UCLouvain

FIAL
Un éventail de formations, une multitude d'horizons

Faculté de philosophie, arts et lettres

EFLSemLex : une ressource lexicale graduée et désambiguïsée pour les apprenants de l'anglais en L2

Enrichir les ressources lexicales pour l'apprentissage des langues étrangères par la désambiguïsation lexicale

Auteur : LANNOO David
Promoteur : Pr. FRANÇOIS Thomas
Lecteur : Dr. ALFTER David
Année académique 2021 – 2022
Master [120] en linguistique – finalité : traitement automatique du langage

« Bien que je ne sois pas ingrat je n'ai pas envie de vous dire merci, parce qu'au fond, ce que j'ai, ici, je l'ai conquis »
– Kery James

Mais puisqu'aucune conquête ne se fait seul, j'aimerais tout d'abord remercier mon promoteur, le Professeur Thomas François : sans son esprit scientifique rigoureux et toujours intransigeant, sans ses commentaires précis et toujours pertinents, et – enfin – sans son soutien inflexible et toujours bienveillant, ce mémoire n'aurait tout simplement jamais (permettons-nous ici la classique personnification) vu la lumière du jour.

Dans la foulée, je remercierai l'équipe du CENTAL, qui m'a ouvert les bras il y a maintenant deux étés pour mon stage et qui m'a offert un aperçu du monde de la recherche. Merci en particulier à Mme. Eva Rolin, pour les moments de partage qui m'ont montré que cet univers ne correspond pas toujours à l'idée que j'aie pu m'en faire. Je ne peux continuer sans évoquer les remarques carrées du Docteur Patrick Watrin qui, malgré lui, s'est imposé comme un Némésis auquel j'ai eu à cœur de montrer que, quels que fussent mes choix, l'adjectif « fainéant » ne me correspondait sûrement pas. Je remercierai aussi le Docteur David Alfter, pour sa relecture attentive de mon mémoire et les éclaircissements prodigués lors de notre entretien.

À Lulu, pour l'héritage humain qu'elle m'a laissé, et tout ce qu'elle prodigue avec passion à ses élèves d'année en année – toujours fidèle à elle-même et à son *freestyle*.

Au professeur Philippe Hambye, qui m'a sévèrement réappris à écrire et a su, dès son tout premier cours, me transmettre sa passion, son amour pour la linguistique, j'adresse ma plus profonde gratitude.

Je pense également à mes parents et ma grand-mère, Véronique, Yannick & Blanche, qui, loin de la science mais proche du cœur (et des cordons de la bourse), m'ont fermement encouragé durant toute la durée de mes études de langues modernes puis de traitement automatique du langage.

À Gaëlle, qui me fait part depuis des mois de sa volonté d'être LA relectrice officieuse de mon mémoire, je réservais une mention zeugmatique « merci pour tes remarques éclairantes, et surtout d'avoir été là tout du long ». Promesse non tenue *vs*. promesse tenue, je m'en tiendrai à « merci d'avoir toujours été là et pour l'écoute de mes complaintes ».

Mes derniers remerciements vont à Masashi Kishimoto, dont les personnages déterminés et jusqu'au-boutistes m'ont prouvé qu'une féroce volonté pouvait venir à bout de tous les obstacles. Ce mémoire marque la fin d'une période de transition parfois difficile, et sa réalisation aurait sans l'ombre d'un doute été plus morose sans le sourire fictionnel de Naruto.

Table des matières

Introduction

La sagesse surréaliste du peintre belge René Magritte, souvent résumée – pour le grand public – par la célèbre légende « Ceci n'est pas une pipe » (*La Trahison des Images*, 1929), découle d'une profonde réflexion sur le langage. Cette réflexion, entamée par le peintre à partir de l'année 1926, aspire à saisir les rapports entre les mots et les représentations. Ni le mot « pipe », ni le dessin d'une pipe, ne coïncident avec la chose, l'objet qu'est la « pipe ». Le mot sur la page, ou la peinture sur la toile, activent, dans l'esprit du locuteur ou de l'observateur, une représentation mentale de la chose, un concept, stocké quelque part dans la mémoire de ce dernier.

Bien entendu, cette idée de la « pipe » pointe du doigt vers le réel, faute de quoi le langage ne serait qu'abstraction. Dans le sens contraire, le monde matériel peut aussi appeler le langage. Pensons, par exemple, au nourrisson qui, voyant son grand-père fumer, articulerait fièrement : « Pipe !! »

Le phénomène ici décrit a été mieux formalisé en 1923, sous la forme d'un triangle (Ogden & Richards, 1923). Ce triangle est dit « sémantique », parce qu'il modélise le sens. La première pointe correspond au *symbole*, ou *signifiant* : il s'agit d'une forme acoustique, écrite ou visuelle (\pip\, ***p i p e***, ou encore  , si l'on sort du champ strict de la linguistique). La seconde pointe représente le *concept*, l'idée qu'un locuteur peut se faire ; il s'agit de ce que Saussure, fondateur de la linguistique moderne, nomme *signifié*. Ces deux dimensions du triangle appartiennent purement au domaine du langage. En particulier, la conjugaison de ces deux aspects incarne le *signe* saussurien, soit l'unité d'expression de base au cœur du langage. La troisième pointe équivaut à la *chose*, l'objet, l'action, l'émotion, etc. à laquelle le *signe* renvoie symboliquement. Le lien entre le *symbole* et la *chose* est indirect : « ceci n'est pas une pipe, mais ceci *représente* une pipe » (Figure 1)[1].

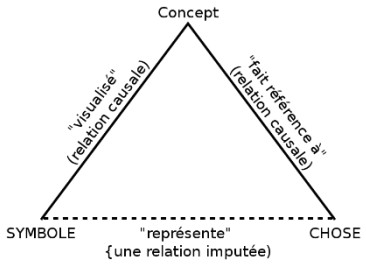


Figure 1 - le triangle sémantique d'Ogden & Richards (1923)

Ce schéma présente l'avantage de donner clairement à voir les trois parties de la signification. La forme triangulaire illustre bien que la langue ne se limite pas à une simple nomenclature : les *mots* ne sont pas des étiquettes que l'on peut directement appliquer sur les *choses*.

Cependant, au grand dam des enfants, des apprenants, et des spécialistes en traitement automatique du langage – nous le verrons – ce triangle théorique s'érode assez rapidement lorsqu'il est confronté à la réalité de la langue. En effet, le caractère clos de la figure 1 laisse supposer l'existence d'un lien unique associant un *symbole*, un *concept* et une *chose*.

[1] Pour faciliter la navigation du lecteur, les renvois aux figures, tableaux et sections du texte sont cliquables (☜).

D'emblée, la limite de la représentation affleure : un même *symbole* peut permettre de *visualiser* plusieurs *concepts*, qui *font eux-mêmes référence* à des *choses* diverses. À ce titre, l'exemple de la « pipe » est assez évocateur. Les lettres ***p i p e*** feront tantôt appel à l'idée d'un objet oblong utilisé pour fumer, tantôt – dans le registre trivial – au concept de stimulation orale du sexe masculin. Pour chacune de ces images mentales, la réalité peut prendre des formes différentes : en tant qu'acte ou en tant qu'objet, aucune pipe n'est identique. Au demeurant, le *signifié* de la forme ***pipe*** prend une coloration supplémentaire encore dans la locution populaire « Casser sa pipe » (mourir)[2].

Ainsi, confronté à l'usage effectif, le triangle éclate rapidement ; il faut alors plutôt l'envisager comme une structure en arbre, dont la racine serait le *symbole*, les branches seraient les *concepts,* et enfin les feuilles seraient les *choses*. Là encore, la vision arborescente ne se suffit pas à elle-même, et les branches peuvent, en sens inverse, mener à des racines différentes. En d'autres termes, un même *concept* (*e.g.* acte sexuel oral) peut aussi être activé par plusieurs *symboles* différents (*e.g.* ***pipe***, ***fellation***) : les arbres sont interconnectés. Pour résumer, la multiplicité des associations entre symboles, concepts et choses peine à être réduite à une forme triangulaire : le tableau est, dans les faits, bien plus complexe.

Ce bref détour par les théories linguistiques de la première heure nous aura permis de soulever le point central de notre travail. Pour en faire part simplement, les mots ont plusieurs sens ; il s'agit d'une caractéristique inhérente à toutes les langues. Bien qu'ayant des allures de lapalissade, cette assertion pose, à ce jour, encore de nombreux problèmes et questionnements pour le traitement automatique et l'apprentissage des langues.

Dans le traitement automatique du langage (ci-après TAL ou NLP[3]), la possibilité de créer une machine capable de déterminer correctement le sens des mots a d'abord été rejetée en bloc, mais, au fil du temps, de nombreux auteurs se sont penchés sur la question, élaborant des systèmes avec les ressources et les technologies dont ils disposaient. Ces systèmes sont dits de « désambiguïsation lexicale ». À l'heure actuelle, la question de la désambiguïsation lexicale automatique n'est toujours pas résolue, mais des algorithmes de plus en plus efficaces voient le jour, si bien que les facultés sémantiques humaines pourraient, à l'avenir, se voir dépassées par les ordinateurs.

La seconde composante essentielle de notre travail concerne les listes de vocabulaire. Il ne paraît pas exagéré de supposer que toute personne ayant bénéficié au moins d'un enseignement primaire, ou ayant suivi des cours de langue étrangère, soit plus ou moins familière avec ces listes de mots à connaître. Ces dernières peuvent également jouer un rôle clé dans d'autres domaines que l'apprentissage : ainsi, les ressources lexicales peuvent être instrumentalisées par la psycholinguistique, ou le NLP, par exemple.

Comment ces listes – ou ressources lexicales – sont-elles construites ? Sur quels critères sont-elles fondées ? Comment une ressource lexicale peut-elle discriminer le vocabulaire important du vocabulaire secondaire ? Ce genre de questions sera discuté dans les chapitres à venir.

---

[2] La discussion de ce cas nous emmènerait malheureusement trop loin du sujet de notre travail.
[3] *Natural language processing.*

En particulier, nous chercherons à savoir comment les systèmes de désambiguïsation lexicale permettraient d'enrichir les ressources de vocabulaire destinées aux apprenants. En quoi le fait de déterminer le sens des mots permettrait-il d'avoir un meilleur aperçu de ce qui rend une langue – spécifiquement les mots d'une langue – complexes ou difficiles ?

Récemment, de nombreuses ressources lexicales graduées ont permis de mieux investiguer la question de la complexité lexicale des L1 ou des L2. Une ressource graduée consiste en un inventaire de mots pour lesquels sont calculées des distributions de fréquence dans des matériaux pédagogiques authentiques et sur chacun des niveaux, des grades, de l'apprentissage de la langue. Ces ressources graduées peuvent être crées dans le but d'assister les instituteurs pour l'enseignement d'une langue maternelle (dans les écoles primaires), ou être destinés aux apprenants d'une langue étrangère.

À cet effet, l'échelle la plus connue est le Cadre européen commun de référence pour les langues (CECRL ou CEFR[4] - Conseil de l'Europe, 2001). Il s'agit d'un cadre général qui fournit une description exhaustive des types de discours (écrits ou oraux) qu'un apprenant devrait pouvoir produire et comprendre à chaque stade de compétence.

Au départ de cette échelle, le projet CEFRLex a établi des ressources lexicales graduées pour plusieurs langues européennes. Ces ressources sont, pour la plupart, basées sur un corpus de tâches de compréhension à la lecture collectées dans des manuels classifiés selon le CEFR. D'autres s'appuient sur des productions écrites d'apprenants. Par conséquent, les inventaires produits par le projet CEFRLex sont révélateurs de ce qui devrait être compris ou produit par les apprenants d'une langue seconde à chaque niveau de compétence.

La ressource CEFRLex pour le néerlandais – NT2LEX (Tack et al., 2018) – a mis en exergue la possibilité de développer des ressources lexicales graduées *et* désambiguïsées. Les auteurs de NT2Lex soulignent l'importance de prendre en compte la variété sémantique dans les ressources lexicales. Il semble, en effet, pour reprendre l'exemple précédent, que les deux sens du mot ***pipe*** ne sont pas supposés être connus et produits au même stade de l'apprentissage par les apprenants. Les autres ressources du projet CEFRLex – et les ressources lexicales en général – ne prennent pas en compte cet aspect.

Dans ce mémoire, nous avons cherché à créer une ressource de type CEFRLex – graduée **et désambiguïsée** – pour la langue anglaise (EFLSemLex - *English as a Foreign Language Semantic Lexicon*). Nous nous sommes également demandé si la ressource créée permettrait d'appréhender les phénomènes de complexité lexicale mieux que la ressource CEFRLex existant déjà pour l'anglais (EFLLex – Dürlich & François, 2018). Ce faisant, nous avons cherché à répondre à la question suivante : « En quoi la désambiguïsation lexicale peut-elle potentiellement améliorer les ressources lexicales destinées aux apprenants (d'une langue étrangère ?) »

Le présent travail se présente en deux parties. La première, d'ordre théorique, s'ouvre sur un état de l'art de la désambiguïsation lexicale. Dans ce premier chapitre, nous faisons le tour des différents systèmes ayant existé jusqu'à ce jour, en mettant l'accent sur le type de

[4] *Common European Framework of Reference for Languages.*

données utilisées pour distinguer automatiquement le sens des mots (Chapitre 1). Ensuite, seront évoquées les principales ressources lexicales en anglais. Les différents critères permettant de construire ces ressources s'y trouvent détaillés. Ce second chapitre se clôturera par une description plus avancée du projet CEFRLex, dans le sillage duquel s'inscrit notre travail (Chapitre 2). Enfin, plusieurs arguments théoriques en faveur de la désambiguïsation des ressources lexicales sont présentés, ainsi que les quelques ressources désambiguïsées connues à ce jour, dont NT2Lex (Chapitre 3). La seconde partie, afférente à la création de la ressource EFLSemLex, sera, elle aussi, subdivisée en trois chapitres. Nous expliciterons la méthodologie respectée (Chapitre 1), avant de décrire la ressource obtenue au terme des étapes décrites (Chapitre 2). Finalement, l'utilité concrète de l'étape de désambiguïsation – qui est la spécificité principale d'EFLSemLex par rapport à EFLLex – sera remise en question et discutée (Chapitre 3).

Attendu que notre travail est centré sur la création d'une ressource en langue anglaise, nous nous focaliserons uniquement sur les travaux et ressources pour cette langue.

Partie I) Désambiguïsation et ressources lexicales : état de l'art

Chapitre 1. Ambiguïté et désambiguïsation lexicale : inculquer le sens aux machines

Schématiquement, la sémantique est l'étude du sens : elle touche les significations qui sont transmises par les mots et les énoncés, le « fond » du langage. Étudier la sémantique revient à étudier *ce dont* les locuteurs parlent. Cette branche de la linguistique peut être confrontée à celles qui s'intéressent à la forme de la langue : orthographe, phonétique, syntaxe, etc. Étant donné que la sémantique sera le domaine central et transversal de ce travail, nous y accordons logiquement le premier chapitre.

Celui-ci consistera d'abord en un dégrossissement de quelques notions de sémantique (ambiguïté, homonymie, polysémie – A). Ensuite, nous nous intéresserons plus particulièrement à la question de la gestion sémantique dans le cadre du NLP (B). Après une définition formelle de la « désambiguïsation lexicale » (B.1), seront présentées les principales applications (B.2) et difficultés associées (B.3) à cette dernière. Les données et ressources utilisées pour la désambiguïsation (B.4) feront office d'armature pour une synthèse globale de l'histoire de la tâche (B.5). Le chapitre se clôturera sur une comparaison quantitative des meilleurs systèmes de désambiguïsation actuels (B.6).

A. Définir l'ambiguïté lexicale

Une forme lexicale donnée peut renvoyer à des concepts différents. Par exemple, *brightness* fera tantôt référence à la qualité d'être lumineux, tantôt à celle d'être intelligent. Les mots sont susceptibles d'avoir plusieurs sens ; pour désigner cette caractéristique fondamentale des langues naturelles, on utilise le terme « ambiguïté lexicale »[5].

En anglais, plus de 80% des mots communs comptent plus d'une entrée au dictionnaire (Rodd et al., 2002), et chaque mot de la liste de substantifs les plus courants de l'anglais dispose, en moyenne, de 7.8 sens dans WordNet (Turdakov, 2010). Autrement dit, le récepteur d'un message linguistique doit presque systématiquement choisir la signification correcte – contextuellement appropriée – des mots qu'il rencontre. Pourtant, malgré l'ampleur de l'ambiguïté lexicale, les locuteurs sont rarement conscients de l'abondance d'autres sens possibles pour les mots rencontrés : le système de compréhension de la langue est naturellement efficace pour résoudre les ambiguïtés (Degani & Tokowicz, 2010b).

Certaines situations utilisent expressément l'ambiguïté lexicale, à des fins artistiques ou humoristiques, notamment. Les jeux de mots, par exemple, sont souvent délibérément construits afin de référer à plusieurs sens d'un mot : ces traits d'esprits sont l'une des rares situations dans lesquelles les locuteurs sont amenés à prendre conscience de l'ambiguïté lexicale (« *Tu connais l'histoire de l'armoire ? Elle n'est pas commode…* »). Communément, les récepteurs perçoivent uniquement la signification contextuellement appropriée des mots (Rodd, 2018). Les locuteurs ne sont donc pas nécessairement conscients des différents types d'ambiguïtés existant, dont homonymie et polysémie sont les principaux représentants.

[5] Certains auteurs y préfèrent l'étiquette « ambiguïté sémantique », qui rendrait mieux compte du fait que le *sens* du mot rend celui-ci ambigu, plutôt que ses propriétés grammaticales (Vitello & Rodd, 2015).

i. *Homonymie*

Dans certains cas, les différentes significations d'une même forme ne sont pas sémantiquement reliées et ne partagent pas une racine commune au sein de l'histoire de la langue : dans ces cas, on parle d'*homonymie*. Les homonymes sont le résultat de contingences dans l'évolution de la langue, qui expliquent qu'une même forme finisse par correspondre à deux significations distinctes (*e.g. bark* [aboiement] et *bark* [écorce d'un arbre], exemple tiré de Rodd, 2018). Généralement, la dissociation sémantique s'accompagne d'une dissociation lexicographique : les homonymes se retrouvent sous des entrées différentes dans les dictionnaires. Les cas d'homonymie complète (homophonie/homographie) sont assez rares. Seuls 7% des mots anglais peuvent être classifiés comme tels (Rodd et al., 2002).

ii. *Polysémie*

Une autre forme – plus commune – d'ambiguïté lexicale concerne les mots *polysémiques*, dont les significations sont sémantiquement reliées. Bien que les sens d'un mot polysémique se recouvrent, il est nécessaire pour le récepteur de savoir exactement quelle définition est visée par l'émetteur (*e.g.* « *The athlete runs down the track* » / « *The mayor runs for election* » / « *The film runs at the cinema* ». Rodd, 2018). La polysémie permet d'utiliser les mots de manière flexible, avec de subtiles variantes sémantiques, et la diversité des usages d'un même mot n'est probablement pas toujours capturée par la liste des sens dudit mot qui se retrouve dans les dictionnaires (Rodd, 2018).

Les méthodes de désambiguïsation lexicale, que nous présenterons en détails dans la section B.5, se concentrent uniquement sur les cas d'homonymie et de polysémie. Notons déjà que, pour un ordinateur, la polysémie – en ce qu'elle est plus subtile – est plus difficile à désambiguïser que les homonymes (Agirre & Edmonds, 2007, p. 4)[6].

D'autres auteurs évoquent aussi une forme d'*ambiguïté inter-linguistique*, qui toucherait les locuteurs multilingues (*e.g. room* - EN. chambre, espace *vs*. NL. crème). Ce type d'ambiguïté est limité par les caractéristiques individuelles des langues (sonorités, alphabet, etc.) et les indices contextuels sur la langue utilisée, mais peut entraver la compréhension chez les locuteurs bilingues (Degani & Tokowicz, 2010b), en particulier pour les formes écrites des langues ayant une origine commune. D'ailleurs, les enseignants ont bien connaissance du phénomène : ils réfèrent à ces cas d'ambiguïté par le terme « *false friends* ».

B. Résoudre automatiquement l'ambiguïté : la désambiguïsation lexicale

L'identification du sens spécifique d'un mot – nous l'avons dit – pose généralement peu de problèmes aux locuteurs. Cependant, l'étape de résolution de l'ambiguïté n'est simple qu'en apparence : pour une machine, cette dernière implique de traiter et structurer de l'information textuelle complexe avant de pouvoir déterminer la signification de chaque unité lexicale. L'identification automatique du sens des mots en contexte est appelée « désambiguïsation lexicale (ou sémantique) » (*word sense disambiguation* - *WSD*).

[6] Dans le cadre des travaux sur la désambiguïsation lexicale, la différence entre homonymie et polysémie est communément enclose dans l'opposition plus générale entre distinction sémantiques « fines » ou « raffinées » (*fine-grained*) et distinctions « grossières » (*coarse-grained*).

B.1. Définition de la tâche

Formellement, la désambiguïsation lexicale peut être décrite comme la tâche d'assignation d'une (ou plusieurs) significations appropriées aux mots d'un texte, en fonction du contexte environnant[7]. Traditionnellement, la tâche est découpée en deux étapes : (1) détermination des sens des mots (sur la base d'un inventaire sémantique, cf. B.4.i) ; (2) assignation du sens approprié à chaque occurrence d'un mot.

La plupart des auteurs conçoivent cette deuxième étape comme une tâche de *classification* : les sens des mots sont des classes, et une méthode de classification automatique est utilisée pour assigner chaque mot à une ou plusieurs classes, en se basant sur le *contexte* (cf. B.3.ii) et sur des *sources de connaissances externes* (ressources lexicales, réseaux sémantiques, corpus, etc. - cf. B.4).

Néanmoins, certains auteurs (Agirre & Edmonds, 2007, p. 18) avancent que les significations d'une phrase sont interdépendantes (ce qui est clairement pris en compte dans les premiers systèmes – tels que Lesk, 1986), et que ces interdépendances pourraient également être modelées et traitées comme un problème d'*optimisation*. Cette notion d'interdépendance est reprise par d'autres systèmes plus récents qui abordent la WSD comme un problème de *classification séquentiel* (Huang et al., 2019; Yap et al., 2020).

En outre, deux variantes de la tâche générique[8] de désambiguïsation lexicale se distinguent : la *WSD ciblée* implique de désambiguïser quelques mots d'un texte seulement – typiquement un par phrase ; tandis que la *all-words WSD*[9] suppose du système qu'il soit capable de désambiguïser tous les mots pleins dans un texte.

B.2. Une fonction intermédiaire ? Applications de la WSD

Initialement, le problème de la désambiguïsation sémantique a été présenté dans le contexte de la traduction automatique (Weaver, 1949 - cf. B.5.i). D'emblée, la désambiguïsation apparaît donc comme « intermédiaire » dans le cadre du traitement automatique du langage (Wilks & Stevenson, 1996). Ainsi, la désambiguïsation sémantique ne serait vue que comme un moyen permettant d'aboutir à d'autres fins diverses. En effet, la WSD octroie des « compétences » sémantiques à de nombreuses applications liées au langage (cf. Ide & Véronis, 1998 ; Navigli, 2009 ; Turdakov, 2010) :

* **Traduction automatique**. La compréhension du sens permet de traduire correctement en contexte. Par exemple, le mot français 'CADRE' pourra se traduire en anglais par 'ENVIRONMENT' (*contexte de travail*), 'EXECUTIVE' (*responsable d'une entreprise*), 'FRAME' (*pour une photo*) ou encore 'SCOPE' (*la portée d'un projet, d'une enquête*).

[7] La désambiguïsation lexicale concerne les mots appartenant à une même catégorie grammaticale, car les distinctions inter-catégorielles sont gérées par l'analyse syntaxique (POS-tagging, parsing, etc.)

[8] Par opposition à la tâche de WSD *domain-specific* (focalisée sur un domaine d'expertise donné), qui ne fait pas l'objet du présent travail.

[9] Remarquons que l'utilité de l'approche *all-words* peut être questionnée : « *Nonetheless, we think that disambiguating all content words sometimes proves to be an academic exercise […] Again, it might not be necessary to disambiguate all the words […] but rather a substantial subset of them, that is, those conveying the real content of the resource* » (Navigli, 2009, p. 57).

- **Recherche d'information**. Dans une recherche d'information basée sur des mots-clés, il peut être intéressant de sélectionner les documents contenant certains mots uniquement dans le sens voulu. Par exemple, dans une recherche liée au droit, les documents reprenant le mot 'COUR' dans le sens '*cour de récréation*' ne présentent pas d'intérêt.
- **Analyse de contenu**. L'objectif de l'analyse de contenu est d'analyser la distribution de catégories de mots dans des ensembles de textes, par exemple les mots liés à un concept, à un sujet donné, etc. Le dépassement de la forme par la prise en compte du sens permet de construire des distributions plus exactes.
- **Traitement de la parole**. La désambiguïsation peut être utile pour la production correcte des mots en synthèse vocale (*e.g.* 'DESERT' : /dɪˈzɜːt/ - déserter vs. /ˈdez.ət/ - le désert). En reconnaissance vocale, la WSD aide à segmenter les mots et à différencier les homophones.
- **Analyse grammaticale**. La désambiguïsation est parfois utile pour le POS-tagging : dans la phrase *L'étagère plie sous les livres*, « livres » est un nom masculin, mais si le sens de « livres » est mal interprété, il pourrait être taggué comme féminin (« une livre » étant également une unité de mesure). La WSD réduit, en outre, le nombre d'analyses syntaxiques possibles.
- **Traitement de texte**. La WSD peut être utilisée pour améliorer les méthodes de classification et de regroupement de textes, pour corriger des fautes d'orthographe sophistiquées (*e.g. Une personne *censée est pleine de bon sens*), pour analyser des textes, pour avoir un accès lexical aux langues sémitiques (dans lesquelles les voyelles ne sont pas écrites), etc.
- **Lexicographie**. La lexicographie actuelle est basée sur des corpus. La désambiguïsation lexicale et la lexicographie peuvent donc travailler de concert, avec la WSD offrant de nouveaux groupements de sens aux lexicographes, et les lexicographes fournissant de meilleurs corpus et inventaires sémantiques aux développeurs de WSD.
- **Web sémantique**. Le Web sémantique, par définition, requiert des outils de désambiguïsation lexicale afin de favoriser l'interopérabilité entre les systèmes et les utilisateurs.

Bien qu'il soit évident que la prise en compte d'une phase de désambiguïsation sémantique soit nécessaire dans plusieurs domaines, la collaboration n'est pas toujours aisée :

> And yet, explicit WSD has not yet been convincingly demonstrated to have a significant positive effect on any application (Agirre & Edmonds, 2007, p. 3)

> Since WSD systems now work fairly well, it is time to employ them in other applications too, e.g., boosting semantic intensive downstream tasks such as Machine Translation, Semantic Role Labeling, and Question Answering. (Bevilacqua et al., 2021, p. 4336)

Au-delà de ses utilisations implicites, l'information d'ordre intermédiaire est souvent essentielle pour comprendre le fonctionnement du langage naturel, pour infirmer ou confirmer des théories linguistiques (Wilks & Stevenson, 1996). Certains auteurs, arguant que la polysémie demeure l'un des phénomènes les plus complexes, s'éloignent de la distinction "intermédiaire – final" et positionnent la désambiguïsation lexicale aux premières lignes du traitement automatique du langage (Turdakov, 2010 ; Bevilacqua et al., 2021).

L'histoire de la désambiguïsation lexicale (B.5), dont on peut mesurer l'ampleur par les quelques 15.000 résultats qu'une recherche pour « *word sense disambiguation* » dans la base d'articles *ACL Anthology*[10] renvoie, indique bien l'importance de ce champ d'investigation dans le cadre du traitement automatique du langage.

B.3. Difficultés inhérentes à la désambiguïsation lexicale

La WSD a été décrite comme un problème *AI-complete* (Mallery, 1988), soit un problème dont la difficulté est équivalente à celle de problèmes centraux de l'intelligence artificielle (qui visent à rapprocher les compétences des systèmes informatiques à celles des humains). La difficulté inhérente à la désambiguïsation lexicale trouve sa source dans une variété de facteurs (Navigli, 2009).

i. *Définir le sens : une tâche compliquée*

L'un des paramètres les plus restrictifs est la représentation du sens des mots et la granularité de l'inventaire sémantique utilisé (Turdakov, 2010). D'une part, il n'existe pas de définition stricte et consensuelle du sens lexical. Le langage est soumis à la variation et à l'interprétation, c'est pourquoi de nombreuses incohérences subsistent entre les dictionnaires, mais également au sein des dictionnaires. Les lexicographes ne s'accordent pas toujours sur le nombre de sens d'un mot ou la correction des définitions. D'autre part, la granularité (ou degré de précision dans l'inventaire des significations) est un problème central pour la désambiguïsation : des distinctions sémantiques trop subtiles (*fine-grained*), qui, de surcroit, ne correspondent pas toujours à la perception des locuteurs (Agirre & Edmonds, 2007, p. 9), rendent la sélection trop complexe pour les systèmes et font exploser le nombre de combinaisons de sens possibles ; en revanche, des divisions trop grossières (*coarse-grained*) – bien que donnant lieu à des systèmes plus performants et généralisables (Navigli, 2009) – ne suffisent pas toujours pour certaines tâches de NLP. Deux problèmes différents peuvent, en effet, requérir des degrés de granularité différents.

ii. *Représentation du contexte*

La notion de contexte n'est pas univoque. Celle-ci peut être réduite aux mots environnants – *contexte local* –, mais également prendre en considération la thématique du texte ou de ses sous-parties – *contexte topical*. Le *domaine* de spécialité est une autre dimension du contexte (Ide & Véronis, 1998, p. 18). En outre, le contexte peut être envisagé comme un ensemble de mots non ordonné (approches dites '*bag-of-words*'), mais l'information relationnelle (distance, relations syntaxiques et sémantiques, etc.) n'est pas systématiquement exclue des systèmes de WSD. Pour finir, les caractéristiques contextuelles efficaces diffèrent selon la classe grammaticale des mots à désambiguïser : en ce qui concerne les noms, l'approche par une fenêtre formelle (*bag-of-words*) fonctionne mieux ; tandis que pour les verbes, les méthodes qui prennent en compte d'autres types de relations sont plus efficaces (Ide & Véronis, 1998, p. 19).

[10] https://aclanthology.org/. Recherche effectuée le 16 janvier 2022.

iii. *Knowledge acquisition bottleneck*

La WSD repose fortement sur des données : chaque méthode requiert une forme ou l'autre d'information. Ces données vont de textes sémantiquement annotés ou non à des ressources plus structurées telles que des dictionnaires ou des réseaux sémantiques (cf. B.4). La principale limitation liée au besoin de données réside dans la laboriosité de la création de ressources exploitables (*knowledge acquisition bottleneck* - Gale et al., 1992). Cette étape doit, en outre, être répétée à chaque fois que les modalités de la désambiguïsation changent (nouveaux inventaires de sens, nouvelles langues, nouveaux domaines).

En un mot, la tâche de WSD peut difficilement être rendue universelle en raison de questions fondamentales : représentation du sens et du contexte, approche générique vs. *domain-specific*, approche ciblée vs. *all-words*, etc. Il s'est d'ailleurs avéré difficile, jusqu'au premier séminaire Senseval en 1998 (Kilgarriff & Palmer, 2000), de déterminer un étalon pour l'évaluation des modèles de WSD (cf. B.6). La pauvre disponibilité de ressources utilisables représente un autre frein important au développement d'outils efficaces.

B.4. Ressources utilisées pour la désambiguïsation lexicale

La recherche en WSD a produit, utilisé et rendu disponible de nombreuses ressources. Celles-ci se déclinent en deux types : (i) les inventaires sémantiques, dont les données sont structurées ; et (ii) les corpus, dont les données non structurées peuvent être brutes ou étiquetées sémantiquement. Pour chacune de ces catégories, nous décrirons en détails les ressources principales et présenterons plus succinctement les ressources secondaires.

i. *Inventaires sémantiques*

La plupart des inventaires sémantiques connus inventorient tous les sens possibles d'un mot (approche énumérative[11]). Les informations encodées dans ces inventaires sémantiques peuvent, par ailleurs, être utilisées dans le cœur des processus de désambiguïsation.

i.a. *Machine readable dictionaries* (MRDs)

Dans les années 80, les dictionnaires ont été rendus disponibles au format électronique. Attendu qu'ils incarnent des sources d'informations sémantiques fécondes, de nombreux chercheurs ont cherché à en incorporer l'information dans leurs systèmes. Les dictionnaires présentent deux désavantages majeurs : premièrement, ils sont conçus pour les humains et présentent des incohérences internes ; deuxièmement, ils contiennent peu d'information d'ordre pragmatique.

Parmi les dictionnaires les plus utilisés pour la désambiguïsation en langue anglaise, nous retrouvons *the LONGMAN DICTIONARY OF CONTEMPORARY ENGLISH* (LDOCE) ; *the OXFORD DICTIONARY OF ENGLISH* (OED) ; *the COLLINS ENGLISH DICTIONARY* et *the OXFORD ADVANCED LEARNER'S DICTIONARY OF CURRENT ENGLISH* (OALD). Plus récemment, le dictionnaire collaboratif *WIKTIONARY* a également suscité un intérêt dans la communauté scientifique (Blevins et al., 2021), malgré ses distinctions sémantiques trop subtiles.

[11] Notons l'existence d'une autre approche, générative, qui consiste à dériver des lexiques au moyen de règles aspirant à capturer les régularités dans la création des sens (Pustejovsky, 1995). L'approche générative présente l'avantage de tenir compte du caractère non statique de la sémantique. Dans le cadre de l'approche générative, Buitelaar (1998) créé CoreLex, un inventaire sémantique qui identifie tous les sens systématiquement reliés et permet un *tagging* sémantique « sous-spécifié » (puisque les sens ne sont pas *énumérés*). Ces méthodes n'ont pas suscité davantage d'expérimentations par la suite.

i.b. Thésaurus

Les thésaurus offrent un riche réseau d'associations lexicales (synonymie et antonymie, surtout) potentiellement exploitables pour le traitement du langage. Toutefois, les thésaurus – à l'instar des dictionnaires – sont destinés aux humains, et ne représentent pas une source idéale d'informations lexico-sémantiques. Concrètement, les thésaurus n'ont pas été très utilisés en désambiguïsation lexicale.

- *ROGET'S INTERNATIONAL THESAURUS* (Roget, 1911). Le *Roget's* est le thésaurus le plus fréquemment utilisé pour la désambiguïsation lexicale (v. Masterman, 1957; Patrick, 1985; Yarowsky, 1992). La version la plus récente contient 250.000 entrées organisées en six classes hiérarchiques et près de 1000 catégories. Chaque occurrence d'un mot dans une catégorie représente un sens différent de ce mot. Les plus hauts niveaux de la hiérarchie du *Roget's* sont généralement considérées comme trop larges pour établir des catégories sémantiques intéressantes.

i.c. Réseaux sémantiques

Les réseaux sémantiques – ou « lexiques computationnels » - encodent un riche ensemble de concepts sémantiques. Pour cette raison, ils sont – à l'heure actuelle – largement plus populaires que les dictionnaires informatisés et thésaurus. Au vu du consensus sur son utilisation, WordNet peut être considéré *de facto* comme le standard pour la WSD en langue anglaise. BabelNet est également couramment utilisé.

- *WORDNET* (WN - Fellbaum, 1998; Miller et al., 1990). Dans cette ressource lexicale, les mots de l'anglais ne sont pas seulement répertoriés, mais organisés dans un graphe dont les nœuds sont des *synsets* (*i.e.* des groupes de synonymes contextuels). Chaque synset représente – approximativement – un sens. Autrement dit, les occurrences des mots dans un *synset* donné correspondent à un sens possible de ce mot. Les *synsets* de WN sont associés selon de nombreux liens conceptuo-sémantiques (hyponymie, hypéronymie, antonymie, partie, substance, causalité, etc.). Pour chaque *synset*, WN offre aussi d'autres types de connaissances lexicales, notamment des définitions (ou gloses) et des exemples d'usage.

Les travaux les plus récents utilisent la version 3.0 de WN (parue en 2006). Cette dernière compte 117.659 synsets[12] et 206.879 occurrences de mots. Le contre-pied de cette large couverture réside dans sa précision en termes de définition des sens. La granularité sémantique de WN est très fine, ce qui engendre une surabondance de *synsets* qui pourraient – par ailleurs – être rassemblés. Bien que très commun, le verbe *to give* ("donner"), par exemple, y compte à lui seul 44 sens.

- *BABELNET* (Navigli & Ponzetto, 2012). Ce réseau sémantique multilingue a été obtenu en modélisant semi-automatiquement plusieurs ressources, telles que WordNet, des versions multilingues de WordNet ainsi que Wikipédia, notamment. Comme pour WN, les nœuds du graphe sont des synsets multilingues, et les liens entre ces nœuds incarnent plusieurs types de relations sémantiques. La version 5.0 de BabelNet (Navigli et al., 2021) couvre 500 langues, s'appuie sur 51 sources d'informations et contient plus de 20M de *synsets*.

[12] Récemment, English WordNet 2020 (McCrae et al., 2020) a étendu la version originale de WN en y ajoutant près de 3000 nouveaux synsets, comprenant de l'argot et des néologismes.

ii. *Corpus*

Les corpus sont des collections de textes utilisées pour l'apprentissage automatique. Ceux-ci contiennent de l'information fréquentielle et co-occurrentielle qui ne peut être extraite depuis les inventaires sémantiques. Les corpus peuvent être classifiés en deux groupes : les corpus bruts, et les corpus annotés sémantiquement.

ii.a. Corpus bruts

Les corpus bruts sont pratiques pour les approches de désambiguïsation non supervisées. Les trois corpus les plus fréquemment repris dans la recherche en WSD sont le BROWN CORPUS (Kučera & Francis, 1967), le BRITISH NATIONAL CORPUS (Clear, 1993) et le WALL STREET JOURNAL CORPUS (Paul & Baker, 1992)[13].

- Le *BROWN CORPUS* est un ensemble « équilibré » de 500 textes publiés aux États-Unis en 1961. Celui-ci contient notamment des articles de presse, des articles scientifiques et de la fiction, pour un total d'un million de mots. Le corpus annoté SEMCOR (cf. infra) repose sur le BROWN CORPUS.

- Le *BRITISH NATIONAL CORPUS* compte 100M de mots extraits de l'anglais britannique parlé et écrit. Le BNC, en traitement automatique du langage, est souvent utilisé pour collecter des fréquences lexicales ou identifier des relations grammaticales.

- Le *WALL STREET JOURNAL (WSJ) CORPUS* a été largement utilisé en traitement automatique du langage. Il comprend près de 30M de mots tirés du WSJ et est notamment à l'origine du corpus manuellement annoté DSO.

ii.b. Corpus annotés sémantiquement

Les corpus étiquetés, à l'inverse des corpus bruts, sont utilisés par les systèmes dits « supervisés ». Ceux-ci peuvent être obtenus par l'annotation manuelle ou automatiquement. Pour la plupart des corpus annotés, l'inventaire sémantique utilisé est tiré des différentes versions de WN.

Les deux principaux corpus annotés manuellement sont le corpus SEMCOR (Miller et al., 1993) et le corpus DSO (Ng & Lee, 1996).

- *SEMCOR*. SEMCOR est le corpus le plus utilisé pour la désambiguïsation lexicale. Celui-ci est composé de 352 textes (700.000 *tokens*) du BROWN CORPUS, pour un total d'environ 230.000 occurrences annotées sémantiquement. En revanche, SEMCOR ne recouvre que 22% de tous les synsets de WN, et repose sur des textes de 1967 qui ne reflètent plus exactement le paysage linguistique contemporain. Pour augmenter la couverture des annotations, de récents travaux (Bevilacqua & Navigli, 2020; Vial et al., 2019) ont commencé à utiliser le WORDNET GLOSS CORPUS (WNG) comme source d'information complémentaire. Celui-ci contient des définitions et des exemples de WN annotés à la fois manuellement et automatiquement, et couvre plus de 50% des sens de WN.

[13] À titre informatif, nous pouvons citer d'autres corpus repris dans les historiques de la WSD (Agirre & Edmonds, 2007; Navigli, 2009) : le REUTERS NEWS CORPUS, le AMERICAN NATIONAL CORPUS et le GIGAWORD CORPUS.

▸ *DSO*. Ce corpus a été compilé par la *Defense Science Organisation* (DSO) à Singapour. Il contient des textes du BROWN CORPUS et du WSJ. Le DSO consiste en 192.800 annotations pour 121 noms et 70 verbes qui, selon les auteurs, représenteraient les mots les plus fréquents et ambigus de la langue anglaise.

Plusieurs autres corpus étiquetés manuellement existent. Parmi ceux-ci, nous pouvons évoquer MULTISEMCOR, OPEN MIND WORD EXPERT, LINE-HARD-SERVE CORPUS, INTEREST CORPUS, PENN TREEBANK et PROPBANK.

Plus récemment, les méthodes automatiques se sont montrées efficaces pour la création de corpus annotés. Ces derniers présentent l'avantage d'être plus longs et moins difficiles à concevoir, mais demeurent encore peu utilisés : parmi ceux-ci, nous évoquerons OMSTI (Taghipour & Ng, 2015), TRAIN-O-MATIC (Pasini & Navigli, 2019) et UWA (Loureiro & Camacho-Collados, 2020).

▸ *OMSTI* (*One-million sense-tagged instances*). OMSTI (1M d'occurrences annotées) a été obtenu grâce à l'alignement – sur l'anglais – de l'annotation sémantique semi-automatique d'un corpus parallèle chinois-anglais. Les méthodes de désambiguïsation sur lesquelles reposent ce corpus sont datées, mais les annotations en résultant sont précises à 83.7% (Chan & Ng, 2005; Ng et al., 2003). OMSTI a déjà démontré son potentiel en améliorant les performances des systèmes supervisés qui l'ont ajouté à leurs données d'entraînement.

▸ *TRAIN-O-MATIC* (Pasini & Navigli, 2019). Grâce à une méthode complètement automatisée basée uniquement sur une version améliorée de WN, TRAIN-O-MATIC permet de générer – pour plusieurs langues – des millions de phrases d'entraînement annotées. La ressource TRAIN-O-MATIC pour l'anglais comporte 12.722.530 occurrences annotées (couverture : 51.396 sens).

▸ *UWA* (*Unambiguous Word Annotations*). Les couvertures d'OMSTI et de TRAIN-O-MATIC restent aussi limitantes que celle de SEMCOR (moins de 25% des synsets). Loureiro & Camacho-Collados (2020) proposent une méthode automatique pour l'annotation des mots non ambigus en corpus[14]. Cette ressource contient 6.111.453 occurrences annotées en 98.494 sens, et inclut plus de la moitié des sens de WN (56.7% des synsets sont couverts).

B.5. Classification et histoire des recherches en WSD

Précédemment, nous avons donné à voir différentes déclinaisons de la tâche de désambiguïsation lexicale (ciblée vs. *all-words*, générique vs. *domain-spécific*, etc.). Toutefois, les principales synthèses de la désambiguïsation lexicale (Ide & Véronis, 1998; Agirre & Edmonds, 2007; Navigli, 2009; Bevilacqua et al., 2021) s'appuient plus volontiers sur le type de données utilisées par les systèmes en vue de classifier ceux-ci. La figure 2 présente graphiquement les distinctions usuelles entres les systèmes de WSD.

[14] La prise en compte des mots non ambigus dans les données d'entraînement n'est pas anodine, car la présence de ceux-ci peut avoir un impact sur les mots à désambiguïser. De plus, la majorité des lemmes de WN sont non ambigus (116.000 vs. 30.000).

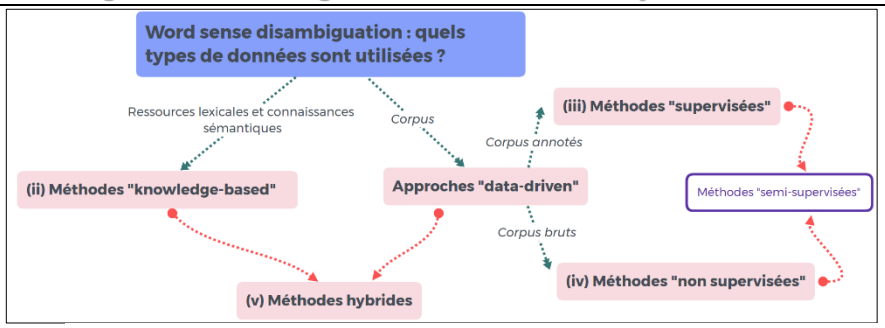


Figure 2 - Principales méthodes de désambiguïsation lexicale
(sur les lignes pointillées : types de données ; en rose : types d'approches)

Après un rapide aperçu de l'origine de la désambiguïsation lexicale (i), nous en explorerons les principales approches (Figure 2) d'hier à aujourd'hui. Cette section s'appuie sur les états de l'art principaux de la WSD évoqués dans le paragraphe précédent. Nous avons choisi de ne pas nous arrêter sur les détails mathématiques des systèmes abordés, afin de présenter au lecteur un aperçu global et clair des méthodes de désambiguïsation existantes.

i. *Naissance d'une problématique*

La question de la désambiguïsation remonte à l'époque des balbutiements du traitement automatique du langage. Initialement, la nécessité de la désambiguïsation lexicale a été soulevée par Weaver (1949), dans le cadre de la traduction automatique. Un mot identique dans une langue source (*e.g.* EN. *bass*) peut avoir des traductions différentes dans une langue cible (FR. *basse* [instrument] ou *bar* [poisson]). Par conséquent, il sera nécessaire de pouvoir désambiguïser automatiquement ces mots. À cet effet, l'importance de l'environnement linguistique direct était déjà bien comprise :

> If one examines the words in a book, one at a time as through an opaque mask with a hole in it one word wide, then it is obviously impossible to determine, one at a time, the meaning of the words .... But if one lengthens the slit in the opaque mask, until one can see not only the central word in question but also say N words on either side, then if N is large enough one can unambiguously decide the meaning of the central word .... The practical question is: "What minimum value of N will, at least in a tolerable fraction of cases, lead to the correct choice of meaning for the central word?" (Weaver, 1949)

Kaplan (1955) a montré que la décision prise était généralement aussi bonne avec une « loupe » de valeur N=2 (de chaque côté), que si le lecteur dispose de la phrase entière.

Weaver a également évoqué le rôle du domaine du texte dans la désambiguïsation : « […] *each word, within the general context of a mathematical article [for instance], has one and only one meaning* ».

Par ailleurs, Weaver postule que les études sémantiques devraient être abordées d'abord et surtout sous l'angle statistique. Le chercheur prépare ainsi le terrain pour les méthodes basées sur le *machine learning* et les réseaux de neurones éminemment complexes que la WSD connaît aujourd'hui. Néanmoins, le père de la désambiguïsation lexicale reconnaît également l'importance des connaissances encyclopédiques sur les mots et la structure de la langue. Ainsi, Weaver ouvre la voie du développement des méthodes de désambiguïsation lexicales « *knowledge-based* ».

Il est intéressant de remarquer que les approches et entraves majeures liées à la désambiguïsation lexicale étaient déjà esquissées par Weaver. Toutefois, les idées de ce

dernier sont longtemps restées intestables, par manque de ressources d'envergure. D'ailleurs, la connaissance des difficultés inhérentes à la désambiguïsation a causé l'avortement des tentatives de traduction automatiques dans les années 60 (« *no existing or imaginable program will enable an electronic computer to determine that the word pen in the given sentence within the given context has the second of the above meanings* » Bar-Hillel, 1960).

ii. *Méthodes knowledge-based*

Les méthodes basées sur les connaissances ne requièrent pas de données annotées, ce qui, au vu du coût de création de ces données, représente un avantage en soi. Au surplus, les méthodes *knowledge-based* disposent souvent d'une couverture plus importante que leurs homologues supervisés – puisqu'il est impossible que les annotations sémantiques des corpus couvrent tout le vocabulaire ou tous les sens. Ainsi, les systèmes *knowledge-based* sont – au départ – plus facilement applicables pour les tâches de *all-words* WSD.

Les premières méthodes de désambiguïsation (des années 60 aux années 80) étaient toutes basées sur des connaissances sémantiques et syntaxiques de la langue. Ces méthodes s'ancraient essentiellement dans des systèmes globaux visant à modéliser la compréhension humaine du langage. L'apparition, dans les années 80, de grandes bases de données exploitables (dictionnaires, thésaurus, lexiques sémantiques) marque un tournant important pour la WSD. Le focus a, dès lors, pu être mis sur l'extraction et l'exploitation de données, plutôt que sur leur création.

Agirre & Edmonds (2007) distinguent quatre types d'approches *knowledge-based*, selon le type de connaissances exploitées : les gloses des dictionnaires, les réseaux sémantiques, les préférences sélectives et des propriétés générales du langage humain.

ii.a. Recours aux gloses lexicographiques

L'algorithme de Lesk (Lesk, 1986) est certainement le plus célèbre de ce type. La désambiguïsation appliquée par cet algorithme consiste en la sélection des sens dont la définition (glose) présente le plus grand nombre de *chevauchements* avec les mots du *contexte* – sur une certaine fenêtre syntagmatique. L'inventaire de sens utilisé par Lesk est relativement raffiné (issu du *Oxford Advanced Learner's Dictionary of Current English*). Bien que très sensible à la formulation des gloses – l'absence ou la présence de tel ou tel mot peut radicalement altérer les résultats –, l'algorithme de Lesk a servi de base pour la plupart des travaux de désambiguïsation basés sur les MRDs.

De nombreux auteurs ont affiné l'algorithme de Lesk afin d'obtenir de meilleurs résultats, notamment par l'utilisation de réseaux neuronaux (Veronis & Ide, 1990), par l'utilisation des *subject codes* des dictionnaires (ingénierie, économie, etc. - Guthrie et al., 1991) par l'utilisation du « recuit-simulé » (*simulated annealing*) pour éviter l'explosion des combinaisons de sens possibles (Cowie et al., 1992) ou en considérant également les définitions des mots associés (Adapted Lesk Algorithm - Banerjee & Pedersen, 2002).

De nos jours, les gloses sont encore fréquemment utilisées pour la WSD, mais sont généralement incorporées en tant qu'information complémentaire à des systèmes plus complexes (cf. section v.a).

ii.b. Recours aux réseaux sémantiques

Voorhees (1993) fut le premier à exploiter WordNet pour la WSD dans le cadre de l'extraction d'information. La technique utilisée consiste à trouver le plus grand sous-graphe connecté à un mot donné afin de représenter des catégories de sens. La même année, Sussna (1993) postule que, pour un ensemble de termes apparaissant ensemble, choisir la combinaison de sens minimisant la distance dans le graphe sémantique permettrait de sélectionner les sens appropriés.

En effet, pour que le langage soit cohérent, les mots d'un discours doivent être reliés sémantiquement (Halliday & Hasan, 1976). Ceci est une contrainte puissante pour la WSD, et a donné naissance à la notion de similarité sémantique. Mesurer la similarité d'un mot aux autres mots du contexte permet de désambiguïser les textes. Plusieurs auteurs ont développé différentes mesures de similarité sémantique. Ainsi, Resnik (1995) propose que plus le plus « bas » commun hypéronyme de deux mots est spécifique, plus les mots sont sémantiquement proches. Cette idée sera reprise par Jiang & Conrath (1997) et Lin (1998). D'autres auteurs réduisent la similarité à la synonymie (Stetina & Nagao, 1998) ou à la distance dans le graphe (Leacock & Chodorow, 1998). Hirst & St-Onge (1998) introduisent la notion de direction : pour que deux concepts soient proches, le graphe ne doit pas changer de direction trop fréquemment entre les deux. Enfin, Mihalcea & Moldovan (1999) choisissent de ne plus se baser sur des graphes réels, mais sur les gloses pour mesurer un chemin virtuel entre différents concepts.

Un autre moyen d'utiliser la structure des réseaux sémantiques est de chercher – dans le contexte textuel – des chaînes de sens associés. Ces chaînes lexicales seront retenues si elles excèdent une longueur donnée (Okumura & Honda, 1994; Harabagiu, 1999; Galley & McKeown, 2003). Mihalcea et al. (2004) et Erkan & Radev (2004) approfondissent l'idée de chaînes lexicales dans une méthode non supervisée basée sur l'ordonnancement de graphes.

ii.c. Recours aux préférences sélectives

Wilks (1975) développe la « sémantique préférentielle », qui prend en compte des préférences (ou restrictions) sélectives. Celles-ci capturent des connaissances encyclopédiques au sujet des mots et de leurs relations – *e.g.* manger-nourriture ; boire-boisson – afin de déterminer un ensemble de mots cohérents pour tous les mots d'une phrase. Bien qu'intuitive, cette approche souffre d'une forme de circularité bloquante : la WSD requiert de grands ensembles de préférences sélectives, qui elles-mêmes requièrent de bonnes connaissances sémantiques. Pour dépasser le problème, des outils ont été développés pour acquérir automatiquement les liens préférentiels (McCarthy & Carroll, 2003)

Les premiers systèmes de sémantique préférentielle observaient des associations mot-à-mot, puis des auteurs ont cherché à rendre les préférences sélectives progressivement plus globales : mot-à-classe sémantique (Resnik, 1993), puis classe-à-classe (Agirre & Martinez, 2001).

Par la suite, les préférences sélectives ont pu être incluses dans des systèmes de WSD plus larges, utilisant également des données grammaticales (POS) et d'autres sources de connaissances (Stevenson & Wilks, 2001). Cependant, les systèmes utilisant les préférences sélectives peinent à dépasser les performances d'une *baseline* Lesk ou MFS (cf. ii.d).

ii.d. Recours aux propriétés du langage humain

Paradoxalement, certaines connaissances fondamentales à propos du langage n'ont été exploitées que relativement tardivement dans l'histoire de la désambiguïsation lexicale.

La désambiguïsation par « *most frequent sense* » (MFS) démontre des performances qui se sont longtemps avérées difficiles à dépasser pour les chercheurs. Cette approche s'appuie sur l'observation largement admise selon laquelle un des sens du mot se démarque en général assez bien en termes de fréquence (Zipf, 1949). Les performances du modèle MFS sont utilisées comme *baseline* pour comparer les systèmes de désambiguïsation depuis Gale et al. (1992a). Cette méthode nécessite des données de distribution sémantiques, qui ne sont pas disponibles pour toutes les langues et dépendent du corpus dont elles découlent.

D'autres heuristiques s'appuient sur l'idée qu'un mot préserverait le même sens tout au long d'un même texte (*one-sense-per-discourse* - (Gale et al., 1992b), ainsi qu'au sein d'une même collocation (*one-sense-per-collocation* - Yarowsky, 1993). En revanche, plus les distinctions sémantiques reconnues sont fines, plus ces postulats perdent en vigueur (Krovetz, 2000; Martinez & Agirre, 2000, respectivement).

ii.e. Systèmes *knowledge-based* modernes

Les méthodes actuelles emploient souvent des algorithmes de graphes sur les réseaux sémantiques : marche aléatoire (Agirre et al., 2014 - UKB), approximation de clique (Moro et al., 2014 - Babelfy), ou encore théorie des jeux (Tripodi & Navigli, 2019).

Les approches les plus efficaces sont SyntagRank (Scozzafava et al., 2020) et $SREF_{KB}$ (Wang & Wang, 2020). Le premier modèle est purement basé sur des graphes, et applique l'algorithme *Personalized PageRank* sur la portion de BabelNet issue de WordNet, améliorée avec des relations du corpus WNG et SyntagNet (Maru et al., 2019), une ressource manuelle faisant état de relations entre des synsets dont les sens forment une collocation. Par son association avec BabelNet, SyntagRank a pu être utilisé sur différentes langues, tandis que le second modèle ($SREF_{KB}$) n'a été testé que sur l'anglais. $SREF_{KB}$ est une approche vectorielle tirant parti de représentations de mots contextualisées et d'*embeddings* sémantiques. Ces derniers sont obtenus au moyen de BERT (Devlin et al., 2019) sur des exemples et gloses de WN ainsi que d'autres contextes extraits automatiquement du Web. Bevilacqua et al. (2021) avancent que l'utilisation des exemples d'usages de WN représente une forme de supervision, étant donné que ces exemples sont créés manuellement.

iii. *Méthodes supervisées*

Ce n'est que dans les années 80[15] que l'utilisation de modèles statistiques basés sur des corpus annotés sémantiquement se répand dans le champ de la linguistique. Les méthodes supervisées atteignent leur apogée dans les années 1990 et sont – depuis lors – largement considérées comme les plus efficaces (Bevilacqua et al., 2021, p. 4332).

Weiss (1973) montre que les règles de désambiguïsation peuvent être apprises sur base de corpus sémantiquement labellisés par des annotateurs humains. Toutes les approches supervisées supposent donc l'utilisation d'un inventaire de sens préexistant. Les travaux de Kelly & Stone (1975), de Brown et al. (1991) et de Black (1988) anticipent les approches statistiques et *machine learning* qui deviendront centrales dans les années 2000.

[15] Avec le développement des ressources computationnelles et des capacités de stockage des machines.

Les procédures supervisées peuvent être triées selon les types d'induction qu'elles utilisent : (iii.a) méthodes probabilistes, (iii.b) méthodes *exemplar-based*, (iii.c) méthodes par règles, (iii.d) méthodes d'ensemble, (iii.e) classifications linéaires et approches par kernels. Nous terminerons la présentation des méthodes supervisées par une présentation des réseaux de neurones actuels (iii.f) et des méthodes dites « semi-supervisées » (iii.g).

iii.a. Méthodes probabilistes

L'application de modèles de statistiques bayésiennes permet d'assigner à une occurrence d'un mot le sens qui maximise la probabilité conditionnelle étant donné les caractéristiques du contexte (*Naive Bayes Classifier*).

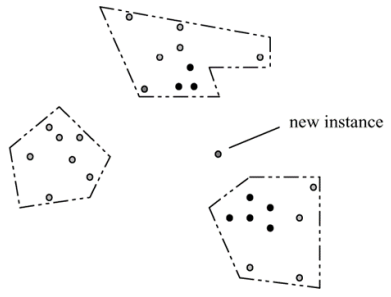
Le postulat d'indépendance des propriétés du contexte est inhérent à l'utilisation du Théorème de Bayes. Bruce & Wiebe (1994) s'éloignent de ce postulat d'indépendance, mais trop de paramètres sont alors à estimer. Malgré le postulat d'indépendance, les *Naive Bayes Classifiers* ont obtenu les meilleures performances jusqu'au début des années 2000 (Navigli, 2009).

D'autres travaux ont également fait usage de l'entropie maximale (Berger et al., 1996) à des fins de désambiguïsation lexicale.

iii.b. Méthodes exemplar-based

En représentant l'information sémantique sous forme de vecteurs, il devient possible de mesurer la similarité entre une occurrence et un exemple de prototypique de sens dans un espace vectoriel multi-dimensionnel. L'algorithme de ce type le plus connu est celui des k-plus-proches-voisins (*k-nearest-neighbors*, ou kNN).

Une recherche de *k* exemples proches de l'occurrence à analyser est effectuée, puis les vecteurs sémantiques des plus proches voisins sont moyennés pour prédire le sens. Plusieurs mesures de similarité entre les vecteurs existent (cosinus[16], distance de Hamming, *modified value difference metric* [Cost & Salzberg, 1993]). La meilleure valeur de *k* et les paramètres à représenter dans les vecteurs sémantiques sont également à déterminer empiriquement. Les modèles « *exemplar-based* » représentaient l'état de l'art à l'aube des années 2010 : Navigli (2009) mentionne notamment Hoste et al. (2002) et Decadt et al. (2004).



La figure 3 représente visuellement la logique des algorithmes kNN : les exemples associés à un même sens sont entourés de pointillés et les points noirs représentent les plus proches voisins de l'occurrence à désambiguïser. Celle-ci sera assignée à la classe du dessous, comptant cinq exemples proches.

Figure 3 - fonctionnement d'un algorithme kNN

L'efficacité de l'approche kNN a été réappuyée par LLMS (Loureiro & Jorge, 2019) qui ont montré qu'une simple méthode 1-NN basée sur les *embeddings* d'ELMo et de BERT (cf. section iii.f) surpassait tous les réseaux de neurones de l'époque.

[16] Le cosinus entre deux vecteurs correspond au produit scalaire de ces vecteurs divisé par le produit de la norme des vecteurs.

iii.c. Méthodes par règles

Les listes de décision (Figure 4) consistent en un ensemble de règles « *if-then-else* » pondérées à l'aide d'un corpus d'entraînement. Lors de la désambiguïsation d'un nouveau mot, les règles sont vérifiées les unes après les autres par ordre décroissant d'importance, et la première règle qui correspond aux propriétés du mot à désambiguïser en sélectionne le sens. Les listes de décision figurent parmi les systèmes plus efficaces lors de Senseval-1 (Yarowsky, 2000).

Les règles en question peuvent également être représentées sous la forme d'une structure parcourue récursivement (arbres de décision – Figure 5). Les prédictions sur le sens des mots sont effectuées lorsque l'algorithme atteint une feuille de l'arbre.

Table III. An Example of Decision List

Feature	Prediction	Score
account with bank	Bank/FINANCE	4.83
stand/V *on*/P . . . bank	Bank/FINANCE	3.35
bank *of blood*	Bank/SUPPLY	2.48
work/V . . . bank	Bank/FINANCE	2.33
the left/J bank	Bank/RIVER	1.12
of the bank	-	0.01

Figure 4 - liste de décision

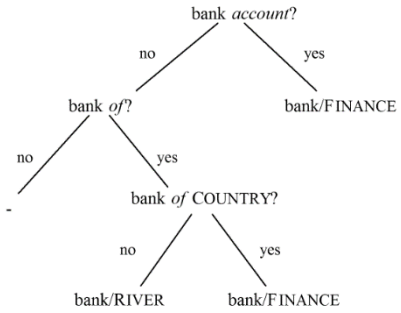


Figure 5 - arbre de décision

L'étude comparative de Mooney (1996) a montré que les arbres de décision ne sont pas les mieux adaptés à la tâche de WSD, mais ceux-ci peuvent toutefois s'allier dans des méthodes d'ensemble.

iii.d. Méthodes d'ensemble

Les méthodes d'ensemble combinent un ou plusieurs types d'algorithmes d'apprentissage afin d'aborder les données d'entraînement depuis différents points de vue. Klein et al. (2002) et Florian et al. (2002) ont étudié la combinaison de méthodes supervisées. Les méthodes d'ensemble représentaient l'état de l'art – pour la désambiguïsation d'échantillons lexicaux – au moment de Senseval-2 (Edmonds & Cotton, 2001)[17].

Les classifieurs peuvent s'agencer ensemble selon plusieurs stratégies : vote majoritaire (le vote le plus fréquent parmi les classifieurs est retenu), mixture probabiliste (les résultats de plusieurs classifieurs probabilistes sont sommés), combinaison de rankings ou encore AdaBoost (*adaptative boosting* - Freund & Schapire, 1997).

AdaBoost est un moyen de construire un classifieur « fort » par la combinaison linéaire de plusieurs règles de classification simples et peu précises (autant de classifieurs « faibles »). Ces règles sont apprises séquentiellement sur le corpus, et les occurrences mal classées sont mises en évidence pour les phases suivantes d'apprentissage : l'algorithme est adaptatif. Escudero et al. (2000) illustre l'utilisation d'*AdaBoost* pour la WSD.

[17] Pour la WSD *all-words* non supervisées, les méthodes d'ensemble dépassent l'état de l'art au moment de Navigli (2009).

iii.e. Classifications linéaires et approches par kernels

Les classifieurs linéaires (binaires) reposent sur la volonté de classer des exemples d'entraînement positifs et négatifs selon un hyper-plan linéaire dans un espace vectoriel multi-dimensionnel. Une telle classification est permise notamment par les machines à vecteurs de support (SVM). Dans ces systèmes, les occurrences positives et négatives du corpus d'entraînement sont représentées par des vecteurs sémantiques (vecteurs de support). Le classifieur optimal est celui qui maximise la distance entre les occurrences positives et négatives. L'intuition géométrique sous-jacente aux SVM est représentée par la figure 6.

Il arrive également que les occurrences ne puissent être séparées de manière linéaire, il faut alors autoriser une marge d'erreur pour déterminer le meilleur hyperplan (*soft margin),* ou encore utiliser une fonction non-linéaire (*kernel*) pour classifier les instances.

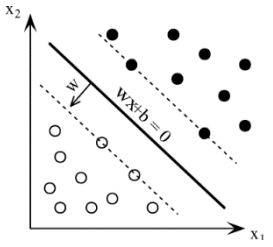
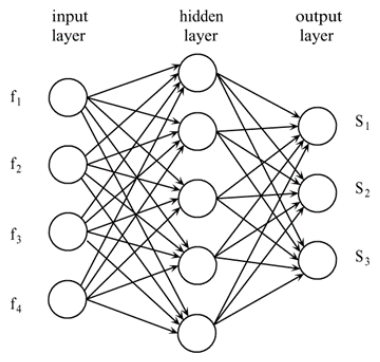


Figure 6 - principe des SVM

Les premières utilisations des SVM en WSD remontent au début du XXIe siècle (Murata et al., 2001; Lee & Ng, 2002). Plusieurs travaux ont ensuite montré les bonnes performances de modèles SVM lors de Senseval-3 (Snyder & Palmer, 2004), notamment Strapparava et al. (2004), Agirre & Martinez (2004), Cabezas et al. (2004) et Escudero et al. (2004). Le modèle de désambiguïsation *It Makes Sense* (IMS - Zhong & Ng, 2010) – utilisant les SVM avec trois types des données (POS environnant le mot cible, mots environnants et collocations locales) – a été longtemps considéré comme l'un des meilleurs systèmes supervisés, et servira de base pour introduire les *embeddings* dans le champ de la désambiguïsation lexicale (Iacobacci et al., 2016, cf. infra).

iii.f. Aujourd'hui : l'ère neuronale

De nos jours, les modèles supervisés comptent presque unanimement sur les architectures neuronales. Ces dernières sont caractérisées par une architecture interconnectée et multi-couche d'unités linéaires (ou neurones). Les réseaux de neurones sont une alternative pour modéliser des fonctions non-linéaires complexes. L'approche connexionniste sur laquelle reposent – philosophiquement – les réseaux de neurones était déjà utilisée dans les années 80 et 90 pour représenter les relations sémantiques sous forme de réseaux (cf. section B.4.i.c). À la fin du XXe siècle, Towell & Voorhees (1998) présentent un réseau de neurones simple (*feed-forward*) et supervisé pour désambiguïser des mots hautement ambigus, mais la révolution neuronale en WSD devra attendre les années 2015.



La figure 7 représente un réseau de neurones très simple de type « *feed-forward* » (l'information ne circule que dans un seul sens). Grossièrement, nous pouvons expliquer que des paires [`caractéristiques d'entrée, sortie attendue`] sont données au programme, et que les poids inter-neuronaux sont ensuite progressivement mis à jour jusqu'à ce que la sortie voulue ait un plus haut taux d'activation que les autres sorties possibles.

Figure 7 - architecture d'un réseau de neurones feed-forward (Navigli, 2009)

En 2016, Iacobacci et collègues observent une augmentation significative des performances en utilisant les *embeddings sémantiques*[18] comme caractéristique supplémentaire pour un modèle WSD basé sur les SVM (cf. section iii.e). À la différence des représentations vectorielles présentées jusqu'ici, les *embeddings* sont *appris* par modélisation neuronale du langage (NLM). Les meilleures représentations sémantiques ont été longtemps associées aux mots (word2vec – Mikolov et al., 2013 ; fastText – Bojanowski et al., 2017), mais les développement plus tardifs en NLM ont permis de représenter les contextes, ce qui s'avère particulièrement efficace en désambiguïsation lexicale (context2vec – Melamud et al., 2016 ; ELMo – Peters et al., 2018 ; BERT – Devlin et al., 2019). Le séisme causé par ce dernier en linguistique computationnelle[19] n'a pas épargné la désambiguïsation lexicale, et la plupart des travaux subséquents feront usage des *embeddings* pré-entraînés de BERT.

La plupart des méthodes exploitent l'apprentissage par transfert[20], et requièrent un transformeur[21]/*embedder* pré-entraîné pour atteindre les meilleures performances. Dans les *token taggers* (cf. section v.), les représentations contextuelles sont données tantôt à un réseau de neurones *feed-forward* (GLU – Hadiwinoto et al., 2019) ou a à une pile de transformeurs (SVC – Vial et al., 2019). Ces approches ont produit une large amélioration par rapport aux modèles initialisés au hasard (Raganato, Delli Bovi, et al., 2017). Toutefois, les performances sont limitées par l'entropie croisée catégorielle souvent utilisée pour l'entraînement. Conia & Navigli (2021) ont prouvé que l'entropie croisée binaire est plus efficace comme fonction de perte, attendu qu'elle permet des annotations multiples pour les occurrences du corpus d'entraînement, et – par conséquent – offre une plus grande flexibilité aux systèmes.

Parmi les principaux défauts des réseaux de neurones, nous pouvons citer la difficulté d'interprétation des résultats, le besoin de grandes quantités de données d'entraînement et l'ajustement complexe des paramètres du modèle.

iii.g. Méthodes semi-supervisées

Pour les modèles supervisés, le *knowledge acquisition bottleneck* se fait particulièrement ressentir : les données annotées sémantiquement sont rares (d'autant plus en dehors du champ anglophone), et l'obtention de corpus sémantiquement annotés représente un travail humain de longue haleine. Afin de pallier la modestie des corpus disponibles[22], certains auteurs ont développé des méthodes automatiques pour élaborer des données

[18] « *Embedding* » peut se traduire par « plongement » ou « enchâssements », mais l'idée serait mieux rendue par « vectorisation » (https://fr.wikipedia.org/wiki/Word_embedding). Pour éviter toute ambiguïté, nous utiliserons le terme d'origine.

[19] BERT a permis d'établir un nouvel état de l'art pour 11 tâches distinctes en NLP (Devlin et al., 2019).

[20] Visant à exploiter les connaissances acquises via une (ou plusieurs) tâches sources pour gérer une tâche cible.

[21] Modèle neuronal utilisant une architecture d'encodeur-décodeur et des mécanismes d'attention pour gérer des données séquentielles.

[22] En réalité, il y a tellement de sens qu'il est virtuellement impossible d'établir un corpus d'entraînement – aussi large soit-il – recouvrant tout le vocabulaire d'une langue, et tous les sens possibles pour ce vocabulaire. Les techniques de lissage permettent de s'assurer que les événements non observés n'aient pas une probabilité nulle. Par ailleurs, certains auteurs proposent de regrouper des mots (1) appartenant à la même catégorie sémantique (*class-based models* - Yarowsky, 1992; Resnik, 1993) ou (2) sémantiquement similaires (Dagan et al., 1993), en supposant que ces mots aient des fonctionnements et des contextes similaires.

d'entraînement étiquetées sémantiquement. Les classifieurs basés sur cette stratégie sont dits « semi-supervisés », en ce sens qu'ils requièrent moins d'intervention humaine.

À cet égard, les méthodes de *bootstrapping* ont l'avantage d'augmenter graduellement leurs connaissances : sur la base de quelques occurrences annotées manuellement (typiquement entre 10 et 30 *seeds*), l'algorithme construit un modèle de classification simple qui taggue ensuite d'autres contextes, qui – au-delà d'un seuil de confiance – seront utilisés comme données d'entraînement lors de l'itération suivante, et ainsi de suite jusqu'à ce qu'une grande quantité de données aient pu être étiquetées.

Le *bootstrapping* peut consister en une comparaison de vecteurs sémantiques similaires (Schütze, 1993; Mihalcea & Faruque, 2004) ou s'appuyer sur des corpus alignés (Gale & Church, 1993), mais l'algorithme de Yarowsky (1995) reste l'approche semi-supervisée la plus exemplaire (Figure 8). Celle-ci s'appuie sur quelques exemples annotés et sur une *decision list* pour étiqueter un grand nombre d'exemples. Quelques propriétés fondamentales du langage sont au cœur du processus d'apprentissage et permet de labelliser de nouvelles occurrences avec certitude : *one-sense-per-discourse* et *one-sense-per-collocation*.

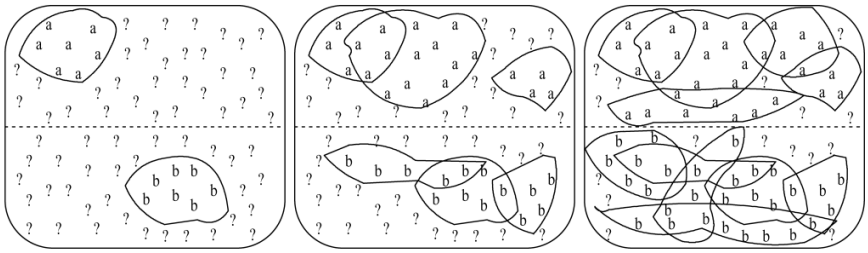


Figure 8 - trois itérations de l'algorithme de Yarowsky (1995). De plus en plus de données sont étiquetées et réutilisées.

La génération automatique de données d'entraînement n'a pas été abordée exclusivement sous l'angle du *bootstrapping* : Leacock et al. (1998) utilisent la technique des *monosemous relatives* (identification de synonymes non ambigus) pour trouver des instances d'entraînement dans un grand corpus brut. Enfin, l'apprentissage actif (*active learning*) aspire à trouver les exemples les plus informatifs à tagguer, en vue de diminuer le coût d'acquisition des données d'entraînement (Dagan & Engelson, 1995; Argamon-Engelson & Dagan, 1999).

iv. *Méthodes non supervisées*

L'hypothèse sous-jacente aux systèmes non supervisés est que des sens similaires tendent à apparaître dans des contextes similaires (hypothèse distributionnelle). Dès lors, il possible de grouper les usages – d'un corpus brut – selon une signification « virtuelle », puis d'induire cette signification. D'un côté, les méthodes non supervisées ne dépendent pas d'un recueil de sens ou de données annotées – elles sont donc plus adaptables et indépendantes de la langue, mais d'un autre côté, il est ardu de cartographier les significations induites dans un

inventaire sémantique humainement compréhensible. Cette étape requiert généralement une intervention manuelle[23].

En un mot, les approches non supervisées n'assignent pas un sens donné aux mots, mais groupent les mots selon les contextes dans lesquels on les retrouve, avant d'associer *a posteriori* un sens à chacun de ces contextes. La tâche de *classification* se mue – dans le cadre des approches non supervisées – en une tâche de *discrimination* des sens, il est donc plus difficile d'évaluer les systèmes non supervisés selon les cadres traditionnels. Les méthodes de désambiguïsation non supervisées diffèrent tant des approches supervisées qu'on y réfère parfois par le terme *word sense induction* (WSI).

Il est intéressant de remarquer que ni l'état de l'art de Raganato, Camacho-Collados, et al. (2017), ni celui de Bevilacqua et al. (2021) ne font mention des approches non-supervisées. Deux hypothèses peuvent être émises à ce sujet : soit les performances de ces modèles ne sont plus suffisantes par rapport aux méthodes supervisées (ou hybrides) développées à présent, soit les tâches de WSD et WSI sont devenus trop dissemblables. Dans la suite, nous présentons somme toute les quatre approches principales de désambiguïsation non supervisée : (iv.a) regroupement de contexte (*context clustering*), (iv.b) regroupement de mots (*word clustering*), (iv.c) graphes de co-occurrences et (iv.d) clustering probabiliste.

iv.a. Context clustering

Un premier ensemble de méthodes non supervisées fait appel à la notion de *context clustering*. Dans cette configuration, chaque occurrence des mots du corpus est représentée comme un *vecteur de contexte*. Ces vecteurs peuvent être de premier ordre, c'est-à-dire représenter directement certaines caractéristiques lexicales et syntaxiques du contexte (Pedersen & Bruce, 1997), ou de second ordre, c'est-à-dire être conçus comme la moyenne des vecteurs de mots qui apparaissent dans ce contexte (Schütze, 1998). Ces vecteurs de contexte peuvent être vus comme une approximation du contexte sémantique. Les vecteurs de contexte sont ensuite regroupés (*clustered*) au moyen d'un *clustering algorithm* qui mesure la similarité entre des vecteurs : *context group discrimination* (Schütze, 1998), *clustering agglomératif* (Pedersen & Bruce, 1997). Chacun des *clusters* ainsi créés identifie l'un des sens du mot cible. Désambiguïser une nouvelle occurrence revient alors à comparer le vecteur de contexte de cette occurrence à ceux des différents *clusters* établis lors de l'entraînement.

Une grande quantité des données d'entraînement est requise pour déterminer une distribution significative des co-occurrences. Cela peut être réglé en augmentant les vecteurs de mots avec les mots pleins que l'on retrouve dans leurs gloses (Purandare & Pedersen, 2004). Cette approche est toutefois plutôt hybride, en ce sens qu'elle utilise un inventaire de sens afin de faire fonctionner un système non supervisé.

Une autre tradition des approches non supervisées consiste à utiliser l'information des corpus bilingues alignés (P. F. Brown et al., 1991; Dagan et al., 1991; Gale et al., 1992a), en

[23] Une autre solution peut être de se baser sur certains mots d'un *cluster* (*e.g.* un des sens de *ligne* pourrait être défini par [*téléphone, occupé, appeler*…]). Une autre solution (McCarthy et al., 2004) serait de revenir aux mesures de similarité sémantiques évoquées (cf. section ii.b) pour trouver – dans WordNet – le sens majoritairement associé aux plus proches voisins du *cluster* contextuel.

supposant que les traductions d'un mot dans une langue cible dépendent du sens du mot dans la langue source (cf. section i.) [24]. Ide et al. (2001) s'inscrivent dans cette mouvance en utilisant des vecteurs contextuels multilingues pour déterminer le sens des mots.

iv.b. Word clustering

Dans la section précédente, nous avons vu comment les contextes pouvaient être regroupés pour discriminer les sens. Une seconde approche consiste à identifier et grouper les mots sémantiquement similaires – qui peuvent donc véhiculer un sens particulier.

Parmi les méthodes de *word clustering*, nous pouvons citer le *clustering by committee* (CBC - Lin & Pantel, 2002). Pour chaque mot cible, un ensemble de mots – éventuellement synonymes – est trouvé, et les relations sont encodées dans une matrice de similarité. Pour pouvoir calculer ces similarités, chaque mot est représenté sous forme d'un vecteur de caractéristiques syntaxiques[25]. En second lieu, des ensembles de mots similaires (comités de mots) sont établis récursivement avec un algorithme standard de *clustering*. À la fin de cette procédure, chaque mot du corpus est inclus dans un comité, si bien que les comités correspondent à des sens distincts. Chacun des mots de ce comité dispose d'un vecteur de caractéristiques, et le centroïde du comité peut donc être calculé. Pour la désambiguïsation, chacun des mots est assigné au comité dont il est le plus proche (par mesure de la similarité entre le vecteur de caractéristiques du mot et le centroïde du comité).

Une autre approche opère la désambiguïsation sur base de triplets de mots (Bordag, 2006). Cette méthode repose sur le postulat « *one-sense-per-collocation* » et regroupe des co-occurrences de triplets en utilisant leurs intersections comme caractéristiques. La désambiguïsation qui en résulte est « hautement précise » (Navigli, 2009)

iv.c. Graphes de co-occurrences

Ces méthodes sont liées aux approches par *word clustering*, mais construisent et analysent un graphe de co-occurrences lexicales afin d'identifier un ensemble de sens pour un mot donné. Les co-occurrences sont tantôt obtenues au départ de relations grammaticales (Widdows & Dorow, 2002) ou de fréquences de contiguïté (HyperLex – Véronis, 2004) dans un paragraphe ou un contexte plus large.

Toutefois, des approches efficaces telles que HyperLex – un algorithme de graphe basé sur l'identification des sens associés à un mot dans un graphe de co-occurrences –, ou encore celle de Agirre et al. (2006) – une adaptation de l'algorithme de graphe PageRank à la WSD – sont confrontés au réglage laborieux d'un grand nombre de paramètres. Avec un paramétrage optimal, HyperLex et PageRank atteignent des performances proches des systèmes supervisés pour la désambiguïsation – de substantifs – évalués lors de Senseval-3 (Agirre et al., 2006).

Pour éviter la problématique du paramétrage, Navigli & Crisafulli (2010) proposent un algorithme de graphe basé sur des patrons graphiques simples. Ces patrons carrés et triangulaires ont pour objectif d'identifier les sens en utilisant les propriétés structurelles locales du graphe de co-occurrences.

---

[24] Cette approche présente l'avantage d'être cohérente dans le contexte de la traduction automatique. En revanche, l'hypothèse de départ n'est pas toujours vérifiable : il arrive que la polysémie soit conservée d'une langue à l'autre (*e.g.* EN. *mouse* – FR. *souris*).

[25] Pour utiliser le CBC, une phase de *parsing* est donc requise en amont.

iv.d. Clustering probabiliste

Certains auteurs (Brody & Lapata, 2009) choisissent d'adopter une approche probabiliste bayésienne, et formalisent la désambiguïsation non supervisée dans un modèle génératif. D'abord, une distribution des sens est établie pour chaque mot ambigu. Par la suite, plusieurs mots de contexte sont générés selon cette distribution. De cette manière, différents sens peuvent être obtenus, et chacun de ces sens suit une distribution différente selon les mots.

Toutefois, cette approche se concentre plutôt sur l'estimation des distributions sens-contextes. Une fois la distribution des sens des mots calculée, le système applique une annotation sémantique par MFS.

v. *Méthodes hybrides*

L'ambition de combiner différents systèmes et sources de données n'est pas récente (v. Yarowsky et al., 2001; Florian et al., 2002; Stevenson & Wilks, 2001; Montoyo et al., 2005). Cependant, les possibilités de combinaisons explorées à travers le temps sont trop nombreuses pour être présentées exhaustivement dans ce travail. Retenons, à ce stade, que les méthodes « hybrides » font en général appel à un système supervisé amélioré par d'autres types de connaissances lexicales. Pour Agirre & Edmonds (2007), il ne suffit pas de combiner : il faut également tirer parti de l'information disponible le plus judicieusement possible.

Depuis la révolution neuronale (~ 2015), les approches supervisées – souvent hybrides[26] - ont largement assis leur domination, si bien qu'elles remettent en question la distinction ternaire entre *knowledge-based*, supervisé et non supervisé. À l'heure actuelle, les systèmes supervisés neuronaux sont plus volontiers subdivisés en trois familles : les classifieurs d'occurrences (*token tagger*), les méthodes *1-NN vector based* (s'appuyant sur l'algorithme des k-plus-proches-voisins), et les classifieurs de séquences. À ces trois familles se greffent encore quelques méthodes purement *knowledge-based*, systématiquement moins efficaces.

Selon Bevilacqua et al. (2021), le type d'informations additionnelles que les systèmes sont capables d'utiliser est plus pertinent pour trier ces derniers que les modalités des architectures neuronales. Dans cette section, nous présenterons trois types d'hybridations productives – et les systèmes associés les plus actuels : (v.a) modèles supervisés exploitant les gloses, (v.b) modèles supervisés exploitant les relations des graphes sémantiques et (v.c) modèles supervisés exploitant d'autres types de connaissances.

v.a. Modèles supervisés exploitant les gloses

Les définitions offrent une manière simple et compréhensible de clarifier les distinctions sémantiques. Ces définitions – ou gloses – ont montré (au moins depuis Lesk, 1986) leur utilité en WSD, et de nombreuses façons d'exploiter celles-ci ont, depuis lors, été explorées dans la littérature. Les gloses peuvent être encodées sous forme de vecteurs en moyennant les représentations contextuelles de leurs mots constitutifs. Ces vecteurs (*embeddings* de sens/sémantiques) peuvent alors être incorporés dans les approches par 1-NN (1-plus-proche-voisin) ou les *token taggers*.

[26] Les architectures connexionnistes permettent d'exploiter plus aisément divers types d'informations lexico-sémantiques.

En particulier, les approches par 1-NN bénéficient grandement de la concaténation des vecteurs de gloses aux représentations (*embeddings*) supervisés (Loureiro & Jorge, 2019). D'autres approches 1-NN plus sophistiquées recourent aussi aux gloses : SensEmBERT (Scarlini et al., 2020a), ARES (Scarlini et al., 2020b), et SREF (Wang & Wang, 2020). Ces derniers diffèrent dans leur manière d'extraire automatiquement des contextes additionnels pour construire la partie supervisée des *embeddings* sémantiques. ARES atteint les meilleurs résultats en tirant parti des relations de collocations entre les sens pour trouver de nouvelles phrases d'entraînement dans Wikipédia. Berend (2020) a montré que les *embeddings* sémantiques peuvent être rendus creux[27] (*sparse*) en appliquant le codage parcimonieux (*sparse coding*).

Une autre utilisation des *embeddings* sémantiques (incluant l'information des gloses) est de fournir des poids pour la couche de classification des architectures des *token taggers*. EWISE (Kumar et al., 2019) créé des représentations sémantiques en entraînant un encodeur de gloses avec une fonction de coût par triplet sur WordNet. EWISE est le meilleur modèle ne tirant pas profit *d'embeddings* pré-entraînés. EWISER (Bevilacqua & Navigli, 2020), pour sa part, adapte des *embeddings* sémantiques basés sur des modèles linguistiques pré-entraînés (SensEmBERT et LMMS), et atteint des résultats proches de l'état de l'art. Enfin, BEM (Blevins & Zettlemoyer, 2020) tire pleinement profit de l'idée d'entraîner en parallèle des représentations textuelles et sémantiques, et met ceci en pratique en faisant appel à deux transformeurs distincts pour encoder – d'une part – le contexte du mot cible et – d'autre part – la définition de ce mot.

Par ailleurs, les gloses peuvent être utilisées par les systèmes qui abordent la désambiguïsation comme un problème de *classification séquentiel* (Huang et al., 2019 - GlossBERT ; Yap et al., 2020). Plutôt que d'observer chaque mot dans son contexte, la classification de séquence considère en premier lieu le contexte, et classifie ensuite tous les membres de ce contexte. L'objectif de cette méthode est de privilégier la cohérence des annotations entre elles. Bien que ces systèmes atteignent des performances compétitives, ils restent moins efficaces que les *token taggers*, car ils doivent traiter une même phrase pour chacun de ses mots et chacune des définitions possibles. Barba et al. (2021 - ESCHER) résout ce problème en abordant la désambiguïsation comme un problème d'extraction d'intervalle : étant donné un mot et toutes ses définitions concaténées, un modèle doit trouver l'intervalle (dans cette concaténation) qui correspond le mieux à l'utilisation du mot dans la phrase.

Une variante générative de la tâche de classification de séquences a été introduite par Bevilacqua et al. (2020 – Generationary). Au lieu de sélectionner le sens le plus adéquat, le système génère lui-même une description du sens qui sera ensuite comparée aux gloses d'un inventaire sémantique. Ces systèmes présentent l'avantage de ne pas recourir à un inventaire de sens établi, ce qui donne plus de flexibilité au système (pour désambiguïser des termes argots ou néologiques, par exemple). Ce dernier système illustre toute l'importance des gloses en WSD : on ne cherche plus seulement à les utiliser, mais également à les produire.

[27] À l'inverse des *vecteurs denses*, les *vecteurs creux* présentent souvent la valeur 0 dans leurs dimensions.

v.b. Modèles supervisés exploitant les relations

Dernièrement, plusieurs systèmes supervisés – 1-NN et *token taggers* – ont mis en évidence la possibilité d'utiliser les relations lexico-sémantiques du graphe de WordNet en tant qu'informations additionnelles.

Par exemple, LMMS créé des représentations pour les sens absents de SemCor en moyennant les *embeddings* de leurs voisins dans WordNet. SREF fait appel aux relations d'hypéronymie et d'hyponymie pour élaborer un mécanisme *try-again* qui affine les prédictions du modèle. Vial et al. (2019 – SVC) réduisent le nombre de classes de sortie en associant chaque sens à un ancêtre dans la taxonomie de WordNet.

Parmi les *token taggers*, EWISE utilise la structure de graphe de WordNet pour entraîner son *embedder* de gloses ; tandis que EWISER montre que l'entièreté du graphe de WordNet peut être intégrée à l'architecture du modèle. Le système de Conia & Navigli (2021) inclut aussi de l'information relationnelle : la fonction de coût par entropie croisée binaire retient tous les sens liés au sens principal comme pertinents.

En définitive, l'utilisation d'informations relationnelles est devenue un lieu commun dans la désambiguïsation supervisée, mais leur intégration dans les méthodes de classification séquentielles n'a pas encore été investiguée.

v.c. Modèles supervisés exploitant d'autres types de connaissances

Il a longtemps été postulé que les traductions permettraient d'améliorer les performances des systèmes de désambiguïsation. Cette hypothèse a récemment été prouvée par Luan et al. (2020), qui ont réussi à affiner les résultats d'un système de WSD arbitraire grâce aux traductions de BabelNet. Pour ce faire, les traductions des sens trouvés (BabelNet) sont comparées avec les traductions du mot cible (telles qu'obtenues par un système de traduction neuronale).

Dans une autre direction, Calabrese et al. (2020 – EVilBERT) utilisent des images de la base de données BabelPic pour construire des vecteurs de gloses multimodaux, qui sont plus efficaces que les *embeddings* traditionnels pour initialiser les poids de la couche de classification d'une architecture neuronale.

Pour finir, des contextes de Wikipédia ou de recherches Web peuvent également être considérés comme des sources d'informations additionnelles pour construire des *embeddings* de sens (Scarlini et al., 2020a, 2020b; Wang & Wang, 2020).

B.6. Évaluation de la désambiguïsation lexicale

i. *Senseval / SemEval*

Il a longtemps été difficile d'établir un cadre d'évaluation standard pour les divers modèles de désambiguïsation lexicale. Une discussion lors d'un séminaire sponsorisé par le ACL Special Interest Group on the Lexicon (SIGLEX) à propos de l'évaluation des *taggers* sémantiques (Resnik & Yarowsky, 1997) a pointé du doigt cette faillite, et donné le jour aux séminaires Senseval. Ceux-ci seront plus tard renommés SemEval, car ils aborderont progressivement d'autres problématiques liées à la sémantique (*e.g.* la substitution lexicale).

Avant Senseval, aucune échelle clairement définie ne permettait d'évaluer les performances de manière homogène. Les évaluations antérieures étaient construites variablement d'article en article (mots tests, corpus d'entraînements, inventaires sémantiques, mesures d'efficacité… différents). En outre, les évaluations étaient fréquemment menées à petite échelle et pour des distinctions sémantiques grossières. Le projet Senseval a permis d'homogénéiser l'évaluation, et a encouragé le développement de la recherche en désambiguïsation lexicale en se présentant sous la forme de compétitions périodiques.

Pour commencer, Senseval a établi une limite inférieure (*baseline*) et une limite supérieure pour mesurer les compétences. La première correspond souvent à l'heuristique du MFS, qui fournit déjà des résultats élevés (75% de décisions correctes pour une ambiguïté binaire – Gale et al., 1992a) ; l'algorithme de Lesk fait également partie des *baselines* fréquentes. La seconde consiste généralement en une mesure des compétences humaines (96.8% de précision pour une ambiguïté binaire – Gale et al., 1992a). Les modèles sont évalués en termes de précision et de rappel. La première mesure détermine la qualité des décisions prises par le système (les erreurs grossières étant plus pénalisantes que les erreurs plus subtiles), tandis que la seconde mesure la couverture des systèmes.

Néanmoins, ces mesures n'ont pas pu être mises en application sans le développement de corpus de tests annotés manuellement par rapport à un inventaire sémantique consensuel. Ainsi, les séminaires Senseval/SemEval ont également fourni de précieuses données d'évaluation. La WSD en anglais bénéficie de la suite d'évaluation de Raganato, Camacho-Collados, et al. (2017), qui regroupe cinq *gold-standards* d'évaluation *all-words* : Senseval-2 (Edmonds & Cotton, 2001), Senseval-3 (Snyder & Palmer, 2004), SemEval-2007 (Pradhan et al., 2007), SemEval-2013 (Navigli et al., 2013) et SemEval-2015 (Moro & Navigli, 2015).

L'influence de Senseval/SemEval ne se limite pas uniquement à l'évaluation. Par exemple, le Senseval-3 aura certainement suscité une vague de désintérêt pour la désambiguïsation lexicale. En effet, les scientifiques présents se sont globalement accordés pour observer que la désambiguïsation avait atteint un palier, et ne générerait probablement plus de recherches fondamentalement nouvelles dans les années à venir. Une telle observation peut, de façon plus optimiste, être symptomatique d'une nécessité d'explorer de nouvelles pistes. De la même manière, les meilleurs modèles actuels – tels qu'évalués par le cadre d'évaluation défini précédemment – atteignent des performances proches, voire supérieures, à la compétence humaine. Ceci révèle, selon Bevilacqua et al. (2021) l'inadéquation des références d'évaluation, et le besoin de construire de nouveaux ensembles de test plus ambitieux.

ii. *Comparaison des approches actuelles*

Dans la section B.5, nous avons sporadiquement montré quels types de systèmes obtenaient les meilleurs résultats sur les premières éditions de Senseval/SemEval. Au vu du nombre de systèmes de WSD ayant jalonné les vingt dernières années, il nous est bien entendu impossible de tout comparer ; il a semblé plus à propos de reprendre visuellement une comparaison des meilleurs systèmes actuels, et de considérer les systèmes cités par

Bevilacqua et al. (2021) comme l'état de l'art de la désambiguïsation lexicale en anglais de nos jours (Figure 9).[28]

	Kind	System	ALL	S2	S3	S7	S13	S15
KB	(❄)	[Scozzafava *et al.*, 2020, SyntagRank]	71.7	71.6	72.0	59.3	72.2	75.8
	(➡ 📖 ❄ 🔵)	[Wang and Wang, 2020, $SREF_{KB}$]	**73.5**	72.7	71.5	61.5	76.4	79.5
Vector-based 1-nn	(➡ 📖 ❄)	[Loureiro and Jorge, 2019, LMMS]	75.4	76.3	75.6	68.1	75.1	77.0
	(➡ 📖)	[Berend, 2020]	76.8	77.9	77.8	68.8	76.1	77.5
	(➡ 📖)	[Scarlini *et al.*, 2020b, ARES]	**77.9**	78.0	77.1	71.0	77.3	83.2
	(➡ ❄)	[Conia and Navigli, 2020, Conception]	76.4	77.1	76.4	70.3	76.2	77.2
	(➡ 📖 🅰)	[Luan *et al.*, 2020]	76.4	77.2	77.1	69.2	76.1	77.2
	(➡ 📖 ❄ 🔵)	[Scarlini *et al.*, 2020a, SensEmBERT]	-	-	-	-	78.7	-
	(➡ 📖 ❄ 🔵)	[Wang and Wang, 2020, SREF]	77.8	78.6	76.6	72.1	78.0	80.5
Token Classifier	(🤖)	[Hadiwinoto *et al.*, 2019, GLU]	74.1	75.5	73.6	68.1	71.1	76.2
	(🤖 ❄)	[Vial *et al.*, 2019, SVC]	76.7	76.5	77.4	69.5	76.0	78.3
	(🤖 📖 ❄)	[Kumar *et al.*, 2019, EWISE]	71.8	73.8	71.1	67.3	69.4	74.5
	(🤖 📖)	[Blevins and Zettlemoyer, 2020, BEM]	**79.0**	79.4	77.4	74.5	79.7	81.7
	(🤖 📖 🖼)	[Calabrese *et al.*, 2020a, EViLBERT]	75.1	-	-	-	-	-
	(🤖 📖 ❄)	[Bevilacqua and Navigli, 2020, EWISER]	78.3	78.9	78.4	71.0	78.9	79.3
	(🤖 ❄)	[Conia and Navigli, 2021]	77.6	78.4	77.8	72.2	76.7	78.2
Seq. Classif.	(🚌 📖)	[Huang *et al.*, 2019, GlossBERT]	77.0	77.7	75.2	72.5	76.1	80.4
	(🚌 📖)	[Bevilacqua *et al.*, 2020, Generationary]	76.7	78.0	75.4	71.9	77.0	77.6
	(🚌 📖)	[Yap *et al.*, 2020]	78.7	79.9	77.4	73.0	78.2	81.8
	(🚌 📖)	[Barba *et al.*, 2021, ESCHER]	**80.7**	81.7	77.8	76.3	82.2	83.2

Table 1: F1 performance figures of recent WSD systems in the literature. We consider results on the evaluation sets (S)enseval-(2)/(3), (S)emEval 200(7)/20(13)/20(15), and on the concatenation of all of them (ALL). All supervised systems (bottom three blocks) use SemCor *only* as training corpus. The leftmost column indicates the kind of system, i.e., 🦇 knowledge-based, ➡ vector-based 1-nn classifier, 🤖 token tagger, 🚌 sequence tagger and, in parentheses, the additional resources leveraged by each model: WordNet glosses (📖), relational information (❄), text from Web (🔵), automatic translations (🅰), or visual information (🖼).

Figure 9 - comparaison des systèmes de WSD les plus récents (Bevilacqua et al., 2021)

L'intégralité des systèmes repris dans cette table, ainsi que la typologie des systèmes utilisée par Bevilacqua et collègues, ont été mentionnés dans la section B.5. Les performances des autres systèmes présentés ne sont pas pour autant dénuées d'intérêt, et permettent notamment d'avoir un aperçu d'autres *baselines* utilisées, ou encore de mieux saisir les nombreux basculements qui se sont opérés dans le champ de la désambiguïsation lexicale. Pour avoir un aperçu quantitatif des différents modèles que nous avons cités, nous ne pouvons qu'encourager le lecteur curieux à s'orienter vers les papiers relatifs à ces derniers, et/ou aux états de l'art successifs de la WSD : notamment Ide & Véronis (1998), Agirre & Edmonds (2007), Navigli (2009), Raganato, Camacho-Collados, et al. (2017), et enfin Bevilacqua et al. (2021).

[28] L'état de l'art de Bevilacqua et al. (2021) peut, selon nous, être considéré comme faisant autorité en mai 2022.

Chapitre 1) – À retenir :

L'ambiguïté sémantique est une caractéristique inhérente au langage humain : les mots peuvent disposer de plusieurs significations en fonction des contextes. Deux grands types d'ambiguïtés existent : l'homonymie et la polysémie.

En traitement automatique des langues, la tâche qui vise à déterminer automatiquement le sens des mots en contexte s'appelle « désambiguïsation lexicale ». Il s'agit d'une tâche compliquée dont les principales difficultés ont été soulevées rapidement après son émergence autour des années 1950.

L'une des principales complications associées à la désambiguïsation lexicale est connue sous le nom de « *knowledge-acquisition bottleneck* » : tous les systèmes de WSD requièrent de l'information, quelle qu'en soit la forme (structurée : réseaux sémantiques, dictionnaires, etc. ou non structurée : corpus annotés sémantiquement ou non).

Les sources d'information sont si importantes pour la désambiguïsation que la typologie des systèmes la plus répandue se base sur ce critère (voir Figure 2). De nous jours, les algorithmes reposent souvent sur des réseaux de neurones, et les représentations vectorielles (*embeddings*) permettent à ces modèles d'encoder plusieurs types d'informations, brouillant ainsi les lignes de la classification ordinaire par types de données.

Les séminaires Senseval et SemEval ont établi des standards d'évaluation incontournables, et les meilleurs systèmes actuels atteignent des niveaux de précision de l'ordre de 80% sur ces ensembles d'évaluation, ce qui s'approche de la compétence humaine. Les résultats de ces systèmes sont si bons que le besoin de forger de nouvelles normes d'évaluation se fait progressivement ressentir.

Maintenant que les systèmes de désambiguïsation sont devenus aussi performants, plusieurs autres types d'applications peuvent les utiliser afin de distinguer les sens des mots et, de cette manière, gagner en efficacité.

Dans les chapitres suivants, nous ferons le tour des principales ressources lexicales en langue anglaise (Chapitre 2), avant de réfléchir à ce en quoi la désambiguïsation lexicale pourrait enrichir les ressources lexicales, et de chercher à savoir s'il existe déjà des ressources lexicales désambiguïsées en anglais (Chapitre 3).

Chapitre 2. Des ressources lexicales anglophones

L'acquisition du vocabulaire est essentielle pour l'apprentissage d'une langue maternelle ou étrangère. Pour le prouver, il suffit de rappeler, par exemple, qu'entre 95% et 98% des mots d'un texte – selon que la lecture soit guidée ou libre – doivent être connus par son lecteur pour une compréhension correcte du contenu (Laufer & Ravenhorst-Kalovski, 2010). Par conséquent, les listes de vocabulaire – et toute forme de ressource lexicale, par extension – représentent un outil efficace pour l'apprentissage (Nation & Chung, 2009, p. 545). Bien que, pour certains auteurs, l'apprentissage explicite des mots sous la forme de listes ne soit qu'une première étape de l'acquisition lexicale, qui requiert d'autres formes d'exposition implicites (lecture, audition, oral) et explicites (collocations, grammaire, etc.), ces ressources n'en restent pas moins un instrument exploitable à tous niveaux de compétence (Nation & Waring, 1997).

En effet, disposer de listes de vocabulaire *triées* (selon quelques critères) permet aux enseignants de classer le vocabulaire et de sélectionner les mots à enseigner prioritairement. De la même manière, ces ressources font office de support pour l'élaboration de matériaux didactiques : cours, manuels, syllabus et textes destinés aux apprenants (Kwary & Jurianto, 2017; Leech et al., 2001). Divers exemples, dans plusieurs langues, attestent du bien-fondé des ressources lexicales comme support à l'acquisition lexicale et, *a fortiori*, comme source pour l'élaboration des programmes d'apprentissage.

En primaire, la langue maternelle peut être enseignée par le prisme de ces ressources : celles-ci seront principalement adressées aux instituteurs et concerneront en particulier la maîtrise écrite. Pour l'anglais, *The basic spelling vocabulary list* (Graham et al., 1993), par exemple, liste 850 des mots les plus fréquents dans les usages écrits des enfants, et donne aux instituteurs un cadre pour l'enseignement de l'orthographe par année. Un répertoire des 500 mots les plus utilisés dans les usages écrits d'enfants australiens existe aussi (*Oxford Wordlist* - Bayetto, 2017), et peut notamment s'employer pour soutenir le développement du vocabulaire et de la lecture durant les trois premières années du primaire. Pour le français, l'échelle Dubois-Buyse (Dubois & Buyse, 1952) ou les moins anciennes *Listes orthographiques de base du français* (Catach et al., 1984) servent le même objectif : répertorier les mots les plus fréquents pour soutenir le développement de la langue écrite chez l'enfant. Enfin, les listes d'idéogrammes (kanjis), prescrites par le Ministère de l'Éducation japonais, présentent les principaux signes à acquérir par année scolaire – en primaire (*gakushū kanji*, « kanjis d'étude ») et en secondaire (*jōyō kanji,* « kanjis à usage commun »).

Dans le cadre de l'apprentissage d'une langue étrangère ou seconde (L2), remarquons que selon Zapata Monge (2013), les méthodes basées sur des listes de mots correspondent à une première étape chronologique de l'enseignement, qui aurait prévalu jusqu'aux années 70 et au développement d'une approche communicative du lexique. Néanmoins, il semble aussi que la mémorisation de listes de vocabulaire s'impose souvent encore à l'apprenant comme un souvenir marquant. La liste *Oxford 3000*, par exemple, est explicitement destinée à soutenir l'apprentissage de l'anglais : celle-ci recense les 3000 mots « les plus importants » de la langue, selon des critères de fréquence, de portée, de familiarité

(pour le locuteur anglophone), et de pertinence. Le recours soutenu et durable à des listes de mots pour l'acquisition des L2 témoigne, une fois encore, de l'intérêt de ces ressources lexicales.

L'élaboration de telles listes n'est pas une tâche neuve : dans ce second chapitre, nous retracerons l'évolution des principales ressources lexicales disponibles pour l'anglais. Nous commencerons par donner un aperçu des ressources fondées sur base d'informations fréquentielles (A), puis passerons en revue les arguments pour remettre en cause suprématie de la fréquence (B.1). Nous observerons ensuite plusieurs autres types de normes lexicales existant (B.2), ce avant de nous concentrer sur les ressources disponibles pour l'apprentissage de l'anglais en L2 (C).

A. Focus sur la L1 : exploitation de mesures fréquentielles

Il est depuis longtemps largement admis que la fréquence des mots est une variable importante dans la mémorisation, la reconnaissance et le traitement des mots (Brysbaert, Buchmeier, et al., 2011; Brysbaert & New, 2009). L'effet facilitateur de la fréquence sur les performances en reconnaissance des mots a d'abord été mentionné par Cattell (1885), puis rapporté par Preston (1935), confirmé empiriquement par Howes & Solomon (1951) ainsi que, plus récemment, par Monsell (1991) et Brysbaert et al. (2000).

Cet effet fréquentiel s'explique généralement par le fait que les mots communs du lexique mental sont plus faciles d'accès que les mots moins communs (par le fait d'un seuil d'activation moins élevé, ou d'un niveau d'activation de base plus élevé), ou parce que la stratégie de recherche dans le lexique mental est fréquentielle (Brysbaert et al., 2000).

En ce qui concerne l'apprentissage, au-delà des implications directes des hypothèses susmentionnées, l'observation – longuement établie (Bongers, 1947; Nation, 1990) – qu'un étudiant qui connaîtrait une quantité relativement réduite des mots les plus fréquents de la langue (approximativement 3.000 pour l'anglais) pourrait déjà comprendre 80 à 90% d'un texte, aurait également assis la domination des fréquences dans le domaine des ressources lexicales (Honeyfield, 1977).

Quoi qu'il en soit, plusieurs auteurs ont créé des listes de mots sur la base de cette norme, en partant du postulat que les mots fréquents devaient être les plus « accessibles » pour les apprenants, et *vice versa*. Cependant, qualifier ces ressources de « purement fréquentielles » relèverait de l'erreur, attendu que chez la plupart des auteurs passés en revue se retrouve la conscience que la *portée* ou *dispersion* des mots (le nombre de sources dans lesquelles ils apparaissent) mérite également d'être mentionnée, ce pour éviter des biais imputables à la sur-représentation de mots dans certaines sources du corpus[29]. Parmi ces listes, nous distinguons entre les ressources pédagogiques et les ressources appuyées sur du contenu pour adulte.

[29] Par exemple, retrouver « *quotient* » 121 fois dans un corpus d'un million de mots n'en fait pas un mot commun, parce qu'il ne se retrouvera que dans deux sources spécialisées en mathématiques.

A.1. Ressources pédagogiques : aider l'enfant à l'école primaire

i. *The Teacher's Word Book (Thorndike, 1921, 1931)*

La première liste de vocabulaire de grande portée a été produite au début du XX^{e} siècle. Il s'agit d'une liste alphabétique des 10.000 lemmes apparaissant les plus fréquemment dans un corpus de 4,565 millions de mots issus de diverses sources écrites : Bible, classiques de la littérature anglaise, quotidiens, correspondances, livres « de ménagère » (80%), contre seulement 20% de manuels de primaire et de littérature pour enfants. Cette liste avait pour objectif de permettre aux instituteurs d'estimer l'importance et le caractère plus ou moins commun des mots, et ainsi de mieux sélectionner les mots à apprendre aux élèves. Pourtant, la sous-représentation du contenu pour enfants dans l'élaboration d'une ressource destinées aux instituteurs paraît immédiatement paradoxale et contre-productive.

Dans la première version, les fréquences brutes ne sont pas disponibles : les auteurs proposent un « numéro de crédit » qui indique l'importance du mot selon deux informations : sa fréquence et sa portée. Sur base de cette valeur, les mots sont classés par ordre d'importance.

Une première extension de ce livre sera publiée en 1931, comprenant 10.000 mots supplémentaires et intégrant cette fois plus de textes pour enfants et adolescents. Une dernière version augmentée paraît en 1944 (Thorndike & Lorge, 1944), qui inclut cette fois 30.000 lemmes, dont les fréquences ont été estimées sur un corpus de 18 millions de mots.

Bien que ces listes soient aujourd'hui caduques, François et al. (2014) indiquent que les travaux de Thorndike ont contribué à l'utilisation de données statistiques dans le domaine pédagogique. Thorndike (1921) fut, en effet, l'un des premiers à avancer que plus un mot est fréquent, plus celui-ci est adapté aux jeunes lecteurs (François et al., 2014).

ii. *A basic vocabulary of elementary school children (Rinsland, 1945)*

À la différence des listes de Thorndike, celle-ci s'appuie sur une collection de 100.000 textes (poèmes, lettres, histoires, etc.) effectivement produits *par* des élèves américains des niveaux scolaires 1 à 8 (c'est-à-dire de de 6 à 14 ans). Plus de 6 millions de mots ont été examinés, résultant en une liste des fréquences brutes pour 25.632 formes.

Rinsland (1945) est le premier à avoir fourni des fréquences réparties par année scolaire : pour chaque mot, le nombre d'occurrence de ce mot dans chacun des niveaux est donné (ainsi, huit valeurs sont données pour un mot).

Cette approche a d'ailleurs permis de mesurer empiriquement la diversification du vocabulaire avec l'âge. En effet, parmi les 5.099 mots ayant une valeur non-nulle pour le premier niveau, les 2000 plus fréquents représentent 98% des mots produits par ces élèves, tandis que cette valeur n'est plus que de 90% pour le huitième niveau.

Toutefois, une autre voie intéressante ouverte par ce travail – se baser sur les textes écrits par des enfants pour compiler des mesures de fréquence - ne semble pas avoir été poursuivie dans les ressources fréquentielles en L1.

iii. *The American Heritage Word Frequency Book (Carroll et al., 1971)*

La base de données réalisée par Carroll et ses collaborateurs s'attache à des matériaux textuels destinés aux écoliers : la liste comprenant 86.741 mots est basée sur un corpus de 5,09 millions de mots extraits de textes écrits en usage dans les écoles américaines. Il est intéressant de noter que l'information fréquentielle est donnée pour chacune des 17 matières ainsi que pour les 8 niveaux d'enseignement (niveaux 3 à 9, soit de 7 à 15 ans). Autrement dit, pour un mot, 25 valeurs fréquentielles sont données. La focalisation de cette liste sur des textes scolaires ainsi que la répartition des fréquences sur plusieurs niveaux d'enseignement en représentent certainement l'intérêt principal (Nation & Waring, 1997).

Les perspectives statistiques présentées par les auteurs représentent un attrait supplémentaire de cet ouvrage : en plus des fréquences brutes, trois indices statistiques « originaux » (Carroll, 1972) et supérieurs à ceux des comptes de fréquence précédents (Darnell & Howes, 1972) sont donnés pour chaque mot. D'abord, l'entropie relative des mots définie sur les 17 matières (*D)* mesure la *dispersion* des fréquences. Les mots dont le *D* tend vers 0 ne se retrouvent que dans quelques catégories ; les mots dont le *D* s'approche de 1 sont plus équitablement répartis. En d'autres termes, la mesure *D* remplace la *portée* de Thorndike (1921), moins sophistiquée. Cette valeur permet notamment des sélectionner les mots qui apparaissent dans des contextes restreints (Manelis, 1972). Ensuite, la valeur *U* est une estimation de la fréquence "réelle" (fréquence par millions de mots - *fpmw*) du mot dans un corpus théoriquement infini. Cette estimation est réalisée en fonction de la dispersion : pour une même fréquence, les mots avec un *D* plus élevé auront également un *U* plus élevé, et inversement. La dernière mesure calculée est l'index de fréquence standard (*SFI*), résultant d'une transformation logarithmique de *U*. Cette mesure indique la probabilité d'occurrence d'un mot (*type*) et peut se révéler efficace pour établir des catégories de fréquences (Manelis, 1972).

Les contributions de Carroll et al. à la théorie mathématique de la distribution des fréquences lexicales est conséquente (Darnell & Howes, 1972). Les auteurs ont, en effet, montré théoriquement que la distribution des mesures *SFI* était approximativement normale, ce qui implique que le modèle log-normal prédirait assez bien la distribution empirique des mots d'un corpus, mieux encore que le modèle Zipf (Zipf, 1949 – cf. A.2.vi).

Cependant, en plus de s'être envieillie, cette base de données présente plusieurs faiblesses qui ont été rapidement identifiées (Darnell & Howes, 1972). Premièrement, le corpus n'est équilibré ni en termes de niveaux, ni en termes de matières : on trouve, par exemple, deux fois plus de mots pour le niveau scolaire sept que pour le niveau neuf, et 100 fois plus de mots dans le domaine scientifique (plus d'un demi-million de mots) que dans le domaine de la religion (moins de 5000 mots). Deuxièmement, la base de données inclut des redondances : *the*, *The* et *THE* représentent trois entrées différentes. Troisièmement, de la construction d'une base de données aussi large découle une sur-représentation de *types* peu fréquents, étant donné que les 1000 mots les plus fréquents représentent 74% du corpus, et que ce pourcentage s'élève à 90% si les 5000 mots les plus fréquents sont considérés. Enfin, le calcul des mesures U et SFI a été remis en question (voir Darnell & Howes (1972) pour une discussion statistique plus avancée).

iv. *The Educator's Word Frequency Guide – WFG (Zeno et al., 1995)*

Selon Keuleers et al. (2010), la base de données WFG est la plus utilisée en ce qui concerne les fréquences pour enfants. Elle est basée sur un corpus de 17 millions de mots issus de textes écrits pour les élèves américains de la primaire à l'école secondaire (niveaux 1 à 12)[30]. La liste est constituée des fréquences – réparties sur ces douze niveaux d'apprentissage[31] - pour environ 155.000 mots, qui correspondent, selon les auteurs, aux mots que les élèves sont susceptibles de rencontrer de la maternelle au début des études supérieures. Les statistiques répertoriées dans le WFG sont identiques à celles proposées par Carroll et al. (1971) : *D* (dispersion), *U* (*fpmw*) et *SFI* (index de fréquence standard).

Il est intéressant de remarquer que seuls 19.468 mots dans la base de données ont une fréquence d'au moins une occurrence par million de mots, ce qui signifie que les mots peu fréquents sont sur-représentés dans la liste (87% des mots apparaissent moins d'une fois tous les millions de mots).

La liste WFG jouit d'une large influence dans le domaine, ce quels que soient les groupes et les tâches concernées (Balota et al., 2004). Par ailleurs, Zevin & Seidenberg (2002) affirment que le WFG est une base de données efficace pour toute expérience concernant l'âge d'acquisition des mots, attendu que celle-ci fournit des mesures de fréquences pour chaque niveau de l'apprentissage de l'anglais comme langue maternelle. L'hypothèse défendue par Zevin & Seidenberg (2002) est que plus un mot est appris tôt, plus la fréquence cumulée de celui-ci sera élevée, c'est-à-dire qu'il est rencontré plus fréquemment durant le parcours scolaire.

v. *Metametrics*

Les normes de fréquence MetaMetrics (MetaMetrics, Inc., 2003) le prouvent, les listes de mots ne sont pas l'apanage unique des académiques. Ces fréquences proviennent d'un corpus de 350 millions de mots répartis sur 21.000 textes de fiction, de non-fiction ou encore tirés de manuels scolaires à destination d'élèves du primaire et du secondaire (Balota et al., 2004).

Bien que les normes MetaMetrics soient privées, nous les incluons dans cette section pour deux raisons : premièrement, Metametrics Inc. est une société spécialisée dans la recherche en matière d'éducation. La compagnie a notamment développé le cadre de référence *Lexile* pour la lecture, dont les mesures se basent en partie sur la fréquence lexicale ; deuxièmement, deux travaux notables en psycholinguistique (Balota et al., 2004; Brysbaert & New, 2009) ont démontré une forte corrélation entre des tâches de décision/dénomination lexicales et les fréquences MetaMetrics, attestant par conséquent du caractère qualitatif de ces dernières.

[30] Manuels scolaires, fiction, non-fiction et littérature scolaire.

[31] Le nombre d'occurrences d'un mot est donné pour chacun des niveaux scolaires. En d'autres termes, à tout mot sont associées douze mesures fréquentielles.

vi. *Children's Printed Words Database – CPWD (Masterson et al., 2010)*

Dans le sillage des listes de fréquences lexicales focalisées sur des textes pour enfants, la CPWD représente, selon van Heuven et al. (2014), l'une des meilleures bases de données fréquentielles pour enfants en anglais *britannique*. En effet, à la différence des listes de 1971, de 1995, et de 2003, celle-ci est centrée sur des livres lus dans les écoles de Grande-Bretagne. Contrairement à ses prédécesseuses, en outre, le dessein de cette base de données ne semble pas être la mesure des fréquences lexicales par niveau scolaire[32].

Les textes choisis sont destinés aux élèves de cinq à neuf ans, c'est-à-dire dans les quatre premières années de l'école primaire[33], que les auteurs de la CPWD estiment les « plus critiques » pour l'acquisition de la langue écrite chez les enfants. Le corpus ainsi créé se veut être un échantillon représentatif des livres lus par les enfants dans ces années. À cet effet, une enquête en milieu scolaire a été réalisée par les auteurs. De cette façon, 1011 livres (995.927 mots) ont été rassemblés, et la liste CPWD inclut les fréquences (*fpmw*) de 12.193 mots. Au surplus, la ressource construite par Masterson et al. (2010) inclut d'autres informations, concernant notamment le voisinage orthographique et phonologique.

Cette liste remplace une autre base de données (Stuart et al., 2003), qui – au contraire – n'est pas représentative, puisqu'elle s'appuie uniquement sur des livres lus par des écoliers de première primaire dans une école du nord de Londres (Masterson et al., 2010)[34]. À l'instar de la base de données de Stuart et collaborateurs, le vocabulaire dans la CPWD est sévèrement biaisé du côté des fréquences faibles. En effet, les 100 mots les plus fréquents dans la liste – c'est-à-dire moins d'1% de cette dernière – représentent près de 52% de l'ensemble du corpus.

vii. *Des listes pour enfants basées sur du contenu écrit par des adultes*

Avant de continuer, remarquons qu'une analyse de la fréquence des mots dans les textes *pour* enfants présuppose nécessairement une dimension prescriptive de la démarche, dont l'objectif est alors de découvrir quels mots doivent être connus et appris, plutôt que quels mots sont réellement connus :

> The wisdom of choosing a corpus of this type may be questioned. What does a count of fifth grade readers represent but the language that current adult writers think is appropriate to children of that level ? (Darnell & Howes, 1972)

Cette citation fait écho à ce que l'un des auteurs écrit dans le manuel d'utilisation : « *Teacher and other professional users of the database will be able to discover which words children need to know (and be taught)*[35] *in order to read at a given level* » (Lovejoy, 2003). Un paradoxe fondamental est ici soulevé : lorsque l'on s'intéresse au langage des adultes, on observe les fréquences dans des textes écrits par ces mêmes adultes. Or, lorsque l'on s'intéresse au

[32] La ressource répertorie toutefois la fréquence de chaque mot dans chaque livre, les fréquences nivelées restent donc accessibles indirectement (cf. infra pour la composition du corpus).
[33] Année R (*Reception Year*), Année 1, Année 2, Année 3.
[34] Pour cette raison, nous ne nous arrêtons pas plus longuement sur la base de données de Stuart et al. (2003).
[35] Nous soulignons.

langage des enfants, on n'observe pas les fréquences dans des textes écrits par ces derniers, mais bien dans des textes qu'un certain nombres d'adultes (experts, éditeurs, etc.) jugent adéquats pour l'enfant. En somme, chacune des ressources pour enfants présentées jusqu'ici, à l'exception de celle de Rinsland (1945), montre ce que l'enfant devrait connaître, et non ce que l'enfant connaît et utilise effectivement à tel ou tel âge : ces dernières ne reflètent pas l'utilisation réelle du langage en bas âge et pourraient être biaisées par la perception adulte de ce qui est adéquat.

A.2. Au-delà de l'aspect pédagogique : normes fréquentielles dans du contenu destiné à l'adulte

L'utilité des fréquences lexicales n'est pas uniquement pédagogique. Celles-ci peuvent aussi être utilisées en psycholinguistique, afin de comprendre le traitement humain du langage oral (discours, écoute) ou de l'écrit (lecture, rédaction) ; dans diverses applications en traitement automatique du langage (évaluation automatique de la lisibilité d'un texte, simplification de textes, extraction d'informations, etc.) ou dans d'autres domaines de recherche, afin d'avoir une idée des thèmes et sujets essentiel d'une branche. Une pléthore de listes de vocabulaire technique basées sur les fréquences existent aussi (*e.g.* Coxhead & Hirsch (2007) pour le vocabulaire scientifique et James et al. (1994) pour l'informatique).

i. *Kučera & Francis (1967)*

L'essor de la linguistique computationnelle a permis de rassembler de plus grands corpus et d'en extraire l'information fréquentielle moins laborieusement. Sur la base de leur première version du BROWN CORPUS (1,014 millions de mots en anglais américain), Kučera & Francis (1967) établissent une nouvelle liste de fréquences lexicales pour l'anglais (50.406 mots[36]). Il s'agit de la première liste fréquentielle obtenue de manière automatique (Leech et al., 2001, p. x).

Le BROWN CORPUS (Kučera & Francis, 1964) se veut représentatif et équilibré : celui-ci se compose de 500 textes, datant tous de 1961, répartis sur 15 catégories textuelles (reportage de presse, fiction romanesque, textes religieux, humour, etc.). Cette caractéristique fait remarquer aux auteurs que les distributions de fréquence sont tributaires du type et du thème des documents : la fréquence d'un mot peut se voir surévaluée parce que ce dernier apparaîtrait de nombreuses fois dans des textes spécifiques, sans pour autant être véritablement « commun ». Deux mesures de portée sont prises en compte : le nombre de genres (max. 15) et le nombre d'échantillons (max. 500) dans lequel le mot apparaît (*ex. French* : 139 occurrences – dans 14 genres – et 71 textes).

Les normes définies par ces auteurs resteront l'une des mesures de fréquence les plus populaires auprès des (psycho)linguistes jusque récemment : en attestent les 215 articles citant encore la base de données (de 1967) en 2008 (Brysbaert & New, 2009), par exemple. La popularité de Kučera & Francis par rapport à Thorndike & Lorge (1944) s'expliquerait parce que les textes choisis étaient plus récents (1961 vs. années 20-30) et uniquement destinés à un public adulte. En outre, les mesures auraient été plus facilement disponibles et mises en avant par certains articles clés (Brysbaert & New, 2009).

[36] Selon Manelis (1972).

Enfin, notons que le BROWN CORPUS fut l'un des premiers à être *POS-taggué,* ce grâce au *TAGGIT tagger* (Greene & Rubin, 1971). L'algorithme avait taggué correctement 77% des unités lexicales[37] et permis d'obtenir des normes fréquentielles basées non plus sur une simple forme orthographique, mais sur une forme associée à son étiquette grammaticale (Francis & Kučera, 1982). L'étiquetage morpho-syntaxique constitue une première étape importante de la description sémantique (ex. : connaître l'étiquette de *bear* - verbe *vs*. nom - suffit à en connaître le sens – « porter, endurer » *vs*. « gros animal dangereux »).

ii. *LOB CORPUS – Comparaison fréquentielle des variétés américaines et britanniques (Hofland & Johansson, 1982)*

Les textes du BROWN CORPUS sont représentatifs de la variété d'anglais en usage en Amérique. Quelques années plus tard, un corpus équivalent au BROWN CORPUS a été assemblé, associé cette fois à l'anglais britannique : le LOB (Lancaster-Oslo/Bergen) CORPUS, et une liste de fréquences similaire à celle de Kučera et Francis a été compilée (Hofland & Johansson, 1982). Le principe fondateur du LOB CORPUS est clair : ce dernier se veut similaire trait pour trait au BROWN CORPUS (représentativité d'une variété de langue, méthode d'échantillonnage identique, 500 textes d'une longueur approximative de 200 mots pour un total d'un million de mots, sources écrites, même année de publication). En revanche, la version imprimée des fréquences de Hofland et Johansson est moins exhaustive que l'équivalent américain : seules 7476 formes y sont répertoriées, correspondant aux formes apparaissant dix fois ou plus dans le corpus.

Dans cet ouvrage se retrouvent également des comparaisons de fréquences pour les formes correspondant en anglais américain et britannique. Ainsi, Hofland & Johansson (1982) introduisent l'idée que les listes de fréquence peuvent être utiles afin de comparer des variétés de langues (Leech et al., 2001). À titre informatif, notons, par exemple, que, parmi les 50 mots les plus fréquents dans le LOB et le BROWN CORPUS, un seul diffère (Hofland & Johansson, 1982, p. 18).

Enfin, Johanson et Hofland, à l'instar des créateurs du BROWN CORPUS, figurent parmi les premiers à avoir publié une ressource fréquentielle dont les entrées sont étiquetées grammaticalement. Le taggeur utilisé atteint, sur le LOB CORPUS, les 96% de précision (Johansson & Hofland, 1989).

iii. *CELEX (Baayen et al., 1993)*

D'autres projets ont intégré l'observation des fréquences en corpus à une description multiaxiale du lexique, c'est le cas notamment de CELEX, qui fournit, pour 52.446 lemmes issus de deux dictionnaires (*Oxford Advanced Learner's Dictionary*[38] et *Longman Dictionary of Contemporary English*[39]), d'une part, les fréquences calculées sur le corpus COBUILD (17.9 millions de mots[40]), et, d'autre part, des informations orthographiques, phonologiques, morphologiques et syntaxiques. Les fréquences sont données (a) brutes, (b) par millions de mots, (c) sur une échelle logarithmique.

[37] Le reste ayant été taggué à la main.
[38] Hornby et al. (1974).
[39] Procter (1978).
[40] L'ensemble des lemmes recouvre 92% du corpus.

L'originalité de la base de données CELEX, de surcroît, réside dans le fait que ses auteurs furent les premiers à véritablement préconiser l'utilisation de la fréquence des lemmes : jusque-là, les fréquences calculées concernaient plutôt les formes telles que rencontrées en corpus (Brysbaert & New, 2009), et éventuellement *pos-taggées* (cf. BROWN & LOB CORPUS). Ainsi, deux bases de données composent CELEX, l'une concernant les formes (160.594), l'autre afférente aux lemmes (52.446). Pour calculer les fréquences lemmatisées, le corpus a été *parsé* et *pos-taggué* semi-automatiquement avec une vérification manuelle sur des échantillons choisis (Brysbaert & New, 2009).

Il est intéressant de noter que les recherches effectuées plus récemment par Brysbaert & New (2009) n'ont pas permis de prouver que la fréquence des lemmes soit plus efficace que la fréquence des formes pour expliquer le degré de connaissance des mots en anglais. Ce débat est intéressant, car – à partir de CELEX – plusieurs auteurs ont fait le choix de mesurer les fréquences des formes lemmatisées (Brysbaert et al., 2012; Burgess & Livesay, 1998; M. Davies & Gardner, 2010; Dürlich & François, 2018; Leech et al., 2001). Indirectement, la position de Brysbaert & New (2009) questionne l'utilité même de la désambiguïsation[41] dans le contexte des ressources lexicales : s'il n'est (même) pas utile de lemmatiser, pourquoi faudrait-il aller jusqu'à désambiguïser ? Nous reviendrons à cette question dans le chapitre 3.

iv. *HAL (Burgess & Livesay, 1998)*

L'essor de l'Internet a permis de rechercher sans peine des corpus plus grands, plus évolutifs et illustrant une langue écrite plus spontanée et naturelle. À cet effet, Burgess & Livesay (1998) ont tiré parti d'un réseau de forums (*Usenet*) pour extraire – durant le mois de février 1995 – des discussions libres sur une large gamme de sujets. Le corpus obtenu de cette manière – le HAL CORPUS (de *Hyperspace Analogue to Language,* nom d'un modèle de mémoire) – comprend 131 millions de mots, et a permis le calcul des fréquences brutes pour 97.261 mots.

Bien que le corpus ne fût pas étiqueté morpho-syntaxiquement et que les fréquences calculées n'aient été rendues disponibles que près de dix ans après la publication de Burgess et Livesay *(*dans le cadre du *English Lexicon Project* - Balota et al., 2007), HAL fait partie des trois corpus américains principaux en ce qui concerne les fréquences lexicales (Brysbaert & New, 2009). En effet, plusieurs expériences avec cette ressource ont généré des résultats positifs. Certains temps de décision lexicale – obtenus par plusieurs expériences – enregistrent une corrélation importante avec les normes HAL (Balota et al., 2007). En particulier, cette corrélation est plus grande avec HAL qu'avec d'autres normes populaires - notamment celles de Kučera & Francis, en particulier pour les mots de basses et moyennes fréquences (Burgess & Livesay, 1998). Cette dernière observation pourrait être associée à la différence de volume des 2 corpus utilisés (131M vs. 1M de mots). Quoi qu'il en soit, vers la fin du XX^{e} siècle, les normes de Kučera & Francis commencent à être sérieusement remises en question, mouvement qui s'amplifiera par la suite avec Zevin & Seidenberg (2002), Balota et al. (2004) et surtout Brysbaert & New (2009).

[41] Dans ce cas, la désambiguïsation n'est assurée que par le POS-tagging, mais la question pourrait également s''étendre à la désambiguïsation sémantique.

v. *BRITISH NATIONAL CORPUS Frequencies – BNC (Leech et al., 2001)*

La liste des réalisée sur base du BNC s'inscrit dans le sillage de CELEX : ses auteurs rendent compte des fréquences pour chacune des formes retrouvées dans le corpus ainsi que pour les formes non-fléchies (lemmes étiquetés morpho-syntaxiquement)[42]. En plus des listes alphabétiques et triées par fréquence - dans lesquelles se retrouvent un score de fréquence normalisé (*fpmw*), un indice de portée et une mesure de dispersion (*Juilliand's D*) pour 678.211 formes –, les auteurs présentent un indice de spécificité (*distinctiveness*) qui leur permet de concevoir des listes supplémentaires vérifiant, par exemple, si les fréquences obtenues pour un mot sont significativement dissemblables à l'oral et à l'écrit, dans l'écrit imaginaire et dans l'écrit informatif, ou encore dans l'oral conversationnel et dans l'oral « pratique » du quotidien[43]. En effet, Leech et al. (2001) sont conscients de la nécessité de dresser un profil fréquentiel aussi pour les différentes variétés de la langue ; les auteurs remarquent notamment qu'entre les 50 mots les plus retrouvés à l'écrit et à l'écrit, seuls 66% sont identiques. Le corpus utilisé représente certainement la clé de voûte de cette entreprise.

Le BRITISH NATIONAL CORPUS est supposé représentatif de l'anglais britannique *oral et écrit* : les nombreuses sources sont échantillonnées selon l'usage réel des locuteurs, de telle sorte que le BNC permet de faire des inférences fiables sur l'anglais britannique en général (Leech et al., 2001). Les quelques 100 millions de mots de ce corpus sont majoritairement issus de sources datant d'entre 1985-1994, et aucune n'est antérieure à 1960, ce qui en faisait, à l'époque, un corpus particulièrement récent.

Le corpus et les fréquences extraites présentent toutefois leurs limitations. D'abord, 90% du BNC provient de l'écrit, et 10% des sources seulement sont orales, or - dans l'usage des locuteurs – l'oral domine indubitablement l'écrit, ce qui remet en question le caractère représentatif du corpus. Ensuite, l'étiquetage morphosyntaxique a été réalisé avec le *CLAWS Tagger*, qui offre une précision suboptimale de ~96,5%. Enfin, les auteurs signalent une faible marge d'erreur dans le calcul des *fpmw*, susceptible de biaiser légèrement les résultats pour les mots les plus fréquents.

vi. *SUBTLEX*

Les bases de données SUBTLEX sont nées par l'observation selon laquelle les relevés fréquentiels précédents étaient souvent effectués principalement sur des sources écrites (livres, journaux, magazines, etc.) qui présentent le désavantage notable d'être édités, c'est-à-dire que la langue qui s'y retrouve se voit – en amont - modifiée, embellie, notamment dans l'optique d'éviter la répétition. De plus, les sujets traités par ces sources s'éloignent souvent des préoccupations quotidiennes des locuteurs.

Pour pallier ce problème, de nombreux travaux – au départ de New et al. (2007), pour le français – ont fait le choix de calculer les fréquences des mots sur base de sous-titres télévisés. Ceux-ci, en plus d'être facilement accessibles, sont l'expression d'une langue

[42] La mesure de fréquence des lemmes valant la somme des fréquences des différentes formes dans le corpus.

[43] *Task-oriented speech* : cours, sermons, appels téléphoniques, commentaires sportifs, émissions télé, interactions en classe, actes de parlement, consultations, etc.

interactive, naturelle, et avec laquelle les locuteurs sont peut-être plus souvent confrontés que la langue écrite.

L'une des forces de ces ressources lexicales SUBTLEX est certainement leur proximité avec la psycholinguistique : la qualité des normes fréquentielles est systématiquement testée en vertu de leur corrélation avec le temps et la précision de tâches de décisions lexicales[44]. Les recherches en anglais américain (Brysbaert & New, 2009) et dans d'autres langues, ont montré que les méthodes basées sur les sous-titres prédisaient mieux les temps de traitement lexicaux que les fréquences relevées sur des livres ou d'autres sources écrites (Boada et al., 2020; Brysbaert, Buchmeier, et al., 2011; Brysbaert, Keuleers, et al., 2011; Cai & Brysbaert, 2010; Cuetos et al., 2011; Dimitropoulou et al., 2010; Ferrand et al., 2010; Keuleers et al., 2010; Mandera et al., 2015; New et al., 2007; van Heuven et al., 2014).

La première ressource créée pour l'anglais (Brysbaert & New, 2009), $SUBTLEX_{US}$, concerne l'anglais américain. Pour commencer, les auteurs déconstruisent la norme fréquentielle préférentielle jusque-là (Kučera & Francis, 1967)[45]. Ensuite, la méthodologie adoptée pour le calcul des fréquences lexicales est détaillée : les fréquences $SUBTLEX_{US}$ sont obtenues[46] sur un corpus de 51 millions de mots (16.1 millions venant de séries télévisées, 14.3 millions de mots venant de films pré-1990 et 20.6 millions de mots venant de films post-1990). Au total, la ressource comprend 74.286 entrées. Bien que la qualité des normes $SUBTLEX_{US}$ soit vérifiée empiriquement comme décrit précédemment, celles-ci ne se départissent pas de certaines faiblesses : par exemple, les fréquences HAL (cf. infra) sont plus efficaces sur des mots longs, ce qui s'explique parce que les mots des sous-titres tendent à être plus courts.

Attendu que l'anglais américain diffère par certains aspects de l'anglais britannique (orthographe de certains mots [*labor vs*. *labour*] et variantes sémantiques [*biscuits*, *pants*]), il s'est avéré pertinent, par la suite, d'établir une ressource similaire pour l'anglais britannique : $SUBTLEX_{UK}$, afin que les chercheurs puissent effectuer des expériences sur le traitement lexical en Grande-Bretagne sans générer certaines imprécisions (van Heuven et al., 2014). Cette version a été construite au départ d'un corpus de 201 millions de mots et contient 106.022 entrées. Cette entreprise s'est trouvée efficace, puisque $SUBTLEX_{UK}$ prédit mieux les temps de décision lexicale du *British Lexicon Project* que $SUBTLEX_{US}$ (van Heuven et al., 2014).

L'originalité de $SUBTLEX_{UK}$ se définit par deux critères. D'une part, certaines mesures de fréquence sont calculées au départ de contenu pour enfants : près de 37% des entrées sont issues de chaînes pour enfants de primaire (CBBC, une chaîne à destination des 6-12, dont le volume correspond à 6,76% du corpus total), et plus de 17% des entrées viennent de chaînes pour les 0-6 (CBeebies, 2,9% du corpus total). Cette partie de $SUBTLEX_{UK}$ s'inscrit

[44] Ces variables sont rendues disponibles par des méga-études telles que ELP (*English Lexicon Project* : Balota et al., 2007) ou encore BLP (*British Lexicon Project* : Keuleers et al., 2012).

[45] L'un des arguments de Brysbaert & New (2009) étant que les mesures de fréquences de Kučera & Francis (1967) sont prises sur un trop petit corpus (~ 1M de mots). Les auteurs présentent une étude statistique réalisée sur le BNC qui montre que la taille d'un corpus devrait être comprise entre 16 et 30 millions de mots pour avoir des normes lexicales de confiance exploitables en psycholinguistique.

[46] Fréquences par millions de mots (*fpmw*) et $log_{10}(fpmw)$.

directement dans le sillage de la CPWD (Masterson et al., 2010 - cf. - 40 -)[47]. D'autre part, les auteurs proposent d'utiliser l'échelle Zipf (Zipf, 1949) pour calculer les fréquences, ce contrairement à Carroll et al. (1971). L'échelle Zipf est une mesure de fréquence lexicale standardisée dont l'interprétation est indépendante de la taille du corpus. Au lieu de faire état de la fréquence par millions de mots – ou d'une transformation logarithmique de celle-ci ($log_{10}(fpmw)$) -, les auteurs utilisent une transformation logarithmique de la fréquence par milliards de mots ($log_{10}(fpbw)$)[48]. Ce choix est essentiellement justifié par le fait que $log_{10}(fpmw) < 0 \ \forall fpmw < 1$; l'objectif étant de déconstruire l'idée préconçue selon laquelle les mots ayant une *fpmw* inférieure à 1 seraient méconnus des locuteurs. De plus, l'échelle Zipf se visualise facilement : la valeur sera nécessairement comprise entre 1 et 7, les valeurs des mots peu fréquents seront inférieures à 2.5-3, et les valeurs des mots de fréquence élevée seront supérieures à 4. Enfin, le rapport entre $log_{10}(fpmw)$ et $log_{10}(fpbw)$ est aisément calculable : $log_{10}(fpbw) = log_{10}(fpmw)+3$. Nous savons, par exemple, que 80% des mots de $SUBTLEX_{UK}$ ont une fréquence inférieure à 1 *pmw*, soit inférieure à 3 Zipf.

Les deux ressources incluent également de l'information quant à la diversité contextuelle des entrées (« dans combien de sources le mot apparaît-il ? »). En outre, les deux corpus ont été *parsés* et *pos-taggués*[49], cependant, les entrées dans SUBTLEX ne sont ni lemmatisées (*bear* ≠ *bears*) ni désambiguïsées selon leur étiquette grammaticale (*beat* [verbe] = *beat* [nom]). Pour chaque forme, c'est-à-dire chaque entrée, toutes les étiquettes (POS) possibles sont données, ainsi que le POS le plus fréquent et la fréquence de celui-ci – en guise de comparaison avec la fréquence totale.

B. D'autres normes lexicales

La multiplication des ressources lexicales faisant usage des normes fréquentielles invite à se questionner sur celles-ci. Bien entendu, l'effet fréquentiel sur le traitement et la connaissance des mots est reconnu depuis longtemps (cf. A). Pour autant, les fréquences lexicales sont-elles, seules, suffisantes pour élaborer des ressources lexicales destinées aux apprenants ? N'existe-t-il pas d'autres normes pouvant être prises en compte ? Dans ce sous-chapitre, nous tenterons de répondre à ces deux questions. Pour commencer, nous évoquerons les principales limitations des normes fréquentielles (B.1), avant de présenter d'autres types de normes lexicales (B.2) praticables pour l'élaboration de ressources lexicales.

B.1. Remise en question des normes fréquentielles

Les travaux de Brysbaert et collègues ont montré que les fréquences lexicales ne représentent qu'une partie de l'explication des divergences dans la connaissance des mots (Brysbaert, Mandera, McCormick, et al., 2018). Ceci se vérifie d'autant plus pour les mots de

[47] Le reste du corpus est extrait de 7 chaînes de télévision anglaises : *BBC1-4, BBC News, BBC Parliement* et *BBC HD* diffusées sur une période de trois ans (janvier 2010 à décembre 2012). Parmi ces chaînes, la *BBC1* est la plus représentée dans le corpus : les autres sont moins populaires et ont des heures plus limitées.

[48] La formule exacte étant $log_{10}(fpbw+1)$: un lissage de Laplace est effectué pour que la mesure ne soit pas égale à - ∞ lorsque des mots absents du corpus sont ajoutés à la ressource ($log_{10}(0) = -\infty$).

[49] À l'aide du *CLAWS Tagger* pour $SUBTLEX_{US}$ (Brysbaert et al., 2012) et du *Stanford Tagger* pour $SUBTLEX_{UK}$ (van Heuven et al., 2014).

basse fréquence. Certains de ces derniers sont globalement bien connus (*toolbar, screenshot, soulmate, uppercase, hoodie*) tandis que d'autres sont quasiment inconnus *(scourage, thunk, whicker* ou *caudle).* Dans le domaine de la psycholinguistique, cette dernière affirmation peut être posée autrement : ni les rapports entre temps de réaction et fréquences, ni la précision des prévisions fréquentielles dans des tâches de décision lexicale ne sont systématiques, en particulier pour les mots de basse fréquence : par conséquent, l'effet fréquentiel ne justifie que 30% à 40% de la variance dans les temps et la précision des réponses dans ces tâches (Brysbaert, Mandera, & Keuleers, 2018).

L'une des principales limites des mesures de fréquence point déjà : celles-ci sont objectives, se basent uniquement sur le contenu d'un corpus, et ne prennent pas en compte l'expérience des locuteurs. La fréquence réelle ne correspond pas exactement aux « mots disponibles », ou aux mots qui sont les mieux connus par la population. Ainsi, certains auteurs ont introduit la notion de prévalence : cette variable correspond au pourcentage de la population qui connaît un mot donné, et permet de pallier ce problème des mesures fréquentielles (Brysbaert et al., 2016; Keuleers et al., 2015).

Ensuite, il est important de rappeler que toutes les mesures de fréquences ne se valent pas. Pour commencer, elles ne se basent pas sur les mêmes matériaux. Or, nous savons que la nature du corpus est susceptible d'influencer les mesures de fréquence (registre, domaine, modalité, etc.). De plus, ces mesures de fréquence ne prennent sens qu'à condition d'être le reflet d'une langue auxquels les locuteurs sont effectivement exposés. Ceci implique que les divergences des profils socio-démographiques du public-cible devraient être prises en compte dans l'élaboration de ressources lexicales : l'entreprise même de cataloguer les « mots les plus fréquents *de la langue* » rencontre peut-être ici l'une de ses principales limites. À cet égard, Brysbaert, Mandera, & Keuleers (2018) remarquent que, pour les étudiants, les mesures de fréquences les plus fonctionnelles s'appuient sur les sous-titres télévisés, les réseaux sociaux et les blogs, tandis que pour les participants plus âgés, les fréquences traditionnelles (basées sur des livres, des journaux, des magazines, etc.) fonctionnent parfois mieux. Ici encore, les fréquences se trouvent limitées par leur caractère objectif. En effet, ces dernières ne peuvent prendre en compte la perception individuelle des fréquences lexicales. Remarquons, en outre, que les fréquences subjectives, bien que perçues par l'intuition de chaque locuteur, ne peuvent être fondamentalement homogènes d'un locuteur à l'autre. En effet, selon Richards (1976, p. 83) : « connaître un mot [dans une langue] signifie connaître le degré de probabilité de rencontrer ce mot à l'oral ou imprimé[50] ». Les mesures de familiarité (cf. B.2.i) recouvrent en partie cette notion de fréquence subjective.

À l'identique, la prise en compte du profil linguistique des locuteurs fait défaut aux normes fréquentielles ; pourtant, l'effet fréquentiel se manifeste différemment chez les locuteurs natifs et les apprenants L2 (Brysbaert, Mandera, & Keuleers, 2018). Nous reviendrons plus tard sur les désavantages des mesures de fréquences pour l'apprentissage en L2 (cf. C).

En outre, les ressources basées sur la fréquence ont été développées selon des méthodologies diverses. En particulier, les auteurs ne s'entendent pas sur la manière de mesurer les fréquences : certains mesurent celles-ci par millions de mots (*fpmw*), d'autres

[50] Notre traduction.

préconisent d'utiliser des mesures logarithmiques (ex. : Zipf). À cet égard, aucun consensus n'a été trouvé, pourtant, il est admis que les statistiques utilisées sont elles-mêmes susceptibles de faire pencher l'interprétation des mesures (Brysbaert, Mandera, & Keuleers, 2018).

Par ailleurs, l'interprétation standard de l'effet fréquentiel comme effet d'apprentissage peut être remise en question, or ce principe d'apprentissage par l'exposition suffit généralement à justifier l'entreprise d'élaboration de ressources lexicales fréquentielles. Pour commencer, la fréquence est hautement corrélée à d'autres variables (âge d'acquisition, familiarité, longueur du mot, similarité avec d'autres mots, etc.). Par conséquent, l'effet « fréquentiel » pourrait être engendré par l'une de ces autres variables. D'autres chercheurs avancent que la diversité contextuelle serait un meilleur prédicteur du traitement lexical, plaçant ainsi théoriquement la fréquence sur un second plan (Adelman et al., 2006). Dans les faits, les mesures de fréquences sont souvent standardisées, et prennent en compte cette diversité contextuelle (cf. A.1.iii). L'effet d'apprentissage, de surcroît, est basé sur l'idée que chaque occurrence, chaque exposition à un mot a le même poids, or cette conception est faussée : certains mots peuvent être retenus à vie après leur première rencontre (*ex.* : un film à propos d'une *licorne* ou d'un *gnome*), tandis que d'autres mots s'oublient assez facilement (*ex.* : *kestrel, hangar, cinch*), de telle sorte que le nombre d'occurrences d'un mot n'en est peut-être pas la meilleure mesure de connaissance (Brysbaert, Mandera, & Keuleers, 2018).

La meilleure conclusion à cet examen des limitations des normes fréquentielles se retrouve peut-être chez Jones & Durrant (2010), qui affirment que l'établissement de ressources lexicales destinées à l'apprentissage doit nécessairement être orienté par une main humaine ; les normes fréquentielles seules ne peuvent suffire dans cette tâche. Si cette assertion met en exergue le caractère indispensable des avis d'experts et de professeurs – auquel nous reviendrons lors du focus sur l'apprentissage en L2 (cf. C) –, nous y retrouvons également une invitation à considérer d'autre types de normes lexicales, reposant sur l'expérience des locuteurs.

B.2. Arguments en faveur des normes psycholinguistiques

D'autres variables sont susceptibles d'affecter le traitement du lexique par les locuteurs. Ces variables se présentent généralement comme des normes « perçues », à la différence des mesures de fréquences, qui sont calculées sur corpus (Stadthagen-Gonzalez & Davis, 2006). En effet, ces autres mesures incluent forcément une part de subjectivité : elles sont généralement obtenues en demandant à un panel de répondants d'estimer, sur base de leur expérience de locuteur, certaines propriétés des mots (le plus souvent au moyen d'une échelle de Likert).

L'une des bases de données lexicales les plus importantes pour la psycholinguistique en anglais est la *MRC*[51] *Psycholinguistic Database* (Coltheart, 1981). Dans celle-ci se retrouvent notamment des mesures de familiarité, d'âge d'acquisition, de concrétude et d'imageabilité. Ces quatre variables, ainsi qu'une cinquième variable introduite par Brysbaert et al. (2016) et Keuleers et al. (2015) - la prévalence - seront présentées dans cette section.

[51] *Medical Research Council*.

Il n'existe que peu de ressources lexicales pour apprenants uniquement basée sur l'une ou l'autre de ces mesures. Ceci s'explique peut-être par leur inhérente subjectivité. Néanmoins, plusieurs études mesurant l'impact de ces dernières dans le traitement des mots ont été menées. Nous pensons que la prise en compte de ces diverses variables puisse être pertinente pour l'élaboration de futures ressources lexicales pour apprenants, et tenterons ici de justifier cette position.

Le principal obstacle à l'utilisation de ces variables réside certainement dans la *rareté* des mesures : évaluer ces données psycholinguistiques à grande échelle (pour beaucoup de mots) est un procédé chronophage, et jusqu'aux années 2010, peu de travaux s'y essayant ont été publiés. Le tableau 1 répertorie les principales bases de données aspirant à établir ces normes pour l'anglais.

i. *Familiarité*

Le concept de familiarité est assez large, et il est difficile de désigner clairement ce à quoi celui-ci correspond : cette variable est notamment fortement corrélée à la fréquence orale et écrite, ainsi qu'à l'âge d'acquisition (Stadthagen-Gonzalez & Davis, 2006). Longtemps, la familiarité a été perçue comme une mesure de la fréquence d'exposition à un mot, si bien que plusieurs auteurs ont avancé que la familiarité s'imposait comme une mesure de fréquence plus acceptable que les fréquences objectives dans des corpus écrits (Gernsbacher, 1984; Gilhooly & Logie, 1980a). Toutefois, l'importance des normes de familiarité s'est rapidement vue relativisée (G. D. A. Brown & Watson, 1987) : la familiarité peut difficilement faire office de mesure de fréquence subjective ; en effet, leur analyse statistique a montré (1) que la familiarité était corrélée plus fortement avec l'âge d'acquisition qu'avec plusieurs mesures de fréquences objectives, (2) que l'âge d'acquisition était un meilleur prédicteur que la familiarité pour des tâches de dénomination lexicale. Plus tard, les auteurs des *Bristol Norms* observent que la familiarité s'impose plutôt comme un estimateur de la fréquence des mots dans le discours oral (Stadthagen-Gonzalez & Davis, 2006, p. 14), et en remettent également en cause l'importance par rapport à d'autres normes telle que l'âge d'acquisition.

De nombreuses études ont montré que les mots plus familiers étaient associés à des temps de réaction moindres dans des tâches de décision lexicale, tant orales qu'écrites (Balota et al., 2002; Connine et al., 1990; Morrel-Samuels & Krauss, 1992). Puisqu'elle joue un rôle notable sur le traitement des mots, la familiarité semble être un facteur à prendre en compte dans l'élaboration de ressources lexicales. Au surplus, la familiarité recouvre certains aspects non pris en compte par la fréquence : en effet, certains référents sont généralement familiers (ex. : *échelle, charnière, désinfectant*) sans pour autant se retrouver fréquemment en corpus (Brysbaert, Mandera, & Keuleers, 2018). Néanmoins, selon d'autres chercheurs, la familiarité est tant associée à la fréquence et à l'âge d'acquisition que l'importance de cette variable semble finalement plutôt minime (Brysbaert & Cortese, 2011; Stadthagen-Gonzalez & Davis, 2006).

Dimension(s)	Nombre de mots	Nombre de participants / mot	Source
CNC	329 noms	?	(Spreen & Schulz, 1966)
CNC, IMAG	925	28 (CNC), 30 (IMAG)	(Paivio et al., 1968)
CNC, IMAG, FAM	2.854	54-65	(Toglia & Battig, 1978)
CNC, IMAG, FAM, AOA	1.944	35-37	(Gilhooly & Logie, 1980a)
CNC, IMAG, FAM, AOA	905	35-37	(Gilhooly & Logie, 1980b)
(désambiguïsation des homographes)			
FAM, AOA, CONC, IMAG	9.392 (FAM)	Fusion statistique des 3 ressources grisées	*MRC Psycholinguistic Database* - (Coltheart, 1981)
	3.503 (AOA)		
	8.228 (CONC)		
	9.240 (IMAG)		
CNC, IMAG	1.080	50	(Friendly et al., 1982)
AOA, IMAG, FAM	297	280 enfants	(Morrison et al., 1997)
IMAG, AOA	2.694	78 (IMAG), 45 (AOA)	(Bird et al., 2001)
IMAG (mots monosyllabiques)	3.000	31	(Cortese & Fugett, 2004)
IMAG, FAM	2.311	16 (IMAG), 47–49 (FAM)	(Clark & Paivio, 2004)
IMAG, FAM, AOA	1.526	20	*Bristol Norms* - (Stadthagen-Gonzalez & Davis, 2006)
AOA (mots monosyllabiques)	3.000	32	(Cortese & Khanna, 2008)
AOA	562	Milliers de questionnaires parentaux (de 8 à 18 mois)	(Goodman et al., 2008)
(AOA)	396-680	Questionnaires parentaux (de 8 à 30 mois)	(Jørgensen et al., 2009)
(AOA)	Corpus de 3.5M de mots	Fréquences dans le langage de l'enfant (de 6 mois à 7 ans : CHILDES)	(Bååth, 2010). CHILDES Project : (MacWhinney, 2000)
AOA	3.460	30	(Khanna & Cortese, 2011)
(désambiguïsation des homographes)			
IMAG (mots disyllabiques)	3.000	35	(Schock, Cortese, & Khanna, 2012)
AOA (mots disyllabiques)	3.000	32	(Schock, Cortese, Khanna, et al., 2012)
AOA	30.124	18-22 pour la plupart	(Kuperman et al., 2012)
CNC	37.058	au moins 25	(Brysbaert, Warriner, et al., 2014)
IMAG, FAM, AOA	629	21 (IMAG), 14 (FAM), 15 (AOA)	(Juhasz et al., 2015)
IMAG, AOA	2.204	277	(S. K. Davies et al., 2016)
(CNC, IMAG, FAM, AOA)	85.942	~~Participants~~ → algorithme de bootstrapping	(Paetzold & Specia, 2016)
CNC, IMAG, FAM, AOA	5.553	~33	*Glasgow norms* - (Scott et al., 2018)
PREV	61.858	388	(Brysbaert, Mandera, McCormick, et al., 2018)

Tableau 1 - Inventaire chronologique des normes lexicales pour l'anglais (repris et adapté de Scott et al., 2018)

* CNC = concrétude / IMAG = imageabilité / FAM = familiarité / AOA = age of acquisition / PREV = prévalence.

** Normes AOA obtenues empiriquement (cf. infra).

*** Complément d'information sur le langage du jeune enfant, dont pourraient dériver d'autres normes AoA.

ii. *Âge d'acquisition*

Récemment, il a été montré que l'âge d'acquisition (AoA) expliquait près de 5% de la variance dans des tâches de décision lexicales pour les mots monosyllabiques, ce en plus des informations fréquentielles (Brysbaert & Cortese, 2011).

La première explication à l'importance des normes d'âge d'acquisition serait que les mesures de fréquences traditionnelles, c'est-à-dire estimée sur un corpus produit par des adultes instruits, sous-estiment les mots utilisés typiquement durant l'enfance. La seconde associe directement âge d'acquisition et structure du lexique mental : l'ordre dans lequel les mots sont appris pourrait influencer la vitesse à laquelle leurs représentations peuvent être activées. Les mots appris en premier sont plus faciles d'accès et plus facilement traités que les mots appris plus tardivement, peut-être parce que le sens de ceux-ci est plus accessible (Brysbaert & Biemiller, 2017; Kuperman et al., 2012).

Typiquement, les normes d'AoA sont obtenues en demandant aux locuteurs à quel âge ils ont appris un certain mot. Une telle approche pose le risque d'être influencée par d'autres facteurs, telle que la longueur du mot, proportionnelle à sa difficulté (Brysbaert & Biemiller, 2017). Cependant, les mesures d'AoA peuvent aussi être basées sur des études empiriques. Ainsi, l'AoA peut s'estimer par le truchement de questionnaires parentaux (cf. Tableau 1), ou encore être dérivée de mesures de familiarité d'enfants d'âges différents avec le vocabulaire (Dale & O'Rourke, 1981 – cf. infra). À cet égard, certains travaux sont particulièrement intéressants, car ils s'inscrivent d'abord dans une volonté de fournir aux enseignants des lignes directrices pour l'acquisition lexicale. En un mot, ce sont des ressources lexicales pour apprenants.

L'intention première de Dale & O'Rourke (1981) n'était pas, en effet, de développer des normes d'AoA, mais d'aider les enseignants à savoir quels mots enseigner à quel niveau scolaire. Pour ce faire, ils ont demandé à des élèves américains de dire s'ils connaissaient ou non 44.000 sens (pour 31.000 formes différentes)[52]. Un sens était assigné à un niveau si entre 67 et 80% des élèves déclaraient connaître celui-ci, sinon il était testé sur les niveaux supérieurs ou inférieurs. Les tests ont été effectués de 1950 à 1980, principalement dans les écoles primaires, et dans deux niveaux en écoles secondaires. Ainsi, chaque sens a été évalué par près de 200 élèves. Plus tard seulement, Brysbaert & Biemiller (2017) ont montré que les normes de connaissance des sens par niveau scolaire fournissaient, au surplus, une estimation satisfaisante de l'AoA : la familiarité des enfants avec les mots permet d'en estimer l'AoA.

Dans la même mouvance, un site d'enseignants américains a publié une liste de 1.461 mots appris du jardin d'enfant à l'année 8.[53] Celles-ci sont basées sur des mesures fréquentielles dans un corpus écrit, et peuvent également être utilisées comme des mesures d'AoA (Brysbaert & Biemiller, 2017).

[52] Cette ressource présente l'avantage supplémentaire de prendre en compte l'information sémantique. Nous y reviendrons dans le chapitre 3 de cet état de l'art.
[53] https://www.flocabulary.com/wordlists/.

iii. *Concrétude et imageabilité*

Certains mots se trouvent essentiellement dans le discours et l'esprit humain, tandis que d'autres mots ont des référents ayant une existence spatio-temporelle, et sont directement appris à travers l'expérience sensorielle des locuteurs. Ces derniers se définissent comme concrets, par rapport aux premiers, qui représentent les mots plus abstraits.

Bien que cette variable ne semble pas directement corrélée aux temps de réaction dans le ELP (cf. note de bas de page 44, p. - 45 -), la concrétude semble avoir un effet significatif sur des bases de données pour d'autres langues, notamment le néerlandais (Brysbaert et al., 2016). En revanche, l'effet observé dans ces expériences va à l'encontre de l'intuition : le temps de réactions sont inférieurs pour les mots abstraits. Cet effet a également été observé par Kousta et al. (2011) et s'expliquerait parce que les mots abstraits tendent à être émotionnellement plus chargés que les mots concrets. Ces observations n'empêchent pas Brysbaert, Stevens, et al. (2014) de proposer que la concrétude soit l'une des variables les plus importantes en ce qui concerne la reconnaissance des mots.

La démonstration de l'importance de la concrétude est plus particulièrement associée à un cadre théorique : la théorie du double codage de Paivio (Paivio, 1971, 2013). D'après cette théorie, les mots concrets seraient plus facilement remémorés parce qu'ils activent les codes de la mémoire perceptive en plus des codes de la mémoire verbale. Dans le même sens, la concrétude a connu un regain d'intérêt lorsque les études en neurosciences ont établi que les mots référant à des entités facilement perceptibles coactivaient les régions du cerveau impliquées dans la perception de ces entités. Des observations similaires ont été rapportées pour les mots « d'action », qui coactivent les régions motrices liées à l'exécution de ces actions (v. Brysbaert, Warriner, et al., 2014).

Outre leur effet supposément réduit, les mesures de concrétude tendent à être basées essentiellement sur la perception visuelle (Lynott & Connell, 2009, 2013), aux dépens d'autres sens. Pour mesurer ce biais, Lynott & Connell (2009) ont demandé aux participants de noter à quel point un adjectif était lié au toucher, au goût, à l'ouïe, etc. Leurs résultats indiquent que les mesures de concrétude étaient surtout corrélées avec les sens de la vision et du toucher. Un problème similaire s'est posé à Brysbaert, Warriner, et al. (2014), qui – malgré des instructions indiquant clairement que la concrétude est basée sur les perceptions de tous les sens ainsi que sur les réponses motrices – observent que leurs mesures de concrétude demeurent principalement centrées sur l'expérience visuelle et haptique.

Quoi qu'il en soit, du fait de ses implications théoriques, il semblerait que la concrétude mérite d'être exploitée dans les ressources lexicales, à la différence des normes d'imageabilité qui – par plusieurs aspects – paraissent moins influentes au regard du traitement des mots (Brysbaert, Stevens, et al., 2014). En effet, l'imageabilité est hautement corrélée à la concrétude (Brysbaert, Stevens, et al., 2014), et la pertinence des mesures d'imageabilité se justifie également par la théorie de Paivio (Cortese & Fugett, 2004). Toutefois, il semblerait que l'imageabilité soit un « doublon » de la concrétude qui insiste trop sur la modalité visuelle (Connell & Lynott, 2012), déjà elle-même sur-représentée par la concrétude (cf. supra).

iv. *Prévalence*

La dernière variable que nous évoquerons est la prévalence, celle-ci réfère au pourcentage de la population indiquant connaître un mot. Celle-ci est exprimée sur une échelle normale standardisée[54]. Les mesures de prévalence sont obtenues en demandant aux participants s'ils connaissent ou non un mot. La prévalence est donc également une mesure subjective.

Il a été montré que cette variable a un effet robuste – le plus important après l'effet fréquentiel - sur le traitement des mots (Brysbaert et al., 2016; Brysbaert, Mandera, McCormick, et al., 2018; Keuleers et al., 2015). En outre, la prévalence est peu corrélée aux autres variables généralement contrôlées (fréquence, AoA, longueur du mot, similarité à d'autres mots), ce qui signifie que les effets observés sur les temps de décision lexicale dans les méga-études sont imputables essentiellement à la prévalence.

Il est important de remarquer que l'effet de la prévalence paraît particulièrement important pour les mots de basse fréquence, car certains sont globalement connus, tandis que d'autres ne le sont quasiment pas (cf. B.1). Cependant, ces effets ne sont pas limités aux mots de basse fréquence (Brysbaert et al., 2016). Par ailleurs, les interprétations de l'effet de prévalence avancées par Brysbaert et al. (2016) en font une solution acceptable aux problèmes qui ont été soulevés à propos des normes fréquentielles. Premièrement, les mots produits plus régulièrement, mais non « capturés » par les corpus sont susceptibles d'être plus connus. Selon cette interprétation, la prévalence serait une mesure complémentaire à la fréquence qui corrigerait les failles de représentativité des corpus sur lesquels sont calculés celles-ci. Deuxièmement, la prévalence pourrait être liée à la familiarité : plusieurs référents sont familiers, mais apparaissent relativement peu en corpus, et les mesures de prévalence parviendraient à saisir cette différence. Enfin, plusieurs mots sont de faible fréquence (et familiarité) sont susceptibles d'être sémantiquement transparents pour les locuteurs (mots composés, dérivés, etc. : *distinctively, antioxidant, microbiologist, antivirus, reusable, legalization*).

En un sens, la liste établie par Dale & O'Rourke (1981 – cf. supra) peut également être considérée comme une étude de la prévalence : ceci, conjugué aux résultats précédemment évoqués, prouve que cette variable est exploitable et pertinente dans le cadre du développement de ressources lexicales. Le cas de Dale & O'Rourke (1981) illustre bien la possible confusion entre familiarité, âge d'acquisition et prévalence : la familiarité est une mesure relativement abstraite, en ce que les manières de la mesurer ne font pas consensus. L'âge d'acquisition et la familiarité étant corrélés, il est possible d'estimer l'âge d'acquisition en mesurant la familiarité d'enfants d'âges différents avec les mots du lexique. Enfin, la mesure de la prévalence s'impose comme une formalisation plus rigoureuse de la familiarité.

---

[54] Ainsi, un mot connu par 2.5% des participants aura une prévalence de -1.96, tandis qu'un mot connu par 97.5% de la population aura une prévalence de 1.96 (Brysbaert et al., 2016).

C. Ressources orientées L2 : classer les mots par niveaux de compétences

Nous l'avons déjà montré dans le début de ce chapitre, au plus la connaissance du vocabulaire est étendue, au mieux les textes seront compris par les apprenants de la langue. Ainsi, l'étendue du vocabulaire est directement proportionnelle au nombre de textes qu'une personne qui étudie l'anglais comme langue seconde sera capable de lire (Dürlich & François, 2018). Néanmoins, il paraît peu probable qu'un apprenant soit capable d'emmagasiner les quelques 150.000 mots qu'un jeune anglophone rencontre supposément tout au long de son enseignement obligatoire (Zeno et al., 1995). En outre, généralement, les apprenants n'ont pas été en contact avec la L2 depuis leur prime enfance et sont moins exposés à cette langue qu'un enfant qui apprendrait sa langue maternelle. Pour cette raison, l'identification des mots principaux à enseigner en langue seconde est peut-être plus importante encore que dans le cadre de l'apprentissage des langues maternelles. Cette identification représente enjeu clé pour les personnes qui conçoivent les programmes d'apprentissage en L2, pour les éditeurs de manuels scolaires et autres documents pédagogiques, ainsi que pour les enseignants (Dürlich & François, 2018).

Pour ce faire, la stratégie la plus communément adoptée reste l'établissement de normes fréquentielles dans des corpus de textes. Dans la section A, nous avons dénombré un certain nombre de ressources se proposant d'identifier, au moyen de listes fréquentielles, les mots les plus importants de la langue anglaise. En plus des manquements de cette approche que nous avons identifiés en B, la méthode fréquentielle ne convient pas à l'enseignement d'une langue seconde : pour commencer, ces listes informent leur utilisateur de la distribution du lexique dans la langue native, ce qui ne correspond que partiellement à la distribution des mots dans la langue apprise aux locuteurs étrangers (que l'on retrouve dans les livres et manuels destinés à ceux-ci) ; ensuite, ces listes ne précisent pas à quel niveau de compétence linguistique tel mot est supposé être appris (Dürlich & François, 2018).

À cet égard, d'ailleurs, nous remarquons que parmi les listes présentées précédemment, très peu se font explicitement fort de proposer une ressource précisément adaptée à l'apprentissage de l'anglais comme langue seconde. L'une des ressources fréquentielles allant le plus dans ce sens est *A Frequency Dictionary of Contemporary American English* de Davies & Gardner (2010). Cette liste des 5.000 lemmes les plus fréquents du corpus COCA[55] est explicitement orientée vers l'apprentissage de l'anglais en tant que langue étrangère :

> We wanted to know which of the vast number of English words *to start with*[56] […] In short, we offer *A Frequency Dictionary of Contemporary American English* with the hope that it will benefit *those who are trying to learn our current mother tongue*[21], as well as for those who desire to assist them. (M. Davies & Gardner, 2010, p. 11)

À cet effet, les auteurs proposent également des listes de vocabulaire thématiques (corps, famille, couleurs, émotions, etc.) ainsi que des informations collocationnelles. Toutefois, les

[55] CORPUS OF CONTEMPORARY AMERICAN ENGLISH. Il s'agit du plus grand corpus pour l'anglais américain : 400 millions de mots au moment de l'élaboration du dictionnaire. Le corpus est équilibré : les mots sont issus de sources diverses (orales, fictionnelles, journalistiques, académiques), au sein desquels des textes de natures et de thèmes divers ont été sélectionnés (M. Davies & Gardner, 2010, p. 3-4).
[56] Nous soulignons.

limitations pointées précédemment restent vérifiées : la liste s'appuie sur des ressources composées par et pour des natifs et n'offre aucune information quant aux niveaux de compétences auxquels les mots devraient être appris.

Selon Jones & Durrant (2010), la construction de ressources pédagogiques devrait être largement informée par des avis d'experts ; il semble qu'à cette époque, l'idée fût déjà profondément ancrée dans le domaine. Au départ de cette hypothèse, des niveaux de référence pour l'apprentissage des langues ont été dressés, et plusieurs ressources lexicales orientées selon ces niveaux « standards » de maîtrise linguistique ont été façonnées. Ces échelons de référence, ainsi que les ressources qui en découlent feront l'objet des trois sous-sections à venir.

C.1. Des niveaux de référence pour l'apprentissage des langues

Lorsqu'un apprenant désire évaluer son niveau d'anglais, une multitude de tests en ligne s'offrent à lui. Il peut également s'inscrire à des tests standardisés (TOEIC, TOEFL ou IELTS). Le point commun entre toutes ces méthodes d'évaluation est qu'ils utilisent ou se réfèrent systématiquement à une échelle de compétences linguistiques de référence : le CECRL (Cadre Européen Commun de Référence pour les Langues), mieux connu sous le nom CEFR (*Common European Framework of Reference for languages).* Le projet, dont le Conseil de l'Europe (2001) marque la publication, avait pour but d'uniformiser l'apprentissage des langues étrangères dans l'Union. Aujourd'hui, la reconnaissance et l'utilisation des échelles CEFR à un niveau international en assure certainement l'hégémonie en tant que cadre de référence pour l'acquisition des L2[57] ; la familiarité des apprenants avec les niveaux définis par le CEFR (allant de A1 à C2) le prouve indéniablement.

L'échelle CEFR consiste en une description des compétences que l'apprenant devrait développer sur six niveaux de compétences : apprenant « élémentaire » (A), « indépendant » (B) et « expérimenté » (C). Un apprenant A1, par exemple, *doit pouvoir* comprendre des mots énoncés simples et des mots du quotidien, tandis qu'un apprenant B1 *doit pouvoir* raconter un événement, un souhait, une émotion, comprendre un texte avec des mots fréquents de la langue et se débrouiller dans la plupart des situations s'il est confronté à des locuteurs natifs. Le dernier niveau, C2, correspond à des compétences proches de celle du locuteur natif : expression spontanée et compréhension de tous types de textes. Néanmoins, les descriptions des compétences requises à chaque niveau restent larges et approximatives – ce qui s'explique notamment par le caractère « multilingue » de l'entreprise – et la mise en pratique problématique de l'échelle CEFR a rapidement été soulignée (North, 2005).

Pour pallier les imprécisions intralinguistiques, des descriptions des niveaux de référence ont été produites pour plus d'une vingtaine de langues européennes (RLDs – *reference level descriptions*). Ces RLDs consistent en des listes de mots, d'expressions, de fonctions ou de structures grammaticales systématiquement associées à

[57] D'autres cadres de référence existent : mentionnons notamment les *Proficiency Guidelines* du *American Council on the Teaching of Foreign Languages* (ACTFL) et l'échelle définie par le *Interagency Language Roundtable* (ILR), essentiellement utilisée par le gouvernement américain. Étant donné que ces échelles sont principalement utilisées à un niveau national (USA) et qu'il n'existe pas – à notre connaissance - de listes lexicales associées, nous en faisons état à titre purement informatif.

l'un des six niveaux définis par le CEFR. Les RLDs ont été créés selon les prescriptions du *Guide for the production of RDL* (2005) émis par la Division des Politiques Linguistiques DG IV du Conseil de l'Europe : elles s'appuient sur des catalogues de fréquences, mais également sur l'étude attentive de grands corpus et de productions d'apprenants, ainsi que sur des avis d'experts (Marello, 2012).

Pour l'anglais, les RLDs ont été construites dans le cadre du projet *English Vocabulary Profile* (EVP), basé à l'université de Cambridge et dirigé par Annette Capel (Capel, 2010, 2012). Attendu que cette ressource lexicale comporte de l'information sémantique, nous y reviendrons à l'occasion du Chapitre 3.C.1.

Une autre échelle de référence existe : la *Pearson Global Scale of English* (GSE). Celle-ci est développée par Pearson, une société mondiale consacrée à l'apprentissage, et se présente comme une extension du CEFR qui identifie, parmi les quatre compétences fondamentales (expression orale, écoute, lecture et écriture), ce qu'un apprenant *doit pouvoir* maîtriser à chaque stade de son apprentissage sur une échelle de 10 à 90. Le cadre GSE est aligné avec les niveaux de référence européens, tout en proposant une approche plus fine et une couverture plus large des compétences (voir Figure 10, à titre illustratif). L'échelle GSE prend en charge les apprenants sous le niveau A1, et étend les niveaux A2 à B2.

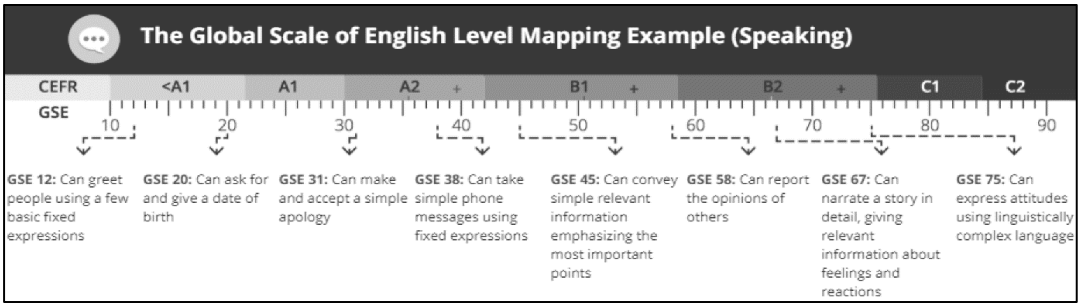


Figure 10 - comparaison des échelles CEFR et GSE en expression orale [58]

Les RLDs associées à l'échelle de Pearson sont basées sur des recherches impliquant plus de 6000 professeurs d'anglais basés dans près de 50 pays. Par rapport aux descriptions associées au CEFR, celles-ci présentent l'avantage notable d'être adaptées au groupe d'apprenants auxquels elles se destinent. Ainsi, en ce qui concerne l'acquisition du vocabulaire, deux ressources ont été élaborées jusqu'ici : l'une d'elles présente un inventaire lexical des mots devant être maîtrisés à chaque niveau de compétence par un apprenant *adulte* (Benigno & de Jong, 2017b) ; l'autre transpose cet objectif à des apprenants plus *jeunes* - de six à onze ans (Benigno & de Jong, 2017a). À l'instar des bases de données développées par Capel, ces dernières prennent en compte le sens des mots, et seront pour cette raison exposées plus amplement dans le Chapitre 3.C.1.

L'une des faillites de ces RLDs, identifiée par Dürlich & François (2018), est qu'elles ne permettent pas, au sein d'un niveau de compétence donné, de distinguer quels sont les mots les plus importants à apprendre. Dans le niveau B1, par exemple, le mot *happiness* est-il plus ou moins important que son hyperonyme *feeling* ? Ni les descriptions du *English Vocabulary Profile*, ni celles de *Pearson* ne permettent d'avancer une réponse justifiée à cette question. En revanche, le projet CEFRLex le peut.

[58] https://www.pearson.co.jp/en/forabrighterfuture.

C.2. CEFRLex

Lancé en 2014, le projet CEFRLex propose une approche originale de l'acquisition du lexique, qui approfondit la conception progressive de l'apprentissage du vocabulaire. À cet égard, les ressources CEFRLex vont au-delà des RLDs, qui supposent que les apprenants d'un certain niveau devraient connaître tous les mots attribués à ce niveau (Dürlich & François, 2018). En effet, le projet CEFRLex décrit la distribution fréquentielle des unités lexicales sur chacun des niveaux du cadre de référence européen.

Ces distributions sont le fruit d'analyses statistiques menées sur des corpus de documents pédagogiques (manuels et livres de lecture) destinés aux apprenants L2, et classifiés selon leur niveau sur l'échelle CEFR par des experts de l'enseignement des L2. Une telle approche offre un aperçu plus précis de l'usage du mot, et permet, par exemple, de savoir qu'un mot habituellement considéré appris à un certain niveau peut déjà apparaître dans des textes pour apprenants de plus bas niveau. Par ailleurs, les fréquences par niveaux permettent – au sein d'un même niveau – de classer les mots par ordre d'importance.

Des ressources lexicales CEFRLex existent pour plusieurs langues européennes. À ce jour, des listes de vocabulaire ont été construites pour le français (François et al., 2014), le suédois (François et al., 2016), l'anglais (Dürlich & François, 2018), le néerlandais (Tack et al., 2018), et l'espagnol (François & De Cock, 2018).

i. *Une ressource CEFRLex pour l'anglais : EFLLex (Dürlich & François, 2018)*

La ressource anglaise s'appuie sur un corpus des ressources produites par des éditeurs populaires tels que *Cambridge University Press*, *Oxford University Press*, ou *Exam English Ltd.* (1.971 textes, issus de 17 manuels, 33 livres de lectures et 7 documents en ligne à destination d'apprenants de l'anglais comme langue seconde, pour un total de 486.178 mots). Tous ces textes ont été classés sur l'échelle CEFR par leurs éditeurs. Disposant de trop peu de textes de niveau C2, les auteurs ont choisi de n'inclure que les niveaux de A1 à C1. De plus, le corpus est biaisé en faveur des plus hauts niveaux (B2 et C1), pour lesquels une plus grande quantité de mots sont disponibles, car les textes tendent à être plus longs dans ces niveaux (Dürlich & François, 2018, p. 874).

Calculer les fréquences des formes fléchies pourrait causer des problèmes d'interprétation (les verbes ont de nombreuses formes fléchies, mais d'autres mots sont invariables, comme les adverbes et les prépositions). Pour cette raison, les auteurs ont choisi d'observer les fréquences d'occurrence des formes lemmatisées. En outre, il paraît probable que les apprenants soient le plus souvent capables de relier les formes fléchies à leur lemme. Nous l'avons déjà dit, l'étiquetage morpho-syntaxique permet également de désambiguïser les homographes correspondant à différentes étiquettes. Pour cette tâche, le *NLP4J Tagger* a été choisi pour sa précision (95.88%) ainsi que pour sa capacité à détecter des expressions polylexicales (ex. : mots composés – *shopping center* et verbes à particule – *pick up*)[59]. Ainsi, chacune de 15.280 entrées d'EFLLex résulte de la combinaison d'un lemme et d'une étiquette morpho-syntaxique.

[59] Qui sont d'importance capitale pour l'apprentissage en L2 mais absentes de plusieurs des listes de fréquences développées pour l'anglais (Dürlich & François, 2018, p. 876).

Pour chacune de ces entrées, les fréquences EFLLex sont normalisées par millions de mots (*fpmw*). Pour ce faire, les auteurs ont utilisé un indice de dispersion basé sur celui de Carroll et al. (1971), qui a permis de pondérer les fréquences brutes et de prendre en compte la diversité contextuelle.

Le corpus sur lequel est basé CEFRLex est assez peu volumineux. Il convient pour cela de ne pas tirer de conclusions hâtives sur les mots qui semblent apparaître peu fréquemment dans EFLLex. Toutefois, les auteurs font remarquer que les fréquences EFLLex sont fortement corrélées aux fréquences du corpus BNC (Leech et al., 2001), ce qui assure que les normes EFLLex soient un bon estimateur de l'usage réel des mots.

En outre, EFLLex inclut tous les mots rencontrés dans le corpus, or certains de ceux-ci peuvent être d'importance mineure pour l'acquisition lexicale en L2. Il est donc nécessaire de prendre du recul sur les mesures fréquentielles avant développer des objectifs d'apprentissage sur cette base (Dürlich & François, 2018, p. 877). De plus, les auteurs rappellent que la ressource lexicale se concentre uniquement sur le vocabulaire réceptif, puisque tous les documents du corpus sont associés à une tâche de compréhension.

Finalement, l'application de notre précédent exemple (*happiness vs. feeling,* niveau B1) sur la base de données EFLLex nous apprend (1) que le mot *happiness* apparaît déjà dans des textes destinés à des apprenants de niveau inférieur - A1 – et qu'au contraire, feeling apparaît plus fréquemment dans textes destinés aux apprenants de niveau C1 ; (2) au sein-même des textes de niveau B1, la ressource EFLLex indique aussi que *feeling* apparaît plus fréquemment que *happiness*. Pour deux mots, les résultats se prêtent bien à la représentation graphique (Figure 11) [60].

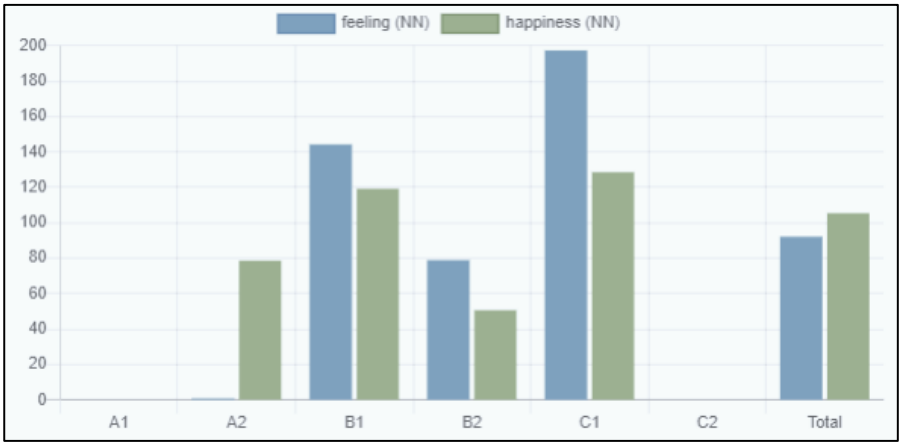


Figure 11 - résultats de l'outil de recherche EFLLex pour 'feeling' *et* 'happiness'

L'un des objectifs du présent travail étant de développer une ressource parallèle à EFLLex – EFLSemLex (*English as a Foreign Language Semantic Lexicon*), nous montrerons, dans le prochain chapitre, les avantages de la prise en charge d'informations sémantiques dans le contexte de l'élaboration de ressources lexicales, et présenterons les listes de vocabulaire ayant déjà exploré cette voie, en anglais et dans d'autres langues.

[60] UCLouvain, https://cental.uclouvain.be/cefrlex/search.

Chapitre 2) – À retenir :

Depuis longtemps, l'apprentissage explicite du vocabulaire sous forme de listes est reconnu comme une méthode efficace – bien que non suffisante – pour l'acquisition du lexique d'une langue étrangère (ou maternelle, pour les enfants de primaire).

Dans ce chapitre, nous avons vu que l'effet facilitateur de la fréquence pour l'apprentissage du lexique est largement reconnu dans la communauté scientifique. Par conséquent, plusieurs listes de vocabulaire (ou ressources lexicales) fréquentielles ont été conçues depuis le début du XX^{e} siècle.

Néanmoins, nombreuses sont les listes fréquentielles qui présentent le désavantage de s'adresser aux enfants (L1) ou aux apprenants (L2) en s'appuyant sur des corpus produits par des locuteurs natifs adultes. D'autres ressources lexicales fréquentielles s'affranchissent du caractère pédagogique, et sont – entre autres – utilisées dans des applications en NLP ou dans des expériences psycholinguistiques.

En outre, les fréquences ne peuvent être considérées comme l'unique explication des divergences dans la connaissance des mots. Certains mots de base fréquence sont en effet bien connus (e.g. *soulmate*). Les diverses limites, dont notamment le caractère objectif (mesuré), des normes fréquentielles ont mené plusieurs chercheurs à se pencher sur d'autres normes subjectives (perçues) pour l'analyse de la difficulté du lexique : familiarité, âge d'acquisition, concrétude, imageabilité et prévalence.

Enfin, pour la L2, les simples normes fréquentielles n'informent pas les utilisateurs à propos du niveau de compétence linguistique auquel les mots devraient être appris. Certaines ressources établissent un classement des mots par niveau de compétence (EVP, GSE). À cet effet, l'échelle de compétence la plus connue est celle du CEFR, qui établit six niveaux d'apprentissage, du plus élémentaire au plus complexe (A1, A2, B1, B2, C1, C2).

Le projet CEFRLex propose des distributions fréquentielles réparties sur les 6 niveaux CEFR, et ce pour plusieurs langues. Cela permet de connaître l'importance relative des mots dans chacun de ces niveaux.

La ressource CEFRLex pour l'anglais est EFLLex, celle-ci a été décrite en profondeur, car la ressource que nous avons créée, EFLSemLex, en est très proche, idéologiquement et méthodologiquement. La principale différence entre EFLLex et EFLSemLex est le caractère désambiguïsé de cette dernière.

Avant de passer à la Partie II), concernant l'élaboration d'EFLSemLex, nous verrons en quoi cette phase de désambiguïsation pourrait être utile dans les ressources pour apprenants, et aborderons quelques-unes des principales ressources lexicales désambiguïsées connues (Chapitre 3).

Chapitre 3. Vers des ressources lexicales désambiguïsées

Quels sont les avantages, pour les apprenants, de disposer de ressources lexicales désambiguïsées ? Cette question peut également être posée comme suit : en quoi les mots ambigus sont-ils susceptibles d'avoir un caractère bloquant dans l'apprentissage et la pratique d'une langue seconde ?

Dans ce dernier chapitre théorique, nous proposons des réponses orientées selon les axes de l'apprentissage et de la compréhension. Pour ce faire, une approche *bottom-up* est adoptée : pour commencer, nous nous demanderons si les mots ambigus, individuellement, sont plus difficiles que les mots univoques (A) ; ensuite, nous verrons comment l'ambiguïté impacte la compréhension phrastique et textuelle (B). Pour finir, nous nous arrêterons sur les ressources lexicales qui existent pour l'anglais (C.1) et d'autres langues (C.2).

A. Difficultés sémantiques au niveau du mot

Le besoin de prendre en considération les variations sémasiologiques pour estimer la difficulté d'un mot remonte, selon Tack et al. (2018), aux travaux de Tharp (1939). Celui-ci avait déjà soulevé la faiblesse de la mesure de fréquence formelle comme estimateur de la difficulté lexicale : des formes similaires peuvent, en effet, avoir des significations différentes. Le mot n'est donc pas connu entièrement, qu'il soit polysémique (l'apprenant n'a alors pas une connaissance de tous les sens « dérivés » d'un mot), ou homonymique (l'apprenant peut très bien associer une forme à son sens, mais ignorer que la même forme peut être associée à un sens tout à fait différent).

En règle générale, les mots sémantiquement ambigus sont traités différemment des mots non ambigus (Eddington & Tokowicz, 2015). Rodd (2018) cite de nombreuses études montrant qu'il est difficile pour les enfants d'assigner de nouveaux sens aux mots déjà connus. L'idée que cette observation convoie est que les associations univoques entre une forme et un sens (*one-to-one mapping*) sont plus simples à acquérir que les associations d'une même forme à plusieurs sens (*one-to-many mapping*)[61].

Traditionnellement, ceci s'explique par le principe d'exclusivité mutuelle : les enfants ont tendance à préférer les associations uniques entre un référent et son étiquette. Lors de l'acquisition d'une langue seconde, chaque mot univoque dispose d'une association relativement claire entre une forme orthographique et/ou phonologique et une représentation sémantique en mémoire. Ceci n'est pas vrai pour les mots ambigus ; ainsi, ceux-ci demandent plus de ressources cognitives pour être appris. Bien entendu, les enfants ne sont pas éternellement bloqués par leur difficulté à passer au-dessus de la signification principale d'un mot, et les enfants bilingues n'ont d'ailleurs pas le choix que d'intégrer rapidement qu'un même référent peut avoir des labels différents (Degani & Tokowicz, 2010a).

[61] Rodd (2018) signale toutefois qu'il demeure plus facile d'apprendre de nouvelles significations pour des mots familiers, que d'apprendre de nouveaux mots (ce qui demande d'intégrer à la fois une nouvelle forme *et* un nouveau sens).

Si l'ambiguïté est particulièrement difficile à acquérir pour l'enfant, elle ne l'est pas moins pour l'adulte. Degani & Tokowicz (2010a), partant du constat de la pauvreté des recherches sur l'apprentissage des mots ambigus chez l'apprenant adulte, interprètent les difficultés rencontrées par ces derniers comme une conséquence de l'organisation du réseau lexico-sémantique : le stockage des mots ambigus implique une forme d'interférence ou de compétition active entre les significations alternatives.

Sur cette base, il a été suggéré que les différents types d'ambiguïtés seraient traités et représentés différemment dans l'esprit des locuteurs (Eddington & Tokowicz, 2015). En psycholinguistique, il est généralement admis que les homonymes sont stockés dans des endroits différents du lexique mental, tandis que les polysème seraient organisés dans une entrée unique, éventuellement sous-spécifiée sémantiquement[62]. Des concepts mieux reliés sémantiquement sont plus fortement connectés dans le réseau lexico-sémantique que des concepts peu – voire pas – associés sémantiquement. Par conséquent, la polysémie serait généralement plus facile à intégrer en mémoire que l'homonymie, du fait de la proximité sémantique entre les sens associés. En effet, des significations associées partagent un nombre important de caractéristiques communes, ce qui a pour conséquence une meilleure imprégnation des sens nouveaux dans le réseau sémantique des locuteurs.

L'avantage de la polysémie ne concerne pas uniquement l'apprentissage, mais également le traitement lexical. De nombreuses études (v. Eddington & Tokowicz, 2015) ont découvert un effet facilitateur de la polysémie dans des tâches de décision lexicale, par exemple. Cette observation pourrait être imputable à l'activation sémantique additionnelle provenant des sens similaires, ou encore au caractère sous-spécifié du sens dans l'entrée mentale du polysème. Cet avantage n'est pas observé dans les cas d'homonymie, ce qui souligne encore l'importance du degré de proximité sémantique dans le traitement de l'ambiguïté. Eddington & Tokowicz proposent même que l'homonymie inhiberait la reconnaissance.

De plus, les locuteurs compétents d'une langue sont en règle générale conscients de légères nuances sémantiques (ou de registre) pour les mots ambigus. Cependant, ces subtiles différences de sens doivent être apprises, et les débutants d'une L2 ne disposent d'ordinaire pas de l'expérience nécessaire pour acquérir et appliquer ces nuances de manière appropriée. Beaucoup de manuels pour débutants ne rendent d'ailleurs pas ces distinctions clairement visibles (Degani & Tokowicz, 2010a). EFLSemLex se présente comme un moyen pour combler cette lacune :

> [D]irectly mapping L2 words to their meanings appears to assist the learner in reducing the challenges of learning translation ambiguous words. (Degani & Tokowicz, 2010a, p.311)

En plus d'associer explicitement chaque mot à ses définitions, EFLSemLex fournit également de l'information quant à la distribution fréquentielle des sens. Nous justifions la pertinence de cette entreprise dans les paragraphes suivants.

Précédemment, nous avons observé que le *sens principal* du mot fait souvent obstacle à l'acquisition de sens nouveaux chez l'enfant. Nous pourrions, intuitivement, supposer, que le sens principal d'un mot est celui auquel l'enfant est confronté en premier lieu, celui qui *doit*

[62] Le sens du mot polysémique serait, alors, actualisé et précisé – notamment grâce aux indices contextuels – au moment de la lecture.

être appris avant les autres. Une telle définition peine à rendre compte de la plasticité du réseau lexico-sémantique des locuteurs, et occulte les mécanismes à l'œuvre dans l'évolution de celui-ci. Au plus un apprenant est confronté aux différents sens d'un mot, au plus les connections entre la forme et chacune des significations sont établies et renforcées (Eddington & Tokowicz, 2015). La fréquence des sens, à la lumière de cette assertion, prend un rôle crucial : cette fréquence change avec l'expérience, de telle sorte que le sens dominant peut être relégué au rang de sens subordonné, et *vice versa*. Par exemple, le mot « *service* » pourrait avoir comme sens dominant « ce que l'on fait pour quelqu'un, avantage qu'on lui procure bénévolement », jusqu'à ce que l'individu décide d'intégrer un club de tennis et que le sens sportif de « *service* » devienne – pour cet individu – dominant.

Il se pourrait, en outre, que les mots ambigus soient plus difficiles à apprendre parce que chacune de leurs alternatives sémantiques sont rencontrées moins souvent que les mots non ambigus. Selon Degani & Tokowicz (2010a), il est probable que cela se vérifie dans l'apprentissage des langues naturelles, puisque la fréquence d'occurrence formelle des mots ambigus est « morcelée » en ses multiples significations.

Pour ces deux raisons, il nous semble particulièrement important de prendre en considération la fréquence des sens dans le cadre de l'acquisition d'une langue seconde. C'est ce que nous proposons avec EFLSemLex. Premièrement, cette ressource permet au professeur d'avoir un aperçu du sens dominant et des sens subordonnés à chaque niveau du CEFR, et de mieux comprendre, par conséquent, les problèmes associés à l'apprentissage des sens à chaque stade de l'apprentissage. Deuxièmement, la ressource pourrait, en elle-même, également lever partiellement les difficultés des apprenants grâce à l'illustration explicite de la fréquence des sens des mots ambigus ainsi que grâce à la comparaison visuelle de ces fréquences avec celles des mots non ambigus.

En définitive, les mots ambigus sont plus difficiles à apprendre et à traiter que les mots non ambigus. Degani & Tokowicz (2010a), qui ont travaillé sur le traitement de l'ambiguïté dans le cadre de la SLA, ont montré que les mots ambigus de la langue cible sont produits et reconnus – comme traductions correctes – moins rapidement et moins précisément que les mots univoques. Ces chercheurs ont montré qu'entraîner les mots ambigus aussi souvent que les mots non ambigus ne suffit pas à réduire ce désavantage, et que les *one-to-many mappings* représentent, *per se*, un défi pour l'apprentissage.

Bien sûr, un apprenant pourra toujours apprendre le sens des mots ambigus en prêtant une attention particulière aux contextes dans lesquels ces mots surviennent (Rodd, 2018). Cela dit, l'ambiguïté est un phénomène si prévalent dans la langue que chercher à l'enseigner implicitement par le biais de la seule lecture paraît utopique :

> Therefore, it would be advantageous to identify teaching methods that are specifically tailored to these ambiguities, to provide learners with a better starting point when they try to communicate in L2. (Degani & Tokowicz, 2010a, p.311)

Le présent travail, et la ressource que nous avons développée dans le cadre de ce dernier, s'offrent comme une réponse possible à cette proposition.

B. Difficultés sémantiques au niveau de la phrase

La connaissance du vocabulaire réceptif est importante pour la compréhension à la lecture d'une langue seconde (cf. introduction du chapitre 2). Bien que l'importance de prendre en compte les associations entre les formes et les sens ait été clairement démontrée dans le cadre de la lecture en L2, les distinctions sémantiques ont souvent été ignorées dans les approches lexicales (Tack et al., 2018).

En 1985, Groebel, au moyen d'une étude comparative d'étudiants américains (L1) et d'étudiants israéliens (L2), a montré que la résolution d'ambiguïtés (isolées) et la compréhension à la lecture sont positivement corrélées. L'auteur postule que de bonnes compétences en résolution d'ambiguïtés affecterait positivement la compréhension à la lecture. Les difficultés liées à la résolution d'ambiguïtés seraient – au moins partiellement – associées à des représentations lexico-sémantiques faibles (Henderson et al., 2013). Selon cette perspective, le traitement de l'ambiguïté apparaît comme élément central à la maîtrise d'une langue. D'autres études ont obtenu des résultats similaires (Rodd, 2018). Les auteurs de celles-ci s'accordent pour affirmer que l'incapacité à comprendre une phrase sémantiquement ambigüe bloque la compréhension fluide et rapide d'une langue.

Selon Groebel (1985), il serait salutaire pour l'apprentissage que ces observations mènent à des remaniements des matériaux d'apprentissage. Nous pensons qu'il est intéressant d'offrir aux apprenants et aux professeurs une ressource qui offrirait un aperçu « fixé » du sens des mots et de leurs fréquences, afin d'augmenter éventuellement les compétences de désambiguïsation des apprenants, ainsi que leurs compétences métalinguistiques relatives au phénomène même d'ambiguïté.

Qian (1999) souligne aussi que, dans l'interaction entre la taille du vocabulaire et la compréhension à la lecture, la notion de profondeur du vocabulaire joue un rôle important. Cette dernière notion inclut – au-delà des niveaux de surface (orthographe, phonologie, morphologie) – une maîtrise profonde des aspects *sémantiques*, collocationnels, contextuels et discursifs du mot. Les aspects sémantiques et collocationnels, en particulier, se sont avérés efficace pour ajouter de la profondeur au vocabulaire et améliorer les scores de compréhension.

Plusieurs études ont cherché à paramétriser l'étendue des variations sémasiologiques – et onomasiologiques – dans le cadre de l'évaluation de la *lisibilité textuelle*. Pour ce faire, des caractéristiques de polysémie, d'hypéronymie, et d'autres liens conceptuo-lexicaux de WordNet ont été explorés (Tack et al., 2018). Dans ce contexte, la contribution la plus remarquable est Coh-Metrix[63] (Graesser et al., 2004).

Toujours dans le champ de la lisibilité, Bailin & Grafstein (2016) ont montré que l'ambiguïté était susceptible d'interférer avec la compréhension, même si le lecteur n'a pas conscience de l'ambiguïté. Pour ces auteurs, mesurer la difficulté des mots décontextualisés d'un texte ne permet pas de lever le voile sur ce qui rend un texte plus ou moins facile à comprendre :

[63] Coh-Metrix est un système qui analyse diverses mesures textuelles (cohérence, cohésion, paramètres lexicaux) afin de fournir une mesure de difficulté pour un texte.

> It is important to look at how the words are shaped by the text and the assumptions needed to understand them. (Bailin & Grafstein, 2016, p.114)

Un mot en lui-même peut ne pas être intrinsèquement difficile, mais tout de même être problématique. Par exemple, « grammaire » renvoie à (1) l'évaluation de ce qui est correct dans la langue, ou (2) la description formelle des composants d'une phrase (noms, verbes, adjectifs, etc.). Ces définitions pourraient égarer un lecteur lisant Chomsky sans avoir une connaissance préalable des notions de linguistique chomskienne et, en particulier, de « *grammaire générative* ». Le mot « *grammaire* » n'en serait pas rendu plus difficile pour autant.

Généralement, le sens doit être pris et compris dans une fenêtre contextuelle. Dans ce cas précis, le sens du mot ne serait pas compréhensible grâce au contexte direct : le lecteur doit avoir une certaine connaissance du monde et/ou du domaine de spécialité du texte qu'il lit. La question n'est donc plus purement lexicale : qu'est-ce qui mène le lecteur à ajuster ses interprétations selon le contexte interne et externe du texte ? Quels sont les mécanismes qui gèrent ces réajustements ? Cette question a largement été discutée dans le domaine de la psycholinguistique.

En psycholinguistique, le processus de désambiguïsation sémantique est considéré comme la récupération – ou activation – de la signification la plus appropriée d'un mot dans un texte[64]. Trois modèles d'accès lexical sont évoqués par Ishida (2019) dans le cadre du traitement des ambiguïtés : l'accès sélectif, l'accès multiple et l'accès ordonné. Premièrement, le modèle d'accès sélectif propose que seule la signification contextuellement appropriée serait activée dans l'esprit du locuteur. Deuxièmement, le modèle d'accès multiple suppose que toutes les significations sont d'abord activées, la sélection du sens étant alors perçue comme un processus autonome subséquent. Troisièmement, l'accès ordonné avance que le sens le sens le plus fréquent serait activé en premier ou, plus exactement, que les sens seraient tous activés, mais que le sens dominant serait directement privilégié.

Le modèle d'accès multiple est majoritairement favorisé dans les études psycholinguistiques. Toutefois, la fréquence du sens jouerait aussi un rôle important dans l'activation. Les prochains paragraphes explorent plus profondément les mécanismes de désambiguïsation selon le modèle d'accès lexical multiple, ainsi que les implications de ce modèle dans le cadre de notre travail.

Pour Rodd (2018), les locuteurs sont capables d'extraire en parallèle – rapidement et automatiquement – tous les sens qu'ils connaissent pour un mot, et ensuite, en une centaine de millisecondes, de sélectionner la signification jugée la plus adéquate. Dans cette optique,

[64] Les études à ce sujet sont nombreuses. Au niveau neurolinguistique, notamment, les zones cérébrales les plus associées au traitement des phrases contenant des mots ambigus ont pu être mises en évidence : gyrus frontal inférieur gauche (plus connu comme « Aire de Broca »), le gyrus frontal inférieur droit, et le cortex temporal postérieur gauche (Rodd, 2018).

l'accès aux unités ambigües est considéré comme « exhaustif » : tout est d'abord activé, puis un choix très rapide est effectué[65].

D'autres études (cf. Rodd, 2018) ont montré que si les mots précédents sont fortement liés au sens dominant – le plus fréquent – d'un mot, seul ce sens sera activé. Cette observation ne récuse pas le modèle d'accès multiple, mais amène à reconnaître que les niveaux d'activation sont influencés par deux facteurs clé : le contexte linguistique, et la fréquence du sens. Par conséquent, les sens hautement fréquents et les sens appelés par le contexte sont plus aisément accessibles.

D'ordinaire, l'information textuelle additionnelle permet de lever les ambiguïtés. Si, lorsque le mot ambigu est rencontré, le contexte est neutre et ne permet pas de décider avec certitude du sens, la décision est tout de même prise en quelques centaines de millisecondes, car les capacités humaines de traitement de la phrase ne permettent pas de maintenir plusieurs interprétations concurrentes en parallèle. Dans ce cas, la décision du locuteur est basée sur la fréquence des sens. Malheureusement, l'heuristique fréquentielle n'est pas exempte d'erreurs. La suite de la phrase ou du texte invite alors le locuteur à revisiter sa prime interprétation ; ce processus de réinterprétation peut être couteux en ressources cognitives, et impacter notamment le temps de lecture (Rodd, 2018).

Il arrive, par exemple, qu'aucun indice ne permette, dans la phrase, de prendre une décision définitive : « *The father turned around and saw a **bat** flying through the air* »[66]. Supposons que le sens perçu comme plus fréquent – et donc plus probable – soit « un animal nocturne », mais que le texte contienne plus tard « *that heavy **object*** … » : le lecteur serait alors certainement confus, et devrait revoir son interprétation initiale du texte.

Il peut, selon nous, être intéressant de fournir aux apprenants des connaissances « méta » sur les sens qu'ils sont susceptibles de rencontrer le plus fréquemment, afin que les prises de décisions se basent sur des mesures objectives, plutôt que sur des impressions subjectives potentiellement inexactes. En effet, les connaissances fréquentielles subjectives sont susceptibles d'être plus imprécises pour les apprenants que pour les locuteurs natifs. EFLSemLex – grâce à ses mesures de fréquences sémantiques réparties selon les niveaux CEFR – offrirait aussi aux enseignants la possibilité de savoir quels sens les apprenants ont le plus de chances de rencontrer dans chaque niveau de textes. Corollairement, notre ressource pourrait soutenir la création de listes de vocabulaire adaptées par niveaux. La distribution par niveau de difficulté peut effectivement différer de la distribution globale.

En outre, il a été reconnu que la fréquence des mots a moins d'influence en L1 qu'en L2. Une moindre exposition à une langue pourrait mener à des effets de fréquence amplifiés. L'écart entre les mots très et peu fréquents a un plus grand impact sur les apprenants que sur les locuteurs natifs, en particulier si lesdits apprenants sont peu confrontés à la langue. Par conséquent, la fréquence des sens ambigus pourrait également avoir un plus grand effet sur le traitement des ambiguïtés en L2 (Ishida, 2019). Cette dernière hypothèse est importante, car elle implique que la fréquence des sens – telle que perçue par des apprenants – pourrait

[65] Cette conception est appuyée par de nombreuses expériences psycholinguistiques telles que le priming sémantique *cross-modal*, l'utilisation du masque sémantique, la complétion de tâches parallèles non reliées à l'interprétation de la phrase ambigüe, etc. (v. Rodd, 2018, pour plus de précisions).
[66] Cet exemple montre que la décidabilité n'est pas une fonction de la proximité des sens.

causer des mésinterprétations sémantiques, due à une favorisation trop prononcée du sens le plus fréquent. Une ressource telle que CEFRLex permettrait aux apprenants d'améliorer leurs compétences réceptives en offrant une meilleure connaissance des sens inhabituels et – éventuellement – une relativisation du sens dominant.

Notons que d'autres informations peuvent également influencer l'activation lexicale (Rodd, 2018) : l'expérience récente (les locuteurs sont également plus susceptibles de récupérer le sens le plus récemment rencontré, même si celui-ci est moins fréquent) et la variété de langue (les locuteurs sont capables d'observer quels usages sémantiques reviennent souvent au sein d'un même registre ou dialecte).

Bien que les chercheurs s'accordent à dire qu'un choix est fait très rapidement parmi toutes les alternatives, il n'y pas d'accord clair sur le devenir des sens non retenus. Il n'est pas clairement établi que ces sens soient complètement supprimés pour les empêcher d'interférer avec le traitement de la phrase, ou qu'ils gardent un faible niveau d'activation en cas de nécessité de réinterprétation.

Par ailleurs, les dissemblances entre les locuteurs natifs et les apprenants en termes de processus de désambiguïsation ne font pas consensus. Certains auteurs (Love et al., 2003) reportent des mécanismes différents entre les locuteurs en L1 et en L2 ; d'autres (Elston-Güttler & Friederici, 2007) affirment que les apprenants et les locuteurs natifs présentent des fonctionnements de désambiguïsation fondamentalement similaires. Les résultats des recherches récentes en L2 appuient cette seconde hypothèse, bien que l'activation du sens adéquat paraisse plus lente et plus éphémère chez les apprenants (Ishida, 2019).

Enfin, Rodd (2018) indique que, dès quatre ans, les enfants se débrouillent relativement bien pour désambiguïser une phrase dont un seul mot de contexte est significatif (« *Snoopy chased/swung the* ***bat*** »). Les enfants ne se basent pas uniquement sur des associations lexicales, mais également sur des évaluations plus globales à propos des sens les plus probables en contexte. Néanmoins, les enfants font plus d'erreurs que les adultes dans la compréhension de phrases ambiguës.

Pour conclure, il apparaît que les difficultés associées à la résolution des ambiguïtés sont inhérentes aux langues naturelles, tant à cause de la complexité propre aux mots ambigus, qu'en raison de leur présence dans les textes, multipliant par leur ambivalence les interprétations possibles. Puisque l'ambiguïté est omniprésente dans le langage, et que les obstacles associés affectent indubitablement l'apprentissage, de nombreux auteurs proposent d'adapter les ressources d'enseignement à cette problématique. Dans la section suivante, nous présentons des bases de données lexicales qui prennent en compte l'information d'ordre sémantique.

C. Ressources lexicales désambiguïsées

Nous venons de soulever quelques-unes des principales difficultés liées à l'information d'ordre sémantique. Existe-t-il un grand nombre de ressources lexicales pour l'apprentissage expressément désambiguïsées ? Nous répondrons d'abord à cette question pour l'anglais, puis évoquerons quelques ressources notables existant dans d'autres langues.

C.1. Pour l'anglais

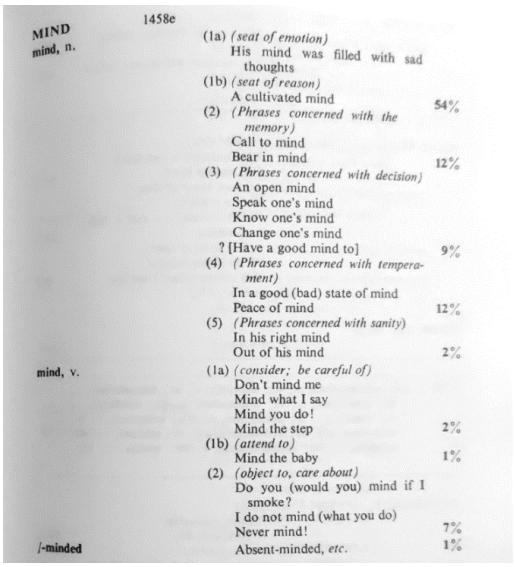
Nous avons soulevé, précédemment, l'apparente nécessité d'introduire dans le matériel d'apprentissage des informations d'ordre sémantique. À l'exception des dictionnaires pour apprenants, il existe peu de ressources lexicales à visée pédagogique qui intègrent de l'information d'ordre sémantique[67]. Nous excluons d'emblée les dictionnaires de notre propos, car ceux-ci ne réalisent aucun des deux objectifs que nous poursuivons avec EFLSemLex : (1) observer la répartition des sens sur les niveaux européens de maîtrise d'une langue, (2) présenter des comptes de fréquences sémantiques dans une optique d'acquisition de l'anglais.

Dans la suite, nous présenterons chronologiquement cinq ressources notables qui se donnent au moins l'un des objectifs précités. Il est difficile de trouver d'autres ressources de ce genre, ce qui atteste une fois encore du bien-fondé de la démarche entreprise avec EFLSemLex.

i. *A General Service List of English (West, 1953)*

La *General Service List* (GSL) établit un inventaire du vocabulaire de base – 2000 mots – pour l'apprentissage de l'anglais. Toutefois, la première version de celle-ci (Faucett et al., 1936) laisse beaucoup de place au jugement de l'enseignant, qui doit choisir quel sens du mot enseigner en priorité.

West propose de dépasser ce jugement en améliorant la liste au moyen de la fréquence sémantique des mots (Lorge & Thorndike, 1938). Ces fréquences ont été obtenues manuellement sur base d'un corpus de 5M de mots issus principalement de l'anglais écrit - extraits d'encyclopédies, de magazines, de manuels, de romans, d'essais, de biographies, de livres scientifiques, de poèmes, etc.[68] Les distinctions sémantiques utilisées sont tirées du *Oxford Dictionary of English*.


1458e

MIND
mind, n.
(1a) *(seat of emotion)*
His mind was filled with sad thoughts
(1b) *(seat of reason)*
A cultivated mind 54%
(2) *(Phrases concerned with the memory)*
Call to mind
Bear in mind 12%
(3) *(Phrases concerned with decision)*
An open mind
Speak one's mind
Know one's mind
Change one's mind
? [Have a good mind to] 9%
(4) *(Phrases concerned with temperament)*
In a good (bad) state of mind
Peace of mind 12%
(5) *(Phrases concerned with sanity)*
In his right mind
Out of his mind 2%

mind, v.
(1a) *(consider; be careful of)*
Don't mind me
Mind what I say
Mind you do!
Mind the step 2%
(1b) *(attend to)*
Mind the baby 1%
(2) *(object to, care about)*
Do you (would you) mind if I smoke?
I do not mind (what you do)
Never mind! 7%

/-minded
Absent-minded, *etc.* 1%

Figure 12 - Entrée de la GSL de West (1953)

Dans le compte sémantique initial, les fréquences étaient exprimées en nombre d'occurrences pour 1000 mots ; dans la liste de West, celles-ci sont exprimées en pourcentages relatifs, afin que la ressource soit exploitable au mieux par les enseignants. Une telle présentation indique en effet clairement quel sens du mot est dominant.

Il apparaît que peu de ressources similaires aient été créées par la suite. Cela peut s'expliquer par le temps et les ressources nécessaires pour établir base de données d'une

[67] Nous n'évoquons pas les listes de vocabulaire spécialisé, qui – en un sens – possèdent de l'information sémantique, représentée dans ce cas par l'appartenance du mot à un domaine. Il semble, en effet, qu'au sein d'un même champ d'expertise, un mot disposera rarement de plusieurs sens.

[68] Les auteurs sont conscients du manque de sources orales pour le comptage. Les fréquences obtenues tendent, par conséquent, à surestimer certains sens purement « littéraires ».

telle envergure. Actuellement, la désambiguïsation lexicale automatique permettrait d'établir de tels comptes plus aisément. Toutefois, les algorithmes ne sont pas parfaitement précis. Pour autant, l'information sémantique n'est nécessairement mise au ban : d'autres ressources, plus ou moins récentes, exploitent les mesures sémantiques fréquentielles.

ii. *The Living Word Vocabulary (Dale & O'Rourke, 1981)*

Dans l'optique d'aider les instituteurs à décider de quels mots enseigner à quels niveaux scolaires, Dale & O'Rourke ont demandé à des élèves américains s'ils connaissaient ou non 44.000 sens de la langue anglaise (pour 31.000 formes différentes). Cette recherche s'est étendue de 1950 à 1980 dans les écoles primaires et dans les deux premières années de l'enseignement secondaire. Ce faisant, chaque sens a été évalué par près de 200 élèves issus de milieux socio-économiques divers.

En règle générale, un sens est associé à un niveau (*i.e.* à une année de l'enseignement), si 67% à 80% des élèves de cette année ont déclaré le connaître, faute de quoi il est testé sur les niveaux inférieurs ou supérieurs. Certains sens peuvent néanmoins faire exception à cette règle (cf. Figure 13)[69].

Grade	Score	Word — Word Meaning
16	78%	**abruption** — a sudden breaking off
08	71%	**abscess** — wound with pus
12	31%	**abscind** — to cut apart
16	72%	**abscissa** — horizontal coordinate
16	84%	**abscond** — run away and hide
04	67%	**absence** — being away
06	91%	**absence** — not having something
04	84%	**absent** — not here

Figure 13 - quelques entrées de Dale & O'Rourke (1981). Les entrées sont désambiguïsées ('absence') *et sont associées à des pourcentages de connaissance dans chaque année scolaire*

De cette manière, via une approximation de la familiarité d'élèves de primaire et de secondaire avec les sens, les auteurs ont été capables de produire une ressource lexicale graduée (selon les années d'enseignement) et désambiguïsée. Cette démarche a été reprise et approfondie par Biemiller (2010).

iii. *English Vocabulary Profile (EVP - Capel, 2010, 2012)*

Le *EVP* se présente comme une ressource pour les enseignants et les concepteurs de matériaux pédagogiques. Cette ressource offre de l'information concernant le niveau CEFR des mots (du *sens* des mots, en particulier) et expressions en langue anglaise. La figure 14 montre que différents sens d'un mot peuvent être associés à des niveaux de compétences différents.

[69] Nous supposons qu'il s'agit de sens trop peu connus dans les niveaux inférieurs, mais trop bien connus dans les niveaux supérieurs. Pour *absence*, par exemple, le second sens (score : 91%) serait assigné à l'année 6, parce que c'est dans celle-ci que le taux de connaissance serait le plus proche de la fourchette 67%-80%. N'ayant pas eu accès à la ressource intégrale, nous devrons nous en tenir à cette hypothèse.

Base Word	Guideword	Level	Part of Speech	Topic
match	BE THE SAME	B1	verb	
match	BE AS GOOD AS	C1	verb	
match	SUITABLE	C2	noun	
match	STICK	B2	noun	
match	COMPETITION	A2	noun	shopping
match	CHOOSE	B1	verb	people: actions

Figure 14 - quelques entrées pour 'match' dans le EVP

Cette ressource compte près de 7.000 entrées (Capel, 2012). Au lieu d'adopter une attitude normative quant à ce qui devrait être connu à quel niveau, les auteurs ont préféré observer quels sens les apprenants connaissent effectivement à tel ou tel niveau. Les mesures obtenues sont basées sur le CAMBRIDGE LEARNER CORPUS (50M de tokens), le CAMBRIDGE ENGLISH CORPUS OF FIRST LANGUAGE USE (1.2b de tokens), ainsi que d'autres corpus pour les niveaux supérieurs C1 et C2.

Dans l'EVP, les mots polysémiques sont traités en profondeur, et le projet a cherché à savoir dans quel ordre les diverses significations sont acquises. La ressource montre également que les sens moins fréquents continuent d'être acquis au cours de l'apprentissage (Figure 15). Certains sens – s'ils sont moins fréquents ou plus spécialisés – ne sont pas repris dans la ressource.

Headword	Meanings at A1–B2 levels	Meanings at C levels	Learner example for C level entry?
click (verb)	COMPUTER (A2)	IDEA (C2)	Yes
		SOUND (C2)	Yes
		PEOPLE (C2)	Yes
force (noun)	POWER (B2)	INFLUENCE (C2)	Yes
	GROUP (B2)		
Memory	COMPUTING (A2)	MIND (C2) (the part of your mind that stores what you remember – *He recited the poem from memory.*)	No
	ABILITY TO REMEMBER (B1)		
	EVENT REMEMBERED (B1)		
plain (adjective)	SIMPLE (B1)	OBVIOUS (C2)	Yes
	NOT MIXED (B1)	PERSON (C2)	Yes
rough	NOT SMOOTH (B1)	DANGEROUS (C1)	Yes
	NOT EXACT (B1)		
	SEA/WEATHER (B2)		
	DIFFICULT (B2)		
Serve	PROVIDE FOOD/DRINK (A2)	BE USEFUL (C1)	Yes
		WORK (C1)	Yes
	SHOP (B1)	PRISON (C2)	Yes
Within	TIME (B1)	INSIDE (C1)	Yes
	DISTANCE (B1)		
	LIMIT (B2)		

Figure 15 – les sens des mots sont progressivement ajoutés aux niveaux CEFR (Capel, 2012)

Bien que l'information sémantique et l'ordre d'acquisition des sens soient des informations capitales pour l'enseignement, l'absence de données fréquentielles au sein d'un même niveau ne permet pas de connaître l'importance relative des mots au sein de ce niveau. Par exemple, *happiness* et *feeling* (cf. Chapitre 2.C.2.i) sont tous deux compris dans le niveau B1. Pour autant, les descriptions ne permettent pas de savoir lequel de ces deux mots devrait être appris en priorité.

iv. *PHaVE List (Garnier & Schmitt, 2015)*

La *PHaVE List* est une liste pédagogique de verbes à particules (*phrasal verbs*) associés à leurs sens les plus fréquents. Pour son aspect « localisé », la *PHaVE List* nous a semblé pertinente à évoquer ici : cette dernière illustre les nombreuses potentialités de l'information sémantique fréquentielle.

À l'image de la GSL, l'objectif est ici d'aider les enseignants à sélectionner les sens principaux des verbes à particules. Les recherches ont montré que, généralement, les verbes à particules sont hautement polysémiques : ces derniers disposent en moyenne de 5.6 sens, dont plusieurs sont plutôt marginaux.

Garnier & Schmitt ont établi une liste des 150 *phrasal verbs* les plus fréquents dans les corpus BNC et CORPUS OF CONTEMPORARY AMERICAN ENGLISH (COCA). Cette liste donne également les fréquences relatives de chacun des sens principaux de ces verbes. Les sens sont accompagnés de définitions et d'exemples d'utilisation accessibles aux apprenants. Pour chaque verbe de la *PHaVE,* les sens donnés couvrent, ensemble, plus de 75% des occurrences de ce verbe dans le COCA.

Les comptes de fréquences de cette liste sont estimés sur base de 200 lignes de concordances tirées aléatoirement du COCA et annotées manuellement. L'inventaire sémantique utilisé résulte de la combinaison de plusieurs dictionnaires importants (*Cambridge, Oxford, Merriam-Webster, Collins, MacMillan*) et du réseau lexical WordNet.

Il est intéressant de noter, finalement, que la *PHaVE List* s'inspire explicitement de la méthode de la *General Service List* de 1953, ce qui atteste de l'influence et de la longévité de la GSL dans le cadre de l'EFL.

v. *Pearson Global Scale of English - GSE (Benigno & de Jong, 2017a, 2017b)*

L'objectif principal de la *Global Scale of English* est de déterminer ce qui doit être maîtrisé à quel niveau d'apprentissage pour divers types de tâches linguistiques (écoute, lecture, vocabulaire, etc.) Parmi ces « tâches », l'acquisition du lexique joue un rôle majeur, et est traitée avec précision.

Le pan lexico-sémantique de la GSE a été développé afin de fournir de l'information prescriptive concernant le niveau des sens en l'anglais : « Tel sens devrait être connu au niveau B2[70] ». Les prescriptions du GSE sont issues d'une large enquête regroupant plus de 6000 professeurs d'anglais basés dans près de 50 pays, et menée par la société privée *Pearson*. À l'instar d'EVP, la GSE s'intéresse aux sens, mais l'approche d'EVP est descriptive, tandis que celle de *Pearson* est normative.

Par ailleurs, les auteurs de la GSE sont convaincus de la nécessité de dresser des listes de vocabulaire basées sur le sens plutôt que sur la forme :

> Each lexical unit is a distinct word meaning - in line with research evidence demonstrating that vocabulary learning gradually develops from basic to complex and that, therefore, not all meanings of a word are learned at once. Take, for example, two different meanings of *can*: our analysis suggests that the verb *can* (to express ability) should be learned before the noun *can* (as in a can of coke). (Benigno & de Jong, 2017a, p. 3)

Une force supplémentaire de la GSE est qu'elle n'est pas uniquement basée sur des comptes de fréquences en corpus, mais également sur le jugement de professeurs. Celui-ci intervient dans la décision d'associer un tel sens à un tel niveau, mais également lors de la sélection des sens les plus utiles, communicativement parlant.

En outre, la GSE présente l'avantage d'être adaptée au public-cible : à ce jour, pour l'acquisition du vocabulaire, deux ressources ont été compilées. La première présente un inventaire lexical de plus de 37.000 sens devant être maîtrisés à chaque niveau par un

[70] La GSE a également établi sa propre catégorisation (linéaire) des niveaux de compétences, mais celle-ci peut directement être associée à l'échelle CEFR (cf. Figure 10).

apprenant *adulte* (Benigno & de Jong, 2017b) ; la seconde fait de même, mais pour des apprenants plus *jeunes* – de six à onze ans (Benigno & de Jong, 2017a). Cette dernière ne compte que 3.000 sens, plus essentiels, afin de ne pas surcharger la mémoire des enfants. Les auteurs prévoient qu'une troisième ressource devrait également voir le jour, pour des enfants/adolescents plus âgés.

Vocabulary	Topic	Grammatical Category	GSE	CEFR
stream	Geographic features Geographic features Natural or built features and structures	noun	43	B1 (43-50)
Definition: a very small river		*Collocations:* a mountain stream an underground stream a stream runs somewhere		
stream	Moving Physical phenomena Portion or piece of	noun	71	B2+ (67-75)
Definition: a flow of liquid, gas, smoke etc		*Collocations:*		
stream	Driving People or things Things that happen	noun	72	B2+ (67-75)
Definition: a long series of people, vehicles, ideas etc coming continuously or one after another		*Collocations:* a steady stream a revenue stream		

Figure 16 - quelques entrées sémantiques pour 'stream' *dans le « GSE Teacher Toolkit »*[71] (version adultes)

La version basique (pour apprenants adultes) de la GSE est basée sur trois corpus : LONGMAN CORPUS NETWORK (330M de mots), UKWAC (2 milliards de mots) et les sources orales du COCA (90M de mots). La version pour jeunes enfants, quant à elle, s'appuie sur les sources orales du BNC (~25.000 mots), sur le corpus CHILDES (~160.000 mots – v. Tableau 1), sur $SUBTLEX_{UK}$ (13M de tokens – cf. Chapitre 2.A.2.vi), ainsi que sur des manuels pour jeunes apprenants, sur la liste de vocabulaire de WFG (cf. Chapitre 2.A.1.iv) de Zeno et al. (1995), et enfin sur la liste de 3.000 mots simples de Dale & Chall (1948).

Les auteurs de la *Global Scale of English*, tout comme ceux du *English Vocabulary Profile*, n'expliquent pas précisément comment a été réalisée la désambiguïsation des mots des corpus.

C.2. Pour d'autres langues

Il n'est pas possible de passer exhaustivement en revue toutes les ressources lexicales désambiguïsées (ou intégrant de près un de loin de l'information sémantique) dans toutes les langues. Pour cette raison, nous avons retenu deux ressources ayant en commun la caractéristique suivante : chacune illustre et affirme les bénéfices de la sémantique pour les ressources lexicales.

i. *Néerlandais : NT2Lex (Tack et al., 2018)*

Cette ressource, développée dans le cadre du projet CEFRLex (cf. Chapitre 2.C.2), présente des distributions de fréquences pour plus de 17.000 mots et expressions du néerlandais. Les particularités de NT2Lex sont similaires à celles que nous souhaitons donner à EFLSemLex. Premièrement, il s'agit d'une ressource graduée : l'information fréquentielle est exposée selon les niveaux de maîtrise européens. Deuxièmement, il s'agit d'une ressource désambiguïsée : les fréquences *de sens* sont préférées aux fréquences *des formes*.

[71] https://www.english.com/gse/teacher-toolkit/user/vocabulary.

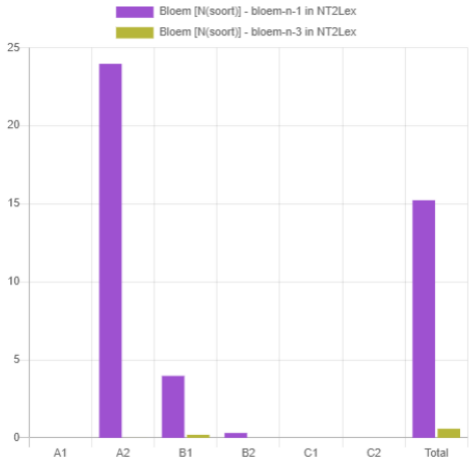


Figure 17 - NT2Lex : fréquence de bloem *('fleur' – en mauve) vs.* bloem *('farine' – en vert)*

Le corpus sur lequel est basé NT2Lex se compose de 461.088 *tokens* issus de livres de lecture et manuels de néerlandais L2, distribués des niveaux A1 à C1. Les écrits proviennent aussi bien de la variété belge du néerlandais, que de la variété néerlandaise (ou « hollandaise »). Puisqu'elle se base sur des productions expertes, la ressource NT2Lex donne un bon aperçu de ce qui devrait *a priori* être compris à tel ou tel niveau de maîtrise. Cependant, la constitution du corpus ne permet pas l'observation des usages réels des apprenants, ce que les auteurs identifient comme une faiblesse.

L'aspect véritablement novateur de NT2Lex réside dans le fait que les entrées lexicales soient désambiguïsées sémantiquement et liées aux synsets du WordNet néerlandais (*Open Dutch WordNet*). Cette étape de désambiguïsation est présentée comme le moyen d'obtenir une meilleure compréhension de la complexité lexico-sémantique des productions d'apprenants de la langue.

Les étapes de création de NT2Lex trouvent un parfait parallèle dans celles que nous avons mis en place pour l'anglais : choix d'un corpus gradué, choix d'une méthode de désambiguïsation de ce corpus, compte des fréquences sémantique sur les différents niveaux de compétence langagière. Au surplus, la finalité pédagogique de la ressource se retrouve également dans EFLSemLex. En un mot, NT2Lex a fortement inspiré le présent travail.

ii. *Français : ReSyf (Billami et al., 2018)*

Bien qu'elle ne présente ni comptes de fréquences explicites ni répartition sur les niveaux du CEFR, la ressource ReSyf demeure intéressante à aborder : elle constitue un autre bon exemple de ce que la désambiguïsation automatique peut apporter aux ressources lexicales.

Resyf est une ressource de synonymes ordonnancés selon leur difficulté à être lus et compris par des locuteurs natifs du français. Les synonymes sont issus du réseau lexical JeuxDeMots (Lafourcade, 2007), dans lesquels les mots sont liées par leurs relations de synonymie. Ces synonymes sont ensuite désambiguïsés par une méthode fondée sur la similarité vectorielle, et enfin ordonnancés par un modèle de *pairwise SVM ranking* utilisant les trois niveaux de difficulté de Manulex (Lété et al., 2004) comme données d'entraînement, et associant à chaque sens un vecteur de caractéristiques[72].

[72] Les caractéristiques retenues sont d'ordre orthographiques, fréquentielles, sémantiques et morphologiques.

Au total, ReSyf compte 57.589 entrées, et est composée majoritairement de substantifs monosémiques. Les ensembles de synonymes désambiguïsés et ordonnancés de ReSyf ont été confrontés à des annotations humaines, et, dans 91% des cas, l'ordonnancement automatique concorde avec l'ordonnancement manuel.

Les auteurs soulignent que l'information fournie par la ressource est pertinente dans le contexte éducatif : pour soutenir des productions écrites claires et efficaces, mais également pour adapter les matériaux d'apprentissage. En outre, ReSyf peut être utilisée pour la simplification lexicale (LS)[73].

Une ressource désambiguïsée similaire à ReSyf existe pour l'espagnol : CASSAurus (Baeza-Yates et al., 2016). En revanche, les synonymes qui s'y trouvent sont répartis en deux classes seulement : simple et complexe, ce qui est peut-être moins approprié dans le cadre de la LS.

Mentionnons, avant de continuer, une autre ressource de vocabulaire désambiguïsée *pour le suédois*. David Alfter nous a fait part de l'existence de celle-ci lors d'un entretien personnel (D. Alfter, communication personnelle, 11 mai 2022). Malheureusement, cette ressource n'a pas été publiée et n'est pas publiquement accessible[74]. Cette information prouve encore que, bien que peu populaire, la question de la désambiguïsation des ressources lexicales n'est pas inopportune ou hors de propos.

Chapitre 3) – À retenir :

De nombreux travaux – notamment en acquisition des langues (étrangères) et en psycholinguistique – ont montré que l'ambiguïté sémantique pourrait rendre les mots et les phrases plus complexes, et donc impacter l'acquisition lexicale et la compréhension de textes.

Les recherches s'accordent à reconnaître l'influence de la fréquence sémantique. Les sens les plus fréquemment rencontrés domineraient les autres, et seraient ainsi appris et récupérés plus facilement que les sens subordonnés. À cet égard, la ressource EFLSemLex permettra d'avoir un bon aperçu des significations principales et secondaires à chaque stade de l'apprentissage.

Les difficultés posées par l'ambiguïté ont amené certains auteurs à proposer d'adapter les matériaux pédagogiques pour mieux tenir compte des aspects sémantiques. En ce sens, nous avons identifié quelques ressources lexicales désambiguïsées pour l'anglais (depuis les années 50). Celles-ci demeurent toutefois peu nombreuses. Par ailleurs, la ressource CEFRLex pour le néerlandais (NT2Lex – Tack et al., 2018) aura particulièrement retenu notre attention, car il s'agit de la seule ressource CEFRLex désambiguïsée publiée à ce jour.

La méthodologie respectée pour la création de NT2Lex est très similaire à celle qui sera décrite ci-après. Nous invitons maintenant le lecteur à prendre connaissance, dans la Partie II) de ce mémoire, des modalités pratiques relatives à la concrétisation d'EFLSemLex.

[73] Remplacement des mots compliqués par une alternative plus simple dans les textes.

[74] Elle est uniquement utilisée dans quelques universités, notamment l'université de Gothenburg, en Suède.

Partie II) Compilation de la ressource EFLSemLex

Chapitre 1. Méthodologie

Le développement de EFLSemLex s'inspire de la méthodologie utilisée dans le cadre du projet CEFRLex, en particulier de celles de EFLLex (Dürlich & François, 2018) et NT2Lex (Tack et al., 2018). La ligne directrice est la suivante : sur la base d'un corpus de textes classés selon leur niveau CEFR par des annotateurs humains, la fréquence de chaque mot est estimée et normalisée pour chacun des niveaux CEFR. Ce procédé permet d'obtenir une distribution fréquentielle des mots sur l'échelle CEFR.

Dans le cadre de l'élaboration de EFLSemLex, nous n'avons pas cherché à obtenir une distribution fréquentielle des mots, mais une distribution fréquentielle des sens. Le glissement opéré de la forme vers le sens – inspiré de NT2Lex – a requis la mise en place de certaines solutions « locales ».

Dans ce chapitre, nous décrirons d'abord quantitativement le corpus sur la base duquel nous avons obtenu notre distributions sémantique (A). Ensuite, nous présenterons et proposerons une évaluation de l'algorithme de désambiguïsation lexicale utilisé (B), avant de décrire la technique d'estimation des fréquences (C). Le chapitre se poursuivra sur la piste explorée face au problème de granularité de l'inventaire sémantique utilisé (D), pour se conclure sur la description d'une expérience menée afin d'évaluer l'apport de la désambiguïsation pour la ressource (E).

A. Description quantitative du corpus

Le corpus utilisé pour EFLSemLex est largement inspiré du corpus collecté par Dürlich et François dans le cadre de l'élaboration de EFLLex (2018). Il est toutefois important de noter que notre corpus est significativement plus réduit que le corpus utilisé pour EFLLex : telle quelle, la ressource EFLSemLex ne peut donc pas être comparée directement à EFLLex (2018).

En effet, notre corpus compte 857 textes en moins (1.971 vs. 1.114), et comporte moins de la moitié des mots du corpus initial (486.176 vs. 199.527). Le nombre de mots de notre corpus, par ailleurs, est à comprendre « par excès » : pour les compter, nous avons utilisé *TreeTagger* (Schmid, 1994), tandis que les textes de EFLLex ont été taggués au moyen du *tagger* inclus dans le module *NLP4J* (Choi, 2016). Nous avons remarqué que, pour un même texte, le *TreeTagger* tend à segmenter plus que le *tagger NLP4J*. À *tagger* identique, notre corpus compterait donc moins de 199.527 *tokens*[75].

Les textes extraits proviennent d'onze manuels et sept ressources en lignes. Ces cinq sources sont prévues pour les apprenants de l'anglais en langue seconde, et proviennent de quelques-uns des éditeurs les plus connus (*Cambridge University Press, Linguapress, Exam English Ltd*)[76]. Les textes retenus pour ce corpus sont associés à des tâches de

[75] Cela s'explique par notre utilisation du *tagger* : à l'inverse de Dürlich et François (2018), nous ne tenons pas compte des expressions polylexicales (*multi-word expressions*), ni des verbes à particules (*phrasal verbs*). Cette décision est justifiée par l'importance secondaire de la phase de *POS-tagging* dans ce travail : le véritable *tagging* des unités lexicales est effectué par le module de désambiguïsation (cf. B).

[76] À titre de comparaison, le corpus initial d'EFLLex est tiré de 17 manuels, 7 ressources en ligne et 33 ouvrages classés par niveaux (*graded readers*).

compréhension à la lecture, et ont été classifiés sur l'échelle CEFR par les équipes d'édition. Attendu que très peu de textes sont disponibles pour le niveau C2, notre corpus ne couvre que les cinq premiers niveaux CEFR (A1, A2, B1, B2, C1).

Genres	A1		A2		B1		B2		C1		Total	
Ad	14	*841*	11	*1571*	15	*2309*	7	*1465*	8	*1389*	55	*7575*
Dialogue	43	*2952*	52	*5395*	62	*8099*	45	*9858*	27	*5610*	229	*31914*
E-Mail / Mail	1	*23*	18	*1489*	22	*2118*	24	*4771*	9	*1084*	74	*9485*
Informative	22	*2088*	39	*7509*	64	*16736*	90	*29029*	103	*32945*	318	*88307*
Narrative	24	*2477*	11	*1838*	15	*2948*	24	*5964*	11	*2589*	85	*15816*
Recipe	\	\	\	\	2	*180*			\	\	2	*180*
Sentences	54	*3271*	35	*3143*	46	*4136*	17	*2240*	20	*2815*	172	*15605*
Various	16	*1319*	37	*4757*	51	*8327*	38	*8638*	37	*7604*	179	*30645*
Total	**174**	**12971**	**203**	***25702***	**277**	***44853***	**245**	***61965***	**215**	***54036***	**1114**	***199527***
(Mean text length)	74,55		126,61		161,92		252,92		251,33		179,11	

Tableau 2 – Distribution du nombre de textes et du nombre de mots dans le corpus par niveau et par type de texte

Le tableau 2 donne à voir la distribution des mots et textes par niveaux de maîtrise. Ce tableau décrit également la répartition des textes selon leur genre (*e.g.* dialogues, e-mails, textes narratifs, etc.), soulignant ainsi la variété de textes considérés. Il est à noter que les niveaux supérieurs (B2 et C1) représentent le plus large sous-corpus. Ceci s'explique par la plus grande disponibilité de textes avancés, ainsi que par une longueur moyenne des textes plus élevée pour ces deux niveaux.

La plupart des mots du corpus sont « lexicaux », *i.e.* ils appartiennent aux classes grammaticales ouvertes (noms, verbes, adjectifs et adverbes). Parmi ceux-ci, une grande majorité peut être considérée comme polysémiques[77]. Le tableau 3 offre une description précise du corpus en termes de *tokens*, de *types*, de nombres de sens, et de catégories grammaticales.

Avant de continuer, nous évoquerons deux failles importantes qui affleurent à la suite de la description de notre corpus. Premièrement, Brysbaert & New (2009) ont identifié une longueur idéale pour établir des normes fréquentielles pertinentes :

> […] for most practical purposes, a corpus of 16–30 million words suffices for reliable word frequency norms. In particular, there is no evidence that a corpus of 3 billion words is much better than a corpus of 30 million words. For these sizes, it becomes more important to know where the words of the corpus came from. (Brysbaert & New, 2009, p. 980)

Que l'on considère la version du corpus utilisée dans cette ressource ou le corpus utilisé par Dürlich & François (2018), le nombre de mots (*tokens*) est largement en deçà de la fourchette idéale suggérée par la précédente citation. En outre, ladite fourchette concerne les fréquences formelles ; puisqu'il est admis qu'il y a plus de sens que de mots[78], il paraît raisonnable de supposer que, dans l'optique de mesurer des fréquences sémantiques, la taille du corpus devrait être revue à la hausse (afin d'accroître le nombre de sens observables).

[77] Selon la typologie de WordNet 3.1. : nous considérons un mot comme étant polysémique si, *au sein d'une même classe grammaticale*, plusieurs synsets contiennent ce mot.
[78] Le substantif moyen compte 1.23 sens, et le verbe moyen compte 2.16 sens (Jurafsky & Martin, 2021).

	A1		A2		B1		B2		C1		Total	
N° tokens	12971		25702		44853		61965		54036		199527	
• lexicaux	6926	53,40%	13590	52,88%	23801	53,06%	33377	53,86%	30081	55,67%	107775	**54,02%**
ꝏ polysémiques	5620	81,14%	11027	81,14%	19053	80,05%	26727	80,08%	23612	78,49%	86039	**79,83%**
ʘ monosémiques	1306	18,86%	2563	18,86%	4748	19,95%	6650	19,92%	6469	21,51%	21736	**20,17%**
• grammaticaux	6045	46,60%	12112	47,12%	21052	46,94%	28588	46,14%	23955	44,33%	91752	**45,98%**
ꝏ polysémiques	2615	43,26%	4694	38,75%	7594	36,07%	9994	34,96%	7879	32,89%	32776	**35,72%**
ʘ monosémiques	3430	56,74%	7418	61,25%	13458	63,93%	18594	65,04%	16076	67,11%	58976	**64,28%**
◊ Verbes	2475	35,73%	4759	35,02%	8555	35,94%	11674	34,98%	9846	32,73%	37309	**34,62%**
◊ Noms (communs)	2955	42,67%	5592	41,15%	9526	40,02%	13826	41,42%	12892	42,86%	44791	**41,56%**
◊ Adjectifs	792	11,44%	1848	13,60%	3228	13,56%	4456	13,35%	4477	14,88%	14801	**13,73%**
◊ Adverbes	704	10,16%	1391	10,24%	2492	10,47%	3421	10,25%	2866	9,53%	10874	**10,09%**
N° types	1945		3907		5941		8501		9011		16506	
• lexicaux	1504	77,33%	3078	78,78%	5092	85,71%	7502	88,25%	8152	90,47%	14063	**85,20%**
ꝏ polysémiques	1017	67,62%	2001	65,01%	3270	64,22%	4726	63,00%	5100	62,56%	7967	**56,65%**
ʘ monosémiques	487	32,38%	1077	34,99%	1822	35,78%	2776	37,00%	3052	37,44%	6096	**43,35%**
• grammaticaux	441	22,67%	829	21,22%	849	14,29%	999	11,75%	859	9,53%	2443	**14,80%**
ꝏ polysémiques	164	37,19%	284	34,26%	283	33,33%	328	32,83%	298	34,69%	648	**26,52%**
ʘ monosémiques	277	62,81%	545	65,74%	566	66,67%	671	67,17%	561	65,31%	1795	**73,48%**

Tableau 3 – caractéristiques du corpus d'EFLSemLex

Deuxièmement, la fin de la citation de Brysbaert et New rappelle l'importance de l'origine des mots du corpus. Au sujet de l'approche du projet CEFR, Billami et al. (2018) indiquent que la couverture des corpus utilisés est limitée. En effet, tous les textes exploités sont issus de ressources pédagogiques : cette procédure ne donne pas d'informations sur d'autres synonymes plus complexes qui pourraient tout autant être rencontrés. Bien qu'il semble opportun de s'appuyer sur des textes destinés aux apprenants pour créer une ressource lexicale destinée aux apprenants, il est indéniable que certains mots en seront – de ce fait – systématiquement exclus.

B. Désambiguïsation automatique du corpus

L'une des étapes les plus cruciales de la création d'EFLSemLex est la désambiguïsation automatique du corpus. En effet, un algorithme de désambiguïsation peu efficace fournira nécessairement des résultats peu représentatifs[79]. Dans ce sous-chapitre, nous justifierons d'abord (B.1) le choix de l'algorithme utilisé. Nous détaillerons ensuite le fonctionnement de celui-ci (B.2), avant de présenter les résultats de l'évaluation de l'algorithme sur notre corpus (B.3).

B.1. Choix d'un modèle de désambiguïsation

Dans la Partie I) Chapitre 1. B.6, nous avons présenté, pour chacun des types de systèmes de WSD, les meilleurs algorithmes existant (cf. Figure 9). Pour ce travail, nous avons choisi d'utiliser EWISER (Bevilacqua & Navigli, 2020). Bien que les résultats d'EWISER avoisinent ceux des systèmes les plus performants, il n'est plus – à ce jour – à considérer comme état de l'art dans le domaine de la désambiguïsation. Notre décision s'explique par deux facteurs pratiques que nous présentons dans les paragraphes suivants.

Premièrement, le projet EWISER – contrairement à d'autres modèles – dispose d'un plugin *spacy* (Honnibal & Montani, 2017) qui rend le programme facilement exploitable. Ce plugin nous a permis de désambiguïser automatiquement l'ensemble de textes constituant le corpus (1,114). Au vu de la complexité des algorithmes de désambiguïsation actuels, l'utilisation d'un module pré-entraîné nous a semblé être le meilleur moyen de se garder d'éventuels dysfonctionnements.

[79] Cela est d'autant plus crucial que, comme évoqué dans la section III.A., la taille du corpus utilisé est loin d'être optimale.

Deuxièmement, EWISER était encore à la pointe de la technologie durant l'été 2020. Bien que ce système ait été dépassé par des travaux ultérieurs, il ne nous semble pas que les améliorations connues soient fondamentalement révolutionnaires en termes de performances. EWISER offre 78,3 % de précision sur tous les ensembles d'évaluation de SemEval. Cela ne représente qu'une différence de 0,7% par rapport au meilleur *token classifier* (BEM, Blevins & Zettlemoyer, 2020), et de 2,4% en comparaison au meilleur système, toutes catégories confondues (ESCHER, Barba et al., 2021).

Pour ces deux raisons, nous avons retenu EWISER comme algorithme de désambiguïsation lexicale. Si ce dernier n'est pas totalement optimal, il nous semble que ce choix est tout de même acceptable dans le cadre de ce travail. Cette hypothèse sera vérifiée dans la sous-section 3. D'abord, nous présentons plus en détails la procédure de désambiguïsation sur laquelle EWISER est fondé.

B.2. EWISER – un système hybride pour couvrir tous les sens

Comme tous les modèles de désambiguïsation lexicale actuels, EWISER[80] repose sur une architecture neuronale. L'originalité de la méthode proposée réside dans son emploi de la variété de relations lexico-sémantiques encodées dans le graphe de WordNet pour reconnaître les sens : « *which is often neglected due to the non-trivial integration of data of this kind into neural architectures* » (Bevilacqua & Navigli, 2020, p. 2854). Cette méthode hybride (basée sur des connaissances linguistiques et des *embeddings* pré-établis, ainsi que sur un corpus d'entraînement) permet de prédire des synsets qui ne sont pas présents dans l'ensemble d'entraînement. Ainsi, EWISER couvre – *i.e.* peut reconnaître – tous les sens (synsets) présents dans WordNet.

EWISER aborde la tâche de désambiguïsation comme un problème de classification de *tokens*. Formellement, ce problème peut se définir comme suit : chaque mot (en contexte) est représenté sous la forme d'un vecteur **h**. Ce vecteur est utilisé pour générer une distribution de probabilité **z** sur tous les labels possibles de ce mot, c'est-à-dire, sur tous ses sens possibles[81]. Pour plusieurs instances à classifier, les vecteurs **h** sont stockés dans une matrice **H**, et les scores **z** sont stockés dans une matrice **Z**. L'architecture de base d'EWISER est relativement simple :

$$\begin{aligned} B &= B_{-4} + B_{-3} + B_{-2} + B_{-1} \\ H_0 &= \text{BatchNorm}(B) \\ H_1 &= \text{swish}(H_0 W + \mathbf{b}) \\ Z &= H_1 O \end{aligned} \qquad \text{(a)}$$

Pour chaque mot à désambiguïser, le réseau prend comme entrée la somme des outputs des 4 dernières couches de BERT Large-cased (de $\mathbf{B_{-4}}$ à $\mathbf{B_{-1}}$). Et utilise un réseau *feed-forward* à deux couches ($\mathbf{H_0}$ et $\mathbf{H_1}$) pour calculer les scores **Z**. $\boldsymbol{W}$ et **b** sont des paramètres du modèle, et « swish »[82] correspond à la fonction d'activation utilisée (Ramachandran et al., 2018). $\boldsymbol{O}$ est la matrice des poids de sortie (*output layer weights*). Cette dernière est généralement initialisée aléatoirement et entraînée avec le modèle, mais, dans EWISER, $\boldsymbol{O}$ est initialisé différemment (cf. infra).

---

[80] *Enhanced WSD Integrating Synset Embeddings and Relations*.

[81] L'ensemble des sens possibles ($S(w_i)$) dépend du lemme et du POS.

[82] $swish(x) = x.sigmoïd(\beta x)$. Les auteurs justifient le choix de « swish » parce que « cette dernière a montré des résultats prometteurs en NLP » (Bevilacqua & Navigli, 2020, p. 2856. Notre traduction).

Typiquement, **Z** est transformé en une distribution de probabilité au moyen d'une fonction d'activation *softmax* standard. La prédiction du modèle correspond alors au synset $\hat{s}_i$ qui a la plus grande probabilité parmi l'ensemble de synsets possibles pour le mot w_i ($S(w_i)$).

$$\hat{s}_i = \underset{s \in S(w_i)}{\operatorname{argmax}} Z_{i,s} \qquad \text{(b)}$$

Néanmoins, nous savons qu'EWISER peut injecter dans l'architecture l'information relationnelle d'un graphe lexico-sémantique, et utiliser cette information pour l'entraînement et la prise de décision. Pour cela, EWISER fait appel au mécanisme de « *structured logits* » : au lieu des scores **Z,** une nouvelle matrice de « *logit scores* » **Q** peut être obtenue au moyen d'une matrice d'adjacence **A.** Cette matrice **A** de dimension *S x S* indique le poids des relations entre l'ensemble S de synsets. Par exemple, si $A_{s1, s2} = 0$, il n'y aucun lien entre les synsets s1 et s2. Étant donné que les relations du WordNet ne sont pas pondérées, $A_{s', s}$ est fixé à 1/N, où N est le nombre de connexions entrantes. La nouvelle matrice **Q** correspond dès lors au produit scalaire des scores cachés **Z** et de la transposée $\mathbf{A^T}$, auquel on additionne Z[83] :

$$Q = Z \cdot A^T + Z \qquad \text{(c)}$$

Finalement, la fonction *softmax* est appliquée à Q pour obtenir la distribution de probabilité, et la formule pour la prise de décision devient :

$$\hat{s}_i = \underset{s \in S(w_i)}{\operatorname{argmax}} Q_{i,s} \qquad \text{(d)}$$

Plusieurs relations peuvent être incluses dans la matrice **A** : similarité, relation dérivationnelle, hyponymie, hypéronymie, etc. Les expériences les plus concluantes sont celles qui ont inclus les relations d'hyponymie et d'hypéronymie.

Une autre caractéristique importante d'EWISER est l'utilisation de *synset embeddings* pour l'initialisation de ***O*** (matrice des poids de sortie). La justification de ce choix, ainsi que les raffinements complémentaires de la matrice ***O*** lors de l'entraînement d'EWISER dépassent toutefois le cadre de ce travail[84]. Retenons que les meilleurs résultats d'EWISER ont été observés en utilisant les *embeddings* sémantiques de SensEmBert (Scarlini et al., 2020)[85] pour l'initialisation de ***O***.

Les diverses évaluations d'EWISER ont montré que les meilleurs résultats sont obtenus en incluant uniquement les relations d'hypéronymie dans la matrice d'adjacence A, et en utilisant quatre sources d'entraînement : le corpus SemCor, les gloses de WordNet, les gloses de WordNet annotées sémantiquement, et les exemples de WordNet[86]. L'architecture sémantique ainsi entraînée peut être récupérée sous forme d'un *checkpoint* (*.pt*) sur la page Github du projet EWISER[87], et employée telle quelle. La figure 18 atteste du bon

[83] Les auteurs ne justifient pas l'addition de *Z* pour obtenir ***Q***.
[84] Nous invitons le lecteur curieux à se référer à la publication de Bevilacqua & Navigli (2020, p. 2857-2858) pour plus d'informations à ce sujet.
[85] Les *embeddings* de SensEmBert s'appuient eux-mêmes sur ceux de LMMS (Loureiro & Jorge, 2019), en les améliorant par l'exploitation de BabelNet et Wikipédia.
[86] Pour plus d'information sur les données et les hyper-paramètres d'entraînement, nous renvoyons le lecteur à Bevilacqua & Navigli (2020, p. 2858).
[87] https://github.com/SapienzaNLP/ewiser

fonctionnement local d'EWISER, puisque les résultats obtenus correspondent à ceux de l'article de Bevilacqua & Navigli (2020)[88].

```
(ewiser) david@Turing:~/ewiser$ time python bin/eval_wsd.py --checkpoints ewiser.semcor+wngt.pt --xmls ${WSD_FRAMEWORK}/Evaluat
ion_Datasets/ALL/ALL.data.xml ${WSD_FRAMEWORK}/Evaluation_Datasets/semeval2007/semeval2007.data.xml
Loading checkpoints: /home/david/ewiser/ewiser.semcor+wngt.pt
res/WSD_Evaluation_Framework/Evaluation_Datasets/ALL/ALL.data.xml
P: 0.8009099682889839   R: 0.8009099682889839   F1: 0.8009099682889839  N/T:7253/7253   Y/N/M/S: 5809/1444/0/0
res/WSD_Evaluation_Framework/Evaluation_Datasets/semeval2007/semeval2007.data.xml
P: 0.7516483516483516   R: 0.7516483516483516   F1: 0.7516483516483515  N/T:455/455     Y/N/M/S: 342/113/0/0
```

Figure 18 – duplication des résultats obtenus dans l'article d'EWISER (Bevilacqua & Navigli, 2020)

Pour la désambiguïsation du corpus d'EFLSemLex, nous avons utilisé le code du plugin *spacy* quasiment tel quel. La figure 19 montre la création du « *Desambiguator* ». Il est intéressant de noter que la longueur du contexte pris en compte est de 100 *tokens* autour du mot-cible – au maximum.

```
location = "/home/david/ewiser/corpus"                  # emplacement des fichiers à désambiguïser (.xml)
write_location = "/home/david/ewiser/corpus_desamb"     # emplacement des fichiers désambiguïsés (.txt)
checkpoint = "ewiser.semcor+wngt.pt"                    # modèle entraîné sur les 4 sources
device = "cpu"                                          # tâche effectuée sur le cpu

# création du "Disambiguator"
wsd = Disambiguator(checkpoint, maxlen=100, batch_size=5, save_wsd_details=False).eval()
wsd = wsd.to(device)
nlp = spacy.load('en', disable=['ner', 'parser'])

# ajout du Disambiguator au pipeline de traitement
# (NB : le pipeline 'nlp' inclut de base le tokenizer et le tagger)
nlp.add_pipe(wsd)

# exécution de la désambiguïsation des documents
# (dure environ une nuit sur le corpus EFL2Lex - 1.114 docs)
apply_desamb_to_corpus(location=location, new_location=write_location, wsd_model=nlp)
```

*Figure 19 - "*main process*" pour la désambiguïsation automatique du corpus*

Nous remarquons également que les étapes de *tokenization* et de *POS-tagging* sont effectuées par le *pipeline* « nlp » issu de *spacy*. Nous n'avons pas remanié ces deux étapes du traitement. Les implications de cette décision seront discutées dans la sous-section suivante, axée sur l'évaluation de notre méthode de désambiguïsation lexicale.

B.3. Évaluation quantitative et qualitative de la WSD

Pour commencer, notons que EWISER affiche un taux de désambiguïsation relativement bon : au total, 82.64% de tous les *tokens* lexicaux du corpus (noms, verbes, adjectifs et adverbes) ont été désambiguïsés ; cela correspond à 88.99% des types. Cela signifie que certains mots ont pu n'être pas traités dans certains (con)textes, mais tout de même désambiguïsés en d'autres endroits. Les mots monosémiques sont également « désambiguïsés » (en ce sens qu'ils sont, eux aussi, associés à un sens donné).

[88] Notons d'ailleurs que le résultat du *checkpoint* sur l'ensemble d'évaluation de SemEval (*ALL.data.xml*) est supérieur aux valeurs fournies par la Figure 9 (80.1% vs. 78.3%). L'historique de Bevilacqua et al. (2021) compare les systèmes récents pour le même corpus d'entraînement. Les performances du *checkpoint* utilisé ne sont donc qu'à 0.2% du meilleur système de désambiguïsation lexicale connu à ce jour.

Le tableau 4 expose les taux d'annotations par niveau de difficulté. Nous observons que l'algorithme semble avoir plus de mal pour la désambiguïsation des textes plus courts (A1, A2). Cela déséquilibre davantage la ressource EFLSemLex, puisque nous avons observé que les niveaux plus avancés représentaient déjà une majeure partie du corpus brut.

	A1	A2	B1	B2	C1	**Total**
***Tokens* désambiguïsés (%)**	74,86	78,52	81,95	84,34	84,94	**82,64**
***Types* désambiguïsés (%)**	89,25	87,18	94,68	95,09	94,88	**88,99**

Tableau 4 - pourcentage de types et de tokens du corpus annotés sémantiquement

En termes de quantité d'annotations, EWISER peut être considéré comme un algorithme efficace. À titre de comparaison, la méthode de désambiguïsation employée pour le néerlandais par Tack et al. (2018) annotait 76% des unités lexicales distinctes du corpus.

Avant l'analyse qualitative des annotations, remarquons que, le plus souvent, les mots exclus de l'annotation correspondent à des verbes très généraux, ainsi qu'aux inflexions de ceux-ci (*is, was, do, have, are, be, had, has, were, find, want, make*, etc.).

Dans l'optique d'évaluer la qualité des annotations obtenues, nous avons demandé à quatre annotateurs[89] de prendre une décision ternaire quant à la définition donnée par EWISER pour un mot (« Correct », « Erroné » ou « Indécidable »). Pour ce faire, nous avons extrait ~1% des annotations sémantiques, remis celles-ci dans leur contexte phrastique, et donné en parallèle la définition proposée par EWISER. Le nombre de mots extraits par document est proportionnel à la longueur de celui-ci. Une vérification préalable a été effectuée afin que les mots retenus soient polysémiques. Au total, 1401 mots ont été ainsi testés.

Les résultats de cette évaluation locale sont analogues à l'évaluation d'EWISER sur l'ensemble de test de SemEval (82.53% d'annotations correctes, 15.97% d'annotations incorrectes, et 1.5% de cas indécidables). L'accord inter-annotateurs peut être considéré comme modéré ($\kappa_{Fleiss} = 0.578$), avec la valeur la plus élevée pour la catégorie « Correcte » (0.612 ; 0.60 et 0.049, respectivement). Bien que ces résultats montrent que l'algorithme est efficace sur notre corpus, il paraît important de rappeler ici que le système reste imparfait, et que près de 20% des annotations du corpus d'EFLSemLex sont erronées[90].

Par ailleurs, nous avons également mentionné le choix de ne pas altérer les étapes de *tokenization* et de *POS-tagging*. Cette décision est discutable, car elle fait fi de des expressions polylexicales ainsi que des verbes à particule, dont nous savons qu'ils sont importants dans l'acquisition d'une langue seconde (Dürlich & François, 2018, p. 875). La ressource EFLSemLex ne portera par conséquent pas d'information sur les fréquences de ces derniers. Cette absence est d'autant plus dommageable que les parties de certaines de ces expressions ou verbes composés se sont parfois cristallisées, figées et – partant – ont perdu leur sens propre ; l'algorithme de désambiguïsation aura pourtant trouvé un sens pour ces parties, qui sera nécessairement erroné[91].

[89] Ces annotateurs ont le français pour langue maternelle, mais sont titulaires d'un bachelier en langue et littérature anglaise. Tous poursuivent – ou sont titulaires – d'un master dans ce même domaine.
[90] À notre échelle, une correction manuelle des entrées aurait représenté un travail trop considérable.
[91] Nous n'avons pas d'information quant à la quantité d'annotations incorrecte que la non-prise en compte des unités composées aura engendré.

C. Estimation des fréquences lexico-sémantiques

Après avoir obtenu l'ensemble des entrées lexico-sémantiques, nous avons calculé leurs fréquences à travers les cinq niveaux CEFR attestés dans le corpus. Conformément aux autres ressources de CEFRLex, les statistiques suivantes ont été calculées pour chaque entrée lexico-sémantique au sein de chaque niveau.

C.1. Fréquence brute par niveau (*FBN*)

D'abord, la *FBN* correspond au nombre de fois où une entrée sémantique apparaît dans un niveau donné. Pour cela, il suffit de faire la somme – sur toutes les sources ***I*** au sein d'un même niveau – du nombre d'occurrences du sens ***w*** (f).

$$FBN = \sum_{i=1}^{I} f_i \qquad \text{(a)}$$

Toutefois, l'utilisation de la fréquence brute n'est pas suffisante pour établir des normes de distribution lexicale (Carroll et al., 1971).

C.2. Dispersion (*D*)

Gries (2008) indique que la maîtrise du langage écrit est intimement liée à la dispersion de ce mot à travers les textes : plus un mot apparaît dans différents textes, plus il sera maîtrisé par les apprenants. Pour la compréhension du langage écrit, une dispersion élevée peut aussi être considérée comme un bon indicateur de ce qui est compris à chaque niveau de maîtrise (Tack et al., 2018).

Une mesure de dispersion s'établit par catégories. Dans notre cas, les catégories seront les différentes sources textuelles dont est issu le corpus. Mathématiquement, les unités (sens, dans notre cas) dont le *D* tend vers 0 n'apparaissent que dans quelques catégories. À l'inverse, les mots dont le *D* s'approche de 1 sont plus équitablement répartis.

Par raisonnement contraire, la mesure de dispersion permet également de sélectionner les mots qui apparaissent dans des contextes restreints. Mieux encore, la *dispersion* permet de limiter la surestimation des fréquences des sens très spécifiques (peu fréquents dans le langage courant), mais très utilisés dans un petit groupe de textes. En effet, comme le notent Francis & Kučera (1982), les mots moins fréquents tendent à être spécifiques à un contexte particulier : ils n'apparaissent que dans certains textes, mais avec des fréquences relativement hautes. D'après Dürlich & François (2018), l'utilisation d'une mesure de dispersion est d'autant plus pertinente dans le cadre d'une estimation basée sur des manuels pour apprenants. Effectivement, les thèmes utilisés dans ces derniers sont parfois relativement spécifiques[92], et il arrive que certains de ces thèmes soient même transversaux au sein d'une leçon ou d'une unité.

Nous postulons que ces observations – établies dans le domaine lexical – puissent être extrapolées au domaine sémantique. Par exemple, bien qu'en français, *servir* ou *service* correspondent rarement au sens de « *lancer/lancement une balle pour engager le jeu* », la présence de l'un ou l'autre textes spécialisés dans le tennis pourraient occasionner une

[92] Les créateurs disposent d'une grande liberté concernant les thèmes abordés, étant donné que l'objectif des manuels est d'enseigner des compétences linguistiques générales (Dürlich & François, 2018).

surévaluation de ces sens. En outre, le fait que ce sens soit moins *dispersé* pourrait également être révélateur de son caractère secondaire.

Pour ces deux raisons, les fréquences lexico-sémantiques que nous cherchons à utiliser pour mieux appréhender la compréhension d'une langue seconde devraient être ajustées pour tenir compte de la dispersion lexicale. À cet effet, nous avons utilisé – à la suite de Dürlich & François (2018) – un indice de dispersion *D* adapté de celui de Carroll et al. (1971) :

$$D_{w,k} = \frac{\log(\sum p_i) - \frac{\sum p_i \log(p_i)}{\sum p_i}}{\log(I)} \qquad \text{(b)}$$

Dans un corpus de K niveaux de difficulté, contenant chacun I sources, l'indice de dispersion $D_{w,k}$ d'un mot *w* pour un niveau *k* est calculé sur base de deux sources d'information :

- p_i : la probabilité qu'un mot w apparaisse dans la i[ème] source. Celle-ci est calculée comme suit : $p_{i=f_i/N_i}$, où N_i correspond au nombre de mots dans la i[ème] source [93] ;
- I : le nombre total de sources.

Il est intéressant de remarquer que la mesure de dispersion utilisée par Tack et al. (2018) pour NT2Lex est plus fidèle à Carroll et al. (1971) ; dans l'équation (b), f_i est préféré à p_i :

$$D_{w,k} = \frac{\ln(\sum f_i) - \frac{\sum f_i \ln(f_i)}{\sum f_i}}{\ln(I)} \qquad \text{(c)}$$

En d'autres termes, l'équation (c) répond à la question : « un [mot] se retrouve-t-il dans plusieurs textes ? », tandis que l'équation (b) répond à la question : « un [sens] a-t-il de grandes chances de se retrouver dans plusieurs textes ? »[94]. La seconde proposition diffère de la première en ce qu'elle tient compte de la longueur des sources/catégories. Dans le cas de catégories équitablement réparties, le résultat sera similaire ; ceci n'est pas vrai dans le cadre de catégories de longueurs différentes – comme on en trouve dans notre corpus :

> Some parts-based measures [among which Carroll's D_2] require the corpus parts for which a dispersion measure is computed to be identically large […] For some measures, workarounds have been proposed […] Juilland's D, for example, can be adjusted for unequal corpus sizes [by using the observed relative frequencies instead of the observed frequencies]. (Gries, 2008, p. 410-411).

Puisque nos sources sont de longueurs différentes, et que la référence dont est tirée cette citation évoque la possibilité de recourir aux valeurs de p_i plutôt qu'à celles de f_i pour calculer la dispersion, nous retiendrons l'équation (b), issue de Dürlich & François (2018), comme étant la plus adéquate[95].

---

[93] Si $p_i = 0$, $p_i \log(p_i)$ est considéré comme étant égal à 0.

[94] Les implications du choix de la fonction '*log*' (b) *vs.* '*ln*' (c) ne relèvent pas de notre sujet et ne seront pas discutées ici.

[95] En revanche, la mesure utilisée – qu'elle soit fréquentielle ou probabiliste – gagnerait à être lissée, car « le D2 de Carroll pénalise les parties du corpus ne contenant pas le [mot] *a* en question » (Gries, 2008, p. 413. Notre traduction). Une telle transformation dépasse toutefois le cadre de ce travail. Nous l'évoquons à titre informatif.

Dans le cas du sens de *service* évoqué précédemment, la présence de ce sens dans 50% des occurrences d'une source relativement courte (~100 mots) aura un impact important sur la dispersion mesurée selon (b). En revanche, cette fréquence de 50 mots, d'après la formule (c), aura un impact plus modéré sur la dispersion, si les autres sources comptent plusieurs milliers de mots.

C.3. Fréquence ajustée (*U*)

En combinant l'indice de dispersion (*D*) et les fréquences brutes par niveau (*FBN*), nous obtenons *U*, la fréquence ajustée, exprimée par million de mots :

$$U_{w,k} = \frac{1.000.000}{N_k}\left(FBN * D_{w,k} + \left(1 - D_{w,k}\right) * f_{min}\right) \qquad \text{(d)}$$

N_k est défini comme le nombre total de mots dans le niveau k, et f_{min} est défini comme :

$$f_{min} = \frac{1}{N_k}\sum_{i=1}^{I} f_i s_i \qquad \text{(e)}$$

Où s_i correspond au nombre total de mots dans la source *i*.

Les valeurs de *D* et *U* ont été calculées grâce à Python (v. 3.7). Nous avons formalisé le fonctionnement du programme sous-forme de pseudo-codes (Figure 20 & Figure 21).

```
entry = {}
# pour chaque mot w du vocabulaire total W :
    # pour chaque niveau k de la liste de niveaux K :

        # I = nombre de textes
        # sum_pi & sum_pilogpi sont calculés au moyen d'une boucle sur l'ensemble des sources I de k

        if sum_pilogpi == 0.0:
            entry[level] = 'NA'
        else:
            D = (math.log(sum_pi) - (sum_pilogpi / sum_pi)) / math.log(I)
            entry[level] = D

    # D[w][k] = entry
```

Figure 20 - pseudo-code pour le calcul de D

```
# pour chaque mot w du vocabulaire total W :
    # pour chaque niveau k de la liste de niveaux K :
        wordFreq[w][k] = # fréquence brute du mot dans le niveau k

        if wordFreq[w][k] > 0.0:
            sum_FiSi = 0
            # pour chaque source i dans k :

                Fi = # fréquence du mot dans la source i
                Si = # nombre total de mots dans la source i
                sum_FiSi += Fi * Si
            Fmin = (1 / N_k) * sumFiSi
            U = (1000000 / N_k) * (wordFreq[w][k] * D[w][k] + (1.0 - D[w][k]) * Fmin)

        else:
            U = 0.0
```

Figure 21 - pseudo-code pour le calcul de U

Enfin, il n'est pas possible de présenter les estimateurs fréquentiels sans évoquer une dernière lacune liée à la typologie des sens de WordNet. La dernière section de ce chapitre méthodologique expose la solution adoptée pour contrer ce problème.

D. Nettoyage des entrées

La distinction – ou granularité – des sens dans WordNet est trop fine. D'une part, l'abondance de sens que l'on peut y retrouver pour un même mot ne correspond pas toujours à ce que les locuteurs sont capables de percevoir. D'autre part, ces distinctions nombreuses augmentent inutilement le nombre d'entrées dans EFLSemLex, délitant ainsi l'information fréquentielle, qui – en conséquence – perd en représentativité.

Forts de ce constat, de nombreux auteurs ont cherché à regrouper les sens de WordNet afin que l'inventaire sémantique soit plus « grossier » (*coarse-grained*), et donc plus en phase avec les perceptions sémantiques réelles des locuteurs. Le recours à un inventaire sémantique plus restreint permet également d'observer des améliorations dans plusieurs tâches de NLP ayant recours à WordNet.

Par exemple, Snow et al. (2007) proposent une méthode de regroupement des sens basée sur un classifieur supervisé entraîné au départ d'une grande variété de caractéristiques de la structure de WordNet, ainsi que sur d'autres informations provenant de corpus et de ressources lexicales. Toutefois, la ressource créée par Snow et collègues concerne une version ancienne de WordNet (2.1), et nous n'avons pas pu la mettre à profit.

La méthode de regroupement (*clustering*) que nous avons utilisée est celle de Vial et al. (2019). Celle-ci fait appel aux diverses relations entres les sens dans WordNet (synonymie, hypéronymie, hyponymie, méronymie, holonymie, antonymie, instance, etc.) pour en compresser le vocabulaire.

En particulier, deux modèles de regroupement ont été proposés. Le premier s'appuie uniquement sur les relations d'inclusion (hypéronymie et hyponymie). Celui-ci permet de « remonter » dans la structure arborescente de WordNet afin d'éliminer les sens trop spécialisés, et de ne garder que les synsets suffisamment discriminatoires. Ce modèle réduit les 117.659 synsets de WordNet en 39.147 synsets essentiels. Le second rassemble l'entièreté des synsets de WordNet en 11.885 synsets en tenant compte de toutes les relations lexico-sémantiques qui s'y trouvent. La table 4 donne à voir quelques exemples des « réductions » proposées.

Deux observations importantes sont à tirer de ce tableau : premièrement, les adjectifs et les adverbes ne sont pas inclus dans les relations d'inclusion. Pour cette raison, les entrées adverbiales et adjectivales ne seront pas modifiées par le clustering par hypéronymie/hyponymie. Deuxièmement, le clustering par toutes les relations lexico-sémantiques engendre de nombreuses erreurs de transformation, relatives tant à la classe grammaticale qu'au sens. Dès lors, nous avons retenu le premier modèle de clustering et l'avons appliqué aux entrées de EFLSemLex.

Avant de passer à la description de la ressource lexico-sémantique EFLSemLex obtenue en suivant la méthodologie présentée ci-dessus, nous faisons remarquer que Vial et al. (2019) ont également développé un algorithme de désambiguïsation lexicale basé sur le système

Modèle	Lemme	POS	Sens		Nouveau POS	Nouveau sens
Inclusion (hypéronymie - hyponymie)	individuality	N	the quality of being an individual	→	N	a distinguished feature of your personal nature
	hotelier	N	an owner or manager of hotels	→	N	someone who administers a business
	clam	V	gather clams, by digging in the sand by the ocean	→	V	assemble or get together
	eviscerate	V	remove the contents of	→	V	make void or empty of contents
	exchangeable	ADJ	permetting substitution without loss of function or suitability	→	AJD	permetting substitution without loss of function or suitability
	wholly	ADV	to a complete degree or to the full extent	→	ADV	to a complete degree or to the full or entire extent
Toutes les relations lexico-sémantiques	individuality	N	the quality of being an individual	→	* ADJ	* going directly from one point to another without veering or turning aside
	hotelier	N	an owner or manager of hotels	→	N	* aquatic plant with deep green foliage useful to oxygenate an aquarium
	clam	V	gather clams, by digging in the sand by the ocean	→	V	* systematize, as by classifying and summarizing
	eviscerate	V	remove the contents of	→	V	* clear out the cinders and clinker from
	exchangeable	ADJ	permetting substitution without loss of function or suitability	→	ADJ	* lacking importance; not mattering one way or the other
	wholly	ADV	to a complete degree or to the full extent	→	* N	* the wood of the Port Orford cedar tree

*Tableau 5 – exemple de regroupements de synsets d'après Vial et al. (2019). Les * indiquent des erreurs de transformation.*

BiLSTM de Raganato, Delli Bovi et al. (2017). En utilisant cet algorithme avec l'inventaire sémantique « *coarse-grained* » obtenu par les relations d'inclusion, Vial et al. (2019) démontrent une performance de 79.0% sur l'ensemble de tâches de désambiguïsation lexicale de SemEval. Les développements ultérieurs de EFLSemLex gagneront certainement à explorer et évaluer cette méthode.

E. Expérience : EFLSemLex *vs*. EFLLex

Dans une tentative de réponse empirique à la question : « La désambiguïsation sémantique offre-t-elle une meilleure compréhension de la langue destinées aux apprenants dans le cadre d'une ressource lexicale ? », nous avons mené une expérience mêlant le corpus, la ressource EFLSemLex, et un algorithme d'identification de mots complexes (CWI).

Le recours à l'identification de mots complexes pour attester du bien-fondé de notre démarche se justifie par l'ambition même de cette dernière : développer une ressource lexicale graduée pour faire avancer les recherches quant à la complexité d'une langue seconde, et – particulier – la complexité sémantique de la L2 en question.

Pour ce faire, nous avons fait appel à l'algorithme de CWI « SEQ » développé par Gooding & Kochmar (2019). Celui-ci a été choisi parce qu'il surpasse l'état de l'art établi lors de la pénultième[96] *shared task* en CWI (Yimam et al., 2018), et s'impose à ce jour encore comme l'un des meilleurs systèmes de CWI binaire (simple / complexe). En utilisant la statistique d'évaluation de la *shared task* de 2018, SEQ atteint une précision de 71% ; le meilleur système participant à la *shared task* (CAMB – Gooding & Kochmar, 2018) obtenait une précision de 64%.

SEQ aborde l'identification de mots complexes comme une tâche d'annotation de séquences (*sequence labelling*). Contrairement à ses prédécesseurs, le système SEQ fonctionne donc *en contexte* et ne requiert pas l'analyse d'une foule de paramètres lexicaux pour capturer la complexité lexicale (POS-tag, nombre de syllabes, nombre de synsets, etc.) Ces deux

[96] Depuis 2018, une autre *shared task* en CWI a eu lieu à l'occasion de SemEval-21 (Shardlow et al., 2021). L'estimation de la complexité lexicale dans cette tâche n'était pas binaire, mais établie sur une échelle de Lickert à cinq niveaux ; cela aurait pu constituer des résultats intéressants, mais les meilleurs systèmes construits pour cette tâche ne semblent pas accessibles au public.

caractéristiques permettent de prédire correctement la complexité pour les mots polysémiques. Le système élaboré par Gooding & Kochmar (2019) repose sur une architecture neuronale (BiLSTM) et sur les *embeddings* lexicaux de GLOVE (Pennington et al., 2014)[97] pour procéder à l'identification de mots complexes. Si la probabilité en sortie du système est supérieure à un seuil (0.5), le mot est considéré comme complexe (1) ; dans le cas contraire, il est considéré comme simple (0).

Pour commencer, nous avons repris le modèle SEQ (entraîné sur les données de la *CWI shared task 2018*), et avons appliqué ce dernier au corpus d'EFLSemLex, afin d'en identifier les mots complexes. Seuls les mots pleins, ou lexicaux, ont été pris en compte.

Nous avons ensuite mis en parallèle la variable qualitative binaire obtenue (« complexité ») avec les fréquences de EFLSemLex et EFLLex. Pour chacune des ressources, nous avons ensuite procédé au test statistique U de Mann-Whitney[98] afin de vérifier si les fréquences moyennes des mots/sens simples et des mots/sens complexes, dans chacun des niveaux CEFR, sont significativement différentes (test U de Mann-Whitney). Nous avons ensuite, pour chacune des deux ressources et pour tous les niveaux CEFR, calculé la corrélation bisérielle de point (r_{pb}) entre la variable qualitative binaire (« complexité ») et la variable quantitative (« fréquence »).

Les tests précités ont été menés selon deux axes distincts. Dans la première configuration de l'expérience, ont été retenus – pour chaque niveau – tous les mots/sens lexicaux ayant une fréquence supérieure à zéro dans ce niveau. Ceci nous permettra d'avoir un aperçu général de la distribution des fréquences moyennes. Dans la seconde configuration de l'expérience, ont été retenus – pour chaque niveau – tous les mots/sens lexicaux dont la première occurrence se trouve dans ce niveau. En effet, Billami et al. (2018) démontrent que, au départ d'une ressource fréquentielle graduée, la « règle de la première occurrence » est efficace pour assigner un niveau de difficulté précis à une entrée donnée.

À partir de cette expérience, nous pouvons émettre trois hypothèses :

(1) En général (*config. 1*), les mots/sens labellisés comme « simples » devraient avoir des fréquences plus élevées que les mots/sens « complexes », puisque ces derniers seraient globalement moins courants, et ce dans chacun des niveaux du CEFR.

(2) Pour les mesures par première occurrence (*config. 2*), les mots/sens « simples » devraient avoir des fréquences plus élevées dans les niveaux élémentaires, tandis que les mots/sens « complexes » devraient avoir des fréquences plus élevées dans les niveaux avancés. Une intuition élémentaire indique en effet que les unités lexicales (et sémantiques) seront plutôt simples si elles sont spécifiques au niveau A1, plutôt complexes si elles sont spécifiques au niveau C1.

Ces deux premières hypothèses ne permettent pas d'évaluer telle quelle la pertinence de la désambiguïsation lexicale, mais des tendances divergentes pour les mots d'EFLLex et les

[97] Il est intéressant de noter que les *embeddings* de GLOVE ne sont pas désambiguïsés. Le fait que les mots polysémiques puissent être évalués différemment par SEQ est permis par la prise en compte du contexte environnant (grâce à l'architecture neuronale), plutôt que par les propriétés des *embeddings* utilisés.

[98] Ce test a été choisi parce que les échantillons ne peuvent pas être considérés comme issus d'une distribution normale (vérification effectuée par le test de Shapiro-Wilk).

sens d'EFLSemLex seront porteuses d'information sur la spécificité du recours aux sens par rapport au recours aux mots.

(3) Si les estimations fréquentielles désambiguïsées (EFLSemLex) ont un meilleur rapport de corrélation (r_{pb}) avec la variable qualitative « complexité » que les estimations fréquentielles lexicales (EFLLex), nous pourrons postuler que les fréquences désambiguïsées sont plus aux prises avec la complexité lexicale, et donc plus pertinentes pour l'apprentissage.

Ce dernier postulat permettra véritablement de comparer EFLLex et EFLSemLex en termes d'utilité pratique. Les résultats de cette expérience seront donnés et discutés dans le chapitre 3.

Chapitre 1) – À retenir :

Établir une ressource lexicale graduée requiert d'abord un corpus gradué. Pour cela, EFLSemLex a partiellement réutilisé le corpus de textes pour apprenants rassemblé dans le cadre d'EFLLex. Le corpus utilisé dans ce travail est toutefois largement inférieur en taille à ce dernier : cette longueur est loin d'être idéale pour l'établissement de fréquences sémantique fiables.

Puisqu'EFLSemLex est un ressource graduée *sémantique*, il ensuite fallu désambiguïser automatiquement le corpus. À cet effet, nous avons utilisé l'algorithme pré-entraîné "EWISER". Celui-ci, en effet, montre des performances très similaires aux meilleurs systèmes de WSD actuels. Sur notre corpus, la précision de l'algorithme a été évaluée à 82.53% par quatre annotateurs. Outre cette performance imparfaite, le principal défaut de l'algorithme EWISER est qu'il ne tient pas compte des expressions composées et des *phrasal verbs*, qui sont pourtant cruciaux pour l'apprentissage de l'anglais.

Au départ des associations forme-sens obtenues au moyen d'EWISER, les fréquences (par million de mots) de chacune de ces associations ont été calculées pour les cinq premiers niveaux de l'échelle CEFR. Les estimations fréquentielles appliquées tiennent compte de la diversité contextuelle (*dispersion*) des unités sémantiques.

Puisque les sens de WordNet sont discriminés de manière trop fine, nous avons finalement regroupés certaines entrées grâce à une ressource rassemblant les sens de WordNet selon leurs relations d'inclusions (hypéronymie et hyponymie). Ce faisant, nous avons "nettoyé" les entrées pour ne retenir, pour chaque forme, que les distinctions sémantiques fonctionnelles.

L'originalité d'EFLSemLex réside dans le caractère désambiguïsé de ses entrées. Afin d'évaluer l'utilité de l'étape de désambiguïsation – autrement dit, afin d'examiner l'intérêt d'avoir recours à une ressource sémantique plutôt que simplement lexicale – nous avons conçu une expérience pour déterminer si les fréquences d'EFLSemLex pourraient (ou non) mieux prédire la complexité des unités lexicales que celles d'EFLLex. Avant de passer en revue les aboutissements de cette expérience (Chapitre 3), nous décrivons la ressource obtenue en suivant la méthodologie décrite dans ce premier chapitre (Chapitre 2).

Chapitre 2. Description de la ressource

À la suite des différentes étapes décrites dans le chapitre précédent, nous obtenons une première version d'EFLSemLex. Ci-après, nous en offrons description plus précise. En premier lieu (A), nous détaillons les caractéristiques quantitatives de la ressource, en mettant l'accent sur les différences ou similitudes entre EFLSemLex et d'autres ressources lexicales. Ensuite (B), nous nous enquérons de la justesse des estimations fréquentielles mesurées. La description de la ressource se termine par une succincte observation de la polysémie au sein de celle-ci (C).

A. Dans la lignée du projet CEFRLex

EFLSemLex est similaire aux autres lexiques gradués développés dans le projet CEFRLex (cf. Partie I) Chapitre 2. C.2 ; Partie I) Chapitre 3. C.2. i). Nous avons vu qu'NT2Lex-CGN+ODWN (Tack et al., 2018) est la ressource la plus proche de celle que nous avons compilée. Quelques exemples d'entrées EFLSemLex sont donnés par le tableau 6. Un aperçu comparatif du nombre d'entrées dans EFLSemLex et dans autres ressources se trouve dans le tableau 7.

Lemma	Offset	*U - A1*	*U - A2*	*U - B1*	*U - B2*	*U - C1*	*U - Total*	Définiton
write	v1744611	578,59	281,14	153,8	177,62	273,95	222,23	*have (one's written work) issued for publication*
translate	v959827	0	0	0	35,52	39,14	13,51	*restate (words) from one language into another language*
sorry	a1150475	192,86	1311,97	461,42	142,1	39,14	241,33	*feeling or expressing regret or sorrow or a sense of loss over something done or undone*
diverse	s2067719	0	0	0	35,52	78,27	19,13	*many and different*
early	r100082	1542,91	187,42	51,27	213,14	78,27	134,59	*before the usual time or the time expected*
immediately	r48739	0	187,42	358,88	284,19	313,09	248,77	*without delay or hesitation; with no time intervening*
lifetime	n15140405	578,59	0	0	142,1	156,54	80,51	*the period during which something is functional (as between birth and death)*
minute	s1393483	0	0	0	0	39,14	3,22	*infinitely or immeasurably small*
minute	n15180528	192,86	562,27	153,8	106,57	156,54	171,45	*an instant of time*
note	n6624161	0	93,71	102,54	35,52	0	31,93	*a written message addressed to a person or organization*
note	n6643408	0	0	51,27	35,52	0	12,92	*an indication that makes something evident*

Tableau 6 - quelques entrées dans EFLSemLex. Une colonne « Définition » a été ajoutée à des fins d'illustration.

Pour commencer, il est notable que les entrées – contrairement aux autres ressources CEFRLex[99] – ne sont pas définies par leur lemme et leur POS, mais par leur lemme et leur *offset*[100]. Les quatre dernières lignes représentent, en quelques sortes, l'objectif de la ressource. Comme dans les autres ressources CEFRLex, un même mot est parfois assigné à une catégorie grammaticale différente (*minute*), mais la véritable originalité d'EFLSemLex réside dans ce qu'un mot puisse être associé à une même classe, et enregistrer tout de même des entrées différentes – correspondant à des sens différents (*note*). Rappelons toutefois que l'algorithme de désambiguïsation ne se défait pas de certaines erreurs d'assignation. Les fréquences ajustées *U* qui sont données pour les niveaux CEFR, ainsi que pour la totalité du corpus, sont donc à considérer sinon avec méfiance, au moins avec prudence.

La première version d'EFLSemLex compte 14662 entrées. Cela est légèrement inférieur aux ressources construites pour l'anglais (EFLLex) et le suédois (SVALex), et plus minime encore par rapport aux bases de données relatives au français (FLELex) et au néerlandais (NT2Lex). La principale raison à cette dissimilarité réside certainement dans la taille du corpus utilisé.

[99] À l'exception de la version désambiguïsée de NT2Lex (*CGN+ODWN*).
[100] L'*offset* est une clé unique pour chaque sens de WordNet. Celle-ci indique l'emplacement du synset dans le graphe et inclut le *POS-tag*.

Ressource	EFLLex				EFLSemLex				NT2Lex				SVALex		FLELEx		EVP
version	\				Base [+ regroupement par POS]				CGN+ODWN				\		CRF		
# entrées	15 281				14 662 [8 297]				17 743				15 681		17 871		-
lexicales	14 857				14 662 [8297]				16 884				15 291		17 404		-
grammaticales	424				0				400				390		467		-
polylexicales	3 852				0				459				1 450		2 038		-
niveaux	#	Nouvelles # (%)	Hapax	> 10	#	Nouvelles # (%)	Hapax	> 10	#	Nouvelles # (%)	Hapax	> 10	#	Nouvelles # (%)	#	Nouvelles # (%)	Nouvelles #
A1	2 395	2 395 (1,00)	893	509	1 726 [1 012]	1 726 (1,00) [1 012 (1,00)]	910 [399]	80 [117]	1 189	1 189 (1,00)	427	228	1 157	1 157 (1,00)	4 976	4 976 (1,00)	601
A2	4 205	2 478 (0,59)	1 633	1 000	3 381 [2 021]	2 278 (0,67) [1 287 (0,64)]	1 721 [865]	177 [232]	7 630	6 580 (0,86)	3 073	1 386	3 327	2 432 (0,73)	6 995	3 516 (0,5)	925
B1	5 607	2 740 (0,49)	2 366	1 003	5 571 [3 349]	3 212 (0,58) [1 801 (0,54)]	2 760 [1 398]	322 [376]	10 160	5 571 (0,55)	4 739	1 128	6 554	4 332 (0,66)	10 780	4 970 (0,46)	1 429
B2	8 228	3 935 (0,48)	3 580	1 571	8 067 [5 452]	4 067 (0,5) [2 886 (0,53)]	4 794 [2 572]	394 [514]	9 366	3 998 (0,43)	5 092	619	8 728	4 553 (0,52)	7 349	1 653 (0,22)	1 711
C1	9 232	3 733 (0,40)	4 254	1 591	8 531 [4 882]	3 379 (0,4) [1 311 (0,27)]	4 210 [2 094]	471 [539]	1 841	405 (0,22)	1 282	62	7 564	3 160 (0,41)	8 348	2 122 (0,25)	-
C2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	7 433	636 (0,09)	-

Tableau 7 - comparaison de EFLSemLex avec d'autres lexiques en termes de nombre (#) d'entrées (ci-inclus nouvelles entrées, hapax et entrées à la fréquence supérieure à 10), ainsi qu'en termes d'unités lexicales, grammaticales et polylexicales (MWE)

Plusieurs observations sont à faire au départ du tableau 6. Pour commencer, remarquons l'absence totale de mots grammaticaux et de *multi-word expressions* dans EFLSemLex. Ce manque est imputable, nous l'avons dit, au module EWISER utilisé pour la désambiguïsation lexicale. L'absence de *multi-word expressions* – ci-inclus les verbes à particules – est présentée par Dürlich & François (2018) comme une faiblesse notable dans la plupart des listes fréquentielles de l'anglais. Dans l'optique d'offrir une ressource plus qualitative aux apprenants, cette question devrait donc être résolue par de futurs travaux à ce sujet.

Par ailleurs, le nombre total d'entrées augmente avec l'avancée dans les niveaux, ce qui était prévisible au vu des tailles relatives des sous-corpus. Néanmoins, le nombre d'entrées n'est pas exactement proportionnel au nombre de mots par sous-corpus. Nous postulons que cette évolution est partiellement justifiable par une diversification du vocabulaire dans les niveaux plus avancés. De la même manière, si l'on tient compte du nombre de nouvelles entrées par niveau, une augmentation modérée émerge également. Dürlich & François (2018) notent que les mots de EFLLex sont rencontrés plus rapidement[101] que dans SVALex. Au contraire, les niveaux élémentaires de FLELex comptent plus de nouvelles entrées encore qu'EFLLex. L'hypothèse de Dürlich & François (2018) est que ces motifs s'expliquent par la variation dans la taille du corpus. Les entrées de EFLSemLex sont introduites moins rapidement que celles d'EFLLex ; ce résultat confirme une fois de plus qu'un corpus plus réduit entraîne une introduction plus tardive (*i.e.* dans les niveaux plus avancés) des entrées lexicales.

Il est également intéressant de comparer EFLSemLex au référentiel EVP. Comme on le voit dans le tableau ci-dessus, EVP introduit beaucoup moins de mots par niveau. Selon Dürlich & François, ce phénomène pourrait avoir deux causes. D'abord, EVP aspire à couvrir le vocabulaire de production des apprenants, qui est nécessairement plus réduit que le vocabulaire réceptif couvert par EFLSemLex. De plus, EFLSemLex inclut des sens qui appartiennent au vocabulaire central à l'apprentissage, mais aussi des sens périphériques qui, quoique rencontrés, devraient certainement être appris seulement plus tard dans l'apprentissage. En d'autres termes, à l'instar de EFLLex, EFLSemLex privilégie la couverture à la précision. Par conséquent, bien qu'EFLSemLex soit un outil pratique pour saisir les usages dans les matériaux d'apprentissage, la question de sa pertinence à des fins pédagogiques reste en suspens.

[101] C'est-à-dire que plus de mots sont introduits dès les niveaux élémentaires (A1 et A2).

B. Des estimations fréquentielles représentatives ?

Les questionnements sur l'application concrète d'EFLSemLex auprès d'un public d'apprenants sont d'autant plus vivaces que le corpus sur lequel sont estimées les fréquences est de taille relativement limitée, comme nous l'avons fait remarquer dans le Chapitre 1.A.

Cette observation se trouve reflétée dans le tableau 7 : il y a plus de sens qui n'apparaissent qu'une fois par niveau dans EFLSemLex (14.395 *hapax*) que de mots qui n'apparaissent qu'une fois par niveau dans EFLLex (12.996 *hapax*). Le problème d'informativité du corpus est plus manifeste encore avec le nombre d'unités ayant une fréquence supérieure à 10 : en moyenne, EFLLex compte 4.5x plus d'entrées répondant à ce critère qu'EFLSemLex. Pour

Outre la taille réduite du corpus, la phase de désambiguïsation, qui délaye l'information fréquentielle contenue dans les *tokens* du corpus (puisque ces derniers ont généralement plusieurs sens) est également susceptible d'engendrer la pauvreté des données observée. Afin d'investiguer les effets respectifs de ces deux caractéristiques, nous avons rassemblé les entrées d'EFLSemLex en regroupant les lemmes ayant le même POS[102]. Ce faisant, nous excluons la perte d'information imputable à la désambiguïsation. Une fois les entrées ainsi agglomérées, le nombre d'*hapax* diminue (2.02x moins d'*hapax* en moyenne), mais ceci n'est pas significatif, car le nombre d'entrées diminue lui-même également (1.77x moins d'entrées en moyenne). Par conséquent, nous pouvons affirmer que la désambiguïsation n'est pas à l'origine d'une majorité des *hapax*. De plus, le nombre d'entrées ayant une fréquence supérieure à 10 a augmenté avec le rassemblement des entrées, mais reste en moyenne 3.47x inférieur aux valeurs d'EFLLex. Ces chiffres nous permettent d'affirmer qu'entre la taille du corpus et l'étape de désambiguïsation, la première a certainement un impact négatif plus important sur la fiabilité de ressource.

Quoi qu'il en soit, il reste vrai, comme l'indiquent déjà Dürlich & François pour EFLLex, que les conclusions sur l'usage réel des entrées infréquentes d'EFLSemLex ne doivent pas être tirées hâtivement[103]. Cette précaution est d'autant plus capitale pour EFLSemLex, puisque les entrées infréquentes y sont beaucoup plus nombreuses.

Afin de mieux cerner le problème de la non-représentativité des estimations fréquentielles, Dürlich & François (2018) ont calculé la corrélation entre leurs estimations et celles de normes fréquentielles de référence pour l'anglais (BNC et MRC). Nous avons fait de même sur base de comptes fréquentiels désambiguïsés : (1) les fréquences de SEMCOR (Miller et al., 1993) et (2) les fréquences d'OMSTI (Taghipour & Ng, 2015)[104]. Ces corrélations ont été calculées uniquement pour les sens aux fréquences non-nulles dans chacun de ces deux corpus, ce qui équivaut respectivement à 8630 et 8678 entrées d'EFLSemLex.

SEMCOR est un corpus assez restreint en termes d'annotations sémantiques (~230.000), et ne couvre que 33.362 sens (22% des *synsets* de WordNet). En outre, le corpus dont est issu SEMCOR (BROWN CORPUS) est représentatif des usages de l'anglais américain des années 60.

[102] Pour rappel, le POS est donné par l'*offset* du sens.
[103] En particulier parce que les fréquences calculées ne sont pas lissées (cf. Partie II) Chapitre 1. C).
[104] Il s'agit de normes fréquentielles désambiguïsées sur les synsets de WordNet que nous avons déjà utilisées dans le cadre de notre stage au CENTAL, durant l'été 2020.

Bien que le BROWN CORPUS contienne toutes sortes de textes (articles de presse, articles scientifiques, fiction, dissertations philosophiques, etc.), l'ancienneté de ces sources pourrait partiellement justifier les faibles corrélations obtenues.

Le coefficient de corrélation de Pearson entre EFLSemLex et SEMCOR est assez faible ($r = 0.23$; $p < 0.0001$). Étant donné qu'il est largement admis que SEMCOR – malgré sa taille et sa vétusté – soit une référence dans le domaine de la sémantique, ce coefficient apparaît comme une raison supplémentaire d'aborder nos estimations fréquentielles d'un œil critique.

Le corpus OMSTI présente l'avantage d'être plus long (1M de *tokens* annotés), mais garde une couverture similaire à celle de SEMCOR. Les annotations d'OMSTI sont plus actuelles, car le corpus a été collecté entre 2000 et 2009. Ceci ne signifie pas nécessairement que lesdites annotations soient représentatives des usages sémantiques des locuteurs de l'anglais. En effet, lesdites annotations sont issues d'un corpus parallèle de documents *officiels* des Nations Unies. Gardons aussi à l'esprit que ce corpus a été obtenu grâce à des méthodes de désambiguïsation semi-automatiques précises à 83.7% seulement. La corrélation de Pearson calculée est plus faible encore pour OMSTI que pour SEMCOR ($r = 0.17$; $p < 0.0001$).

Les résultats obtenus lèvent le voile sur un réel problème de représentativité au cœur de notre corpus. D'une part, il est possible que ceci provienne du caractère pédagogique de celui-ci. Il peut paraître normal que les fréquences issues d'un corpus pédagogique ne soient que faiblement corrélées avec des corpus « généralistes ». Cette hypothèse est néanmoins mise à mal par les corrélations obtenues entre EFLLex et la liste BNC ($r = 0.97$; $p < 0.0001$) et, dans une moindre mesure, entre EFLLex et MRC ($r = 0.53$; $p < 0.0001$). À cet égard, l'importante dissimilarité entre les résultats d'EFLSemLex et EFLLex est révélatrice d'une nécessité d'amplifier le corpus pour une estimation de l'usage des sens (par rapport à l'usage des mots), comme nous le proposions dans le Chapitre 1.A.

C. La polysémie dans EFLSemLex

L'innovation principale d'EFLSemLex est son caractère désambiguïsé. C'est d'ailleurs la raison pour laquelle, au départ d'un corpus deux fois plus restreint qu'EFLLex, nous obtenons une ressource lexicale – lexico-sémantique, plus exactement – au nombre d'entrées avoisinant celui d'EFLLex. Malgré les défauts soulevés précédemment, EFLSemLex compte une large gamme de concepts : la table 7 montre que 7.034 sens distincts sont répartis sur les 14.662 entrées. Pour le dire autrement, les 14.662 entrées se « partagent » 7.034 concepts uniques ; cela implique que pour chaque sens, il y a en moyenne 2,08 lexicalisations.

	A1		A2		B1		B2		C1		**Total**	
Entrées	1.726		3.381		5.571		8.067		8.531		**14.662**	
Sens distincts	1.313		2.387		3.493		4.561		4.860		**7.034**	
Mots polysémiques (% d'entrées)	816	(0,89)	1.607	(0,88)	2.623	(0,87)	3.772	(0,86)	4.153	(0,85)	**5.953**	**(0,84)**
Mots hautement polysémiques (% d'entrées)	331	(0,47)	587	(0,45)	866	(0,42)	1.090	(0,38)	1.143	(0,36)	**1.395**	**(0,35)**
Lemme avec le + de sens	go	(20)	work	(22)	make	(21)	make	(28)	take	(29)	**make**	**(36)**

Tableau 8 - nombre d'entrées polysémiques et de sens distincts dans EFLSemLex. Le nombre de mots (hautement) polysémiques est calculé pour chaque association lemme-pos unique ; le pourcentage adjacent est calculé pour chaque association lemme-offset unique. Par exemple, pour deux entrées polysémiques (2 lemmes = ; 2 POS = ; 2 offsets ≠), le nombre de mots polysémiques serait « 1 », mais le pourcentage serait 100%.

En outre, cette table montre la distribution des mots polysémiques (> 1 sens) et hautement polysémiques (> 5 sens) à travers tous les niveaux CEFR et pour l'entièreté du corpus. Le fait qu'une majorité des entrées soient polysémiques (84%) soutient, selon nous, le postulat de la pertinence de l'approche sémantique pour l'établissement de normes fréquentielles graduées.

Nous observons par ailleurs que certains mots comptent un nombre très important d'entrées, et ce malgré l'étape de *clustering* des sens décrite dans le Chapitre 1.D. Les prolongements ultérieurs de EFLSemLex pourraient éventuellement chercher à diminuer encore le nombre d'entrées pour les mots ayant autant de significations. Néanmoins, ces cas particuliers ne devraient pas représenter la majorité de la ressource, puisque seules 35% des entrées disposent de plus de 5 sens (Tableau 8). Dans NT2Lex, à titre informatif, l'entrée la plus polysémique ne compte que dix sens (*pakken*, qui peut se traduire par « prendre »/« saisir »/« vaincre »/ *etc*.)

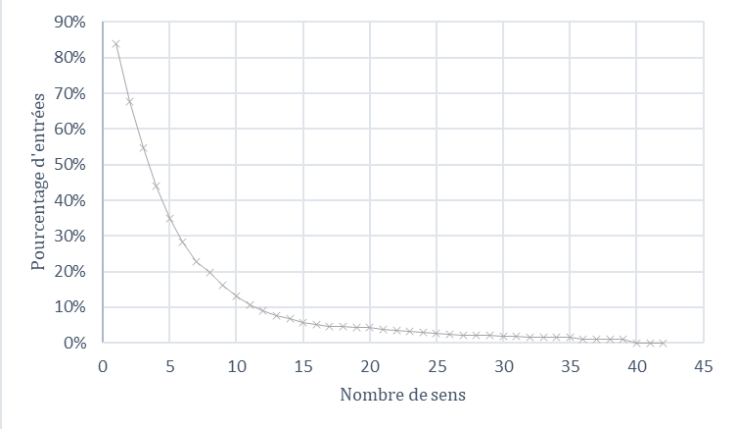


Figure 22 - pourcentage d'entrées en fonction du nombre de sens dans EFLSemLex

Pour aller plus loin dans l'analyse de la polysémie dans la ressource, la figure 22 représente le pourcentage d'entrées en fonction du nombre de sens de l'entrée. Nous ne serons pas surpris d'observer que les entrées avec peu de sens constituent une grande partie de la ressource, et que le pourcentage diminue au fur et à mesure que le nombre de sens augmente : beaucoup d'entrées ont peu de sens, très peu d'entrées ont beaucoup de sens. Au demeurant, observons que ces entrées aux sens nombreux correspondent à des mots que l'intuition linguistique perçoit immédiatement comme très généraux (*go, work, take, make, etc.*) ; cela confirme l'hypothèse de Zipf selon laquelle plus un mot est fréquent, plus il tendra à avoir de sens (Zipf, 1949).

Chapitre 2) – À retenir :

Bien que fondée sur un corpus au moins deux fois inférieur en taille, la ressource EFLSemLex – grâce à son aspect désambiguïsé – compte quasiment autant d'entrées que son homologue EFLLex.

Cette information, si elle paraît – de prime abord – de bonne augure, cache en réalité plusieurs problèmes : beaucoup de sens ne sont observés qu'une seule fois par niveau, et très peu ont une fréquence supérieure à 10. Partant, nous avons pu montrer, au moyen de corrélations statistiques avec des corpus désambiguïsés connus, que les fréquences EFLSemLex manquent globalement, en l'état, de fiabilité. Cette faillite a été identifiée comme la conséquence directe de la principale faiblesse de notre méthode : la taille du corpus utilisé est insuffisante pour l'établissement de normes lexico-sémantiques de référence.

Au surplus, les entrées d'EFLSemLex ne prennent pas en compte le vocabulaire productif des apprenants, ce qui ajourne encore la possibilité d'utiliser EFLSemLex comme ressource pédagogique.

Somme toute, notre ressource présente l'avantage de couvrir un large éventail de concepts. Des estimations fréquentielles plus en phase avec la réalité des apprenants de l'anglais permettront indubitablement à ces derniers de faire usage de la ressource EFLSemLex pour se familiariser avec un grand nombre d'associations forme-sens qu'ils seront susceptibles de rencontrer au cours de leur apprentissage.

Enfin, le fait qu'une majorité des entrées d'EFLSemLex soient polysémiques s'impose comme un argument non négligeable supplémentaire en faveur de la désambiguïsation des ressources lexicales. Dans le dernier chapitre de ce mémoire, nous chercherons d'autres arguments, d'autres preuves, tangibles, numériques, en faveur des ressources sémantiques (Chapitre 3).

Chapitre 3. EFLSemLex *vs.* EFLLex : faut-il désambiguïser ?

Durant tout ce mémoire, nous avons tenu pour acquis qu'il était forcément avantageux de mettre à dispositions des apprenants des ressources lexicales désambiguïsés. Cependant, la plupart des ressources lexicales recourent à la forme des mots (et souvent à leur catégorie grammaticale), plutôt qu'à leur sens. Malgré de nombreux arguments théoriques en faveur de la désambiguïsation, peu de ressources appliquent concrètement celle-ci. Quelles sont les raisons à ce phénomène contre-intuitif ?

Dans la Partie II) Chapitre 1. E, nous avons présenté une expérience comparant EFLSemLex et EFLLex en recourant à une analyse des mots complexes dans le corpus. Dans ce dernier chapitre, nous exposons les résultats de cette expérience (A), et discutons de la potentielle valeur ajoutée aux ressources lexicales par l'étape de désambiguïsation (B).

A. Résultats

Pour rappel, l'expérience réalisée consiste à confronter la complexité (binaire) des mots ou sens aux fréquences d'EFLSemLex et EFLLex afin de chercher à savoir si des normes fréquentielles désambiguïsées permettent – ou non – de mieux cerner le phénomène de complexité lexicale que des normes fréquentielles classiques.

Pour ce faire, nous avons comparé les fréquences moyennes – pour chacune des deux ressources, et chaque niveau CEFR – d'un échantillon de mots complexes et d'un échantillon de mots simples (U de Mann-Whitney). Nous avons ensuite cherché à voir – au moyen d'une corrélation bisérielle de point (r_{pb}) – lesquelles des fréquences désambiguïsées ou classiques sont le plus associées à la mesure de complexité, afin de désigner les plus pertinentes. Les résultats des analyses menées se retrouvent dans le tableau 9. Les statistiques ont été calculées sur deux ensembles de données différents : premièrement, pour tous les mots pleins ayant une fréquence non-nulle dans les niveaux donnés (*config. 1*) ; deuxièmement, pour les mots pleins dont la première occurrence se trouve dans le niveau donné (*config. 2*).

		EFLSemLex						EFLLEX					
	Mots considérés	A1	A2	B1	B2	C1	Total	A1	A2	B1	B2	C1	Total
Mann-Whitney's U	Tous les mots (lexicaux, fréquence ≠ 0)	57190747.5	177281151.5	356415682	540105239	520220896.5	1070232803.5	374796553.5	623253372	910452563	1107650161.5	1143219685	1334607708
$p < 0{,}0001$		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
$\lvert r_{pb} \rvert$		0.08	0.157	0.158	0.163	0.165	0.171	**0.165**	**0.214**	**0.276**	**0.273**	**0.282**	**0.287**
N° mots simples		27 943	38 132	45 593	49 544	47 768	62 893	54 430	58 605	61 814	62 782	62 665	64 477
Moyenne "simple"		1 431.72	922.98	771.91	629.24	569.69	470.287	1 270.62	1 186.43	1 063.43	952.15	873.57	975.4
N° mots complexes		3 540	6 979	11 291	16 058	16 097	24 564	8 741	12 912	17 382	20 612	21 252	24 091
Moyenne "complexe"		903.19	331.89	206	152.86	150.03	73.34	193.32	148.53	111.65	83.35	81.89	76.91
% unités complexes		11.24	15.47	19.85	24.44	25.2	28.09	13.84	18.05	21.95	24.72	25.33	27.2
Moyenne globale		1 372.29	831.54	659.58	512.83	463.91	358.88	1 121.56	999.05	854.53	737.42	673.08	731.01
Mann-Whitney's U	Tous les mots (lexicaux, fréquence ≠ 0, première occurrence dans ce niveau)	57190747.5	33303838.5	32809197	22323563.5	7760228.5	1070232803.5	374796553.5	15084550.5	8277108.5	2176792.5	277849	1334607708
$p < 0{,}0001$		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
$\lvert r_{pb} \rvert$		0.08	0.09	0.064	0.041	0.048	0.171	**0.165**	**0.096**	**0.068**	**0.124**	**0.147**	**0.287**
N° mots simples		27 943	13 659	10 468	6 879	3 944	62 893	54 430	4 916	3 012	1 865	254	64 477
Moyenne "simple"		1 431.72	246.45	121.52	67.13	58.31	470.287	1 270.62	11.42	13.36	3.44	5.67	975.41
N° mots complexes		3 540	4 426	5 877	5 589	4 132	24 564	8 741	5 203	4 934	3 465	1 748	24 091
Moyenne "complexe"		903.19	191.23	107.45	72.36	63.55	73.64	193.32	7.75	10.87	5.84	2.96	76.91
% unités complexes		11.24	24.47	35.96	48.92	51.16	28.09	13.84	51.42	62.09	65	87.31	27.2
Moyenne globale		1372.29	232.94	116.46	69.69	60.99	358.88	1 121.56	9.54	11.81	5	3.3	731.01

Tableau 9 - statistiques comparées des mots "simples" et "complexes" dans EFLSemLex et EFLLex

Le pourcentage de mots « complexes » - identifiés comme tels par l'algorithme – dans le corpus (28.09%) peut, à première vue, sembler invraisemblable. Toutefois, ce pourcentage n'est pas révélateur d'une complexité textuelle hors norme : lors de la *CWI shared task 2018*, approximativement 40% des mots ont été considérés – par des annotateurs humains – comme complexes dans les textes (articles de presse professionnels, amateurs et Wikipédia)(Gooding & Kochmar, 2018).

Tout d'abord, dans chacune des configurations et pour chacun des niveaux du CEFR, le résultat des tests U de Mann-Whitney indiquent que la fréquence moyenne des unités simples est significativement différente de la fréquence moyenne des unités complexes ($p < 0.0001$). Cette divergence prouve que la complexité et la fréquence sont liées : dans le cas contraire, les fréquences moyennes ne différeraient pas (autant). L'intensité de cette relation sera examinée ci-après.

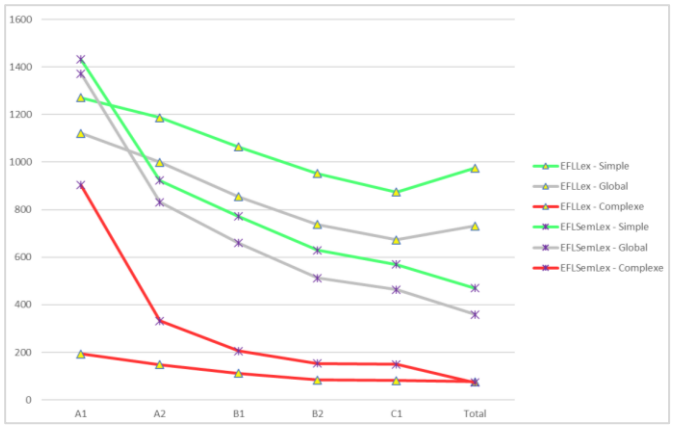


Figure 23 - fréquence des unités simples vs. complexes dans EFLSemLex vs. EFLLex (config. 1 : tous les mots pleins de fréquence ≠ 0)

Dans les deux ressources, pour la plupart des niveaux, les unités simples sont plus fréquentes que les unités complexes. Seule la configuration 2 fait exception à cette règle dans les niveaux plus avancés. Cela est rendu visible sur la figure 24.

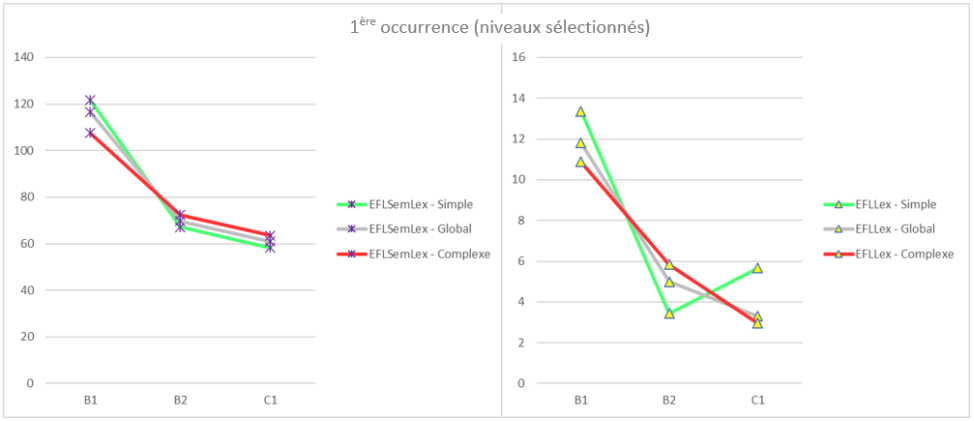


Figure 24 - les mots complexes dont la 1ère occurrence se situe dans les niveaux plus avancés ont, en moyenne, une fréquence supérieure à leurs homologues simples (B2, C1 pour EFLSemLex ; B2 pour EFLLex)

La figure 23 (*config. 1*) nous permet de vérifier visuellement la première hypothèse de cette expérience : la fréquence des unités simples est, pour tous les niveaux de chacune des ressources, nettement supérieure à celle des unités complexes. Il est intéressant de remarquer que les fréquences moyennes des unités complexes se situe bien en-dessous de la courbe des fréquences moyennes pour toutes les unités du corpus (mots ou sens). Les unités complexes sont donc généralement plus rares que les unités simples. Retenons que cette tendance, presque intuitive au niveau lexical, s'observe également sur le plan sémantique.

Ensuite, la figure 24 (*config. 2*) indique que, dans les niveaux du CEFR plus avancés, la fréquence moyenne des mots/sens complexes spécifiques à ce niveau (1ère occurrence) surpasse la fréquence moyenne des mots/sens simples. Pour EFLSemLex, cette inversion est observée dans les niveaux B2 et C1. Dans le cas d'EFLLex, le phénomène ne se manifeste qu'en B2. Il est vrai que ce renversement peut sembler infime et – partant – fortuit, mais il est à considérer de pair avec une augmentation marquante du pourcentage d'unités complexes spécifiques à chaque niveau (pour EFLSemLex, 11.24% des mots propres au niveau A1 sont complexes *vs*. 51.16% au niveau C1 ; pour EFLLex, cette augmentation est encore plus claire : 13.84% *vs*. 87.31%). Conjointement, ces deux observations attestent d'une complexification progressive du vocabulaire : il y a plus d'unités complexes dans les niveaux supérieurs, donc la fréquence de ces dernières augmente au fur et à mesure des échelons du CEFR.

Pour finir, l'objectif principal de notre expérience était de mesurer la corrélation (r_{pb}) entre les fréquences d'EFLSemLex / EFLLex (variable quantitative) et la complexité des unités du corpus (variable qualitative binaire). Pour les mots aux fréquences non-nulles dans un niveau (*config. 1*), les coefficients obtenus se situent entre 0.08 et 0.287, et attestent donc d'une faible association[105] entre les fréquences et la complexité, quels que soient la ressource et le niveau CEFR pris en compte. Pour la seconde configuration, les coefficients de corrélation sont parfois bien inférieurs à 0.1[105], indiquant que – pour certains niveaux – les fréquences des mots dont la première occurrence se trouve dans ce niveau ne sont que peu corrélées avec la complexité. Tant pour EFLSemLex que pour EFLLex, toutefois, les fréquences totales sont liées (faiblement, une fois encore) à la complexité. Ainsi, ces deux normes ne permettent que partiellement d'appréhender les mécanismes de la complexité lexicale, étant donné que les corrélations trouvées entre ces variables sont légères.

En revanche, le tableau 9 indique que les coefficients de corrélation sont plus élevés pour les fréquences issues de EFLLex que pour les fréquences issues de EFLSemLex. Pour la configuration 1, la différence moyenne entre les relations « f_{EFLLex}-complexité » et « $f_{\text{EFLSemLex}}$ – complexité » est de 0.097 (9.7%). Dans la configuration 2, l'écart entre les ressources se réduit (0.055 – 5.5%). En termes de fréquences totales, EFLLex est mieux associé (0.287) à la complexité qu'EFLSemLex (0.171). Autrement dit, les fréquences EFLLex rendent mieux compte des différences en complexité que les fréquences EFLSemLex.

Puisque les mesures de complexité sont – dans la majorité des cas de figure observés – moins corrélées avec les fréquences sémantiques qu'avec les fréquences lexicales, nous pouvons conclure que ces dernières sont plus aux prises avec le phénomène de complexité lexicale. Les implications de cette observation seront explorées dans la sous-section suivante.

[105] Une corrélation faible étant comprise entre 0.1 et 0.3 (Cohen, 1988).

B. Discussion

Dans la Partie I) Chapitre 3, nous avons évoqué plusieurs justifications théoriques – liées notamment à la (S)LA – appuyant l'ambition de désambiguïser les ressources lexicales à destination des apprenants. Tant au niveau du mot qu'au niveau de la phrase, les caractéristiques sémantiques des unités lexicales sont susceptibles d'entraver l'apprentissage.

En pratique, certains auteurs ont proposé que le recours aux sens plutôt qu'aux formes pourrait être bénéfique pour l'analyse de la complexité lexicale (Gala et al., 2013), mais ces auteurs sont peu nombreux. David Alfter (D. Alfter, communication personnelle, 11 mai 2022) indique avoir observé qu'il est difficile, pour l'heure, de démontrer que la prise en compte du sens (et du contexte) amélioreraient les systèmes d'identification de mots complexes. D'un point de vue applicatif, à ce jour, il serait donc hasardeux de certifier que la désambiguïsation permettrait de mieux comprendre la complexité lexicale.

Récemment, toutefois, certains chercheurs se sont intéressés à l'intégration de critères sémantiques dans les ressources lexicales (Capel, 2010, 2012 ; Garnier & Schmitt, 2015 ; Benigno & de Jong, 2017a, 2017b ; Tack et al. 2018 ; Billami et al., 2018), présupposant ainsi qu'une telle approche permettrait d'obtenir un meilleur aperçu de la difficulté d'une langue, et – *a fortiori* – d'en aider les apprenants. Aflter & Volodina (2019) affirment, par exemple, que ces ressources pourraient améliorer la génération d'exercices adaptés à des apprenants de différents niveaux.

En résumé, la question : « Est-il intéressant de désambiguïser les ressources lexicales ? » est encore sujette à débat dans la communauté scientifique. Pour Alfter, il serait intéressant, voire nécessaire de poursuivre les investigations à ce sujet, afin d'éventuellement atteindre un point de convergence entre la théorie – qui répond par l'affirmative – et la pratique – qui manque de preuves tangibles (D. Alfter, communication personnelle, 11 mai 2022). Dans le cadre restreint des ressources de vocabulaire à destination d'apprenants, il sera certainement à-propos, à l'avenir, d'évaluer empiriquement les progrès concrets d'un public ayant fait usage de ressources désambiguïsées.

L'élaboration et les expériences menées sur EFLSemLex ne permettent pas de statuer irrévocablement sur la précédente question. En effet, nous avons observé que les fréquences sémantiques étaient moins associées aux mesures de complexité que les fréquences lexicales. Bien que ce résultat prête vivement à reconsidérer les avantages concrets de la désambiguïsation lexicale dans le cadre de ressources pensées pour aborder la complexité des mots d'une langue, il faut ici rappeler les problèmes inhérents à la constitution d'EFLSemLex. En effet, plusieurs éléments mettent en évidence le caractère lacunaire des normes fréquentielles EFLSemLex.

Premièrement, nous avons vu que le corpus utilisé n'est pas suffisamment grand pour établir des fréquences sémantiques optimales, et en particulier que le corpus d'EFLSemLex est largement inférieur en taille au corpus d'EFLLex (cf. Chapitre 1.A). Pour cette raison, nous avons jugé périlleux de faire des comparaisons directes entre EFLSemLex et EFLLex ; si les résultats exposés précédemment remettent en cause l'utilité de la désambiguïsation

lexicale dans le cadre de notre travail, ceux-ci sont à relativiser, et ne peuvent en aucun cas être compris comme une clôture définitive du débat.

Deuxièmement, l'algorithme de désambiguïsation utilisé est imparfait (cf. Chapitre 1.B.3 – précision estimée : 82.53%), et les entrées d'EFLSemLex – contrairement à celles d'EFLLex – ne sont pas corrigées manuellement, ce qui biaise les normes fréquentielles, mais impliquerait une quantité de travail trop importante ici.

Probablement en conséquence à ces deux défectuosités, les fréquences EFLSemLex ne sont que faiblement corrélées avec d'autres normes fréquentielles générales (SEMCOR / OMSTI – cf. Chapitre 2.B). Avant de pouvoir interpréter correctement les résultats de notre comparaison entre EFLLex et EFLSemLex, il sera nécessaire de revoir la méthodologie suivie pour cette dernière ressource. En outre, remarquons que, malgré de nombreux défauts méthodologiques, les estimations fréquentielles EFLSemLex restent légèrement corrélées avec la complexité lexicale, ce qui ne peut être qu'encourageant pour les travaux à venir.

Ce premier jet d'EFLSemLex, et les résultats mitigés qui y sont associés, nous permettent également de tirer un enseignement central pour l'avenir des ressources lexico-sémantiques : puisque les mots ont généralement plusieurs sens, la désambiguïsation lexicale « dilue » l'information contenue dans les mots d'un corpus, il est donc capital de recourir à un corpus plus large afin d'obtenir des fréquences sémantiques qualitatives. Selon Alfter, il n'est ni opportun ni suffisant d'utiliser le même corpus pour les fréquences formelles et les fréquences désambiguïsées. En effet, à corpus égaux, l'étape de désambiguïsation cause indubitablement des problèmes de pauvreté des données (*data sparseness*), rendant ainsi les ressources moins fiables (D. Alfter, communication personnelle, 11 mai 2022). Une proposition avancée par Alfter pour combler cette lacune méthodologique est de considérer et d'inclure en parallèle des données provenant de ressources externes, tels que des dictionnaires, pour contrebalancer la dilution des données subie.

En l'état, néanmoins, les résultats obtenus ne nous inclinent pas à affirmer que la désambiguïsation soit une étape contre-productive dans la genèse des ressources lexicales. Apprendre et comprendre un mot ne se limite jamais à l'apprentissage et à la compréhension d'une forme : le sens – au même titre que la forme – est une composante essentielle des mots. De plus, l'impossibilité d'affirmer que tous les sens d'un mot sont acquis simultanément est communément admise : même les natifs ne connaissent pas certains sens des mots de leur langue. Par conséquent, il est réducteur de ne tenir compte que des formes dans les ressources lexicales (D. Alfter, communication personnelle, 11 mai 2022)

Pour cette raison, ainsi que celles évoquées dans le chapitre 2 de notre mémoire, nous continuons de penser – dans le prolongement d'une école qui remonte au moins à West (1953) et continue d'exister de nos jours (cf. *supra*) – que les ressources lexicales sémantiques devraient complémenter et accompagner les ressources lexicales formelles.

Chapitre 3) – À retenir :

Nous avons mené une expérience aspirant à comparer EFLSemLex et EFLLex en termes de coïncidence avec la complexité lexicale (telle que mesurée par l'un des meilleurs systèmes de CWI actuels). Cette expérience nous a fourni plusieurs résultats notables.

Pour commencer, dans EFLLex comme dans EFLSemLex, le comportement fréquentiel des unités « simples » et « complexes » nous a permis de révéler certaines similarités fonctionnelles entre les mots et les sens : il y a – communément – plus d'unités simples que d'unités complexes, mais avec la progression dans l'apprentissage, les nouvelles unités auxquelles seront confrontées les apprenants tendront à se complexifier.

Ensuite, et surtout, les rapports de corrélation r_{pb} indiquent que les fréquences EFLSemLex sont moins associées à la complexité lexicale que les fréquences EFLLex. Ce résultat, de prime abord, peut sembler peu prometteur, mais il témoigne plutôt des limites méthodologiques de notre travail (relatives en particulier à la taille insuffisante du corpus), que de l'inutilité de l'étape de désambiguïsation sémantique.

Des recherches complémentaires restent à mener dans l'optique de démontrer l'intérêt pratique de la désambiguïsation lexicale, mais les arguments théoriques issus du champ de la *(second) language acquisition*, ainsi que les ressources sémantiques existant, invalident le rejet du postulat de bénéfice de la désambiguïsation pour les ressources lexicales.

Conclusion

Dans ce travail, nous avons cherché à découvrir ce que la désambiguïsation lexicale pouvait apporter aux ressources lexicales destinées aux apprenants. Dans ce contexte, nous avons développé une ressource basée sur des textes issus de manuels pour apprenants classés selon les stades d'apprentissage définis par l'échelle CEFR. EFLSemLex s'inscrit dans la lignée des autres ressources CEFRLex, qui donnent des distributions fréquentielles des mots pour chacun des niveaux du CEFR. L'originalité de la base de données EFLSemLex réside dans le regroupement des mots du corpus par sens (au moyen d'un algorithme de désambiguïsation automatique), plutôt que par catégorie grammaticale. De cette manière, nous avons pu estimer des fréquences sémantiques, qui – en un sens – sont plus précises que les fréquences formelles, et pourraient, à l'avenir, favoriser une meilleure compréhension des mécanismes à l'origine de la complexité des langues.

En effet, s'il est bien établi que la forme des mots (nombre de lettres, nombres de syllabes, fréquence, voisins orthographiques et phonologiques…) influence la difficulté des textes, l'impact du sens (nombre de sens, polysémie, nombre de synonymes, caractère général ou spécifique du sens…) est également reconnu de longue date dans le domaine de l'apprentissage des langues étrangères.

Nous avons commencé par définir la tâche de « désambiguïsation lexicale » (WSD), qui consiste à donner à un ordinateur la capacité d'assigner aux mots d'un texte leur signification appropriée en fonction du contexte environnant (Partie I) Chapitre 1). Cette tâche est globalement méconnue du grand public, car les locuteurs d'une langue ne sont pas toujours conscients de la résolution des ambiguïtés qu'ils opèrent quasiment immédiatement en lisant ou en écoutant un texte. De plus, la WSD est souvent utilisée en profondeur et de manière invisible, pour soutenir d'autres types d'applications.

Parmi les nombreux obstacles auxquels est confrontée la désambiguïsation lexicale, nous pouvons retenir d'abord la difficulté de trouver un inventaire de sens universel. Les distinctions sémantiques de ces inventaires sont tantôt trop subtiles, ce qui affecte les performances de désambiguïsation, tantôt trop grossières, ce qui n'est pas suffisant pour certaines tâches. Par ailleurs, tous les algorithmes de désambiguïsation reposent sur des données, dont une quantité colossale est nécessaire. Ce besoin de données est longtemps demeuré bloquant pour l'évolution de la désambiguïsation lexicale, d'autant que les méthodes supervisées, généralement reconnues comme les plus efficaces, requièrent des corpus annotés sémantiquement. Ces annotations représentent une charge de travail supplémentaire non négligeable.

Nous ne repassons pas ici sur l'histoire des systèmes de désambiguïsation depuis la moitié du XX^{e} siècle. Il ne semble toutefois pas approprié d'aller plus loin sans préciser que, comme beaucoup de domaines du NLP, la désambiguïsation lexicale a connu un tournant avec la révolution neuronale, vers les années 2015. À ce jour, les méthodes de désambiguïsation à la pointe fonctionnent avec des réseaux de neurones. Ceux-ci sont capables d'intégrer de l'information de toutes sortes, et ont – logiquement – conduit à une hybridation productive des systèmes.

Les séminaires Senseval/SemEval ont permis d'établir un étalon pour l'évaluation des algorithmes de désambiguïsation lexicale. À l'heure actuelle, les performances des meilleurs systèmes (évaluées sur l'ensemble de test de SemEval) côtoient les performances humaines (80% de précision). Cela amène les chercheurs à se demander s'il n'est pas nécessaire d'établir de nouveaux critères d'évaluation, sans quoi il s'avérera certainement difficile de continuer à faire évoluer la recherche.

Dans le chapitre suivant (Partie I) Chapitre 2), nous avons fait le tour des ressources lexicales en langue anglaise. La plupart de celles-ci s'appuient sur des estimations fréquentielles en corpus afin de déterminer quel vocabulaire est plus important qu'un autre. De telles normes fréquentielles, malheureusement, sont indubitablement fonction du corpus duquel elles sont tirées. Paradoxalement, une partie considérable des ressources lexicales destinées aux enfants ou aux apprenants sont en réalité fondées sur du contenu produit par des adultes natifs, et destinés à ces derniers. Les textes utilisés ne représentent donc pas toujours la langue à laquelle seront effectivement confrontés les apprenants. Cette lacune manifeste ne se retrouve pas dans EFLSemLex : la ressource est destinée aux apprenants, et les comptes fréquentiels sont estimés au départ de manuels conçus spécialement pour ces derniers.

De plus, il a été montré que les fréquences lexicales n'expliquent qu'une partie des divergences dans la connaissance des mots. Les différentes limites des comptes de fréquence ont mené à la recherche d'autres types de normes, basées plutôt sur les perceptions des locuteurs. Ainsi, nous avons montré que des estimations de familiarité, d'âge d'acquisition, de concrétude et de prévalence permettraient d'ajouter à la compréhension du rapport des locuteurs – et des apprenants – aux mots d'une langue.

Pour l'apprentissage des langues, en particulier, l'avis de professeurs et d'experts devrait également être considéré pour déterminer les mots à connaître à chaque niveau d'apprentissage. Au demeurant, il a été souligné que la plupart des ressources lexicales pour apprenants classent les mots par niveaux de compétences. Le projet CEFRLex va plus loin : pour chaque mot, les fréquences d'occurrence sont données pour tous les niveaux du curriculum. Notre ressource, EFLSemLex, pousse davantage encore la méthodologie : les fréquences d'occurrence ne sont plus données pour les mots, mais pour les sens.

Ce choix a été justifié (Partie I) Chapitre 3) par l'évocation des entraves que les mots ambigus peuvent représenter en termes d'acquisition lexicale ou de compréhension de textes. Plusieurs travaux ont montré qu'il est difficile, pour les enfants, d'assigner de nouveaux sens aux mots déjà connus ; le même problème a pu être observé chez les apprenants adultes. Communément, le sens principal d'un mot – c'est-à-dire le sens le plus fréquent, ou le plus fréquemment observé par un locuteur – fait obstacle à l'acquisition de sens nouveaux. Il a également été montré que la résolution des ambiguïtés au moment de la lecture d'un texte pouvait être influencée par le caractère dominant (fréquent, ou perçu comme tel) d'un sens par rapport aux autres. Au regard de ces deux observations, le dessein d'EFLSemLex – dresser des distributions de sens à connaître à chaque étape d'apprentissage – paraît particulièrement pertinent.

Bien que les difficultés associées à la sémantique aient mené certains auteurs à souligner l'importance de prendre en compte les sens dans les matériaux destinés à l'apprentissage, les ressources lexicales adoptant cette approche restent relativement rares. Pour l'anglais, nous avons néanmoins été en mesure de trouver des ressources sémantiques à partir des années 1950, ce qui atteste d'une conscience de la question inscrite depuis plusieurs années dans le champ de l'apprentissage des langues. La ressource CEFRLex pour le néerlandais (Tack et al., 2018), enfin, dispose d'une version désambiguïsée qui a largement inspiré notre travail.

Dans la deuxième partie de ce mémoire, nous avons d'abord décrit en détails la méthodologie poursuivie pour la compilation d'EFLSemLex (Partie II) Chapitre 1). Premièrement, la description du corpus a permis de mettre en évidence la principale limitation de notre ressource : puisque les mots ont plusieurs sens, le corpus utilisé pour établir des fréquences de sens devrait être plus large que pour des fréquences de formes, or le corpus utilisé pour EFLSemLex ne correspond qu'à une portion du corpus collecté à l'occasion d'EFLLex (Dürlich & François, 2018). Deuxièmement, nous avons justifié le choix de l'algorithme EWISER (Bevilacqua & Navigli, 2020) pour la désambiguïsation du corpus. L'estimation locale de la performance d'EWISER a montré que l'algorithme fonctionne relativement bien sur notre corpus (82.53% de précision). Troisièmement, les formules utilisées pour les estimations fréquentielles ont été données. Enfin, nous avons montré qu'en raison d'une trop fine granularité de l'inventaire sémantique de WordNet, il a été nécessaire de rassembler les entrées obtenues. Avant le chapitre suivant, nous avons décrit l'expérience qui nous a permis de comparer EFLLex et EFLSemLex, via un algorithme d'identification des mots complexes.

Les informations essentielles de la description de la ressource (Partie II) Chapitre 2) sont rapidement résumées : les estimations fréquentielles d'EFLSemLex sont assez peu fiables. En raison d'un corpus de taille trop réduite et de l'étape de désambiguïsation – qui déconstruit l'information contenue dans les mots du corpus, puisque lesdits mots peuvent disposer de plusieurs sens –, les fréquences EFLSemLex sont assez peu corrélées avec d'autres normes fréquentielles sémantiques reconnues (SEMCOR / OMSTI). En revanche, la ressource couvre un vaste éventail de concepts (7.034 sens distincts pour 14.662 entrées EFLSemLex). En un mot, bien que sa concrétisation ait été – dans notre cas – plutôt problématique, nous ne pouvons qu'encourager à reprendre l'idée d'une ressource lexicale désambiguïsée pour l'anglais et à en retravailler la méthodologie pour aboutir à un résultat plus exploitable que cette première version d'EFLSemLex.

Pour terminer d'entériner le caractère non-exploitable de notre ressource, nous avons cherché à découvrir lesquelles – des fréquences EFLSemLex ou EFLLex – sont le mieux corrélées avec la complexité au sein du corpus utilisé (Partie II) Chapitre 3). Sans surprise, les fréquences EFLLex se sont avérées mieux adaptées – à cette fin – que celles d'EFLSemLex. Nous restons convaincus que ce résultat n'est que provisoire, et que les améliorations ultérieures apportées à EFLSemLex pourraient permettre de montrer que, sans supplanter les fréquences formelles, les fréquences sémantiques méritent de complémenter les premières.

À l'avenir, il pourrait notamment être intéressant de s'enquérir des répercussions directes de l'utilisation de ressources sémantiques auprès d'un public d'apprenants. En effet, les ressources CEFRLex, nous l'avons suffisamment dit, proposent d'aborder le phénomène de complexité lexicale en la ramenant à l'échelle CEFR. Nous nous demandons, à la suite de Billami et al. (2018), si les niveaux pédagogiques utilisés dans ces ressources établissent des distinctions suffisamment fines et précises en regard de la complexité lexicale.

Un autre prolongement potentiel d'EFLSemLex, qui n'a été que très sporadiquement évoqué dans ce mémoire, consisterait en l'adresse du vocabulaire productif des apprenants. Si EFLSemLex donne un bon aperçu de ce qui devrait théoriquement être compris à chaque stade d'apprentissage, notre ressource ne permet pas, en effet, de voir ce qui est effectivement produit par les apprenants de tel ou tel niveau.

Bibliographie

Adelman, J. S., Brown, G. D. A., & Quesada, J. F. (2006). Contextual Diversity, Not Word Frequency, Determines Word-Naming and Lexical Decision Times. *Psychological Science*, *17*(9), 814-823.

Agirre, E., & Edmonds, P. (2007). *Word Sense Disambiguation : Algorithms and Applications*. Springer Science & Business Media.

Agirre, E., López de Lacalle, O., & Soroa, A. (2014). Random Walks for Knowledge-Based Word Sense Disambiguation. *Computational Linguistics*, *40*(1), 57-84.

Agirre, E., & Martinez, D. (2001). *Learning class-to-class selectional preferences*.

Agirre, E., & Martinez, D. (2004). Smoothing and Word Sense Disambiguation. In J. L. Vicedo, P. Martínez-Barco, R. Muńoz, & M. Saiz Noeda (Éds.), *Advances in Natural Language Processing* (p. 360-371). Springer.

Agirre, E., Martínez, D., López de Lacalle, O., & Soroa, A. (2006). Two graph-based algorithms for state-of-the-art WSD. *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, 585-593.

Alfter, D. (2022, mai 11). *Discussion sur l'intérêt de la désambiguïsation pour les ressources de vocabulaire* [Communication personnelle].

Argamon-Engelson, S., & Dagan, I. (1999). Committee-Based Sample Selection for Probabilistic Classifiers. *Journal of Artificial Intelligence Research*, *11*, 335-360.

Bååth, R. (2010). *ChildFreq : An Online Tool to Explore Word Frequencies in Child Language.*

Baeza-Yates, R., Rello, L., & Dembowski, J. (2016). CASSAurus : A Resource of Simpler Spanish Synonyms. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 951-955.

Bailin, A., & Grafstein, A. (2016). Meaning in Words and Sentences. In *Readability : Text and Context*, 219. Palgrave MacMillan.

Balota, D. A., Cortese, M. J., Sergent-Marshall, S. D., Spieler, D. H., & Yap, M. (2004). Visual word recognition of single-syllable words. *Journal of Experimental Psychology. General*, *133*(2), 283-316.

Balota, D. A., Yap, M. J., Hutchison, K. A., Cortese, M. J., Kessler, B., Loftis, B., Neely, J. H., Nelson, D. L., Simpson, G. B., & Treiman, R. (2007). The English Lexicon Project. *Behavior Research Methods*, *39*(3), 445-459.

Banerjee, S., & Pedersen, T. (2002). An Adapted Lesk Algorithm for Word Sense Disambiguation Using WordNet. In A. Gelbukh (Éd.), *Computational Linguistics and Intelligent Text Processing*, 136-145.

Barba, E., Pasini, T., & Navigli, R. (2021). ESC : Redesigning WSD with Extractive Sense Comprehension. *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4661-4672.

Bar-Hillel, Y. (1960). The Present Status of Automatic Translation of Languages. In F. L. Alt (Éd.), *Advances in Computers* (Vol. 1), 91-163. Elsevier.

Bayetto, A. (2017). *Oxford Wordlist*. Oxford University Press.

Benigno, V., & de Jong, J. H. A. L. (2017a). Developing the Global Scale of English Vocabulary for Young Learners (6 to 11). *Global Scale of English Research Series*.

Benigno, V., & de Jong, J. H. A. L. (2017b). Developing the GSE Vocabulary. *Global Scale of English Research Series*.

Berend, G. (2020). Sparsity Makes Sense : Word Sense Disambiguation Using Sparse Contextualized Word Representations. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 8498-8508.

Berger, A. L., Pietra, V. J. D., & Pietra, S. A. D. (1996). A maximum entropy approach to natural language processing. *Computational Linguistics*, *22*(1), 39-71.

Bevilacqua, M., Maru, M., & Navigli, R. (2020). Generationary or "How We Went beyond Word Sense Inventories and Learned to Gloss". *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 7207-7221.

Bevilacqua, M., & Navigli, R. (2020). Breaking Through the 80% Glass Ceiling : Raising the State of the Art in Word Sense Disambiguation by Incorporating Knowledge Graph Information. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2854-2864.

Bevilacqua, M., Pasini, T., Raganato, A., & Navigli, R. (2021). Recent Trends in Word Sense Disambiguation : A Survey. *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, 4330-4338.

Biemiller, A. (2010). *Words worth teaching : Closing the vocabulary gap*. McGraw-Hill SRA.

Billami, M. B., François, T., & Gala, N. (2018). ReSyf : A French lexicon with ranked synonyms. *Proceedings of the 27th International Conference on Computational Linguistics*, 2570-2581.

Bird, H., Franklin, S., & Howard, D. (2001). Age of acquisition and imageability ratings for a large set of words, including verbs and function words. *Behavior Research Methods, Instruments, & Computers*, *33*(1), 73-79.

Black, E. (1988). An experiment in computational discrimination of English word senses. *IBM Journal of Research and Development*, *32*(2), 185-194.

Blevins, T., Joshi, M., & Zettlemoyer, L. (2021). FEWS : Large-Scale, Low-Shot Word Sense Disambiguation with the Dictionary.

Blevins, T., & Zettlemoyer, L. (2020). Moving Down the Long Tail of Word Sense Disambiguation with Gloss-Informed Biencoders.

Boada, R., Guasch, M., Haro, J., Demestre, J., & Ferré, P. (2020). SUBTLEX-CAT : Subtitle word frequencies and contextual diversity for Catalan. *Behavior Research Methods*, *52*(1), 360-375.

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching Word Vectors with Subword Information.

Bongers, H. (1947). *The History and Principles of Vocabulary Control as it Affects the Teaching of Foreign Languages in General and of English in Particular*. Wocopi.

Bordag, S. (2006). Word Sense Induction : Triplet-Based Clustering and Automatic Evaluation. *11th Conference of the European Chapter of the Association for Computational Linguistics*, 137-144.

Brody, S., & Lapata, M. (2009). Bayesian Word Sense Induction. *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, 103-111.

Brown, G. D. A., & Watson, F. L. (1987). First in, first out : Word learning age and spoken word frequency as predictors of word familiarity and word naming latency. *Memory & Cognition*, *15*(3), 208-216.

Brown, P. F., Della Pietra, S. A., Della Pietra, V. J., & Mercer, R. L. (1991). Word-Sense Disambiguation Using Statistical Methods. *29th Annual Meeting of the Association for Computational Linguistics*, 264-270.

Bruce, R., & Wiebe, J. (1994). A New Approach to Word Sense Disambiguation. *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*. HLT 1994.

Brysbaert, M., & Biemiller, A. (2017). Test-based age-of-acquisition norms for 44 thousand English word meanings. *Behavior Research Methods*, *49*(4), 1520-1523.

Brysbaert, M., Buchmeier, M., Conrad, M., Jacobs, A. M., Bölte, J., & Böhl, A. (2011). The word frequency effect : A review of recent developments and implications for the choice of frequency estimates in German. *Experimental Psychology*, *58*(5), 412-424.

Brysbaert, M., & Cortese, M. J. (2011). Do the effects of subjective frequency and age of acquisition survive better word frequency norms? *The Quarterly Journal of Experimental Psychology*, *64*(3), 545-559.

Brysbaert, M., Keuleers, E., & New, B. (2011). Assessing the Usefulness of Google Books' Word Frequencies for Psycholinguistic Research on Word Processing. *Frontiers in Psychology*, *2*.

Brysbaert, M., Lange, M., & Wijnendaele, I. V. (2000). The effects of age-of-acquisition and frequency-of-occurrence in visual word recognition : Further evidence from the Dutch language. *European Journal of Cognitive Psychology*, *12*(1), 65-85.

Brysbaert, M., Mandera, P., & Keuleers, E. (2018). The Word Frequency Effect in Word Processing : An Updated Review. *Current Directions in Psychological Science*, *27*(1), 45-50.

Brysbaert, M., Mandera, P., McCormick, S. F., & Keuleers, E. (2018). Word prevalence norms for 62,000 English lemmas. *Behavior Research Methods*, *51*(2), 467-479.

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.

Brysbaert, M., New, B., & Keuleers, E. (2012). Adding part-of-speech information to the SUBTLEX-US word frequencies. *Behavior Research Methods*, *44*(4), 991-997.

Brysbaert, M., Stevens, M., De Deyne, S., Voorspoels, W., & Storms, G. (2014). Norms of age of acquisition and concreteness for 30,000 Dutch words. *Acta Psychologica*, *150*, 80-84.

Brysbaert, M., Stevens, M., Mandera, P., & Keuleers, E. (2016). The impact of word prevalence on lexical decision times : Evidence from the Dutch Lexicon Project 2. *Journal of experimental psychology - human perception and performance*, *42*(3), 441-458.

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, *46*(3), 904-911.

Buitelaar, P. (1998). CoreLex : An Ontology of Systematic Polysemous Classes. *In Proceedings of FOIS98, International Conference on Formal Ontology in Information Systems*, 6-8.

Burgess, C., & Livesay, K. (1998). The effect of corpus size in predicting reaction time in a basic word recognition task : Moving on from Kučera and Francis. *Behavior Research Methods, Instruments, & Computers*, *30*(2), 272-277.

Cabezas, C., Bhattacharya, I., & Resnik, P. (2004). The University of Maryland Senseval-3 system descriptions. *Proceedings of SENSEVAL-3, the Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, 83-87.

Cai, Q., & Brysbaert, M. (2010). SUBTLEX-CH : Chinese Word and Character Frequencies Based on Film Subtitles. *PLoS ONE*, *5*(6).

Calabrese, A., Bevilacqua, M., & Navigli, R. (2020). *EViLBERT : Learning Task-Agnostic Multimodal Sense Embeddings*. *1*, 481-487.

Capel, A. (2010). A1–B2 vocabulary : Insights and issues arising from the English Profile Wordlists project. *English Profile Journal*, *1*(1), 1-11.

Capel, A. (2012). Completing the English Vocabulary Profile : C1 and C2 vocabulary. *English Profile Journal*, *3*, 1-14.

Carroll, J. B. (1972). A New Word Frequency Book. *Elementary English*, *49*(7), 1070-1074.

Carroll, J. B., Davies, P., & Richman, B. (1971). *The American Heritage Word Frequency Book*. Houghton Mifflin.

Catach, N., Jejcic, F., & Équipe H.E.S.O. (France). (1984). *Les listes orthographiques de base du français (LOB) : Les mots les plus fréquents et leurs formes fléchies les plus fréquentes*. F. Nathan.

Cattell, J. M. (1885). The inertia of the eye and brain. *Brain*, *8*, 295-312.

Chan, Y., & Ng, H. (2005). *Scaling Up Word Sense Disambiguation via Parallel Texts.* 1037-1042.

Choi, J. D. (2016). Dynamic Feature Induction : The Last Gist to the State-of-the-Art. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (NAACL'16)*.

Clark, J. M., & Paivio, A. (2004). Extensions of the Paivio, Yuille, and Madigan (1968) norms. *Behavior Research Methods, Instruments, & Computers*, *36*(3), 371-383.

Clear, J. H. (1993). The British National Corpus. *The Digital World*, 163-187.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.).

Coltheart, M. (1981). The MRC psycholinguistic database. *The Quarterly Journal of Experimental Psychology A: Human Experimental Psychology*, *33A*(4), 497-505.

Conia, S., & Navigli, R. (2021). Framing Word Sense Disambiguation as a Multi-Label Problem for Model-Agnostic Knowledge Integration. *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 3269-3275.

Connell, L., & Lynott, D. (2012). Strength of perceptual experience predicts word processing performance better than concreteness or imageability. *Cognition*.

Conseil de l'Europe. (2001). *Common European Framework of Reference for Languages : Learning, Teaching, Assessment*. Cambridge University Press.

Cortese, M. J., & Fugett, A. (2004). Imageability ratings for 3,000 monosyllabic words. *Behavior Research Methods, Instruments, & Computers*, *36*(3), 384-387.

Cortese, M. J., & Khanna, M. M. (2008). Age of acquisition ratings for 3,000 monosyllabic words. *Behavior Research Methods*, *40*(3), 791-794.

Cost, S., & Salzberg, S. (1993). A weighted nearest neighbor algorithm for learning with symbolic features. *Machine Learning*, *10*(1), 57-78.

Cowie, J., Guthrie, J., & Guthrie, L. (1992). Lexical Disambiguation using Simulated Annealing. *COLING 1992 Volume 1: The 14th International Conference on Computational Linguistics*. COLING 1992.

Coxhead, A., & Hirsch, D. (2007). *A pilot science-specific word list*.

Cuetos, F., Glez-Nosti, M., Barbón, A., & Brysbaert, M. (2011). SUBTLEX-ESP : Spanish word frequencies based on film subtitles. *Psicológica*, *32*(2), 133-143.

Dagan, I., & Engelson, S. P. (1995). Committee-Based Sampling For Training Probabilistic Classifiers. In A. Prieditis & S. Russell (Éds.), *Machine Learning Proceedings 1995*, 150-157. Morgan Kaufmann.

Dagan, I., Itai, A., & Schwall, U. (1991). Two Languages Are More Informative Than One. *29th Annual Meeting of the Association for Computational Linguistics*, 130-137.

Dagan, I., Marcus, S., & Markovitch, S. (1993). Contextual Word Similarity and Estimation From Sparse Data. *31st Annual Meeting of the Association for Computational Linguistics*, 164-171.

Dale, E., & Chall, J. S. (1948). A Formula for Predicting Readability. *Educational Research Bulletin*, *27*(1), 11-28.

Dale, E., & O'Rourke, J. (1981). *The Living Word Vocabulary : A National Vocabulary Inventory*. World Book-Childcraft International.

Darnell, D. K., & Howes, D. H. (1972). Review of The American Heritage Word Frequency Book [by J. B. Carroll, P. Davies, & B. Richman]. *Research in the Teaching of English*, *6*(2), 222-246. JSTOR.

Davies, M., & Gardner, D. (2010). *A frequency dictionary of contemporary American English : Word sketches, collocates, and thematic lists*. Routledge.

Davies, S. K., Izura, C., Socas, R., & Dominguez, A. (2016). Age of acquisition and imageability norms for base and morphologically complex words in English and in Spanish. *Behavior Research Methods*, *48*(1), 349-365.

Decadt, B., Hoste, V., Daelemans, W., & van den Bosch, A. (2004). GAMBL, genetic algorithm optimization of memory-based WSD. *Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, 108-112.

Degani, T., & Tokowicz, N. (2010a). Ambiguous words are harder to learn. *Bilingualism: Language and Cognition*, *13*(3), 299-314.

Degani, T., & Tokowicz, N. (2010b). Semantic Ambiguity within and across Languages : An Integrative Review. *Quarterly Journal of Experimental Psychology, 63*(7), 1266-1303.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT : Pre-training of Deep Bidirectional Transformers for Language Understanding.

Dimitropoulou, M., Duñabeitia, J. A., Avilés, A., Corral, J., & Carreiras, M. (2010). Subtitle-Based Word Frequencies as the Best Estimate of Reading Behavior : The Case of Greek. *Frontiers in Psychology, 1*.

Dubois, F., & Buyse, R. (1952). Échelle Dubois-Buyse (Originally published 1940). *Bulletin de la Société Alfred Binet, 405*.

Dürlich, L., & François, T. (2018). EFLLex : A Graded Lexical Resource for Learners of English as a Foreign Language. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. LREC 2018, Miyazaki, Japan.

Eddington, C. M., & Tokowicz, N. (2015). How meaning similarity influences ambiguous word processing : The current state of the literature. *Psychonomic Bulletin & Review, 22*(1), 13-37.

Edmonds, P., & Cotton, S. (2001). SENSEVAL-2 : Overview. *Proceedings of SENSEVAL-2 Second International Workshop on Evaluating Word Sense Disambiguation Systems*, 1-5.

Elston-Güttler, K. E., & Friederici, A. D. (2007). Ambiguous words in sentences : Brain indices for native and non-native disambiguation. *Neuroscience Letters, 414*(1), 85-89.

Erkan, G., & Radev, D. R. (2004). LexPageRank : Prestige in Multi-Document Text Summarization. *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, 365-371.

Escudero, G., Màrquez, L., & Rigau, G. (2000). Boosting Applied to Word Sense Disambiguation. In R. López de Mántaras & E. Plaza (Éds.), *Machine Learning : ECML 2000*, 129-141.

Escudero, G., Màrquez, L., & Rigau, G. (2004). TALP system for the English lexical sample task. *Proceedings of SENSEVAL-3, the Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, 113-116.

Faucett, L., Palmer, H., Thorndike, E. L., & West, M. (1936). *Interim report on vocabulary selection*. PS King and Son, Ltd.

Fellbaum, C. (1998). Towards a Representation of Idioms in WordNet. *Usage of WordNet in Natural Language Processing Systems*.

Ferrand, L., New, B., Brysbaert, M., Keuleers, E., Bonin, P., Méot, A., Augustinova, M., & Pallier, C. (2010). The French Lexicon Project : Lexical decision data for 38,840 French words and 38,840 pseudowords. *Behavior Research Methods, 42*(2), 488-496.

Florian, R., Cucerzan, S., Schafer, C., & Yarowsky, D. (2002). Combining Classifiers for word sense disambiguation. *Natural Language Engineering, 8*(4), 327-341.

Francis, W. N., & Kučera, H. (1982). *Frequency Analysis of English Usage*. Boston: Houghton Mifflin.

François, T., & De Cock, B. (2018). *ELELex : A CEFR-graded lexical resource for Spanish as a foreign language*. PLIN Day 2018, Louvain-la-Neuve.

François, T., Gala, N., Watrin, P., & Fairon, C. (2014). *FLELex : A graded lexical resource for French foreign learners*. International conference on Language Resources and Evaluation (LREC 2014).

François, T., Volodina, E., Pilán, I., & Tack, A. (2016). SVALex : A CEFR-graded Lexical Resource for Swedish Foreign and Second Language Learners. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, 213-219.

Freund, Y., & Schapire, R. E. (1997). A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *Journal of Computer and System Sciences*, *55*(1), 119-139.

Friendly, M., Franklin, P. E., Hoffman, D., & Rubin, D. C. (1982). The Toronto Word Pool : Norms for imagery, concreteness, orthographic variables, and grammatical usage for 1,080 words. *Behavior Research Methods & Instrumentation*, *14*(4), 375-399.

Gala, N., François, T., & Fairon, C. (2013). Towards a French lexicon with difficulty measures : NLP helping to bridge the gap between traditional dictionaries and specialized lexicons. *Proceedings of eLex 2013*.

Gale, W. A., & Church, K. W. (1993). A program for aligning sentences in bilingual corpora. *Computational Linguistics*, *19*(1), 75-102.

Gale, W. A., Church, K. W., & Yarowsky, D. (1992a). Using bilingual materials to develop word sense disambiguation methods. *In Proceedings of the International Conference on Theoretical and Methodological Issues in Machine Translation*, 101-112.

Gale, W. A., Church, K. W., & Yarowsky, D. (1992b). A method for disambiguating word senses in a large corpus. *Computers and the Humanities*, *26*(5), 415-439.

Galley, M., & McKeown, K. (2003). *Improving Word Sense Disambiguation in Lexical Chaining*. 3.

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.

Gernsbacher, M. A. (1984). Resolving 20 years of inconsistent interactions between lexical familiarity and orthography, concreteness, and polysemy. *Journal of Experimental Psychology. General*, *113*(2), 256-281.

Gilhooly, K. J., & Logie, R. H. (1980a). Age-of-acquisition, imagery, concreteness, familiarity, and ambiguity measures for 1,944 words. *Behavior Research Methods & Instrumentation*, *12*(4), 395-427.

Gilhooly, K. J., & Logie, R. H. (1980b). Meaning-dependent ratings of imagery, age of acquisition, familiarity, and concreteness for 387 ambiguous words. *Behavior Research Methods & Instrumentation*, *12*(4), 428-450.

Gooding, S., & Kochmar, E. (2018). CAMB at CWI Shared Task 2018 : Complex Word Identification with Ensemble-Based Voting. *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, 184-194.

Gooding, S., & Kochmar, E. (2019). Complex Word Identification as a Sequence Labelling Task. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 1148-1153.

Goodman, J. C., Dale, P. S., & Li, P. (2008). Does frequency count? Parental input and the acquisition of vocabulary. *Journal of Child Language*, *35*(3), 515-531.

Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-Metrix : Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers*, *36*(2), 193-202.

Graham, S., Harris, K., & Loynachan, C. (1993). *The basic spelling vocabulary list*.

Greene, B. B., & Rubin, G. M. (1971). *Automatic Grammatical Tagging of English*. Department of Linguistics, Brown University.

Gries, S. (2008). Dispersions and adjusted frequencies in corpora. *International Journal of Corpus Linguistics*, *13*, 403-437.

Groebel, L. (1985). Ambiguity and Second Language Learning. *IRAL : International Review of Applied Linguistics in Language Teaching*, *23*(2), 149-158.

Guthrie, J. A., Guthrie, L., Aidinejad, H., & Wilks, Y. (1991). Subject-Dependent Co-Occurrence and Word Sense Disambiguation. *29th Annual Meeting of the Association for Computational Linguistics*, 146-152.

Hadiwinoto, C., Ng, H. T., & Gan, W. C. (2019). Improved Word Sense Disambiguation Using Pre-Trained Contextualized Word Representations.

Halliday, M. A. K., & Hasan, R. (1976). *Cohesion in English*. Pearson.

Harabagiu, S. M. (1999). From lexical cohesion to textual coherence : A data driven perspective. *International Journal of Pattern Recognition and Artificial Intelligence*, *13*(02), 247-265.

Henderson, L., Snowling, M., & Clarke, P. (2013). Accessing, integrating, and inhibiting word meaning in poor comprehenders. *Scientific Studies of Reading*, *17*(3), 177-198.

Hirst, G., & St-Onge, D. (1998). Lexical chains as representations of context for the detection and correction of malapropisms. In *WordNet : An Electronic Lexical Database*. MIT Press.

Hofland, K., & Johansson, S. (1982). *Word frequencies in British and American English*. Norwegian computing centre for the humanities.

Honeyfield, J. (1977). Word Frequency and the Importance of Context in Vocabulary Learning. *RELC Journal*, *8*(2), 35-42.

Honnibal, M., & Montani, I. (2017). *spaCy 2 : Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing*.

Hornby, A. S., Cowie, A. P., & Lewis, J. W. (1974). *Oxford advanced learner's dictionary of current English*. Oxford University Press.

Hoste, V., Hendrickx, I., Daelemans, W., & Van den Bosch, A. (2002). Parameter Optimization for Machine-Learning of Word Sense Disambiguation. *Natural Language Engineering*, *8*.

Howes, D. H., & Solomon, R. L. (1951). Visual Duration Threshold as a Function of Word-Probability. *Journal of Experimental Psychology*, *41*(6), 401.

Huang, L., Sun, C., Qiu, X., & Huang, X. (2019). GlossBERT : BERT for Word Sense Disambiguation with Gloss Knowledge.

Iacobacci, I., Pilehvar, M. T., & Navigli, R. (2016). Embeddings for Word Sense Disambiguation : An Evaluation Study. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 897-907.

Ide, N., Erjavec, T., & Tufis, D. (2001). Automatic sense tagging using parallel corpora. *Proceedings of the 6th Natural Language Processing Pacific Rim Symposium*, 83-89.

Ide, N., & Véronis, J. (1998). Introduction to the special issue on word sense disambiguation : The state of the art. *Computational Linguistics*, *24*(1), 2-40.

Ishida, T. (2019). The Effects of Meaning Dominance in the Time-Course of Activation of L2 Lexical Ambiguity Processing. *Journal of Psycholinguistic Research*, *48*(6), 1269-1284.

James, G., Davison, R., Cheung, A. H. Y., & Deerwester, S. (Éds.). (1994). *English in computer science : A corpus-based lexical analysis*. Longman, for the Language Centre, Hong Kong University of Science and Technology.

Jiang, J. J., & Conrath, D. W. (1997). Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy.

Johansson, S., & Hofland, K. (1989). *Frequency Analysis of English Vocabulary and Grammar : Tag frequencies and word frequencies*. Clarendon Press (2 volumes).

Jones, M., & Durrant, P. (2010). What can a corpus tell us about vocabulary teaching materials? In A. O'Keeffe & M. McCarthy (Éds.), *The Routledge handbook of corpus linguistics* (p. 387-400). Routledge.

Jørgensen, R. N., Dale, P. S., Bleses, D., & Fenson, L. (2009). CLEX : A cross-linguistic lexical norms database. *Journal of Child Language*, *37*(2), 419-428.

Juhasz, B. J., Lai, Y.-H., & Woodcock, M. L. (2015). A database of 629 English compound words : Ratings of familiarity, lexeme meaning dominance, semantic transparency, age of acquisition, imageability, and sensory experience. *Behavior Research Methods*, *47*(4), 1004-1019.

Jurafsky, D., & Martin, J. (2021). Word Senses and WordNet. In *Speech and Language Processing : An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition* (draft). Prentice Hall.

Kaplan, A. (1955). An experimental study of ambiguity and context. *Mech. Transl. Comput. Linguistics*.

Kelly, E., & Stone, P. J. (1975). *Computer recognition of English word senses* (vol. 13). North-Holland.

Keuleers, E., Brysbaert, M., & New, B. (2010). SUBTLEX-NL : A new measure for Dutch word frequency based on film subtitles. *Behavior Research Methods*, *42*(3), 643-650.

Keuleers, E., Stevens, M., Mandera, P., & Brysbaert, M. (2015). Word knowledge in the crowd : Measuring vocabulary size and word prevalence in a massive online experiment. *Quarterly Journal of Experimental Psychology*, *68*(8), 1665-1692.

Khanna, M. M., & Cortese, M. J. (2011). Age of acquisition estimates for 1,208 ambiguous and polysemous words. *Behavior Research Methods*, *43*(1), 89-96.

Kilgarriff, A., & Palmer, M. (2000). Introduction to the Special Issue on SENSEVAL. *Computers and the Humanities*, *34*(1), 1-13.

Klein, D., Toutanova, K., Ilhan, H. T., Kamvar, S. D., & Manning, C. D. (2002). Combining Heterogeneous Classifiers for Word Sense Disambiguation. *Proceedings of the ACL-02 Workshop on Word Sense Disambiguation: Recent Successes and Future Directions*, 74-80.

Kousta, S.-T., Vigliocco, G., Vinson, D. P., Andrews, M., & Del Campo, E. (2011). The representation of abstract words : Why emotion matters. *Journal of Experimental Psychology: General*, *140*(1), 14-34.

Krovetz, R. (2000). *More than One Sense Per Discourse*.

Kučera, H., & Francis, W. N. (1964). *A Standard Corpus of Present-Day Edited American English, for use with Digital Computers (Brown Corpus)*. Department of Linguistics, Brown University.

Kučera, H., & Francis, W. N. (1967). *Computational Analysis of Present Day American English* (1st edition). Brown University Press.

Kumar, S., Jat, S., Saxena, K., & Talukdar, P. (2019). Zero-shot Word Sense Disambiguation using Sense Definition Embeddings. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 5670-5681.

Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods*, *44*(4), 978-990.

Kwary, D. A., & Jurianto. (2017). Selecting and creatinf a word list for English language teaching. *Teaching English with Technology*, *17*(1), 60-72.

Lafourcade, M. (2007). *Making people play for Lexical Acquisition with the JeuxDeMots prototype*.

Laufer, B., & Ravenhorst-Kalovski, G. C. (2010). *Lexical threshold revisited : Lexical text coverage, learners' vocabulary size and reading comprehension*.

Leacock, C., & Chodorow, M. (1998). Combining local context and WordNet similarity for word sense identification. In *WordNet : An Electronic Lexical Database*. MIT Press.

Leacock, C., Chodorow, M., & Miller, G. A. (1998). Using Corpus Statistics and WordNet Relations for Sense Identification. *Computational Linguistics*, *24*(1), 147-165.

Lee, Y. K., & Ng, H. T. (2002). An Empirical Evaluation of Knowledge Sources and Learning Algorithms for Word Sense Disambiguation. *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, 41-48.

Leech, G., Rayson, P., & Wilson, A. (2001). *Word Frequencies in Written and Spoken English : Based on the British National Corpus*.

Lesk, M. (1986). Automatic sense disambiguation using machine readable dictionaries : How to tell a pine cone from an ice cream cone. *Proceedings of the 5th annual international conference on Systems documentation*, 24-26.

Lété, B., Sprenger-Charolles, L., & Colé, P. (2004). MANULEX : A grade-level lexical database from French elementary school readers. *Behavior Research Methods, Instruments, & Computers*, *36*(1), 156-166.

Lin, D. (1998). Extracting Collocations from Text Corpora. *In First Workshop on Computational Terminology*, 57-63.

Lin, D., & Pantel, P. (2002). Concept Discovery from Text. *COLING 2002: The 19th International Conference on Computational Linguistics*. COLING 2002.

Lorge, I., & Thorndike, E. L. (1938). *A semantic count of English words*. Institute of educational research, Teachers college, Columbia university.

Loureiro, D., & Camacho-Collados, J. (2020). Don't Neglect the Obvious : On the Role of Unambiguous Words in Word Sense Disambiguation.

Loureiro, D., & Jorge, A. (2019). Language Modelling Makes Sense : Propagating Representations through WordNet for Full-Coverage Word Sense Disambiguation.

Love, T., Maas, E., & Swinney, D. (2003). The influence of language exposure on lexical and syntactic language processing. *Experimental Psychology, 50*(3), 204-216.

Lovejoy, S. (2003). *Children's printed word database : Manual & documentation*.

Luan, Y., Hauer, B., Mou, L., & Kondrak, G. (2020). Improving Word Sense Disambiguation with Translations. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 4055-4065.

Lynott, D., & Connell, L. (2009). Modality exclusivity norms for 423 object properties. *Behavior Research Methods, 41*(2), 558-564.

Lynott, D., & Connell, L. (2013). Modality exclusivity norms for 400 nouns : The relationship between perceptual experience and surface word form. *Behavior Research Methods, 45*(2), 516-526.

MacWhinney, B. (2000). *The Childes Project : Tools for Analyzing Talk* (3rd edition). Psychology Press.

Mallery, J. C. (1988). Thinking About Foreign Policy : Finding an Appropriate Role for Artificially Intelligent Computers. *Master's thesis, M.I.T. Political Science Department*.

Mandera, P., Keuleers, E., Wodniecka, Z., & Brysbaert, M. (2015). Subtlex-pl : Subtitle-based word frequency estimates for Polish. *Behavior Research Methods, 47*(2), 471-483.

Manelis, L. (1972). *« The American Heritage Word Frequency Book » and Its Relation to the Communication Skills Lexicon. Technical Note No. 2-72-38*.

Marello, C. (2012). *Word lists in Reference Level Descriptions of CEFR (Common European Framework of Reference for Languages)*. 328-335.

Martinez, D., & Agirre, E. (2000). One Sense per Collocation and Genre/Topic Variations. *2000 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, 207-215.

Maru, M., Scozzafava, F., Martelli, F., & Navigli, R. (2019). SyntagNet : Challenging Supervised Word Sense Disambiguation with Lexical-Semantic Combinations. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3534-3540.

Masterman, M. (1957). The thesaurus in syntax and semantics. *Mechanical Translation, 4*(1-2), 35-43.

Masterson, J., Stuart, M., Dixon, M., & Lovejoy, S. (2010). Children's printed word database : Continuities and changes over time in children's early reading vocabulary. *British Journal of Psychology*, *101*(2), 221-242.

McCarthy, D., & Carroll, J. (2003). Disambiguating Nouns, Verbs, and Adjectives Using Automatically Acquired Selectional Preferences. *Computational Linguistics*, *29*(4), 639-654.

McCarthy, D., Koeling, R., Weeds, J., & Carroll, J. (2004). Using automatically acquired predominant senses for Word Sense Disambiguation. *Proceedings of SENSEVAL-3, the Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, 151-154.

McCrae, J. P., Rademaker, A., Rudnicka, E., & Bond, F. (2020). English WordNet 2020 : Improving and Extending a WordNet for English using an Open-Source Methodology. *Proceedings of the LREC 2020 Workshop on Multimodal Wordnets (MMW2020)*, 14-19.

Melamud, O., Goldberger, J., & Dagan, I. (2016). context2vec : Learning Generic Context Embedding with Bidirectional LSTM. *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, 51-61.

MetaMetrics, Inc. (2003). *MetaMetrics word frequency counts [Database]*. Unvavailable._

Mihalcea, R., & Faruque, E. (2004). SenseLearner : Minimally supervised Word Sense Disambiguation for all words in open text. *Proceedings of SENSEVAL-3, the Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, 155-158.

Mihalcea, R., & Moldovan, D. I. (1999). A Method for Word Sense Disambiguation of Unrestricted Text. *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, 152-158.

Mihalcea, R., Tarau, P., & Figa, E. (2004). PageRank on Semantic Networks, with Application to Word Sense Disambiguation. *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, 1126-1132.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. *Advances in Neural Information Processing Systems*, *26*.

Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., & Miller, K. J. (1990). Introduction to WordNet : An On-line Lexical Database. *International Journal of Lexicography*, *3*(4), 235-244.

Miller, G. A., Leacock, C., Tengi, R., & Bunker, R. T. (1993). A Semantic Concordance. *Human Language Technology: Proceedings of a Workshop Held at Plainsboro, New Jersey, March 21-24, 1993*. HLT 1993.

Monsell, S. (1991). The nature and locus of word frequency effects in reading. *Basic Processes in Reading: Visual Word Recognition. Lawrence Erlbaum Associates Inc., Hillsdale, NJ.*, 148-197.

Montoyo, A., Suarez, A., Rigau, G., & Palomar, M. (2005). Combining Knowledge- and Corpus-based Word-Sense-Disambiguation Methods. *Journal of Artificial Intelligence Research*, *23*, 299-330.

Mooney, R. J. (1996). Comparative Experiments on Disambiguating Word Senses : An Illustration of the Role of Bias in Machine Learning.

Moro, A., & Navigli, R. (2015). SemEval-2015 Task 13 : Multilingual All-Words Sense Disambiguation and Entity Linking. *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, 288-297.

Moro, A., Raganato, A., & Navigli, R. (2014). Entity Linking meets Word Sense Disambiguation : A Unified Approach. *Transactions of the Association for Computational Linguistics*, *2*, 231-244.

Morrison, C. M., Chappell, T. D., & Ellis, A. W. (1997). Age of acquisition norms for a large set of object names and their relation to adult estimates and other variables. *The Quarterly Journal of Experimental Psychology A: Human Experimental Psychology*, *50A*(3), 528-559.

Murata, M., Utiyama, M., Uchimoto, K., Ma, Q., & Isahara, H. (2001). Japanese Word Sense Disambiguation using the Simple Bayes and Support Vector Machine Methods. *Proceedings of SENSEVAL-2 Second International Workshop on Evaluating Word Sense Disambiguation Systems*, 135-138.

Nation, P. (1990). *Teaching & Learning Vocabulary* (1st edition). Heinle ELT.

Nation, P., & Chung, T. (2009). Teaching and Testing Vocabulary. *The Handbook of Language Teaching*, 543-559. John Wiley & Sons, Ltd.

Nation, P., & Waring, R. (1997). Vocabulary Size, Text Coverage and Word Lists. Schmitt, N. & McCarthy, M. (Éds.). *Vocabulary: Description, Acquisition, and Pedagogy, Cambridge University Press, Cambridge*, 6-19.

Navigli, R. (2009). Word Sense Disambiguation : A Survey. *ACM Comput. Surv.*, *41*.

Navigli, R., Bevilacqua, M., Conia, S., Montagnini, D., & Cecconi, F. (2021). Ten Years of BabelNet : A Survey. *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, 4559-4567.

Navigli, R., & Crisafulli, G. (2010). Inducing Word Senses to Improve Web Search Result Clustering. *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, 116-126.

Navigli, R., Jurgens, D., & Vannella, D. (2013). SemEval-2013 Task 12 : Multilingual Word Sense Disambiguation. *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, 222-231.

Navigli, R., & Ponzetto, S. P. (2012). BabelNet : The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, *193*, 217-250.

New, B., Brysbaert, M., Veronis, J., & Pallier, C. (2007). The use of film subtitles to estimate word frequencies. *Applied psycholinguistics*, *28*(4), 661-677.

Ng, H. T., & Lee, H. B. (1996). Integrating Multiple Knowledge Sources to Disambiguate Word Sense : An Exemplar-Based Approach. *34th Annual Meeting of the Association for Computational Linguistics*, 40-47.

Ng, H. T., Wang, B., & Chan, Y. S. (2003). Exploiting Parallel Texts for Word Sense Disambiguation : An Empirical Study. *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, 455-462.

North, B. (2005). The CEFR levels and descriptor scales. In Association of Language Testers in Europe (Éd.), *Multilingualism and assessment : Achieving transparency, assuring quality, sustaining diversity : Proceedings of the ALTE Berlin Conference,* 21-66. Cambridge University Press.

Ogden, C. K., & Richards, I. A. (1923). *The Meaning of Meaning. A study of the influence of language upon thought and the science of symbolism*. Harcourt, Brace & World, Inc.

Okumura, M., & Honda, T. (1994). Word Sense Disambiguation and Text Segmentation Based on Lexical Cohesion. *COLING 1994 Volume 2: The 15th International Conference on Computational Linguistics*. COLING 1994.

Paetzold, G., & Specia, L. (2016). Inferring Psycholinguistic Properties of Words. *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 435-440.

Paivio, A. (1971). *Imagery and Verbal Processes*. Holt, Rinehart and Winston.

Paivio, A. (2013). Dual coding theory, word abstractness, and emotion : A critical review of Kousta et al. (2011). *Journal of Experimental Psychology: General*, *142*(1), 282-287.

Paivio, A., Yuille, J. C., & Madigan, S. A. (1968). Concreteness, imagery, and meaningfulness values for 925 nouns. *Journal of Experimental Psychology*, *76*(1), 1-25.

Pasini, T., & Navigli, R. (2019). Train-O-Matic : Supervised Word Sense Disambiguation with No (manual) effort. *Artificial Intelligence*, *279*.

Patrick, A. B. (1985). *An exploration of an abstract thesaurus instantiation*. Univeristy of Kansas, Computer science.

Paul, D. B., & Baker, J. M. (1992). The Design for the Wall Street Journal-based CSR Corpus. *Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992*. HLT 1992.

Pedersen, T., & Bruce, R. (1997). Distinguishing Word Senses in Untagged Text. *Second Conference on Empirical Methods in Natural Language Processing*.

Pennington, J., Socher, R., & Manning, C. (2014). GloVe : Global Vectors for Word Representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532-1543.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep Contextualized Word Representations. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 2227-2237.

Pradhan, S., Loper, E., Dligach, D., & Palmer, M. (2007). SemEval-2007 Task-17 : English Lexical Sample, SRL and All Words. *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, 87-92.

Preston, K. A. (1935). The speed of word perception and its relation to reading ability. *Journal of General Psychology*, *13*, 199-203.

Procter, P. (1978). *Longman dictionary of contemporary English*. Longman.

Purandare, A., & Pedersen, T. (2004). Improving Word Sense Discrimination with Gloss Augmented Feature Vectors. *Proceedings of the Workshop on Lexical Resources for the Web and Word Sense Disambiguation*, 123-130.

Pustejovsky, J. (1995). *The Generative Lexicon*. MIT Press.

Qian, D. (1999). Assessing the Roles of Depth and Breadth of Vocabulary Knowledge in Reading Comprehension. *The Canadian Modern Language Review*, *56*(2), 282-308.

Raganato, A., Camacho-Collados, J., & Navigli, R. (2017). Word Sense Disambiguation : A Unified Evaluation Framework and Empirical Comparison. *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, 99-110.

Raganato, A., Delli Bovi, C., & Navigli, R. (2017). Neural Sequence Learning Models for Word Sense Disambiguation. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 1156-1167.

Ramachandran, P., Zoph, B., & Le, Q. V. (2018). *Searching for Activation Functions*.

Resnik, P. (1993). Semantic Classes and Syntactic Ambiguity. *Human Language Technology: Proceedings of a Workshop Held at Plainsboro, New Jersey, March 21-24, 1993*. HLT 1993.

Resnik, P. (1995). Using Information Content to Evaluate Semantic Similarity in a Taxonomy.

Resnik, P., & Yarowsky, D. (1997). A Perspective on Word Sense Disambiguation Methods and Their Evaluation. *Tagging Text with Lexical Semantics: Why, What, and How?*

Richards, J. C. (1976). The Role of Vocabulary Teaching. *TESOL Quarterly*, *10*(1), 77-90.

Rinsland, H. D. (1945). *A basic vocabulary of elementary school children*. Macmillan.

Rodd, J. (2018). Lexical ambiguity. In *The Oxford Handbook of Psycholinguistics*. Oxford University Press.

Rodd, J., Gaskell, G., & Marslen-Wilson, W. (2002). Making Sense of Semantic Ambiguity : Semantic Competition in Lexical Access. *Journal of Memory and Language*, *46*(2), 245-266.

Roget, P. M. (1911). *The New Thesaurus of English Words and Phrases Classified and Arranged So as to Facilitate the Expression of Ideas and Assist in Literary Composition.* Current Literature Publishing Company.

Scarlini, B., Pasini, T., & Navigli, R. (2020a). SensEmBERT : Context-Enhanced Sense Embeddings for Multilingual Word Sense Disambiguation. *Proceedings of the AAAI Conference on Artificial Intelligence*, *34*(05), 8758-8765.

Scarlini, B., Pasini, T., & Navigli, R. (2020b). With More Contexts Comes Better Performance : Contextualized Sense Embeddings for All-Round Word Sense Disambiguation. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 3528-3539.

Schmid, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. *Proceedings of International Conference on New Methods in Language Processing*.

Schock, J., Cortese, M. J., & Khanna, M. M. (2012). Imageability estimates for 3,000 disyllabic words. *Behavior Research Methods*, *44*(2), 374-379.

Schock, J., Cortese, M. J., Khanna, M. M., & Toppi, S. (2012). Age of acquisition estimates for 3,000 disyllabic words. *Behavior Research Methods*, *44*(4), 971-977.

Schütze, H. (1993). Word Space. *Advances in Neural Information Processing Systems*, *5*.

Schütze, H. (1998). Automatic Word Sense Discrimination. *Computational Linguistics*, *24*(1), 97-123.

Scott, G. G., Keitel, A., Becirspahic, M., Yao, B., & Sereno, S. C. (2018). The Glasgow Norms : Ratings of 5,500 words on nine scales. *Behavior Research Methods*, *51*(3), 1258-1270.

Scozzafava, F., Maru, M., Brignone, F., Torrisi, G., & Navigli, R. (2020). Personalized PageRank with Syntagmatic Information for Multilingual Word Sense Disambiguation. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 37-46.

Shardlow, M., Evans, R., Paetzold, G. H., & Zampieri, M. (2021). SemEval-2021 Task 1 : Lexical Complexity Prediction. *SEMEVAL*.

Snow, R., Prakash, S., Jurafsky, D., & Ng, A. Y. (2007). Learning to Merge Word Senses. *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, 1005-1014.

Snyder, B., & Palmer, M. (2004). The English all-words task. *Proceedings of SENSEVAL-3, the Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, 41-43.

Spreen, O., & Schulz, R. W. (1966). Parameters of abstraction, meaningfulness, and pronunciability for 329 nouns. *Journal of Verbal Learning & Verbal Behavior*, *5*(5), 459-468.

Stadthagen-Gonzalez, H., & Davis, C. J. (2006). The Bristol norms for age of acquisition, imageability, and familiarity. *Behavior Research Methods*, *38*(4), 598-605.

Stetina, J., & Nagao, M. (1998). General Word Sense Disambiguation Method Based on a Full Sentential Context. **自然言語処理**, *5*(2), 47-74.

Stevenson, M., & Wilks, Y. (2001). The Interaction of Knowledge Sources in Word Sense Disambiguation. *CL*.

Strapparava, C., Gliozzo, A., & Giuliano, C. (2004). Pattern abstraction and term similarity for Word Sense Disambiguation : IRST at Senseval-3. *Proceedings of SENSEVAL-3, the Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, 229-234.

Stuart, M., Dixon, M., Masterson, J., & Gray, B. (2003). Children's early reading vocabulary : Description and word frequency lists. *British Journal of Educational Psychology*, *73*(4), 585-598.

Sussna, M. (1993). Word sense disambiguation for free-text indexing using a massive semantic network. *Proceedings of the second international conference on Information and knowledge management*, 67-74.

Tack, A., François, T., Desmet, P., & Fairon, C. (2018). NT2Lex : A CEFR-Graded Lexical Resource for Dutch as a Foreign Language Linked to Open Dutch WordNet. *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, 137-146.

Taghipour, K., & Ng, H. T. (2015). One Million Sense-Tagged Instances for Word Sense Disambiguation and Induction. *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, 338-344.

Tharp, J. B. (1939). The Measurement of Vocabulary Difficulty. *The Modern Language Journal*, *24*(3), 169-178.

Thorndike, E. L. (1921). *The Teacher's Word Book*. Teachers College, Columbia University.

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*. Teachers College, Columbia University.

Thorndike, E. L., & Lorge, I. (1944). *The Teacher's Word Book of 30,000 words*. Teachers College, Columbia University.

Toglia, M. P., & Battig, W. F. (1978). *Handbook of semantic word norms* (p. vii, 152). Lawrence Erlbaum.

Towell, G., & Voorhees, E. M. (1998). Disambiguating highly ambiguous words. *Computational Linguistics*, *24*(1), 125-145.

Tripodi, R., & Navigli, R. (2019). Game Theory Meets Embeddings : A Unified Framework for Word Sense Disambiguation. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 88-99.

Turdakov, D. Yu. (2010). Word sense disambiguation methods. *Programming and Computer Software*, *36*(6), 309-326.

van Heuven, W. J. B., Mandera, P., Keuleers, E., & Brysbaert, M. (2014). Subtlex-UK : A New and Improved Word Frequency Database for British English. *Quarterly Journal of Experimental Psychology*, *67*(6), 1176-1190.

Véronis, J. (2004). HyperLex : Lexical cartography for information retrieval. *Computer Speech & Language*, *18*(3), 223-252.

Veronis, J., & Ide, N. M. (1990). Word Sense Disambiguation with Very Large Neural Networks Extracted from Machine Readable Dictionaries. *COLING 1990 Volume 2: Papers presented to the 13th International Conference on Computational Linguistics*. COLING 1990.

Vial, L., Lecouteux, B., & Schwab, D. (2019). Sense Vocabulary Compression through the Semantic Knowledge of WordNet for Neural Word Sense Disambiguation.

Vitello, S., & Rodd, J. (2015). Resolving Semantic Ambiguities in Sentences : Cognitive Processes and Brain Mechanisms. *Language and Linguistics Compass*, *9*(10), 391-405.

Voorhees, E. M. (1993). Using WordNet to disambiguate word senses for text retrieval. *Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval*, 171-180.

Wang, M., & Wang, Y. (2020). A Synset Relation-enhanced Framework with a Try-again Mechanism for Word Sense Disambiguation. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 6229-6240.

Weaver, W. (1949). Translation. *Proceedings of the Conference on Mechanical Translation*. EarlyMT 1952, Massachusetts Institute of Technology.

Weiss, S. F. (1973). Learning to disambiguate. *Information Storage and Retrieval*, *9*(1), 33-41.

West, M. A. (1953). A General Service List of English words, with semantic frequencies and a supplementary word-list for the writing of popular science and technology.

Widdows, D., & Dorow, B. (2002). A Graph Model for Unsupervised Lexical Acquisition. *COLING 2002: The 19th International Conference on Computational Linguistics*. COLING 2002.

Wilks, Y. (1975). A preferential, pattern-seeking, Semantics for natural language inference. *Artificial Intelligence*, *6*(1), 53-74.

Wilks, Y., & Stevenson, M. (1996). The Grammar of Sense : Is word-sense tagging much more than part-of-speech tagging?

Yap, B. P., Koh, A., & Chng, E. S. (2020). Adapting BERT for Word Sense Disambiguation with Gloss Selection Objective and Example Sentences.

Yarowsky, D. (1992). Word-Sense Disambiguation Using Statistical Models of Roget's Categories Trained on Large Corpora. *COLING 1992 Volume 2: The 14th International Conference on Computational Linguistics*. COLING 1992.

Yarowsky, D. (1993). One Sense per Collocation. *Human Language Technology: Proceedings of a Workshop Held at Plainsboro, New Jersey, March 21-24, 1993*. HLT 1993.

Yarowsky, D. (1995). Unsupervised Word Sense Disambiguation Rivaling Supervised Methods. *33rd Annual Meeting of the Association for Computational Linguistics*, 189-196.

Yarowsky, D. (2000). Hierarchical Decision Lists for Word Sense Disambiguation. *Computers and the Humanities*, *34*(1), 179-186.

Yarowsky, D., Cucerzan, S., Florian, R., Schafer, C., & Wicentowski, R. (2001). The John Hopkins SENSEVAL-2 System Descriptions. *Proceedings of SENSEVAL-2 Second International Workshop on Evaluating Word Sense Disambiguation Systems*, 163-166.

Yimam, S. M., Biemann, C., Malmasi, S., Paetzold, G., Specia, L., Štajner, S., Tack, A., & Zampieri, M. (2018). A Report on the Complex Word Identification Shared Task 2018. *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, 66-78.

Zapata Monge, A. L. (2013). L'évolution de l'enseignement du vocabulaire dans la classe de L2. *Revista de Lenguas Modernas, 19*, 437-447.

Zeno, S., Ivens, S. H., Millard, R. T., & Duvvuri, R. (1995). *The Educator's word frequency guide*. Touchstone Applied Science Associates.

Zevin, J. D., & Seidenberg, M. S. (2002). Age of Acquisition Effects in Word Reading and Other Tasks. *Journal of Memory and Language, 47*(1), 1-29.

Zhong, Z., & Ng, H. T. (2010). It Makes Sense : A Wide-Coverage Word Sense Disambiguation System for Free Text. *Proceedings of the ACL 2010 System Demonstrations*, 78-83.

Zipf, G. K. (1949). *Human behavior and the principle of least effort.* Addison-Wesley.

UNIVERSITÉ CATHOLIQUE DE LOUVAIN
Faculté de philosophie, arts et lettres
Place Blaise Pascal, 1 bte L3.03.11, 1348 Louvain-la-Neuve, Belgique | **www.uclouvain.be/fial**