens :



6. Exemple de tâche d'appariement forme écrite- définition/ synonyme en anglais :



7. Exemple de tâche de repérage du mot dans une phrase (certaines phrases contiennent *growth*, d'autres *gross*) :



8. Exemple de production de la forme écrite à partir d'un contexte oral (le mot *claim* a été rendu inaudible dans les extraits proposés) :