UNIVERSITÉ DU QUÉBEC À MONTRÉAL

STRATÉGIES EFFICACES POUR L'APPRENTISSAGE DES MOTS D'UN DICTIONNAIRE : UNE APPROCHE BASÉE SUR LES GRAPHES

MÉMOIRE

PRÉSENTÉ

COMME EXIGENCE PARTIELLE

DE LA MAÎTRISE EN INFORMATIQUE

PAR

JEAN-MARIE POULIN

JANVIER 2019

REMERCIEMENTS

Tout d'abord, je désire remercier profondément mon directeur de maîtrise, Alexandre Blondin Massé, pour son ouverture d'esprit envers un candidat au profil non conventionnel. Ses connaissances, ses grandes qualités pédagogiques, sa patience, sa disponibilité, les encouragements qu'il a su me prodiguer au cours de ces quelques années m'auront permis de mener à terme ce travail de longue haleine.

Merci aussi aux professeurs qui ont su partager leurs connaissances et qui m'ont permis d'approfondir des aspects pour moi inexplorés de l'informatique : Guy Tremblay, Fatiha Sadat et Étienne Gagnon de l'UQAM, ainsi que Pascal Vincent, de l'Université de Montréal

Je tiens aussi à remercier personnellement Étienne Harnad. Je me considère privilégié d'avoir eu la chance de côtoyer un chercheur de sa trempe.

Merci infiniment à Hélène, ma compagne, mon égérie, mon inspiratrice. Sans son appui indéfectible, sa patience, sa contribution même, je n'aurais sans doute pas réussi à venir à bout de la rédaction de ce mémoire.

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

RÉSUMÉ

Nous nous intéressons dans ce mémoire à la structure des dictionnaires, plus spécifiquement à l'entrecroisement des liens qui unissent entre eux les mots à travers leurs définitions. À quelques exceptions près, tous les mots utilisés pour construire les définitions sont définis quelque part ailleurs dans le dictionnaire. Toutes ces références entre les mots créent entre eux un réseau de relations pouvant être représenté par un graphe, rendant ainsi possible l'étude des dictionnaires avec les algorithmes de la théorie des graphes.

Plusieurs façons de faire ont déjà été avancées pour étudier les caractéristiques des dictionnaires à travers les propriétés de leurs graphes associés. Dans cette perspective, nous présentons une nouvelle piste de recherche, consistant à utiliser l'apprentissage comme outil d'investigation : Étant donné un dictionnaire ou un lexique, quelle serait la meilleure stratégie à adopter pour apprendre tous ses mots ? À l'aide de concepts simples de la théorie des graphes, nous proposons un modèle formel et des algorithmes qui permettent de répondre à cette question. Nous évaluons plusieurs stratégies d'apprentissage différentes en comparant le rythme auquel l'apprentissage se déroule et le taux de rendement final par rapport à huit dictionnaires monolingues de langue anglaise.

Il s'avère que le facteur qui influence le plus la performance des différentes stratégies d'apprentissage est leur capacité à briser la circularité des définitions. Autrement dit, les stratégies possédant le plus haut taux de rendement sont celles qui parviennent à briser le plus rapidement possible les boucles de définitions entre les mots. Nous montrons qu'une stratégie algorithmique très simple, basée uniquement sur le degré extérieur des sommets – le nombre de définitions où les lexèmes sont utilisés –, améliore de façon importante le processus d'apprentissage par rapport à diverses stratégies d'apprentissage psycholinguistiques. Nous avançons aussi l'hypothèse qu'une telle approche pourrait représenter une solution de rechange à l'utilisation de *corpus* pour la création des « listes de mots » utilisées en didactique des langues.

Mots-clés : Dictionnaires ; Lexiques ; Mots ; Lexèmes ; Apprentissage ; Stratégies.

INTRODUCTION

> *« Dictionnaire : Recueil des mots d'une langue ou d'un domaine de l'activité humaine, réunis selon une nomenclature d'importance variable et présentés généralement par ordre alphabétique, fournissant sur chaque mot un certain nombre d'informations relatives à son sens et à son emploi et destiné à un public défini. »*
>
> – Trésor de la langue française (TLFi, 2019)

Que ce soit sous la forme de tablettes d'argile, de papyrus, de manuscrits, de livres imprimés, de page Web ou de tablettes électroniques, les dictionnaires existent depuis l'Antiquité (Boulanger, 2003). Depuis cette époque, ils sont utilisés communément comme ouvrages de référence dans tous les domaines de connaissance liés à la langue. Ils représentent des ressources indispensables pour tout ce qui touche la lecture, l'écriture, la traduction de textes et même l'acquisition de connaissances générales.

Avec l'apparition de l'imprimerie au début de la Renaissance, et surtout avec, à la fin du 20e siècle, le développement des ordinateurs et la représentation numérique des connaissances, les dictionnaires ont subi de profondes métamorphoses. En dépit de cela les dictionnaires, les lexiques et les encyclopédies de toutes sortes conservent encore aujourd'hui toute leur pertinence. Les plateformes ouvertes, comme le Wiktionnaire ou Wikipédia, ou encore les versions Web de dictionnaires commerciaux, comme le Larousse (Larousse, 2019) ou le Merriam-Webster (Merriam-Webster, 2019a), connaissent une popularité sans cesse croissante.

L'un des facteurs déterminants de ce succès est sans doute l'intégration des notions d'hypertexte et d'hyperlien. Ces nouvelles technologies permettent aux lexicographes

d'établir facilement des liens de diverses natures entre les mots et les concepts d'un même ouvrage, ou même de diriger l'utilisateur vers des ressources Web externes. Il devient alors possible de naviguer facilement d'un mot à l'autre, sans avoir à feuilleter laborieusement les milliers de pages d'un ouvrage papier. Le mode d'utilisation des dictionnaires s'en trouve profondément modifié. La richesse des relations entre les mots prend autant d'importance que les informations fournies sur les mots eux-mêmes.

Récemment, aidés en cela par les développements en psychologie cognitive et en traitement automatique des langues, des chercheurs ont commencé à s'interroger sur la manière dont ces liens entre les mots des dictionnaires sont organisés. Existe-t-il des invariants ou des motifs communs à tous les dictionnaires ? Ce questionnement a fait l'objet de plusieurs articles ayant comme thème l'analyse de la structure des dictionnaires.

Dans un des premiers travaux sur le sujet, Clark a examiné le vocabulaire de contrôle [1] des dictionnaires LDOCE et CIDE (Clark, 2003; Procter, 1995, 1978). Il a démontré que les mots provenant du vocabulaire de contrôle possèdent des propriétés particulières : ils sont en majorité plus abstraits et leur définition est plus longue et plus complexe que celle des autres mots. Par la suite, Steyvers et Tenenbaum (Steyvers et Tenenbaum, 2005) ont poursuivi en analysant la structure des graphes associés au réseau sémantique WordNet et au Thesaurus de Roget (Fellbaum, 1998; Roget, 1911).

Continuant dans la même veine, des chercheurs ont publié ces dernières années une série d'articles dans le but d'explorer la structure interne des dictionnaires (Blondin Massé *et al.*, 2008; Picard *et al.*, 2009, 2010, 2013; Vincent-Lamarre *et al.*, 2016). Ces travaux ont comme point commun d'utiliser un *modèle formel de lexique*, basé sur la théorie des graphes. Cette approche permet d'appliquer aux dictionnaires des algorithmes classiques de traitement des graphes et de déduire une foule d'informations pertinentes d'un point de vue linguistique. De leur analyse de plusieurs dictionnaires numériques de langue anglaise, il ressort que ceux-ci possèdent tous une structure commune et contiennent les

1. Certains dictionnaires, comme LDOCE et CIDE, sont construits à l'aide d'un vocabulaire de contrôle, c'est-à-dire un sous-ensemble prédéfini de mots servant à construire les définitions.

mêmes composantes de base (Picard *et al.*, 2013), soit :

Un noyau, qui correspond à un sous-ensemble de mots du dictionnaire permettant de définir tous les autres mots. Le noyau peut à son tour être subdivisé en une série de sous-composantes de taille variable, formées de groupes de mots étroitement reliés entre eux.

Un cœur, qui est la sous-composante du noyau comprenant le plus grand nombre de mots. Dans tous les dictionnaires numériques étudiés, le cœur est considérablement plus volumineux que les autres sous-composantes du noyau.

Un ensemble d'ancrage minimal, formé d'un sous-ensemble de mots plus petit que le noyau, obtenu en combinant judicieusement des éléments du cœur et des autres sous-composantes du noyau. C'est le plus petit groupe de mots qui permet de définir tous les autres mots.

De plus, il s'avère que le noyau possède des caractéristiques psycholinguistiques particulières (Vincent-Lamarre *et al.*, 2016) :

- Les mots du noyau sont appris plus tôt, sont plus concrets et sont plus fréquemment utilisés que les autres mots du dictionnaire.
- Il existe une forte corrélation entre ces différentes variables psycholinguistiques mesurant l'âge d'acquisition, le degré d'abstraction, ainsi que la fréquence d'utilisation des mots.
- À l'intérieur du noyau lui-même, l'on remarque une gradation marquée de ces mêmes mesures selon que l'on évalue les mots du noyau, du cœur, ou d'un ensemble d'ancrage minimal.

Parmi ces observations sur la structure des dictionnaires et des graphes associés l'élément principal à retenir est sans doute la question de l'*ensemble d'ancrage minimal* (Minset). Dans (Vincent-Lamarre *et al.*, 2016), les auteurs établissent un lien direct entre cet ensemble et le « problème de l'ancrage symbolique » – en anglais *Symbol Grounding Problem* – (Minset), décrit par Harnad (Harnad, 1990). L'on peut résumer sommairement ce problème de la façon suivante. Lorsque nous consultons la définition d'un mot dans un

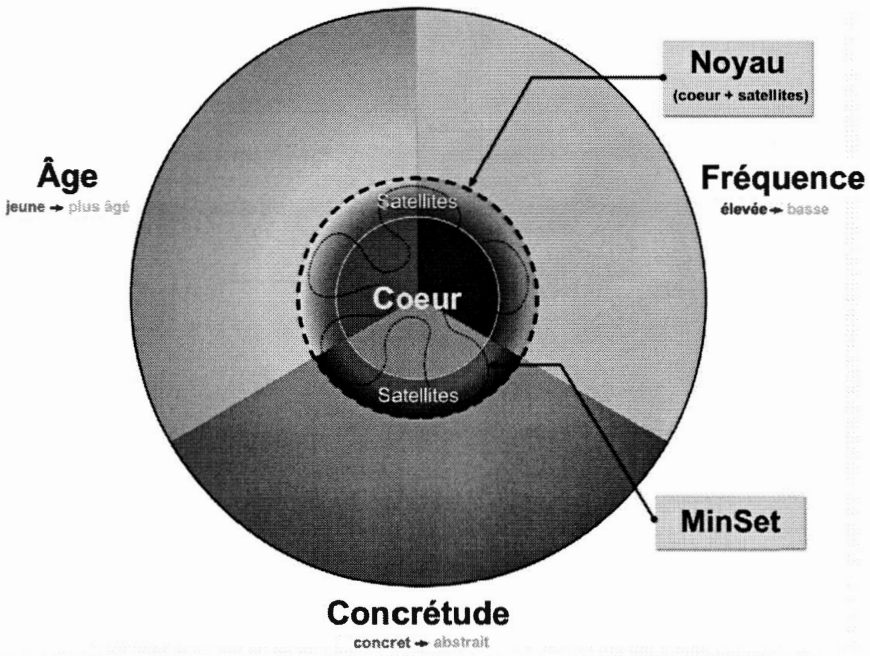


Fig. 1: Structure interne d'un dictionnaire. Source : (Vincent-Lamarre *et al.*, 2016)

dictionnaire, nous constatons que cette définition est construite à l'aide d'autres mots. Si ces mots ne sont pas connus, nous pouvons bien entendu les chercher à leur tour dans le dictionnaire. Mais, sous peine de tourner en boucle indéfiniment, la signification de certains d'entre eux doit être connue et ancrée dans l'expérience sensorimotrice : "[...] it cannot be dictionary look-ups all the way down!" (Vincent-Lamarre *et al.*, 2016).

Le problème de l'ancrage symbolique se pose de façon particulièrement aiguë lors de l'apprentissage d'une autre langue. Lorsque l'on commence à apprendre une nouvelle langue, il faut non seulement se familiariser avec la grammaire et la syntaxe, mais également acquérir du vocabulaire et apprendre suffisamment de nouveaux mots pour pouvoir comprendre et se faire comprendre. L'on doit pouvoir associer la forme extérieure d'un mot écrit ou parlé, avec son sens, sa signification dans le contexte de la communication. Selon Schmitt, le lien entre la forme et le sens des mots est l'élément lexical le plus important à acquérir (Schmitt, 2010) :

> "[the] form-meaning link is the first and most essential lexical aspect which

must be acquired".

Quelle est la meilleure façon d'apprendre tous ces nouveaux mots ? Existe-t-il en didactique des langues des façons de faire ou des stratégies privilégiées ? Dans de nombreux ouvrages, comme ceux de Prince (Prince, 1996), de Schmitt (Schmitt, 2008) et de Joyce (Joyce, 2015), les auteurs mettent en parallèle deux approches traditionnellement utilisées pour l'enseignement du vocabulaire d'anglais langue seconde. La première méthode, appelée "L1 translation"[2], consiste à expliquer les nouveaux mots d'anglais dans la langue maternelle de l'étudiant. Par exemple, si l'étudiant est hispanophone, le professeur lui fournit une explication ou une définition du mot anglais *cat* en espagnol, c.-à-d. *gato, felino.* Avec la deuxième approche, nommée selon les auteurs "L2 context" ou "L2 definition", l'étudiant doit déduire par lui-même le sens d'un nouveau mot en se servant du contexte dans lequel le mot est introduit ou d'une explication en anglais. On pourrait par exemple expliquer à Jacques, un étudiant francophone, le mot anglais OWN avec une définition comme : "*to have or hold as property*".

Joyce (Joyce, 2015) évalue ces deux approches ainsi :

- l'approche "L1 translation" est privilégiée pour les étudiants dont le niveau de compétence est moins élevé. ("[...] L1 translations for intentional vocabulary learning is seen as being most effective for students at lower proficiency levels").
- l'approche "L2 definition" est la plus efficace pour le développement du vocabulaire ("for the purposes of general language development, learning through an L2 definition is favoured").

Un simple dictionnaire de langue peut donc être un moyen étonnamment efficace pour comprendre et mémoriser les nouveaux mots, les situer dans leur contexte cognitif. Il y a cependant un prérequis important. L'apprenant doit maîtriser au préalable un sous-ensemble « de base » des mots de la nouvelle langue. Ce n'est que de cette façon qu'il sera en mesure d'utiliser avec profit un dictionnaire.

---

2. Dans le contexte de l'enseignement d'une langue seconde, L1 fait référence à la langue maternelle de l'étudiant et L2 fait référence à la langue étudiée.

Reprenons l'exemple précédent de l'étudiant Jacques, qui rencontre dans un texte anglais le mot OWN, qu'il ne connaît pas. Il consulte donc le Merriam-Webster (Merriam-Webster, 2003) et trouve la définition suivante : "*to have or hold as property*". Supposons qu'il connaisse déjà le sens des mots TO, HAVE, OR, HOLD et AS, mais pas celui du mot PROPERTY. Jacques est un étudiant persévérant. Il continue sa recherche dans le dictionnaire et trouve la définition suivante pour PROPERTY : *"something that is owned by a person"*. Même s'il connaît bien les mots SOMETHING, THAT, IS/BE, BY et PERSON, il n'a pas résolu son problème. Il est confronté à ce que l'on appelle une *boucle de définition*. Il faudrait qu'il connaisse le sens de owned/OWN pour comprendre le sens de PROPERTY, alors qu'au départ c'est ce même mot OWN qu'il voulait apprendre. La seule manière de s'en sortir, c'est d'apprendre d'une autre façon l'un des 2 mots pour briser la boucle, par exemple en interrogeant son professeur. L'on retrouve ici la même difficulté évoquée précédemment, c'est-à-dire le « problème de l'ancrage des symboles ». Les définitions d'un dictionnaire ne sont pas suffisantes par elles-mêmes pour apprendre les mots d'une langue. La signification d'un sous-ensemble « de base » des mots d'une langue doit être connue et ancrée par un moyen quelconque dans l'expérience sensorimotrice.

Ces différentes questions sont au cœur de nos préoccupations dans ce mémoire. Nous y étudions la relation étroite qui existe entre la structure des dictionnaires monolingues et la façon dont les mots de cette langue peuvent être appris.

Précisons maintenant la démarche utilisée lors de la rédaction de ce document. Dans l'article de Vincent-Lamarre *et al* (Vincent-Lamarre *et al.*, 2016), les auteurs abordent de manière analytique l'exploration de la structure interne des dictionnaires. Ils cherchent à savoir s'il existe des groupes de mots qui possèdent des qualités en termes de structure de graphes ou de caractéristiques psycholinguistiques. Pour ce faire, ils évaluent les relations de définition entre tous les mots pour déterminer s'il est possible de découvrir des regroupements de mots possédant des propriétés particulières du point de vue de l'ancrage symbolique.

Nous poursuivons dans la même ligne de pensée, en faisant appel à une démarche com-

plémentaire. Tout d'abord, nous construisons des listes de mots, appelés « stratégies d'apprentissage », établies à partir de séquences de mots possédant des caractéristiques psycholinguistiques particulières ou de calculs algorithmiques standards de la théorie des graphes. Nous évaluons ensuite de manière comparative comment ces différentes stratégies se comportent par rapport à l'exécution d'une tâche de référence consistant à « apprendre » tous les mots d'un dictionnaire. Nous déterminons avec quel degré d'efficacité elles parviennent à briser les boucles de définition dans les dictionnaires, évitant ainsi le « problème de l'ancrage des symboles ».

Dans le premier chapitre, nous introduisons la terminologie linguistique et les quelques notions de base de la théorie des graphes qui ont servi à établir notre modèle formel de lexique. Nous proposons ensuite une façon de représenter un lexique sous la forme d'un graphe orienté.

Dans le second chapitre, nous décrivons la notion de « stratégie d'apprentissage ». Tout d'abord, nous examinons plus en détail le problème de l'ancrage des symboles. Puis, nous abordons la question des listes de mots, ces outils d'enseignement fréquemment utilisés par les professeurs de langue. Par la suite, nous proposons un modèle d'apprentissage formel ainsi que des algorithmes nous permettant d'évaluer le taux de rendement des diverses stratégies par rapport à la tâche consistant à « apprendre » tous les mots d'un lexique.

Dans le chapitre III, nous présentons l'environnement d'expérimentation utilisé. Nous y indiquons la provenance des dictionnaires numériques et des listes de mots basées sur des variables psycholinguistiques. Puis nous décrivons les deux types de stratégies d'apprentissage développées :

- les stratégies algorithmiques, construites à l'aide d'algorithmes basés sur la théorie des graphes ;
- les stratégies psycholinguistiques, basées sur des listes de mots ordonnés selon des propriétés psycholinguistiques spécifiques.

Le quatrième chapitre détaille le contenu des expérimentations réalisées. Nous y décrivons de quelle façon nous avons mesuré le rendement des stratégies d'apprentissage en fonction des dictionnaires évalués, ainsi que les différents types de données colligées. Nous présentons ensuite les résultats obtenus sous la forme de tableaux comparatifs et de graphiques. Après quoi, une analyse rapide fait ressortir les éléments les plus importants.

Pour terminer, nous concluons notre exposé en mettant en relief les observations les plus significatives et en proposant quelques avenues pour élargir la recherche dans le futur.

Mentionnons finalement que le sujet traité dans ce mémoire a fait l'objet d'un article accepté et présenté à la conférence "Cognitive 2018" (Poulin *et al.*, 2018). Il y a obtenu une récompense dans la catégorie "Best Papers".

CHAPITRE I

DICTIONNAIRES, LEXIQUES ET GRAPHES

Afin de mieux cerner le sujet de notre étude, reprenons la définition de « dictionnaire », telle que nous l'avons présentée en introduction.

> « Dictionnaire : Recueil des mots d'une langue ou d'un domaine de l'activité humaine, réunis selon une nomenclature d'importance variable et présentés généralement par ordre alphabétique, fournissant sur chaque mot un certain nombre d'informations relatives à son sens et à son emploi et destiné à un public défini. »(TLFi, 2019)

Cette description correspond assez bien à la vision traditionnelle que la plupart des gens ont d'un dictionnaire. Cependant, en y regardant de plus près, un élément central de cette formulation, le terme *mots*, mérite que l'on s'y attarde plus longuement. Regardons un premier exemple, tiré de Polguère, où l'on voit *mot* utilisé de façon ambiguë, avec deux sens différents (Polguère, 2016, p. 47).

Exemple 1.

a) « "Parce que" s'écrit en deux mots ». Ici, *mot* correspond directement à la définition courante en langage écrit, « un segment de discours compris entre deux espaces blancs »(Arrivé, 1986).

b) « "Parce que" est un mot qui se traduit en anglais par *because* ». Dans cette phrase, *mot* fait référence au groupe « Parce que » au complet.

Voici un autre exemple, qui illustre de façon différente l'ambivalence du terme *mot*.

Exemple 2.

a) Nous avons trouvé un *chat* sur la galerie.

b) Il y a beaucoup de *chats* dans le voisinage.

Par rapport à l'exemple 1, le problème se pose de façon différente. Est-ce que *chat* et *chats* sont deux mots différents ? Si l'on applique encore une fois à la lettre la définition d'Arrivé (1986), nous sommes ici en présence de deux *mots*. Cependant, grâce à notre connaissance du français, nous comprenons facilement que dans les deux cas, c'est du même « petit animal domestique carnassier [...] »(TLFi, 2019) dont il est question. Dans la phrase 2 b), c'est le même *mot* « chat » qui a été accordé au pluriel pour indiquer que l'on parle de plusieurs animaux.

Par contre, l'ambiguïté de *mot* représente une difficulté importante pour notre travail : ce n'est pas un terme assez précis. Nous devons trouver une façon de distinguer ces différents cas de figure afin d'effectuer des traitements automatisés sur les dictionnaires. C'est pourquoi nous introduisons d'abord une terminologie linguistique précise, nous permettant de réduire le plus possible l'imprécision du vocabulaire.

Nous mettons ensuite à profit cette terminologie pour proposer une définition formelle d'un lexique. Puis, après avoir rappelé quelques notions élémentaires de la théorie des graphes, nous examinons comment il est possible de représenter un lexique sous la forme d'un graphe orienté.

1.1 Terminologie

Il n'existe pas vraiment de consensus entre les différents auteurs et écoles de pensées linguistiques quant à la terminologie à employer. Nous proposons dans cette section une nomenclature des notions linguistiques nécessaires pour décrire notre modèle formel d'apprentissage des mots d'un dictionnaire, tout en facilitant la compréhension de notre document.

1.1.1 Lexique

D'un point de vue linguistique, quelle est la différence entre un lexique et un dictionnaire ? En anglais, le terme *lexicon* est un synonyme très répandu pour *dictionary*. Selon le Merriam-Webster et le *Handbook of Linguistics*, c'est un livre contenant une liste de mots accompagnés de leur définition, présentés en ordre alphabétique (Merriam-Webster, 2019c; Cruse, 2002). Par contre en français le terme « lexique », envisagé comme synonyme de dictionnaire, est considéré comme vieilli (TLFi, 2019, LEXIQUE). Pour les fins de notre mémoire, nous utilisons le terme **lexique** dans son sens linguistique, c'est-à-dire : « l'entité théorique qui correspond à l'ensemble de toutes les lexies d'une langue ou encore [du lexique mental] d'un individu »(Polguère, 2016, p. 109). Remarquons que cette définition fait référence à un « ensemble de lexies », et non à un « ensemble de mots ». Nous examinons plus loin la distinction qu'il y a lieu d'apporter entre ces 2 termes.

Pour mieux faire ressortir la différence entre un dictionnaire et un lexique, ajoutons les précisions suivantes :

1. Un **dictionnaire** est un modèle, une représentation particulière du lexique d'une langue, qui met l'accent sur l'aspect descriptif, la définition des mots – les lexèmes –.
2. Dans un **lexique**, les relations entre les mots sont aussi importantes que les *mots* eux-mêmes : ce n'est pas une simple liste séquentielle de mots. On peut aussi voir un lexique comme une toile – en anglais *web* –, où les mots sont associés entre eux par un réseau complexe de relations de diverses sortes.

Parmi les nombreuses relations différentes que les *mots* peuvent entretenir entre eux, considérons quelques exemples :

Exemple 3.

a) Dans la phrase « Le chat est un animal domestique », CHAT et ANIMAL sont reliés entre eux à la fois par une relation d'hyponymie et d'hyperonymie. CHAT est un hyponyme d'ANIMAL, alors qu'en sens inverse, ANIMAL est un hyperonyme de

CHAT.

b) Dans la phrase « J'ai vu un chat errant », CHAT et ERRANT sont reliés par une relation d'un autre ordre. ERRANT est ici une qualité qu'il est habituel d'appliquer au *mot* CHAT (Beauchesne *et al.*, 2009). Cependant, le qualificatif *violet*, comme dans « J'ai vu un chat violet », n'est pas approprié pour un chat, à moins d'être dans un contexte particulier, par exemple dans une bande dessinée.

c) Dans la définition de CHAT tirée du Petit Robert (Robert *et al.*, 1979), « Petit mammifère familier à poil doux, ... », les *mots* PETIT, MAMMIFÈRE, etc., entretiennent avec CHAT une relation différente encore. Ils contribuent à décrire, à définir ce qu'est un chat.

C'est ce dernier type de relation, qualifiée de relation « définitionnelle », que nous employons pour explorer la structure des lexiques.

1.1.2 Mots, lexèmes et autres

Voyons maintenant les différents éléments qui composent notre terminologie. La figure 1.1 illustre sous forme de diagramme entité-relation les liens réciproques qui existent entre les termes linguistiques présentés. Ces termes, ainsi que les conventions d'écriture associées, sont fortement inspirés de Polguère (Polguère, 2016).

mot-forme :

Un **mot-forme** est un signe linguistique, un « segment de discours », possédant des qualités particulières d'autonomie et de cohésion (Arrivé, 1986). Pour la langue anglaise, le *Oxford Dictionary* définit un mot-forme – en anglais *word form* – comme : "a (particular) form of a word; especially each of the possible forms taken by a given lexeme, typically distinguished by their grammatical inflections" (Oxford, 2019).

Sans entrer plus avant dans la théorie linguistique, nous disons simplement que *chat* et *chats* sont deux mots-formes différents du même mot « chat », faisant tous les deux référence ~~au~~ à l'idée générale de <chat>.

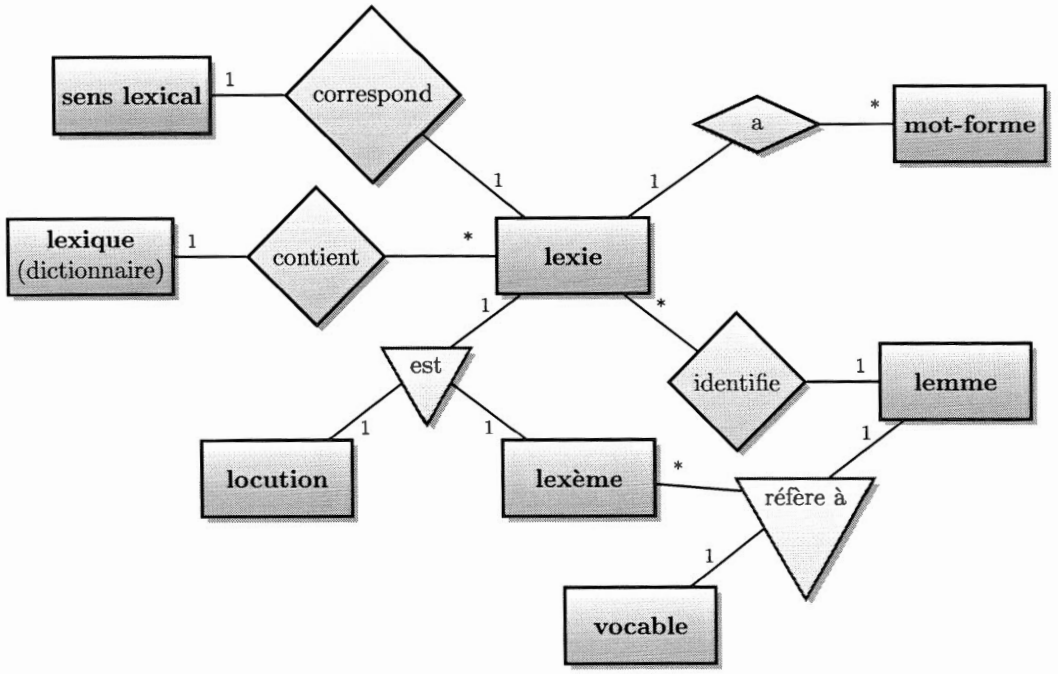


Fig. 1.1: Diagramme entité-relation des termes linguistiques

Avec cette définition, nous pouvons reformuler l'exemple 1.a, à la page 9, comme suit :

Exemple 4.

« "Parce que" s'écrit avec deux *mots-formes* ».

Convention d'écriture :
Les **mot-formes** sont notés en italique, par exemple *chats*.

lexie :

La **lexie** – en anglais *lexical item* ou *headword* – est l'unité de base du lexique, l'équivalent d'une entrée dans un dictionnaire. « Une lexie, aussi appelée unité lexicale, est soit un lexème, soit une locution. Chaque lexie (lexème ou locution) est associée à un sens donné [...] » (Polguère, 2016, p. 69)

Par exemple, les mots « pomme de terre » et « chat » sont tous deux des lexies.

« Chat » correspond au lexème CHAT, une lexie simple constituée d'un seul mot-forme, tandis que « pomme de terre » est une locution, une lexie composée de mots-formes associés.

locution :

Si l'on examine plus en détail le cas de « Parce que » dans l'exemple 1 à la page 9, l'on constate que, bien qu'ils soient écrits de façon séparée, les mots-formes *Parce* et *que* ne forment qu'une seule entrée dans le dictionnaire. *Parce* pris de façon isolée ne veut rien dire. Le sens lexical est associé à la combinaison des deux mots-formes. C'est la même chose pour « pomme de terre », une lexie constituée des mots-formes *pomme*, *de* et *terre*. « Pomme de terre » n'a pas le même sens que « pomme » ou « terre » pris isolément. C'est ce qui s'appelle une **locution**.

Nous pouvons reformuler l'exemple 1.b, à la page 9, comme suit :

Exemple 5.

« "Parce que" est une *locution* qui se traduit en anglais par *because* ».

Cela dit, il reste que la tâche visant à déterminer correctement si un groupe de mots-formes correspond à une locution est en soi un processus complexe, qui ne fait pas partie de la portée du mémoire. Nous ne prenons donc pas en compte les locutions dans notre analyse ; elles sont éliminées lors de la transformation des dictionnaires en graphes de lexèmes.

lexème :

Considérons de nouveau l'exemple 2 à la page 10 . Dans les deux phrases, l'on utilise deux formes graphiques distinctes, deux mots-formes, *chat* et *chats*, qui font référence à la même idée, au même noyau de sens, le **sens lexical** <chat>. Ces deux mots-formes sont en fait deux « formes fléchies » [1] du même lexème CHAT.

---

1. Terme linguistique, dérivé de flexion : « Changement morphologique dans la finale d'un mot (nom, pronom, participe, adjectif) selon la fonction qu'il occupe dans la phrase ou dans la proposition, par l'adjonction d'un affixe ou désinence [...] » (TLFi, 2019)

Polguère définit un **lexème** comme « [...] une généralisation du signe linguistique de type mot-forme : chaque lexème de la langue est structuré autour d'un sens exprimable par un ensemble de mots-formes que seule distingue la flexion »(Polguère, 2016). En d'autres mots, nous pouvons nous représenter un lexème comme une façon d'identifier un **sens lexical** précis, auquel sont associées une série de variations grammaticales, représentées par des **mots-formes** différents. La même idée s'applique pour l'accord grammatical des verbes. Pour Spencer (Spencer, 2002), les mots-formes {*write*, *writes*, *written*, ... } sont des formes grammaticales différentes du même lexème WRITE.

Convention d'écriture :
Les **lexèmes** sont écrits en petites majuscules, comme dans CHAT.
Ils peuvent aussi être présentés sous une forme plus complexe, comme dans $\text{CHAT}^{1}_{\text{N}}$, où l'exposant « ^{1} » indique le résultat de la désambiguïsation et où l'indice « $_{\text{N}}$ » représente la **partie du discours**.

En outre, comme nous l'avons mentionné précédemment, les locutions ne sont pas prises en compte dans notre travail. Nous pouvons donc simplifier la terminologie en faisant l'économie du terme **lexie**. Nous pouvons voir à la figure 1.2 une version simplifiée du diagramme entité-relation. Donc, dans notre vocabulaire, nous utilisons **lexème** en remplacement de **lexie** à titre d'entrée dans un lexique – en anglais *headword* –.

sens lexical :

Dans notre travail, nous choisissons d'employer **sens lexical** pour faire référence à l'idée, à l'image mentale à laquelle un lexème renvoie. « Le sens lexical renvoie à un concept mental qui est associé à une unité lexicale pour permettre d'exprimer une idée. » (Wenski-Béthoux, 2005)

Le terme sens lexical peut, selon les disciplines et les auteurs, être mis en parallèle avec les notions apparentées de *concept* : « Entre tous les individus ainsi reliés par le langage, il s'établira une sorte de moyenne : tous reproduiront [...] les mêmes signes unis aux mêmes concepts » (De Saussure, 1989), de *catégorie*, en philosophie

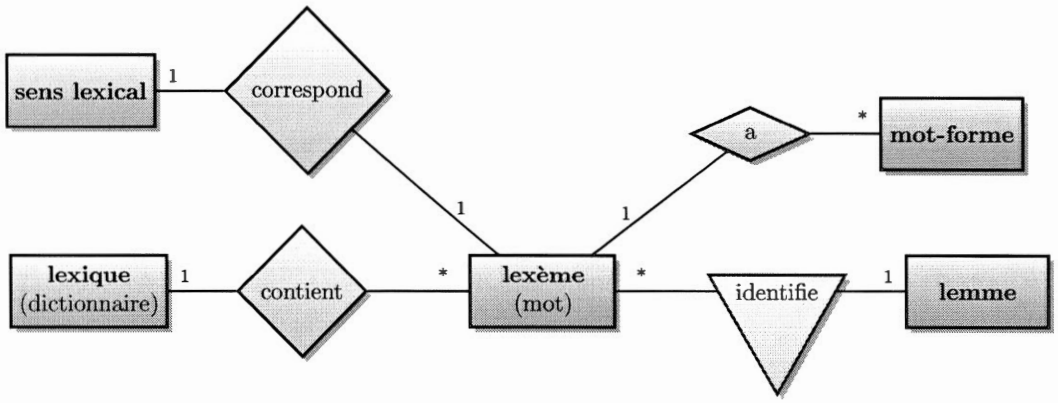


Fig. 1.2: Diagramme entité-relation des termes linguistiques
(version simplifiée)

et en psychologie cognitive, ou encore de *signifié*, en sémiotique.

Convention d'écriture :
Le **sens lexical** d'un lexème est noté avec des chevrons simples. Par exemple, <chat> est le sens lexical associé au lexème CHAT.

vocable

Le terme **vocable** fait référence à un ensemble de lexèmes qui partagent le même mot-forme canonique – lemme, comme nous le voyons plus loin, – mais qui n'ont pas « exactement » le même sens. Les lexèmes d'un même vocable conservent toutefois une parenté sémantique plus ou moins rapprochée, comme THÉ[1] et THÉ[2] dans l'exemple 6.

La question des vocables polysémiques est traitée à la section 1.1.3, page 19.

Exemple 6.

a) THÉ[1] : pour désigner les feuilles séchées du théier, comme dans la phrase « As-tu acheté du thé ? »

b) THÉ[2] : pour désigner la boisson elle-même, comme dans la phrase « Voudrais-tu un thé ? »

En revanche, dans l'exemple 7 qui suit, les lexèmes LOUER^{I} et LOUER^{II} ne font pas partie *au sens strict* du même vocable. Puisqu'ils sont d'origine étymologique différente (TLFi, 2019), l'on parle dans ce cas d'homonymes.

Exemple 7.

a) LOUER^{I} : Adresser un compliment, du latin *laudare*

b) LOUER^{II} : Louer un logement, du latin *locare*

Dans le contexte de notre travail, il n'est pas utile de faire cette distinction entre vocables polysémiques et homonymes. Nous simplifions là aussi la terminologie (voir la figure 1.2, à la page 16) en intégrant la notion de vocable à celle de lemme.

lemme

Selon Polguère, un **lemme** est le mot-forme canonique employé pour désigner un vocable (Polguère, 2016, p. 135). En français par exemple, l'on utilise l'infinitif présent pour représenter un verbe, le masculin singulier pour représenter un nom, etc.. Dans notre nomenclature, nous disons que c'est le mot-forme qui a été choisi pour identifier un ou plusieurs lexèmes.

Convention d'écriture :

1. Un **lemme** est écrit en police non proportionnelle, comme par exemple CHAT.
2. Pour distinguer les lexèmes associés à un même **lemme**, nous utilisons un exposant compris entre 1 et n, par exemple CHAT^{1}, CHAT^{2}, ..., CHAT^{n}.

En traitement automatique des langues, on appelle lemmatisation l'opération qui consiste à identifier le lemme qui correspond aux différents mots-formes d'un lexème. Par exemple : le lemme ALLER est le résultat de la lemmatisation des mots-formes {*vais*, *vas*, *va*, *allons*, ... }.

La distinction entre les termes **lexème**, **vocable** et **lemme** peut parfois sembler difficile à établir. Pour aider à les démarquer, il suffit de se rappeler que lexème et vocable sont plutôt liés au sens, à la sémantique, alors que lemme est plus lié à la

forme, à la morphologie.

partie du discours

Selon Polguère, les **parties du discours** sont « des classes générales dans lesquelles sont regroupées les lexies de la langue en fonction de leurs propriétés grammaticales. » (Polguère, 2016).

Pour les besoins de notre exposé, nous considérons la partie du discours comme étant un attribut d'un lexème, résultant de la classification des lexèmes selon leurs propriétés grammaticales et morphologiques. Tous les lexèmes font partie de l'une des 5 parties du discours suivantes :

- **noms**,
- **verbes**,
- **adjectifs**,
- **adverbes**,
- **mots fonctionnels**,

Les quatre premières classes – nom, verbe, adjectif et adverbe – regroupent la très grande majorité des lexèmes. La cinquième classe, celle des mots fonctionnels – en anglais *stop words* –, regroupe tous les autres lexèmes dont la valeur sémantique est plus pauvre.

Convention d'écriture :
La **partie du discours** d'un lexème est représentée par un indice accompagnant le lexème : « $_{N}$ » pour un nom, « $_{V}$ » pour un verbe, « $_{A}$ » pour un adjectif, « $_{R}$ » pour un adverbe et « $_{S}$ » pour un mot fonctionnel.
Par exemple CHAT_{N} indique que le lexème CHAT est un nom.

En traitement automatique des langues (TAL), on appelle étiquetage morphosyntaxique – en anglais *POS-tagging* – l'opération qui consiste à déterminer, pour chacun des lexèmes d'une phrase ou d'un texte, la partie du discours à laquelle il appartient. Nous disons alors qu'un lexème est étiqueté morpho-syntaxiquement, ou plus simplement, étiqueté.

1.1.3 Polysémie et désambiguïsation

Dans cette section, nous proposons un bref survol de la question de la polysémie. À la section 1.1.2, page 16, nous avons rapidement mentionné qu'un même vocable pouvait, selon les cas, correspondre à plus d'un lexème. À ce sujet, les chercheurs en linguistique font habituellement la nuance entre deux situations différentes (Duchacek, 1962) :

- l'*homonymie*, lorsque les lexèmes sont d'origine étymologique différente :

  Exemple 8.

  a) LOUER^{I} : Adresser un compliment, du latin *laudare* «louer, approuver, vanter» (Gaffiot, 2000)

  b) LOUER^{II} : Louer un logement, du latin *locare* «donner à loyer, à ferme» (Gaffiot, 2000)

- la *polysémie*, lorsqu'il s'agit d'acceptions [2] différentes d'un même vocable.

  Exemple 9.

  a) CAFÉ^{1} : la boisson

  b) CAFÉ^{2} : les grains de la plante

Pour la langue anglaise, l'exemple suivant tiré de (Jurafsky et Martin, 2009) illustre bien cette multiplicité des sens possibles d'un mot :

Exemple 10.

a) "Instead, a *bank* can hold the investments in a custodial account in the client's name."

b) "But as agriculture burgeons on the east *bank*, the river will shrink even more."

c) "The *bank* is on the corner of Nassau and Witherspoon."

---

2. Les différents lexèmes d'un vocable sont parfois aussi appelés des acceptions de ce vocable (Polguère, 2016, p. 70)

Pour comprendre ces phrases, il faut pouvoir distinguer quel sens est le plus approprié parmi les choix possibles pour le lemme BANK, :

$\text{BANK}_{\text{N}}^{1}$: "financial institution",

$\text{BANK}_{\text{N}}^{2}$: "building belonging to a financial institution",

$\text{BANK}_{\text{N}}^{3}$: "sloping mound",

Pour les phrases a) et b) de l'exemple 10, comme le contexte est assez différent, il est relativement facile de discriminer les sens. Dans un cas, il est question de INVESTMENT, ACCOUNT et CLIENT, alors que dans le second l'on parle d' AGRICULTURE et RIVER, etc.. Le domaine sémantique est carrément différent. Nous sommes donc en présence d'un cas simple d'homonymie.

Par contre, la phrase c) est plus compliquée à analyser. Nous ne disposons pas de beaucoup d'indices provenant du contexte pour orienter le choix. Il faut savoir ou imaginer que NASSAU and WITHERSPOON sont des noms de rues, pour en arriver à déduire qu'il est question d'un édifice, donc de la succursale d'une BANK.

C'est ce processus complexe de distinction de sens que l'on appelle « désambiguïsation lexicale » – en anglais *Word Sense Disambiguation (WSD)* – ou simplement *désambiguïsation.* Pour un humain, la distinction se fait naturellement, sans effort apparent. Par contre, c'est une tout autre paire de manches pour un algorithme ou un programme informatique :

> " The reason that lexical polysemy causes so little actual ambiguity is that, in actual use, context provides information that can be used to select the intended sense. Although contextual disambiguation is simple enough when people do it, it is not easy for a computer to do " (Miller, 1986)

La question de la désambiguïsation lexicale en intelligence artificielle demeure encore en 2018, selon Corrêa, Lopes et Amancio, un problème non résolu (Correa Jr *et al.*, 2018). Pour plusieurs auteurs, il est même considéré comme « IA-complet » – en anglais *AI-complete* –, c'est-à-dire un problème qui, par analogie avec les problèmes NP-complets

en théorie de la complexité, est aussi difficile à résoudre que la création d'une véritable intelligence artificielle (Navigli, 2009; Yampolskiy, 2013).

Il n'existe pas de méthode fiable, permettant de désambiguïser à coup sûr le sens des mots dans une phrase. Toutefois, l'on remarque que les lexicologues ordonnent habituellement [3] les sens selon l'usage, du plus fréquent au moins fréquent : CIDE : Procter (1995, p. *ix*), LDOCE : van Sterkenburg (2003), WORDSMYTH :Wordsmyth (2017). Ainsi, en se basant simplement sur l'ordre des définitions, « l'heuristique du premier sens » donne généralement des résultats satisfaisants. Cette méthode constitue encore souvent un point de référence – en anglais *baseline* – difficile à surpasser : "The first sense heuristic [...] outperforms many of these systems which take surrounding context into account" (McCarthy *et al.*, 2004). Pour ces raisons, ainsi que pour fins de simplicité, nous utilisons dans ce travail l'heuristique du 1er sens comme méthode de désambiguïsation.

1.2 Définition formelle d'un lexique

Comme nous l'avons vu précédemment, un lexique peut être décrit d'un point de vue linguistique comme un ensemble de lexèmes accompagnés de leurs définitions et de toute autre information nécessaire à leur utilisation (Cruse, 2002).

Cependant, pour notre analyse, il nous faut pousser plus loin en termes de formalisme mathématique. En procédant par raffinements successifs, nous proposons dans cette section la définition formelle d'un *lexique complet.*

Définition 1.2.1 (Lexique). Un *lexique* est un quadruplet $X = (\mathcal{A}, \mathcal{P}, \mathcal{L}, \mathcal{D})$, où :

(i) $\mathcal{A}$ est un *alphabet*, dont les éléments sont appelés *lettres.*

3. Il y a toutefois des exceptions, par exemple le Merriam-Webster qui présente les sens selon l'ordre « historique » d'apparition des sens dans la langue anglaise, cf. : Merriam-Webster (2019b). Cette exception est aussi relevée par Lew (2013) : "[...] *the American dictionary publisher Merriam Webster's Incorporated has insisted on the application of the historical principle in its range of general dictionaries, including the popular Merriam-Webster's Collegiate Dictionary. This dictionary was found inferior for US college students compared with other dictionaries aimed at college students or advanced learners of English* ". Par ailleurs, Kipfer (1984) examine en détail la question de l'ordre des sens.

(*ii*) $\mathcal{P}$ = {N, V, A, R, S} est un ensemble non vide d'éléments appelés *parties du discours* – en anglais *part of speech (POS)*. Les éléments correspondent aux cinq *parties du discours* décrites à la section 1.1.2.

(*iii*) $\mathcal{L}$ est un ensemble fini de triplets $\ell = (w, i, p)$, appelés *lexèmes* et notés $\ell = w_p^i$, où $w \in A^*$ est un mot-forme, $i \geq 1$ est un entier, et $p \in \mathcal{P}$. Nous disons alors que (w, i, p) est le i-ème sens du mot-forme étiqueté (w, p) :

- S'il n'existe pas de $(w, i, p) \in \mathcal{L}$ avec $i > 1$, alors w_p^1 est alors dénoté simplement w_p et est dit *monosémique*. De plus, si tous les $(w, i, p) \in \mathcal{L}$ sont monosémiques, nous disons alors que X est *monosémique*.
- S'il existe un $(w, i, p) \in \mathcal{L}$ avec $i > 1$, nous disons que le mot-forme étiqueté (w, p) et le lexique X sont *polysémiques*.
- Pour rendre la numérotation cohérente, nous prenons comme hypothèse que si $(w, i, p) \in \mathcal{L}$ et que $i > 1$, alors $(w, i-1, p) \in \mathcal{L}$ aussi.
- Si $p = \text{S}$, alors $\ell = w_{\text{S}}^i$ est appelé un *lexème fonctionnel*.

(*iv*) $\mathcal{D}$ est une fonction qui associe à chaque lexème $\ell \in L$ une séquence finie $D(\ell) = (d_1^{(\ell)}, d_2^{(\ell)}, \ldots, d_k^{(\ell)})$, où $d_i^{(\ell)} \in A^*$ pour $i = 1, 2, \ldots, k$, appelée la *définition* de ℓ.

Nous voyons ci-après à l'exemple 11 un cas de lexique polysémique.

Exemple 11. Soit $X = (\mathcal{A}, \mathcal{P}, \mathcal{L}, \mathcal{D})$ un lexique tel que :

- $\mathcal{A} = \{a, b, \ldots, z\}$
- $\mathcal{P} = \{\text{N}\}$, où N indique que la partie du discours est un NOM,
- $\mathcal{L}$ et $\mathcal{D}$ sont tels que définis dans le tableau 1.1, à la page 23.

Définition 1.2.2 (Lexique lemmatisé)**.** Soit une fonction $lemma(w)$, qui associe à un mot-forme $w \in A^*$ sa forme canonique, son lemme. Si nous remplaçons dans la définition 1.2.1 (*iv*) $D(\ell) = (d_1^{(\ell)}, d_2^{(\ell)}, \ldots, d_k^{(\ell)})$ par $D(\ell) = (lemma(d_1^{(\ell)}), lemma(d_2^{(\ell)}), \ldots, lemma(d_k^{(\ell)}))$, alors $D(\ell)$ est appelée une *définition lemmatisée* de ℓ.

Nous disons alors que X est un *lexique lemmatisé*.

Tableau 1.1: Lexèmes et définitions d'un lexique polysémique

ℓ	$D(\ell)$
$\text{FRUIT}_{\text{N}}^{1}$	(*plant, part, that, has, seed, and, edible, flesh*)
$\text{FRUIT}_{\text{N}}^{2}$	(*the, result, of, work, or, action*)
$\text{FLESH}_{\text{N}}^{1}$	(*the, edible, part, of, a, fruit, or, vegetable*)
$\text{FLESH}_{\text{N}}^{2}$	(*the, part, of, an, animal, used, as, food*)
$\text{SEED}_{\text{N}}^{1}$	(*the, small, part, of, a, plant, from, which, a, new, plant, can, develop*)

Définition 1.2.3 (Lexique étiqueté)**.** Si nous remplaçons la condition $d_i^{(\ell)} \in A^*$ par $d_i^{(\ell)} \in \mathcal{A}^* \times \mathcal{P}$ dans la définition 1.2.2 (*iv*), alors $D(\ell)$ est appelée une *définition étiquetée* de ℓ.

Nous disons alors que X est un *lexique étiqueté.* L'exemple 12 illustre un tel lexique.

Exemple 12. Soit $X = (\mathcal{A}, \mathcal{P}, \mathcal{L}, \mathcal{D})$ un lexique tel que :

- $\mathcal{A} = \{a, b, \ldots, z\}$
- $\mathcal{P} = \{\text{N}, \text{V}, \text{S}\}$, où N $\rightarrow$ NOM, V $\rightarrow$ VERBE, S $\rightarrow$ STOP
- $\mathcal{L}$ et $\mathcal{D}$ sont tels que définis dans le tableau 1.2.

Tableau 1.2: Lexèmes et définitions d'un lexique étiqueté

ℓ	$D(\ell)$
HAVE_{V}	$(to_{\text{S}}, own_{\text{V}}, or_{\text{S}}, possess_{\text{V}})$
OWN_{V}	$(to_{\text{S}}, have_{\text{V}}, in_{\text{S}}, your_{\text{S}}, possession_{\text{N}})$
$\text{POSSESS}_{\text{V}}$	$(to_{\text{S}}, have_{\text{V}}, in_{\text{S}}, its_{\text{S}}, possession_{\text{N}}, to_{\text{S}}, own_{\text{V}})$
$\text{POSSESSION}_{\text{N}}$	$(having/have_{\text{V}}, or_{\text{S}}, owning/own_{\text{V}}, something_{\text{S}})$

Définition 1.2.4 (Lexique désambiguïsé)**.** Si nous remplaçons dans la définition 1.2.3 (*iv*) la condition par $d_i^{(\ell)} \in \mathcal{L}$, alors $D(\ell)$ est appelée une *définition désambiguïsée* de ℓ.

Nous disons que X est un *lexique désambiguïsé.*

Définition 1.2.5 (Lexique complet)**.** Finalement, si nous modifions la définition 1.2.3 pour y ajouter les conditions

(v) $\mathcal{L} = \{d_i^{(\ell)}\}$ où les $d_i^{(\ell)}$ ne sont pas des mots fonctionnels,

(vi) il existe un $D(\ell)$ pour tout $\ell \in \mathcal{L}$

nous disons alors que X est un *lexique complet.*

Exemple 13. Soit $X = (\mathcal{A}, \mathcal{P}, \mathcal{L}, \mathcal{D})$ un lexique tel que :

- $\mathcal{A} = \{a, b, \ldots, z\}$
- $\mathcal{P} = \{\text{N}, \text{V}, \text{S}\}$, où N $\rightarrow$ NOM, V $\rightarrow$ VERBE, S $\rightarrow$ STOP
- $\mathcal{L}$ et $\mathcal{D}$ sont tels que définis dans le tableau 1.3.

Tableau 1.3: Lexèmes d'un lexique complet

ℓ	$D(\ell)$
HAVE_V	$(to_\text{S}, \text{OWN}_\text{V}, or_\text{S}, \text{POSSESS}_\text{V})$
OWN_V	$(to_\text{S}, \text{HAVE}_\text{V}, in_\text{S}, your_\text{S}, \text{POSSESSION}_\text{N})$
POSSESS_V	$(to_\text{S}, \text{HAVE}_\text{V}, in_\text{S}, its_\text{S}, \text{POSSESSION}_\text{N}, to_\text{S}, \text{OWN}_\text{V})$
$\text{POSSESSION}_\text{N}$	$(having/\text{HAVE}_\text{V}, or_\text{S}, owning/\text{OWN}_\text{V}, something_\text{S})$

1.3 Graphes

Dans cette section, nous donnons un aperçu du modèle mathématique utilisé pour notre analyse de la structure des lexiques : la théorie des graphes. Mais tout d'abord, introduisons la notion de réseau sémantique.

Pour de nombreux auteurs spécialisés en intelligence artificielle, un réseau sémantique – en anglais *semantic network* – est une forme particulièrement utile de représentation des connaissances (Sowa, 2000; Hendler et van Harmelen, 2008; Russell *et al.*, 2010).

Lehmann en donne une définition très concise : "A semantic network is a graph of the structure of meaning" (Lehmann, 1992). Dans sa forme traditionnelle, un réseau sémantique représente des objets sous la forme de nœuds, connectés entre eux par des liens pouvant être étiquetés. La figure 1.3 offre un exemple de réseau sémantique simple. Les nœuds et les flèches y représentent un sous-ensemble d'une base de données d'associations libres (Nelson *et al.*, 1999). Dans l'étude en question, les auteurs demandaient à des participants, après leur avoir montré un mot amorce, de nommer le premier mot qui leur venait spontanément à l'esprit. Par exemple, dans le diagramme de la figure 1.3, le mot *volcano* est relié à *explode* par une flèche. Cela signifie que plusieurs participants ont associé spontanément le mot *explode* au mot *volcano* utilisé comme amorce.

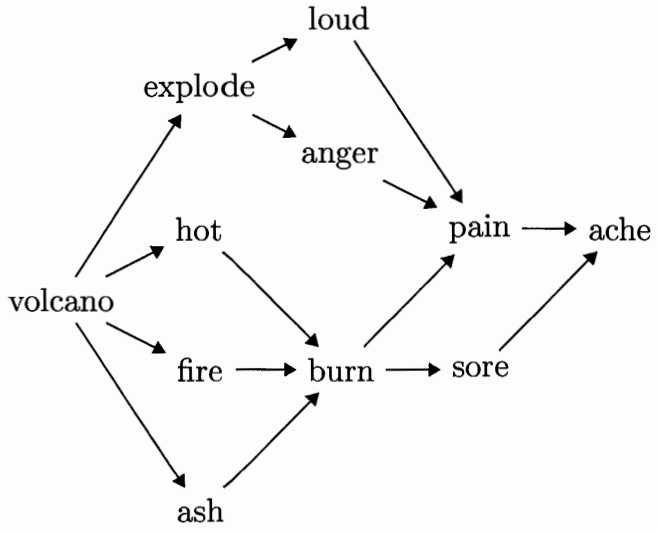


Fig. 1.3: Réseau d'association tiré de (Steyvers et Tenenbaum, 2005).

En utilisant le même genre de représentation, l'on peut facilement imaginer un lexique sous la forme d'un ~~graphique~~ graphe où les lexèmes sont illustrés par des nœuds et où les relations entre les lexèmes sont indiquées par des liens entre les nœuds. À titre d'exemple, reprenons la définition du lexème HAVE_V (exemple 13 à la page 24) :

$$D(\text{HAVE}_\text{V}) = (to_\text{S}, \text{OWN}_\text{V}, or_\text{S}, \text{POSSESS}_\text{V})$$

La figure 1.4 illustre le même lexème HAVE_V et sa définition sous la forme d'un réseau sémantique.

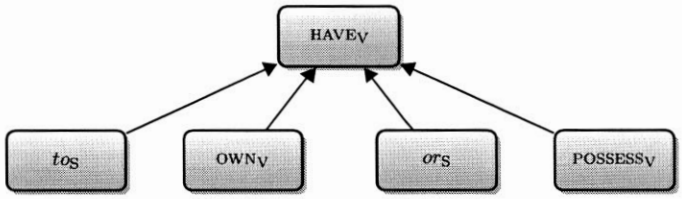


Fig. 1.4: Réseau sémantique représentant les relations définitionnelles

Cette forme de représentation est très semblable à la manière dont Bondy et Murty introduisent la notion de graphe (Bondy *et al.*, 1976, p. 1), i.e. : "[...] a diagram consisting of a set of points together with lines joining certain pairs of these points". Un graphe défini de cette manière ne pourrait toutefois pas représenter adéquatement un lexique. Une arête sans indication de direction dénote une relation entre deux lexèmes, mais elle ne permet pas de préciser le sens de cette relation. Autrement dit, il n'est pas possible de répondre à la question : Lequel des deux lexèmes est utilisé dans la définition de l'autre ? Il faut donc être plus précis et introduire la notion de graphe orienté.

Définition 1.3.1 (Graphe orienté). Un *graphe orienté* D est un couple (V, A) où :

(*i*) V est un ensemble fini de *sommets* – en anglais *vertices*,

(*ii*) $A \subseteq V \times V$ est un ensemble fini d'éléments appelés *arcs* – en anglais *edges*.

Remarque : Si $v_1, v_2 \in V$, alors $(v_1, v_2) \in A$ n'implique pas que $(v_2, v_1) \in A$.

Exemple 14. Soit un graphe orienté $D = (V, A)$, avec

$$V = \{v_1, v_2, v_3, v_4\},$$

$$A = \{(v_2, v_1), (v_3, v_1), (v_1, v_2), (v_3, v_2), (v_4, v_3), (v_2, v_3), (v_2, v_4), (v_1, v_4), (v_3, v_4)\}$$

La figure 1.5 illustre le graphe orienté D.

À partir de cette définition de graphe orienté, nous pouvons définir les notions reliées suivantes :

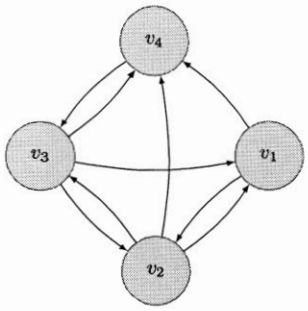


Fig. 1.5: Le graphe orienté D

degré

Soit $D = (V, A)$ un graphe orienté. Pour $u, v \in V$, u est un *prédécesseur* de v si $(u, v) \in A$. L'ensemble des prédécesseurs de v est représenté par $N^-(v)$. Le nombre de prédécesseurs de v est appelé *degré intérieur* – en anglais *in-degree* – de v, représenté par $\deg^-(v)$. De la même façon, nous disons que v est un *successeur* de u si $(u, v) \in A$ et que l'ensemble des successeurs de u est noté $N^+(u)$. Dans ce cas $\deg^+(u) = |N^+(u)|$ est appelé le *degré extérieur* de u.

circuit

$(v_1, v_2, \ldots, v_k) \in V^k$ est un *chemin* de D si $(v_i, v_{i+1}) \in A$ pour $i = 1, 2, \ldots, k-1$. Si en plus $v_1 = v_k$, alors p est appelé un *circuit*.

transversal de circuit

Un *transversal de circuit* de D, – en anglais *feedback vertex set* – (FVS), est un sous-ensemble $U \subseteq V$ de sommets tel que, pour tout circuit c de D, l'ensemble $U \cap c$ est non vide (Vazirani, 2006). C'est-à-dire que U couvre tous les circuits de D.

Le problème du calcul d'un *transversal de circuit* consiste à trouver dans un graphe un *transversal de circuit* minimal (MFVS). Pour un graphe général, c'est un problème NP-difficile, c'est-à-dire qu'il n'existe pas d'algorithme permettant de résoudre ce problème en temps polynomial à moins que P = NP (Karp, 1972). Cependant, en jumelant des opérateurs combinatoires et des techniques de pro-

grammation linéaire (Lin et Jou, 2000; Lapointe *et al.*, 2012), l'article de Vincent-Lamarre *et al.* (2016) a montré qu'il est possible de résoudre le problème pour les lexiques les moins volumineux et de trouver une bonne approximation pour les autres. Cela dit, nous n'allons pas plus avant avec cette question dans ce mémoire.

sous-graphe induit

$D' = (V', A')$ est un sous-graphe de D si $V' \subseteq V$ et $A' \subseteq A$.

De plus, nous disons que D' est un sous-graphe induit de D si pour tout couple $u \in V'$ et $v \in V'$, $(u, v) \in A \implies (u, v) \in A'$ (Diestel, 2000)

composante fortement connexe

Pour $u, v \in V$, soient les relations :

(i) $u \to v$ s'il existe un chemin de u vers v,

(ii) $u \leftrightarrow$ si $u \to v$ et $v \to u$.

Une *composante fortement connexe* (CFC) – en anglais *Strongly Connected Component* – est un sous-graphe de D induit par une classe d'équivalence de la relation $\leftrightarrow$ sur V.

En d'autres mots, lorsqu'il est possible de se déplacer d'un sommet u à un sommet v d'une composante fortement connexe, il est aussi possible d'aller en sens inverse du sommet v au sommet u (Blondin Massé *et al.*, 2008).

1.4 Lexiques et graphes associés

Les graphes orientés sont particulièrement indiqués pour représenter les relations entre les lexèmes d'un lexique. Pour notre analyse de la structure des lexiques, nous considérons uniquement les relations définitionnelles de la forme : le lexème ℓ « participe à la définition » du lexème ℓ'.

Nous représentons un lexique en utilisant les conventions suivantes :

- Les sommets du graphe correspondent aux lexèmes.
- Les arcs entre les sommets correspondent aux relations entre les lexèmes. Par exemple, si un arc va du sommet ℓ vers le sommet ℓ', cela signifie que le lexème ℓ

fait partie de la définition de ℓ'.

- Pour ce qui est des lexèmes fonctionnels – les *stop words* –, nous considérons que leur valeur lexicale est très faible comparée aux lexèmes faisant partie des autres parties du discours (nom, verbe, adjectif et adverbe). Nous ne les représentons donc pas dans les graphes associés et nous n'en tenons pas compte dans notre analyse. Cette façon de faire est très souvent utilisée en TAL (Jurafsky et Martin, 2009), en recherche d'information (RI) (Manning *et al.*, 2008), et en forage de données ("data mining") (Leskovec *et al.*, 2014).

De façon plus formelle, nous définissons un « graphe associé à un lexique » ou plus simplement, « graphe associé », de la façon suivante :

Définition 1.4.1 (Graphe associé)**.**

Soit $X = (\mathcal{A}, \mathcal{P}, \mathcal{L}, \mathcal{D})$ un lexique complet. Alors $G(X)$ est le *graphe associé* à X si :

(*i*) $G(X) = (V, A)$ est un graphe orienté

(*ii*) $V = \mathcal{L}$

(*iii*) Si $\ell \in D(\ell')$ et que ℓ n'est pas un lexème fonctionnel, alors $(\ell, \ell') \in A$

L'exemple 15 qui suit montre le graphe associé au lexique de très petite taille X_{petit}, formé de quatre sommets (quatre lexèmes) et de neuf arcs (neuf relations).

Exemple 15.

Soit $X_{petit} = (\mathcal{A}, \mathcal{P}, \mathcal{L}, \mathcal{D})$ un lexique complet où $\mathcal{L}$ et $\mathcal{D}$ sont décrits par le tableau 1.4. La figure 1.6 représente le graphe associé au lexique X_{petit}.

Tableau 1.4: Lexique complet X_{petit}

ℓ	$D(\ell)$
HAVE_V	$(to_\text{S}, \text{OWN}_\text{V}, or_\text{S}, \text{POSSESS}_\text{V})$
OWN_V	$(to_\text{S}, \text{HAVE}_\text{V}, in_\text{S}, your_\text{S}, \text{POSSESSION}_\text{N})$
POSSESS_V	$(to_\text{S}, \text{HAVE}_\text{V}, in_\text{S}, its_\text{S}, \text{POSSESSION}_\text{N}, to_\text{S}, \text{OWN}_\text{V})$
$\text{POSSESSION}_\text{N}$	$(having/\text{HAVE}_\text{V}, or_\text{S}, owning/\text{OWN}_\text{V}, something_\text{S})$

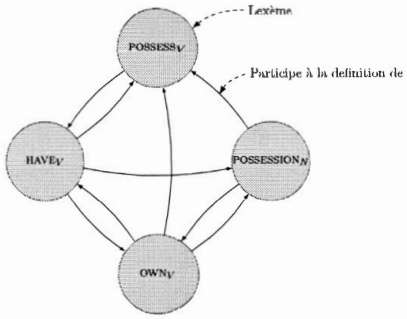


Fig. 1.6: Graphe associé au lexique X_{petit}

L'exemple 16 montre le graphe associé au lexique X_{gros} de plus grande taille. Il comprend 40 sommets et 123 arcs.

Exemple 16.

Soit $X_{gros} = (\mathcal{A}, \mathcal{P}, \mathcal{L}, \mathcal{D})$ un lexique où $\mathcal{L}$ et $\mathcal{D}$ sont décrits dans le tableau 1.5.

Tableau 1.5: Lexique complet X_{gros}

ℓ	$D(\ell)$
$\text{ACCOMPLISH}^1_{\text{V}}$	$(to_{\text{S}}, \text{SUCCEED}^1_{\text{V}}, in_{\text{S}}, doing/\text{DO}^1_{\text{V}}, \text{SOMETHING}^1_{\text{N}})$
$\text{ACTION}^1_{\text{N}}$	$(the_{\text{S}}, \text{PROCESS}^1_{\text{N}}, of_{\text{S}}, doing/\text{DO}^1_{\text{V}}, \text{SOMETHING}^1_{\text{N}})$
$\text{ACTION}^2_{\text{N}}$	$(a_{\text{S}}, \text{THING}^1_{\text{N}}, done/\text{DO}^1_{\text{V}})$
$\text{APART}^1_{\text{R}}$	$(in_{\text{S}}, a_{\text{S}}, \text{DIFFERENT}^1_{\text{A}}, \text{PLACE}^1_{\text{N}}, from_{\text{S}}, \text{SOMETHING}^1_{\text{N}}, \text{ELSE}^1_{\text{R}})$
$\text{CONDITION}^1_{\text{N}}$	$(the_{\text{S}}, \text{PARTICULAR}^1_{\text{A}}, \text{STATE}^1_{\text{N}}, that_{\text{S}}, \text{SOMETHING}^1_{\text{N}}, is_{\text{S}}, in_{\text{S}})$
$\text{DETAIL}^1_{\text{N}}$	$(a_{\text{S}}, \text{THING}^1_{\text{N}}, that_{\text{S}}, does/\text{DO}^1_{\text{V}}, \text{NOT}^1_{\text{R}}, \text{HAVE}^1_{\text{V}}, \text{IMPORTANCE}^1_{\text{N}})$
$\text{DIFFERENT}^1_{\text{A}}$	$(\text{NOT}^1_{\text{R}}, \text{LIKE}^1_{\text{A}}, \text{SOMETHING}^1_{\text{N}}, \text{ELSE}^1_{\text{R}})$
DO^1_{V}	$(to_{\text{S}}, \text{ACCOMPLISH}^1_{\text{V}}, \text{SOMETHING}^1_{\text{N}}, in_{\text{S}}, \text{PARTICULAR}^1_{\text{A}})$
ELSE^1_{R}	$(\text{APART}^1_{\text{R}}, from_{\text{S}}, or_{\text{S}}, \text{INSTEAD}^1_{\text{R}}, of_{\text{S}}, \text{SOMETHING}^1_{\text{N}})$
$\text{GENERAL}^1_{\text{A}}$	$(only_{\text{S}}, the_{\text{S}}, most_{\text{S}}, \text{IMPORTANT}^1_{\text{A}}, things/\text{THING}^1_{\text{N}}, about_{\text{S}}, \text{SOMETHING}^1_{\text{N}},$ $\text{NOT}^1_{\text{R}}, the_{\text{S}}, details/\text{DETAIL}^1_{\text{N}})$
HAVE^1_{V}	$(to_{\text{S}}, \text{OWN}^1_{\text{V}}, or_{\text{S}}, \text{POSSESS}^1_{\text{V}})$
$\text{IMPORTANCE}^1_{\text{N}}$	$(refers/\text{REFER}^1_{\text{V}}, to_{\text{S}}, \text{SOMETHING}^1_{\text{N}}, that_{\text{S}}, is_{\text{S}}, \text{IMPORTANT}^1_{\text{A}})$
$\text{IMPORTANT}^1_{\text{A}}$	$(qualifies/\text{QUALIFY}^1_{\text{V}}, a_{\text{S}}, \text{THING}^1_{\text{N}}, that_{\text{S}}, has/\text{HAVE}^1_{\text{V}}, \text{IMPORTANCE}^1_{\text{N}})$
$\text{INSTEAD}^1_{\text{R}}$	$(in_{\text{S}}, \text{PLACE}^1_{\text{N}}, of_{\text{S}}, \text{SOMETHING}^1_{\text{N}}, \text{ELSE}^1_{\text{R}})$
LIKE^1_{A}	$(qualifies/\text{QUALIFY}^1_{\text{V}}, \text{SOMETHING}^1_{\text{N}}, \text{SIMILAR}^1_{\text{A}}, to_{\text{S}}, \text{SOMETHING}^1_{\text{N}}, \text{ELSE}^1_{\text{R}})$
MAKE^1_{V}	$(to_{\text{S}}, \text{PRODUCE}^1_{\text{V}}, \text{SOMETHING}^1_{\text{N}})$
$\text{MENTION}^1_{\text{V}}$	$(to_{\text{S}}, \text{REFER}^1_{\text{V}}, to_{\text{S}})$
NEED^1_{V}	$(to_{\text{S}}, \text{HAVE}^1_{\text{V}}, to_{\text{S}}, \text{DO}^1_{\text{V}})$
$\text{OTHER}^1_{\text{A}}$	$(\text{APART}^1_{\text{R}}, from_{\text{S}}, \text{SOMETHING}^1_{\text{N}}, \text{ELSE}^1_{\text{R}})$
OWN^1_{V}	$(to_{\text{S}}, have/\text{HAVE}^1_{\text{V}}, in_{\text{S}}, your_{\text{S}}, \text{POSSESSION}^1_{\text{N}})$
$\text{PARTICULAR}^1_{\text{A}}$	$(qualifies/\text{QUALIFY}^1_{\text{V}}, the_{\text{S}}, \text{THING}^1_{\text{N}}, that_{\text{S}}, you_{\text{S}}, are_{\text{S}},$

	$referring/\text{REFER}^1_\text{V}, to_\text{S})$
PLACE^1_N	$(a_\text{S}, \text{PARTICULAR}^1_\text{A}, \text{POSITION}^1_\text{N})$
$\text{POSITION}^1_\text{N}$	$(the_\text{S}, \text{PLACE}^1_\text{N}, of_\text{S}, a_\text{S}, \text{THING}^1_\text{N})$
$\text{POSSESS}^1_\text{V}$	$(to_\text{S}, have/\text{HAVE}^1_\text{V}, in_\text{S}, its, \text{POSSESSION}^1_\text{N}, to_\text{S}, \text{OWN}^1_\text{V})$
$\text{POSSESSION}^1_\text{N}$	$(the_\text{S}, \text{STATE}^1_\text{N}, of_\text{S}, having/\text{HAVE}^1_\text{V}, or_\text{S}, owning/\text{OWN}^1_\text{V})$
$\text{PROCESS}^1_\text{N}$	$(\text{DIFFERENT}^1_\text{A}, actions/\text{ACTION}^1_\text{N}, needed/\text{NEED}^1_\text{V}, to_\text{S}, \text{PRODUCE}^1_\text{V})$
$\text{PROCESS}^1_\text{V}$	$(to_\text{S}, \text{DO}^1_\text{V}, actions/\text{ACTION}^1_\text{N}, to_\text{S}, \text{PRODUCE}^1_\text{V})$
$\text{PRODUCE}^1_\text{V}$	$(to_\text{S}, \text{MAKE}^1_\text{V}, or_\text{S}, \text{DO}^1_\text{V})$
$\text{QUALIFY}^1_\text{V}$	$(to_\text{S}, have/\text{HAVE}^1_\text{V}, all_\text{S}, the_\text{S}, qualities/\text{QUALITY}^1_\text{N}, needed/\text{NEED}^1_\text{V},$ $to_\text{S}, be_\text{S}, a_\text{S}, \text{PARTICULAR}^1_\text{A}, \text{THING}^1_\text{N})$
$\text{QUALITY}^1_\text{N}$	$(\text{SOMETHING}^1_\text{N}, that_\text{S}, makes/\text{MAKE}^1_\text{V}, one_\text{S}, \text{THING}^1_\text{N}, \text{DIFFERENT}^1_\text{A},$ $from_\text{S}, \text{OTHER}^1_\text{A}, things/\text{THING}^1_\text{N})$
REFER^1_V	$(to_\text{S}, \text{MENTION}^1_\text{V}, \text{SOMETHING}^1_\text{N}, \text{ELSE}^1_\text{R})$
$\text{SIMILAR}^1_\text{A}$	$(having/\text{HAVE}^1_\text{V}, \text{LIKE}^1_\text{A}, qualities/\text{QUALITY}^1_\text{N}, as_\text{S}, \text{SOMETHING}^1_\text{N}, \text{ELSE}^1_\text{R})$
$\text{SOMETHING}^1_\text{N}$	$(refers/\text{REFER}^1_\text{V}, to_\text{S}, a_\text{S}, \text{PARTICULAR}^1_\text{A}, \text{THING}^1_\text{N})$
STATE^1_N	$(the_\text{S}, \text{CONDITION}^1_\text{N}, in_\text{S}, \text{GENERAL}^1_\text{A})$
$\text{SUCCEED}^1_\text{V}$	$(to_\text{S}, have/\text{HAVE}^1_\text{V}, \text{SUCCESS}^1_\text{N})$
$\text{SUCCESS}^1_\text{N}$	$(when_\text{S}, \text{SOMETHING}^1_\text{N}, has/\text{HAVE}^1_\text{V}, been_\text{S}, accomplished/\text{ACCOMPLISH}^1_\text{V})$
THING^1_N	$(refers/\text{REFER}^1_\text{V}, to_\text{S}, \text{SOMETHING}^1_\text{N}, in_\text{S}, \text{GENERAL}^1_\text{A})$
THING^2_N	$(refers/\text{REFER}^1_\text{V}, to_\text{S}, my, possessions/\text{POSSESSION}^1_\text{N}, in_\text{S}, \text{GENERAL}^1_\text{A})$
NOT^1_R	$(to_\text{S}, \text{MAKE}^1_\text{V}, \text{NEGATIVE}^1_\text{A})$
$\text{NEGATIVE}^1_\text{A}$	$(qualifies/\text{QUALIFY}^1_\text{V}, \text{SOMETHING}^1_\text{N}, as_\text{S}, \text{NOT}^1_\text{R})$

La figure 1.7 représente le graphe associé au lexique X_{gros}.

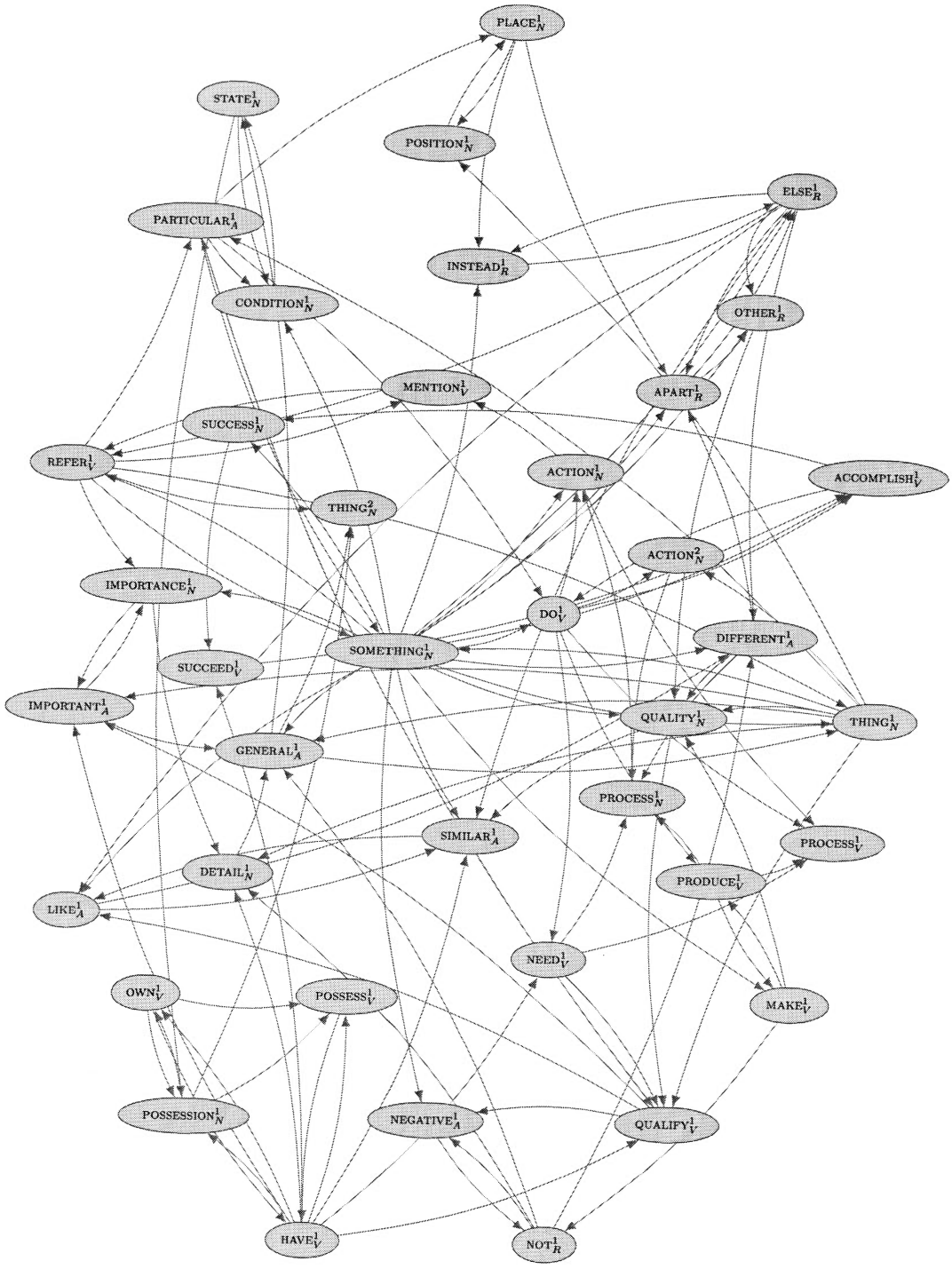


Fig. 1.7: Graphe associé au lexique X_{gros}

CHAPITRE II

STRATÉGIES D'APPRENTISSAGE

Dans ce chapitre, nous examinons les rapports qui existent entre l'apprentissage de nouveaux mots et la structure des dictionnaires.

Tout d'abord, nous précisons ce qui est sous-entendu par l'expression « apprendre de nouveaux mots ». Nous cherchons à comprendre de quelle façon l'on apprend à associer un signe linguistique lu ou entendu – en anglais *token* –, avec un sens. Pour ce faire, nous revenons brièvement sur la question de l'ancrage des symboles (Harnad, 1990). Puis nous introduisons une approche pédagogique traditionnelle visant à remédier à la difficulté soulevée par Harnad, c.-à-d. la construction de listes de mots.

Dans un deuxième temps, nous proposons un modèle de haut niveau pour représenter le processus d'apprentissage de nouveaux mots. Après avoir défini de façon formelle ce que signifie dans notre contexte « apprendre » un mot, nous décrivons différents algorithmes permettant de réaliser cet apprentissage.

2.1 Apprentissage de nouveaux mots

Dans cette section, nous abordons la question de l'acquisition du vocabulaire, en particulier dans le contexte de l'apprentissage d'une langue seconde. Tout d'abord, nous cherchons à identifier les principales difficultés rencontrées lorsque l'on utilise un dictionnaire monolingue pour apprendre le sens des nouveaux mots rencontrés.

2.1.1 Le problème de l'ancrage des symboles

Dans plusieurs articles traitant de ce sujet, Harnad analyse le problème de l'ancrage des symboles, le fameux *Symbol Grounding Problem* (Harnad, 1990, 2003, 2005). Sans trop entrer dans les détails de la linguistique et des sciences cognitives, l'on peut résumer cette question de la façon suivante : d'où provient le sens des mots ? Comment se fait-il qu'un mot que l'on connaît évoque habituellement quelque chose de précis ? Selon Harnad, c'est dû au fait que les mots sont ancrés de façon sensorimotrice :

> "*How are word meanings grounded ? Almost certainly in the sensorimotor capacity to pick out their referents.*" (Harnad, 2005)

Cependant, il est clair que l'apprentissage de nouveaux mots ne se déroule pas de la même façon chez un jeune enfant qui assimile les premiers rudiments de sa langue maternelle que chez un adulte qui étudie une nouvelle langue.

Lorsqu'un étudiant en langues rencontre un mot inconnu, une des façons pour lui de contourner la difficulté peut être de consulter un dictionnaire et d'y retrouver la définition du mot inconnu. Si tout se passe bien, la définition lui permet « d'apprendre » le nouveau mot et de le mémoriser. Illustrons cette situation à l'aide d'un exemple, tiré de (Harnad, 1990) :

(1) Supposons qu'un apprenant connaisse déjà bien le mot *horse*, qui est ancré symboliquement dans son expérience sensorimotrice. Il est en mesure d'identifier facilement un cheval s'il en rencontre un.

(2) Supposons que *striped* soit aussi connu de la même façon.

(3) Alors, avec la seule définition "*striped horse*", quelqu'un qui n'a jamais vu de zèbre est en mesure d'en identifier un s'il le voit. Il peut associer le symbole – le mot *zebra* – avec l'animal qui a l'air d'un cheval et qui est rayé.

Mais les choses se compliquent s'il y a dans la définition beaucoup de mots dont il ne connaît pas le sens. Dans l'article de Blondin Massé *et al.* (2008), les auteurs décrivent cette situation inconfortable où l'on tournerait sans fin dans le dictionnaire, en allant de mots inconnus vers d'autres mots inconnus, sans espoir d'arriver à une

quelconque compréhension des mots et de leurs définitions. Donc, pour que la définition d'un mot dans un dictionnaire soit compréhensible et utile, il faut absolument qu'un nombre suffisant de mots soient déjà « ancrés », c'est-à-dire qu'ils représentent autre chose que des formes abstraites sur du papier ou sur un écran.

Nous n'étudions pas plus avant la façon dont les mots sont ancrés dans l'expérience sensorimotrice. Nous constatons seulement que si les mots de la définition sont connus et suffisamment bien ancrés, cela permet d'apprendre un nouveau mot, de l'ancrer à son tour.

2.1.2 Ensemble d'ancrage minimal

Voyons maintenant de quelle façon la question de l'ancrage des symboles se traduit dans notre modèle formel de lexèmes, de lexique et de graphe associé.

Prenons comme hypothèse que l'on peut apprendre un nouveau lexème – un nouveau mot – seulement si nous connaissons déjà tous les lexèmes qui apparaissent dans sa définition. Nous pouvons alors définir un ensemble d'ancrage comme étant un sous-ensemble des lexèmes d'un lexique qui nous permet d'apprendre tous les autres lexèmes de ce lexique.

Définition 2.1.1 (Ensemble d'ancrage)**.** Soient :

(a) $X = (\mathcal{A}, \mathcal{P}, \mathcal{L}, \mathcal{D})$ un lexique complet,

(b) $G(X) = (V, A)$ son graphe associé,

(c) $U \subseteq V$ un sous-ensemble de V,

(d) L une fonction définie par $L(U) = U \cup \{v \in V \mid N^-(v) \subseteq U\}$.

S'il existe un $k \in \mathbb{Z}^+$ tel que $L^k(U) = V$, nous disons alors que U est un *ensemble d'ancrage* de X et que $\mathcal{L}$ est *k-atteignable* – en anglais *k-reachable*.

En d'autres mots, si U est un ensemble d'ancrage de X, cela signifie que l'on peut apprendre par définition tous les autres lexèmes de X, c'est-à-dire $V \setminus U$.

À partir de l'exemple du lexique X_{gros}, figure 1.7 à la page 33, voyons comment on

peut utiliser cette définition pour valider qu'un sous-ensemble des sommets d'un graphe constitue un ensemble d'ancrage.

Exemple 17.

Prenons comme sous-ensemble de départ $U = \{$ HAVE^1_V, PLACE^1_N, $\text{POSSESSION}^1_\text{N}$, $\text{QUALIFY}^1_\text{V}$, REFER^1_V, STATE^1_N, THING^1_N $\}$. En appliquant de façon récursive la fonction définie au point (d) de la définition précédente, nous obtenons :

$$
\begin{aligned}
L^0(U) &= U \\
L^1(U) &= L^0(U) \cup \{\text{PARTICULAR}^1_\text{A}, \text{POSITION}^1_\text{N}, \text{OWN}^1_\text{V}\} \\
L^2(U) &= L^1(U) \cup \{\text{POSSESS}^1_\text{V}, \text{SOMETHING}^1_\text{N}\} \\
L^3(U) &= L^2(U) \cup \{\text{CONDITION}^1_\text{N}\} \\
L^4(U) &= L^3(U)
\end{aligned}
$$

Dans le diagramme de la figure 2.1, les éléments des ensembles $L^0(U)$, $L^1(U)$, $L^2(U)$, $L^3(U)$, sont respectivement marqués avec les symboles ⓪, ①, ② et ③.

Par exemple, nous pouvons voir que le lexème OWN^1_V est *1-atteignable*, puisqu'il peut être appris à partir des lexèmes $\text{POSSESSION}^1_\text{N}$ et HAVE^1_V. De même, le lexème $\text{POSSESS}^1_\text{V}$ est *2-atteignable*, puisqu'il peut être appris à partir des lexèmes $\text{POSSESSION}^1_\text{N}$ et HAVE^1_V, et OWN^1_V.

De plus, puisque nous avons $L^4(U) = L^3(U)$, il n'est pas possible d'apprendre des lexèmes supplémentaires. U n'est donc pas un ensemble d'ancrage du lexique X_{gros}.

Blondin Massé *et al.* ont démontré qu'il existe aussi une correspondance exacte entre les ensembles d'ancrage d'un lexique X et les transversaux de circuit du graphe associé $G(X)$ (Blondin Massé *et al.*, 2008). Comme nous l'avons expliqué plus tôt à la section 1.3 page 27, le calcul d'un transversal de circuit minimal est, de façon générale, un problème NP-difficile. Il est cependant possible, comme le montre l'article de Vincent-Lamarre *et al.* (2016) d'utiliser des algorithmes et des techniques de programmation

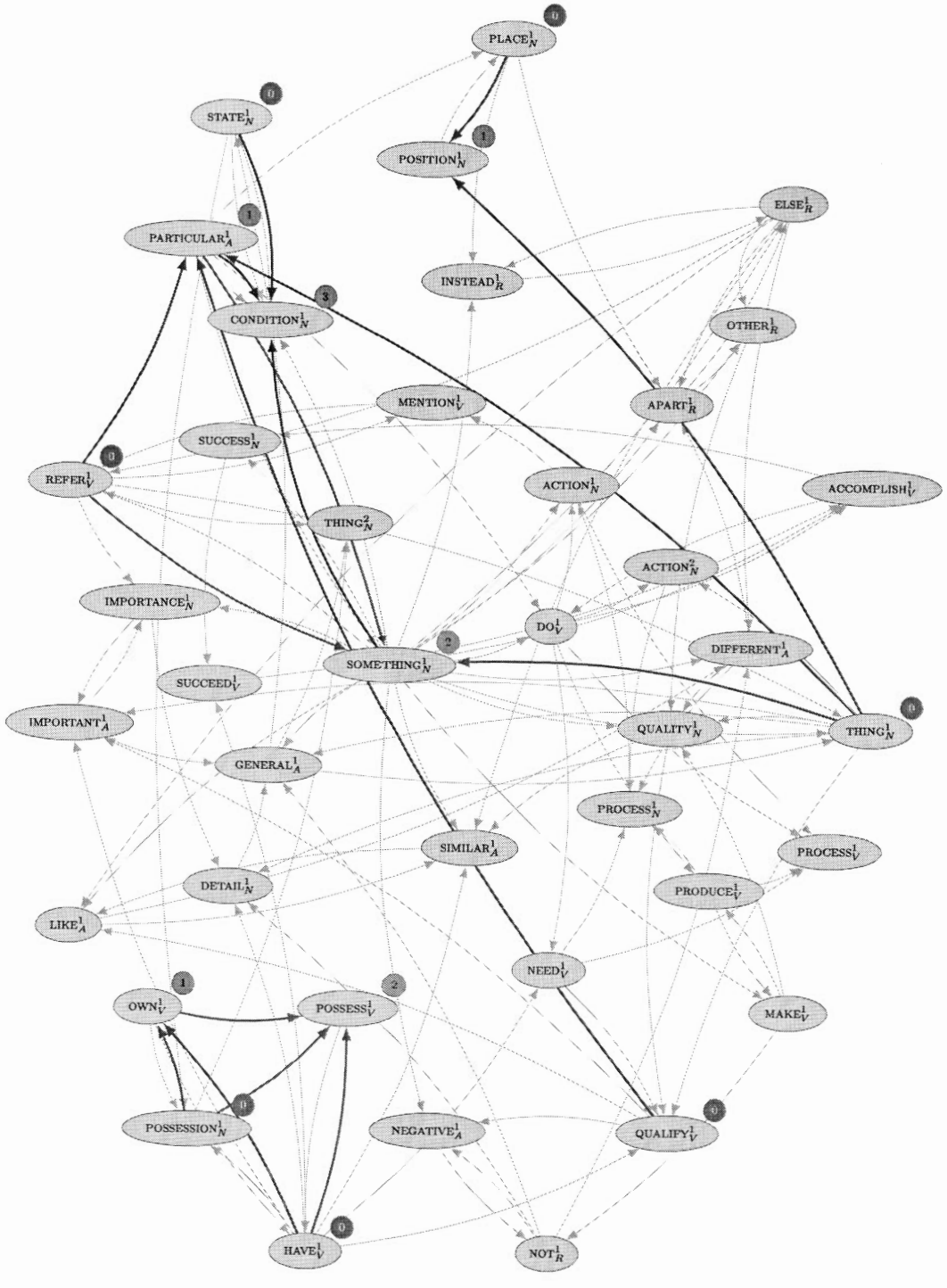


Fig. 2.1: Graphe associé au lexique X_{gros}

(Les lexèmes y sont marqués selon leur *k-atteignabilité* à partir de U).

linéaire permettant dans les meilleurs cas de calculer une solution exacte, ou à tout le moins, de trouver une approximation valable.

Poursuivons avec les mêmes exemples de lexiques complets présentés à la section 1.4, pages 29 et 31, afin d'illustrer le résultat du calcul des ensembles d'ancrage minimaux.

Exemple 18.

Pour le lexique trivial X_{petit} de la figure 2.2, il est facile de trouver par recherche exhaustive un ensemble d'ancrage minimal, par exemple :

{ HAVE^1_{V}, $\text{POSSESSION}^1_{\text{N}}$, }

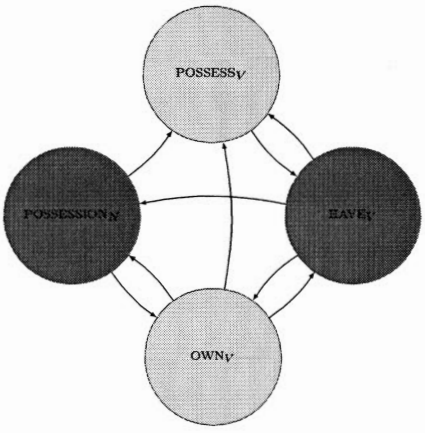


Fig. 2.2: Graphe associé au lexique X_{petit}
(Les lexèmes de l'ensemble d'ancrage minimal sont marqués en rouge).

Exemple 19.

Par contre, pour le lexique X_{gros}, de taille très restreinte par rapport à un dictionnaire réel, nous constatons rapidement que la méthode « manuelle » ne suffit pas pour trouver un ensemble d'ancrage minimal. La figure 2.3 illustre le résultat obtenu en utilisant la méthode décrite dans Vincent-Lamarre *et al.* (2016). L'ensemble d'ancrage minimal trouvé contient les lexèmes :

{ $\text{ACCOMPLISH}^1_{\text{V}}$, HAVE^1_{V}, $\text{IMPORTANT}^1_{\text{A}}$, LIKE^1_{A}, MAKE^1_{V}, $\text{PLACE}^1_{\text{N}}$, $\text{POSSESSION}^1_{\text{N}}$,
$\text{QUALIFY}^1_{\text{V}}$, $\text{REFER}^1_{\text{V}}$, $\text{STATE}^1_{\text{N}}$, $\text{THING}^1_{\text{N}}$, NOT^1_{R}, ELSE^1_{R} }

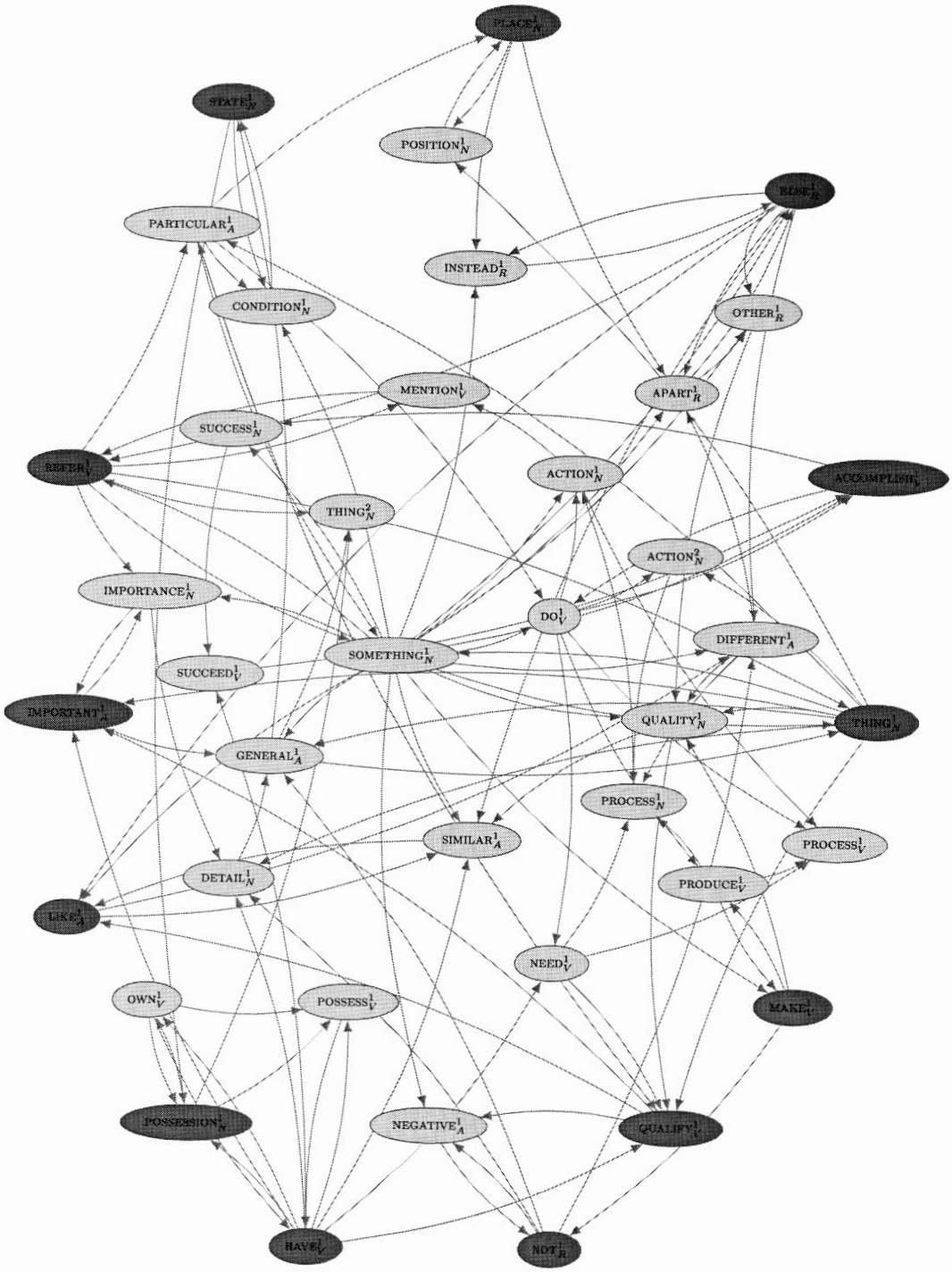


Fig. 2.3: Graphe associé au lexique X_{gros}

(Les lexèmes de l'ensemble d'ancrage minimal sont marqués en rouge).

2.1.3 Les listes de mots

Observons maintenant comment les concepts précédents d'ancrage symbolique et d'ensemble d'ancrage minimal peuvent être rapprochés de techniques utilisées pour l'enseignement des langues.

L'importance accordée à l'enseignement du vocabulaire dans les classes de langue seconde a varié au cours des ans, suivant en cela l'évolution des théories et des approches en didactique des langues (Monge, 2013). Mais il n'en reste pas moins que pour les étudiants, l'acquisition d'un large vocabulaire demeure essentielle pour la maîtrise d'une langue. Les professeurs et les chercheurs dans ce domaine ont donc cherché depuis longtemps les meilleurs moyens pour faciliter à leurs étudiants l'apprentissage de nouveaux mots. Dans ce contexte, l'on peut comprendre leur intérêt pour les *listes de mots* – en anglais *word lists* –.

Les listes de mots sont des regroupements de mots représentatifs d'une langue ou d'un domaine spécialisé que les étudiants doivent maîtriser le plus tôt possible afin de devenir autonomes dans leur étude. Ils disposent alors d'une base de mots connus – ancrés – qui leur permet d'utiliser de manière indépendante les dictionnaires et les divers autres outils d'aide à l'apprentissage. Selon Nation, "Word lists lie at the heart of good vocabulary course design" (Nation, 2016).

Dès les années 1930, Charles Ogden a introduit son "Basic English", une version de l'anglais possédant une grammaire et un vocabulaire simplifié (Ogden, 1930). Le "Basic English" devait selon Ogden devenir une langue universelle, un peu comme l'esperanto. Plusieurs listes différentes, comprenant entre 850 et 2000 mots, ont par la suite été construites (Ogden, 2019) afin de faciliter l'apprentissage de cet « anglais de base ».

Dans les années 1950, West a proposé sa "General Service List" (GSL), contenant environ 2000 mots fréquemment utilisés en anglais (West, 1953). La GSL a été pendant de longues années une référence incontournable : "There has been no comparable replacement for the GSL up to now." (Coxhead, 2000).

Récemment, Brezina et Gablasova (2013) et Browne (2014) ont proposé des versions améliorées de la liste de West, nommées dans les deux cas "New General Service List" (NGSL). Browne ajoute aussi, en plus de la NGSL, 3 listes complémentaires (Browne *et al.*, 2019b) :

- La "New Academic Word List" (NAWL) ;
- La "TOEIC Service List" (TSL) ;
- La "Business Service Lists" (BSL).

Mais comment ces listes sont-elles construites ? Dans les cas recensés plus haut, elles ont toutes été créées à partir de corpus. Grosso modo, la méthode la plus utilisée consiste à comptabiliser la fréquence relative des mots apparaissant dans une collection de documents pertinents, puis à agencer ces mots dans une *liste de mots* selon leur fréquence et leur importance pour l'auteur de la liste. Dans un ouvrage récent, Nation (2016) présente même une description détaillée de ces techniques de construction de listes à partir de corpus.

Dans ce mémoire, nous introduisons une approche différente, selon nous inédite. Elle consiste à utiliser un dictionnaire, ou plus précisément un lexique, pour construire de façon algorithmique des listes de mots. Pour ce faire, nous développons tout d'abord une représentation d'un lexique sous forme de graphe orienté. Puis, nous utilisons des algorithmes propres à la théorie des graphes pour construire des listes de mots qui permettent « d'apprendre » l'ensemble des mots du lexique.

Toutefois, étant donné le cadre restreint de notre mémoire, cette nouvelle façon de faire n'est pas complètement développée et demeure encore relativement conceptuelle. Elle fait entre autres complètement abstraction des critères linguistiques et des contraintes supplémentaires dont les spécialistes du domaine doivent tenir compte afin de faciliter l'apprentissage d'une langue.

2.2 Modèle d'apprentissage

Dans un article de Picard *et al.* (2010), les auteurs mettent de l'avant l'hypothèse qu'il existe deux façons d'apprendre de nouveaux mots ou de nouveaux sens lexicaux : par instruction verbale – en anglais *verbal instruction* – et par expérience sensorimotrice – en anglais *direct sensorimotor induction* –.

Nous nous appuyons sur cette prémisse pour fonder notre *modèle formel d'apprentissage.* Nous disons qu'un nouveau lexème peut être appris de deux manières :

Par apprentissage direct : Avec cette approche, le lexème et le sens lexical sont ancrés directement par expérience sensorimotrice; par exemple, lors d'une visite à la ferme quelqu'un explique à un enfant que l'animal en face de lui s'appelle un cheval.

Pour garder notre modèle simple, nous ne nous préoccupons pas de la façon dont ce lien s'établit ni de ce qui est mis en œuvre au niveau mental et sensorimoteur. Nous nous en tenons au fait que c'est en général une opération complexe, qui requiert souvent l'intervention d'une tierce personne pour expliquer ou montrer ce dont il est question : il faut sortir de « l'univers des mots ». Nous considérons que c'est donc un processus relativement coûteux.

Par définition : Dans ce cas, l'on utilise une source d'information lexicale pour établir le lien entre le sens et le lexème; par exemple, un étudiant consulte un dictionnaire pour y retrouver la définition de *zebra* : "*striped horse*". Nous prenons comme hypothèse que cette forme d'apprentissage est comparativement peu coûteuse par rapport à l'apprentissage direct. Elle ne nécessite pas de déplacement de personnes ou d'objets, ni de participation d'un tiers pour fournir des explications; l'on reste dans la sphère exclusive des mots et du sens. Néanmoins, pour éviter de tomber dans le problème de l'ancrage des symboles, nous posons comme condition qu'un lexème ne peut être appris par définition que lorsqu'il est *complètement défini*, c'est-à-dire que tous les lexèmes qui forment sa définition sont déjà connus. Pour notre modèle, la source de donnée utilisée est un lexique monolingue.

L'on pourrait argumenter qu'il est intuitivement plus simple d'apprendre le nom d'un objet à partir d'une illustration qu'en cherchant une définition dans un dictionnaire. Reprenons l'exemple du cheval et du zèbre :

(a) Le sujet connait déjà les mots cheval et rayure

(b) Pour apprendre le nouveau mot zèbre à partir de l'image d'un animal qui ressemble à un cheval mais qui a des rayures, il faut que quelqu'un lui montre l'image et lui dise « C'est un zèbre », ou qu'il retrouve lui-même le mot zèbre dans un dictionnaire illustré.

Mais cette approche ne peut fonctionner que pour les noms de choses concrètes et tangibles. Elle n'est pas utile pour apprendre le nom des concepts plus abstraits. La simplicité n'est donc qu'apparente. Il faut nécessairement se rabattre sur les explications verbales, revenir aux domaine des *mots*.

Avec notre modèle, nous nous fixons l'*objectif d'apprentissage* suivant : en partant d'une situation initiale où l'on ne connaît le sens d'aucun lexème, en arriver à connaître le sens de tous les lexèmes d'un lexique. Pour ce faire, le *processus d'apprentissage* consiste à apprendre les lexèmes un par un selon le déroulement suivant :

1. S'il y a dans le lexique un lexème inconnu, mais dont tous les éléments de la définition sont déjà connus, l'apprendre *par définition*
2. Sinon, apprendre de façon directe le prochain lexème indiqué par une *stratégie d'apprentissage*.
3. Répéter les étapes précédentes jusqu'à ce que le lexique au complet soit appris.

2.2.1 Stratégie d'apprentissage

Voyons maintenant la définition formelle d'une stratégie d'apprentissage.

Définition 2.2.1 (Stratégie d'apprentissage). Soit $X = (\mathcal{A}, \mathcal{P}, \mathcal{L}, \mathcal{D})$ un lexique complet.

(*i*) Une *stratégie d'apprentissage* S est une séquence ordonnée d'éléments de $\mathcal{L}$.

(*ii*) Si $\mathcal{S} \subseteq \mathcal{L}$ l'ensemble associé à S est un ensemble d'ancrage de X, nous disons alors que S est *exhaustive*.

(*iii*) Sinon, nous disons que S est *non exhaustive*.

En d'autres mots, une stratégie d'apprentissage pour un lexique correspond tout simplement à une liste de lexèmes de ce lexique triés dans l'ordre selon lequel on doit les apprendre. Comme nous le verrons au chapitre suivant, cette liste peut être dérivée d'une liste de mots externe, par exemple la liste de fréquence d'utilisation de Brysbaert et New (2009), ou elle peut être déterminée à l'aide d'un algorithmique. C'est une stratégie exhaustive si elle nous permet d'apprendre tous les lexèmes du lexique.

En tenant compte des deux façons d'apprendre décrites précédemment nous constatons intuitivement que l'effort d'apprentissage lié à une stratégie sera minimisé si cette stratégie fait en sorte qu'il soit nécessaire d'apprendre de façon directe le moins de lexèmes possible. Sans perte de généralité, nous posons comme hypothèse que le coût pour apprendre un lexème de façon directe est égal à 1 et qu'il est égal à 0 pour un apprentissage par définition. Nous disons alors que la stratégie S_1 est plus efficace que la stratégie S_2 si S_1 nous permet d'apprendre complètement le lexique à moindre coût que S_2.

2.2.2 Algorithmes d'apprentissage

À partir de notre modèle d'apprentissage, nous pouvons maintenant énoncer trois algorithmes qui vont nous permettre de calculer le coût d'une stratégie d'apprentissage et de déterminer si elle est exhaustive.

Algorithme 1 : Coût d'une stratégie d'apprentissage

La fonction COÛT calcule le coût associé à l'utilisation d'une stratégie d'apprentissage S pour apprendre complètement ou partiellement les lexèmes d'un lexique X. Comme nous l'avons vu précédemment, les stratégies peuvent être :

- *exhaustives* : elles permettent d'apprendre tous les lexèmes du lexique. Le coût calculé par la fonction COÛT correspond alors au coût total d'apprentissage du lexique.
- *non exhaustives* : dans ce cas, la fonction calcule uniquement le coût partiel

associé aux lexèmes appris et retourne la portion restante du lexique qui correspond aux lexèmes n'ayant pu être appris.

Algorithme 1

```
1: fonction CoûT(S : stratégie, X : lexique) : (coût, lexique)
2:     coût ← 0
3:     tant que S ≠ ∅ and X ≠ ∅ faire
4:         ℓ ← S.POP()                          ▷ Obtenir le prochain lexème
5:         si ℓ ∈ X alors                       ▷ Présent dans X ?
6:             coût ← coût + 1                  ▷ Apprendre ℓ à coût 1
7:             Retirer ℓ de X
8:             tant que il existe un ℓ' ∈ X avec deg⁻(ℓ') = 0 faire
9:                 Retirer ℓ' de X              ▷ Apprendre ℓ' à coût 0
10:            fin tant que
11:        fin si
12:    fin tant que
13:    retourner (coût, X)
14: fin fonction
```

La fonction retourne comme résultat le couple $(coût, X')$, où :

- *coût* est le coût occasionné par la stratégie S pour apprendre complètement ou partiellement X
- X' est la portion restante de X qui n'a pu être apprise avec S. X' peut être utilisé de la façon suivante afin de déterminer si S est *exhaustive* :
  - Si le lexique X' est vide, alors la stratégie S est *exhaustive* et la variable *coût* correspond au coût total.
  - Si le lexique X' n'est pas vide, alors la stratégie S est *non exhaustive*. Il faut alors utiliser une stratégie de repli pour compléter l'apprentissage du lexique et obtenir le coût total.

Analysons maintenant l'algorithme afin d'évaluer sa complexité :

- Structures de données :
  - S est stocké dans une liste de taille $s \leq n$
  - Le graphe associé à X est stocké dans une liste d'adjacence.
  - $n =$ le nombre de sommets de X.
  - $m =$ le nombre d'arcs de X.
- Complexité de temps

  Si nous prenons comme hypothèses que :
  - le retrait d'un sommet se fait en $\mathcal{O}(1)$,
  - à la ligne 8, seulement les voisins des sommets retirés sont pris en compte,

  la complexité de temps est alors de $\mathcal{O}(n+m)$
- Complexité d'espace

  La complexité d'espace requise pour stocker la liste d'adjacence est de $\mathcal{O}(n+m)$

Algorithme 2 : Coût d'apprentissage par degré dynamique

La fonction CoûtDegréDynamique permet de calculer le coût d'apprentissage de tous les lexèmes d'un lexique en se basant sur le calcul dynamique du degré extérieur maximal. À chaque itération, le lexème choisi est celui ayant le degré extérieur le plus élevé parmi les sommets restants du graphe. Autrement dit, le prochain lexème à apprendre de façon directe sera celui qui apparaît dans le plus grand nombre de définitions des lexèmes restants.

Algorithme 2

```
1: fonction CoûtDegréDynamique(X : lexique) : (coût)
2:     coût ← 0
3:     Ordonner les sommets de X selon leur degré extérieur
4:     tant que X ≠ ∅ faire
5:         ℓ ← lexème de X dont le degré extérieur est maximal
6:         coût ← coût + 1
```

```
7:        Retirer ℓ de X                          ▷ Apprendre ℓ à coût 1
8:        tant que il existe un ℓ' ∈ X avec deg⁻(ℓ') = 0 faire
9:            Retirer ℓ' de X                     ▷ Apprendre tous les ℓ' à coût 0
10:       fin tant que
11:   fin tant que
12:   retourner (coût)
13: fin fonction
```

La fonction retourne comme résultat *coût*, qui correspond au coût d'apprentissage par degré dynamique de tous les lexèmes du lexique X reçu en paramètre.

Analysons l'algorithme afin d'évaluer sa complexité :

- Structures de données :
  - Le graphe associé à X est stocké dans une liste d'adjacence.
  - $n =$ le nombre de sommets de X.
  - $m =$ le nombre d'arcs de X.
  - Les sommets de X sont ordonnés par degré extérieur avec une file à priorité implantée à l'aide d'un monceau – en anglais *heap* –.
- Complexité de temps

  Si nous prenons comme hypothèses que :
  - à la ligne 3, la construction de la file à priorité est faite en $\mathcal{O}(n \log n)$
  - à la ligne 5, l'extraction de la liste de priorité est réalisée en $\mathcal{O}(\log n)$.
  - à la ligne 8, seulement les voisins des sommets retirés sont pris en compte,
  - Aux lignes 7 et 9, le retrait d'un sommet se fait en $\mathcal{O}(1)$ et la mise à jour de la liste de priorité est faite en $\mathcal{O}(\log n)$ au pire m fois.

  la complexité de temps est alors de $\mathcal{O}(m \log n)$.
- Complexité d'espace

  La complexité d'espace requise pour stocker la liste d'adjacence est de $\mathcal{O}(n + m)$ et de $\mathcal{O}(n)$ pour la file de priorité, ce qui donne au final $\mathcal{O}(n + m)$

Algorithme 3 : Calcul du coût total d'une stratégie d'apprentissage

La fonction CoûtTotal calcule le *coût total* encouru pour « apprendre » tous les lexèmes du lexique X à l'aide de la stratégie S.

Pour une stratégie exhaustive, le coût total obtenu avec la fonction CoûtTotal est identique à celui obtenu avec la fonction Coût (voir l'algorithme 1). Par contre, pour une stratégie non exhaustive, le coût total correspond à la somme du coût de la stratégie S, plus le coût encouru en appliquant à la portion restante X' du lexique une *stratégie de repli*[1] – en anglais *fallback strategy*.

Algorithme 3

1: **fonction** CoûtTotal(S : stratégie, X : lexique) : coût
2: $(coût, X') \leftarrow$ Coût(S, X)
3: $(coût\ total) \leftarrow coût +$ CoûtDegréDynamique(X')
4: **retourner** *coût total*
5: **fin fonction**

Étant donné les valeurs obtenues pour les algorithmes précédents, la complexité de temps pour l'algorithme 3 est par conséquent de $\mathcal{O}(m \log n)$ et la complexité d'espace de $\mathcal{O}(n + m)$.

1. Il est théoriquement possible de concevoir différentes méthodes qui pourraient être utilisées comme stratégie de repli. Dans le cas présent, nous utilisons la méthode de calcul par degré dynamique CoûtDegréDynamique (voir l'algorithme 2).

CHAPITRE III

DONNÉES D'EXPÉRIMENTATIONS

Dans ce chapitre, nous présentons les données sources à partir desquelles nous avons réalisé notre étude de la structure des dictionnaires. Tout d'abord, nous décrivons les différents dictionnaires numériques ayant servi à construire les lexiques et les graphes associés. Ce sont tous des ouvrages produits par des lexicographes professionnels et publiés en format électronique. Puis, nous examinons les différentes stratégies d'apprentissage développées pour « apprendre » les mots des dictionnaires. Elles sont de deux types :

- les *stratégies psycholinguistiques*, bâties à partir de listes de mots spécialement étiquetés, appelées *normes psycholinguistiques*,
- les *stratégies algorithmiques*, obtenues en analysant la structure des graphes associés aux lexiques.

3.1 Les dictionnaires numériques

Comme base pour notre analyse de la structure des lexiques, nous avons utilisé des dictionnaires monolingues de langue anglaise. Ce sont des documents élaborés par des linguistes professionnels, disponibles en format numérique ou papier, à l'exception de l'un d'entre eux qui est disponible seulement sur le web.

Le "Cambridge International Dictionary of English", ou CIDE, est un dictionnaire de langue anglaise, développé pour les étudiants d'anglais langue seconde (Procter, 1995). Dans la version que nous avons utilisée, il comprend grosso modo 19 000 articles et 47 000 lexèmes différents.

Le "Longman Dictionary of Contemporary English", ou LDOCE, est un dictionnaire de niveau avancé pour des étudiants d'anglais langue seconde. Il a été publié pour la première fois en 1978 (Procter, 1978). Il comprend environ 29 000 articles et 70 000 lexèmes.

Ces dictionnaires, le CIDE et le LDOCE, possèdent une caractéristique commune (Black, 1997; Longman, 2019). Ce sont tous deux des *'monolingual learners' dictionaries* (MLD), des dictionnaires développés spécialement pour les besoins des étudiants d'une langue seconde, dans ce cas-ci l'anglais (Brown et Anderson, 2006, p. 739, Rundell). Chacun d'entre eux a été construit à partir de son propre vocabulaire de contrôle. Autrement dit, toutes les définitions n'utilisent que des mots provenant d'un vocabulaire restreint, facilitant ainsi la compréhension des définitions pour un utilisateur novice. Dans les deux cas, le vocabulaire de contrôle est de l'ordre de 2 000 lexèmes.

Le "Merriam-Webster's Collegiate Dictionary", ou MWC, est le dictionnaire le plus volumineux que nous avons étudié (Merriam-Webster, 2003). La 11e édition comprend plus de 250 000 lexèmes, regroupés en 70 000 articles.

WordNet (WN), quant à lui, n'est pas un dictionnaire au sens propre du terme. C'est plutôt une base de données lexicale de l'anglais (Brown et Anderson, 2006, p. 665, Fellbaum). Il regroupe en *synsets* les différents lexèmes faisant référence à un même « sens » et à une même définition, ou *gloss*. Les *synsets* sont connectés entre eux par diverses relations sémantiques telles l'hyponymie, l'hyperonymie, etc.. La version que nous avons utilisée, WordNet 3.0, contient environ 132 000 lexèmes regroupés en 57 000 *synsets*.

Wordsmyth est, selon ses auteurs, à la fois un dictionnaire et un *thesaurus* (Wordsmyth, 2017). Contrairement à CIDE et LDOCE, il n'utilise pas un vocabulaire de contrôle. Il présente toutefois la particularité d'offrir en plus de la définition d'un mot donné, de l'information sur ses synonymes, ses antonymes et les mots semblables (Wordsmyth, 2017). Il se décline en quatre versions :

- Le "Wordsmyth Educational Dictionary-Thesaurus" (WEDT), le plus complet, a

été développé le premier dans les années 1980. Il comprend 73 000 lexèmes.

- Le "Wordsmyth Illustrated Learner's Dictionary" (WILD) est un dictionnaire illustré pour enfants. Il comprend 4 200 lexèmes.
- Le "Wordsmyth Learner's Dictionary-Thesaurus" (WLDT) est un dictionnaire de niveau intermédiaire. Il comprend 6 000 lexèmes.
- Le "Wordsmyth Children's Dictionary-Thesaurus" (WCDT) est un dictionnaire pour débutants. Il comprend 20 000 lexèmes.

À l'aide d'une série de prétraitements, nous avons transformé tous ces dictionnaires numériques, huit au total, en lexiques désambiguïsés et complets. Pour ce faire, nous avons tout d'abord extrait des dictionnaires uniquement les mots correspondant aux parties du discours considérées : • **nom**, • **verbe**, • **adjectif** et • **adverbe**, en ignorant les **mots fonctionnels**. À l'intérieur des définitions, nous avons encore une fois éliminé les mots fonctionnels, puis lemmatisé et étiqueté les lexèmes restants avec le "Stanford POS-tagger" (Toutanova *et al.*, 2003) Finalement, nous avons désambiguïsé les lexèmes en appliquant l'heuristique du 1er sens.

Le tableau 3.1 présente diverses données statistiques pour les 8 lexiques obtenus :

- Le nombre de lexèmes dans chaque dictionnaire (Lexèmes).
- Le nombre de lemmes (Lemmes).
- Le degré moyen de polysémie, qui est le nombre moyen de lexèmes par lemme (Polysémie).
- Le nombre de lexèmes utilisés dans les définitions (Lexèmes utilisés).
- Le ratio du nombre de lexèmes utilisés dans les définitions par rapport au nombre total de lexèmes (Ratio d'utilisation).

Tableau 3.1: Données statistiques sur les lexiques

Lexique	Lexèmes	Lemmes	Polysémie	Lexèmes utilisés	Ratio d'utilisation
WILD	4 244	3 081	1.377	2 995	0.972
WLDT	6 036	3 433	1.758	2 212	0.644
WCDT	20 128	9 303	2.164	6 597	0.709
CIDE	47 092	18 694	2.519	8 773	0.469
LDOCE	69 204	22 511	3.074	10 074	0.448
WEDT	73 091	28 986	2.522	18 197	0.628
WN	132 547	57 243	2.316	29 600	0.517
MWC	249 137	68 181	3.654	33 533	0.492

Après avoir construit les graphes associés à ces lexiques, nous avons ensuite analysé leur structure. Il existe de nombreuses mesures pouvant être appliquées à des réseaux ou à des graphes. Batagelj et al., entre autres, recense une série de mesures spécialement adaptées aux graphes de dictionnaires (Batagelj *et al.*, 2002). Pour notre analyse, nous avons retenu celles qui permettent de brosser un portrait des graphes. Le tableau 3.2 montre les résultats obtenus à partir des graphes associés aux 8 lexiques :

- Le nombre de sommets (Sommets).
- Le nombre d'arcs (Arcs).
- Le nombre de composantes fortement connexes (CFCs).
- Le nombre de lexèmes contenus dans la CFC la plus volumineuse (<CFC).
- Le diamètre de la CFC la plus volumineuse (Diamètre).

  Le diamètre d'une composante fortement connexe correspond au nombre maximum de sommets que l'on doit traverser pour aller d'un sommet, cf : "a graph's diameter is the largest number of vertices which must be traversed in order to travel from one vertex to another " (Weisstein, 2003, WolframMathWorld).
- Le nombre moyen d'arcs par sommet (Arcs/Sommets)

- La longueur moyenne des plus courts chemins, – en anglais *Average shortest path length* – ou – *Characteristic Path Length* – (LMPCC), calculée pour un graphe $G = (V, E)$ avec la formule (Hagberg *et al.*, 2008, Networkx) :

$$\sum_{u,v \in V} \frac{d(u,v)}{|V|(|V|-1)}$$

Tableau 3.2: Données structurelles des graphes associés aux lexiques

Lexique	Sommets	Arcs	CFCs	<CFC	Diamètre	#Arcs / #Somm.	LMPCC
WILD	4 244	45 789	2 750	1 446	17	10.79	1.75
WLDT	6 036	28 623	5 088	858	25	4.74	1.10
WCDT	20 128	102 657	17 551	2 341	22	5.10	0.87
CIDE	47 092	334 888	45 306	1 702	16	7.11	0.21
LDOCE	69 204	415 052	67 224	1 770	16	6.00	0.16
WEDT	73 091	362 569	67 318	5 056	29	4.96	0.61
WN	132 547	694 067	124 589	7 079	30	5.24	0.50
MWC	249 137	1 155 085	239 478	8 842	29	4.64	0.31

3.2 Les stratégies d'apprentissage

L'on peut théoriquement imaginer un très grand nombre de stratégies pour apprendre tous les mots d'un dictionnaire ou d'un lexique. L'on pourrait par exemple essayer toutes les permutations possibles des lexèmes. Si un lexique contient n lexèmes, il y aurait alors $n!$ façons différentes d'ordonner ces lexèmes pour déterminer l'ordre d'apprentissage. Sauf pour les cas triviaux, il est évidemment impossible d'évaluer toutes ces possibilités. Nous avons donc restreint notre étude à deux genres de stratégies :

- Les stratégies *psycholinguistiques* : Ce sont des stratégies basées sur des listes de mots ordonnées selon des propriétés psycholinguistiques.

- Les stratégies *algorithmiques* : Ce sont des stratégies construites en ayant recours à des algorithmes basés sur la théorie des graphes. Parmi celles-ci, l'on peut encore distinguer les stratégies adaptées, uniquement construites pour un lexique en particulier, et les stratégies globales, basées sur des propriétés structurelles normalisées pour tous les lexiques.

3.2.1 Les stratégies psycholinguistiques

Les chercheurs qui s'intéressent aux aspects cognitifs du langage utilisent depuis longtemps des bases de données standardisées, appelées *normes psycholinguistiques* – en anglais *norms* –, qui regroupent les mots en fonction de leurs propriétés psycholinguistiques (Bonin *et al.*, 2003; Gilhooly et Logie, 1980; Coltheart, 1981; Wilson, 1988). Par exemple, la base de données MRC inventorie 150 837 mots de langue anglaise, pour lesquels 26 propriétés psycholinguistiques différentes sont répertoriées (Wilson, 1988).

Parmi les normes psycholinguistiques récentes, nous en avons retenu 5, rendues disponibles par leurs auteurs en complément de leurs travaux de recherches. Ce sont des listes de mots établies en fonction de *variables psycholinguistiques* fréquemment employées par les chercheurs en psychologie du langage[1] : la *fréquence d'utilisation*, l'*âge d'acquisition* et le degré de *concrétude* des mots (Harley, 2013). La mesure de *fréquence d'utilisation* est sans doute la plus répandue et la plus régulièrement utilisée pour les recherches en psycholinguistique (Brysbaert et New, 2009). Elle consiste en une mesure du taux d'occurrence des mots normalisé à 1 million, à l'intérieur d'un corpus donné. Par *âge d'acquisition*, l'on entend l'âge moyen auquel les enfants sont présumés avoir appris un mot. Quant à la *concrétude*, selon Paivio *et al.*, elle « [...] renvoie au degré avec lequel les mots réfèrent à des individus, des lieux et des objets qui peuvent être vus, entendus, touchés, sentis ou goûtés » (Paivio *et al.*, 1968, cité par (Bonin *et al.*, 2003)).

Le tableau 3.3 présente les cinq sources de données utilisées pour la construction de nos

1. Toutefois, il va sans dire que notre analyse pourrait être facilement étendue à d'autres bases de données utilisant ces mêmes variables ou d'autres, selon la disponibilité des données.

stratégies d'apprentissage, en relation avec les variables psycholinguistiques dont elles sont dérivées.

Tableau 3.3: Variables psycholinguistiques et stratégies d'apprentissage

Variable	Stratégie	Source	# Mots
Fréquence d'utilisation	$\text{FREQ}_{\text{Brysbaert}}$	Brysbaert et New (2009)	74 000
Fréquence d'utilisation	$\text{FREQ}_{\text{NGSL+}}$	Browne *et al.* (2019a,b), Browne et Culligan (2019a,b)	6 600
Âge d'acquisition	$\text{AOA}_{\text{Brysbaert}}$	Kuperman *et al.* (2012)	31 000
Âge d'acquisition	$\text{AOA}_{\text{Childes}}$	MacWhinney (2000)	13 000
Concrétude	$\text{CONC}_{\text{Brysbaert}}$	Brysbaert *et al.* (2014)	37 000

Pour construire nos stratégies d'apprentissage, nous avons tout d'abord lemmatisé et désambiguïsé les mots provenant des bases de données afin de les transformer en lexèmes, puis nous les avons ordonnées en fonction de la variable psycholinguistique considérée. Par exemple, pour une stratégie basée sur l'âge d'acquisition, le premier lexème proposé par la stratégie correspond au mot que les auteurs estiment être appris le plus tôt dans le développement de l'enfant. Puis le 2e lexème suggéré correspond au 2e mot appris et ainsi de suite jusqu'aux lexèmes et aux mots estimés être appris à l'âge le plus avancé.

Une étape d'alignement supplémentaire entre les lexiques et les stratégies s'avère nécessaire. Comme les données psycholinguistiques servant à construire les stratégies proviennent de sources hétérogènes, les lexèmes qu'elles contiennent ne concordent pas nécessairement avec les lexiques. Lorsqu'un lexème proposé par une stratégie n'apparaît pas dans un lexique, nous choisissons de tout simplement l'ignorer [2].

2. Nous ne mesurons pas le degré d'alignement des stratégies psycholinguistiques par rapport aux lexiques, c'est-à-dire la taille des intersections entre les stratégies et les lexiques. C'est l'une des limites de notre analyse. Si nous devions pousser plus loin la recherche, ce calcul pourrait possiblement permettre une évaluation plus fine de la qualité des stratégies.

Lorsque tous les lexèmes d'une stratégie ont été utilisés sans que l'on réussisse à apprendre tout le lexique, l'on se rabat sur une stratégie de repli, comme le décrit l'algorithme 3 à la page 50.

Examinons maintenant comment chacune des différentes stratégies d'apprentissage est développée.

La première stratégie du tableau 3.3, $\text{FREQ}_{\text{Brysbaert}}$, est basée sur la norme de Brysbaert et New (2009). Les auteurs l'ont assemblée à partir du $\text{SUBTLEX}_{\text{US}}$, un corpus de sous-titres de films en anglais américain. Elle comprend 74 000 mots non lemmatisés.

La stratégie $\text{FREQ}_{\text{NGSL+}}$ est dérivée de listes de mots utilisées pour l'apprentissage de l'anglais langue seconde. Bien que les *listes de mots* (section 2.1.3) ne soient pas basées uniquement sur des critères psycholinguistiques, elles représentent tout de même un secteur de recherche important depuis les travaux d'Ogden (1930) et de West (1953). Ce sont des regroupements de mots jugés importants et d'usage fréquent : "[...] high-frequency words that were deemed important for second language learners" (Browne, 2014). Il existe de nombreuses versions de ces listes. Pour ce mémoire, nous avons retenu la "New General Service List" (NGSL), de Browne, Culligan et Phillips (Browne *et al.*, 2019b). C'est une version mise à jour et remaniée de la liste originale de West, contenant 2 800 mots sélectionnés à partir du "Cambridge English Corpus" (CEC). Afin de faire en sorte que la NGSL contienne un nombre de lexèmes comparable aux autres stratégies psycholinguistiques, nous en avons développé une version augmentée : la NGSL+. Cette dernière est obtenue en concaténant à la NGSL trois autres *listes de mots* complémentaires, mises au point par les auteurs de la NGSL à partir de corpus spécialisés :

- La "New Academic Word List" (NAWL) : c'est une liste de 963 mots construite à partir d'un corpus de textes académiques(Browne *et al.*, 2019a).
- La "Business Service List" (BSL) : c'est une liste de 1 700 mots reliés au domaine des affaires et du commerce(Browne et Culligan, 2019a).
- La "TOEIC Service List" (TSL) : c'est une liste de 1 200 mots, complémentaire à la NGSL (Browne et Culligan, 2019b). Elle est destinée aux étudiants désireux de

réussir la certification "Test of English for International Communication" (TOEIC).

Pour construire la stratégie $AOA_{Brysbaert}$, nous avons utilisé la norme basée sur l'âge d'acquisition des mots de Kuperman *et al.* (2012). Comme il n'est pas possible d'obtenir cette information directement des enfants, la méthode la plus fréquemment utilisée consiste à interroger des adultes et à leur demander d'évaluer à quel moment ils pensent avoir appris certains mots. Pour leur recherche, Kuperman, Stadthagen-Gonzalez et Brysbaert ont fait appel à une technique de *crowdsourcing* utilisant le *Amazon Mechanical Turk* (Kuperman *et al.*, 2012). Des participants adultes devaient indiquer à quel âge ils estimaient avoir appris les mots d'une liste. À partir des réponses fournies, les auteurs ont construit une liste de 31 000 mots avec leur âge d'acquisition estimé, compris entre 1 et 21 ans.

L'autre stratégie basée sur l'âge d'acquisition, $AOA_{Childes}$, utilise des données provenant d'une autre source : le projet "Child Language Data Exchange System" (CHILDES) (MacWhinney, 2000). Dans ce cas, une méthode différente est utilisée pour récolter les données. L'âge d'acquisition est estimé à partir de conversations enregistrées d'enfants âgés de un à onze ans. La liste obtenue compte au total 13 000 mots.

Pour notre stratégie $CONC_{Brysbaert}$, nous avons utilisé le travail d'évaluation du degré de concrétude de mots de Brysbaert, Warriner et Kuperman (Brysbaert *et al.*, 2014). Comme pour leur étude concernant l'âge d'acquisition (Kuperman *et al.*, 2012), les auteurs ont utilisé le *crowdsourcing* pour le recrutement des participants. Ces derniers devaient classer des mots selon une échelle de concrétude allant de 1.0 à 5.0, où 1.0 correspond à des mots complètement abstraits et 5.0 correspond aux mots les plus concrets. Par exemple, les mots très concrets BANANA, APPLE et BABY sont de degré 5.0, alors que BELIEF et ALTHOUGH sont respectivement de degré 1.19 et 1.07. La liste utilisée pour notre analyse contient 37 000 mots.

3.2.2 Les stratégies algorithmiques

Les stratégies *algorithmiques* sont des listes de lexèmes calculées à partir des propriétés structurelles des graphes. C'est-à-dire que les lexèmes y sont ordonnés en fonction des résultats d'algorithmes basés sur la théorie des graphes.

Le tableau 3.4 résume les stratégies algorithmiques avec lesquelles nous avons expérimenté. L'on remarque que toutes ces stratégies utilisent directement les algorithmes COÛT ou COÛTDEGRÉDYNAMIQUE sans faire appel à une stratégie de repli. Contrairement aux stratégies psycholinguistiques, les techniques de construction utilisées font en sorte que les lexiques sont entièrement « appris » lorsque les algorithmes de calcul du coût d'apprentissage se terminent.

Tableau 3.4: Stratégies d'apprentissage algorithmiques

Stratégie	Propriété	Algorithme	Nombre
$MFVS_{<lex>}$	Ensemble d'ancrage minimal	COÛT	8 (1 par lexique)
$DD_{<lex>}$	Degré dynamique	COÛTDEGRÉDYNAMIQUE	8 (1 par lexique)
$SD_{<lex>}$	Degré statique	COÛT	8 (1 par lexique)
$MFVS_{mixte}$	Ensemble d'ancrage minimal	COÛT	1
DD_{mixte}	Degré dynamique	COÛT	1
SD_{mixte}	Degré statique	COÛT	1

Avec les 3 premières catégories de stratégies du tableau, $MFVS_{<lex>}$, $DD_{<lex>}$ et $SD_{<lex>}$, nous avons autant de stratégies différentes que de lexiques, chacune d'entre elles étant adaptée à un lexique déterminé. L'indice <lex> représente ici le lexique. Par exemple.s SD_{LDOCE} correspond à la stratégie *Degré statique* pour le lexique LDOCE.

Chacune des stratégies $MFVS_{<lex>}$ est assemblée pour le lexique <lex> correspondant

selon la méthode de calcul décrite à la définition 2.1.1, page 37. Bien que le problème du calcul d'un MFVS soit en général NP-difficile, il a tout de même été possible d'obtenir une solution optimale pour six des huit lexiques et une bonne approximation pour les deux autres. Contrairement aux autres stratégies du tableau, l'ordre d'apparition des lexèmes dans les stratégies $\text{MFVS}_{\texttt{<lex>}}$ n'est pas important.

Avec les stratégies basées sur le *Degré dynamique* $\text{DD}_{\texttt{<lex>}}$, le prochain lexème à apprendre n'est pas choisi à partir d'une liste prédéterminée. Tel que décrit dans l'algorithme 2 à la page 48, il est plutôt calculé à chaque étape en choisissant le sommet du graphe associé au lexique <lex> dont le degré extérieur est le plus élevé[3]. Comme les lexèmes « appris » sont retirés à chaque étape, c'est donc équivalent au fait de choisir chaque fois le lexème qui apparaît dans le plus grand nombre de définitions.

Pour les stratégies basées sur le *Degré statique* $\text{SD}_{\texttt{<lex>}}$, le prochain lexème à apprendre provient d'une liste contenant tous les lexèmes du lexique <lex>. Les lexèmes y sont ordonnés en ordre descendant du degré extérieur des sommets. Contrairement aux stratégies $\text{DD}_{\texttt{<lex>}}$, le degré des sommets est calculé de façon statique lors de la construction du graphe. Ainsi, l'on commence à « apprendre » les lexèmes en partant de celui qui est le plus utilisé dans les définitions, en allant jusqu'au moins utilisé.

Finalement, afin d'estimer le caractère de généralité de ces stratégies algorithmiques, uniquement adaptées à chaque lexique, nous avons aussi développé des stratégies algorithmiques globales, communes à tous les lexiques. Ces stratégies, dites *stratégies mixtes*, ou encore *stratégies algorithmiques génériques*, sont assemblées en fusionnant dans une liste générique commune à tous les lexiques les lexèmes provenant des stratégies algorithmiques spécifiques à chaque lexique.

Par exemple, les 8 stratégies $\text{DD}_{\texttt{cide}}$, $\text{DD}_{\texttt{ldoce}}$, ..., $\text{DD}_{\texttt{wild}}$ sont fusionnés pour former

---

3. S'il y a plusieurs sommets possédant le même degré extérieur, l'on choisit aléatoirement parmi ceux-ci.

une seule stratégie générique DD_{mixte}, qui constitue une forme de généralisation de la stratégie *Degré Dynamique* appliquée à l'ensemble des lexiques.

La méthode que nous avons utilisé [4] pour construire ces *stratégies mixtes* est la suivante :

- choisir de façon aléatoire un lexique parmi les huit,
- sélectionner le prochain lexème de la stratégie qui correspond à ce lexique et l'en retirer,
- si le lexème n'est pas déjà présent dans la liste générique, l'y ajouter,
  sinon, l'ignorer,
- répèter les étapes précédentes jusqu'à ce que toutes les stratégies soient épuisées.

Par exemple, la stratégie DD_{mixte} a été construite en concaténant les lexèmes dans l'ordre suivant :

Tableau 3.5: Stratégies d'apprentissage algorithmiques

No	Lexème	Provenance
1.	BE ;V	DD_{WILD}
2.	HAVE ;V	DD_{WN}
3.	PERSON ;N	DD_{WLDT}
4.	USE ;N	DD_{WN}
...	...	...
5990.	DEALFISH ;N	DD_{MWC}
5991.	PHENYTOIN ;N	DD_{MWC}

4. Il est clair qu'il est possible de produire de cette façon un très grand de *stratégies mixtes* différentes. Les essais effectués nous ont toutefois démontré qu"il existe très peu de variabilité entre ces différentes variations pour ce qui est de leur niveau de performance.

CHAPITRE IV

RÉSULTATS OBTENUS

Dans ce chapitre, nous présentons les résultats obtenus lors de nos expérimentations.

Nous y expliquons tout d'abord les diverses mesures recueillies lors de l'exécution des traitements algorithmiques sur les lexiques. Nous montrons ensuite les résultats comparatifs des diverses stratégies d'apprentissage en fonction des huit lexiques analysés. Puis, nous concluons le chapitre avec une discussion des résultats.

4.1 Les mesures effectuées

Pour permettre de comparer les différentes combinaisons de stratégies et de lexiques, différents indicateurs de performance ont été enregistrés lors des essais.

4.1.1 Mesures détaillées du déroulement de l'apprentissage

Lors de l'apprentissage d'un lexique avec une stratégie, une série de valeurs est enregistrée chaque fois qu'un lexème est appris de façon directe. Cela permet de mesurer le rythme auquel progresse le processus d'apprentissage. Le tableau 4.1 montre, à titre d'exemple, un aperçu des données enregistrées lors d'un cycle d'apprentissage du lexique MWC avec la stratégie $\text{AOA}_{\text{Brysbaert}}$.

Coût cumulatif	Sommets restants	Arcs restants	Degré	Lexème	Repli
1	249 056	1 152 896	2	mama ;n	0
2	249 054	1 152 894	2	mom ;n	0
3	249 053	1 152 892	8	potty ;n	0

Tableau 4.1 – *Suite à la page suivante*

Tableau 4.1 – *Suite de la page précédente*

Coût cumulatif	Sommets restants	Arcs restants	Degré	Lexème	Repli
4	249 051	1 152 884	17	yes ;n	0
5	249 047	1 152 867	1 522	water ;n	0
6	249 039	1 151 337	130	wet ;a	0
7	249 037	1 151 208	33	spoon ;n	0
8	249 036	1 151 175	51	nap ;n	0
9	249 030	1 151 121	2	daddy ;n	0
10	249 028	1 151 119	18	hug ;n	0
11	249 026	1 151 101	212	shoe ;n	0
10	249 028	1 151 119	18	hug ;n	0
11	249 026	1 151 101	212	shoe ;n	0
...	...	...	...	...	...
10113	14	14	1	kakemono ;n	484
10114	12	12	1	stilbestrol ;n	485
10115	10	10	1	ciphertext ;n	486
10116	8	8	1	banderilla ;n	487
10117	6	6	1	amphitryon ;n	488
10118	4	4	1	mannose ;n	489
10119	2	2	1	phenytoin ;n	490

Tableau 4.1: Déroulement de l'apprentissage (partiel)

L'on y retrouve les mesures suivantes :

Coût cumulatif : Indique le coût cumulatif, c'est-à-dire le nombre de lexèmes appris directement depuis le début du cycle d'exécution

Nb. Sommets : Indique le nombre de sommets restants dans le graphe (avant l'apprentissage du lexème)

Nb. Arcs : Indique le nombre d'arcs restants dans le graphe (avant l'apprentissage du lexème)

Degré : Indique le degré extérieur du lexème proposé par la stratégie

Lexème : Lexème proposé par la stratégie (appris directement)

Repli : Indique le coût cumulatif de la stratégie de repli

4.1.2 Mesures de performance globale

Lorsque l'apprentissage d'un lexique prend fin, des indicateurs de performance globaux sont aussi comptabilisés. Le tableau 4.2 montre pour chacun des huit lexiques évalués (de gauche à droite) comment les différentes stratégies d'apprentissage se comparent en termes de coût, de rendement, de pourcentage de mots appris directement, ainsi que de couverture (si applicable).

Stratégie	Mesure	CIDE	LDOCE	MWC	WN	WEDT	WCDT	WLDT	WILD
	# Lexèmes	47 092	69 204	249 137	132 547	73 091	20 128	6 036	4 244
MFVS	Coût	349	484	1 544	1 251	1 365	570	231	340
	Rendement	**134.93**	**142.98**	**161.36**	**105.95**	**53.55**	**35.31**	**26.13**	**12.48**
	Pct	0,74%	0,70%	0,62%	0,94%	1,87%	2,83%	3,83%	8,01%
	Couverture	s. o.	s. o.	s. o.	s. o.	s. o.	s. o.	s. o.	s. o.
DD	Coût	684	843	3 095	2 566	2 389	897	394	574
	Rendement	68.85	82.09	80.50	51.66	30.59	22.44	15.32	7.39
	Pct	1,45%	1,22%	1,24%	1,94%	3,27%	4,46%	6,53%	13,52%
	Couverture	s. o.	s. o.	s. o.	s. o.	s. o.	s. o.	s. o.	s. o.
DS	Coût	687	838	3 081	2 558	2 386	899	394	577
	Rendement	68.55	82.58	80.86	51.82	30.63	22.39	15.32	7.36
	Pct	1,46%	1,21%	1,24%	1,93%	3,26%	4,47%	6,53%	13,60%
	Couverture	s. o.	s. o.	s. o.	s. o.	s. o.	s. o.	s. o.	s. o.
$MFVS_{mixte}$	Coût	704	966	3 077	2 835	2 348	957	398	612
	Rendement	66.85	71.64	80.96	46.75	31.13	21.03	15.17	6.93
	Pct	1,49%	1,40%	1,24%	2,14%	3,21%	4,75%	6,59%	14,42%
	Couverture	s. o.	s. o.	s. o.	s. o.	s. o.	s. o.	s. o.	s. o.
DD_{mixte}	Coût	768	963	3 466	3 002	2 574	987	448	645
	Rendement	61.32	71.82	71.88	44.15	28.39	20.39	13.47	6.57
	Pct	1,63%	1,39%	1,39%	2,26%	3,52%	4,90%	7,42%	15,20%
	Couverture	s. o.	s. o.	s. o.	s. o.	s. o.	s. o.	s. o.	s. o.
DS_{mixte}	Coût	793	988	3 776	3 021	2 721	1 024	454	678
	Rendement	59.32	70.00	65.98	43.87	26.86	19.65	13.30	6.25
	Pct	1,68%	1,43%	1,52%	2,28%	3,72%	5,09%	7,52%	15,98%
	Couverture	s. o.	s. o.	s. o.	s. o.	s. o.	s. o.	s. o.	s. o.

Tableau 4.2 – *Suite à la page suivante*

Tableau 4.2 – *Suite de la page précédente*

Stratégie	Mesure	CIDE	LDOCE	MWC	WN	WEDT	WCDT	WLDT	WILD
$FREQ_{NGSL+}$	Coût	2 813	1 954	5 008	4 127	3 238	1 354	712	1 260
	Rendement	16.74	35.42	49.75	32.12	22.57	14.87	8.48	3.37
	Pct	5,97%	2,82%	2,01%	3,11%	4,43%	6,73%	11,80%	29,69%
	Couverture	97.0%	90.4%	71.2%	73.4%	67.7%	82.9%	97.9%	92.8%
$FREQ_{Brysbaert}$	Coût	6 751	2 170	8 217	7 204	6 555	1 999	960	1 193
	Rendement	6.98	31.89	30.32	18.40	11.15	10.07	6.29	3.56
	Pct	14,34%	3,14%	3,30%	5,44%	8,97%	9,93%	15,90%	28,11%
	Couverture	99.9%	99.3%	96.1%	94.8%	98.7%	99.8%	99.7%	99.6%
$AOA_{Childes}$	Coût	4 971	5 010	7 729	7 284	5 586	3 409	1 585	2 016
	Rendement	9.47	13.81	32.23	18.20	13.08	5.90	3.81	2.11
	Pct	10,56%	7,24%	3,10%	5,50%	7,64%	16,94%	26,26%	47,50%
	Couverture	99.4%	97.7%	82.9%	86.3%	84.3%	97.3%	99.7%	98.3%
$AOA_{Brysbaert}$	Coût	7 105	4 851	10 119	10 340	8 278	2 950	1 284	1 430
	Rendement	6.63	14.27	24.62	12.82	8.83	6.82	4.70	2.97
	Pct	15,09%	7,01%	4,06%	7,80%	11,33%	14,66%	21,27%	33,69%
	Couverture	99.6%	99.2%	95.2%	94.0%	96.7%	97.6%	99.5%	95.7%
$CONC_{Brysbaert}$	Coût	8 900	11 669	16 580	17 037	12 792	6 042	2 373	2 477
	Rendement	5.29	5.93	15.03	7.78	5.71	3.33	2.54	1.71
	Pct	18,90%	16,86%	6,65%	12,85%	17,50%	30,02%	39,31%	58,36%
	Couverture	99.7%	99.6%	96.4%	96.0%	97.5%	98.9%	99.7%	97.6%

Tableau 4.2: Coût, taux de rendement, pourcentage et couverture

L'on y retrouve les indicateurs suivants :

Coût : indique le coût total d'apprentissage de la stratégie, c.-à-d. le nombre total de lexèmes qu'il a été nécessaire d'apprendre de façon directe pour réussir à apprendre le lexique au complet (voir les algorithmes 1 et 2). Par exemple, le coût de la stratégie DD pour le lexique WILD est de 574, ce qui représente le nombre total de lexèmes qu'il a fallu apprendre de façon directe.

Rendement : représente le rapport entre le nombre total de lexèmes et le coût d'apprentissage. Toujours dans le même tableau, nous voyons que le dictionnaire WILD contient 4 244 lexèmes et qu'il a fallu apprendre de façon directe 1 260 lexèmes avec la stratégie $FREQ_{NGSL+}$. Le taux de rendement de la stratégie $FREQ_{NGSL+}$

pour le lexique WILD est donc de 4 244/1 260 ou 3,37.

Nous pouvons interpréter cette mesure comme étant le nombre moyen de lexèmes pouvant être appris par définition pour chaque lexème appris de façon directe. Ce qui revient à dire que chaque fois que l'on apprend un lexème de façon directe, cela nous permet d'apprendre par définition en moyenne 2,37 lexèmes supplémentaires.

Pct : représente le pourcentage des mots appris directement pour une stratégie et un lexique donnés. Cela correspond à la proportion du nombre de lexèmes appris directement, par rapport au nombre total de lexèmes dans le lexique.

Par exemple, pour le lexique WEDT et la stratégie $\text{FREQ}_{\text{NGSL+}}$, le pourcentage des lexèmes appris directement, est de 4,43%, 3 238 par rapport à 73 091.

Couverture : pour les stratégies non exhaustives, permet de mesurer l'efficacité de la stratégie, en pourcentage du coût total, par rapport à la stratégie de repli.

Par exemple, pour le lexique WCDT et la stratégie $\text{FREQ}_{\text{NGSL+}}$, la couverture est de 82,9%. Cela veut dire que, sur un coût d'apprentissage total de 1 354, 1 122 lexèmes, soit 82,9%, ont été appris avec la stratégie $\text{FREQ}_{\text{NGSL+}}$, et que les 232 restants (17,1%) ont été appris avec DD, la stratégie de repli choisie. Par contre, pour ce même lexique WCDT avec cette fois la stratégie $\text{FREQ}_{\text{Brysbaert}}$, le degré de couverture atteint 99,8%.

Étant donné qu'il existe une correspondance exacte entre les transversaux de circuit et les ensembles d'ancrage (Blondin Massé *et al.*, 2008), l'on peut raisonnablement s'attendre à ce que les stratégies basées sur les transversaux de circuit minimaux produisent de bons résultats. Comme prévu, il s'avère que les stratégies les plus efficaces sont celles qui utilisent les ensembles d'ancrage minimaux des lexiques, les stratégies $\text{MFVS}_{\texttt{<lex>}}$.

En second lieu, les stratégies optimisées en fonction du degré des sommets – $\text{DD}_{\texttt{<lex>}}$ et $\text{DS}_{\texttt{<lex>}}$ – se révèlent aussi très efficaces.

Enfin, l'on constate que pour les lexiques les plus volumineux – MWC, WN, WEDT, WCDT – les stratégies $\text{FREQ}_{\text{NGSL+}}$ et $\text{AOA}_{\text{Childes}}$ présentent un taux de couverture plutôt faible de moins de 90%.

4.2 Discussion des résultats

Rythme d'apprentissage

Les figures[1] de cette section permettent de comparer certains aspects du déroulement de l'apprentissage pour les huit lexiques étudiés. Chacune des sous-figures est produite à partir des mesures de performance détaillées – décrites à la page 63 – enregistrées au fur et à mesure de l'apprentissage des lexèmes d'un lexique à l'aide d'une stratégie donnée.

La première série de graphiques de la figure 4.2 à la page 70 compare le rythme d'apprentissage des stratégies algorithmiques par rapport aux stratégies psycholinguistiques.

Pour faciliter la compréhension, la figure 4.1 reprend en format agrandi le graphique 4.2a pour le lexique CIDE. Ce diagramme illustre le rythme d'apprentissage des stratégies algorithmiques $\text{MFVS}_{\text{CIDE}}$ et DD_{CIDE} par rapport aux stratégies psycholinguistiques $\text{FREQ}_{\text{NGSL+}}$ et $\text{FREQ}_{\text{Brysbaert}}$.

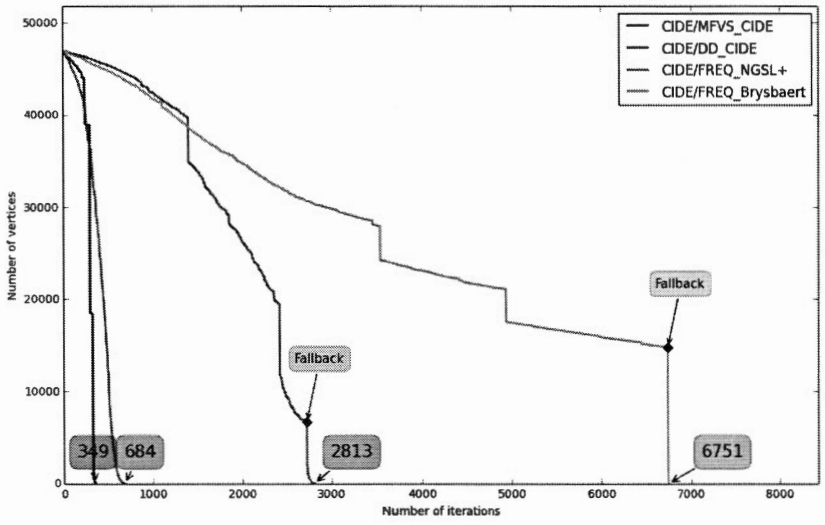
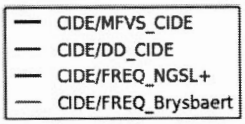


Fig. 4.1: Évolution de l'apprentissage : CIDE

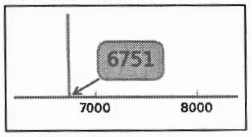
L'on y distingue :

1. À visualiser de préférence en couleurs

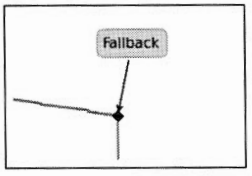
- Les courbes illustrant le rythme d'apprentissage, identifiées avec une couleur différente pour chacune lexique :



- Une pastille de même couleur que la courbe, qui indique pour chaque stratégie le coût total :



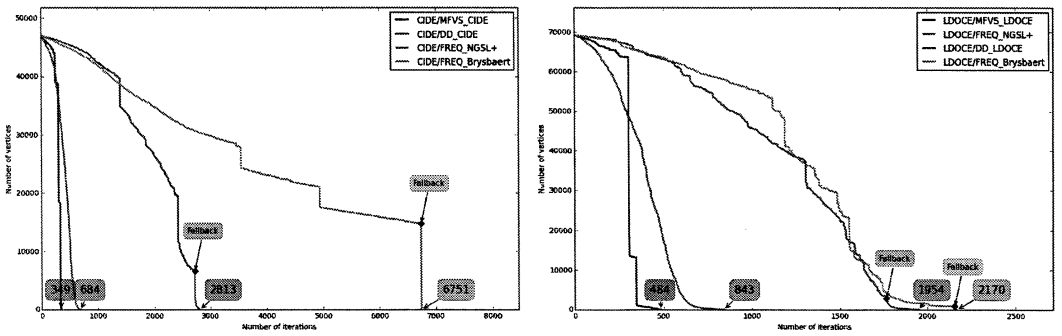
- Une autre pastille, qui montre, pour chaque stratégie non exhaustive, le moment pendant l'apprentissage où il a été nécessaire de recourir à la stratégie de repli :



Pour le lexique CIDE (4.2a) ainsi que pour l'ensemble des lexiques de la figure 4.2, nous voyons que la stratégie MFVS est celle qui offre le meilleur rendement. Cela confirme l'hypothèse voulant qu'apprendre les lexèmes associés aux sommets de l'ensemble d'ancrage minimal fait en sorte de briser rapidement les boucles de définition.

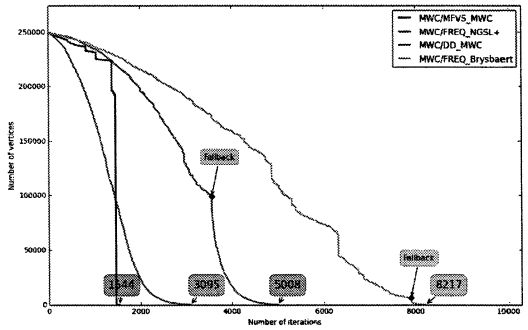
Les graphiques de la figure 4.3 représentent le déroulement de l'apprentissage pour les stratégies utilisant le degré dynamique des sommets par rapport à celles utilisant le degré statique. L'on remarque que ces 2 stratégies algorithmiques, $DD_{\langle LEX \rangle}$ et $SD_{\langle LEX \rangle}$ donnent en pratique des résultats équivalents.

Les graphiques de la figure 4.4, permettent de comparer l'évolution de l'apprentissage de la stratégie algorithmique $DD_{\langle LEX \rangle}$ par rapport aux stratégies psycholinguistiques $FREQ_{NGSL+}$, $FREQ_{Brysbaert}$, $AOA_{Brysbaert}$ et $CONC_{Brysbaert}$. L'on constate que les stratégies psycholinguistiques sont nettement moins efficaces pour briser les boucles de définition. Étant donné que l'ordre d'apparition des lexèmes y est fixé uniquement en fonction de critères psycholinguistiques, de nombreux lexèmes sont appris de façon directe et

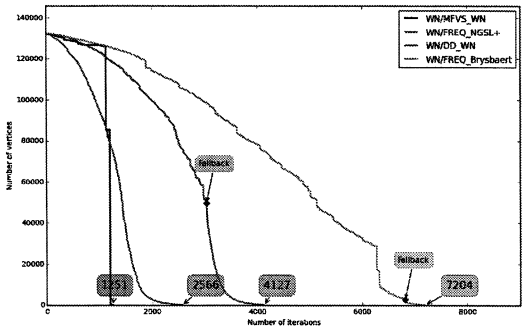


(a) CIDE

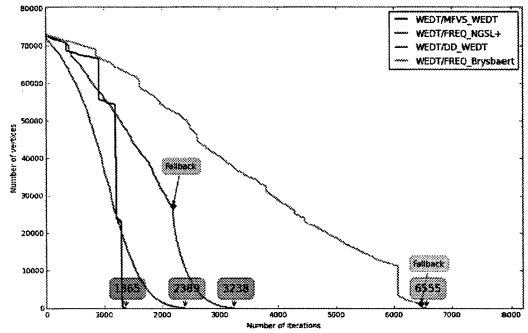
(b) LDOCE



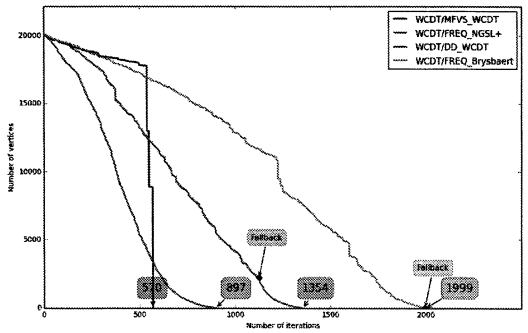
(c) MWC



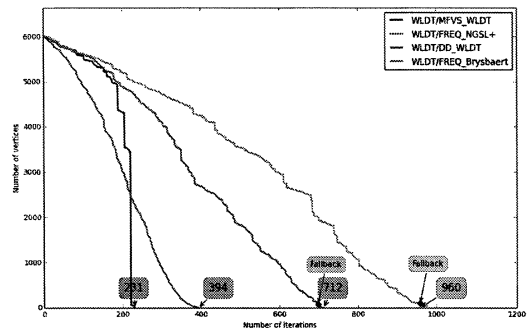
(d) WN



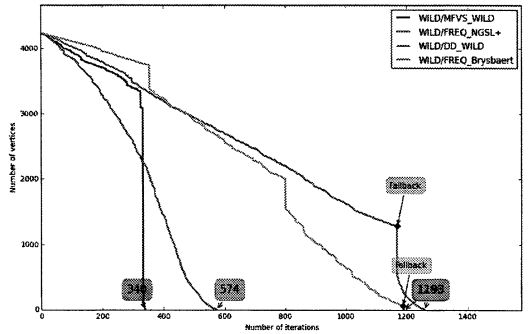
(e) WEDT



(f) WCDT

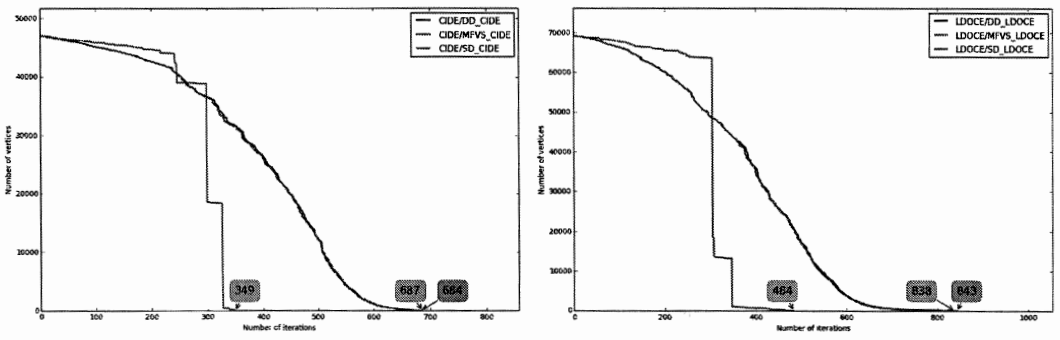


(g) WLDT



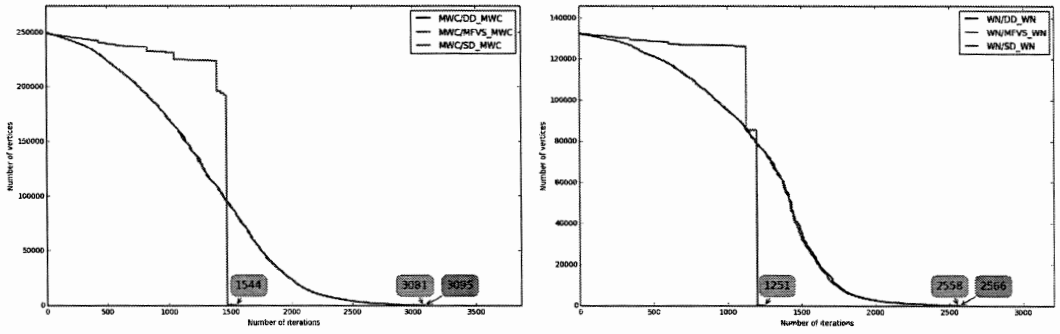
(h) WILD

Fig. 4.2: Évolution de l'apprentissage : Stratégies algorithmiques vs psycholinguistiques



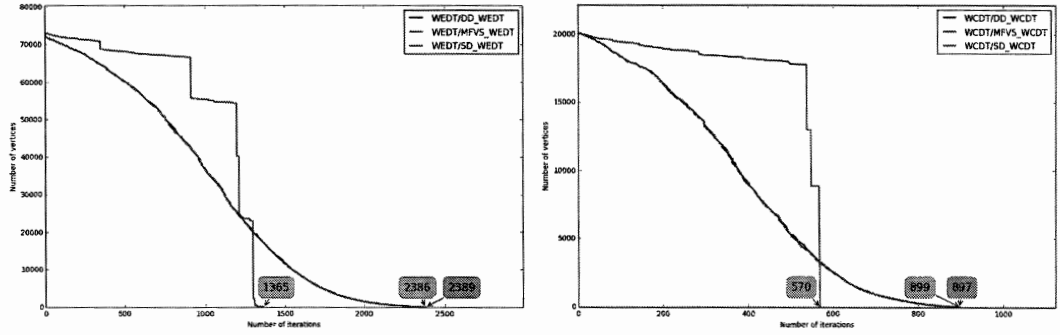
(a) CIDE

(b) LDOCE



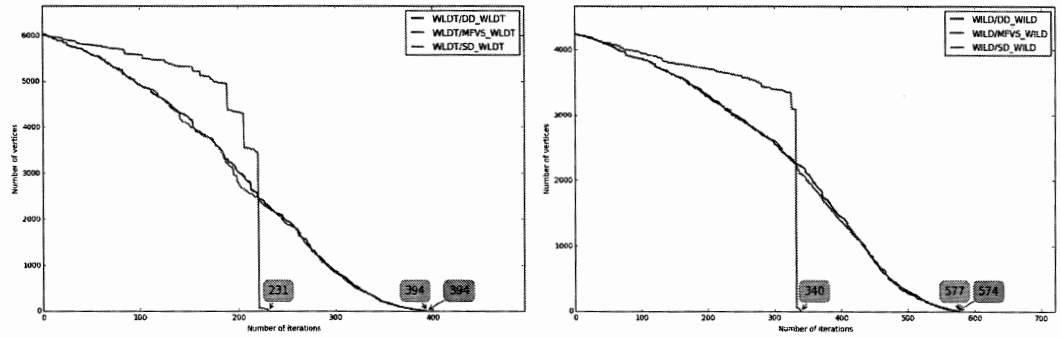
(c) MWC

(d) WN



(e) WEDT

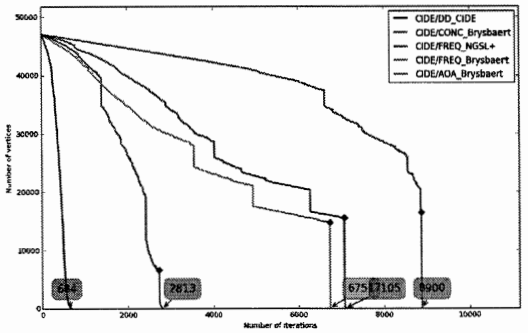
(f) WCDT



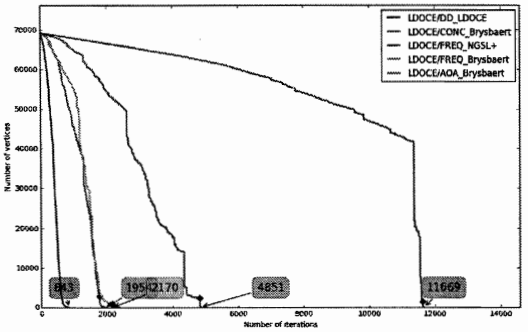
(g) WLDT

(h) WILD

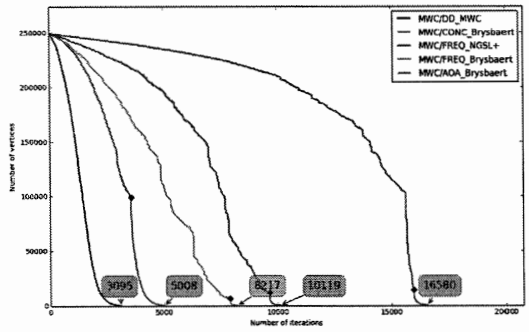
Fig. 4.3: Évolution de l'apprentissage : Stratégies degré dynamique vs degré statique



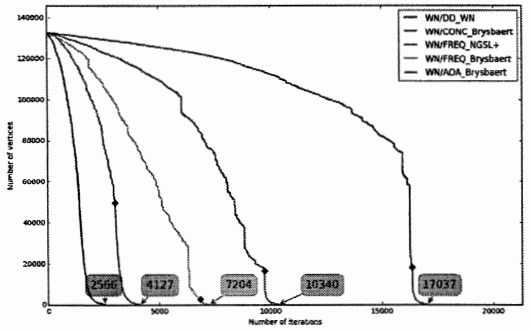
(a) CIDE



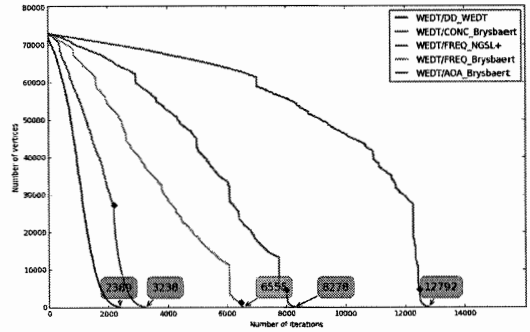
(b) LDOCE



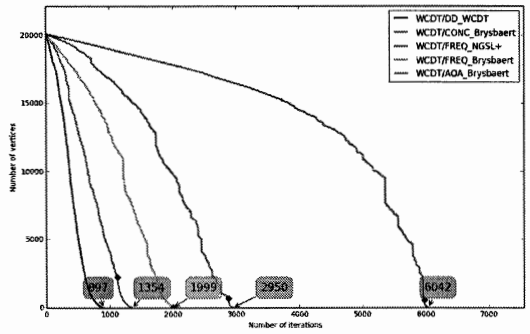
(c) MWC



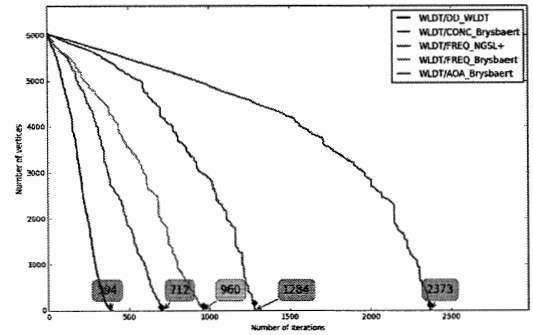
(d) WN



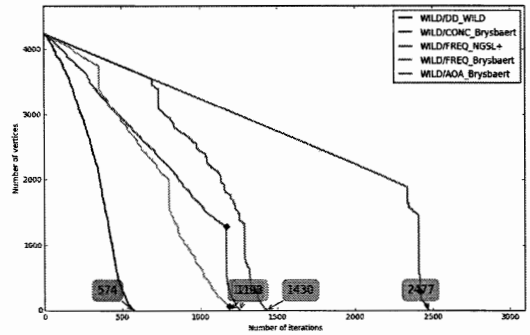
(e) WEDT



(f) WCDT



(g) WLDT



(h) WILD

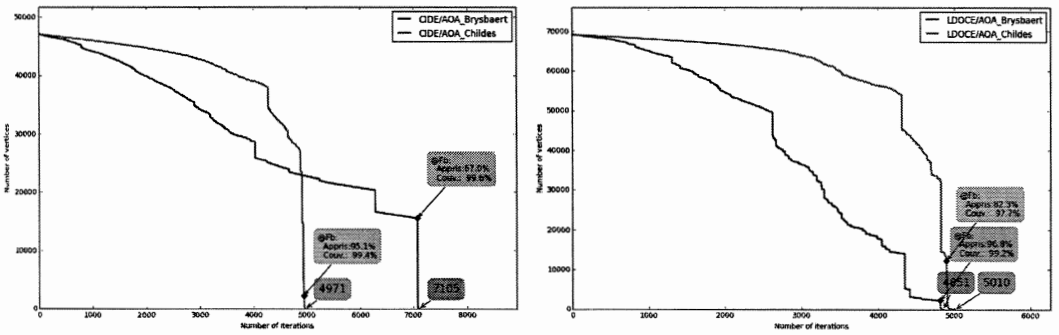
Fig. 4.4: Évolution de l'apprentissage : Fréquence

contribuent à augmenter le coût de la stratégie, alors qu'ils auraient pu être appris par définition – à coût nul – plus tard durant le cycle d'apprentissage. Parmi les stratégies psycholinguistiques, les deux stratégies basées sur la fréquence d'utilisation, $\text{FREQ}_{\text{NGSL+}}$ et $\text{FREQ}_{\text{Brysbaert}}$, sont les plus efficaces, alors que la stratégie $\text{CONC}_{\text{Brysbaert}}$ s'avère nettement moins performante.

De façon intuitive, l'on comprend qu'il n'est pas possible de réussir à apprendre tous les mots d'un dictionnaire en utilisant seulement des mots concrets. Il faut combiner les deux sortes de mots, abstraits et concrets, pour arriver à construire des définitions qui dépeignent bien le sens lexical.

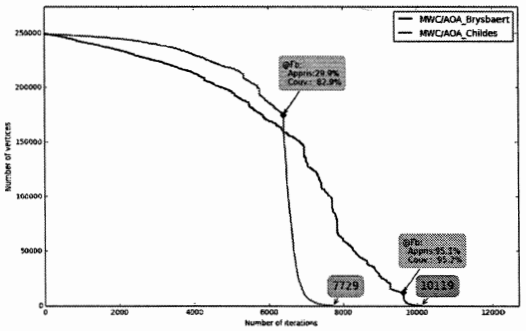
La figure 4.5 met en parallèle le rythme d'apprentissage des stratégies basées sur l'âge d'acquisition. Au premier coup d'œil, la stratégie $\text{AOA}_{\text{Childes}}$ semble offrir un taux de rendement supérieur à $\text{AOA}_{\text{Brysbaert}}$. Toutefois, étant donné que $\text{AOA}_{\text{Childes}}$ contient beaucoup moins de lexèmes que $\text{AOA}_{\text{Brysbaert}}$, son taux de couverture est plus bas. Dans plusieurs cas, l'on atteint le point de repli très rapidement, avant même d'avoir appris 40% des lexèmes. L'on constate que dans ce cas, le fait d'utiliser une stratégie de repli rend difficile la comparaison directe entre $\text{AOA}_{\text{Childes}}$ et $\text{AOA}_{\text{Brysbaert}}$. Cela dit, bien que les stratégies $\text{AOA}_{\text{Childes}}$ et $\text{AOA}_{\text{Brysbaert}}$ ne soient pas parmi les plus performantes, elles démontrent que dans la majorité des cas il est suffisant de connaître moins de 15% des lexèmes pour pouvoir apprendre tous les autres *par définition.*

Les graphiques de la figure 4.6 comparent le rythme d'apprentissage des stratégies algorithmiques mixtes par rapport aux deux stratégies psycholinguistiques les plus efficaces. Dans tous les cas, les stratégies algorithmiques sont clairement plus efficaces pour apprendre un lexique en utilisant le moins de lexèmes possible.

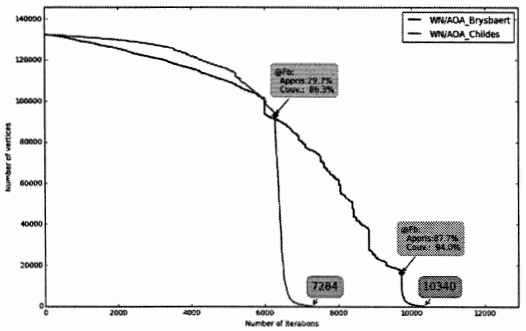


(a) CIDE

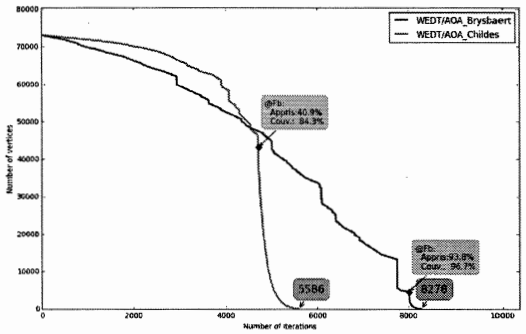
(b) LDOCE



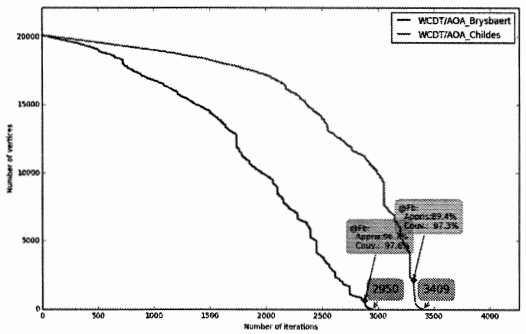
(c) MWC



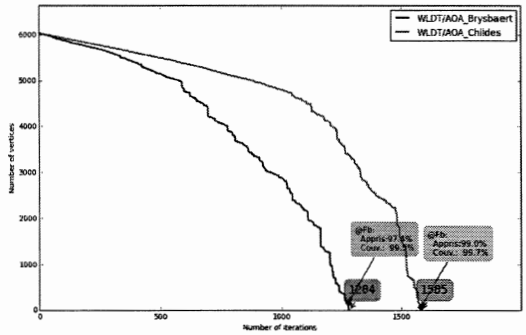
(d) WN



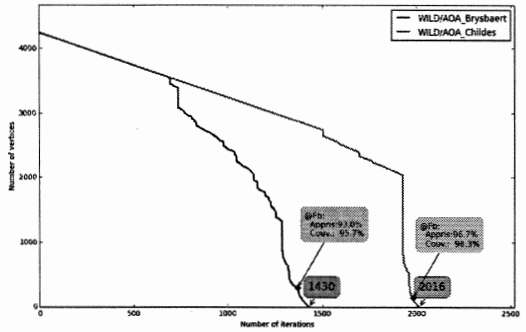
(e) WEDT



(f) WCDT

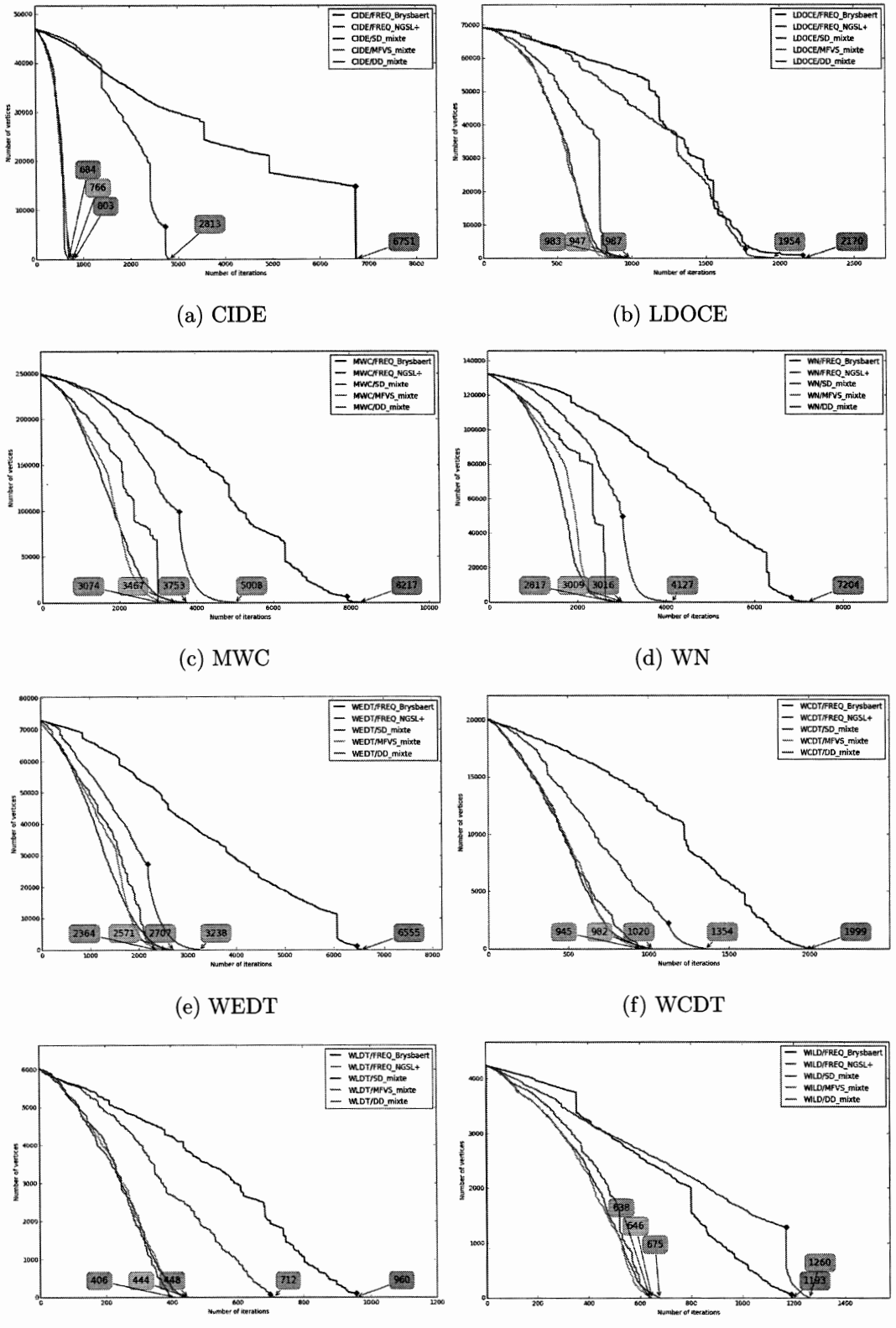


(g) WLDT



(h) WILD

Fig. 4.5: Évolution de l'apprentissage : Stratégies basées sur l'âge d'acquisition



(a) CIDE (b) LDOCE

(c) MWC (d) WN

(e) WEDT (f) WCDT

(g) WLDT (h) WILD

Fig. 4.6: Évolution de l'apprentissage : Stratégies algorithmiques vs mixtes

Taux de rendement

Regardons maintenant les mesures de performance globales des stratégies présentées dans le tableau 4.2 à la page 66. La figure 4.7 présente sous forme de graphique[2] le taux de rendement des lexiques pour chacune des stratégies évaluées. Chacun des huit lexiques est représenté par une courbe et est associé à un code de couleur. Les stratégies sont affichées sur l'axe des X, de gauche à droite en ordre descendant de rendement.

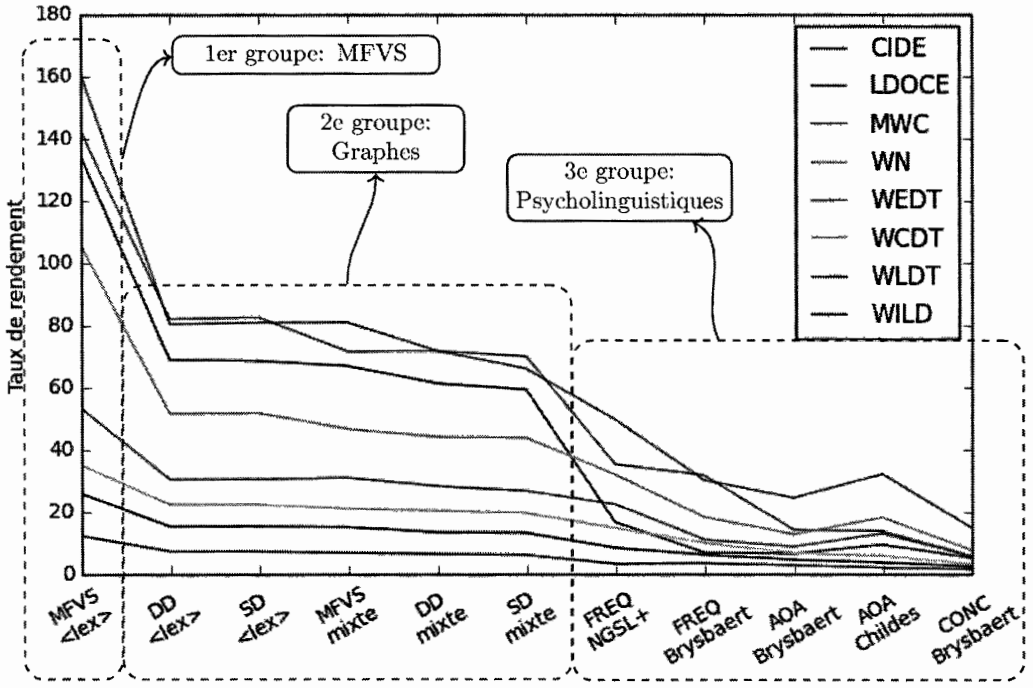


Fig. 4.7: Lexiques : Rendement vs Stratégies

Dans la figure, trois groupes distincts de stratégies sont mis en évidence :

- Le *1er groupe* contient les stratégies algorithmiques $\text{MFVS}_{\texttt{<lex>}}$ optimisées pour chaque lexique. Pour chacun d'eux, elles offrent le taux de rendement le plus élevé.
- Le *2e groupe* comprend les autres stratégies algorithmiques construites à partir des graphes associés. $\text{DD}_{\texttt{<lex>}}$ et $\text{SD}_{\texttt{<lex>}}$ sont construites individuellement pour chaque

2. À visualiser de préférence en couleurs

lexique, alors que $MFVS_{mixte}$, DD_{mixte} et SD_{mixte} sont des stratégies globales pour tous les lexiques. Dans tous les cas, leur rendement est moindre que les stratégies $MFVS_{<lex>}$, mais tout de même appréciable.

- Finalement, le *3e groupe* rassemble les stratégies psycholinguistiques $FREQ_{NGSL+}$, $FREQ_{Brysbaert}$, $AOA_{Childes}$, $AOA_{Brysbaert}$ et $CONC_{Brysbaert}$. Le rendement de ces dernières est manifestement inférieur aux stratégies algorithmiques des deux groupes précédents.

En résumé, ce sont les stratégies optimisées pour chaque lexique, à savoir $MFVS_{<lex>}$, $DD_{<lex>}$ et $SD_{<lex>}$, qui sont les plus efficaces pour l'apprentissage des lexiques. Quant à la question de savoir s'il est possible de construire des stratégies « générales » aussi efficaces que ces stratégies adaptées, les résultats obtenus avec les stratégies mixtes montrent que c'est possible. Les trois stratégies générales $MFVS_{mixte}$, DD_{mixte} et SD_{mixte} sont presque aussi efficaces que les stratégies $MFVS_{<lex>}$, $DD_{<lex>}$ et $SD_{<lex>}$ uniquement optimisées pour chaque lexique. Elles sont dans tous les cas manifestement supérieures aux stratégies basées sur des variables psycholinguistiques.

CONCLUSION

Par définition, un dictionnaire est un système fermé, où les mots utilisés pour bâtir les définitions sont, à de rares exceptions près, définis ailleurs dans le dictionnaire. "The [...] dictionary is a closed system, i.e. words used in definitions are themselves elsewhere defined in the dictionary" (Amsler, 1980, p. *vii*). L'on peut par conséquent construire à partir des mots d'un dictionnaire et des relations définitionnelles qui les lient une structure de graphe complexe. Dans ce mémoire, notre objectif était d'étudier cette structure de graphe en mettant à profit les notions de la théorie des graphes.

Bien que les termes *dictionnaire* et *mot* puissent sembler a priori clairs et sans ambiguïté, leur utilisation dans des contextes très variables de la vie courante les rend difficilement utilisables pour des fins mathématiques. Nous avons convenu de les remplacer dans notre discussion par leurs équivalents plus précis : *lexiques* et *lexèmes*. Puis, nous avons développé une terminologie linguistique rigoureuse qui nous a permis de définir formellement le concept de lexique et de préciser la notion de graphe associé à un lexique.

Nous avons ensuite expliqué comment la notion d'apprentissage des lexèmes – les mots du dictionnaire – peut être utilisée comme outil d'analyse pour explorer la structure même des lexiques. Dans ce contexte, nous avons considéré qu'un lexème peut être appris de deux façons : *par définition*, lorsque tous les lexèmes qui entrent dans sa définition sont déjà connus, ou sinon, *par apprentissage direct*, lorsque l'on doit investir un effort important pour l'apprendre en le reliant à une perception sensorimotrice. Nous avons décrit notre *modèle d'apprentissage*, qui comprend des *stratégies d'apprentissage*, dont l'objectif est de minimiser l'effort – le coût – nécessaire pour apprendre l'ensemble des lexèmes d'un lexique. Nous avons aussi développé les algorithmes nécessaires pour évaluer objectivement le rythme d'apprentissage et le taux de rendement des stratégies lorsqu'elles sont utilisées pour apprendre les lexiques numériques.

Par la suite, nous avons décrit les données sources ayant servi à réaliser nos analyses : les dictionnaires numériques monolingues et les normes psycholinguistiques.

Finalement nous avons exposé les résultats obtenus, qui sont exprimés de deux façons différentes :

- en terme de *rythme d'apprentissage*, qui est une mesure de la rapidité avec laquelle une stratégie progresse vers son objectif d'apprendre tous les lexèmes ;
- en terme de *taux de rendement*, c'est-à-dire la proportion finale entre le nombre de lexèmes appris par définition et le nombre de lexèmes appris directement.

Notre analyse des résultats a confirmé un point d'importance majeure. Les structures circulaires de définitions entre les mots jouent un rôle clé dans l'organisation et la structure interne des dictionnaires.

Si l'on considère un dictionnaire du strict point de vue de son utilité pour le lecteur, la définition d'un mot sera pertinente dans la mesure où ce dernier connaît déjà tous les mots qui composent cette définition, ou à tout le moins, s'il connaît suffisamment de mots pour en comprendre le sens.

> "The usefulness of a dictionary definition depends on its ability to explain a meaning using words the reader already knows" (Bullock, 2010).

Si ce n'est pas le cas, le lecteur doit chercher les mots inconnus. Et dans tous les dictionnaires, il y a nécessairement beaucoup de définitions circulaires :

> "In a typical dictionary, more than a quarter of all definitions are written using words whose definitions ultimately refer back to the word being defined" (Bullock, 2010)

Un lecteur qui ne connaît pas suffisamment la langue va obligatoirement rencontrer rapidement des boucles de définition. Dans ce contexte, notre analyse a montré que les stratégies d'apprentissage les plus efficaces sont celles qui parviennent à briser le plus rapidement les boucles de définition. La meilleure d'entre elles est bien entendu celle qui utilise l'ensemble d'ancrage minimal du lexique, qui correspond au *transversal de circuit*

minimal (MFVS) du graphe. Nous avons vu que pour arriver à cette solution idéale, le problème de la recherche d'un *transversal de circuit* minimal (MFVS) est NP-difficile. Même en faisant appel à des techniques avancées d'approximation, cela demeure un calcul très complexe.

Toutefois, nos résultats montrent aussi qu'avec une stratégie d'apprentissage algorithmique consistant à ordonner les lexèmes à apprendre selon le degré extérieur des sommets qui leur correspondent, l'on arrive à apprendre rapidement tous les lexèmes d'un lexique. En termes lexicographiques, cela correspond à ordonner les mots en fonction du nombre de fois où ceux-ci apparaissent dans la définition d'autres mots. L'on obtient ainsi des résultats qui surpassent de beaucoup, en termes de rendement, les stratégies psycholinguistiques.

En plus de leur utilisation comme outil d'analyse des dictionnaires, les stratégies algorithmiques examinées possèdent un avantage supplémentaire. Elles constituent une approche nouvelle pour le développement de listes de mots semblables à celles utilisées en didactique des langues

Les listes de mots des professeurs d'anglais langue seconde ont traditionnellement été construites en se basant en grande partie sur la fréquence des mots dans un corpus. Comme alternative, nous proposons d'utiliser une stratégie algorithmique simple, basée sur le degré extérieur des sommets.

Bien que l'on ne puisse prétendre que la valeur d'une liste de mots ne tienne qu'à son « taux de rendement », nous croyons que cette approche que nous proposons est intéressante, particulièrement dans les cas où il n'est pas possible d'utiliser une liste de mots existante. Dans ce cas, ou en l'absence d'un corpus établi, l'utilisation d'un lexique numérique ou d'un dictionnaire spécialisé permettrait d'établir facilement une liste de mots ou de concepts pertinents, ainsi que l'ordre selon lequel ils devraient être appris.

Perspectives futures

Comme derniers mots pour conclure ce mémoire, il convient de souligner les nombreuses pistes de recherche laissées inexplorées.

L'on pourrait penser en premier à élargir le domaine d'expérimentation pour prendre en compte de nouvelles sources de données. Que ce soit de nouveaux dictionnaires ou des lexiques numériques, de nouveaux algorithmes pour analyser la structure des graphes, ou encore, de nouvelles normes ou listes psycholinguistiques, les possibilités sont nombreuses.

De la même façon, l'on pourrait explorer l'organisation des dictionnaires dans des domaines spécialisés, comme des dictionnaires médicaux, de mathématiques, de musique, ou autres.

De plus, bien que les ressources dans ce domaine soient très souvent difficiles à obtenir, les langues autres que l'anglais pourraient aussi représenter des avenues de recherche enrichissantes. L'on pourrait imaginer analyser la structure de dictionnaires monolingues de ces langues, ou même des dictionnaires bilingues.

Finalement, un autre perfectionnement possible pourrait consister à développer une approche plus évoluée pour la désambiguïsation lexicale des définitions. L'heuristique du premier sens donne habituellement des résultats satisfaisants et constitue un point de référence, un *baseline*, difficile à surpasser pour cette tâche. Même si cela constitue un domaine de recherche en soi, il serait assurément valable d'explorer l'apport de techniques de désambiguïsation utilisant les réseaux neuronaux ou même l'apprentissage profond – en anglais *deep learning* –. L'on obtiendrait de cette façon un graphe associé sémantiquement plus représentatif du dictionnaire ou du lexique examiné.

RÉFÉRENCES

Amsler, R. A. (1980). *The structure of the Merriam-Webster pocket dictionary.* (Thèse de doctorat). The University of Texas at Austin.

Arrivé, M. (1986). *La grammaire d'aujourd'hui. Guide alphabétique de linguistique française.* Flammarion.

Batagelj, V., Mrvar, A. et Zaveršnik, M. (2002). Network analysis of dictionaries. Dans *Proceeedings of the Third Language Technologies Conference, October 14 - 15, 2002, Jožef Stefan Institute, Ljubljana Slovenia.* University of Ljubljana, Inst. of Mathematics, Physics and Mechanics, Department of Theoretical Computer Science.

Beauchesne, J., Beauchesne, M. et Beauchesne, K. (2009). *Dictionnaire des cooccurrences.* Guérin.

Black, K. (1997). Cambridge international dictionary of english. *The Booklist, 93*(19-20).

Blondin Massé, A., Chicoisne, G., Gargouri, Y., Harnad, S., Picard, O. et Marcotte, O. (2008). How is meaning grounded in dictionary definitions? Dans *Proceedings of the 3rd Textgraphs Workshop on Graph-Based Algorithms for Natural Language Processing*, 17–24. Association for Computational Linguistics.

Bondy, J. A., Murty, U. S. R. *et al.* (1976). *Graph theory with applications.* Oxford, UK : Elsevier Science Ltd.

Bonin, P., Méot, A., Aubert, L., Malardier, N., Niedenthal, P. et Capelle-Toczek, M.-C. (2003). Normes de concrétude, de valeur d'imagerie, de fréquence subjective et de valence émotionnelle pour 866 mots. *L'année Psychologique, 103*(4), 655–694.

Boulanger, J.-C. (2003). *Les inventeurs de dictionnaires [ressource électronique] : De l'eduba des scribes mésopotamiens au scriptorium des moines médiévaux. Canadian*

electronic library. Collection Regards sur la traduction. Ottawa, Ont.] : Presses de l'Université d'Ottawa.

Brezina, V. et Gablasova, D. (2013). Is there a core general vocabulary ? introducing the new general service list. *Applied Linguistics*, *36*(1), 1–22.

Brown, E. K. et Anderson, A. (dir.) (2006). *Encyclopedia of language and linguistics* (2nd ed.. éd.). Amsterdam : Elsevier.

Browne, C. (2014). A new general service list : The better mousetrap we've been looking for. *Vocabulary Learning and Instruction*, *3*(2), 1–10.

Browne, C. et Culligan, B. (2019a). *Business Service List (BSL).* Browne, Charles and Culligan, Brent. Récupéré le 31 janvier 2019 de `http://www.newgeneralservicelist.org/bsl-business-service-list/`

Browne, C. et Culligan, B. (2019b). *TOEIC Service List (TSL).* Browne, Charles and Culligan, Brent. Récupéré le 31 janvier 2019 de `http://www.newgeneralservicelist.org/toeic-list/`

Browne, C., Culligan, B. et Phillips, J. (2019a). *New Academic Word List (NAWL).* Browne, Charles and Culligan, Brent and Phillips, Joseph. Récupéré le 31 janvier 2019 de `http://www.newacademicwordlist.org`

Browne, C., Culligan, B. et Phillips, J. (2019b). *New General Service List (NGSL).* Browne, Charles and Culligan, Brent and Phillips, Joseph. Récupéré le 31 janvier 2019 de `http://www.newgeneralservicelist.org`

Brysbaert, M. et 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., Warriner, A. B. et Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior research methods*, *46*(3), 904–911.

Bullock, D. (2010). Nsm+ ldoce : A non-circular dictionary of English. *International Journal of Lexicography*, *24*(2), 226–240.

Clark, G. (2003). Recursion through dictionary definition space : Concrete versus abstract words. *On WWW at http ://www.ecs.soton.ac.uk/Âharnad/Temp/concreteabstract. pdf. Accessed*, *23*(06).

Coltheart, M. (1981). The Medical Research Council (MRC) psycholinguistic database. *The Quarterly Journal of Experimental Psychology Section A*, *33*(4), 497–505.

Correa Jr, E. A., Lopes, A. A. et Amancio, D. R. (2018). Word sense disambiguation : A complex network approach. *Information Sciences*, *442*, 103–113.

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly*, *34*(2), 213–238.

Cruse, D. A. (2002). The lexicon. In M. Aronoff et J. R. Miller (dir.), *The handbook of linguistics*. Oxford : Blackwell.

De Saussure, F. (1916 (1989)). *Cours de linguistique générale : Édition critique*, volume 1. Otto Harrassowitz Verlag.

Diestel, R. (2000). *Graph theory*. Springer Publishing Company, Incorporated.

Duchacek, O. (1962). L'homonymie et la polysémie. *Vox romanica*, *21*, 49.

Fellbaum, C. (dir.) (1998). *WordNet An Electronic Lexical Database*. Cambridge, MA ; London : The MIT Press.

Gaffiot, F. . (2000). *Le grand Gaffiot dictionnaire latin-français* (nouv. éd. rev. et augm. sous la direction de Pierre Flobert. éd.). Paris : Hachette.

Gilhooly, K. J. et Logie, R. H. (1980). Age-of-acquisition, imagery, concreteness, familiarity, and ambiguity measures for 1,944 words. *Behavior research methods & instrumentation*, *12*(4), 395–427.

Hagberg, A., Swart, P. et S Chult, D. (2008). *Exploring network structure, dynamics, and function using NetworkX.* Rapport technique, Los Alamos National Lab.(LANL), Los Alamos, NM (United States).

Harley, T. A. (2013). *The psychology of language : From data to theory.* Psychology press.

Harnad, S. (1990). The symbol grounding problem. *Physica D : Nonlinear Phenomena*, *42*(1-3), 335–346.

Harnad, S. (2003). Symbol-grounding problem. In *Encyclopedia of cognitive science.* L. Nadel (Ed.).

Harnad, S. (2005). *To Cognize is to Categorize : Cognition is Categorization.* Elsevier.

Hendler, J. et van Harmelen, F. (2008). The semantic web : webizing knowledge representation. *Foundations of Artificial Intelligence*, *3*, 821–839.

Joyce, P. (2015). L2 vocabulary learning and testing : The use of L1 translation versus L2 definition. *The Language Learning Journal*, 1–12.

Jurafsky, D. et Martin, J. H. (2009). *Speech and language processing : an introduction to natural language processing, computational linguistics, and speech recognition* (2nd edition. éd.). Prentice Hall series in artificial intelligence. Pearson Prentice Hall.

Karp, R. M. (1972). *Reducibility among Combinatorial Problems*, Dans *Complexity of Computer Computations : Proceedings of a symposium on the Complexity of Computer Computations, held March 20–22, 1972*, (p. 85–103). Springer US, Miller, Raymond E. and Thatcher, James W. and Bohlinger, Jean D. editors : Boston, MA.

Kipfer, B. A. (1984). Methods of ordering senses within entries. *1984*, 101–108.

Kuperman, V., Stadthagen-Gonzalez, H. et Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 english words. *Behavior Research Methods*, *44*(4), 978–990.

Lapointe, M., Massé, A. B., Galinier, P., Lord, M. et Marcotte, O. (2012). *Enumerating minimum feedback vertex sets in directed graphs*, Dans *Bordeaux Graph Workshop 2012*, (p. 101–102). LaBRI, Université Bordeaux 1.

Larousse. (2019). *Larousse*. Larousse. Récupéré le 31 janvier 2019 de `https://www.larousse.fr/dictionnaires/francais/`

Lehmann, F. (1992). Semantic networks. *Computers & Mathematics with Applications*, *23*(2-5), 1–50.

Leskovec, J., Rajaraman, A. et Ullman, J. D. (2014). *Mining of massive datasets*. Cambridge university press.

Lew, R. (2013). Identifying, ordering and defining senses. *The Bloomsbury companion to lexicography*, 284–302.

Lin, H.-M. et Jou, J.-Y. (2000). On computing the minimum feedback vertex set of a directed graph by contraction operations. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, *19*(3), 295–307.

Longman. (2019). *LDOCE*. Longman. Récupéré le 31 janvier 2019 de `https://pearsonerpi.com/fr/elt/dictionaries/longman-dictionary-of-contemporary-english`

MacWhinney, B. (2000). *The CHILDES Project : Tools for Analyzing Talk*. Mahwah, NJ : Lawrence Erlbaum Associates, third edition.

Manning, C. D., Raghavan, P. et Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press.

McCarthy, D., Koeling, R., Weeds, J. et Carroll, J. (2004). Finding predominant word senses in untagged text. *ACM*.

Merriam-Webster (2003). *Merriam-Webster's Collegiate Dictionary*, (11th éd.).

Merriam-Webster. (2019a). *Merriam-Webster*. Merriam-Webster. Récupéré le 31 janvier 2019 de `https://www.merriam-webster.com`

Merriam-Webster. (2019b). *Merriam-Webster.* Merriam-Webster. Récupéré le 31 janvier 2019 de `https://www.merriam-webster.com/help/explanatory-notes/dict-definitions`

Merriam-Webster. (2019c). *MerriamWebsterThesaurus.* Merriam-Webster. Récupéré le 31 janvier 2019 de `https://www.merriam-webster.com/thesaurus/`

Miller, G. A. (1986). Dictionaries in the mind. *Language and cognitive processes*, *1*(3), 171–185.

Monge, A. L. Z. (2013). L'évolution de l'enseignement du vocabulaire dans la classe de l2. *Revista de Lenguas Modernas*, 437–447.

Nation, I. S. (2016). *Making and using word lists for language learning and testing.* John Benjamins Publishing Company.

Navigli, R. (2009). Word sense disambiguation : A survey. *ACM Computing Surveys (CSUR)*, *41*(2), 10.

Nelson, D. L., McEvoy, C. L. et Schreiber, T. A. (1999). The university of South Florida word association, rhyme, and word fragment norms. Récupéré le 31 janvier 2019 de `http://w3.usf.edu/FreeAssociation/`

Ogden, C. K. (1930). *Basic English : A General Introduction with Rules and Grammar.* Paul Treber & Co Ltd. London.

Ogden, C. K. (2019). *OGDEN's BASIC ENGLISH.* K. Paul, Trench, Trubner. Récupéré le 31 janvier 2019 de `http://ogden.basic-english.org/wordmenu.html`

Oxford. (2019). *Oxford English Dictionary.* Oxford. Récupéré le 31 janvier 2019 de `https://en.oxforddictionaries.com/definition/us/word_form`

Paivio, A., Yuille, J. C. et Madigan, S. A. (1968). Concreteness, imagery, and meaningfulness values for 925 nouns. *Journal of experimental psychology*, *76*(1p2), 1.

Picard, O., Blondin Massé, A. et Harnad, S. (2010). Learning word meaning from dictionary definitions : Sensorimotor induction precedes verbal instruction. Dans *Summer Institute on the Origins of Language, Cognitive Sciences Institute, Université du Québec à Montréal.*

Picard, O., Blondin-Massé, A., Harnad, S., Marcotte, O., Chicoisne, G. et Gargouri, Y. (2009). Hierarchies in dictionary definition space. *arXiv preprint arXiv :0911.5703.*

Picard, O., Lord, M., Blondin-Massé, A., Marcotte, O., Lopes, M. et Harnad, S. (2013). Hidden structure and function in the lexicon. *arXiv preprint arXiv :1308.2428.*

Polguère, A. (2016). *Lexicologie et sémantique lexicale : notions fondamentales* (troisième édition.. éd.). Paramètres. Les Presses de l'Université de Montréal.

Poulin, J.-M., Massé, A. B. et Fonseca, A. (2018). Strategies for learning lexemes efficiently : A graph-based approach. Dans *COGNITIVE 2018 : The Tenth International Conference on Advanced Cognitive Technologies and Applications*, 18–23. ThinkMind.

Prince, P. (1996). Second language vocabulary learning : The role of context versus translations as a function of proficiency. *The modern language journal, 80*(4), 478–493.

Procter, P. (1978). *Longman Dictionary of Contemporary English (LDOCE).* Essex, UK : Longman Group Ltd.

Procter, P. (1995). *Cambridge International Dictionary of English (CIDE).* Cambridge University Press.

Robert, P., Rey-Debove, J. et Rey, A. (1979). *Le petit Robert.* le Robert.

Roget, P. M. (1911). *Roget's Thesaurus of English Words and Phrases...* TY Crowell Company.

Russell, S., Norvig, P. et Intelligence, A. (2010). Artificial intelligence, a modern approach. *Artificial Intelligence. Prentice-Hall, Egnlewood Cliffs, 25*, 27.

Schmitt, N. (2008). Instructed second language vocabulary learning. *Language teaching research, 12*(3), 329–363.

Schmitt, N. (2010). *Researching vocabulary : A vocabulary research manual.* Springer.

Sowa, J. F. (2000). *Knowledge representation logical, philosophical, and computational foundations.* Pacific Grove, Calif. ; Toronto : Brooks/Cole.

Spencer, A. (2002). The handbook of linguistics. In M. Aronoff et J. Rees-Miller (dir.), *The handbook of linguistics* chapitre Morphology, 213–37. John Wiley & Sons, (1st éd.).

Steyvers, M. et Tenenbaum, J. B. (2005). The large-scale structure of semantic networks : Statistical analyses and a model of semantic growth. *Cognitive science, 29*(1), 41–78.

TLFi. (2019). *Trésor de la langue Française informatisé.* ATILF - CNRS et Université de Lorraine. Récupéré le 31 janvier 2019 de `http://www.cnrtl.fr/definition/dictionnaire`

Toutanova, K., Klein, D., Manning, C. D. et Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. Dans *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, 173–180., Stroudsburg, PA, USA. Association for Computational Linguistics.

van Sterkenburg, P. (2003). *A practical guide to lexicography*, volume 6. John Benjamins Publishing.

Vazirani, V. V. (2006). *Algorithmes d'approximation. Traduction de : Algorithms.* Collection IRIS. Paris : Springer.

Vincent-Lamarre, P., Massé, A. B., Lopes, M., Lord, M., Marcotte, O. et Harnad, S. (2016). The latent structure of dictionaries. *Topics in cognitive science, 8*(3), 625–659.

Weisstein, E. W. (2003). *Graph diameter.* Wolfram Research, Inc. Récupéré le 31 janvier 2019 de `http://mathworld.wolfram.com/GraphDiameter.html`

Wenski-Béthoux, C. (2005). *Utilisation de produits multimédia pour la construction de compétences lexicale : analyse linguistique, psycholinguistique et didactique des apports des cédéroms, des sites Internet et du travail en tandem pour l'apprentissage de l'allemand langue seconde.* (Thèse de doctorat). Lyon 2.

West, M. (1953). *A general service list of English words : with semantic frequencies and a supplementary word-list for the writing of popular science and technology.* Addison-Wesley Longman Limited. Récupéré le 31 janvier 2019 de `http://jbauman.com/aboutgsl.html`

Wilson, M. (1988). Medical Research Council (MRC) psycholinguistic database : Machine-usable dictionary, version 2.00. *Behavior research methods, instruments, & computers*, *20*(1), 6–10.

Wordsmyth. (2017). *Wordsmyth.* Wordsmyth. Récupéré le 31 janvier 2019 de `https://www.wordsmyth.net`

Yampolskiy, R. V. (2013). Turing test as a defining feature of ai-completeness. In *Artificial intelligence, evolutionary computing and metaheuristics* 3–17. Springer.