Contributions au Traitement Automatique des Langues et à un domaine d'application, la Communication Alternative et Augmentée

Didier Schwab

Université Grenoble Alpes

Habilitation à diriger des recherches

Spécialité : **Informatique**

Présentée par

Didier Schwab

préparée au sein du **Laboratoire d'Informatique de Grenoble** dans **l'École Doctorale Mathématiques, Sciences et Technologies de l'Information, Informatique**

Contributions au Traitement Automatique des Langues et à un domaine d'application, la Communication Alternative et Augmentée

Thèse soutenue publiquement le **8 décembre 2021** devant le jury composé de :

Madame, Sophie Dupuy-Chessa
Professeure à l'Université Grenoble-Alpes (Présidente)
Monsieur Emmanuel Morin
Professeur à l'Université de Nantes (Rapporteur)
Andrei Popescu-Belis
Professeur à l'HES-SO, Vaud, Suisse (Rapporteur)
Pierre Zweigenbaum
Directeur de recherche au CNRS (Rapporteur)
Pierrette Bouillon
Professeure, Doyenne à l'Université de Genève (Examinatrice)
Laurent Besacier
Scientifique principal à Naver Labs Europe, Grenoble (Examinateur)
Mathieu Lafourcade
Maître de conférences, HDR à l'Université de Montpellier (Examinateur)

Table des matières

Chapitre 1

Introduction générale

Un jour que je courais en forêt, je me suis retrouvé face à un cheval. À quelques mètres, majestueux, une robe magnifique, à la fois marron et scintillante. Nous nous sommes regardés. Je me suis lentement approché, j'ai posé ma main sur son dos, je l'ai caressé. Il a henni. Je l'ai salué. Je suis reparti.

Je me souviens de ma fille, allongée dans son lit, au milieu du petit salon près de la télé. Ses grands yeux me regardaient, l'air repu. Était-ce la même qui braillait 20 minutes auparavant tiraillée par la faim ?

Que feriez-vous si vous étiez privé de la parole, peut-être même incapable de faire le moindre geste ? Cette situation peut arriver à tout le monde. Elle peut arriver simplement après une opération un peu fatigante ou un accident. Dans une moindre mesure, elle peut également arriver si vous êtes dans un pays dont vous ne parlez pas la langue. On vous montrera des images, des pictogrammes pour que vous puissiez faire passer votre message, où vous voulez aller, que vous avez soif, que vous avez mal au ventre. Une telle communication est appelée Communication Alternative et Augmentée (CAA).

Communiquer est un défi. Communiquer touche à la compréhension de l'autre, dans ses différences, ses limites… et beaucoup les nôtres. Nous n'avons pas d'accès direct à la cognition de l'autre [1]. Considérons chaque sujet comme un bloc, un tout qui nous délivre des signes qu'il conviendra d'interpréter pour comprendre le message.

Évidemment, tout cela a ses limites. Quel signe ? Quel message ? Comment ? Qu'est ce que comprendre ? Qu'est ce qu'interpréter ? Pour quoi faire ? Pourquoi

1. En laboratoire, il existe bien entendu des outils pour permettre une certaine analyse des activités mais cela sort de mon domaine de compétences et je ne l'aborde pas ici.

faire ? Autant de sujets abordés ici et déjà largement abordés ailleurs et auxquels nous n'apporterons pas de réponse complètement définitive.

Certaines personnes sont incapables ou difficilement capables de parler. Ces personnes sont empêchées de parler, certes, mais elles ne sont pas pour autant empêchées d'interagir avec leur environnement. Cela même si les gestes sont maladroits voire impossibles. Et surtout, elles ne sont généralement pas empêchées de communiquer. Il existe tout un ensemble de modes de communication langagiers ou non non-langagiers [2]. La figure 1.1 schématise les liens entre interaction, communication, langage et communication non-langagière.

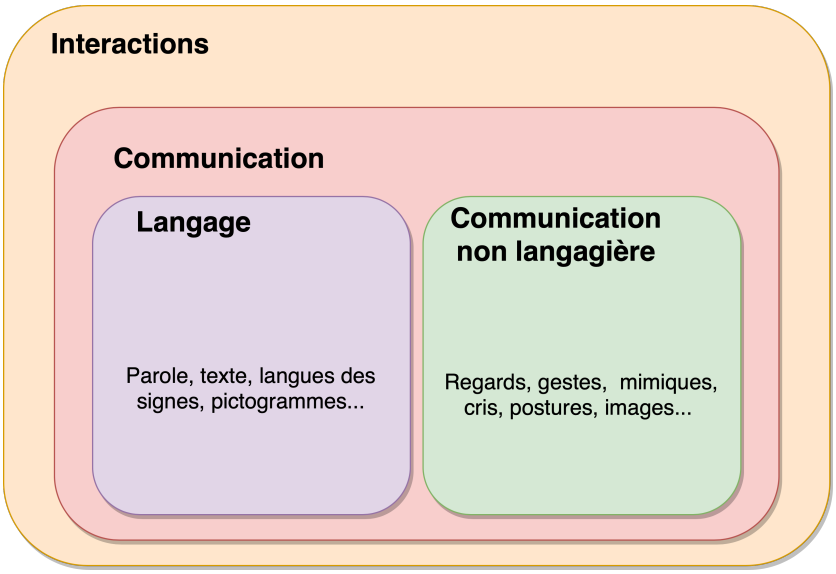

FIGURE 1.1 – Les différents modes langagiers et leur inclusion dans la communication et les interactions. Ici, je considère que l'interaction devient communication lorsqu'il y a message délivré que ce soit volontaire ou non. Cette définition implique nécessairement que le message soit réceptionné et interprété.

Nous en partageons même une partie avec le monde animal et cela peut-être d'autant plus que notre dernier ancêtre commun est proche [3]. Ma rencontre avec le cheval ne m'a-t-elle pas donné l'impression que nous nous étions compris ? J'avais, semble-t-il, parfaitement interprété la faim de ma fille. Aucun des deux ne m'a pourtant parlé… si ce n'est métaphoriquement.

2. Le lecteur pourra consulter, par exemple, le chapitre 3 de (Moeschler, 2020) pour en savoir plus.

3. Voir, par exemple, (Moeschler, 2020, p. 66 et suivantes) ou (Picq, 2008) qui donnent même quelques exemples de ce qui est de la CAA avec des animaux même s'ils ne la nomment pas ainsi.

C'est à cette métaphore que s'intéresse la Communication Alternative et Augmentée. En effet, ces situations de handicap peuvent être plus longues voire définitives et le besoin de communication devient alors plus complexe et peut ainsi recouvrir l'ensemble des objectifs d'une communication classique. C'est la situation que peuvent connaître les personnes touchées par des maladies neurodégénératives, qu'elles soient liées à la démence (syndrome d'Alzheimer, syndrome de Parkinson…) ou non (sclérose latérale amyotrophique – SLA – connue également sous le nom de maladie de Charcot).

Ces handicaps peuvent survenir de manière bien plus précoce dans la vie. Un nourrisson peut être victime d'une naissance prématurée, d'une anormalité chromosomique (syndrome de Rett, d'Angelman, de Prader-Willis…), d'infections ou d'accidents. Il peut avoir un handicap physique sévère – grandes difficultés pour utiliser ses bras et ses jambes – et un handicap cognitif – désordre neurodéveloppemental – les deux généralement liés.

1.1 La Communication Alternative et Augmentée

Selon l'ASHA (*American Speech-Language-Hearing Association*[4]), la Communication Alternative et Augmentée répond aux besoins des individus avec des troubles de communication importants et complexes caractérisés par des déficiences de la parole, que ce soit en production ou en compréhension.

On dit que la CAA est :

— *augmentée* lorsqu'elle est utilisée pour compléter un langage préexistant ;

— *alternative* lorsqu'elle est utilisée en remplacement d'un langage non existant ou dysfonctionnel ;

— *temporaire* lorsqu'elle est utilisée par les patients en postopératoire ou en soins intensifs ;

— *permanente* lorsqu'elle est utilisée par une personne qui aura besoin d'une forme quelconque de CAA tout au long de sa vie.

La CAA utilise une variété de techniques et d'outils :

— objets tangibles (boutons, claviers, images plastifiées…) ;

— gestes comme ceux d'une langue des signes, et aussi des gestes associés au discours ou à des images comme pour le Makaton* ;

— pictogrammes, images correspondant à un ou plusieurs mots ;

4. https://www.asha.org/practice-portal/professional-issues/augmentative-and-alternative-communication/

— logiciels de synthèse vocale qui permettent de synthétiser la parole à partir de texte ou de pictogrammes ;

— oculomètres (eye-trackers) qui permettent aux personnes qui ne peuvent pas utiliser leurs membres d'interagir avec un ordinateur.

Nos travaux traitent plus particulièrement des derniers : grilles de communication avec retour vocal, utilisation de pictogrammes et d'oculomètres.

FIGURE 1.2 – *Divers objets tangibles de la CAA.*

FIGURE 1.3 – *Association de boutons et de pictogrammes.*

FIGURE 1.4 – *Un enfant appuie sur un pictogramme, son appareil synthétise le nom du pictogramme afin que le message puisse être entendu par l'interlocuteur.*

FIGURE 1.5 – *Une enfant joue à GazePlay, l'interaction avec l'ordinateur est réalisée ici grâce à un oculomètre (eye-tracker).*

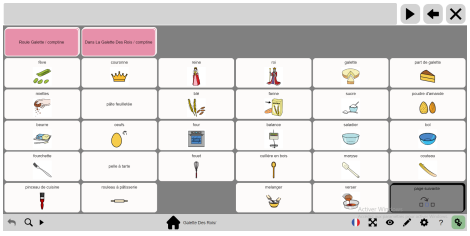

FIGURE 1.6 – *Un exemple de grille de communication avec retour vocal. Les grilles les plus complexes devraient permettre de composer tout message exprimable oralement (exemple avec InterAACtion AugCom dont je coordonne le développement).*

1.2 Cadre de mes recherches

Mes recherches se situent dans le domaine du Traitement Automatique des Langues et de la Parole (TALP). Je m'intéresse en particulier à **la représentation, l'acquisition et l'exploitation du sens pour et par la clarification des données langagières**. Je participe à des travaux autour de l'interopérabilité des bases lexicales, de la désambiguïsation lexicale, de la Traduction Automatique, de l'aide à la détection de plagiat ou de la Recherche d'Information incluant les fondements des modèles et leurs architectures.

Je suis membre du Laboratoire d'Informatique de Grenoble, plus particulièrement de l'équipe GETALP* (Groupe d'Étude en Traduction Automatique/Traitement Automatisé des Langues et de la Parole), équipe née en 2007 quelques mois avant mon arrivée. Issue [5] de l'union vertueuse [6] de chercheurs en traitement de l'écrit et de la parole, le GETALP est une équipe pluridisciplinaire (informaticiens, linguistes, phonéticiens, traducteurs…) dont l'objectif est d'aborder tous les aspects théoriques, méthodologiques et pratiques de la communication et du traitement de l'information multilingue (écrite ou orale).

Je suis informaticien, c'est la discipline pour laquelle j'ai été formé et que j'enseigne à l'Université pour des spécialistes [7] comme pour des non-spécialistes [8] depuis une vingtaine d'années. C'est également comme informaticien [9] que j'ai été élu dans plusieurs instances de l'Université Grenoble 2 fusionnée ensuite dans les différentes versions de l'Université Grenoble Alpes [10].

Jusqu'à ce jour, j'ai participé à l'encadrement de sept doctorantes et doctorants. Six ont soutenu et une est en phase de rédaction pour une soutenance d'ici mars. Pour deux d'entre eux, Hang Le et Zae Myung, je suis officiellement directeur de thèse par autorisation spéciale de l'école doctorale MSTII*. Dans les mois qui

5. Paragraphe largement inspiré des différents rapports et projets nécessitant une présentation de l'équipe, sa rédaction est principalement collective.

6. Par hypothèse de construction, et aussi par ses résultats collectifs, si l'on en croit les rapports de l'HCERES*, l'organisme chargé d'évaluer les structures de recherche publique en France.

7. Étudiants en DUT et BUT d'informatique, Master 2 d'informatique MOSIG* où je suis responsable du module *Speech and Natural Language processing*.

8. Plusieurs masters (Master Technologie et Handicap, Paris 8 ; Master neuropsychologie de l'enfant, UGA ; master AI4oneHealth, UGA ; Master Ingénierie de la finance, Ensimag, Grenoble INP-UGA) auxquels je présente mes recherches sur les vecteurs en TALN et celles concernant la Communication Alternative et Améliorée ainsi que les bases de l'informatique, les bases de données et la conception de sites Web pour des étudiants en DUT et BUT techniques de commercialisation.

9. J'appartiens à la section 27 du Conseil national des universités.

10. Comité d'hygiène, de sécurité et des conditions de travail - 2012-2016 ; Conseil Académique, commission recherche - 2020 - 2023

viennent, il est prévu que je devienne directeur ou co-directeur de plusieurs thèses. Les détails sont donnés en **??**.

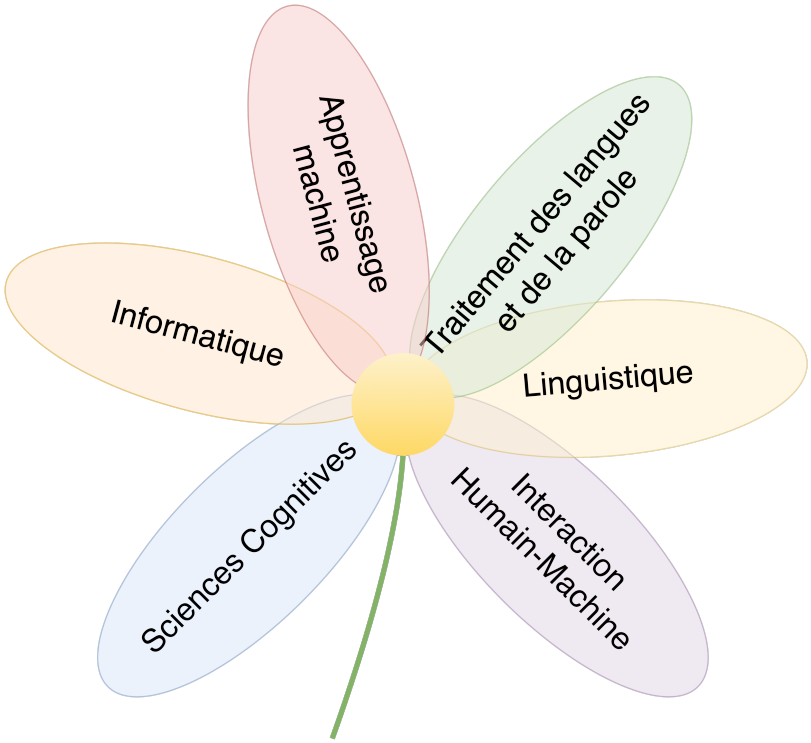

FIGURE 1.7 – Vue à gros grain des domaines liés à mes recherches

Au sein de l'informatique, si mes recherches se situent principalement dans le domaine du Traitement Automatique des Langues et de la Parole, elles touchent aussi d'autres domaines (voir figure 1.7). Je travaille également avec d'autres informaticiens (Recherche d'Information, Interaction Humain-Machine…) et avec des chercheurs d'autres disciplines (linguistes, psycholinguistes, traductologues). Les circonstances m'ont permis, ces dernières années, d'élargir le spectre non seulement aux médecins et aux spécialistes des sciences de la rééducation et de la réadaptation ainsi qu'à d'autres professionnels (orthophonistes, ergothérapeuthes, éducateurs spécialisés et éducatrices spécialisées…) mais aussi avec des utilisateurs dans leur vie quotidienne (enfants et adultes en situation de handicap cognitif et membres de leurs familles).

La découverte de la Communication Alternative et Augmentée m'a rapidement conduit à me demander comment les travaux que nous menions au GETALP pourraient être bénéfiques à la CAA et, suivant un schéma de pensée classique, comment une boucle retour pourrait permettre au Traitement Automatique des Langues

et de la Parole de bénéficier de la CAA.

Certaines de mes recherches sont désormais en TALP mais pas en CAA,celles citées en début de section, et certaines autres sont en CAA mais pas en TALP comme mes travaux sur les jeux sérieux ou les logiciels de scènes visuelles. Enfin, certaines recherches concernent les deux domaines comme ceux concernant l'interopérabilité des pictogrammes, Propicto qui concerne l'étude du passage parole vers séquences de pictogrammes ou la conception automatique de grilles de communication par des corpus transcrits.

Dans ce mémoire, j'essaye ainsi d'expliquer comment mes travaux tout en restant axés sur le sens et sa clarification sont passés petit à petit de l'écrit aux pictogrammes en intégrant la parole et le regard ; comment je me suis intéressé à la désambiguïsation lexicale, aux représentations vectorielles pour le TALP, à la traduction automatique du texte et de la parole jusqu'à des travaux autour du dialogue. C'est ce cheminement qui constitue en quelques sortes le fil rouge de ce récit.

Ces diverses recherches ne se sont pas toujours faites dans le contexte de la CAA puisque c'est un domaine dont je ne connaissais pratiquement rien il y a encore six ou sept ans. Ma deuxième fille, Tess, est née en 2012 avec un syndrome de Rett*. Une fois passé le diagnostic et la sidération qui accompagne ce genre de nouvelle, j'ai commencé à regarder ce qui se faisait pour aider à la communication des enfants comme Tess [11]. J'en ai tiré quelques réflexions qu'il me parait importantes de partager, car elles constituent le socle de ma démarche et de certains des choix que j'ai faits.

1.3 Sortir du cercle vicieux de Vygotski.

Les personnes qui ne parlent pas ou peu sont encore très mal intégrées dans notre société. Cette situation est encore plus vraie lorsqu'elle est associée à une déficience intellectuelle. Dès les années 1930, Lev Vygotski (1896-1934) identifiait un cercle vicieux pour les enfants avec surdité (voir figure 1.8). *« L'éducation sociale s'arrête devant le langage insuffisamment développé, le langage insuffisamment développé entraîne la séparation de la collectivité, et la séparation de la collectivité freine en même temps l'éducation sociale et le développement du*

11. Je demande au lecteur un peu d'indulgence pour avoir introduit ici des éléments personnels mais ils me paraissent importants pour comprendre, et ma démarche, et la place qu'a pris ce sujet dans mes recherches et le lien avec mes recherches initiales. Qu'il sache que je m'interroge toujours sur la légitimité que j'ai à travailler sur un sujet de recherche qui est si important pour ma vie personnelle et s'il ne vaudrait mieux pas être plus distant. Je constate toutefois que ce sujet serait bien moins traité dans l'industrie ou dans l'Université, que ce soit en France ou en Europe, sans les gens impliqués personnellement, que ce soit via des associations ou des fondations.

langage. » (Petitpierre et Barisnikov, 1994). La sortie de ce cercle vicieux passe, d'après Vygotski, par une meilleure connaissance du potentiel de communication des personnes, et par le développement de ce potentiel, via ce qu'il appelait déjà la polyglossie [12] ; c'est-à-dire par la multiplication des formes de développement du langage.

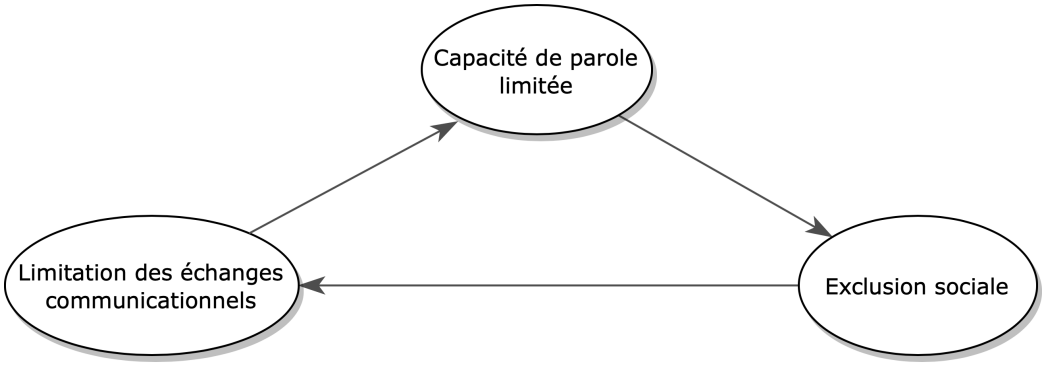

Figure 1.8 – Cercle vicieux décrit par Vygotski

À cause des limites extrêmes liées au polyhandicap, les limites en autonomie (pour pouvoir se déplacer, se nourrir, boire…), les relations avec les autres (communication, compréhension…), il reste très difficile d'évaluer exactement ce que ces personnes peuvent ou ne peuvent pas comprendre. Beaucoup de tests, basés sur la parole, ne sont pas adaptés. Cette situation rend la sortie de ce cercle vicieux d'autant plus difficile.

Néanmoins, la polyglossie décrite par Vygotski s'est tout de même grandement enrichie ces trente dernières années avec le développement de la CAA. Cette dernière, évidemment largement perfectible, a pu faire ses preuves, comme l'ont montré, par exemple, les études de Chirvasiu et Simion-Blândă (2018) ou de Sigafoos *et al.* (2016). L'exploitation du pouvoir du regard permet de proposer des solutions encore plus efficaces comme celles décrites par Gillian S. Townend (2016) ou S. Patachon (2015), qui ont mis en place une CAA basée sur le regard, avec deux jeunes filles atteintes du syndrome de Rett. Enfin, notons l'initiative du C-BiLLT (Geytenbeek *et al.*, 2014), un test destiné à évaluer la compréhension des enfants en situation de polyhandicap et mettant en jeu des oculomètres. Ce test n'a toutefois pas encore été porté en français.

La sortie de ce cercle vicieux ne pourra ainsi passer que par la mise en place d'outils permettant d'aider au mieux la mise en place d'une Communication Al-

12. Voir ce billet sur le blog d'Amélie Rochet-Capellan, membre du groupe InterAACtion https://adorabletoi.com/2019/02/23/vygotsky-un-petit-passage-dont-je-veux-garder-la-trace/.

ternative et Augmentée s'intégrant et se développant en même temps que l'enfant et que son environnement, une sorte de co-évolution*[13] guidée vers la sortie de ce cercle vicieux. Ce problème n'est évidemment pas que technique, comme nous allons le voir dans la suite de ce chapitre.

1.4 La CAA, mythes et réalité

On peut considérer que nos[14] travaux concernent plus particulièrement des personnes en situation de handicap complexe, associant des déficits perceptifs, moteurs et cognitifs. Ces déficits peuvent avoir des origines multiples et affectent la production de la parole aussi bien que sa perception. Parfois, les personnes ne peuvent pas du tout parler ou bien en deçà de ce qu'elles voudraient et auraient la capacité à dire. Dans notre expérience, assise sur un certain nombre de contacts directs, il s'agit non seulement d'enfants, mais également d'adultes, simplement parce que les enfants grandissent et restent longtemps dans les centres[15].

Romski et Sevcik (2005) dressent un état de l'art sur un certain nombre de « *mythes* », d'idées fausses largement répandues, autour de la mise en place d'une Communication Alternative et Augmentée avec des enfants.

Mythe 1 : la CAA est le dernier recours dans la prise en charge de la parole et du langage[16].

Mythe 2 : la CAA entrave ou arrête le développement de la parole[17].

Mythe 3 : les enfants doivent posséder certaines compétences pour pouvoir bénéficier de la CAA[18].

Mythe 4 : les systèmes avec sortie vocale conviennent seulement aux enfants avec une cognition intacte[19].

Mythe 5 : les enfants doivent avoir un certain âge pour bénéficier de la CAA[20].

13. Les termes suivis du caractère * sont définis dans le glossaire.
14. Le « *nous* » ici n'est pas un pluriel de modestie. Ce « *nous* » se réfère à toutes les étudiantes et chercheuses ainsi qu'à tous les étudiants et chercheurs qui ont contribué à ces travaux.
15. Il faut aussi noter que la situation que je décris dans la suite de cette section est également vraie pour les adultes en allant même jusqu'aux maladies neuro-dégénératives souvent liées à l'âge. Les thérapeutes et aidants se retrouvent souvent tout aussi désemparés et nous contactent pour voir si nous pourrions les aider.
16. Myth 1 : AAC is a "last resort" in speech-language intervention.
17. Myth 2 : AAC hinders or stops further speech development.
18. Myth 3 : Children must have a certain set of skills to be able to benefit from AAC.
19. Myth 4 : Speech-generating AAC devices are only for children with intact cognition.
20. Myth 5 : Children have to be a certain age to be able to benefit from AAC.

Mythe 6 : il y a une hiérarchie des représentations des symboles qui va des objets aux mots écrits (orthographe traditionnelle) [21].

Ces mythes sont basés sur des réticences que l'on voit quotidiennement dans les écoles et que véhiculent encore aujourd'hui certains professionnels. Ces mythes constituent le principal frein à la mise en place d'une CAA et confinent les enfants dans le cercle vicieux de Vygotski déjà évoqué. Cependant, mon propos n'est pas de dire que la mise en place d'une CAA va brusquement permettre aux enfants en question de se développer normalement.

Nous savons que permettre certaines interactions comme celles entre le regard et la machine permet immédiatement à beaucoup de personnes d'accéder à des jeux vidéos (Schwab *et al.*, 2018, 2020a). De plus, s'ils sont à la fois sérieux et ludiques, ces jeux permettent d'acquérir certaines compétences rapidement. Je pense en particulier à la compétence de sélection que l'on retrouve, par exemple, dans les grilles de communication. Les personnes sélectionnent un pictogramme et un retour vocal énonce le texte associé. Les joueurs comprennent rapidement le principe des jeux qui utilisent cette sélection comme, par exemple, celui qui consiste à casser un œuf en le regardant fixement quelques secondes. À l'inverse, la composition de messages avec une grille de pictogrammes peut prendre plusieurs mois de bain langagier* qui consiste en l'observation et en l'écoute [22] de l'entourage pour comprendre l'organisation de la grille et la façon de composer un message compréhensible.

Il existe d'autres réticences sociétales sur lesquelles nous ne nous étendrons pas ici. La France est, par exemple, épinglée par l'ONU* pour le manque d'accessibilité des bâtiments ou l'exclusion *de facto* des enfants handicapés de l'école (Devandas-Aguilar, 2019). Mon propos ici est simplement de constater que lever ces mythes et mettre en place une CAA permet une amélioration globale de la qualité de vie de la personne en situation de handicap et de son entourage, comme le montre la littérature et en particulier Beukelman et Mirenda (2017) [23].

Il existe ainsi en francophonie en général et en France en particulier une dynamique certaine pour la Communication Alternative et Augmentée. En témoignent les groupes Facebook dédiés dont le plus important rassemble plus de 3 800 personnes [24] ou les groupes qui militent pour plus de CAA dans la société comme le

21. Myth 6 : There is a representational hierarchy of symbols from objects to written words (traditional orthography).

22. Dans les cas les plus favorables où l'ouïe n'est pas déficiente.

23. Ce dernier est la version française d'une précédente édition de (Beukelman et Light, 2020) qui constituent tout deux les ouvrages de référence du domaine.

24. https://www.facebook.com/groups/1542829772615174 – consulté le 9 septembre 2021.

Collectif CAA - Ma voix, Mes droits [25] ou l'association ISAAC francophone [26]. On voit aussi essaimer, ici ou là, des initiatives locales de création d'écoles comme le centre des possibles [27] ouvert depuis février 2021 à Plomeur (Morbillan).

Le secteur industriel est par contre relativement concentré avec en particulier Tobii, entreprise suédoise de taille intermédiaire [28]. Dans le monde francophone, seule semble émerger la société belge Jabbla [29]. Les entreprises en France sont de petite taille voire de très petite taille (TPE) et se concentrent uniquement sur la revente de matériel. Ainsi, Cenomy, qui semble de loin la plus importante, ne compte que 14 employés [30].

La technologie n'est pourtant pas la solution parfaite clé en main. On ne compte plus les familles qui ont dépensé parfois plusieurs dizaines de milliers d'euros grâce à des dons, des week-ends à organiser des lotos, des courses caritatives, pour acheter du matériel qui finit dans le placard, faute de support humain, alors qu'il existe du matériel similaire à des prix dix fois moindres [31]. La même constatation peut être faite avec les institutions.

La formation sur la Communication Alternative et Augmentée est clairement un problème dans notre pays. Les orthophonistes n'ont pas de doctorat spécifique, alors que c'est pourtant le cas depuis longtemps dans d'autres pays. Il en est de même pour les ergothérapeuthes qui, jusqu'à présent, devaient faire une autre spécialisation : en France, les ergothérapeuthes qui ont fait des thèses de doctorat se comptent presque sur les doigts d'une main. Il existe heureusement quelques micro-entreprises de formation, de conseil et d'accompagnement, mais leur nombre, inférieur à une quinzaine [32], est bien trop faible.

25. https://www.facebook.com/groups/561807155202555 – consulté le 9 septembre 2021.
26. Association Internationale pour la Communication Alternative et Améliorée https://www.isaac-fr.org – consulté le 9 septembre 2021.
27. https://www.descarresdansdesronds.com/copie-de-au-centre-des-possibles
28. La société Tobii fournit des oculomètres et conçoit des logiciels. Elle regroupe environ 1000 personnes selon le rapport d'activité 2020 (Tobii, 2020) pour 40% du marché mondial des logiciels de CAA et 70% des ventes d'oculomètres. Une fission de l'entreprise qui aura pour conséquence l'indépendance des activités autour de la CAA est actuellement en cours, et est censée arriver avant la fin de l'année 2021. La nouvelle division CAA devrait alors avoir 400 à 600 employés.
29. https://www.jabbla.com/fr/ consulté le 12 septembre 2021.
30. Voir https://youtu.be/zDOn4YycBoI à 10'44 – consulté le 18 septembre 2021.
31. J'ai écrit ce petit texte explicatif à destination des famille et institutions à l'été 2018, je le mets à jour régulièrement en suivant les sorties logicielles ou matérielles. – https://lig-membres.imag.fr/schwab/2018/08/03/vers-une-democratisation-de-la-communication-alternative-et-amelioree-un-tobii-pour-moins-de-600-euros/
32. Estimation personnelle basée sur les groupes FaceBook et mes connaissances du terrain.

Partant de ce constat, il m'a semblé naturel que le Traitement Automatique des Langues et de la Parole puisse apporter sa pierre à l'édifice. Grâce au soutien financier initial de l'équipe GETALP et celui du LIG, Benjamin Lecouteux (UGA, LIG, GETALP) et moi avons pu commencer à étudier l'état de l'art et à développer quelques premiers prototypes et outils. En 2018, nous avons rencontré Amélie Rochet-Capellan (CNRS, GipsaLab) et Marion Dohen (Grenoble INP, UGA, GipsaLab) et créé tous les quatre le groupe InterAACtion*. En parallèle, ces outils ont permis l'émergence d'une communauté d'utilisateurs prêts à contribuer, par exemple, par leurs retours critiques, par leurs traductions, ou encore en codant et en mettant en place des outils d'aide au développement. Par les caractéristiques que j'ai présentée ci-dessus, la CAA est un domaine qui se prête bien à ce type de collaboration. De manière plus générale, certaines stratégies précédemment utilisées en TALP peuvent s'adapter à l'organisation et au financement des recherches autour de la CAA. Il s'agit ainsi, par exemple :

— d'exploiter et de constituer des ressources de type FAIR (**F**aciles à (re)trouver, **A**ccessibles, **I**nteropérables, **R**éutilisables) et librement accessibles.

— de créer, de manière plus générale, des outils et des services libres et accessibles : créés par la recherche et l'éducatif et pour la recherche et l'éducatif. Ce sont des outils que chacun peut utiliser ; chacun peut aussi participer à leur développement.

— de développer la science participative c'est-à-dire la science qui cherche à associer les citoyens intéressés directement ou indirectement par le sujet en bénéficiant de leur point de vue formel ou informel sur les fonctionnalités, les modes d'évaluation, d'acquisition et/ou l'éthique.

— de recueillir des données d'usage de la CAA, en étant le plus clair possible sur leur destination et leur anonymisation [33]. Après traitement, ces données peuvent ensuite être mises à disposition de la communauté.

— de développer une visibilité en formation et en recherche permettant de développer des partenariats avec des universités et des institutions (hôpitaux, centres spécialisés, associations, *etc.*) afin de financer nos recherches, et de rassembler des étudiants, professionnels et personnes en situation de handicap.

33. En d'autres termes, pas seulement en demandant d'accepter les cookies sur un site Web comme on le voit trop souvent.

1.5 Structure de ce mémoire

Ce mémoire est constitué de 4 chapitres principaux sans oublier deux annexes et d'un glossaire dans lequel je me suis efforcé de rendre le vocabulaire utilisé le plus cohérent possible. Il s'agit d'une tâche complexe et je ne doute pas qu'il contient des cohérences approximatives et des points de vue discutables. Il est conçu à partir du glossaire réalisé pour ma thèse (Schwab, 2005), complété année après année, et surtout au fur et à mesure de la rédaction de ce présent document.

La rédaction de ce mémoire m'a fait prendre conscience de l'instabilité du vocabulaire du Traitement Automatique des Langues et de la Parole. Je l'évoque à plusieurs reprises : des termes apparaissent brusquement, d'autres tombent rapidement en désuétude, d'autres encore semblent apparaître alors qu'un terme équivalent semble déjà exister. Le vocabulaire autour des représentations vectorielles est, à ce titre, assez caractéristique. Qu'est-ce qui différencie et rapproche dans mes articles ce que j'ai pu appeler successivement vecteurs, modèle vectoriel, plongement de mots, représentations continues, représentation continues et distribuées, modèles de langue, *fundation models* ? L'augmentation certaine du nombre de publications chaque année et l'arrivée de chercheurs issus de cultures différentes explique sans doute en partie ces changements mais je laisserai d'autres que moi en discuter.

C'est essentiellement ce changement de vocabulaire, et probablement une certaine appétence pour l'écriture, qui m'ont conduit à rédiger ces pages, inédites pour la plupart. On y retrouvera évidemment certains propos ou idées développées autrement dans certaines de mes publications antérieures. J'essaye de le préciser autant que possible dans ce cas là.

La figure 1.9 présente un schéma général de mes recherches et des éléments présentés ici.

Passée cette introduction, le deuxième chapitre détaille le cadre de mes recherches et en particulier l'écosystème dans lequel elles se situent. Je présente ensuite la méthodologie habituelle de travail de l'équipe de recherche GETALP et discute de ses applications, en particulier à la Communication Alternative et Augmentée. Enfin, seront abordées les recherches qui me paraissent les plus utiles pour développer la CAA, en mettant en relief certains des projets parmi les plus importants auxquels j'ai participé.

Le troisième chapitre est consacré à mes contributions à la constitution de données langagières statiques comme les corpus et les bases lexicales, qu'elles soient manuellement ou automatiquement créées, quel que soit leur mode langagier (écrit, oral, pictographique...). Je passerai en revue certaines de ces ressources et discute-

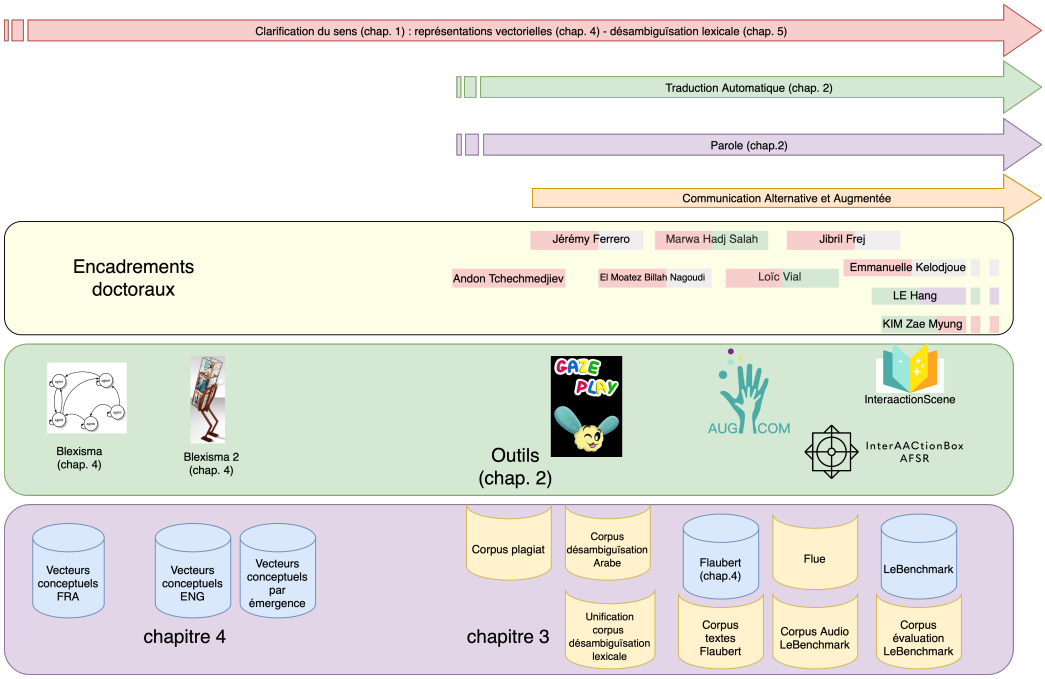

FIGURE 1.9 – Schéma général de mes recherches et des éléments présentés ici.

rai les hypothèses qui ont gouverné leur constitution. J'en ai déjà évoqué quelques-unes comme, par exemple, suivre les idées du *FAIR data*, mais je discuterai aussi de l'importance des langues, de l'alignement de données et de leur description fine.

Le quatrième chapitre traite des modèles pré-entraînés à base de vecteurs comme moyen de représenter et d'exploiter certains aspects du sens. Je présenterai ici un état de l'art ponctué des travaux que j'ai menés sur le sujet en collaboration avec une bonne vingtaine de personnes depuis mon master jusqu'aux tous derniers jours de l'été 2021.

Le cinquième chapitre revient sur un aspect particulier de la clarification de sens, la désambiguïsation lexicale. Je dresse ici un état de l'art ponctué de nos travaux et réflexions menés sur le sujet au cours des 10 dernières années. Je montre comment nous avons traité les langues moins dotées que l'anglais, exploité conjointement des connaissances issues de réseaux (connaissances explicites) et des connaissances issues de vecteurs (connaissances implicites) nous permettant d'obtenir des méthodes à l'état de l'art sur la plupart des langues.

Les premiers résultats obtenus conjointement à la Traduction Automatique et à la reconnaissance de la parole permettent de penser que l'unification d'un certain nombre des sujets abordés dans ce mémoire sera exploitable dans nos projets

autour de la Communication Alternative et Augmentée. Je reviendrai sur ce point dans une conclusion dans laquelle je présenterai également les pistes de travail et problèmes de recherches auxquels je souhaite m'attaquer dans le futur, toujours dans le cadre de coopérations multiples.

Chapitre 2

Cadre des recherches

Sommaire

Mes recherches se situent dans le domaine du traitement automatique des langues et de la parole (TALP). Je m'intéresse en particulier à **la représentation, l'acquisition et l'exploitation du sens pour et par la clarification des données langagières.**

Dans ce chapitre, je présente plus particulièrement et à grands traits l'écosystème dans lequel se placent mes recherches. Je présente ensuite la démarche scientifique que j'essaye de suivre avant de présenter plusieurs des recherches dans lesquelles j'ai été ou je suis fortement impliqué. Je me limite ici à celles qui me paraissent avoir une application directe dans la Communication Alternative et Augmentée. D'autres recherches comme celles que j'ai effectuées sur la détection du plagiat ou la recherche d'information neuronales sont présentées dans les annexes.

2.1 Écosystème de mes recherches

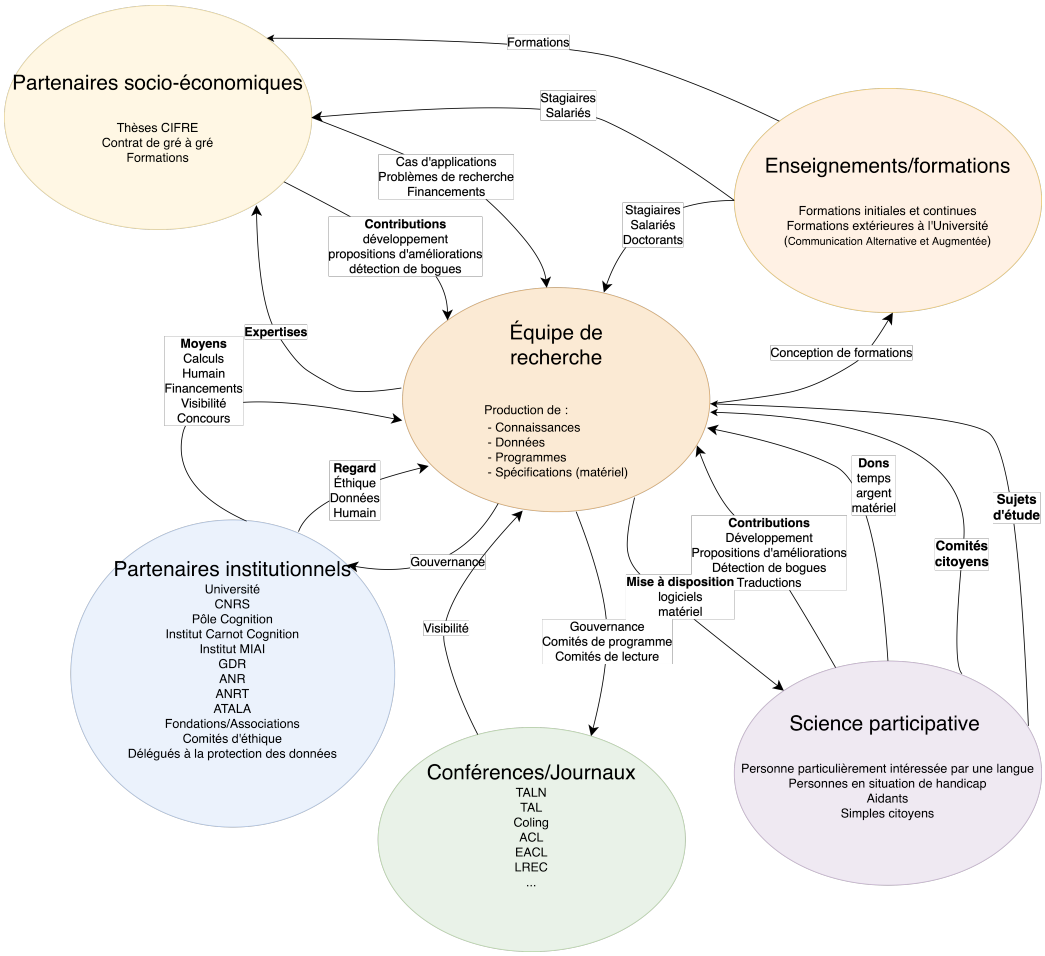

FIGURE 2.1 – Schéma général de l'écosystème de mes recherches et de leurs relations.

La figure 2.1 présente l'organisation globale de l'écosystème dans lequel se déroulent mes recherches. J'essaye de le voir comme un tout, un continuum où les différents aspects se complètent et se renforcent. Par beaucoup de côtés, il est très classique (l'articulation enseignement-recherche par exemple) et je ne suis pas le seul chercheur à essayer d'insérer mes recherches dans des collaborations similaires. D'autres collaborations me paraissent plus singulières comme, par exemple, celles avec des utilisateurs privés, en particulier celles avec des familles de personnes en situation de handicap dans le cadre de la science participative.

Partenaires socio-économiques : je distingue 3 sortes de collaborations :

— les laboratoires publics ou privés où les collaborations se déroulent autour d'un sujet d'intérêt commun, avec généralement l'implication de doctorants (NaverLabs, FAIR, GIPSA…) ;

— les entreprises ne disposant pas de laboratoire de recherche et ayant besoin de conseils pour développer des solutions nécessitant du TALP (OLFEO à Paris, TKM à Voiron…) ;

— des institutions comme des hôpitaux ou des centres spécialisés dans le handicap, qui cherchent à faire progresser la communication des personnes qu'elles accueillent (CHU Genève, IME les sources à Meylan, IME Le Plovier à Valence…).

Dans les 3 trois cas, outre le financement généralement associé à la collaboration, ces partenariats nous permettent de tester nos hypothèses de recherche de manière écologique et d'en formuler de nouvelles (voir section 2.2).

Enseignements/formations : les deux acceptions universitaires du mot formation sont ici concernées. On l'a vu ci-dessus, j'interviens dans plusieurs enseignements. Je délivre également dans le cadre du groupe InterAACtion* des formations dont le sujet principal est la Communication Alternative et Augmentée, formations issues de nos recherches [1] et qui sont testés en situation écologique dans les centres. Un de nos objectifs à moyen terme est de délivrer un Diplôme Universitaire axé sur la communication Alternative et Augmentée, qui manque clairement en France à l'heure actuelle.

Partenaires institutionnels : je regroupe sous ce vocable les organisations publiques ou d'intérêt public liées à mes recherches. Il s'agit ainsi de fournisseurs de moyens humains comme l'Université ou le CNRS ; de fournisseurs de calcul comme l'IDRIS* qui gère, entre autres, la grille de calcul Jean Zay* ; de financements ou de visibilité à travers des séminaires ou des capsules vidéo destinées aux autres chercheurs ou à de potentiels partenaires comme le pôle Grenoble Cognition* ou l'institut Carnot Cognition*. Derniers partenaires mais pas des moindres, les comités d'éthique et de protection des données. Indispensables aujourd'hui dans les recherches qui touchent aux individus, ce sont des partenaires présents en particulier pour mes recherches concernant le handicap et celles concernant le domaine médical. J'ai participé ou je participe à la gouvernance de plusieurs de ces institutions [2].

1. Que ce soit pour l'état de l'art de la recherche sur le sujet qu'à propos des outils que nous concevons et des recherches que nous faisons.

2. En septembre 2021, je suis membre du conseil d'administration de l'ATALA*, du conseil académique, de la commission recherche de l'Université Grenoble Alpes, correspondant de l'institut Carnot Cognition* après avoir été membre du conseil de laboratoire du LIG, du LIRMM ou encore représentant de la direction du LIG auprès de l'institut tremplin Carnot Cognition*. L'**??** présente l'ensemble de mes fonctions depuis le début de ma carrière.

Conférences/Journaux : il s'agit des conférences et journaux dans lesquels nous publions nos résultats et principalement de ceux du domaine. On peut citer, en premier lieu, la conférence TALN et son pendant jeunes chercheurs RECITAL, lieu de rencontre de la communauté, que je fréquente régulièrement depuis ceux de Nancy en 2002. Cette conférence internationale, sélective pour une conférence francophone, m'a permis de rencontrer la communauté et m'a accueilli dans une ambiance à la fois chaleureuse, cordiale et sérieuse. Aujourd'hui, je suis membre du comité permanent de la conférence, organisme chargé de superviser son organisation.

Je suis particulièrement fier d'avoir remporté le prix du meilleur article de TALN 2019 à Toulouse avec Loïc Vial et Benjamin Lecouteux.

Viennent ensuite Coling et les **ACL* [3], les grandes conférences internationales du domaine, et enfin LREC (*International Conference on Language Resources and Evaluation*), la conférence sur les ressources langagières, devenue une conférence importante parce qu'elle assure une visibilité certaine aux ressources publiées et parce qu'il reste difficile de publier sur les ressources, même si la tendance semble un peu moins vraie avec l'introduction aujourd'hui bien plus systématique de sessions dédiées.

Enfin, par mes travaux dans le domaine de la santé et le handicap, mes publications commencent à dépasser le cadre des conférences de mon domaine principal. Je n'oublie pas non plus les journaux avec plus particulièrement TAL (Traitement Automatique des Langues), le journal francophone de la communauté du traitement de l'écrit dont j'ai intégré le comité de rédaction en 2020. Du fait de l'accélération très importante des publications ces dernières années, avec l'émergence des dépôts comme ArXiv [4], il me semble clair que des moments de respiration et de mise en perspective des résultats et des conclusions à travers une publication plus approfondie comme l'est un journal sont indispensables [5].

Science participative : la science participative regroupe « *les formes de production de connaissances scientifiques auxquelles des acteurs non-scientifiques-professionnels — qu'il s'agisse d'individus ou de groupes — participent de façon active et délibérée* » (Houllier et Merilhou-Goudard).

Dans mes recherches, il s'agit plus précisément en majorité de personnes intéressées par un de nos sujets parce qu'il représente une part importante de leur

3. ACL, EACL, NAACL…

4. L'exemple de l'article sur BERT (Devlin *et al.*, 2019) en est un bon exemple. Les modèles et l'article sont diffusés en octobre 2018. L'article est meilleur papier à NAACL 2019, déjà cité plus de 600 fois dont une cinquantaine à cette même conférence.

5. Le même raisonnement doit s'appliquer également à la rédaction d'un livre ou d'une HDR.

identité propre et/ou parce qu'il pourrait constituer une amélioration sensible de leur qualité de vie.

Ce sont ainsi des personnes particulièrement intéressées par une langue (généralement minoritaire ou en danger) ou par la mise en place d'outils de Communication Alternative et Augmentée car elles sont elles-mêmes touchées ou parce que c'est le cas d'un ou d'une proche. Il s'agit ainsi de chercher à impliquer le plus possible ces personnes à nos recherches dans une démarche de science participative.

L'utilisation d'outils libres et gratuits, nous le verrons, constitue un élément essentiel de cette démarche : les personnes utilisent les outils et en font des retours critiques afin de les améliorer, ou bien participent directement à leur développement dans le cas où il s'agit d'informaticiens. L'implication peut également se faire en intégrant un comité citoyen dans un projet, comme nous le ferons dans le projet AAC4All mené par Jean-Yves Antoine (Université de Tours) et financé par l'Agence Nationale de la Recherche (2022-2025) [6] ou même en organisant et en coordonnant des récoltes de données comme dans le cadre du projet ParticipAACtion [7] qui consiste à créer et analyser un corpus audio-visuel de la communication des personnes qui ne peuvent pas s'exprimer par la parole, filmées dans leur quotidien par leurs proches ou dans le cadre du projet InterAACtionBox [8] dont l'objectif final est de concevoir un outil évolutif permettant de rendre possible la communication des personnes qui ne peuvent pas communiquer pleinement par la parole de manière transitoire ou permanente.

2.2 Méthodologie de travail

La méthodologie de travail du GETALP s'appuie sur des allers-retours continus entre collectes de données, investigations fondamentales, développements de systèmes opérationnels, applications et évaluations expérimentales (voir figure 2.2).

Je place ma démarche clairement dans celles de l'équipe GETALP qui développe dans plusieurs de ces travaux un tel cycle de développement [9]. Il s'agit ainsi,

6. Il est prévu 8 personnes dont 2 personnes en situation de handicap, 2 aidants*, 2 thérapeutes et 2 citoyens sans lien avec ces recherches. Son objectif est d'aider à l'analyse des besoins réels et de donner un avis sur les orientations prises et les résultats obtenus. L'hypothèse sous-jacente est que l'ajout de 2 citoyens ordinaires, qui ne sont pas des acteurs directs de la question de la Communication Alternative et Augmentée, permettra de mener une large réflexion sur les enjeux éthiques de cette recherche, tout en rapprochant le monde du handicap et le grand public.
7. Voir http://www.ParticipAACtion.com – voir section 2.4.1
8. http://interaactionbox.fr – voir section 2.4.2
9. Voir les rapports HCERES ou les présentations de l'équipe, on y retrouvera ces idées.

dans un premier temps, d'étudier un problème, son état de l'art et les outils qui sont proposés (études fondamentales) puis de spécifier et développer une première version d'un système opérationnel voire d'une d'application diffusables s'attaquant à ce problème avant de l'évaluer et de collecter des données puis de les étudier et, tout en suivant l'état de l'art qui évolue par ailleurs, d'en tirer les conséquences pour un nouveau cycle.

Les systèmes opérationnels sont plutôt des prototypes destinés à être utilisés par des utilisateurs aguerris, ou pour construire des applications, tandis que les applications sont destinées aux utilisateurs finals et généralement réalisés en collaboration avec des partenaires socio-économiques. La partie Applications reste ainsi parfois tout à fait optionnelle et nécessite un temps de travail dédié. En ce qui me concerne, je n'ai dirigé ce travail que dans le cas de la Communication Alternative et Augmentée (CAA) grâce au financement d'un ingénieur d'étude dédié au développement des applications, Sébastien Riou (voir section 3.2.2.2).

Notons bien que ce cycle n'est pas le cycle de développement d'un logiciel avec tests internes, intégration continue, *etc.* même s'il s'en rapproche par certains aspects. Il s'agit bien d'un cycle à considérer sur plusieurs mois voire plusieurs années et impliquant un bon nombre d'acteurs et, généralement, de financements.

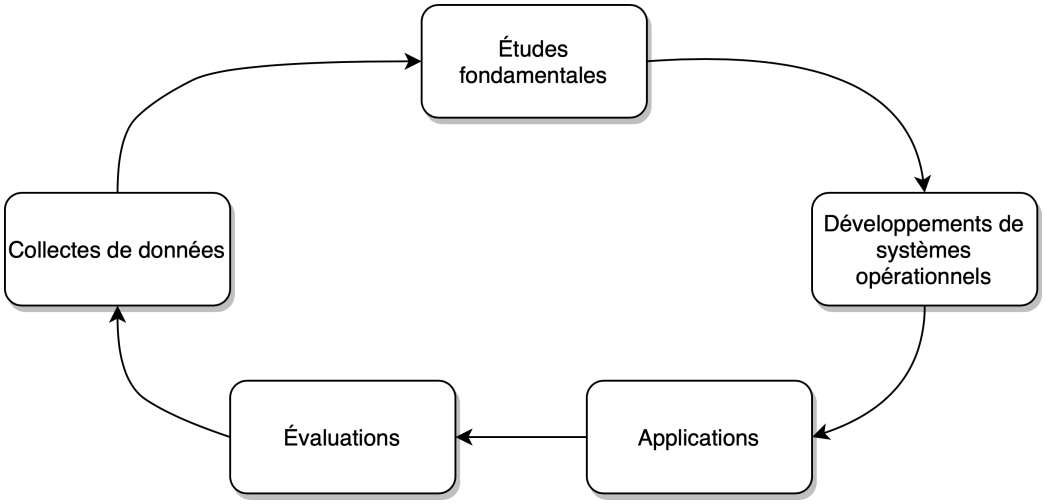

FIGURE 2.2 – Schéma général de la méthodologie de l'équipe GETALP également suivie dans mes recherches.

Dans le cadre du projet FlauBERT, un premier ensemble de modèles a été publié en décembre 2019 (Le *et al.*, 2020f,e) appliquant au français plusieurs méthodes à l'état de l'art. Ils ont été utilisés et étudiés dans différentes recherches – par exemple, Haley (2020); Aloui *et al.* (2020); Ghannay *et al.* (2020) – certaines

de leurs limites conceptuelles ont été mises en évidence [10] et font actuellement l'objet d'un nouveau projet.

Dans le cadre de la Communication Alternative et Augmentée, le développement du logiciel AugCom [11] a commencé par des études comparatives des différents logiciels destinés aux personnes en situation de déficit de la parole, sur les besoins exprimés par leurs familles et les professionnels qui travaillent avec eux. Un premier ensemble de fonctionnalités a été proposé avant d'être testé en instituts spécialisés ou dans des familles. Nous collectons alors les données d'utilisation (retours directs, données d'utilisation) afin d'améliorer nos connaissances sur les utilisations et d'améliorer le logiciel en retour.

Le projet le plus avancé à ce sujet est GazePlay, un ensemble de jeux sérieux développé depuis 2016, dont l'objectif est d'aider les enfants en situation de handicap cognitif à acquérir en jouant un certain nombre de compétences dont celle de savoir utiliser les interactions des logiciels de communication (voir section 2.4.2). GazePlay [12] en est à sa 10ème version majeure (nommée 1.10). La 9 (1.9), après avoir été testée par tout un ensemble de personnes en situation écologique (famille, thérapeutes, étudiants en neuropsychologie de l'enfance [13]) aura été testée plus de 8 mois et téléchargée plus de 1400 fois. Ses petits frères, AugCom et InterAACtionScene suivront le même chemin mais n'en sont qu'au tout début, aux premières évaluations.

2.3 L'obje(c)t(if) de mes recherches

Dès 2005, cherchant à trouver un titre couvrant non seulement mes travaux de thèse (Schwab, 2005) mais également l'ensemble de mes perspectives futures, j'ai proposé de dire que mes recherches concernent plus particulièrement **la représentation, l'acquisition et l'exploitation du sens pour et par la clarification des données langagières**.

Ces recherches, commencées dès mon Master, se sont peu à peu enrichies de nouveaux contextes d'utilisation. Elles ont, certes, bénéficié de l'évolution des ressources matérielles mais ont surtout bénéficié de l'évolution des ressources langagières.

10. L'architecture XLM choisie à l'époque ne permet pas, par exemple, de dresser la liste des mots les plus probables selon le modèle pour une phrase dont un mot est masqué. Ainsi dans « *le <mask> aboit.* », on s'attend à avoir une liste où ‹‹chien›› figure en bonne position.
11. http://augcom.net
12. http://gazeplay.net
13. Oriana Orlandi (2020), Aurane Malassagne (2020), Louis Lambert (2021), tous 3 en stage dans des institutions spécialisées principalement encadrés par Amélie Rochet-Capellan.

Ces deux dernières décennies, les capacités de calcul ont considérablement augmenté. Mes expériences débutées sur un simple PC de bureau se sont peu à peu déplacées sur des serveurs multicœurs, à des grappes de machines grâce à des systèmes multi-agents, puis à des grilles de dizaines de machines pour arriver aujourd'hui à des méthodes d'apprentissage profond distribuées sur grilles de milliers de GPU.

Dans le même temps, et c'est là un point essentiel pour moi, chercheur en traitement automatique des langues et de la parole, nous sommes passés d'une approche individuelle, un monde où chacun gardait ses ressources langagières pour lui, à une approche plus collective, monde où les ressources (données, codes, modèles, méthodes) sont bien plus partagées. Ainsi, si à l'époque, nous utilisions des données propriétaires, faute de mieux, j'utilise et produis désormais des données et des programmes libres d'accès, ce qui n'interdit plus la reproductibilité des recherches [14].

Dans le même temps, les méthodes et les intérêts évoluant, je suis peu à peu passé de travaux portant uniquement sur l'écrit, à des travaux sur l'oral puis à d'autres formes de communication, par exemple celles impliquant les pictogrammes.

La figure 2.3 présente les principaux domaines auxquels je m'intéresse. Dans la suite de ce chapitre, je rattache plus précisément à ces domaines certains de mes projets. Les deux chapitres suivants, l'un consacré aux ressources langagières et l'autre consacré aux modèles, font référence à ces différents travaux.

2.3.1 La clarification du sens

Mes travaux se placent dans le cadre de l'analyse sémantique d'énoncés écrits ou oraux vue comme une tâche générique par opposition à une analyse sémantique dans le cadre d'une application unique comme la traduction automatique, la catégorisation de texte ou la recherche d'information (Schwab, 2005).

Dans ce second cas, pour la création d'outils spécialisés, l'analyse sémantique n'est généralement pas explicite et l'analyse sémantique est alors généralement réalisée de manière indirecte. Si on considère la désambiguïsation lexicale, en traduction automatique, il s'agit alors d'exploiter les informations contrastives entre les langues (si on ne considère que ses acceptions principales, la ‹‹souris›› se traduit en anglais toujours par ‹‹mouse›› tandis qu'en malais, il se traduira par ‹‹tikus›› pour l'animal ou par ‹‹tetikus›› pour le périphérique électronique). Pour le même type d'ambiguïté, en recherche d'information, le co-texte suffit souvent également.

14. Nous y reviendrons, la reproductibilité des recherches n'est pas facile à garantir (Cohen *et al.*, 2018) mais il est certain qu'elle nécessite, en TALP, des données.

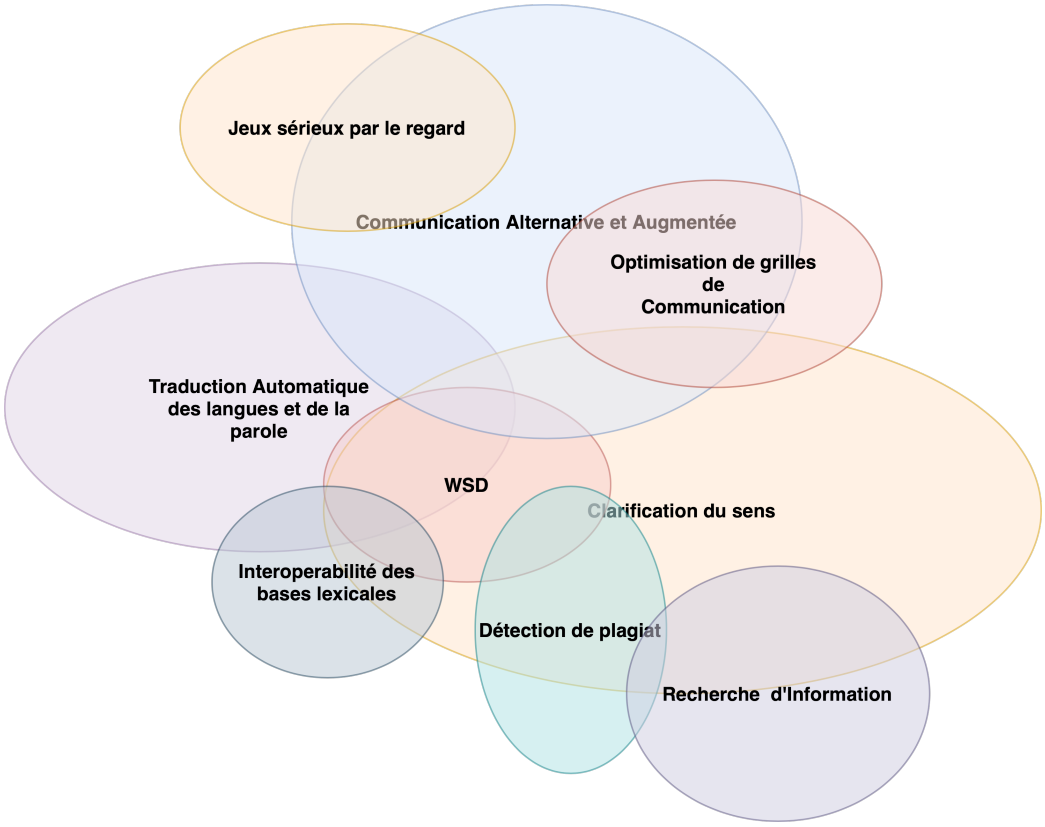

FIGURE 2.3 – Vue à à gros grain de mes recherches.

Ainsi, un utilisateur choisira ‹‹animal››, ‹‹rongeur››, ‹‹grise›› ou ‹‹ordinateur››, ‹‹fil››, ‹‹informatique›› pour clarifier sa requête.

2.3.1.1 Analyse sémantique

Dans ma thèse (Schwab, 2005), je décris l'analyse sémantique comme le calcul mais aussi le résultat de l'analyse d'un texte en langue naturelle. Il s'agit ainsi de donner une représentation du sens global du texte, sous forme vectorielle donc sous une forme continue [15] mais également sous une forme discrète avec un graphe dont l'objectif est de désambiguïser finement le texte. Je distinguais alors 5 types d'ambiguïté :

1) Le phénomène de l'ambiguïté lexicale, phénomène bien connu, qui nécessite d'identifier pour un mot donné, son interprétation la plus raisonnable comme dans la phrase « *La poule mange des grains* » où *poule*/**volatile** est bien plus probable

15. Ce terme est un abus de langage. Informatiquement, il s'agit d'une approximation de forme continue puisque les valeurs sont représentées dans un domaine qui reste fini.

que l'acception *poule*/**groupe d'équipes**.

2) La recherche des références qui consiste à identifier les mots que l'on ne peut interpréter qu'en fonction d'un autre mot du texte. On trouve dans cette catégorie, l'anaphore lorsqu'un pronom fait référence à un autre élément du texte comme dans « *L'homme marcha sur la queue du chien, il aboya.* », où ‹‹il›› est une anaphore de ‹‹chien››. Dans les références, on trouve également la relation d'identité qui lie les éléments d'un texte qui font référence à la même entité comme c'est le cas dans « *Le chat est monté sur la chaise. L'animal s'assoupit.* » pour ‹‹chat›› et ‹‹animal››.

3) La recherche des rattachements prépositionnels où il s'agit de chercher le lien de dépendance entre un groupe prépositionnel et sa tête syntaxique (verbe, nom, adjectif). Par exemple, dans la phrase « *L'homme regarde la fille avec un télescope* », est-ce l'homme qui utilise un télescope pour regarder la fille ou est-ce la fille qui porte le télescope ?

4) La recherche des Fonctions lexicales (FL) pour l'analyse qui permettent de modéliser les connaissances du monde et celles qui permettent de modéliser les connaissances lexicales (Schwab, 2005, p. 196). Il existe deux types de fonctions lexicales, les fonctions lexicales paradigmatiques qui formalisent les relations sémantiques qui existent entre les items lexicaux (synonymie, antonymie, méronymie, hyperonymie, ...) et les fonctions lexicales syntagmatiques c'est-à-dire les fonctions lexicales qui modélisent les combinaisons d'items lexicaux qui prévalent sur d'autres sans qu'il ne semble n'y avoir de motif logique (« *dormir profondément* » et non *« *dormir intensément* »). Pour les connaissances du monde, les FL permettent par exemple de lier les classes aux instances (Inst(‹*Homme politique*›) = ‹*François Hollande*› et Class(‹*François Hollande*›) = ‹*Homme politique*›.

5) La recherche des chemins interprétatifs c'est-à-dire les interprétations possibles pour un texte donné. Des exemples jouets classiques sont, par exemple, « *L'avocat est véreux.* ». Les deux interprétations « *L'homme de loi est malhonnête* » et « *Le fruit est rempli de vers* » sont plus raisonnablement possibles même si « *L'homme de loi est rempli de vers* » reste possible (pour des raisons médicales ou au cimetière) ou « *Le fruit est malhonnête* » dans une bande dessinée. Dans des cas plus réels, un manque de contexte (volontaire ou involontaire par l'auteur) peut entraîner une ambiguïté comme dans la phrase « *Elle est bien seule* » qui peut être la constatation de la solitude d'une personne ou qu'une personne apprécie d'être seule.

Notre hypothèse, raisonnable je crois, est que ces phénomènes sont liés entre eux et la résolution, même partielle, de l'un d'eux peut aider à résoudre les autres. Ainsi dans la phrase « *L' homme marcha sur la queue du chien, il aboya.* », résoudre l'ambiguïté lexicale du mot ‹‹chien›› (*chien*/**animal** plutôt que *chien*/**fusil**) aide à ré-

soudre l'anaphore du ‹‹il››, de même, savoir que ‹‹aboya›› est une action réalisable par un *chien*/**animal** et pas par un *chien*/**fusil** aide à la résolution de l'ambiguïté lexicale pour ‹‹chien››. De même, reconnaître la fonction lexicale d'instrument S_{inst} [16] entre ‹‹regarde›› et ‹‹télescope›› dans la phrase « *L'homme regarde la fille avec un télescope* » peut indiquer une préférence pour l'interprétation de l'homme qui utilise le télescope pour regarder la fille.

Cette thématique a couvert l'essentiel de mes travaux tout au long de ma carrière et ce n'est pas par hasard qu'elle occupe une place centrale dans mes recherches (voir section 2.3).

2.3.1.2 Projets liés à la clarification du sens

Deux des projets qui ont occupé mon temps depuis deux ans sont le projet FlauBERT et, le projet FLUE initialement lié au précédent.

Suivant une idée de Laurent Besacier et d'Alexandre Allauzen, j'ai participé à la rédaction de demande d'heures sur la grille de calcul Jean Zay. J'ai ensuite animé le groupe et co-encadré Hang Le, alors étudiante en master, avec Laurent Besacier. Elle a développé le code nécessaire basé sur la librairie XLM de Facebook [17] afin de pouvoir utiliser Jean Zay, en rodage à l'époque. Dans ce projet, largement informel et basé essentiellement sur une volonté commune de faire avancer certaines connaissances du domaine, mon rôle consiste à organiser la tenue des réunions, convier les participants, prendre part à plusieurs sollicitations scientifiques et grand public. Il faut noter que ce projet nous a également permis plusieurs collaborations et qu'il est devenu central à beaucoup de nos travaux.

L'idée de départ est que les modèles de langue pré-entraînés* étaient désormais indispensables pour obtenir des résultats à l'état-de-l'art dans de nombreuses tâches du TALN. Tirant avantage de l'énorme quantité de textes bruts disponibles, ils permettaient d'extraire des représentations continues des mots, contextualisées au niveau de la phrase.

L'efficacité de ces représentations pour traiter plusieurs tâches de TALN avait été démontrée pour l'anglais (Devlin *et al.*, 2019). Nous avons présenté et partagé FlauBERT, un ensemble de modèles de complexités différentes appris sur un corpus français hétérogène et de taille importante [18]. Dans un contexte de crise sanitaire, j'ai continué cette animation et une suite à ce projet a aujourd'hui commencé mais les résultats sont trop préliminaires pour que nous en parlions ici.

16. Voir (Schwab, 2005, p. 333) pour la liste complète des fonctions lexicales pour l'analyse que nous proposions alors.
17. https://github.com/facebookresearch/XLM
18. 70 Go de données - voir section 3.2.1.4

De son côté, le projet FLUE [19] nous a paru indispensable pour faciliter l'évaluation des vecteurs FlauBERT ou tout autre modèle sur des tâches du Traitement automatique de la langue française écrite. Il consiste à rassembler des corpus préexistants (classification de textes, paraphrase, inférence en langage naturel, analyse syntaxique, désambiguïsation automatique) et à faciliter leur utilisation (voir section 3.3.3).

FlauBERT & FLUE

Rôle Animateur du projet, gestion du site Web
Période Depuis 2019
Encadrement/Collaborations Hang Le (Master), Loïc Vial (doctorant), Jibril Frej (doctorant), Vincent Segonne (doctorant), Maximin Coavoux (CNRS, LIG), Benjamin Lecouteux (UGA, LIG, GETALP), Alexandre Allauzen (PSL Research University), Benoît Crabbé (Universite Paris Diderot), Laurent Besacier (UGA, LIG, GETALP)
Financement(s) principal(aux) Heures sur la grille de calcul Jean Zay du CNRS
Valorisation :
Ressources langagières Corpus en français nettoyés et normalisés (70 Go), FLUE [a]
Modèles FLAUBERT [b]
Références Le *et al.* (2020c,d)
Notes http://fluebenchmark.com

a. https://github.com/getalp/Flaubert/tree/master/flue
b. https://github.com/getalp/Flaubert

J'ai également encadré trois thèses liées à la clarification du sens, à commencer par celle d'Andon Tchechmedjiev (Tchechmedjiev, 2016) qui proposait d'exploiter des méthodes de désambiguïsation lexicales pour assurer l'interopérabilité sémantique multilingue des ressources lexicales. Les deux autres thèses étudiaient la désambiguïsation lexicale multilingue et ses liens avec la traduction automatique.

Grâce à la thèse de Loïc Vial, nous avons unifié l'ensemble des corpus en anglais annotés en sens [20] et proposé tout un ensemble de modèles pour la désambiguïsation lexicale. La thèse de Loïc a été effectuée au moment où les modèles d'apprentissage profonds devenaient prépondérants. Nous avons été ainsi parmi les premiers à en proposer pour la désambiguïsation lexicales jusqu'à (Vial *et al.*, 2019c), longtemps à l'état de l'art, qui proposait une manière originale de combi-

19. fluebenchmark.com
20. En sens issus de WordNet

ner apprentissage profond et informations issues d'un réseau lexical (WordNet).

Interopérabilité sémantique multilingue des ressources lexicales

Rôle Co-encadrant de thèse
Période 2012-2016
Encadrement/Collaborations Andon Tchechmedjiev (doctorant), Gilles Sérasset (UGA, LIG, GETALP), Jérôme Goulian (UGA, LIG, GETALP)
Financement(s) principal(aux) Allocation de recherche
Valorisation :

Ressources langagières Intégrées à DBNary [a]

Modèles *Lexema* [b], aujourd'hui, repris et intégré à *disambiguate* [c] de Loïc Vial

Références Tchechmedjiev *et al.* (2014), Tchechmedjiev (2016)

Notes ∅

a. http://kaiko.getalp.org/about-dbnary/
b. https://github.com/twktheainur/lexsema
c. https://github.com/getalp/disambiguate

Désambiguïsation lexicale de l'arabe pour et par la traduction automatique

Rôle Co-encadrant de thèse
Période 2015-2018
Encadrement/Collaborations Marwa Hadj-Salah (doctorante), Hervé Blanchon (UGA, LIG, GETALP), Mounir Zrigui (Université de Monastir, Tunisie)
Financement(s) principal(aux) Allocation de recherche (Tunisie) - Financements GETALP
Valorisation :

Ressources langagières UFSAC-FRA [a], UFSAC-ARA [b]

Modèles Hadj Salah (2018); Hadj Salah *et al.* (2016, 2018b,a)

Références ∅

Notes

a. https://zenodo.org/record/3549806#.YQKeCS0itpQ
b. https://github.com/getalp/UFSAC-ara

Dans la thèse de Marwa Hadj Salah, nous nous sommes plus particulièrement

intéressés à la désambiguïsation lexicale des langues moins dotées que l'anglais [21]. Les ressources langagières étaient ainsi particulièrement centrales et le manque de corpus tant pour évaluer que pour apprendre était important. Pour l'arabe, les seuls corpus que nous avons trouvé est un corpus de taille relativement modeste et annoté en Ontonote*. En parallèle, nous avons étudié à quel point il était possible de créer automatiquement des corpus annotés en sens pour n'importe quelle langue, en tirant parti de la grande quantité de corpus anglais annotés en sens issus du *Princeton WordNet* (Les UFSAC de Loïc Vial), et en utilisant un système de traduction automatique (voir section 3.2.2.2). Pour connaître la qualité des systèmes fabriqués grâce à ces ressources, nous avons proposé deux méthodes. Une première, de type *in vitro**, a consisté à utiliser une méthode semi-automatique inspirée de celles de (Tchechmedjiev, 2016) pour aligner les sens *Princeton WordNet* et ceux issus de l'Ontonote pour fabriquer un corpus annoté en sens *Princeton WordNet*. Une seconde, de type *in vivo**, consiste à analyser leur contribution à la performance d'une tâche de traduction automatique.

Modèles joints de clarification de texte et de traduction automatique statistique

Rôle Co-encadrant de thèse
Période 2016-2020
Encadrement/Collaborations Loïc Vial (doctorant), Benjamin Lecouteux (UGA, LIG, GETALP)
Financement(s) principal(aux) Allocation de recherche
Valorisation :

Ressources langagières Unification de corpus annotés en sens (UFSAC [a]), vecteurs de définitions du *Princeton WordNet* [b]

Modèles Disambiguate-Translate [c] (Boite à outils pour la désambiguïsation lexicale et la traduction automatique), désambiguïsateur neuronal à l'état de l'art [d]

Références Vial (2016, 2020); Vial *et al.* (2016, 2017a,b,d, 2018a, 2019a,c)

Notes Prix du meilleur article à TALN 2019 (Vial *et al.*, 2019b)

a. https://github.com/getalp/UFSAC
b. https://perscido.univ-grenoble-alpes.fr/datasets/DS117
c. https://github.com/getalp/disambiguate-translate
d. Aujourd'hui encore proche de l'état de l'art – https://github.com/getalp/disambiguate

21. Soit toutes les autres langues.

2.3.2 La traduction du texte et de la parole

Dès ma thèse (Schwab, 2005), j'identifiais la traduction automatique comme une application de mes travaux sur la représentation et l'exploitation du sens mais cette réflexion restait alors assez théorique. Mon postdoc dans un laboratoire spécialiste de traduction automatique, puis mon arrivée au GETALP m'ont permis de travailler de plus en plus dans ce domaine autour duquel s'articulent une grande partie des thèses que j'ai coencadrées et autour desquelles tournent celles que j'encadre [22]. Outre les thèses de Loïc Vial et Marwa Hadj Salah centrées sur la clarification du sens et la désambiguïsation lexicale en particulier pour et par la traduction automatique, j'encadre le travail de thèse de Hang Le, qui introduit la parole, et celle de Zae Myung Kim.

Dans sa thèse, Loïc Vial s'attaque de front à ces deux tâches historiques du TALP que sont la désambiguïsation lexicale (DL) et la traduction automatique (TA). Réalisée à une époque où les méthodes à base de réseaux de neurones et les représentations vectorielles des mots devenaient prépondérantes, cette thèse exploite les architectures neuronales de type *transformer* et les nouveaux modèles de langue pré-entraînés. Il s'agissait ainsi non seulement de développer des systèmes de DL et de TA plus performants, mais surtout de réunir de façon plus directe les deux tâches grâce à des modèles neuronaux conjoints, permettant de faciliter leurs interactions.

Hang LE, dans le cadre d'une thèse portant sur la traduction de l'oral vers un ensemble de langues écrites étudie, en particulier, les modèles pour exploiter au mieux les ressources disponibles tant en termes d'efficacité computationnelle (nombre de paramètres) qu'en termes de score (score BLEU). Trois publications ont été acceptées en septembre 2021. Une première concerne un décodeur dual qui effectue conjointement la reconnaissance automatique de la parole (ASR) et la traduction multilingue de la parole (Le *et al.*, 2020a). Une seconde publication concerne l'utilisation d'adaptateurs (Houlsby *et al.*, 2019) comme alternative à l'ajustement fin qui consiste à geler les paramètres pré-entraînés d'un modèle et à injecter des modules légers entre les couches, ce qui se traduit par l'ajout d'un petit nombre de paramètres entraînables spécifiques aux tâches (Le *et al.*, 2020a). Enfin, un dernier article est la participation à la campagne IWSLT 2021 (Le *et al.*, 2021a).

22. Les annexes précisent mon rôle officiel dans chaque encadrement (voir **??**).

Modèles polyglottes pour la traduction de parole des langues peu dotées

Rôle Directeur de thèse
Période Hang Le (doctorante), Juan Pino (FAIR*), Changhan Wang (FAIR), Jiatao Gu (FAIR), Laurent Besacier [a] (UGA, LIG, GETALP), Benjamin Lecouteux (UGA, LIG, GETALP)
Encadrement/Collaborations 2020-…
Financement(s) principal(aux) Contrat de recherche de gré à gré avec *Facebook Artificial Intelligence Research*
Valorisation :
Ressources langagières ∅
Modèles Contributions au dépôt du logiciel libre FairSeq* développé initialement par FAIR [b].
Références Le *et al.* (2020b), Le *et al.* (2021b)
Notes ∅

a. Laurent Besacier a quitté le LIG et le projet depuis début 2021.
b. https://github.com/formiel/fairseq/blob/master/examples/speech_to_text/README.md

Située dans le cadre de la traduction automatique de l'écrit massivement multilingue, la thèse de KIM Zae Myung a commencé en octobre 2020. Elle a déjà donné lieu à une publication (Kim *et al.*, 2021) étudiant plus particulièrement les têtes d'attention* dans un modèle traduisant plusieurs langues vers une langue (l'anglais) et montrant qu'un tiers des têtes peuvent être supprimées sans perte de qualité, ce qui peut réduire la taille et l'efficience* des modèles.

Apprentissage de représentations multilingues pour le traitement automatique des langues

Rôle Directeur de thèse
Période 2020-…
Encadrement/Collaborations KIM Zae Myung (doctorant), Laurent Besacier (NaverLab) et Vassilina Nikoulina (NaverLab)
Financement(s) principal(aux) Contrat de recherche de gré à gré avec *NAVER LABS Europe*
Valorisation :
Ressources langagières ∅
Modèles ∅
Références Kim *et al.* (2021)
Notes ∅

Ce domaine sera intensivement exploité dans le projet Propicto (voir section 2.4.3) qui peut être vu, pour partie, comme une traduction automatique de l'oral vers des séquences de pictogrammes.

2.3.3 Les travaux sur l'oral

J'ai longtemps travaillé uniquement sur l'écrit, mais deux événements m'ont conduit à m'intéresser petit à petit à l'oral. Le premier est local. Nous l'avons vu, l'équipe GETALP accueille des spécialistes de l'écrit mais également des spécialistes de l'oral. Fréquenter ces spécialistes au quotidien a logiquement conduit à l'établissement de ponts entre les recherches.

Le second événements est plus global. À une époque où on voit une convergence des modèles vers les réseaux neuronaux et les méthodes auto-supervisées, il est bien plus simple de passer d'un mode à un autre.

Les premières recherches sont dues au contexte local. Je travaillais depuis plus de 10 ans sur l'analyse sémantique et les algorithmes à colonies de fourmis. Avec de nombreux échanges, Benjamin Lecouteux et moi les avons appliqués à la Reconnaissance Automatique de la Parole* (Lecouteux et Schwab, 2014, 2015). De tels systèmes nécessitaient dans leur version pré-neuronales de décoder des graphes. L'application d'un modèle de langage d'ordre supérieur à un graphe nécessite son extension afin de construire des historiques correspondants à chaque nouvel n-gramme observé. Cette extension peut rapidement engranger des calculs lourds et une consommation de mémoire conséquente. De plus, cette approche permet d'envisager des décodages massivement parallélisables pour un même graphe ainsi

qu'un contrôle strict du temps de calcul et de la mémoire.

En 2018, au moment où nous commencions à travailler sur le projet Propicto dont l'objectif principal est de convertir de l'oral vers des séquences de pictogrammes, Benjamin Lecouteux et moi avons rencontré Nathalie Henrich Bernardoni (CNRS, GipsaLab). Nathalie est spécialiste de la voix et s'intéresse en particulier au *human-beatbox**. Le *human-beatbox* est un art vocal utilisant les organes de la parole pour produire des sons percussifs et imiter les instruments de musique. Il peut également être utilisé dans le cadre de rééducation orthophonique pour des exercices de diction.

Le projet cherche à étudier à quel point il est possible de passer de la voix d'un beatboxer à une séquence de pictogrammes équivalents. La transcription utilisée s'appuie sur un système d'écriture spécifique aux beatboxers, appelé Vocal Grammatics (VG). Ce système d'écriture s'appuie sur les concepts de la phonétique articulatoire. Nous avons proposé ensemble un système de reconnaissance des sons de beatbox s'inspirant de la reconnaissance automatique de la parole (Evain *et al.*, 2019, 2020, 2021a). Nous nous appuyons sur la boîte à outils Kaldi* très utilisée dans le cadre de la reconnaissance automatique de la parole (RAP). Notre corpus est composé de sons isolés produits par deux beatboxers et se compose de 80 sons différents. Nous nous sommes concentrés sur le décodage avec des modèles acoustiques monophones, à base de HMM-GMM. Développé dans le cadre d'un projet art-science qui devait faire intervenir des artistes *beatboxers* et un mur visuel, cet événement a été repoussé pour cause de crise sanitaire.

Mes travaux sur l'oral se complètent aujourd'hui des travaux autour de la thèse de Hang Le que nous avons déjà évoqués (voir section 2.3.2), dans le projet Propicto dont nous parlerons en fin de chapitre, ainsi que dans le projet LeBenchmark.

Sur le modèle du projet FLUE/FlauBERT axés sur l'écrit, le projet LeBenchmark vise à rassembler un ensemble de corpus pré-existants et à faciliter leur utilisation (voir section 3.3.4). Ces corpus sont destinés à évaluer plusieurs tâches (Reconnaissance automatique de la parole – parole vers texte ; compréhension automatique de la parole ; reconnaissance automatique d'émotions : Traduction automatique multilingue – parole vers texte). Un ensemble de modèles de langue auto-supervisés est également proposé (Evain *et al.*, 2021b). Chacun a été appris sur tout ou partie des quelques 3000 heures de corpus oraux du français que nous avons pu rassembler (voir section 3.2.1.5). Mon rôle dans ce projet très collaboratif consiste à participer aux réunions et aux discussions, à l'encadrement de Hang Le ainsi qu'à la gestion et au contenu du site Web.

LeBenchmark

Rôle Participant, directeur de thèse de Hang Le, gestion du site Web
Période 2020-...
Encadrement/Collaborations Solène Evain (Doctorante), Ha Nguyen (Doctorant), Hang Le (Master), Marcely Zanon Boito (Doctorante), Salima Mdhaffar (Doctorante), Sina Alisamir (Doctorant), Ziyi Tong (Master), Natalia Tomashenko (LIA), Marco Dinarelli (CNRS, LIG, GETALP), Titouan Parcollet (LIA), Alexandre Allauzen (PSL Research University), Yannick Estève (LIA), Benjamin Lecouteux (UGA, LIG, GETALP), François Portet (UGA, LIG, GETALP), Solange Rossato (UGA, LIG, GETALP), Fabien Ringeval (UGA, LIG, GETALP), Laurent Besacier (NaverLab Europe)
Financement(s) principal(aux) Heures sur la grille de calcul Jean Zay du CNRS
Valorisation :
Ressources langagières Corpus normalisés pour le français oral (2937 heures), Corpus d'évaluation du français oral [a]
Modèles LeBenchmark [b]
Références Evain *et al.* (2021b)
Notes http://LeBenchmark.com

a. https://github.com/LeBenchmark
b. https://huggingface.co/LeBenchmark

Ces travaux sur l'oral ainsi que ceux portant sur la clarification du sens ou ceux sur la traduction automatique ont une application directe dans la Communication Alternative et Augmentée.

2.4 Projets liés à la CAA

Dans le cadre du Groupe InterAACtion [23], le groupe de recherche et de formation sur la Communication Alternatives et Augmentée inter-disciplinaire de l'université de Grenoble Alpes constitué avec Marion Dohen (Grenoble INP-UGA, GipsaLab), Benjamin Lecouteux (UGA, LIG, GETALP), Amélie Rochet Capellan (CNRS, GipsaLab), plusieurs projets sont actuellement menés.

23. http://interaaction.com

2.4.1 ParticipAACtion

Le premier est le projet ParticipAACtion [24] qui vise à répertorier les modalités et les compétences communicatives des personnes en situation de handicap complexe, associant des déficits perceptifs, moteurs et cognitifs. Ces déficits peuvent avoir des origines multiples et affectent la production de la parole ainsi que sa perception. Parfois, les personnes ne peuvent pas du tout parler ou bien en deçà de ce qu'elles voudraient et auraient la capacité à dire. Les recherches bibliographiques du groupe InterAACtion ont permis de constater un manque d'observations et d'analyses de la communication des personnes concernées en contexte : très peu de recherches en sciences du langage ont été dédiées à ces personnes. Les vidéos et les témoignages des aidants montrent des compétences qui vont au delà de ce que l'on peut trouver dans les articles de recherche qui se limitent souvent à des évaluations sur la base de questionnaires ou des observations dans des contextes de prise en soin.

ParticipAACtion propose une approche participative et interdisciplinaire visant à créer et analyser un corpus audio-visuel de la communication des personnes filmées dans leur quotidien par leurs proches. En effet, les limites adaptatives et le manque d'observations "positives" des personnes imposent de revoir les méthodes d'investigation et d'opérer un retour à l'observation en contexte. La constitution d'une base de données comportementale anonyme enrichira la connaissance des compétences des personnes. Les analyses permettront d'identifier les outils et les contextes qui favorisent la communication, de valoriser les compétences des personnes concernées et des aidants. Le projet a aussi pour objectif de contribuer au développement de l'usage de la CAA en France avec les personnes en situation de handicap complexe.

ParticipAACtion est d'abord dédié à la communication des personnes avec des handicaps dits sévères tel que celui induit par le syndrome d'Angelman. Cette démarche a été soutenue par la Fondation des Maladies Rares et le projet implique différents partenaires. Cependant, les personnes avec des handicaps plus modérés, comme c'est le cas de la trisomie 21 par exemple, ont aussi des difficultés spécifiques avec la parole et souffrent du manque d'usage de CAA. Le choix d'un outil ou d'une méthode de CAA n'étant pas lié à l'étiologie du trouble mais aux compétences de la personne et de son environnement, il nous a paru incontournable de ne pas nous limiter à un syndrome spécifique afin de favoriser le transfert des connaissance et l'universalisation des usages.

24. http://ParticipAACtion.com/ – les informations sur le projet proviennent d'ailleurs de ce site.

ParticipAACtion

Rôle Organisation, membre du comité d'experts
Période 2020-…
Encadrement/Collaborations Amélie Rochet-Capellan (resp. projet, CNRS, Gipsa-Lab), Marion Dohen (resp. projet, Grenoble INP, UGA, Gipsa-Lab), dizaines de bénévoles, AFSA [a]
Financement(s) principal(aux) Fondation des Maladies Rares
Valorisation :
Ressources langagières ∅
Modèles ∅
Références ∅
Notes http://ParticipAACtion.com/

a. Association Française du Syndrome d'Angelman – https://www.angelman-afsa.org

2.4.2 Logiciels et outils pour la CAA

Comme présenté dans la section 2.2, nos recherches reposent sur la mise à disposition d'outils conçus à partir de l'état de l'art et conçus pour le faire progresser. Un des objectifs du projet ParticipAACtion sera de nous apporter des clés, des idées, des concepts que nous pourrons introduire dans les logiciels et matériels en source ouverte que nous présentons ensuite.

GazePlay [25] (Schwab *et al.*, 2018, 2020a) : une plateforme de jeux utilisables par oculométrie* en développement depuis 2016 et en constante évolution. Elle propose à ce jour plus de 60 jeux créatifs, ludiques ou sérieux permettant d'aider à l'acquisition de compétences (action-réaction, sélection, littératie, mémorisation, etc.). Bien qu'éloigné de mes recherches en TALP, GazePlay s'est vite trouvé au centre de mes réflexions sur la Communication Alternative et Augmentée. En effet, pour des enfants qui partent par nature de zéro, qui, de plus, ont de telles difficultés, c'est déjà un grand défi d'acquérir et de soutenir les connaissances de base que cette communication requiert. Le regard est ainsi souvent considéré comme l'un des moyens les plus naturels et les plus faciles à mettre en place pour permettre aux personnes en situation de polyhandicap d'interagir avec leur environnement.

Les jeux sont souvent considérés comme un bon moyen d'apprendre. Les jeux destinés à être utilisés avec des oculomètres (eye-trackers), c'est-à-dire des dispositifs électroniques capables d'estimer la position du regard, peuvent être un

25. http://gazeplay.net

bon moyen d'améliorer les compétences requises comme la fixation et la poursuite oculaire, ainsi que des conventions comme les récompenses ou les interactions de fixation (dwell interaction) souvent utilisées dans les outils de CAA. En parallèle, grâce à des étudiants en neuropsychologie de l'enfant et en collaboration avec le CHU de Grenoble et des institutions spécialisés locales, nous développons une composante évaluation (GazePlay-Eval).

InterAACtionScene [26], un logiciel interactif et configurable de scènes visuelles pour apprendre le vocabulaire de base aux enfants, tout en faisant un premier pas vers la Communication Alternative et Augmentée.

Augcom [27], un outil de grille de Communication Alternative et Augmenté, basé sur des études comparatives des différents logiciels existants, sur les besoins exprimés par les aidants et thérapeuthes et qui continuera à s'améliorer en proposant notamment un lexique et une organisation optimale basée sur la recherche. Le fondement principal de ces outils repose sur la génération vocale réalisée à partir de pictogrammes : elle permet de composer un message à partir d'un ensemble de pictogrammes afin de les associer entre eux pour qu'une synthèse vocale énonce le message au destinataire. Ce logiciel nous permettra de récolter les messages composés par les aidants ou thérapeuthes dans le cadre du projet Propicto (voir ci-dessous).

Enfin, le projet **InterAACtionBox** a pour objectif de créer dispositif intégré en source ouverte permettant la Communication Alternative et Augmentée pour tous. Réalisé grâce au soutien de l'AFSR [28], il a remporté plusieurs prix qui nous ont largement aidés dans le développement grâce au financement d'un ingénieur d'étude pendant 18 mois. L'idée est de faciliter au maximum la vie des utilisateurs en proposant un système alliant matériel et logiciels qui soit à la fois robuste et simple d'utilisation. Il intègre ainsi :

— Ordinateur ou tablette, l'outil permet des interactions oculaires et tactiles

— Système d'exploitation InterAACtionBoxOs qui gère l'InterAACtionBox

— Logiciels de CAA (GazePlay, InterAACtionScene, Augcom) ainsi que des lecteurs inclusifs (Gaze MediaPlayer, un lecteur audio video qui permet la connexion à diverses plateformes comme YouTube [29], Spotify [30]...)

— Outils d'analyse des résultats de l'utilisateur

26. http://interaactionscene.net
27. http://augcom.net
28. Association Française du Syndrome de Rett (http://afsr.fr)
29. https://www.youtube.com
30. https://www.spotify.com

— Outils d'analyse de l'utilisation (pour l'amélioration de l'outil et la compréhension des usages)

Le projet bénéficie de l'apport indéniable des idées et de la motivation de nos étudiants, particulièrement sensibles à la cause. L'équipe de développement a accueilli en 2021, 1 ingénieur d'étude (18 mois), 1 stage découverte de la recherche pour une étudiante en prépa INP - 2 mois - implantation de grilles pré-existantes dans AugCom), 1 licence informatique (4 mois - Gazeplay), 4 Master 1 informatique (stages de recherche, 4 à 6 mois [31]), 1 ingénieur en cognitique (2 mois - conception d'outils d'analyse pour les jeux) 2 étudiants en master informatique pour 6 mois. Ces deux derniers ont été recrutés en CDD suite à leur stage.

InterAACtionBox

Rôle Responsable du projet et des sous-projets
Période 2016-…
Encadrement/Collaborations Groupe InterAACtion, Sébastien Riou (ingénieur d'étude), nombreux étudiant(e)s
Financement(s) principal(aux) LIG, AFSR
Valorisation :
Ressources langagières ∅
Modèles https://github.com/InterAACtionGroup
Références Schwab (2018); Schwab *et al.* (2018, 2020a); Chasseur *et al.* (2020)
Notes http://interAACtionbox.com, http://gazeplay.net, http://augcom.net
Prix (fondation Free, fondation Klesia, fondation Afnic)

Ce projet est au cœur de nos recherches et formations. C'est avec cette InterAACtionBox ou certains de ses composants, les logiciels présentés ci-dessus, que les institutions avec lesquelles nous collaborons travaillent. Il en sera de même avec les parents de l'AFSR d'ici la fin 2021. Ce dispositif sera également au cœur du projet Propicto.

31. Sur l'adaptation automatique du niveau des jeux de GazePlay, sur des interaction alternatives avec le regard ainsi que des sujets clairement liés au TALP comme la création automatique de grilles de pictogrammes à partir de corpus transcrits.

2.4.3 Propicto

Le projet Propicto (PRojection du langage Oral vers des unités PICTOgraphiques) concerne la transcription automatique de la parole française sous forme pictographique. En effet, composer un message en CAA est réputé comme au moins 10 fois plus lent que la parole. Pourtant, associer des pictogrammes au discours correspondant est essentiel au bain langagier dont nous avons déjà souligné la nécessité. Cette association se complique lorsque l'on souhaite projeter une représentation textuelle complexe sous la forme d'une suite de pictogrammes. Il s'agit de permettre une certaine polyglossie comme évoquée en 1.3. Comme il s'agit de la transposition d'un énoncé linguistique en une représentation visuelle complexe, il ne faut pas seulement associer un pictogramme à un mot, mais plus largement trouver comment des représentations linguistiques peuvent être représentées visuellement. Commencé comme un projet financé par le LIG avec Benjamin Lecouteux (stage de master 2 Recherche de Céline Vaschalde), il a ensuite été poursuivi par un financement CNRS puis de l'AFSR menés par Benjamin Lecouteux. Ce problème a des liens assez étroits avec la traduction automatique et la clarification du sens.

Les premières études théoriques nous ont également permis d'identifier plusieurs problèmes dont la nécessité de :

— simplifier les messages, par exemple en passant les formes passives en active ;

— pallier l'absence de certains pictogrammes ;

— représenter certains phénomènes [32].

Les premières études pratiques nous ont permis d'étudier les liens avec la clarification du sens, la désambiguïsation lexicale sur laquelle nous focalisions nos travaux à l'époque. Ainsi les premiers prototypes de Céline Vaschalde exploitaient d'un côté les résultats de la thèse de Marwa Hadj Salah pour le français avec les modèle de désambiguïsation lexicale de Loïc Vial.

La rencontre avec Pierrette Bouillon (Université de Genève) et Hervé Spechbach (Hôpitaux Universitaires de Genève) qui avaient un projet similaire mais lié à l'accueil des personnes allophones* dans les services d'urgence nous a conduits à commencer une collaboration. Celle-ci a commencé en 2019, dans un premier temps autour du co-encadrement du master de Lucía Ormaechea Grijalba et maintenant à travers le projet financé par l'Agence Nationale de la Recherche et le Fond National Suisse franco-suisse, Propicto. Commencé en 2021, et prévu pour durer jusqu'en 2024, le travail a concerné pour le moment les ressources (voir section 3.2.1.2) . Je vais codiriger la thèse de Cécile Macaire qui commence d'ici la fin

32. Comment représenter le pluriel par exemple. En doublant le pictogramme ? En modifiant le pictogramme par un symbole particulier.

2021 avec Benjamin Lecouteux (GETALP, LIG, UGA) et Emmanuelle Esperança-Rodier (GETALP, LIG, UGA).

Propicto

Rôle Implication dans la plupart des lots, coordination avec les autres projets CAA, coordination science participative
Période 2018-ldots
Encadrement/Collaborations Pierrette Bouillon (Université de Genève), Hervé Spechbach (Hôpitaux Universitaires de Genève), Benjamin Lecouteux (UGA, LIG, GETALP)
Financement(s) principal(aux) LIG, AFSR, Projet ANR-FNS Franco-Suisse (2021-2024)
Valorisation :
Ressources langagières UPSAC
Modèles ∅
Références Vaschalde *et al.* (2018a,b); Schwab *et al.* (2019)
Notes `https://propicto.unige.ch`

2.5 Conclusions du chapitre

Dans ce chapitre, j'ai présenté la façon dont mes recherches en Traitement Automatique des Langues et de la Parole s'articulent et comment mes recherches en Communication Alternative et Augmentée s'y sont intégrées de manière assez naturelle. La méthodologie habituelle de travail de l'équipe GETALP s'y prête particulièrement bien. Nous fabriquons des outils et systèmes qui sont ensuite testées en situation écologique chez des thérapeutes et des particuliers. De leurs retours, par discussions, enquêtes ou par observation automatisée, nous tirons des enseignements qui sont ensuite utilisés pour améliorer les systèmes et outils.

Maintenant que nous avons présenté ces recherches et leur cadre, voyons plus précisément comment nous concevons les outils de TALP et de CAA. Ce sera l'objet des 3 prochains chapitres. Je commence par ce qui est au centre de cette conception, les ressources langagières.

Chapitre 3

Contributions à la constitution de données langagières

Sommaire

En Traitement Automatique des Langues et de la Parole (TALP), Witt *et al.* (2009) définissent deux types de ressources langagières*, (1) les ressources statiques*, c'est-à-dire les données proprement dites, et (2) les ressources dynamiques*, à savoir les outils qui permettent de traiter les ressources statiques.

Dans ce chapitre, je présente plusieurs de mes contributions à la création de ressources langagières statiques. Dans un premier temps, nous verrons les hypothèses posées pour la constitution des ressources langagières statiques [1] puis quelques exemples des contributions auxquelles j'ai participé. Mes recherches actuelles et futures dans ce cadre sont esquissées tout au long de ce chapitre.

3.1 Hypothèses de constitution des ressources langagières

3.1.1 Langues

C'est une lapalissade que de dire que la langue d'un corpus* est souvent liée au contexte de l'application ou de la tâche visée lors de sa constitution. Notre équipe se situe en France, dans un milieu social, culturel et même économique qui utilise en très large majorité la langue française. S'intéresser aux données pour le français semble donc assez naturel. Parallèlement, pour être publié dans les principales conférences anglophones, il faut souvent, encore aujourd'hui, appliquer son travail à l'anglais (Joshi *et al.*, 2020), langue qui dispose de largement plus de ressources que les autres pour la plupart des tâches et applications. À une époque où notre domaine était moins en vogue, certains collègues ne nous suggéraient-ils pas d'abandonner nos travaux sur les autres langues ? Ce n'étaient évidemment pas des gens qui s'intéressaient de près ou de loin à « *la langue* », c'est-à-dire aux langues dans leur variété.

Depuis ses origines, il y a plus de soixante ans (Léon, 2002), le GETALP s'est intéressé à d'autres langues et a une longue tradition de collaboration tournée d'abord vers l'URSS, l'Europe et les États-Unis, puis vers le Canada, l'Asie, l'Afrique et plus récemment l'Australie. Moi-même, avant d'arriver à Grenoble, j'ai effectué plusieurs séjours scientifiques en Malaisie [2], pays culturellement multilingue [3].

1. Dans la suite de ce chapitre, nous ne parlerons plus que des ressources langagières statiques. Le qualificatif statique ne sera plus précisé.

2. Un séjour d'un mois en décembre 2001-janvier 2002 puis un poste de chargé de recherche entre 2006 et 2007 ; j'ai effectué d'autres visites plus courtes depuis.

3. Quatre langues y sont prépondérantes, le malais, l'anglais, le tamoul ou les langues chi-

Nous ne reviendrons pas ici sur les arguments classiques et fondamentaux autour de la préservation de la culture et des langues pour s'attarder sur des arguments plus informatiques. Je m'intéresse en particulier depuis mon retour en France en 2007 à l'usage des outils informatiques pour des personnes non spécialistes [4]. Étudiants en commerce, en linguistique, en psychologie, parents d'enfants en situation de handicap, thérapeutes sont des publics que je fréquente au quotidien par mes recherches et mes enseignements, formations et conseils, dans les universités, les institutions spécialisées, les associations, les fondations, sur Internet et autres réseaux sociaux. J'essaye de leur donner de bonnes pratiques et des explications sur les concepts fondamentaux de l'informatique, devenus aujourd'hui souvent des usages de l'informatique du quotidien* comme, par exemple, ceux qui concernent le partage des documents en particulier et le partage des données en général. Ainsi, je cherche à faire en sorte que les outils soient le plus facile possible à utiliser pour eux, y compris, et surtout, en les adaptant au mieux à leurs besoins.

En effet, si les personnes doivent un peu s'adapter aux outils parce que, comme avec tout artefact, il y a des caractéristiques qu'elles doivent comprendre, qu'elles doivent appréhender, il ne faut pas que l'écart soit, pour elles, un fossé trop difficile à franchir. Les utilisateurs apprennent de l'outil et grâce à l'outil, mais l'outil doit également s'adapter, évoluer avec les utilisateurs, c'est la notion de coévolution* déjà évoquée. C'est donc aussi aux outils de combler cet écart en allant vers les utilisateurs, en particulier en utilisant la ou les langues d'usage habituel dans leur quotidien ou leur travail. En effet, considérer que les gens vont apprendre les langues dominantes par le contexte social n'est certainement pas un raisonnement tenable. Parfois, ces personnes n'en sont tout simplement même pas capables, par exemple parce qu'elles viennent d'arriver dans le pays ou par handicap. Ajouter des langues, c'est inclure plus de personnes, c'est s'ouvrir à plus d'applications, c'est s'ouvrir à plus d'utilisateurs, c'est s'ouvrir à plus de contextes d'usage. Ces deux aspects, humain et machine, permettent ainsi d'avoir plus de retours et au final, une ressource globalement de bien meilleure qualité.

D'autres raisons plus fondamentales me semblent pousser à considérer que les ressources langagières devraient être aussi multilingues* que possible. Une première raison est de nature algorithmique. Pour être certain qu'un algorithme généralise de manière appropriée, qu'il est indépendant de la langue, il faut le tester sur un maximum de langues, issues de familles différentes, des langues typologiquement éloignées comme l'argumente Bender (2011).

---

noises avec en particulier sur l'île de Penang où je vivais, le hokkien, la langue principale de la communauté chinoise locale (Tan, 2001; Leclerc, 2016; Lütkebohle, 2021) ; voir aussi `https://en.wikipedia.org/wiki/Penang_Hokkien`, consulté le 30 juin 2021.

4. La phrase « *Les ordinateurs ne m'aiment pas* » est peut-être celle que j'ai le plus entendue dans ma vie.

Une seconde raison est de nature plus linguistique. On le sait, les langues sont fortement liées entre elles, elles connaissent des échanges réguliers. Comme les espèces animales et végétales, elles ont parfois des ancêtres communs [5]. Mais, contrairement aux espèces qui ne peuvent définitivement plus du tout fusionner lorsqu'elles sont trop éloignées génétiquement, les langues ne connaissent pas ce type de limites et les échanges restent toujours possibles. Les échanges principaux sont lexicaux et reflètent relativement bien l'Histoire. Les échanges entre le français et l'anglais sont, par exemple, extrêmement bien documentés [6]. Aujourd'hui, les spécialistes recensent quelques 2500 mots anglais en français dans le Petit Robert, ce qui fait de la langue anglaise, la seconde contributrice au français derrière le grec ancien et devant l'italien (Jean, 2020). Mais ces quelques 4,5% du vocabulaire français ne sont rien en comparaison des quelques 2/3 du vocabulaire anglais provenant du français [7].

On peut multiplier les exemples d'échanges entre les langues. En suivant les routes commerciales tout comme les religions et les idées, les langues s'influencent en suivant les Routes de la Soie (Frankopan, 2019) ou en suivant les conquêtes européennes et les expansions coloniales. Ainsi, nos chiffres viennent d'Inde via les peuples arabes ; le français ‹‹sarbacane›› vient du malais par le persan et l'espagnol, le mot ‹‹cacao›› est arrivé au français via la langue nahuatl des Aztèques puis l'espagnol, ‹‹thé›› via le malais *te* et l'anglais *tea* , ‹‹shampooing›› par l'anglais *shampoo* qui est lui-même venu de l'hindi *chāmpo* ; ‹‹tiare›› et ‹‹monoï›› ont directement été pris au tahitien (Walter, 1997). L'influence importante de l'anglais sur le malais et du néerlandais sur l'indonésien est également bien documentée (Jones, 2007).

Régulièrement, le TALP cherche à exploiter les similitudes entre les langues.

5. On sait que certains langues dérivent initialement d'une langue commune, comme c'est le cas du français ou de l'italien avec le latin. En revanche, la question d'une langue mère unique apparue il y a des centaines de milliers d'années ou de l'apparition de plusieurs foyers indépendants reste ouverte Sagart (2008).

6. Accompagnant Guillaume le conquérant et les quelques 20 000 nobles venus s'installer en Angleterre après la bataille d'Hastings en 1066, des mots ont évolué avant même parfois, pour certains, de faire le chemin inverse des siècles plus tard. Pendant toute la période du Moyen Âge qui suit et jusqu'à la Renaissance, les Anglais ont des possessions sur le continent. À son apogée au XIIème siècle, les souverains d'Angleterre contrôlent tout le tiers ouest de la France actuelle ; Calais ne tombe que le 8 janvier 1558 ; la Dunkerque anglaise n'est qu'une parenthèse de quelques quatre années jusqu'à son rachat par la couronne en 1662.

Entre-temps, via le normand, le picard ou directement, le dialecte francilien devenu le français a transmis un bon nombre de mots (Walter, 2012). Le flux s'est inversé durant les siècles qui ont suivi accompagnant la liberté et le modernisme que l'Angleterre et les États-Unis représentaient pour la France des Lumières et ses héritiers (Lodge, 1997).

7. Selon Henriette Walter. Il y a, en particulier, 1700 vrais amis, des termes absolument identiques en signifiant et en signifié dans les deux langues – https://www.lexpress.fr/culture/livre/honni-soit-qui-mal-y-pense_804257.html

Ainsi, la communauté a créé ou exploité des ponts entre les ressources en se basant sur des expertises humaines (par exemple, Dbnary (Sérasset, 2015a) qui se base sur le Wiktionnaire pour extraire les données pour plus de 20 langues ; BabelNet (Navigli et Ponzetto, 2012a) qui aligne plusieurs ressources dont *Princeton WordNet* ou Wikipedia pour plus de 500 langues) ; des corpus alignés qu'ils soient uniquement écrits (Tiedemann, 2012a) ou, depuis quelques années, écrits et oraux (Zanon Boito *et al.*, 2020; Di Gangi *et al.*, 2019; Wang *et al.*, 2021).

La communauté du TALP a montré, qu'avoir accès à de telles ressources, comportant autant de langues interconnectées permettait de généraliser un apprentissage sur des couples de langues non directement rencontrés (*zero-shot learning*) (Aharoni *et al.*, 2019; Philip *et al.*, 2020; Kumar *et al.*, 2019), ainsi qu'une amélioration des systèmes, y compris pour des langues déjà bien dotées (Navigli et Ponzetto, 2012a; Tchechmedjiev, 2012; Le *et al.*, 2020b).

Lorsque l'on cherche à constituer des corpus, il n'est bien entendu pas demandé de le faire pour toutes les langues. C'est matériellement impossible même pour de grosses équipes. Seule une collaboration la plus large possible et sur le long terme peut permettre la constitution efficace et pérenne des ressources. Un grand nombre d'acteurs, y compris les GAFAM*, l'a aujourd'hui compris si on croit le nombre de ressources publiées chaque année.

Pour un acteur individuel, même géant, il s'agit d'être conscient de ces limites en mettant en place des moyens pour constituer de manière efficace et pérenne des ressources :

— *en créant des ressources manuellement* pour ces langues grâce à des locuteurs natifs comme nous l'avons fait pour le bengali avec Mohammad Nasiruddin (Nasiruddin *et al.*, 2014) ou avec Marwa Hadj Salah pour l'arabe (Hadj Salah, 2018) ou des locuteurs avertis [8] dans le cadre de la traduction de textes en pictogrammes.

— *en fabriquant des outils permettant de récolter des données* dans le respect des règles éthiques comme nous le ferons, par exemple, dans le cadre du projet Propicto [9] avec la voix, le texte et les pictogrammes.

— *en concevant des méthodologies assez génériques pour être mises en place pour un maximum de langues*, par exemple en exploitant intensivement Wikipedia ; la création d'outils peut également être assez générique comme ce sera le cas dans le cadre du projet Propicto avec les librairies de pictogrammes.

8. On ne peut pas réellement parler de locuteurs natifs en ce qui concerne l'usage des pictogrammes.

9. http://Propicto.unige.ch, page consultée le 30 juin 2021.

— *en établissant des ponts entre les ressources* comme ceux que nous créons entre des bibliothèques de pictogrammes et WordNet dans le cadre de la CAA (Schwab *et al.*, 2019, 2020b) ou dans les travaux menés avec Andon Tchechmedjiev [10] (Tchechmedjiev, 2016). Ces travaux s'inscrivent dans l'initiative des Données Linguistiques Liées Ouvertes (*Linguistic Linked Open Data* - LLOD) en appliquant les principes du Web des données (*linked data*). C'est ce que l'on appelle l'interopérabilité*, c'est un des points principaux que nous développerons dans la suite.

— *en unifiant les formats* comme nous l'avons fait en proposant l'OAPGF (*Open AAC Pictogram Grid Format*), un format ouvert pour décrire les grilles de pictogrammes ou en unifiant les corpus de données annotées en sens avec UFSAC (Unification of Sense Annotated Corpora and Tools) ou en pictogrammes avec UFPAC (*Unification of Pictograms Annotated Corpora*), autant de travaux que nous présenterons dans la suite de ce chapitre.

— *en intégrant des bibliothèques et entrepôts de données* comme ceux d'*Hugging Face* [11] (Wolf *et al.*, 2020) ou ceux dédiés à la recherche d'information de l'*Allen Institute for AI* [12] (MacAvaney *et al.*, 2021).

Mon rapport aux données a continuellement évolué depuis 20 ans époque où, comme d'autres, nous utilisions tout ce que nous trouvions sur Internet sans vraiment nous poser de questions. Petit à petit, j'ai été sensibilisé à un certain nombre de bonnes pratiques pour les données jusqu'à l'arrivée des principes du FAIR dans lesquels je me retrouve aujourd'hui. Principe que je présente ci-dessous.

3.1.2 FAIR data

En 2016, dans un contexte où les données sont de plus en plus présentes et produites dans notre environnement, a été publié (Wilkinson *et al.*, 2016) l'article fondateur sur le FAIR. Reprenant un ensemble de principes préexistants, il les regroupe grâce à un acronyme simple à retenir [13] : **F**aciles à (re)trouver, **A**ccessibles, **I**nteropérables, **R**éutilisables.

Faciles à (re)trouver : les données doivent être faciles à (re)trouver que ce soit pour les humains ou pour les machines grâce à des entrepôts de données. Ce principe est basé sur les métadonnées qui doivent avoir des identifiants uniques et

10. Dans le cadre de sa thèse de doctorat soutenue en octobre 2016 et co-encadrée avec Gilles Sérasset et Jérôme Goulian (Université Grenoble Alpes).
11. https://huggingface.co, page consultée le 30 juin 2021.
12. https://ir-datasets.com, page consultée le 30 juin 2021.
13. https://www.go-fair.org/fair-principles/

persistants au niveau mondial (*Persistent URLs*, *Digital Object Identifier*, *Uniform Resource Identifier*...).

Accessibles : les (méta)données peuvent être retrouvées par leur identifiant grâce à un protocole gratuit, libre, ouvert et standard. Les métadonnées expliquent clairement comment accéder aux données (gestion éventuelle d'authentifications) et quels sont les droits d'utilisation associés. Notons que si le principe n'exige pas la gratuité des données, c'est un mode d'accès que je privilégie autant que possible et que je négocie [14] avec mes partenaires.

Interopérables : l'interopérabilité est la capacité d'éléments matériels ou logiciels à échanger des données et à utiliser les données échangées par le recours à des standards de communication ouverts [15].

Réutilisables : objectif ultime du FAIR, la réutilisabilité permet d'exploiter les données dans d'autres contextes que pour celui où elles avait été originellement conçues. Il s'agit ainsi d'utiliser des formats standard mais également de préciser les conditions légales du partage (licences), la manière dont elles ont été collectées ou si elles ont été générées ou produites manuellement.

Je considère, de plus, que les données devraient être librement disponibles comme je l'explique ci-dessous.

3.1.3 Données librement accessibles

Je vois plusieurs raisons pour lesquelles les données devrait être librement disponibles.

Pour la facilité d'accès. Nous avons vu dans la section 3.1.2, consacrée au FAIR, que le F correspondait à la facilité d'accès aux données du point de vue légal. Les données peuvent être faciles d'accès mais impossible à utiliser pour des questions de licence. Utiliser une licence libre quel que soit le contexte (même industriel), garantit une facilité d'utilisation pour tous les acteurs. C'est de cette facilité d'accès que découlent la plupart des éléments suivants.

Pour la reproductibilité des recherches. Reproduire une recherche dans tous ses aspects (reproductibilité d'une conclusion, reproductibilité d'une constatation, reproductibilité d'une valeur), comme proposé par Cohen *et al.* (2018) n'est pas

14. aujourd'hui souvent facilement et sans contrepartie particulière tant cet élément est rentré dans les mœurs.

15. Cette définition, très inspirée de celle de Duponchelle (2015) qui étudie l'interopérabilité du point de vue juridique, implique que les spécifications des interfaces de ces éléments sont intégralement connues et peuvent ainsi fonctionner avec d'autres sans restriction d'accès ou de mise en œuvre.

toujours facile même si on l'a montée sois-même (Mieskes *et al.*, 2019). Sans les données qui ont permis de la mener, la reproductibilité d'une valeur et la reproductibilité d'une constatation est impossible. Il peut en être de même pour la reproductibilité d'une conclusion en particulier lorsque peu de données autres que celles de la recherche initiale sont disponibles.

Pour l'évaluation des algorithmes. Évaluer un algorithmes dans le plus grands nombre de contextes possibles permet de l'évaluer sur un maximum d'aspects (langues, domaines, locuteurs,...). Disposer facilement des ressources permet de telles évaluations.

Pour la démocratisation des tâches et des langues. Le manque de données décourage la recherche et les applications pour lesquelles il manque des données car le ticket d'entrée est alors prohibitif ou simplement parce qu'il est difficile de penser des applications avant d'avoir un embryon des données nécessaires. Le succès actuel du domaine de l'Intelligence Artificielle est clairement lié à cette démocratisation par l'état d'esprit de partage manifesté par un grand nombre d'entreprises y compris parmi les plus grosses (Le Cun, 2019; Pratap *et al.*, 2020; Wang *et al.*, 2021). L'exemple du récent BigScience semble également assez caractéristique. Rassemblant des centaines de chercheurs [16] du secteur privé et public, il s'agit de créer un grand modèle génératif multilingue sur la grille de calcul du CNRS Jean Zay.

Pour ne pas réinventer la roue. Ne pas diffuser ses données peut obliger les autres à en obtenir d'autres, équivalentes. On peut y voir du point de vue global, un gaspillage de cerveaux qui pourraient être utiles à autre chose et un gaspillage de ressources (production de $C0_2$). Dans un contexte de réchauffement climatique, peut-on encore se permettre ce genre de luxe ?

Pour que d'autres travaillent sur les données et développent ou améliorent des algorithmes qui bénéficieront à la communauté et donc en retour au créateur de la donnée.

Pour le prestige. Mettre à disposition d'une communauté une ressource, c'est également pointer un projecteur vers les chercheurs et les organisations publiques ou privées qui ont permis ces travaux. On cite, on parle des travaux, de la personne et de l'organisation impliqués. Sortir Word2Vec, a mis l'accent en 2013 sur les travaux de l'équipe de Mikolov chez Google (Mikolov *et al.*, 2013). En ce qui me concerne, publier les modèles de FlauBERT m'a fait rencontrer plusieurs journalistes (Journal du CNRS, Science et Vie...) et participer à plusieurs séminaires invités. Nos camarades de l'équipe *CamemBERT* ont eu les honneurs du Monde,

16. Plus de 400 à la fin mai 2021.

le grand quotidien du soir français [17]. On pourrait multiplier les exemples…

Pour permettre de communiquer sur du long terme. Aujourd'hui, les ressources créées avec la plupart des logiciels disponibles pour la CAA ne sont pas facilement lisibles avec un autre logiciel que celui qui a permis de les concevoir et sont même souvent cryptées. Les aidants ou les thérapeutes passent un temps considérable à concevoir des grilles pour les enfants afin de les aider à s'exprimer en fonction du monde qui les entoure et de leurs spécificités individuelles comme, par exemple, les lieux qu'ils fréquentent plus particulièrement, les livres ou dessins animés qu'ils aiment, la présence d'animaux ou d'objets, les textures appréciées, *etc.* Si un jour, le logiciel n'est plus disponible [18], la personne perdra tout le travail accompli et perdra son seul moyen de communiquer. Avec des formats ouverts, d'autres logiciels peuvent permettre d'ouvrir ces documents et d'éviter de perdre des données si précieuses.

Si c'est mon opinion dans un contexte universitaire, je pense aussi que ce raisonnement peut-être prolongé au secteur industriel. On comprend aisément qu'une entreprise ne souhaite pas libérer l'ensemble de ses données car les rendre librement accessibles pourrait offrir un avantage concurrentiel évident dans certain cas de figure. Pour une entreprise, la stratégie consiste ainsi à libérer une partie suffisante de ses données pour obtenir les avantages énoncés ci-dessus, sans pour autant offrir des avantages décisifs à des concurrents existants ou potentiels. Il s'agira alors, par exemple, de publier des corpus d'évaluation ou encore une partie plus ou moins importante des corpus d'apprentissage. Une autre piste pourrait être de

17. https://www.lemonde.fr/sciences/article/2019/11/18/intelligence-artificielle-dorenavant-les-machines-maitrisent-la-grammaire-francaise_6019639_1650684.html, article consulté le 30 juin 2021

18. Parce que la personne n'a pas/plus les moyen de se le procurer financièrement ou légalement, parce que l'éditeur a fermé, a été racheté, a choisi de ne plus éditer le logiciel, a modifié sa licence ou le système d'exploitation de l'utilisateur a changé et que le logiciel n'est plus compatible. L'histoire de l'informatique a longtemps été truffée d'exemples similaires et la situation s'est largement améliorée depuis quelques années, par exemple avec les traitements de texte. Les formats ouverts comme ceux utilisés dorénavant par *Microsoft Word* permettent une lisibilité quasi-complète du texte et relativement bonne du formatage sur d'autres logiciels comme ceux des autres grands acteurs ou des logiciels libres. Dans le monde de la CAA des dernières années, les acteurs se rachètent, se scindent, les rapprochements avortent ou sont interdits par les autorités de la concurrence, les logiciels changent régulièrement de bannière. Par exemple, le logiciel *Communicator* a longtemps été connu comme *Viking Communicator* puisqu'il appartenait à la société *Viking*, rachetée par *Tobii* en 2007 fait aujourd'hui partie de la division *Dynavox* (elle-même ancienne entreprise américaine rachetée en 2014) de *Tobii AB*. En 2018, l'arrêt de l'ajout de nouvelles fonctionnalités et une fin du support à l'horizon de 5 ans avaient été annoncés mais les changements industriels (interdiction du rachat de *Smartbox* par l'autorité britannique de la concurrence) semblent lui a avoir offert une nouvelle vie, mais jusqu'à quand ? *Tobii AB* vient d'annoncer sa séparation en deux entités d'ici la fin 2021, il est trop tôt pour estimer quelles en seront les conséquences pour les consommateurs.

publier des données synthétiques créées à partir des données originelles (Claveau *et al.*, 2021).

Certains ont pourtant peur de partager une partie de leurs ressources. Si cette réflexion est certainement bien moins pertinente dans le monde de la tech aujourd'hui, il subsiste une forte résistance dans de nombreuses entreprises et domaines. Ce phénomène constitue, par exemple, un frein très clair à la recherche dans les domaines comme celui de la Communication Alternative et Augmentée (CAA) où la plupart de données et logiciels sont propriétaires et fermées. Le domaine de la Communication Alternative et Augmentée est un domaine où les ressources sont rares, souvent propriétaires, chères et respectent rarement les normes permettant un partage facile.

Pour certains, c'est simplement que leur système est le meilleur, qu'il ne s'agit pas de discuter ou de critiquer. Je n'insisterai pas sur cet argument qui relève, par la manière dont il est exprimé, plus de la conviction ou du marketing maladroit que de la science.

Une autre crainte exprimée, est que la ressource pourrait être mal comprise et donc mal utilisée par les personnes. Il me semble qu'il faut pourtant bien avoir un peu confiance dans ses logiciels et surtout dans ses compétences. La vraie valeur du travail en CAA n'est pas dans les ressources produites, elle se trouve dans les connaissances apportées sur la manière d'utiliser ces ressources. La valeur n'est pas dans telle ou telle organisation du vocabulaire, la valeur est dans les explications, dans les principes qui ont conduit à une telle organisation, à tel ou tel mode d'interactions, dans l'accompagnement des familles et des institutions au quotidien. Aujourd'hui et, semble-t-il depuis de nombreuses années pour certains, on voit tel ou tel parent, telle ou telle institution qui a copié un des grands systèmes de communication. Ces personnes se cachent de peur de se faire attaquer, par exemple, pour atteinte au droit d'auteur [19]. Comme tout système de communication, en particulier quand il est en vase clos, il s'éloigne petit à petit du canon, du système originel. Sans chercher à répondre à la question de savoir si c'est un mal ou un bien, cela montre que considérer que la fermeture d'un système permet de garder la main sur son contrôle est un leurre. En revanche, en ouvrant, il serait permis d'agrandir la communauté, de réfléchir collaborativement aux améliorations et de les mener bien plus facilement.

19. Je n'ai, à ce jour, aucune idée de la validité juridique de cette crainte mais elle est brandie très régulièrement sur les réseaux sociaux.

3.1.4 Description fine des données

Les données doivent être décrites le plus finement possible. En domaine, en genre, en type d'accent, qualité de l'enregistrement bien entendu mais également de la manière dont elles ont été originellement constituées. C'est souvent la condition *sine qua non* pour pouvoir analyser des données, et aussi pour comprendre certaines limites des systèmes appris sur les données, pour pouvoir identifier ou expliquer des biais (Bender et Friedman, 2018).

Ainsi, de nombreux systèmes de désambiguïsation lexicale sont basés sur le SemCor (Miller *et al.*, 1993), lui-même établi sur un sous-ensemble du *Brown corpus* (Francis et Kučera, 1964; Kucera et Francis, 1967; Francis et Kucera, 1979) qui est constitué de textes publiés en 1961. Il faut être conscient que l'utilisation de données constituées à une époque où John Kennedy arrivait au pouvoir, où les journalistes écrivaient leurs textes sur des machines à écrire et où très peu de monde avait accès à des ordinateurs, et uniquement de manière professionnelle, ne sera pas forcément suffisante dans des conditions écologiques* actuelles et probablement pas lorsque l'on traite de nouvelles technologies, par exemple.

3.1.5 Bilan

Ainsi, mes recherches sur la création des ressources langagières s'attachent à respecter autant que possible les hypothèses de travail suivantes. Les ressources doivent être :

- multilingues ou multi-monolingues ;
- librement disponibles ;
- accompagnées de descriptions de leur constitution et de leur destination originelle les plus précises possibles ;
- respectueuses des principes du FAIR :
  - Faciles à (re)trouver,
  - Accessibles,
  - Interopérables,
  - Réutilisables.

Respecter l'ensemble de ces principes n'est pas simple et prend du temps. Suivant les partenaires, il faut parfois négocier de manière âpre, de manière directe

(accord bilatéral) ou indirecte (preuve de concept, mise en place d'une alternative libre concurrente [20]...).

Après l'exposé de mes hypothèses de travail, voici quelques-unes de mes contributions à la constitution de ressources langagières.

3.2 Des corpus pour faire apprendre

3.2.1 Données naturelles

3.2.1.1 Des corpus en anglais annotés en sens

Pour la tâche de la désambiguïsation lexicale, les corpus annotés en sens sont souvent essentiels pour évaluer un système et indispensables pour atteindre de bonnes performances. Il existe aujourd'hui une vingtaine de corpus anglais annotés en sens, dans des formats variés, et utilisant différentes versions de WordNet. L'hypothèse principale de ce travail est qu'il devrait être possible de construire un système de désambiguïsation en utilisant n'importe lequel de ces corpus pendant la phase d'entraînement ou pendant la phase d'évaluation, indépendamment de leur objectif initial (apprentissage ou évaluation). L'UFSAC a fourni à la communauté l'ensemble des corpus en anglais annotés en sens que nous connaissions en 2018, dans ce format unifié, lorsque la licence le permet, avec des clés de sens converties à la dernière version de WordNet.

Grâce à Loïc Vial [21], nous avons produit et normalisé tout un ensemble de corpus annotés en sens issus du *Princeton WordNet*, UFSAC : Unification of Sense Annotated Corpora and Tools https://github.com/getalp/UFSAC. UFSAC comprend également le code source permettant de construire ces corpus à partir de leurs données originales, ainsi qu'une API Java complète permettant de manipuler les corpus dans ce format. Cette unification a largement facilité non seulement nos travaux sur la désambiguïsation lexicale de l'anglais, mais aussi des autres langues. La figure 3.1 présente un extrait de l'UFSAC.

Dans le cadre du projet Propicto, une mise à jour de ces ressources sera effectuée. Nous pensons en particulier au passage de l'ensemble des corpus du *Princeton WordNet* à l'*Open English WordNet* [22] et à l'ajout d'éventuels nouveaux corpus

20. C'est clairement un des buts pour lesquels nous construisons des outils de Communication Alternative et Augmentée. Montrer la richesse des possibilités offertes par l'utilisation de données gratuites, libres mais surtout interopérables.

21. Dans le cadre de sa thèse de doctorat soutenue en juillet 2020 et co-encadrée avec Benjamin Lecouteux (Université Grenoble Alpes).

22. https://github.com/globalwordnet

```
<corpus>
  <document id="d001" >
    <paragraph>
      <sentence id="d001.s001" >
       <word surface_form="Your" pos="PRP$" />
       <word surface_form="Oct" lemma="oct" pos="NN" />
       <word surface_form="." pos="." />
       <word surface_form="6" pos="CD" />
       <word surface_form="editorial" lemma="editorial"
            pos="NN" wn21_key="editorial%1:10:00::"
            wn30_key="editorial%1:10:00::"
            id="d001.s001.t001" />
        <word surface_form="&quot;" pos="&apos;&apos;" />
       <word surface_form="The" pos="NNP" />
       <word surface_form="Ill" lemma="ill" pos="JJ"
       wn21_key="ill%3:00:01::"
       wn30_key="ill%3:00:01::" id="d001.s001.t002" />
       <word surface_form="Homeless" lemma="homeless"
       pos="NNP" wn21_key="homeless%1:14:00::;
       homeless%1:18:00::" wn30_key=
       "homeless%1:14:00::;homeless%1:18:00::"
       id="d001.s001.t003" />
```

Figure 3.1 – Extrait de l'UFSAC – Les premières lignes de SemEval 2007 (Navigli et al., 2007) correspondant à « Your Oct. 6 editorial, the ill homeless… *»*

qui seraient mis à disposition par d'autres équipes.

3.2.1.2 Des corpus en anglais annotés en pictogrammes

Dans le cadre du projet Propicto, nous proposons UFPAC (*Unification of Pictograms Annotated Corpora*) suivant le raisonnement développé précédement à propos des corpus annotés en sens (voir section 3.2.1.1). UFPAC est un format unifié pour les corpus textuels annotés en pictogrammes. S'il existe des outils qui permettent de fabriquer des séquences de pictogrammes [23] à partir de texte comme Araword [24] ou Pictoselector [25], leur seules sorties possibles sont des formats peu propices à l'interopérabilité, comme les formats PDF et JPEG.

UFPAC contient les versions des corpus UFSAC annotés en pictogrammes *ARASAAC* [26]. Cette ressource a été produite par l'exploitation des liens mis en

23. La recherche est très sommaire, verbes à écrire à l'infinitif, affichage du premier pictogramme trouvé pour un terme,…
24. https://unadys.org/index.php/support-informatiques/logiciels-gratuits/item/3-araword
25. https://www.pictoselector.eu/fr/
26. https://arasaac.org

place entre le WordNet anglais et la base de pictogrammes *ARASAAC* (Schwab *et al.*, 2019, 2020b).

Dans le futur, on peut envisager, sous certaines conditions d'anonymisation et d'accord éclairé des utilisateurs et utilisatrices, d'exploiter également les textes créés à partir du logiciel *InterAACtion AugCom* [27] actuellement développé au LIG au sein du groupe InterAACtion [28]. Ce logiciel nous permettra dans des conditions écologiques de recueillir les échanges entre des personnes en situation de handicap cognitif et leurs aidants.

3.2.1.3 Corpus pour la recherche d'information

Historiquement, la Recherche d'Information textuelle est une tâche qui exploite essentiellement des données non annotées. Par conséquent, le domaine ne bénéficie pas de grandes quantités de requêtes résolues et libres de droit. De telles données sont pourtant indispensables pour développer, entraîner et évaluer des modèles d'apprentissage profond efficaces et réplicables pour la RI *ad hoc* Frej (2021).

Nous avons proposé [29] *WikIR* (Frej *et al.*, 2020a), une boîte à outils (*toolkit*) en libre accès pour construire des collections de RI basées sur *Wikipédia* avec de grandes quantités de requêtes résolues. Il s'agit de tirer parti de la nature semi-structurée de *Wikipédia* pour construire de façon automatique des collections de RI. *WikIR* [30].

Nous avons ainsi créé WikIR78k et WikIRS78k pour l'anglais qui sont des collection d'environ 78 000 requêtes résolues en utilisant l'ensemble des articles de *Wikipédia* (2,4 millions d'articles). Le titre (pour WikIR78k) ou la première phrase de chaque article (pour WikIRS78k) est choisi comme requête, le reste de l'article est choisi comme document. Nous attribuons une pertinence de 2 si la requête et le document ont été extraits du même article. Nous attribuons une pertinence de 1 s'il existe un lien hypertexte entre l'article du document et l'article de la requête.

Nous proposons ainsi :

— une collection avec des requêtes courtes et bien définies car basées sur les titres dans Wikipedia (WikIR78k)

— une collection avec des requêtes longues et bruitées car basées sur les premières phrases des articles dans Wikipedia (WikIRS78k) pour permettre d'étudier la robustesse des modèles de RI face à de telles requêtes.

27. http:\augcom.net
28. http://www.interaaction.com
29. Dans le cadre de la thèse de doctorat de Jibril Frej soutenue en février 2021 et co-encadrée avec Jean-Pierre Chevallet (Université Grenoble Alpes).
30. https://github.com/getalp/wikIR

MLWikir Frej *et al.* (2020b), est la version multilingue de Wikir pour une dizaine de langues issues de plusieurs familles (anglais, français, japonais, chinois, suédois, hollandais, russe, italien, espagnol, allemand). Elle rajoute 176 000 requêtes et 5 millions de documents. MLWikir serait facilement extensible à d'autres langues et est intégré au concentrateur *IR-dataset* dédié à la recherche d'informations de l'*Allen Institute for AI*[31] (MacAvaney *et al.*, 2021). Dans le cadre de la collaboration entre les équipes GETALP et MRIM du LIG, MLWikir devrait être mises à jour et étendu à plus de langues.

3.2.1.4 Corpus normalisés pour le français écrit

Dans le cadre du projet FlauBERT (Le *et al.*, 2020e,f), nous avons cherché à rassembler un maximum de corpus écrits bruts existants en français et librement disponibles afin d'obtenir une variété maximale de genres et de domaines.

Nous avons ainsi agrégé 24 sous-corpus de types divers (Wikipedia, livres, *Common Crawl*...).

Nos trois sources principales sont (1) les textes monolingues des campagnes d'évaluation WMT19 (Li *et al.*, 2019, 4 sous-corpus), (2) les textes en français de la collection OPUS (Tiedemann, 2012b, 8 sous-corpus), (3) le projet Wikimedia [32] (8 sous-corpus).

Pour chacun de ces corpus, nous fournissons les scripts permettant de les nettoyer et de les normaliser (passage en Unicode/UTF-8, unification des espaces, des retours à la ligne...). Ces scripts utilisent largement les librairies libres de la boite à outils Moses. Les corpus bruts font 270Go et les corpus normalisés font environ 70Go pour un peu moins de 489 millions de phrases et environs 12,8 milliards de mots.

Outre le projet FlauBERT, ces corpus ont été utilisés pour la création d'autres modèles comme BARThez [33] Eddine *et al.* (2021) ou dans le cadre du projet arts sciences [34] mené par le metteur en scène et docteur en informatique, Nicolas Zlatoff [35] à la manufacture de Lausanne et financé par le Fond National Suisse. Dans la pièce, les comédiens déclament les textes générés par un modèle transformeur génératif (GPT) appris sur les corpus bruts normalisés pour le français réalisés

31. https://ir-datasets.com
32. https://meta.wikimedia.org/w/index.php?title=Data_dumps&oldid=19312805
33. https://huggingface.co/moussaKam/barthez
34. https://www.manufacture.ch/download/docs/rwkf9mrh.pdf/Recherche%20-%20Chatbot%20-%20Le%20Courrier_01.04.21.pdf
35. https://www.manufacture.ch/fr/1695/Nicolas-Zlatoff

pour le projet FlauBERT avec un apprentissage supplémentaire sur des corpus de théâtre.

3.2.1.5 Corpus normalisés pour le français oral

Suivant des idées similaires à celles qui ont gouvernées le projet FlauBERT, nous avons cherché à rassembler un maximum de corpus oraux pour le français dans le projet LeBenchmark (Evain *et al.*, 2021b). Ici aussi, une variété maximale de genres et de domaines a été visée, ainsi que des singularités propres à l'oral comme le fait qu'il y ait plusieurs contextes d'enregistrement (studio, écologique, téléphone), différents accents, différents sexes, et cela qu'il s'agisse de lecture, d'interprété ou bien de spontané, les deux derniers pouvant véhiculer des émotions.

Dans ce travail, nous avons rassemblé de grandes quantités de données vocales en français, en nous concentrant sur la couverture de différents corpus, listés en annexe. Nous couvrons différents accents français (MLS, African Accented Speech, CaFE), spontanés (CFPP2000, ESLO2, MPF, TCOF), émotionnels joués (GEMEP, CaFE, Att-Hack), dialogues téléphoniques joués (PORTMEDIA), lus (MLS, African Accented French), et émissions de radio (EPAC). Le corpus rassemble également une grande variété de corpus de parole en français qui couvrent différents accents (MLS, African Accented Speech, CaFE), des émotions jouées (GEMEP, CaFE, Att-Hack), des dialogues téléphoniques (PORTMEDIA), lus (MLS, African Accented French) et des phrases spontanées (CFPP2000, ESLO2, MPF, TCOF), ainsi que de la parole radiodiffusée (EPAC).

Les meta-données permettent d'identifier les caractéristiques suivantes : un peu plus de 1,1 million d'énoncés pour 2 937 heures de parole, dont 1 115 heures de parole lue, 1 626 heures de discours diffusé, 127 heures de parole spontanée, 38 heures de dialogues téléphoniques joués et 29 heures de parole émotionnelle jouée. Enfin, il y a 1 824 heures de parole de locuteurs masculins, 1 034 heures de locuteurs féminins et 77 heures de locuteurs de sexe non précisé.

3.2.2 Données synthétiques

3.2.2.1 Principe

Les données synthétiques sont des données langagières générées artificiellement. Elles ont plusieurs avantages :

— Les données originelles sont privées[36] et ne peuvent à ce titre pas être librement partagées. On peut alors remplacer ces données privées par des

36. Pour des raisons liées au respect de la RGPD (Règlement général sur la protection des données), au droit d'auteur, à la confidentialité des données médicales…

données synthétiques qui doivent avoir des caractéristiques les plus similaires possibles aux données originelles tout en restant en surface les plus lointaines possibles des données originales.

— Les données manquent, parce qu'elles sont trop chères, trop compliquées à produire ou à collecter. C'est souvent le problème auquel nous avons été confrontés dans nos travaux parce que la langue considérée est moins dotée de manière générale (bengali, arabe) ou pour une tâche donnée (désambiguïsation lexicale par exemple).

En ce qui concerne les données textuelles, nous explorons actuellement des approches liées aux modèles génératifs dans la lignée des travaux de Vincent Claveau et de ses collègues (Claveau *et al.*, 2021; Claveau, 2020) dans des travaux liées au domaine de la défense ou des méthodes basées sur la rétrotraduction dans des cas de classification de documents avec très peu données dans le cas d'une collaboration avec l'entreprise Ixiade.

Notre principale contribution à la création de données textuelles de synthèse concerne le projet UFSAC-Mult.

3.2.2.2 Données créées à partir de données issues d'une autre langue : UFSAC-Mult

En désambiguïsation lexicale, avoir des corpus annotés en sens est crucial. Ainsi, l'état de l'art sur l'anglais montre que les meilleurs systèmes exploitent largement ces données annotées en sens (Navigli *et al.*, 2007; Navigli, 2009; Vial *et al.*, 2019c; Pasini *et al.*, 2021). Le coût de fabrication manuel de telles ressources est très élevé et doit être reconsidéré si on choisit un autre inventaire de sens, un autre domaine et bien entendu, une autre langue.

À partir de 2014, nous avons travaillé sur la traduction automatique et le portage des annotations pour créer automatiquement des corpus annotés dans une langue cible. Nos méthodes sont génériques et applicables dès que nous avons accès à un système de traduction de la langue originelle des corpus (généralement l'anglais) vers la langue cible. Trois approches ont été successivement tentées dans l'équipe suivant l'état de l'art en matière de traduction automatique de l'époque.

1) La traduction automatique statistique traditionnelle, donc fondée sur des modèles probabilistes (modèle de langage, modèle de traduction ainsi qu'un modèle de réordonnancement). Il s'agissait d'exploiter les tables de traduction pour porter les annotations vers la langue cible, en utilisant la boîte à outils Moses disponible sous licence libre LPL (Licence publique générale limitée GNU). La méthode a été appliquée au SemCor (Miller *et al.*, 1990) vers le bengali et le français (Nasiruddin *et al.*, 2015).

2) La deuxième version est une amélioration de la précédente corrigeant de nombreux problèmes de redondance et ordonnancement. Cette fois l'application, réalisée par Marwa Hadj Salah a été réalisée sur le SemCor vers l'arabe et le français.

3) La dernière version a été appliquée sur les corpus UFSAC ce qui a permis non seulement d'augmenter le nombre de mots annotés mais surtout d'intégrer la traduction automatique neuronale devenue entre-temps l'approche à l'état de l'art : application aux corpus UFSAC-ENG vers l'arabe et le français.

Les expériences réalisées sur la désambiguïsation lexicale du français et de l'arabe ont systématiquement montré une amélioration des résultats à mesure que la méthode de synthèse s'améliorait.

Les corpus UFSAC français et arabe sont actuellement disponibles librement pour la communauté lorsque la licence du corpus originel le permet. Ces corpus permettent d'obtenir des résultat à état de l'art sur ces deux langues et ont étés exploités dans nos travaux sur la génération automatique de pictogrammes (Projet ANR PROPICTO - 2021-2025) ou pour améliorer la traduction automatique neuronale par Marwa Hadj Salah lors de sa thèse [37]. Nous les exploitons également dans les travaux entamés début 2021 avec Rakia Saidi, doctorante au sein du Laboratoire d'Informatique, de Modélisation et de Traitement de l'Information et des Connaissances LIMTIC de Tunis (Tunisie) sous la direction de M. Fethi Jarray.

3.3 Des corpus pour évaluer

3.3.1 Un corpus d'évaluation pour la désambiguïsation lexicale de l'arabe : alignement de l'*OntoNote* avec le *Princeton WordNet*

Un des problèmes principal de la désambiguïsation lexicale de l'arabe est le manque de corpus annotés en sens pour l'évaluation. L'état de l'art du domaine, réalisé lors de la thèse de Marwa Hadj Salah (Hadj Salah, 2018) nous a montré que l'on trouvait uniquement des systèmes évalués sur des corpus différents, réalisés en interne et non rendus disponibles pour la communauté. L'évaluation comparative des approches était rendue de ce fait impossible.

Nous avons ainsi proposé une mise à jour de la partie arabe de *OntoNotes Release 5.0* d'une manière semi-automatisée pour obtenir des mots annotés en sens

37. https://github.com/getalp/UFSAC-ara

avec le *Princeton WordNet* 3.0. Le tableau 3.1 présente la description d'*OntoNotes Release 5.0* ainsi que le nombre de correspondances$_{WordNet}$ uniques ajoutées.

	#Lemmes	**#Lemmes uniques**	**#Sens uniques**	**#Correspondances$_{WordNet}$ uniques**
Verbes	3 990	150	642	4 182
Noms	8 534	111	463	1 376
Total	12 524	261	1 105	5 558

Table 3.1 – Description d'OntoNotes Release 5.0 après l'ajout des correspondances vers le Princeton WordNet 3.0

Pour des raison de droits, le corpus sera accessible gratuitement après signature d'un accord dans les prochains mois. Des négociations sont actuellement en cours avec ELRA/ELDA.

3.3.2 Un corpus multilingue, multi-genre et multi-granularité pour l'évaluation de la détection du plagiat translingue

Constitué dans le cadre de la thèse CIFRE de Jérémy Ferrero en partenariat avec la société Compilatio, ce corpus a l'intérêt de proposer une solution au problème du manque de données pour l'évaluation de la détection du plagiat translingue (Ferrero *et al.*, 2016). Les corpus préexistants sont principalement issus de corpus comparables et couvrent un domaine unique (littéraire, législatif, légendes d'images). Ce manque de diversité nous a semblé insuffisant pour permettre une évaluation précise de ces méthodes. Pour pallier ce manque, nous avons proposé un corpus multilingue, multi-genre et multi-granularité.

Ce corpus se focalise sur le français, l'anglais et l'espagnol, les langues qui intéressaient Compilatio, notre partenaire industriel.

Il est constitué des corpus parallèles et comparables précédemment constitués comme ceux issus de l'*International Competition on Plagiarism Detection*, d'*Europarl* ou ceux de la *Cross-Lingual Sentiment (Webis-CLS-10)* enrichis d'articles publiés en français dans TALN puis en anglais dans une des conférences rassemblées dans l'*ACL Anthology*.

Ses caractéristiques sont les suivantes (Ferrero *et al.*, 2016; Ferrero, 2017) :

— multilingue : il contient des alignements en français, anglais et espagnol ;

— aligné à différentes granularités (taille de textes) : syntagme nominal, phrase et document ;

— basé à la fois sur des corpus parallèles et des corpus comparables ;

— contenant des textes traduits manuellement et automatiquement ;

— contenant des passages altérés automatiquement (c'est-à-dire ayant subi une opération volontaire pour rendre la détection de plagiat plus compliquée), des passages altérés involontairement (présence de balises HTML ou de fautes d'orthographe, par exemple) et d'autres passages sans bruit ;

— hétérogène avec des documents écrits par différents types d'auteurs, allant du néophyte au professionnel, en passant par le contributeur du Web ;

— traitant plusieurs sujets et thèmes différents (littérature, science, avis d'internautes, etc.)

Ce corpus est disponible sous licence GNU à l'adresse https://github.com/FerreroJeremy/Cross-Language-Dataset.

3.3.3 Unification de corpus d'évaluation pour le français écrit : le projet FLUE

Le projet FLUE (**F**rench **L**anguage **U**nderstanding **E**valuation) est un projet visant à rassembler plusieurs corpus d'évaluation pour le français (Le *et al.*, 2020e,f). FLUE est le fruit de la collaboration entre :

— le Laboratoire d'informatique de Grenoble (Université Grenoble Alpes, CNRS) – Hang Le, Loïc Vial, Jibril Frej, Maximin Coavoux, Benjamin Lecouteux, Laurent Besacier, Didier Schwab ;

— Le Laboratoire de Linguistique Formelle – Université Paris Diderot, Université de Paris, CNRS – Vincent Segonne, Benoît Crabbé

— E.S.P.C.I, CNRS LAMSADE, PSL Research University – Alexandre Allauzen.

Le référentiel d'évaluation FLUE est actuellement composé de 7 tâches correspondant à différents niveaux d'analyse (syntaxique, sémantique) du traitement automatique du français.

— *Classification de texte*, qui consiste ici à établir si un texte est positif, neutre ou négatif quant à son sujet ;

— *Identification de paraphrases*, tâche qui consiste à identifier si des paires de phrases sont sémantiquement équivalentes ou non ;

— *Reconnaissance d'implications textuelles*, tâche qui considère une prémisse (p) et une hypothèse (h) et consiste à déterminer si p implique, contredit ou ni n'implique ni ne contredit h ;

— *Analyse syntaxique et étiquetage morphosyntaxique*, deux tâches qui consistent à analyser en constituants ou en dépendances les textes ;

— *Désambiguïsation lexicale des verbes et des noms*, deux tâches dont l'objectif consiste à assigner un sens, parmi un inventaire donné, à des mots d'une phrase.

Il s'agit d'un benchmark similaire au benchmark GLUE (Wang *et al.*, 2018) pour l'anglais. L'existence de tels référentiels d'évaluation sont très utiles pour stimuler des recherches reproductibles. Les bonnes performances obtenues avec des modèles contextuels préentrainés sur la plupart des tâches de TALN couvertes par GLUE ont conduit à son extension, SuperGLUE (Wang *et al.*, 2019) qui est un nouveau référentiel construit sur les mêmes principes, incluant un ensemble de tâches plus difficiles et variées. Une version chinoise de GLUE (Park *et al.*, 2021) [38] et une version coréenne (Xu *et al.*, 2020) sont aussi développées pour évaluer la performance des modèles sur ces langues.

FLUE est à ce jour et à notre connaissance le seul référentiel pour le français. FLUE est disponible en ligne sur un GitHub [39] ainsi que dans le concentrateur d'*Hugging Face* [40] (Lhoest *et al.*, 2021).

3.3.4 Unification de corpus d'évaluation pour le français oral : Le projet LeBenchmark

Le projet LeBenchmark est un projet visant à rassembler plusieurs corpus d'évaluation pour le français oral (Evain *et al.*, 2021b). LeBenchmark est le fruit de la collaboration entre :

— le Laboratoire d'informatique de Grenoble (Université Grenoble Alpes, CNRS) – Solène Evain, Ha Nguyen, Hang Le, Marcely Zanon Boito, Sina Alisamir, Ziyi Tong, Marco Dinarelli, Yannick Estève, Benjamin Lecouteux, François Portet, Solange Rossato, Fabien Ringeval, Didier Schwab ;

— le Laboratoire d'Informatique d'Avignon (Avignon Université) - Ha Nguyen, Natalia Tomashenko, Titouan Parcollet ;

— Naver Labs Europe, Grenoble – Laurent Besacier ;

— E.S.P.C.I, CNRS LAMSADE, PSL Research University – Alexandre Allauzen.

Le référentiel d'évaluation LeBenchmark est actuellement composé de corpus d'évaluation correspondant à plusieurs types de traitements :

38. https://github.com/chineseGLUE/chineseGLUE
39. https://github.com/getalp/Flaubert
40. https://huggingface.co/datasets/flue

— *la reconnaissance automatique de la parole (parole vers texte)*, tâche qui consiste à transcrire automatiquement de l'oral ;

— *la compréhension automatique de la parole*, tâche qui consiste à identifier certains éléments du discours dans des domaines précis (ex : action à effectuer, destination, type de restaurant…).

— *la reconnaissance automatique d'émotions* qui consiste à identifier un type d'émotion/d'affect véhiculé par la parole (ex : peur, tristesse, agacement…).

— *la Traduction automatique multilingue (parole vers texte)*, tâche qui consiste à traduire automatiquement de l'oral vers le texte correspondant dans plusieurs langues.

Inspiré de qui avait été fait avec FLUE pour le français écrit, LeBenchmark est, à ce jour et à notre connaissance le seul référentiel d'évaluation pour le français oral.

LeBenchmark est disponible en ligne sur un GitHub [41] ainsi que dans le concentrateur d'*Hugging Face* [42].

3.4 Conclusions du chapitre

Dans ce chapitre, j'ai présenté les hypothèses de construction des ressources langagières que je m'efforce de suivre. Il s'agit particulièrement de suivre les idées du *FAIR data*, tout en accordant une grande importance aux langues, à l'alignement de données et à la finesse de leur description. J'ai également argumenté pour des données libres dans un maximum de contextes, y compris industriels.

J'ai illustré ces hypothèses par des exemples de corpus et de bases lexicales auxquels j'ai contribué, qu'ils soient manuellement ou automatiquement créés, et quel que soit leur mode langagier (écrit, oral, pictographique…).

Dans les chapitres suivants, je me propose de montrer comment nous exploitons ces ressources, en commençant par mes contributions aux modèles préentraînés à base de vecteurs.

41. https://github.com/LeBenchmark
42. https://huggingface.co/LeBenchmark

Chapitre 4

Contributions aux modèles préentraînés à base de vecteurs – des vecteurs d'idées aux vecteurs d'usage

Sommaire

Nous avons vu dans le chapitre précédent quelques une de mes contributions à la constitution de ressources textuelles. Dans ce chapitre, nous présentons mes contributions aux modèles préentraînés basés sur des vecteurs [1].

Nous l'avons vu dans la section 2.3.1, ces travaux se placent dans le cadre de l'analyse sémantique de texte vue comme une tâche générique par opposition à une analyse sémantique dans le cadre d'une application unique comme la traduction automatique, la catégorisation de texte ou la recherche d'information. Ces travaux sont une sorte de fil rouge de mes recherches en TALP. J'ai commencé à étudier le sujet des représentation vectorielles en master en travaillant sur la fonction lexicale* d'antonymie* puis pendant ma thèse en élargissant le spectre aux autres fonctions lexicales et à leur lien avec cette analyse. Je les ai prolongé ensuite lors de mon postdoc à l'UTMK de l'USM en Malaisie puis au LIG à Grenoble. Ils sont encore aujourd'hui plus que jamais au centre de mes recherches avec nos travaux autour des représentations contextualisées et des modèles FlauBERT.

Je vais maintenant suivre ces quelque 20 années en évoquant plusieurs des travaux dans lesquels j'ai été impliqué et en analysant comment ils se placent, selon moi, dans la recherche en TALN.

1. Je choisis ici de ne pas qualifier plus le nom de ces modèles. Il y a peu de temps, Bommasani *et al.* (2021) proposaient le terme de « *Foundation models* » qui pourrait paraître pertinent, en suivant une métaphore issue de la construction d'édifices, puisque l'on peut construire d'autres modèles sur ces modèles. Ce terme est cependant décrié car certains, comme le Prix Turing Judea Pearl, qui trouvent le qualificatif de « *foundation* » trop fort pour ce qu'il est possible de réaliser grâce à ces modèles. Judea Pearl relève, en substance, que ces modèles ne permettent pas de représenter tous les aspects du sens. Il cite, par exemple, comprendre les humains ou produire des raisonnements (Marcus et Davis, 2021). De toutes façons, je ne trouve pas d'équivalent en français. J'ai pensé aux termes de « *modèles de fondation* », « *modèles fondamental* », « *modèles de base* », « *modèles socles* » ou même « *modèles de charpente* » mais aucun ne me satisfait au moment où j'écris ces lignes. Ainsi, je qualifie dans la suite les approches pour constituer ces modèles mais je ne qualifie plus ces modèles eux-mêmes.

4.1 Constitution de modèles préentrainés

On appelle modèles préentraînés* des modèles génériques qui peuvent ensuite être utilisés pour tout un ensemble de tâches (transfert d'apprentissage). Ces modèles peuvent être de différents types et dans mes travaux, ils sont basés sur des vecteurs ou des réseaux de neurones (profonds ou non). Nous mettions déjà à disposition nos modèles préentraînés au LIRMM*[2] mais, dans le Traitement Automatique des Langues et de la Parole, il était bien moins fréquent qu'aujourd'hui de mettre à disposition des modèles préentraînés. Google ne l'a fait qu'à partir de 2013 avec *Word2Vec*. C'est aujourd'hui devenu une norme et il existe même des concentrateurs comme celui d'*Hugging Face*[3].

Ces modèles ne sont pas trivialement interprétables par des humains comme peuvent l'être des textes[4] et ne sont destinés qu'à des machines, c'est pour cette raison que j'ai choisi d'en parler dans ce chapitre et non dans le précédent. Généralement, ces modèles nécessitent de grandes quantités de données et des ressources computationnelles importantes[5].

Suivant la chaîne de traitement que l'on voit se mettre en place ces deux dernières années (voir figure 4.1) et que nous suivons nous-mêmes, ces modèles préentraînés sont généralement appris sur du vocabulaire général, peuvent ensuite bénéficier d'un apprentissage supplémentaire sur des documents du domaine/genre[6] puis appliqué à des tâches par des algorithmes non-supervisées ou supervisées. Dans le cadre supervisé, l'affinage du modèle est réalisé pour résoudre la tâche. Il en existe plusieurs et nous les étudions que ce soit pour nos travaux pour l'écrit mais aussi sur l'oral (Le *et al.*, 2021b) :

— ajustement fin où l'ensemble des poids sont mis à jour en fonction des exemples ;

— ajustement fin où une partie des poids est mise à jour (généralement ceux directement liés à la tâche) ;

2. Les modèles étaient disponibles en ligne via des formulaires Web ou à la demande.
3. https://huggingface.co/models – consulté le 4 septembre 2021.
4. Y compris ceux qui sont artificiellement générés.
5. Dans mes recherches, j'ai toujours exploité au maximum les ressources disponibles à ce moment là. Je suis ainsi passé de machines monoprocesseurs communiquant via le réseau à des machines multiprocesseurs à des grilles de calcul (ex : Grid 5000) puis à des GPU, idéals pour les calculs vectoriels et aujourd'hui des grilles de GPU (Jean Zay). Ces changements rendent compte de la complexité de la construction des modèles.
6. Le vocabulaire ne semble pas totalement stabilisé. Par exemple, certains utilisent affinage (*fine-tuning*) à la fois pour l'opération qui consiste à adapter au domaine (prolongation d'apprentissage, non-supervisé) et pour la résolution de la tâche (affinage sur la tâche, supervisé). J'utilise ici un vocabulaire qui me paraît relativement cohérent mais qui pourrait paraître en partie différent de celui que l'on trouve dans la littérature scientifique.

— ajustement fin où les poids liés à la tâche sont mis à jour durant les premières époques suivi d'un ajustement global ;

— approche par adaptateurs* qui consiste à rajouter des couches dans le modèle, couches qui seront mises à jour alors que les autres restent gelées.

Le choix doit se faire en fonction des ressources disponibles. Généralement, l'ajustement fin où les poids liés à la tâche sont mis à jour durant les premières époques suivi d'un ajustement global fonctionne le mieux (Yosinski *et al.*, 2014) mais pour certaines tâches, il nécessite de grandes ressources matérielles.

Une alternative relativement récente est celle des adaptateurs proposée par (Houlsby *et al.*, 2019). Les adaptateurs sont de petits modules rajoutés entre les couches transformeur du modèle. Lors de l'apprentissage, seuls les adaptateurs sont mis à jour. Ainsi, peu de paramètres sont rajoutés au modèle et le nombre de paramètres à mettre à jour est nettement plus petit. En revanche, le modèle résultat comporte plus de paramètres et en phase d'exploitation, l'inférence est nettement plus longue (Rücklé *et al.*, 2020).

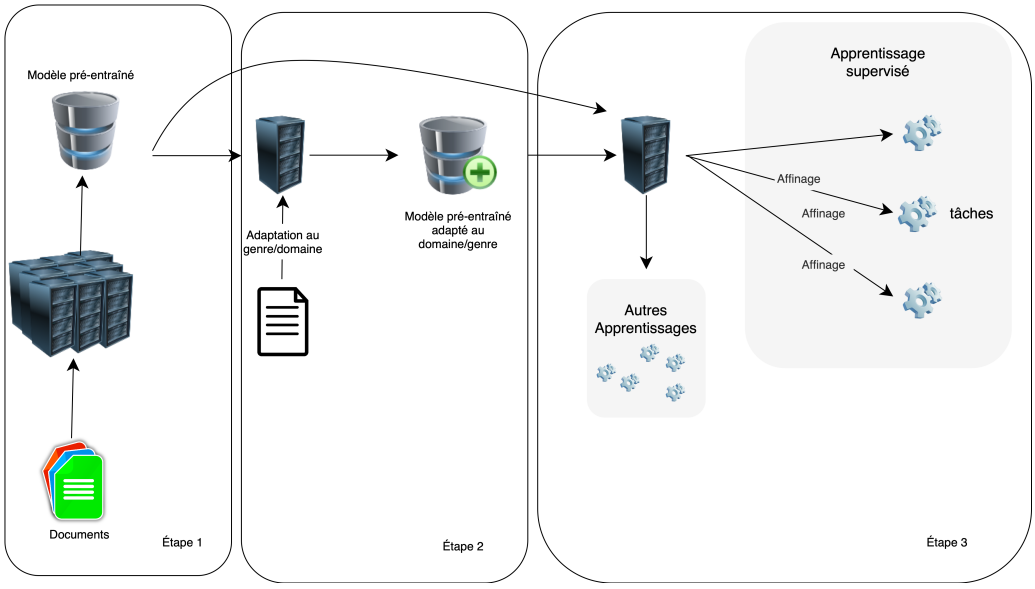

Figure 4.1 – La chaîne de traitement et ses trois étapes devenues classiques aujourd'hui.

La figure 4.1 présente une vision globale de la la chaîne de traitement et de ses 3 étapes :

1. préentraînement de modèles sur un grand ensemble de documents, données régulièrement mises à disposition de la communauté ;

2. étape optionnelle de prolongation de l'apprentissage sur des documents adaptés au domaine, données parfois mises à disposition de la communauté ;
3. apprentissage pour une tâche donnée.

Ces étapes sont séquentielles et les personnes qui s'intéressent aux étapes 1 et 2 mettent souvent leur modèles à disposition de la communauté [7]. Il est important de comprendre que le coût d'entrée n'est absolument pas le même entre les 3 étapes en particulier en ce qui concerne les ressources matérielles nécessaires [8] et il semble actuellement de plus en plus difficile de rivaliser avec les grands acteurs privés sur l'étape 1. La plupart des chercheurs ne s'intéressent qu'à l'étape 3 et parfois à la seconde.

Maintenant que nous avons défini ce que nous entendions par modèle préentrainé, nous allons plus précisément voir ceux qui exploitent des représentations vectorielles.

4.2 Modèles basés sur des vecteurs

Schématiquement, on peut considérer qu'il existe deux types de vecteurs utilisés en TALN, chacun inspiré par une théorie linguistique : ceux issus de la sémantique componentielle, les vecteurs d'idées, et ceux issus de la sémantique distributionnelle, les vecteurs distributionnels. La figure 4.2 présente un schéma général des modèles vectoriels.

Mes premières années (2001-2012) ont été consacrées aux premiers avant que je ne commence à m'intéresser petit à petit aux seconds. J'ai consacré mon master/DEA et ma thèse entière à ce sujet particulier et à ses rapports avec les réseaux lexicaux, sujet que j'ai continué en Malaisie en travaillant sur le malais et l'anglais avec LIM Lian Tze (doctorante) puis à Grenoble à travers 2 projets ANR (OMNIA et VIDEOSENSE) qui ont conduit aux travaux sur la désambiguïsation lexicale.

Mathématiquement, ces vecteurs sont les mêmes objets, de simples vecteurs d'entiers ou de nombre réels qui permettent de représenter certains aspects du sens. Les mêmes opérations mathématiques peuvent leurs être appliquées et certaines d'entre elles ont des interprétations linguistiques raisonnables.

7. En particulier, sur le concentrateur d'*Hugging Face*– https://huggingface.co/models, consulté le 4 septembre 2021.
8. Le coût en calcul est estimé à plusieurs milliers d'euros. De fait, nous avons obtenu des ressources computationnelles correspondant à plus de 200 000 euros sur la grille de calcul Jean Zay.

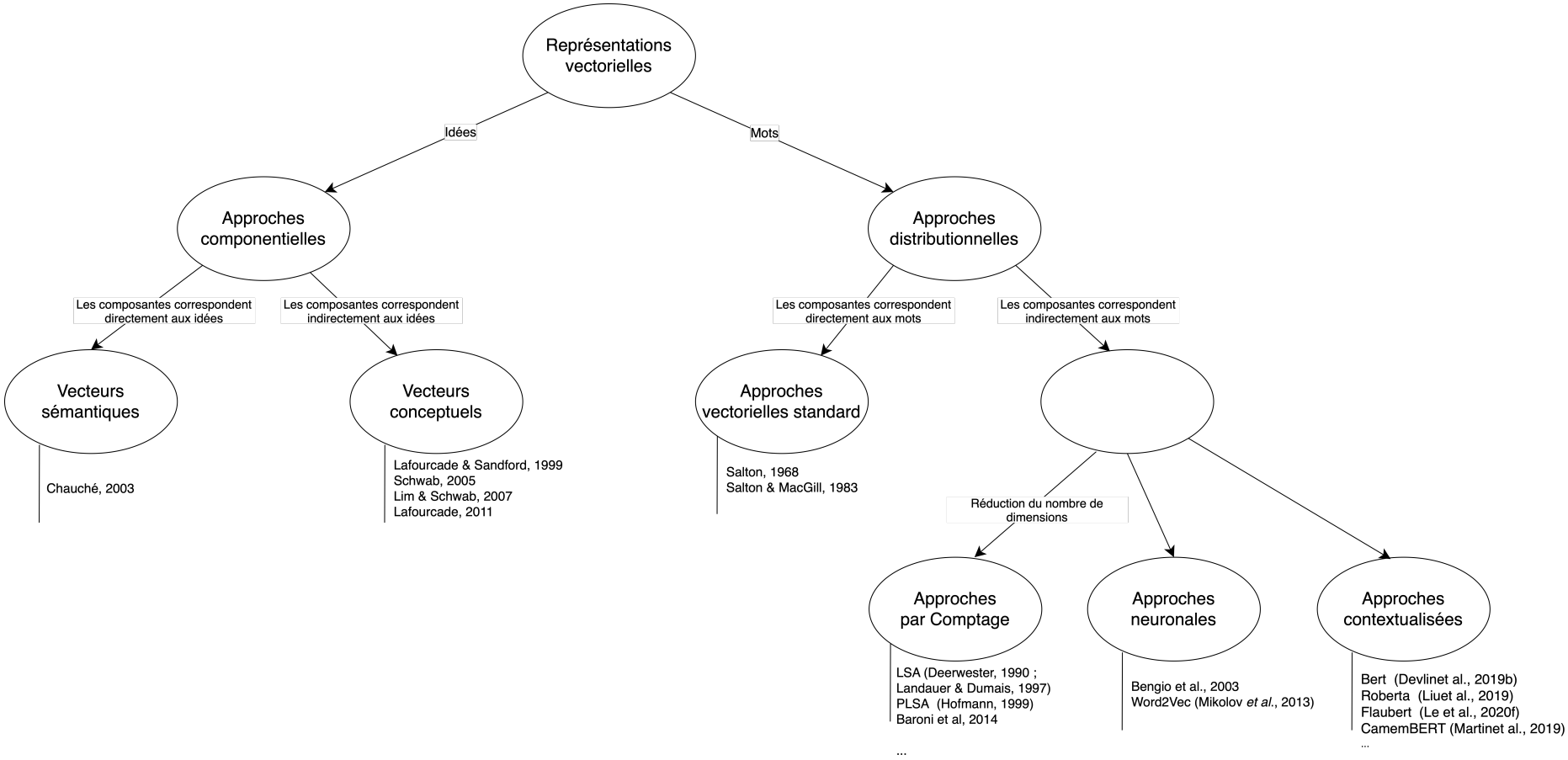

FIGURE 4.2 – Une typologie des vecteurs utilisés en Traitement Automatique des Langues Naturelles

Pour les vecteurs issus de la linguistique componentielle, les composantes correspondent directement ou indirectement à des idées de base sensées permettre par composition de représenter l'ensemble des idées exprimables en langue. Pour ceux issus de la linguistique distributionnelle, les composantes correspondent directement ou indirectement à des mots représentant le contexte dans lequel évolue l'objet linguistique que l'on veut représenter. Ce ne sont pas des linguistiques qui s'opposent et au contraire, elles sont tout à fait compatibles. Représenter les mots grâce aux contextes linguistiques dans lesquels on les trouve donc en fonction du pouvoir d'associativité qu'ils possèdent ou ne possèdent pas entre eux n'interdit pas que chacun soit individuellement décomposables en idées plus basiques. On peut même considérer que ce sont ces idées basiques qui pourraient expliquer ce pouvoir d'associativité.

De fait, si les apprentissages diffèrent, les opérations sont généralement applicables et aux uns et aux autres avec des interprétations linguistiques identiques [9]. Ainsi, il est tout à fait possible d'interpréter des sommes vectorielles* comme des unions d'idées, des produits termes à terme comme des intersections d'idées, les distances angulaires comme des distances thématiques et les proximités comme des voisinages thématiques.

4.3 Vecteurs d'idées et vecteurs conceptuels

4.3.1 Principe

La linguistique componentielle postule que les objets linguistiques* sont composés d'éléments plus simples (Schwab, 2005, p. 42). Ces éléments, qui ne sont pas à considérer comme des atomes [10] c'est-à-dire que l'on pourrait éventuellement encore les découper et qu'ils peuvent se recouper. Les lointains héritiers de Gottfried Wilhelm Leibniz (*1646 - 1716*) et de son alphabet de la pensée les appellent primitives, *primes* (Wierzbicka, 1996), constituants (Greimas, 1984), attributs, sèmes (Rastier, 1991), idées… (Schwab, 2005, p. 42 et suivantes).

Pour les vecteurs d'idées, les composantes correspondent directement (vecteurs sémantiques, 1992–2005) ou indirectement (vecteurs conceptuels, depuis

9. Seule la fonction d'antonymie selon la méthode que nous avions conçue pendant mon master et qui nécessite un ensemble de concepts opposés établi *a priori* me semble incompatible. La conception d'une fonction exploitant la définition basée sur la symétrie que nous proposions alors et calculant des synonymes (et plus exactement une fonction estimant un vecteurs antonyme "idéal") semble possible. Les travaux d'Enrico Santus pendant sa thèse (Santus, 2016; Santus *et al.*, 2014b,a) pourraient consitituer également une piste intéressante.

10. Au sens étymologique du terme, que l'on ne peut découper.

1999) à des idées. Dans leur version française, basée sur le thésaurus Larousse (Larousse, 1992) qui a 873 concepts, les vecteurs ont donc 873 composantes. Ainsi, la première composante peut être considérée comme une indication de l'intensité [11] du concept EXISTENCE, la deuxième de celle d'INEXISTENCE, la 516ème de celle du concept de LIBERTÉ et, enfin, la dernière de celle de JOUETS.

Les vecteurs conceptuels sont généralement normalisés, c'est-à-dire qu'ils font tous la même norme, la même longueur. On peut ainsi considérer que les objets linguistiques* que nous manipulons sont projetés sur une hypersphère et plus précisément sur la partie de l'hypersphère correspondant aux composantes strictement positives. La figure 4.3 est une illustration de cette projection.

Dans ce modèle, la norme n'est pas considérée comme une information qualitative. Les idées s'apprécient uniquement les unes par rapport aux autres non seulement à l'intérieur des vecteurs mais également entre les vecteurs. Ainsi, il est plus pertinent de comparer les proportions des différentes idées à l'intérieur d'un terme, d'un vecteur, plutôt que de les analyser de façon absolue (Schwab, 2005, p. 60). Notons que ça ne signifie pas que toutes les opérations aboutissent à des vecteurs dont la norme est égale à l'unité [12] mais que les vecteurs stockés sont toujours normés. Notons qu'il s'agit d'une spécificité des vecteurs conceptuels que les autres types de vecteurs [13] ne partagent nécessairement pas. Cette normalisation offrait également l'avantage d'accélérer plusieurs calculs nécessitant la norme comme, par exemple, la distance thématique.

4.3.2 Distance thématique

La comparaison entre deux vecteurs se fait grâce à la distance angulaire D_A. Pour deux vecteurs conceptuels A et B,

$$Sim(X,Y) = \cos(\widehat{X,Y}) = \frac{X \cdot Y}{\|X\| \times \|Y\|} \tag{4.1}$$

$$D_A(A,B) = arccos(Sim(A,B)) \tag{4.2}$$

Intuitivement, cette fonction constitue une évaluation de la *proximité thématique* et en pratique la mesure de l'angle entre les deux vecteurs. Empiriquement,

11. Et sont strictement positives.
12. Par exemple, la mise en puissance du vecteur, le produit terme à terme entre deux vecteurs ou la soustraction de vecteurs n'aboutissent pas à un vecteur normé.
13. Y compris les vecteurs sémantique de Jacques Chauché.

nous estimons que pour une distance $D_A(X, Y) \leq \frac{\pi}{4}$ (45°), X et Y sont thématiquement proches et partagent plusieurs concepts. Pour $D_A(X, Y) \geq \frac{\pi}{4}$, la proximité thématique est considérée comme faible et aux alentours de $\frac{\pi}{2}$ (90°), X et Y n'ont aucune relation.

L'exemple suivant permet de mieux comprendre cette distance dans le cas des vecteurs conceptuels mais, encore une fois, on peut également retrouver ce genre de choses avec les autres vecteurs.

D_A('fourmilier', 'fourmilier') = 0 (0°) ; similarité = 1

D_A('fourmilier', 'animal') = 0.45 (26°) ; similarité = 0.9

D_A('fourmilier', 'train') = 1.18 (68°) ; similarité = 0.38

D_A('fourmilier', 'mammifère') = 0.36 (21°) ; similarité = 0.94

D_A('fourmilier', 'quadrupède') = 0.42 (24°) ; similarité = 0.91

D_A('fourmilier', 'fourmis') = 0.26 (15°) ; similarité = 0.97

Le premier résultat a une interprétation directe, ‹*fourmilier*› ne peut être plus proche d'autre chose que de lui même. Le fait qu'un ‹*fourmilier*› soit un ‹*mammifère*› explique le deuxième résultat. Un ‹*fourmilier*› n'a que peu de rapport avec un ‹*train*› ce qui explique l'angle plus important. Dans le dernier exemple, l'angle peu important entre ‹*fourmilier*› et ‹*fourmi*› se comprend si on se rappelle que D_A est une distance thématique et non une distance ontologique. L'examen de la définition de fourmilier, « *mammifère qui se nourrit de fourmis* », explique le résultat.

4.3.3 Le voisinage thématique, une vision continue de la thématique

La fonction de voisinage thématique permet de connaître les items lexicaux voisins d'un item lexical donné. On définit $\mathcal{V}$, la fonction de voisinage qui renvoie les k items les plus proches en termes de distance angulaire D_A d'un texte Z dans une base vectorielle. Soit :

$$\forall X \in \mathcal{V}(D_A, Z, k), \quad \forall Y \notin \mathcal{V}(D_A, Z, k), \quad \mathrm{D}_A(X, Z) \leq \mathrm{D}_A(Y, Z) \tag{4.3}$$

avec $|\mathcal{V}(D_A, Z, k)| = k$. Par exemple, les 7 termes proches et ordonnés par distance thématique croissante du nom ‹*mort*› peuvent être :

$\mathcal{V}(D_A$, ‹*mort*›, 7) = (‹*mort*› 0) (‹*meurtre*› 0.367) (‹*tueur*› 0.377) (‹*âge de la vie*› 0.481) (‹*tyrannicide*› 0.516) (‹*tuer*› 0.579) (‹*mort :adj*› 0.582).

La méthode de voisinage peut être utilisée lors de l'apprentissage des vecteurs conceptuels pour vérifier la cohérence globale de la base ou en phase d'exploitation pour trouver le meilleur mot à utiliser dans un énoncé. Ainsi, elle constitue un nouvel outil pour accéder aux mots et à leur sens, complémentaire à ceux décrits dans (Zock, 2002) comme la forme, la morphologie ou la navigation dans un grand réseau lexical. La fonction de voisinage permet ainsi une navigation dans le domaine du continu contrairement aux réseaux sémantiques qui ne permettent qu'une navigation discrète.

La similarité illustre bien l'interprétation linguistique qu'il est possible de réaliser avec des opérations mathématiques appliquées aux vecteurs. Ainsi si toute opération calculable sur un vecteur est possible, seules une partie d'entre elles semble avoir une interprétation linguistique raisonnable. Ainsi, une somme normalisée* entre deux vecteurs correspond à l'union des idées des objets correspondants, le produit terme à terme* à l'intersection de leurs idées et une combinaison particulière des deux $X \oplus (X \odot Y)$ correspond à une forme de contextualisation [14]. En revanche, $X \odot (X \oplus Y)$ ne semble pas avoir d'interprétation linguistique raisonnable évidente.

Avant mon arrivée, (Lafourcade et Prince, 2001a,b), mes futurs directeurs de thèse, avaient défini un ensemble de fonctions de synonymie. Dans mon master (Schwab, 2001) et les mois qui ont suivi, nous avons fait de même avec les fonctions d'antonymie (Schwab *et al.*, 2002a,b,c, 2005) et tout un ensemble d'autres fonctions lexicales. Nous avons analysé les limites des vecteurs seuls et leur complémentarité avec les réseaux lexicaux dans ma thèse (Schwab, 2005) et plusieurs publications (Lim et Schwab, 2008; Schwab *et al.*, 2007).

Voilà quelques exemples de ce qu'il est possible de faire avec des vecteurs conceptuels. Intéressons-nous maintenant à leur apprentissage. Nous avons réalisé 3 expériences principales sur les vecteurs conceptuels (Montpellier 2001-2005, Penang 2006-2007 et Grenoble 2007-2012). Leurs propriétés sont les mêmes mais leur apprentissage, leurs langues se différencient, nous allons les étudier avant de petit à petit basculer vers les vecteurs distributionnels et mes contributions plus récentes.

14. Dite faible pour les vecteurs conceptuels par opposition à la contextualisation dite forte qui met en jeu les vecteurs des acceptions, c'est-à-dire les vecteurs correspondant aux sens des mots.

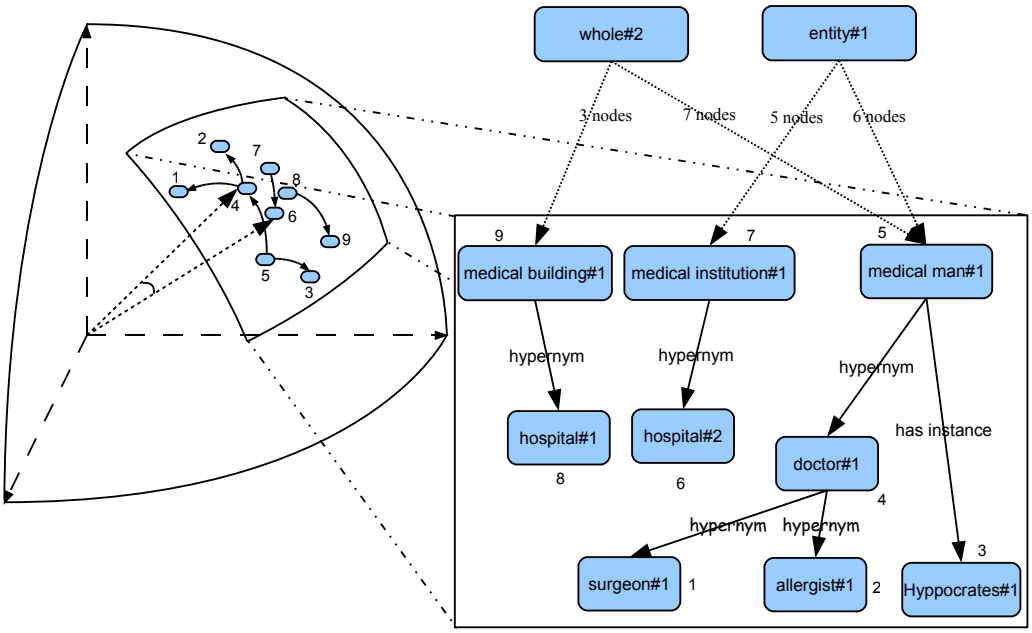

FIGURE 4.3 – Projection d'un réseau lexical (ici celui de WordNet) sur l'hypersphère des vecteurs conceptuels – issu de (Lim et Schwab, 2008). Ce schéma permet de comprendre une des complémentarités entre les réseaux lexicaux et les vecteurs que nous étudions plus particulièrement dans (Schwab et al., 2007). Alors que les différents sens d'«hôpital» («hospital1» et «hospital2») et «chirurgien» sont loin dans la hiérarchie de WordNet leurs vecteurs sont proches puisque leur thématique est proche. Il s'agit d'une autre illustration du problème du Tennis de Fellbaum (1998) qui remarquait que «balle de tennis», «raquette» et «arbitre» apparaissent dans différentes branches de l'arborescence, alors qu'ils sont tous susceptibles d'être nécessaires lorsqu'on parle du sujet qui les réunit, un «match de tennis» (Zock et Schwab, 2014).

4.3.4 Apprentissage des vecteurs conceptuels

Dans ma thèse, je donnais dans le chapitre 5, six hypothèses de construction de tels modèles.

Hypothèse I - Représentation Hybride du sens : approche combinant informations thématiques et informations lexicales : Le sens est représenté dans la base lexicale par des objets lexicaux, composés d'un vecteur conceptuel et d'informations lexicales comme la morphologie, la fréquence d'utilisation, les relations lexicales, etc représentées sous la forme d'un réseau lexical (graphe). Chaque terme du lexique est représenté sous la forme d'un objet lexical appelé ITEM LEXICAL. Cette représentation hybride du sens, cette complémentarité entre vecteurs et réseaux se

justifie par

1. la limitations des vecteurs conceptuels dans la modélisation des fonctions lexicales ;
2. le désir d'allier, dans un but d'analyse sémantique, au fort rappel des vecteurs conceptuels la forte précision des réseaux sémantiques ;
3. le désir d'allier les informations implicites issues des textes des informations explicites issues des réseaux ;
4. une certaine adéquation avec le modèle cognitif humain puisque la gestion des idées et la gestion des formes ne sont pas gérées ensemble dans le cerveau.

Hypothèse II - Utilisation conjointe d'objets lexicaux de type acception et item lexical : Afin de permettre de désambiguïser les termes, nous représentions également le sens que chacun des items lexicaux peuvent prendre dans un objet de type ACCEPTION.

Hypothèse III - Génération automatique : Dans les dictionnaires classiques comme le (Larousse, 2022) ou le (Robert, 2020) pour le français, il y a environ 80 000 termes, dont beaucoup sont polysémiques. Ainsi, dans notre expérience du français lors de ma thèse, portant sur plus de plus de 120 000 entrées, le taux de polysémie est d'environ 55%. Pour les termes polysémiques, il y a en moyenne 5 définitions pour chaque entrée, donc il nous faudrait indexer environ 400 000 ACCEPTIONS, ce qui serait déraisonnable à faire manuellement. Cette automatisation à partir informations extraites de sources hétérogènes telles que les dictionnaires traditionnels, dictionnaires de synonymes et d'antonymes, sites Web, etc. Le troisième type d'objet lexical est défini par cette hypothèse : une LEXIE rassemble toutes les informations extractibles d'une définition.

Hypothèse IV - Analyse multi-source : afin de pallier les lacunes des définitions (couverture du lexique, métalangage, définitions très ambiguës).

Hypothèse V - Apprentissage permanent : Afin de se laisser la possibilité d'apprendre régulièrement de nouveaux termes qui arrivent dans l'actualité [15]

Hypothèse VI - Double boucle : La double boucle est un élément structurel invariant qui permet l'action sur son environnement et qui est un produit de cette

15. Par exemple, le mot ‹‹déconfinement›› a dû rarement être utilisé avant 2020 et *Google trends* l'outil de Google permettant de retrouver les recherche réalisées sur leur moteur semble le montrer (https://trends.google.fr/trends/explore?q=%2Fg%2F11j5nsk4wq&date=all&geo=FR). Il en est de même pour les personnalités qui arrivent brusquement sur le devant de la scène médiatique. Qui a vécu la crise sanitaire commencée à l'hiver 2019-2020 trouvera facilement des exemples.

action (Lecerf, 1997, p. 177). Chez un individu, on retrouve cette structure du niveau le plus bas, la cellule, au niveau le plus haut de la double boucle en passant par le système nerveux central et les neurones qui le constituent [16]. Nous implantions cette hypothèse par un système multi-agent où chaque agent (Analyse sémantique, base lexicale, agent d'antonymie, de synonymie…) pouvait fournir les informations demandées par les autres agents mais également s'automodifier en fonction des informations rencontrées et distribuées sur un ensemble de machines. Deux versions de ce système ont existé. La première, fabriquée *ad hoc* pendant ma thèse est précisément décrite dans (Schwab, 2005; Schwab *et al.*, 2007) et une seconde Blexisma2 (Schwab et Lim, 2008), basée sur une boite à outil de gestion de système multi-agents*, MadKit [17] (Ferber *et al.*, 2004; Mansour et Ferber, 2007).

Les deux premières correspondent à la représentation des objets lexicaux et les 4 dernières à la manière d'apprendre ces représentations.

4.3.5 Calcul des vecteurs par analyse sémantique

L'algorithme principal d'apprentissage des vecteurs est une analyse sémantique dite par « *remonté-redescente* » (Schwab, 2005, p. 83). Son principe général est de se baser sur l'hypothèse de compositionnalité c'est-à-dire que « *le sens du tout est calculable à partir du sens de ses parties* ». Le vecteur d'un texte est donc calculé de façon générale par une fonction ayant pour paramètres l'ensemble des vecteurs des items du texte. Il s'agit ainsi d'une combinaison des vecteurs du texte. Souvent, ces combinaisons sont linéaires mais on peut imaginer d'autres opérations (comme une mise en puissance pour simuler une intensification par exemple ou l'application d'une fonction lexicale d'antonymie pour certaines négations).

Soit s un segment textuel* (texte, paragraphe, phrase, syntagme…) quelconque de longueur n, les n mots du segment sont représentés tels que :

$$s = \{w_1, w_2, w_3, \ldots, w_n\} \tag{4.4}$$

16. Notons que, pour moi, ce concept ne se différencie pas de celui de co-évolution que j'utilise à plusieurs reprises dans ce document pour le niveau le plus haut. La co-évolution concerne l'évolution d'un couple d'entités (agents, programmes, individus…) qui évoluent de manière conjointe alors que la doucle boucle au niveau le plus haut analyse un individu avec son environnement. Ainsi, la co-évolution est constituée de deux doubles boucles considérées au niveau des organismes, la double boucle de l'autre entité faisant partie de l'environnement.

17. `http://www.madkit.net` – consulté le 5 septembre 2021.

avec w_i le ième mot du segment s.

$$V = \sum_{i=1\,,\,w_i \in S}^{n} (vector(w_i)\,.\,\varphi(w_i)) \tag{4.5}$$

où w_i est le ième mot du segment s, $\varphi(w)$ est le poids affecté à ce mot et $vector$ une fonction qui s'applique au mot pour calculer le vecteur correspondant. La fonction $vector$ n'est pas le vecteur brut de l'item lexical correspondant au mot mais peut-être, par exemple, une combinaison linéaire des vecteurs de chaque ACCEPTION fonction du contexte [18] ou l'application d'une fonction lexicale comme antonymie comme nous le faisions pour certaines négations (Schwab, 2005, p. 120).

Il y a plusieurs manières d'estimer le poids pour chacun des mots. Ainsi, Le moyen que nous avons le plus utilisé est de réaliser une analyse syntaxique [19] mais dans notre expérience de Penang (Lim et Schwab, 2008; Schwab *et al.*, 2007), nous avons utilisé la forme logique des définitions du *Princeton WordNet*.

$$\varphi(w) = weight(syntactic\_role) \tag{4.6}$$

où $syntactic\_role$ correspond au rôle syntaxique comme défini dans SYGFRAN (Chauché *et al.*, 2003) et $weight$ est la fonction qui permet d'obtenir son poids. Ainsi, dans le segment « *voile de bateau* », ‹*voile*› est gouverneur syntaxique, son vecteur aura donc un poids plus important que ‹*bateau*›. En revanche, l'inverse sera appliqué pour ‹*bateau à voile*›.

4.3.6 Vecteurs conceptuels par concepts prédéfinis et vecteurs conceptuels par émergence

Globalement, les opérations décrites ci-dessus pour les deux types de vecteurs, ceux construits par concepts prédéfinis (à partir de 1999 et pendant toute ma thèse) comme ceux construits par émergence (à partir de 2006) ainsi que les vecteurs distributionnels. Ces vecteurs diffèrent sur leur apprentissage.

18. Par l'opération de contextualisation forte définie dans (Schwab, 2005, p. 81) qui exploite autant d'informations contextuelles que possible (rôle syntaxique, vecteur du contexte, fréquence de l'ACCEPTION).

19. Sur le français avec SYGFRAN (Chauché, 1984) – `https://www.sygtext.fr` – et plus tard, le *Malt Parser* (Nivre, 2006) – `http://www.maltparser.org` – lorsque nous avons analysé plusieurs langues à la fois.

4.3.6.1 Apprentissage par concepts prédéfinis

L'idée est de partir d'un ensemble considéré comme pertinent de vecteurs, un noyau réduit de 2 000 termes choisis pour leur fréquence dans le corpus d'apprentissage et indexés manuellement grâce à un apprentissage considéré comme cohérent.

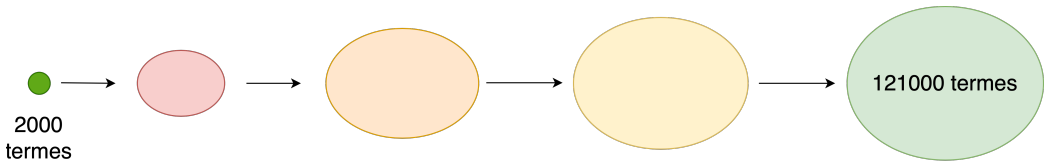

FIGURE 4.4 – L'expérience que j'ai menée entre 2001 et 2005 au LIRMM à Montpellier. Nous sommes partis de 2 000 termes pertinents et grâce à un apprentissage cohérents qui suivait les six hypothèses, nous sommes arrivés à 121 000 termes pertinents.

Pendant tout le début des années 2000, nous avons mené un certain nombre d'expériences avec cette construction des vecteurs. (Schwab, 2005) les retrace jusqu'en 2005. Pourtant, plusieurs limites ont rapidement émergé. Lafourcade (2006) suivi de Lim et Schwab (2008); Schwab *et al.* (2007) les retracent.

On le voit, la construction de ces vecteurs n'était pas simple pour au moins 3 raisons. La première est qu'il est relativement pénible de fabriquer manuellement le noyau car c'est une opération potentiellement longue, sujette à erreur (oubli de sens par exemple) et une source de subjectivité importante par le choix de mots qu'il contient.

La seconde raison est que certaines parties de l'espace se retrouvent densément peuplées [20] et d'autres presques désertes. Ainsi, l'interprétation des mesures (la similarité par exemple) ne devrait pas forcément être la même dans deux zones de l'espace.

Enfin, la dernière raison est de s'apercevoir que ce que dont je présentais à la figure 4.5 était assez imprécis. En effet, la pertinence de l'ensemble des vecteurs n'est pas due à la pertinence initiale des vecteurs mais est uniquement due à la cohérence des vecteurs les uns par rapport aux autres.

En bref, c'est la cohérence des vecteurs entre eux qui fait leur pertinence.

L'idée de l'apprentissage par émergence est la conséquence de cette triple constatation.

20. Celles liées au sexe par exemple pour lequel le français se révèle très imaginatif.

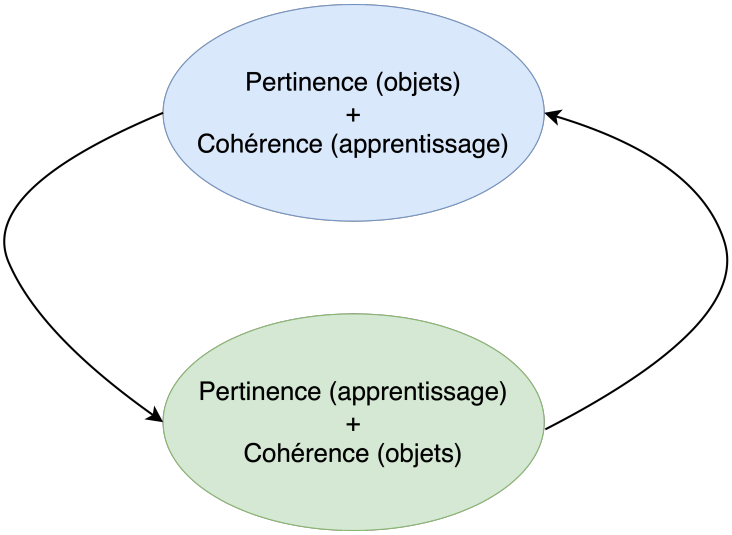

FIGURE 4.5 – L'idée fondatrice des vecteurs par concepts prédéfinis est que la pertinence des objets est assurée par la pertinence du noyau initial et par la cohérence de l'apprentissage, qui lui-même est pertinent en retour.

4.3.6.2 Apprentissage par émergence

Nous avons ainsi cherché des moyens de simplifier l'apprentissage d'un côté en s'affranchissant de certaines hypothèses, la 4 et la 5 mais surtout en redéfinissant la manière d'implanter la 3ème en suivant l'idée initiale de (Lafourcade, 2006) analysée ci-dessus.

L'approche par émergence s'affranchit ainsi de tout ensemble de concepts initiaux et de tout noyau. Seule la taille du vecteur est fixée a priori. Le mode de construction des vecteurs est identique à l'apprentissage précédent à la différence que si un des vecteurs entrant en jeu est inexistant, car non encore calculé, alors ce vecteur est tiré au hasard. Le processus de calcul est itéré jusqu'à convergence de chaque vecteur. Comme nous le montre de façon plus détaillée (Lafourcade, 2006), il y a un certain nombre d'avantages à utiliser ce modèle. Le premier d'entre eux est de pouvoir choisir librement la quantité de ressources que l'on souhaite utiliser en choisissant la taille des vecteurs de façon appropriée. Pour donner une idée de l'importance de ce choix, une base de 500 000 vecteurs de dimension 1 000 fait environ 2Go, de taille 2 000, 4Go [21]...Comme il ne serait pas alors ni raisonnable ni facile de définir un jeu de concepts de la taille choisie, autant chercher une approche nous permettant de nous en passer. De plus, ce qui peut sembler un pis-aller ou au mieux un compromis, s'avère un avantage car la densité lexicale dans l'es-

21. Rappelons que ce sont des tailles importantes en 2006.

pace des mots calculés par émergence est bien plus constante que dans un espace où les concepts sont précalculés. En effet, les ressources, les dimensions de l'espace, ont tendance à être harmonieusement distribuées en fonction de la richesse lexicale.

En revanche, on perd un avantage important, les composantes des vecteurs n'étant pas facilement interprétables. Ainsi, si on pouvait dans l'apprentissage précédent vérifier l'importance de la composante correspondant à OISEAU pour ‹‹mouette›› ou ‹‹chouette››, on ne le peut plus ici et l'utilisation de la similarité devient logiquement l'outil principal de recherche de bugs dans l'apprentissage. Les travaux sur l'interprétabilité comme ceux menés dans l'équipe d'Isabelle Augenstein (Atanasova *et al.*, 2020), arrivés 15 ans plus tard, nous auraient probablement bien aidés.

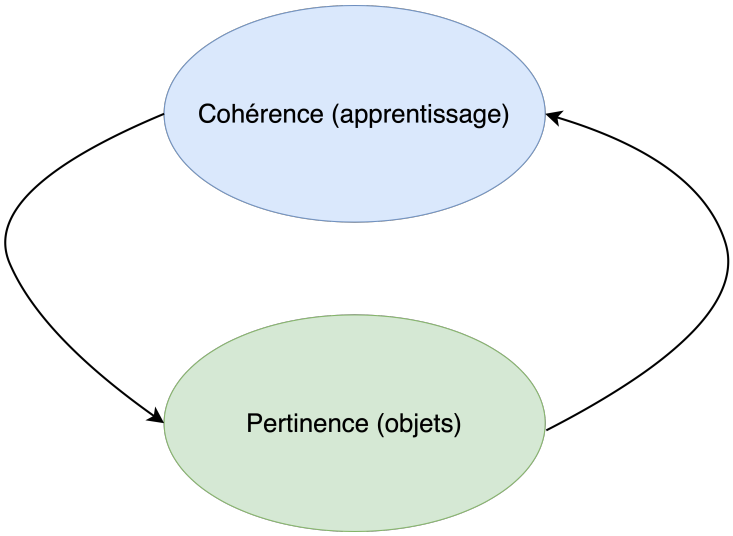

Figure 4.6 – L'idée fondatrice des vecteurs par émergence est que la pertinence des objets est assurée uniquement par la cohérence de l'apprentissage

Cette approche entraîne naturellement une agglomération des vecteurs. Il est donc nécessaire d'augmenter le contraste d'un vecteur à la suite de son calcul. Pour ce faire, on calcule le coefficient de variation* de V. Si ce dernier ne se situe pas à 10% du coefficient de variation* moyen alors le vecteur subit une opération non linéaire d'amplification (la mise à une puissance p de chaque composante puis normalisation), et ce de façon itérée jusqu'à l'obtention d'un coefficient de variation dans la fourchette acceptable. Cette dernière a été estimée à partir des valeurs obtenues dans les expériences avec concepts prédéfinis.

4.3.7 Calcul des vecteurs dans un réseau lexical

Comme je l'ai évoqué à plusieurs reprises, les informations issues des vecteurs et celles issues des réseaux, autrement dit, celles issues des textes (implicites) et celles expertes (explicites) sont indispensables pour réaliser des analyses sémantiques de bonne qualité. Une méthode pour fusionner les deux type d'information est de les représenter sous le même type de représentation mathématique. Nous avons ainsi commencé à donner des vecteurs à des nœuds de réseaux. Il s'agit ainsi ici d'exploiter la complémentarité de la nature des informations en renonçant à la précision des réseaux.

Ces travaux autour de ce qui sera appelé plus tard, des plongements de graphes (*graph embeddings*) que nous retrouverions plus tard (Goyal et Ferrara, 2018; Xu, 2020) en particulier avec les travaux de Jibril Frej (voir section B.2).

La construction d'un vecteur conceptuel est effectuée pour chaque nœud du réseau par simple somme pondérée des vecteurs des nœuds reliés. Soit un nœud N relié a k nœuds $N_1 \dots N_k$, le vecteur de N, $V(N)$ sera égal à $p_1V(N_1)+p_2V(N_2)+ \dots + p_kV(N_k)$ où p_i est le poids du i-ème nœud. Le vecteur somme est ensuite normalisé.

Un dernier problème potentiel est que les vecteurs de deux ensembles distincts (à la fois au sens du réseau lexical et de la thématique) de termes peuvent occuper la même région de l'espace. L'approche du calcul se faisant par activation et les vecteurs étant tirés au hasard à l'initialisation rien n'empêche que cela se produise par accident. Il est donc nécessaire de "séparer" les vecteurs proches mais correspondant pourtant à des parties très différentes du réseau lexical et de la thématique.

La détection de ce phénomène se fait par scrutation du voisinage d'un vecteur conceptuel. Si parmi ses n premiers voisins, la densité de mots n'ayant rien à voir avec le mot étudié est importante alors une action de séparation doit être entreprise.

Cette action de séparation consiste à plonger l'ensemble du réseau dans un champs où les nœuds ont tendance à se repousser. En s'inspirant directement de la physique, une force de répulsion en $1/d^2$ est calculée itérativement entre les nœuds. Pour un nœud donné, on peut ainsi calculer un vecteur déplacement qui va l'éloigner des nœuds dont il se trouve trop près. Les nœuds ne se rapprochant pas par voisinage thématique (lors de la première phase du calcul) mais se trouvant proches "par accident" finissent ainsi naturellement par se séparer.

La dernière expérience a été réalisée avec Gilles Sérasset (UJF [22], LIG, GE-

22. L'Université Joseph Fourrié était l'une des quatre universités grenobloises avant les fusions

	Définitions	Réseaux lexicaux
Concepts prédéfinis	Thésaurus Larousse + diverses sources (Lafourcade et Sandford, 1999; Schwab, 2005)	WordNet + Sumo (Lim et Schwab, 2008)
Émergence	WordNet *Penang-Grenoble* DBNary *Grenoble*	JeuxDeMots (Lafourcade, 2011) WordNet *Penang-Grenoble*

Table 4.1 – Récapitulatif des principales expériences sur les vecteurs conceptuels (2001-2012), à Montpellier (2001-2005), Penang (2006-2007) et Grenoble (2008-2012).

TALP) et Alexandre Labadié, post-doctorant dans le cadre du projet VideoSense. Il faut noter ici la mise en place de Maven* et des tests pour aider au développement, faciliter le développement et assurer sa qualité. Il s'agit d'un élément préfigurateur de ce que nous allions petit à petit faire dans l'équipe. En revanche, cette expérience a été la première qui n'a pas fonctionné pour les vecteurs, ils convergeaient systématiquement vers un point fixe de l'espace. Les circonstances [23] n'ont pas permis d'identifier le problème.

En parallèle, les vecteurs distributionnels commençaient à devenir plus simples à utiliser et sont arrivés sur le devant de la scène et j'ai commencé à bien plus m'y intéresser.

4.4 Les vecteurs distributionnels

Les vecteurs distributionnels constituent l'immense majorité des travaux sur le traitement des langues. Il sont basés sur la linguistique distributionnelle (Harris *et al.*, 1989) qui postule que la sémantique des objets linguistiques peut être décrite en fonction du pouvoir d'associativité qu'ils possèdent ou ne possèdent pas entre eux [24]. La linguistique distributionnelle considère que le sens d'un mot peut être défini à partir de l'ensemble des contextes dans lequel il apparaît, en d'autres termes, par l'ensemble des termes qui lui sont cooccurrents dans un corpus. Comme l'affirme Firth (1957), « *You shall know a word by the company it keeps* » [25] (Léon, 2002; Duvivier-Senis, 2016).

de 2016 et 2020 qui ont donné l'Université Grenoble Alpes.

23. Accroissement de la charge de travail des permanents, besoin d'accorder plus de temps à la vie familiale, fin du contrat d'Alexandre Labadié, fin du projet ANR Videosense.

24. Cette section est adaptée de (Schwab, 2005) et du cours sur les vecteurs que je donne dans diverses formations depuis 2014

25. « *Vous connaîtrez un mot grâce à ceux qui l'accompagnent* ».

Ainsi, selon la linguistique distributionnelle, la sémantique de l'item ‹*lait*› peut ainsi être décrite grâce aux termes ‹*vache*›, ‹*bouteille*›, ‹*fromage*›, ‹*yaourt*›, ‹*allergie*›, ‹*chat*›, ‹*chaton*›, …De même, celle d'‹*ordinateur*› peut l'être grâce aux termes ‹*école*›, ‹*électronique*›, ‹*machine*›, ‹*programmation*›.

On pourra dire que deux mots ont un sens proche s'ils sont employés dans des contextes très voisins. Ce sont ces idées qui ont permis la mise au point des vecteurs saltoniens et de leurs héritiers.

4.4.1 Construction des vecteurs distributionnels

Les vecteurs distributionnels sont donc construits à partir de grands corpus* de texte. Chaque composante correspond à des mots [26] :

— directement : avec le Modèle Vectoriel Standard (Salton, 1968; Salton et McGill, 1983; Salton, 1991, 1971) ;

— indirectement : tous les autres modèles dont les principaux avant 2013, les modèles de type Analyse Sémantique Latente [27] (Deerwester *et al.*, 1990; Landauer et Dumais, 1997).

Soit un corpus textuel contenant n mots* uniques. Avec le modèle saltonien, la taille des vecteurs est de n et le modèle le plus simple est rempli de zéro sauf les composantes correspondant à des mots du cotexte*, où la composante comporte un 1. Plusieurs variantes sont possibles comme, par exemple, le nombre de fois où le mot apparaît dans le cotexte, c'est-à-dire, la fréquence du mot, normalisée avec la fréquence inverse en documents ou non).

Ces modèles souffrent de plusieurs problèmes difficiles techniquement à gérer en particulier à l'époque :

— n la taille de ces vecteurs est importante (*a minima* plusieurs dizaines de milliers) ;

— ils sont particulièrement creux, c'est-à-dire qu'ils contiennent beaucoup de valeurs nulles ;

— la taille des bases est donc importante et/ou les calculs sont particulièrement lourds.

C'est pour ces raisons que les méthodes suivantes se sont attachées à réduire le nombre de dimensions à quelques centaines généralement par des techniques de

26. Ou des sous-mots depuis (Bojanowski *et al.*, 2017).
27. LSA : *Latent Semantic Analysis*.

réduction de matrice par valeurs singulières pour LSA (*Latent Semantic Analysis*) (Deerwester *et al.*, 1990; Landauer et Dumais, 1997), par méthodes probabilistes pour PLSA (*Probabilistic Latent Semantic Analysis*) (Hofmann, 1999). Enfin, on trouve essentiellement depuis 2013 et *Word2Vec*, des méthodes neuronales.

4.4.2 Approches neuronales

Ces approches utilisent des méthodes d'apprentissage qui sont appelées, depuis quelques années, les méthodes autosupervisées . L'idée est (1) de dégrader les données originelles ou (2) d'exploiter une partie des données pour en deviner une autre. On constitue ainsi des paires dont on peut se servir avec des algorithmes supervisés.

Dans le premier cas, on constitue des paires (données originelles - données dégradées), on met alors en entrée les données dégradées, et on essaye de retrouver les données originelles en sortie. Par exemple :

— en masquant un ou plusieurs mots dans une phrase et en cherchant à les retrouver ;
— en prédisant la phrase suivante ;
— en supprimant des espaces et en cherchant à retrouver le segment textuel initial ;
— en supprimant des caractères et en cherchant à retrouver le segment textuel initial ;
— en modifiant (suppression, ajout, échange) aléatoirement des mots et en cherchant à retrouver le segment textuel.

Dans le second cas, Une autre possibilité est de constituer des paires exploitant le corpus originel. Il s'agit alors de :

— prédire le mot w précédant une séquence de mots ;
— prédire le mot w suivant une séquence de mots ;
— prédire la phrase Y précédant la phrase X ;
— prédire la phrase Y suivant la phrase X ;

En 2013, (Mikolov *et al.*, 2013) travaillant chez Google ont publié des modèles préentraînés pour la langue anglaise appris grâce à un réseau de neurone. À l'époque, on est loin d'un réseau profond et, ici, le réseau n'a que 3 couches (voir figure 4.7). Le principe est simple, le réseau va apprendre comment représenter au mieux un mot* grâce à son cotexte* ou en devinant son cotexte. Les auteurs proposent deux alternatives :

— *CBOW* : à partir du cotexte, il s'agit de deviner le mot initial ;

— *Skip-gram* : à partir du mot initial, il s'agit de deviner le cotexte.

Une fois appris pour un mot grâce à l'ensemble du corpus, le vecteur correspond à la couche intermédiaire (notée *Projection* dans figure 4.7). Ce modèle a été entraîné sur environ 100 milliards de mots issus du corpus de nouvelles de Google [28]. La taille du vocabulaire est d'environ 3 millions de mots et les vecteurs ont une dimension de 300. Il y a un seul vecteur par mot et donc l'ensemble des sens est fusionné.

Notons que pour le modèle standard de *Word2Vec*, les mots ont des représentations et pas les items lexicaux et des acceptions comme dans le cas des vecteurs conceptuels. Nous verrons que nous avons étudié une méthode permettant de pallier ce problème dans la suite de ce chapitre (voir section 4.4.3.1).

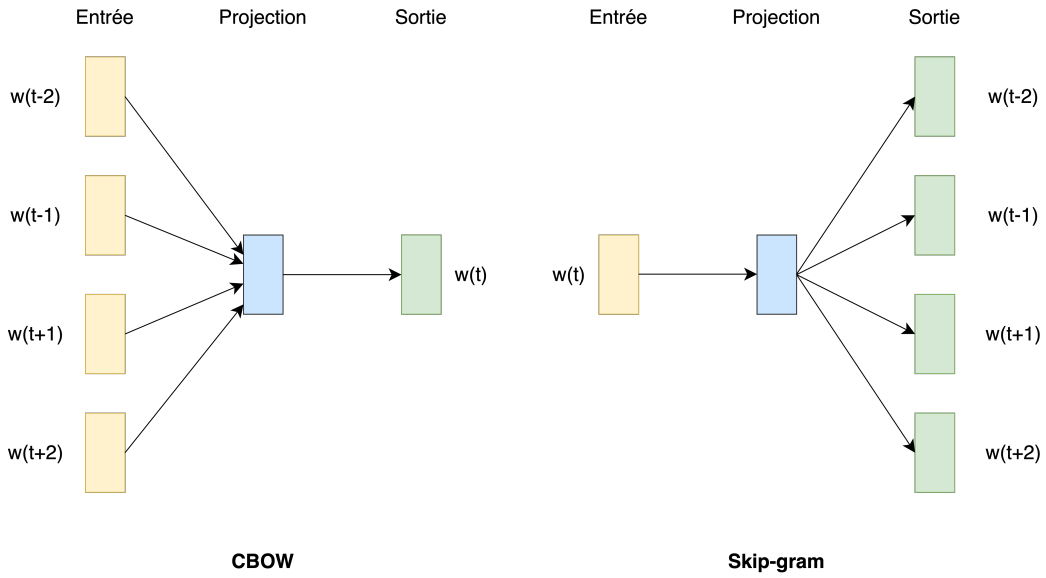

Figure 4.7 – Les deux architectures du modèle Word2Vec : ces modèles sont entraînés sur environ 100 milliards de mots issus du corpus de nouvelles de Google. La taille du vocabulaire est d'environ de 3 millions de mots et les vecteurs ont une dimension de 300.

Longtemps, le modèle CBOW a été considéré comme le plus rapide à apprendre et le modèle skip-gram comme celui qui donnait de meilleurs résultats. Toutefois, à la fin 2020, İrsoy *et al.* (2020) [29] semblent apporter un démenti à cette

28. Corpus non librement disponible.

29. Plus de 8 mois plus tard, cet article n'a été cité qu'une fois et est resté sur ArXiv, preuve s'il en était que la recherche est, pour le moment en tous cas, passée à autre chose.

affirmation en résolvant un bug dans le modèle CBOW découvert dans la bibliothèque initiale en C [30] ou dans une implantation classique en Python GENSIM [31]. Cette nouvelle implantation donne des résultats similaires pour les deux variantes.

Word2Vec marque un essort réel des représentations vectorielles en TALN et l'article est cité quelques centaines de fois moins de deux ans après sa publication ce qui est important pour l'époque. Leur création est vraiment simplifiée et la publicité autour des modèles a fait la suite. La force de frappe d'une société comme Google n'y est évidemment pas pour rien. Il faut dire qu'avec des exemples bien choisis, le modèle permet de s'attaquer à des problèmes comme l'analogie.

$$\text{V('king')} - \text{V('man')} + \text{V('woman')} \approx \text{V('queen')}$$

$$\text{V('woman')}\ \ \text{V('man')} \approx \text{V('aunt')}\ \ \text{V('uncle')}$$

$$\text{V('Rome')} - \text{V('Italy')} \approx \text{V('France')} - \text{V('Paris')}$$

$$\text{V('Iraq')} - \text{V('Violence')} \approx \text{V('Jordan')}$$

Ces résultats sont depuis largement contestés (Linzen, 2016). Il n'est, en effet, pas très difficile de trouver des exemples qui fonctionnent moins bien mais comme souvent avec ces méthodes, c'est le nombre qui fait que globalement les résultats sont satisfaisant dans les applications.

Plusieurs variantes ont été proposées et l'on peut citer, par exemple :

— Levy et Goldberg (2014) qui utilisent comme cotexte les dépendances syntaxiques au lieu du simple voisinage fenêtré de *Word2Vec* ;

— Glove (Pennington *et al.*, 2014) allie méthodes de comptages (factorisation de matrices) et méthodes neuronales ;

— *FastText* (Bojanowski *et al.*, 2017) qui propose de pallier deux problèmes des vecteurs attribués au niveau des mots [32] :

1. Aucun vecteur n'existe pour les mots n'ayant pas été rencontrés dans le corpus d'apprentissage.
2. Les vecteurs ne prennent pas en compte la structure interne des mots. Ainsi les mots «jardin»,«jardiner» et «jardinier» sont trois mots de la même famille et, pourtant, ils sont appris indépendamment et leur racine commune n'entre jamais en compte dans la construction des vecteurs. *FastText* n'apprend ainsi plus les vecteurs au niveau du mot mais

30. https://github.com/tmikolov/word2vec
31. https://github.com/RaRe-Technologies/gensim
32. Considéré comme comme la chaîne de caractères entre deux espaces en français ou en anglais.

au niveau du sous-mot. Les plus fréquents n-grammes de caractères* dans le corpus d'apprentissage sont utilisés pour entraîner un modèle *Word2Vec* de type *skip-gram*. En exploitation, les vecteurs des mots sont fabriqués en combinant les vecteurs des sous-mots qui le composent. Dans le cas de *FastText*, il s'agit d'une somme vectorielle. Cette méthode de représentation au niveau du sous-mot et combinaison des représentations est très efficace et sera reprise par la suite en particulier dans les méthodes dites contextualisées que nous verrons plus loin.

4.4.3 Contributions aux approches vectorielles distributionnelles

Au cours des travaux auxquels j'ai participé, nous nous sommes attaqués à trois questions de recherche importantes que posaient les approches vectorielles distributionnelles :

— Comment calculer le vecteur d'un objet textuel (segment textuel, phrase, texte) ?

— Comment affecter un vecteur à un item lexical ?

— Comment affecter un vecteur à une lexie ?

Nous avons fourni des réponses inspirées des travaux sur les vecteurs conceptuels. Commençons par la première question.

4.4.3.1 Calcul du vecteur d'un objet textuel

Généralement, le vecteur d'un objet textuel est simplement la somme des vecteurs des mots qui le composent, soit :

$$V = \sum_{i=1\,,\,w_i \in S}^{n} (vector(w_i)) \tag{4.7}$$

Pourtant, des travaux, comme les nôtres sur les vecteurs conceptuels ou ceux de Ji et Eisenstein (2013) démontrent que l'intuition que les mots n'ont pas tous la même importance pour détecter paraphrases ou estimer la similarité est correcte. De fait, d'autres recherches utilisent directement des annotations morpho-syntaxiques (Vulić, 2017; Dehouck et Denis, 2017).

Dans (Ferrero *et al.*, 2017b; Billah Nagoudi *et al.*, 2017b), nous avons proposé d'utiliser deux pondérations dans l'Equation 4.7. Une première basée sur la partie

du discours et une seconde basée sur la fréquence inverse en documents* inspirée de (Brychcín et Svoboda, 2016). Il s'agit ainsi d'affiner l'Equation 4.6 utilisée dans l'Equation 4.5, soit :

$$\varphi(w) = weight(pos)^{(1-\alpha)} \times idf(w)^{\alpha} \tag{4.8}$$

où pos est la fonction qui retourne l'étiquette morphosyntaxique (partie du discours) d'un mot, $weight$ est la fonction qui retourne la pondération attribuée à une certaine étiquette morphosyntaxique, idf est la fonction qui retourne la fréquence inverse de document d'un mot et le paramètre α est un moyen de contrôler la contribution morphosyntaxique et fréquentielle dans la formule de pondération. Ainsi, avec $\alpha = 0$, nous retrouvons l'Equation 4.6 [33].

Une différence dans la pondération attribuée aux étiquettes est également apportée. Le choix des poids n'est plus estimé par les chercheurs comme dans le cas des vecteurs conceptuels mais appris sur un corpus. Ainsi, alors que la pondération $idf(w)$ peut être estimée de manière non-supervisée, $weight(pos)$ est estimé de manière supervisée.

Notre objectif est donc de rechercher la meilleure combinaison de poids possible pour obtenir les meilleurs résultats au cours d'une évaluation de détection de similarité sémantique textuelle. Afin d'optimiser au mieux ces variables (soit 12 étiquettes et α), nous avons utilisé l'outil Condor (Berghen et Bersini, 2005).

L'optimisation a été réalisée sur les corpus décrits en 3.3.2. La tableau 4.2 présente les pondération obtenues après normalisation avec leur fréquence d'apparition.

La première remarque que nous pouvons faire est que nous constatons que les noms, verbes, adjectifs, adverbes, numéraux et mots étrangers possèdent des poids plus importants que les mots vides*, comme les prépositions ou les déterminants. Cela peut s'expliquer par le fait que ces mots peuvent être trouvés en grand nombre et facilement dans toutes phrases, et ne peuvent donc pas être considérés comme importants au sein d'une phrase spécifique. D'ailleurs, en Traitement Automatique du Langage Naturel, de nombreuses techniques filtrent ces mots vides en premier lieu avant de passer à une analyse plus profonde. Les adjectifs et les adverbes sont plus importants que les noms et les verbes.

Cette méthode a été appliquée avec succès lors de notre participation à SemEval 2017 [34] où nous avons terminé premiers sur 20 participants et 53 systèmes

33. À ceci près que nous ne bénéficions plus d'un outil de la qualité de SygFran, nous n'avons plus accès qu'à la partie du discours au lieu du rôle syntaxique.

34. *Track 4a – Semantic Textual Similarity* ; épreuve 4a – similarité sémantique de textes.

Étiquettes morphosyntaxiques	corpus syntagmatique	corpus phrastique
VERB - verbes (tous les temps et conjugaisons)	2,00	0,98
NOUN - noms (communs et propres)	0,46	0,79
PRON - pronoms	0,82	0,20
ADJ - adjectifs	1,6	1,97
ADV - adverbes	3,34	2,77
ADP - adpositions (prépositions and postpositions)	0,06	0,08
CONJ - conjonctions	1,00	0,26
DET - déterminants	0,1	0,13
NUM - nombres cardinaux	7,37	6,34
PRT - particules et mots outils	0,00	0,00
X - mots étrangers ou inconnus	26,36	23,62
. - ponctuation	0,67	0,18

TABLE 4.2 – Poids attribués (en %) aux différentes étiquettes morphosyntaxiques après optimisation en normalisant en fonction de la probabilité d'apparition des étiquettes morphosyntaxiques – issu de (Ferrero, 2017).

soumis (Ferrero *et al.*, 2017b; Ferrero, 2017).

Dans (Billah Nagoudi *et al.*, 2017a), nous proposons une variante avec une version de la formule du poids (voir Equation 4.6). Le principe consiste à étiqueter morpho-syntaxiquement le corpus en utilisant un analyseur dédié [35] puis de calculer la fréquence inverse en documents pour chacune des parties du discours.

Dans cette expérience sur l'arabe, nous utilisons l'analyseur de (Gahbiche-Braham *et al.*, 2012) et une collection de 750 paires de phrases tirées du corpus de Microsoft de Recherche et de Description Vidéo (MSR-Video) [36] traduites manuellement vers l'arabe. Ensuite, les phrases ont été manuellement annotées avec un réel compris entre 0 et 1. Le score 0 indique que le sens des deux phrases est totalement différent, et les score 1 que les deux phrases ont exactement le même sens. La table 4.3 présente la corrélation entre le jugement humain et les similarités cosinus entre les vecteurs obtenus selon les différentes méthodes. Nous pouvons constater que la pondération unitaire (Equation 4.7) est nettement moins performante que les autres. La pondération par fréquence inverse en documents et celle basée sur les parties du discours obtiennent des résultats comparables (respectivement +5,87% et +7,18% par rapport à la pondération unitaire), même si la seconde est légèrement meilleure. Enfin, la combinaison de ces deux pondérations donne

35. Nous qualifions notre méthode de non-supervisée dans (Billah Nagoudi *et al.*, 2017a) mais les analyses morpho-syntaxiques sont souvent supervisées donc ce terme est probablement abusif. En revanche, elle demande moins de supervision en se passant d'une tâche annexe comme notre méthode initiale.

36. https://www.microsoft.com/en-us/download/details.aspx?id=52422

largement les meilleurs résultats (+10,43% par rapport à la pondération unitaire).

Méthode	Corrélation
Sans pondération (Equation 4.7)	72,33 %
Pondération avec IDF ($\alpha = 1$)	78,20%
Pondération avec les parties du discours ($\alpha = 0$)	79,51%
Pondération mixte ($\alpha = 0,5$)	82,76%

TABLE 4.3 – Résultats de corrélation sur le corpus MSR-Video (issu de (Billah Nagoudi et al., 2017a) qui montre des résultats similaires sur 3 autres corpus.)

Une autre exploitation de cette méthode est une réponse à la deuxième question posée ci-dessus, comment affecter un vecteur aux sens des items lexicaux ?

4.4.3.2 Calcul de la représentation vectorielle de LEXIE

Nous l'avons vu, un des problèmes des représentations vectorielles distribuées est de ne pas avoir une représentation par sens mais une même représentation pour l'ensemble des sens. En reprenant le vocabulaire introduit dans la partie sur les vecteurs conceptuels, nous nous posons la question : comment affecter un vecteur non plus à un mot mais à une LEXIE [37] ?

Dans (Vial *et al.*, 2017a) et (Vial *et al.*, 2017b), nous présentons une méthode inspirée de l'Equation 4.6 que nous appliquions aux définitions du *Princeton WordNet* pour calculer un vecteur pour chacune de ses définitions. Dans le même mouvement, cette méthode permet également d'affecter un vecteur au niveau des items lexicaux.

Nous avons ici utilisés plusieurs vecteurs préentraînés :

— *Word2Vec* (Mikolov *et al.*, 2013) entraîné sur environ 100 milliards de mots issus du corpus de nouvelles de Google. La taille du vocabulaire est d'environ de 3 millions de mots et les vecteurs ont une dimension de 300 ;

— GloVe, le modèle de Pennington *et al.* (2014) entraîné sur 42 milliards de mots issus de *Common Crawl*. Le vocabulaire est de 2 millions de mots et les vecteurs ont une taille de 300 ;

— Le modèle de Levy et Goldberg (2014) [38], noté *deps* dans la suite. L'apprentissage a été effectué sur Wikipedia, le vocabulaire est de 175 000 mots et la

37. Et non, à une ACCEPTION puisque, nous allons le voir, notre méthode utilise une certaine ressource (*Princeton WordNet* en l'occurence) alors que l'introduction de bien d'autres ressources aurait été possible. La fusion de ces LEXIES aurait alors donné des ACCEPTIONS.

38. https://levyomer.wordpress.com/2014/04/25/dependency-based-word-embeddings/

taille des vecteurs 300.

— Le meilleur des modèles de prédiction de Baroni *et al.* (2014) [39], noté *baroni_p* dans la suite. La taille du vocabulaire est de 300 000 et la taille des vecteurs est de 400 ;

— Et finalement, le meilleur modèle à base de comptage également créé par Baroni *et al.* (2014)[5], noté *baroni_c* dans la suite. La taille de vocabulaire est également de 300 000 et la taille des vecteurs est de 500.

Pour valider notre approche, nous avons intégré les vecteurs obtenus dans une tâche de désambiguïsation lexicale (Vial *et al.*, 2017a,b). Les définitions sont augmentées des mots les plus proches au sens du voisinage thématique (voir section 4.3.3). La méthode d'évaluation utilise comme algorithme local*, l'algorithme de Lesk* qui donne un score correspondant au nombre de mots communs entre définitions et comme algorithme global*, un algorithme de type coucou* (voir section 5.6.4.2).

Système	SemEval 2007 F1 score	SemEval 2015 F1 score
Lesk	68.70%	50.65%
Lesk étendu	78.01%	61.42%
VecLesk (baroni_c)	**75.29**%	58.02%
VecLesk (baroni_p)	73.52%	53.46%
VecLesk (deps)	73.02%	56.40%
VecLesk (GloVe)	73.00%	**59.01**%
VecLesk (*Word2Vec*)	73.30%	57.00%

TABLE 4.4 – Comparaison de nos résultats sur SemEval 2007 et SemEval 2015 pour chacun des modèles de vecteurs de mots par rapport aux étalons Lesk et Lesk étendu (issu de (Vial et al.*, 2017a)).*

Les résultats présentés dans tableau 5.6 montrent que, quels que soit les vecteurs utilisés, notre méthode améliore systématiquement la désambiguïsation lexicale. Nous montrons ainsi qu'il est possible d'affecter des vecteurs au niveau des sens et d'exploiter les opérations permises par les vecteurs pour enrichir les définitions.

Nous pouvons également noter la différence de score obtenue entre les différents modèles de vecteurs de mots. Les meilleurs résultats sur SemEval 2007 utilisent la méthode de Baroni *et al.* (2014) basée sur le comptage. Ceci tend à

39. http://clic.cimec.unitn.it/composes/semantic-vectors.html

montrer que les méthodes de création de vecteurs de mots par comptage sont parfois meilleures que les modèles neuronaux. Cependant, sur SemEval 2015, c'est la méthode GloVe qui obtient les meilleurs résultats. Ces fluctuations dans les résultats sont peut être dues aux différentes natures des méthodes de création des vecteurs de mots, ou bien aussi dues aux différents corpus sur lesquels ils ont été entraînés.

À notre connaissance, il n'existe pas d'étude qui se sont attaqués à ces questions, ces vecteurs étant aujourd'hui bien moins utilisées en tous cas, dans un tel contexte, depuis la vague des approches contextuelles à partir de 2018.

4.5 Les approches neuronales contextualisées : le projet FlauBERT

En 2018 [40], l'introduction de représentations linguistiques profondes contextuelles, obtenues à partir de textes bruts, a conduit à un changement de paradigme pour plusieurs tâches du TALN. Alors que les approches fondées sur des représentations telles que celles que nous venons de présenter dans la section précédente apprennent nativement un vecteur unique pour chaque mot, les approches neuronales contextualisées produisent des représentations dîtes contextuelles puisqu'elles dépendent de la séquence de mots d'entrée complète.

Initialement fondées sur des réseaux neuronaux récurrents (Dai et Le, 2015; Ramachandran *et al.*, 2017; Howard et Ruder, 2018; Peters *et al.*, 2018b), ces approches ont rapidement intégré des modèles *transformeur** (Vaswani *et al.*, 2017b). Dès sa publication en 2018, BERT (Devlin *et al.*, 2019) a permis une avancée de l'état-de-l'art pour de nombreuses tâches du TALN. Il s'agit d'un réseau de neurones profond constitué de plusieurs couches intégrant des têtes d'attentions. Les représentations se font comme dans *FastText* au niveau des sous-mots. Dans l'article original, deux modèles sont présentés :

— $\text{BERT}_{\text{BASE}}$ qui a 12 couches pour des vecteurs de taille 768, 12 têtes d'attention soit un total de 110 millions de paramètres ;

— $\text{BERT}_{\text{LARGE}}$ qui a 24 couches pour des vecteurs de taille 1024, 16 têtes d'attention soit un total de 340 millions de paramètres.

L'apprentissage du modèle se fait par autosupervision en apprenant :

40. Cette section est largement inspirée de (Le *et al.*, 2020e) lui-même résumé de (Le *et al.*, 2020f), de mes cours de master sur le sujet et de diverses présentations que j'ai faites sur le sujet.

— à prédire des sous-mots masqués de façon aléatoire puis à les deviner (objectif MLM [41]) ;

— à prédire si `B` est une phrase qui suit effectivement `A`, étant donné une paire de phrases d'entrée `A, B` (objectif NSP [42]).

La figure 4.8 présente une vue simplifiée du modèle $\text{BERT}_{\text{BASE}}$.

Les avancées dans l'état de l'art de nombreuses tâches ont rapidement conduit à la publication d'autres modèles comme, par exemple, XLNet (Yang *et al.*, 2019), RoBERTa (Liu *et al.*, 2019), ALBERT (Lan *et al.*, 2019) ou T5 (Raffel *et al.*, 2019) pour n'en citer que certains disponibles au moment de nos travaux initiaux [43]. Cependant, ceci est essentiellement montré pour l'anglais, même si des variantes multilingues prenant en compte plus d'une centaine de langues commencent à apparaître comme mBERT (Devlin *et al.*, 2019), LASER (Artetxe et Schwenk, 2019), XLM (Lample et , 2019) ou XLM-R (Conneau *et al.*, 2019). L'utilisation des sous-mots au lieu des mots comme vocabulaire facilite la création de tels modèles en limitant la taille du vocabulaire nécessaire.

À l'époque de nos travaux, plusieurs auteurs ont publié des modèles monolingues disponibles pour d'autres langues que l'anglais. BERT a été entraîné pour plusieurs langues (allemand, chinois, espagnol, finnois, italien, néerlandais, suédois). Pour le français, à la même époque que nous, une équipe jointe INRIA et Facebook a développé CamemBERT (Martin *et al.*, 2019).

Dans (Le *et al.*, 2020f), nous décrivons notre méthodologie pour construire et partager FlauBERT (**F**rench **La**nguage **U**nderstanding via **B**idirectional **E**ncoder **R**epresentations from **T**ransformers), un modèle BERT pour le français.

4.5.1 Apprentissage du modèle FlauBERT

Nous avons déjà présenté en section 3.2.1.4, le corpus sur lequel est appris FlauBERT. Rappelons qu'il s'agit de 71 Go de données normalisées (270 Go de données brutes) issues de 24 sous-corpus de types divers (Wikipedia, livres, *Common Crawl*...).

41. MLM pour *Masked Language Model* – Modèle de langue masqué.
42. NSP pour *Next sentence prediction* – prédiction de la prochaine phrase.
43. De juin à décembre 2019.

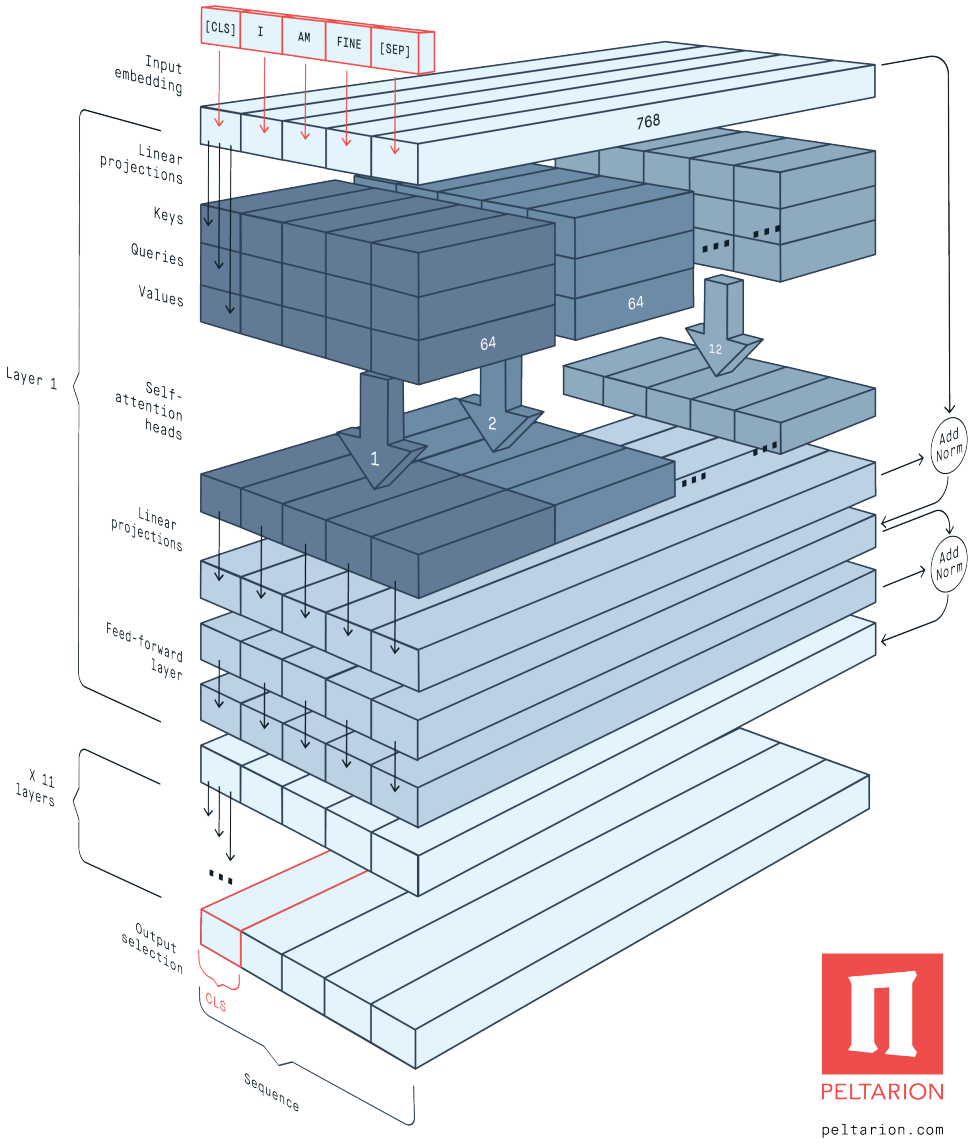

FIGURE 4.8 – Vue simplifiée de l'architecture de $BERT_{BASE}$. La figure ne présente qu'une des 12 couches identiques. Les vecteurs sont de taille 768, il y a 12 têtes d'attention soit un total de 110 millions de paramètres – image issue de https://peltarion.com/knowledge-center/documentation/modeling-view/build-an-ai-model/blocks/english-bert.

4.5.1.1 Objectif de l'entraînement et optimisation

Comme nous l'avons vu, le préentraînement du Bert original consiste en deux tâches autosupervisées : (1) un *modèle de langue masqué* (MLM) et (2) une *pré-*

diction de la prochaine phrase (NSP).

? observe que la suppression de *NSP* nuit considérablement aux performances sur certaines tâches. Cependant, le contraire a été démontré dans des études ultérieures, notamment (Yang *et al.*, 2019), (Lample et , 2019), et (Liu *et al.*, 2019). (Liu *et al.*, 2019) ont même émis l'hypothèse que l'implantation originale de BERT pourrait avoir supprimé la fonction de coût associée au *NSP* tout en conservant le format d'entrée consistant en des paires de phrases. Par conséquent, nous avons utilisé seulement l'objectif *MLM* dans FlauBERT.

Pour optimiser cette fonction objectif, nous avons suivi (Liu *et al.*, 2019) et utilisé l'optimiseur Adam (Kingma et Ba, 2014) avec les paramètres suivants :

— FlauBERT$_{\text{BASE}}$: étapes de mise en route* (ou *warm up*) de 24k, taux d'apprentissage maximal de 6e-4, $\beta_1 = 0,9$, $\beta_2 = 0,98$, $\epsilon = 1e-6$ et perte de poids de 0,01.

— FlauBERT$_{\text{LARGE}}$: étapes de mise en route de 30k, taux d'apprentissage maximal de 3e-4, $\beta_1 = 0,9$, $\beta_2 = 0,98$, $\epsilon = 1e-6$ et perte de poids de 0,01.

4.5.1.2 Modèles et configuration d'apprentissage

Nous utilisons la même architecture que BERT (**?**). Un vocabulaire de 50K unités sous-lexicales est construit en utilisant l'algorithme *Byte Pair Encoding* (Sennrich *et al.*, 2016, BPE). Nous entraînons deux principaux modèles (transformeurs bi-directionnels) : FlauBERT$_{\text{BASE}}$ (12 couches, des vecteurs de dimension 768, 12 têtes d'attention) et FlauBERT$_{\text{LARGE}}$ (24 couches, des vecteurs de dimension 1024, 12 têtes d'attention).

Le critère d'apprentissage est de type *masked language model* : il consiste à prédire des tokens d'une phrase ayant été préalablement et aléatoirement masqués. FlauBERT$_{\text{BASE}}$ est appris sur 32 GPU Nvidia V100 SXM2 32 GB en 410h et FlauBERT$_{\text{LARGE}}$ est appris sur 128 de ces mêmes GPU en 390h grâce à la grille de calcul Jean Zay du CNRS.

La tableau 4.5 compare quatre modèles prototypiques selon plusieurs axes. On peut voir que les modèles diffèrent suivant la taille et la manière de constituer les sous-mots, l'objectif d'apprentissage, la manière d'obtenir les masques, la taille des modèles et bien entendu les données d'apprentissage. Vu la difficulté d'entraîner de tels modèles, il n'est en effet pas facile de tester l'apport de chacune. Bien souvent les gens se contentent simplement de tester les modèles. Cette limite n'est pas propre aux vecteurs contextualisés, on le devine en pointillés dans les sections précédentes mais elles sont ici exacerbées par la quantité de ressources matérielles nécessaires pour fabriquer de tels modèles. Ainsi, notre objectif initial dans le pro-

jet était de comparer plusieurs variantes. À l'époque, nous nous sommes contentés de travailler sur la taille des modèles.

	$\text{BERT}_{\text{BASE}}$	$\text{RoBERTa}_{\text{BASE}}$	CamemBERT	Flaubertbase/Flaubertlarge
Langue	Anglais	Anglais	Français	Français
Données d'apprentissage	13 GB	160 GB	138 GB†	71 GB‡
Objectifs de préentraînement	NSP et MLM	MLM	MLM	MLM
Nombre total de paramètres	110 M	125 M	110 M	138 M/ 373 M
Tokenisation	WordPiece 30K	BPE 50K	SentencePiece 32K	BPE 50K
Masque	Statique + sous-mots	Dynamique + sous-mots	Dynamique + mot entier	Dynamique + sous-mot

†, ‡ : 270 GB before filtering/cleaning.

Table 4.5 – Comparaison entre FlauBERT et d'autres modèles de langue préentraînés – issu de (Le et al., 2020f).

4.5.2 Évaluation sur FLUE

Les modèles FlauBERT ont été évalués sur le référentiel FLUE (**F**rench **L**anguage **U**nderstanding **E**valuation) que nous avons déjà présenté dans la section 3.3.3.

Nous comparons les performances de FlauBERT avec BERT multilingue (Devlin *et al.*, 2019, mBERT) et CamemBERT (Martin *et al.*, 2019) sur toutes les tâches. Nous comparons également avec le meilleur modèle non contextuel pour chaque tâche.

Sur la plupart des tâches de FLUE, les modèles de FlauBERT obtiennent des résultats similaires à ceux de CamemBERT (Martin *et al.*, 2019). Les deux surpassent les résultats des autres vecteurs preuve, d'un côté, que les approches contextualisées fonctionnent mieux que les autres approches distributionnelles et, de l'autre, que les modèles monolingues semblent plus efficaces que les modèles multilingues à l'échelle de la langue.

Tâche	Classification			Paraphrase	NLI	Constituents		Dependences		Désambiguïsation	
Section	Livres	DVD	Musique							Noms	Verbes
Mesure	Acc.	Acc.	Acc.	Acc.	Acc.	F_1	POS	UAS	LAS	F_1	F_1
État de l'art ant.	91,25[c]	89,55[c]	93,40[c]	66,2[d]	80,1/**85,2**[e]	87,4[a]		89,19[b]	85,86[b]	-	43,0[h]
Sans préapprentissage	-	-	-			83,9	97,5	88,92	85,11	50,0	-
FastText	-	-	-			83,6	97,7	86,32	82,04	49,4	34,9
mBERT	86,15[c]	86,9[c]	86,65[c]	89,3[d]	76,9[f]	87,5	98,1	89,5	85,86	56,5	44,9
CamemBERT	**93,40**	**92,70**	94,15	89,8	81,2	88,4	**98,2**	91,37	88,13	56,1	**51,1**
$\text{FlauBERT}_{\text{BASE}}$	**93,40**	92,50	**94,30**	**89,9**	81,3	**89,1**	98,1	**91,56**	**88,35**	54,9/**57,9**[g]	47,4

Table 4.6 – Résultats finals sur les tâches de FLUE (issu de (Le et al., 2020f)). [a] Kitaev et al. (2019). [b] Constant et al. (2013). [c] Eisenschlos et al. (2019, MultiFiT). [d] Chen et al. (2017, ESIM). [e] Conneau et al. (2019, XLM-F BASE/LARGE). [f] Martin et al. (2019). [g] Utilise $\text{FlauBERT}_{\text{LARGE}}$. [h] Segonne et al. (2019).

4.5.3 FlauBERT aujourd'hui et demain

Notons que deux ans plus tard, le même constat peut-être dressé si l'on en croit la littérature. Ainsi dans une étude non encore publiée et menée par Jeongwoo Kang, doctorante que je dirige avec Maximin Coavoux (GETALP, LIG, CNRS), qui compare FlauBERT et CamemBERT dans les 17 études [44] publiées avant juin 2021 on peut constater des performances similaires pour les deux ensembles de modèles.

La seule différence notable est dans (Martin *et al.*, 2020) [45] sur une tâche de réponse à des questions.

	FQuAD1.1-test		FQuAD1.1-dev	
Model	F1	EM	F1	EM
Human Perf.	91.2	75.9	92.1	78.3
$\text{CamemBERT}_{\text{BASE}}$	88.4	78.4	88.1	78.1
$\text{CamemBERT}_{\text{LARGE}}$	**92.2**	**82.1**	**91.8**	**82.4**
$\text{FlauBERT}_{\text{BASE}}$	77.6	66.5	76.3	65.5
$\text{FlauBERT}_{\text{LARGE}}$	80.5	69.0	79.7	69.3
mBERT	86.0	75.4	86.2	75.5
$\text{XLM-R}_{\text{BASE}}$	85.9	75.3	85.5	74.9
$\text{XLM-R}_{\text{LARGE}}$	89.5	79.0	89.1	78.9

Figure 4.9 – Comparaison de plusieurs systèmes sur des corpus d'évaluation d'une tâche de questions-réponse – tableau issu de (Martin et al.*, 2020)*

Une explication possible est la différence entre les deux modèles est que FlauBERT n'a pas de tête de modèle de langue implanté dans *Hugging Face* contrairement à CamemBERT. Une autre pourrait être l'objectif d'apprentissage qui diffèrent (sous-mots pour FlauBERT et mot entiers pour CamemBERT). La conséquence de ces deux raisons est que FlauBERT se revèle bien incapable dans sa version *Hugging Face* utilisée par la plupart des gens d'exploiter ses fonctions génératrices [46]. Considérons l'exemple suivant, « *Le X aboie.* ». Il s'agit de trouver

44. Toute étude, y compris celles qui n'ont pas encore été validées par des pairs (généralement celles diffusées sur arXiv – https://arxiv.org).

45. Cet article n'a pas été publié dans une conférence mais l'expérience, que nous avons reproduite a donné des résultats similaires.

46. Le lecteur pourra le vérifier lui-même en cliquant sur les deux liens suivants https://huggingface.co/flaubert/flaubert_base_uncased?text=Le+%3Cspecial1%3E+aboie pour FlauBERT et https://huggingface.co/camembert-base?text=Le+%3Cmask%3E+aboie – testé le 2 septembre 2021.

les mots les plus susceptibles de remplacer le mot manquant et marqué ici par un X. Les modèles renvoient les réponses suivantes :

— FlauBERT : mal (0.984) ; , (0.004) ; ainsi (0.002) ; ron (0.001) ; avec (0.001)

— CamemBERT : chien (0.424) ; loup (0.092) ; chat (0.083) ; lion (0.051) ; cheval (0.040)

Les résultats de CamemBERT sont plutôt bons [47] tandis que ceux de FlauBERT sont ici catastrophiques.

Une dernière hypothèse, à laquelle je souscris moins, concerne les documents sur lesquels les deux modèles ont été appris (divers genres et domaines pour FlauBERT et uniquement Web pour CamemBERT).

La suite du projet devrait nous permettre d'en savoir plus et de corriger ces limites.

4.6 Conclusions du chapitre

Dans ce chapitre, nous avons vu comment mes travaux sur les vecteurs sont passés des vecteurs d'idées aux vecteurs d'usage et comment il se sont inscrits de manière plus générale dans la recherche durant ces 20 dernières années.

Le prochain chapitre, qui sera le dernier avant la conclusion, traite de l'application principale à laquelle je me suis intéressé avec ces vecteurs, à savoir la clarification du sens et plus particulièrement la désambiguïsation lexicale.

47. Le chien est communément considéré comme l'animal qui aboie mais c'est aussi le cas du loup. En revanche, ni le chat, ni le lion ni le cheval n'aboient contrairement au chevreuil ou à l'otarie si on en croit Wikipedia – https://fr.wikipedia.org/wiki/Aboiement consulté le 2 septembre 2021

Chapitre 5

Contributions aux modèles : l'exemple de la désambiguïsation lexicale

Sommaire

La clarification du sens qui inclut la désambiguïsation lexicale [1] est une tâche centrale à plusieurs applications du TALN comme, par exemple, la traduction automatique ou la recherche d'information. L'équipe GETALP se concentre sur la désambiguïsation lexicale multilingue. Schématiquement, il s'agit de trouver, quelle que soit la langue, quel sens particulier est utilisé pour chacun des mots d'un texte parmi un inventaire de sens prédéfini. Par exemple, dans la phrase « *la souris mange le fromage* », il faudra préférer le sens d'animal plutôt que le sens de dispositif électronique. Dans ses recherches, le GETALP met un accent particulier sur l'enrichissement et l'exploitation des ressources multilingues et sur l'accès multilingue avec un sens garanti.

Nous étudions comment il est possible de clarifier automatiquement un texte en fonction des ressources disponibles pour une langue donnée. Dans ce cadre, les ressources les plus importantes sont les bases lexicales et les corpus annotés en sens. Avant 2016, les corpus en anglais annotés manuellement se présentaient sous des formats hétérogènes et avec différentes versions de bases lexicales. Pour résoudre ce problème, notre équipe a unifié l'ensemble des corpus annotés en sens dans UFSAC pour environ 2 000 000 de mots annotés – voir (Vial *et al.*, 2018b) et section 3.2.2.2.

1. Cette section a initialement été rédigée dans le cadre d'un projet jeune chercheur, jeune chercheuse de l'Agence Nationale pour la Recherche que j'avais déposé en 2013, puis pour les cours donnés en master, réutilisés pour des articles soumis à la conférence TALN ou la revue TAL. Tout ou partie se retrouve par ailleurs, traduit ou non, comme dans les mémoires de master de mes étudiants ou leurs thèses de doctorat. Enfin, elle a été revue suite aux travaux de Loïc Vial au mois d'août 2020 et, à nouveau, en septembre 2021 après la rédaction des premiers chapitres de ce document.

Les autres langues ont très peu de corpus annotés manuellement en sens (au mieux, 10 000 mots). Nous tirons partie de notre corpus anglais unifié et utilisons la traduction automatique pour projeter des annotations dans les langues cibles. Nous avons ainsi publié UFSAC-ara, pour la langue arabe, et UFSAC-fra, pour le français (voir section 3.2.2.2). En ce qui concerne les méthodes, nous utilisons aujourd'hui intensivement des réseaux neuronaux profonds pour WSD et avons proposé plusieurs algorithmes dont ceux de (Vial *et al.*, 2019c) qui permettent d'obtenir des résultats à l'état de l'art sur l'ensemble des langues testées (anglais, arabe et français). Nous avons également appliqué la WSD à la traduction automatique, à la détection du plagiat multilingue et à l'augmentation des ressources lexicales.

5.1 La clarification de sens

La clarification du sens est une tâche centrale à plusieurs applications du TALN comme, par exemple, la traduction automatique et la détection de plagiats. Centrale ne veut pas dire forcément explicite et bien souvent une clarification est indirectement et implicitement réalisée. C'est même le cas dans l'immense majorité des travaux en traduction automatique statistique (classique ou neuronale). Ce sont les exemples qui vont guider la clarification. On notera que, dans ces cas, la clarification ira jusqu'à un certain niveau de granularité. Il sera assez haut pour traduire vers l'anglais depuis le français puisqu'il est inutile d'aller jusqu'à la différence entre l'‹*animal*› et le ‹*dispositif électronique*› pour clarifier « *la souris mange le fromage* » ; plus fin pour traduire d'une de ces langues vers le malais qui, lui, distingue les sens par l'utilisation de deux items lexicaux différents (‹*tikus*› et ‹*tetikus*›).

Dans les travaux que je mentionne dans ce chapitre, la clarification est explicite et est vue comme un ensemble de tâches liées qu'il s'agit de résoudre automatiquement, interactivement ou manuellement pour clarifier au mieux un texte. Dans la section 2.3.1.1, je revenais sur 5 des phénomènes interdépendants à résoudre pour clarifier un énoncé. Le phénomène de l'ambiguïté lexicale (1), la recherche des références (2), la recherche des rattachements prépositionnels, (4) la recherche des fonctions lexicales, 5) la recherche des chemins interprétatifs.

Faute de collaborations sur ces sujets, les 4 derniers ont été petit à petit mis de côté au profit du seul premier. Les futur travaux avec Maximin Coavoux (CNRS, LIG, GETALP) autour de la thèse de doctorat de Jeongwoo Wang sur la clarification de sens dans plusieurs langues, dont l'anglais et le français, devraient me permettre d'avancer sur tout ou partie des autres phénomènes dans les mois et années à venir.

Reste donc la désambiguïsation lexicale que j'ai étudiée intensivement depuis 2010, d'abord seul puis avec plusieurs collègues de l'équipe GETALP, Hervé Blanchon (MCF-HDR à l'UGA), Jérôme Goulian (MCF à l'UGA), Benjamin Lecouteux (MCF à l'UGA) et Gilles Sérasset (MCF à l'UGA) au travers de plusieurs stages de master et thèses de doctorat (Marwa Hadj Salah, Mohammad Nasiruddin, Andon Tchechmedjiev, Loïc Vial). Comment l'avons-nous abordée et comment la traiter quelle que soit la langue ?

5.2 Principes & définitions

La tâche de désambiguïsation lexicale (*Word Sense Disambiguation*) consiste à trouver, pour chaque mot d'un texte, le sens le plus approprié parmi un inventaire de sens prédéfini. Par exemple, dans la phrase « *La souris mange le fromage.* », l'algorithme devrait choisir le sens de ‹‹souris›› qui correspond à l'**animal** plutôt que celui qui correspond au **dispositif électronique**. L'ambiguïté lexicale est un phénomène que l'on retrouve dans la plupart des mots utilisés. Ainsi, les auteurs du corpus DSO dont nous parlerons plus tard estiment, par exemple, dans (Ng et Lee, 1996) que les 121 mots les plus fréquents de la langue anglaise constituent 1 mots sur 5 dans les textes généraux et ont environ 7,8 sens par mot suivant le *Princeton WordNet*. J'ai rappelé dans la section 4.3.4, le chiffre de 55% dans l'expérience sur le français menée pendant ma thèse (Schwab, 2005).

La désambiguïsation lexicale a ainsi pour objectif d'assigner à chaque mot w_i d'un texte de *m* mots, l'un de ses sens $w_{i,j}$. La définition du sens j du mot i est notée $d(w_{i,j})$. L'espace de recherche correspond à toutes les combinaisons de sens possibles pour le texte considéré [2]. Ainsi, une configuration C du problème est représentée par un vecteur d'entiers tel que $j = C[i]$ est le sens $w_{i,j}$.

Pour cette tâche, seul l'anglais peut être considéré comme une langue réellement dotée. En effet, deux types de ressources qui demandent un travail humain particulièrement important peuvent être utilisées pour créer un système de désambiguïsation lexicale : les ressources basées sur des savoirs (dictionnaires, bases lexicales…) et les corpus annotés en sens.

Il convient parfois de distinguer la désambiguïsation lexicale *explicite*, qui est celle que nous venons de décrire, et la désambiguïsation lexicale *implicite*. On retrouve cette dernière dans un certain nombre d'applications du traitement automatique des langues qui l'exploitent grâce au co-texte ou au contexte. C'est en

2. Cet espace de recherche est ainsi rapidement gigantesque comme nous l'avons montré dans (Schwab *et al.*, 2013a). Nous reviendrons sur ce point lorsque nous parlerons des méthodes basées sur les similarités – voir section 5.6.4.

particulier vrai pour les tâches qui disposent d'un grand nombre de données permettant un apprentissage d'autant plus robuste. C'est ainsi le cas de la traduction automatique.

5.3 Applications de la désambiguïsation lexicale

Sans être exhaustif, on peut lister un certain nombre d'applications pour lesquelles la désambiguïsation lexicale explicite pourrait être utile. Cela ne signifie pas forcément que les performances de la désambiguïsation lexicale explicite sont suffisantes y compris en anglais, y compris en 2021, pour améliorer l'ensemble de ces tâches. Nous y reviendrons, car ces tâches peuvent également permettre d'évaluer la désambiguïsation lexicale

Le transfert lexical : un phénomène commun en traduction automatique est que les mots se traduisent souvent différemment en fonction de leur sens. Ainsi, le terme ‹*souris*› se traduit en anglais par ‹*mouse*›, que ce soit pour son sens d'**animal** ou pour son sens de **dispositif électronique**, se traduit en malais par ‹*tikus*› dans le premier cas et par ‹*tetikus*› dans le second. De même, le terme anglais ‹*bank*› peut se traduire en français par ‹*banque*› ou par ‹*berge*›.

Nous avons exploité cette application dans la thèse sur l'interopérabilité des bases lexicales d'Andon Tchechmedjiev (Tchechmedjiev, 2016) ainsi que dans les thèses de Loïc Vial (Vial, 2020) et Marwa Hadj Salah (Hadj Salah, 2018). Il convient de noter que dans les expériences que nous avons menées, les systèmes intégrant des information de sens de manière explicite obtiennent de meilleures performances que ceux qui n'en intègrent pas [3] en particulier pour les paires de langues les moins dotées.

La recherche d'information est une tâche qui consiste à retrouver des documents pertinents en fonction d'une requête. Il s'agit d'un cas clair où la désambiguïsation implicite est classique puisque les utilisateurs sont habitués à poser des requêtes dans leur moteur de recherche clarifiant de fait les termes ambigus (ex : « *souris animal* »). Désambiguïsation implicites et explicites sont exploitées dans les travaux de Jibril Frej (Frej, 2021, p. chap. 6) (FREJ *et al.*, 2020; Frej *et al.*, 2020a) que j'ai co-encadrés avec Jean-Pierre Chevallet (MRIM, LIG, UGA).

La synthèse de la parole est une tâche qui consiste à produire artificiellement de la parole à partir d'un énoncé. Elle peut se révéler très utile pour les personnes en situation de déficience visuelle qui peuvent ainsi accéder au contenu de textes écrits, ou pour celles qui éprouvent des difficultés à parler puisqu'elles peuvent

3. Qui restent ainsi au niveau implicite.

alors écrire leur message sous forme de texte, la machine se chargeant ensuite de le prononcer à leur place. Enfin, on peut la retrouver également dans certains services téléphoniques. Une désambiguïsation est nécessaire lorsque un mot peut se prononcer de plusieurs manières. Ainsi, on prononce ‹‹fils›› différemment dans le titre du film « *Deux fils* » et dans « *fils et laines* », un nom de magasin.

La génération automatique de pictogramme est une application que j'ai proposée avec Benjamin Lecouteux (MCF UGA) en 2017, et qui se poursuit aujourd'hui à travers le projet franco-suisse Propicto (voir section 2.4.3), il s'agit de transcrire un énoncé écrit ou oral en série de pictogrammes. L'un des défis importants de cette application est le manque de données, puisque la taille des corpus alignés *énoncés oraux-pictogrammes* est actuellement nulle et celle des corpus alignés *énoncés écrits-pictogrammes* est de quelques dizaines d'exemples. Un des projets actuellement mené dans notre équipe consiste à aligner manuellement et semi-manuellement une base de pictogrammes classiquement utilisée dans le monde du handicap avec le *Princeton WordNet* afin de pouvoir bénéficier de son écosystème – voir (Schwab *et al.*, 2020b) et section 3.2.1.2. La chaîne de traitement utilise ainsi une désambiguïsation lexicale exploitant WordNet afin de permettre au système de choisir les pictogrammes les plus pertinents.

La lecture active consiste à enrichir un texte avec des informations permettant de mieux en comprendre le sens que ce soit pour une personne étrangère ou dans un domaine qui n'est pas le sien (médical par exemple). L'utilisateur peut ainsi passer sa souris sur un mot pour faire apparaître sa définition, sa traduction ou toute autre information pertinente fonction de l'application visée (Abdellaoui *et al.*, 2018).

5.4 Évaluation de la désambiguïsation lexicale

Pour évaluer un système de désambiguïsation lexicale, deux approches sont possibles, les approches *in vivo* et les approches *in vitro*.

5.4.1 Évaluation *in vivo*

Les évaluation *in vivo* consistent à intégrer les systèmes de désambiguïsation lexicale dans une application particulière (ex : traduction automatique, réponse à des questions…) et à comparer ensuite les performances. Cette idée paraît *a priori* particulièrement séduisante, car conforme à l'idée originelle de la désambiguïsation lexicale, mais ce type d'évaluation souffre de plusieurs problèmes.

— *Problème de génie logiciel* : il n'est pas en général techniquement facile d'intégrer un système de désambiguïsation lexicale à une application. Les

contraintes d'adaptation des formats pourraient ainsi appauvrir les sorties de tel ou tel système. Si l'on imagine, par exemple, un système capable de fournir plusieurs sorties, l'application pourrait n'en nécessiter qu'une seule même si le système de désambiguïsation lexicale en considère plusieurs comme équiprobables.

— *Problème de l'évaluation des applications* : l'évaluation des applications elles-mêmes reste un problème ouvert et sujet à débats. Ainsi, le choix d'une application particulière reporte le problème de la comparaison entre les système sur les mesures d'évaluation de cette application et leurs limites.

— *Problème de domaine* : un système de désambiguïsation lexicale pourrait être tout à fait adapté à une application particulière utilisée dans un certain domaine applicatif, mais pas du tout à une autre. Cette limite se retrouve également dans les évaluations *in vitro* et constitue une limite classique en traitement automatique des langues et de la parole.

— *Complémentarité avec les ressources de l'application* : comme précisé dans la section 5.3, les applications réalisent de fait une désambiguïsation lexicale, désambiguïsation qui peut alors être qualifiée d'implicite par rapport à l'explicite qui nous intéresse plus particulièrement ici. Selon les besoins en désambiguïsation lexicale de l'application et la qualité de la désambiguïsation explicite réalisée, la désambiguïsation explicite est plus ou moins nécessaire. Dans un cas extrême où la désambiguïsation lexicale explicite réalise parfaitement ce qu'il faut en termes de désambiguïsation lexicale, la désambiguïsation lexicale implicite ne peut rien apporter de plus et se révèle inutile. On peut ainsi simplement se demander si une évaluation *in vitro* n'analyserait pas simplement la complémentarité de la désambiguïsation lexicale explicite et des ressources qu'elle utilise avec celles de l'application.

En pratique, l'évaluation *in vivo* a rarement été utilisée, McCarthy (2002) la propose dans le cadre d'une tâche de substitution lexicale et Vickrey *et al.* (2005) pour une tâche de transfert lexical. La première a été mise en place lors de SemEval 2007 et à nouveau en 2019 avec les travaux de Pilehvar et Camacho-Collados (2019) sur les mots en contexte (*Word-in-Context*). On peut le constater, dans tous les cas on peut utiliser de la désambiguïsation lexicale implicite ou explicite et, de fait, les approches utilisées par les participants sont généralement de nature implicite largement adaptées vu la quantité de données disponibles en anglais.

Dans notre équipe, même si les évaluations *in vitro* restent prépondérantes pour les raisons pratiques que nous venons d'évoquer, nous avons conduit plusieurs expériences d'évaluation *in vivo* dans le cadre de la traduction automatique.

Les travaux de Marwa Hadj Salah et Loïc Vial intègrent tous les deux de l'éva-

luation *in vivo* dans le cadre de la traduction automatique. Comme les systèmes de désambiguïsation lexicale qu'ils ont développés sont de même nature, c'est-à-dire qu'ils ont les mêmes caractéristiques logicielles, cette intégration souffre ainsi moins de la limite de génie logiciel évoquée ci-dessus. Il en est probablement différent des autres points, mais nous n'avons pas étudié le sujet lors de ces travaux [4].

5.4.2 Évaluation *in vitro*

En évaluation *in vitro*, les systèmes sont évalués en utilisant des corpus annotés de référence. Il s'agit en particulier de ceux des campagnes d'évaluation Semeval-SensEval qui existent depuis presque 25 ans.

Avec cette dernière approche, les mesures classiques sont la précision P, le rappel R et le score F1 qui correspond à la moyenne harmonique de P et R.

La précision se définit comme :

$$P = \frac{\text{Nombre de mots annotés correctement}}{\text{Nombre total de mots annotés}} \tag{5.1}$$

le rappel :

$$R = \frac{\text{Nombre de mots annotés correctement}}{\text{Nombre total de mots à annoter}} \tag{5.2}$$

le score F1 :

$$F1 = \frac{2 * P * R}{P + R} \tag{5.3}$$

On reviendra dans la section 5.5.2.2 sur les difficultés liées à la construction de corpus annotés en sens, mais intrinsèquement, l'évaluation *in vitro* souffre de plusieurs problèmes. Les deux premiers sont très classiques en évaluation dans le traitement automatique des langues alors que le troisième est plus singulier.

— *Le problème de la langue* est que l'on ne peut évaluer sur une langue uniquement des systèmes capables d'annoter des textes issus de cette même langue ;

— *Le problème du domaine* lié à la représentativité d'un corpus d'évaluation : un système peut-être meilleur sur certains domaines et pas sur d'autres. Ce problème est très lié aux ressources disponibles ;

— *Le problème de l'inventaire de sens* est lié plus directement à la tâche. En effet, on peut seulement évaluer les systèmes capables d'annoter un texte dans l'inventaire de sens choisi. Par exemple, si notre corpus d'évaluation en anglais est annoté avec la dernière version de DbNary basée sur une extraction

4. En particulier car je n'ai repensé la question qu'après leur soutenance.

du Wiktionnaire, il ne sera possible d'utiliser que des systèmes de désambiguïsation lexicale capables d'annoter des corpus avec cette même version de DbNary mais pas avec une ancienne ni même avec le *Princeton WordNet*. Une solution consiste à utiliser les liens entre sens qui existent parfois entre les versions ou les bases lexicales, mais ils n'existent pas toujours car c'est un travail très important à réaliser de manière manuelle. Il y a beaucoup de travaux de recherche sur la création automatisée ou semi-automatisée de ces liens. Nous y avons contribué surtout à travers les travaux d'Andon Tchechmedjiev (Tchechmedjiev, 2016), et aussi ceux de Marwa Hadj Salah dont une partie de la thèse a consisté à établir des liens entre OntoLex et *Princeton WordNet*. Citons également les alignements entre la base de pictogrammes *ARASAAC*[5] et *Princeton WordNet* de (Schwab *et al.*, 2019) – voir section 3.2.1.2.

Si nous avons développé depuis 2010 des systèmes capables d'annoter avec plusieurs inventaires de sens (BabelNet, DbNary, ...), c'est sur les annotations avec *Princeton WordNet* que nous avons principalement concentré nos efforts, puisqu'il s'agit, de loin, de la ressource la plus utilisée[6].

5.5 Ressources génériques utiles pour la désambiguïsation lexicale

En désambiguïsation lexicale, deux types de ressources génériques sont fondamentales et demandent un travail humain important voire considérable s'il doit partir de zéro : (1) les corpus manuellement annotés en sens et (2) les sources de connaissances. Jusqu'à récemment, Les travaux sur l'anglais ont globalement montré une corrélation directe entre la quantité/qualité de corpus annotés et la qualité du système final.

Dans le processus d'informatisation d'une langue, avant de pouvoir construire un corpus annoté manuellement en sens, il faut disposer d'un inventaire de sens. Aucune autre langue que l'anglais ne bénéficie d'autant de corpus manuellement annotés en sens ou de connaissances lexicales. La figure 5.3 illustre l'état des ressources librement disponibles pour la désambiguïsation lexicale pour un certain nombre de langues vers 2015. Un recensement plus précis est difficile à obtenir et il faut donc interpréter les positions des langues les unes par rapport aux autres

5. https://arasaac.org

6. Nous ne revenons pas ici sur ce point mais il s'agit sans conteste d'un pis-aller puisque les découpages des sens sont différents entre les langues. On pourra se rapporter à ma thèse (Schwab, 2005) ou à celle d'Andon Tchechmedjiev (Tchechmedjiev, 2016) pour des discussions sur ce sujet.

plutôt que de manière absolue, sauf pour l'anglais que nous avons placé le plus en haut à droite. Si nous pouvons considérer que la quantité de données annotées est un paramètre quantifiable (par exemple en nombre moyen d'occurrences par terme du lexique), la richesse des sources de connaissances disponibles est, elle, plus difficile à estimer. C'est en particulier le cas entre deux langues différentes puisque la taille de leur vocabulaire est différente. Il faut noter également que certaines langues peuvent bénéficier de données provenant d'autres langues par des alignements (comme c'est le cas dans *BabelNet*, par exemple).

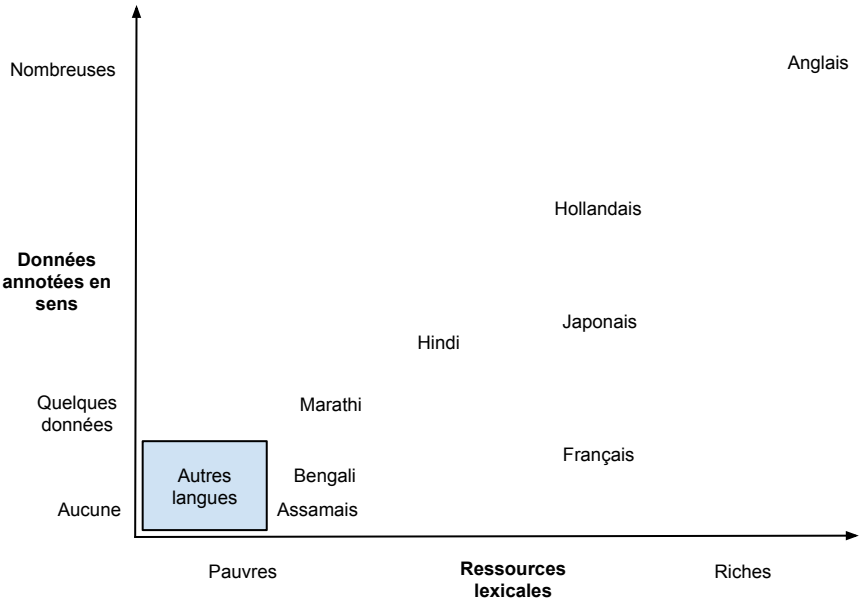

Figure 5.1 – Données disponibles pour la désambiguïsation lexicale en fonction de la langue en 2015

5.5.1 Bases lexicales

Une base lexicale est une source de connaissances structurées qui contient des informations sur les objets lexicaux d'une ou plusieurs langues et qui est accessible par logiciel. On retrouve ainsi parmi les bases lexicales des dictionnaires électroniques, des thésaurus, des ontologies, des graphes de connaissances. Le *Princeton WordNet*, DBNary, ou encore l'ensemble des ressources du *Linguistic Linked Open Data cloud*, sont considérées comme des bases lexicales.

Avant les années 1990, la désambiguïsation lexicale de l'anglais était surtout réalisée à partir de dictionnaires électroniques non librement disponibles. Le *Princeton WordNet* (Miller, 1995), initié au milieu des années 1980, a permis la mise

à disposition d'une ressource utilisable librement. Il a rapidement conduit à la disparition de l'usage des dictionnaires électroniques en désambiguïsation lexicale.

Le *Princeton WordNet* est organisé autour de la notion d'ensemble de synonymes (*synsets*) décrits par une partie du discours (nom, verbe, adjectif, adverbe), une définition et leurs liens (hyperonyme, hyponyme, antonyme...). Chaque sens d'un item lexical (entrée) correspond à un *synset*. La version courante du *Princeton WordNet*, la 3.0, comprend 155 287 items lexicaux pour un total de 117 659 *synsets*. Des versions pour d'autres langues existent mais, faute de moyens humains équivalents, leur qualité est inférieure à celle de l'anglais. Bien souvent, les mots de ces langues sont décrits grâce à des *synsets* du *Princeton WordNet* anglais.

Dans les langues sur lesquelles j'ai plus particulièrement travaillé en désambiguïsation lexicale (français, arabe, bengali), c'est le cas de l'*Arabic WordNet* (Elkateb et Fellbaum, 2006; Abouenour *et al.*, 2013). La *Global WordNet Association* établit la liste des wordnets existants [7].

Il existe d'autres bases lexicales multilingues comme BabelNet (Navigli et Ponzetto, 2012b) ou DBNary (Sérasset, 2015b), que nous avons utilisé dans certaines applications ou campagnes d'évaluation, que je ne présente pas ici.

Dans les travaux futurs, une piste de travail est d'exploiter au mieux les informations parfois complémentaires que l'on peut retrouver dans ces différentes bases. Il serait alors naturel de se tourner vers les données issues de *Linguistic Linked Open Data* [8] (Chiarcos *et al.*, 2013; McCrae *et al.*, 2016) dont l'objectif est d'unifier et de connecter les bases lexicales librement disponibles. Il s'agirait alors de reprendre les travaux menés en 2012-2016 lors de la thèse d'Andon Tchechmedjiev avec Jérôme Goulian (GETALP, LIG, UGA) et Gilles Sérasset (GETALP, LIG, UGA) exploitant les connaissances translingues et que nous avons mis de côté pour nous concentrer sur les méthodes multimonolingues, c'est-à-dire des méthodes considérant individuellement chaque langue en fonction des données disponibles. C'est sur ces méthodes qui ont fait l'objet des thèses de Marwa Hadj Salah et Loïc Vial que j'insisterai plus particulièrement dans la suite.

7. http://globalwordnet.org/wordnets-in-the-world/
8. https://linguistic-lod.org/llod-cloud

5.5.2 Corpus annotés

Selon Benoit Habert, « *un corpus est une collection de données langagières qui sont sélectionnées et organisées selon des critères linguistiques et extra-linguistiques explicites pour servir d'échantillon d'emplois déterminés d'une langue* » (Habert *et al.*, 1998). Généralement, un corpus contient jusqu'à une douzaine de millions de mots et peut être lemmatisé et annoté avec des informations concernant les parties du discours. Parmi ces corpus, nous trouvons le *British National Corpus* (Burnard, 1998) – 100 millions de mots – et le *American National Corpus* (Ide et Macleod, 2001) – 20 millions de mots. Les textes proviennent de diverses sources comme des journaux, des livres, des encyclopédies ou du Web.

5.5.2.1 Exemples de corpus annotés en sens

En désambiguïsation lexicale, plusieurs corpus annotés en sens sont utilisés. Nous pouvons citer, par exemple :

1. La *Defense Science Organisation* (Ng et Lee, 1996; Ng, 1997) a produit un corpus non disponible librement. 192 800 mots ont été annotés avec des *synsets* du *Princeton WordNet*. L'annotation se concentre sur 121 noms (113 000 occurences) et 70 verbes (79 800 occurences) qui ont été choisis parmi les plus fréquents et les plus ambigus de l'anglais. Selon les auteurs, la couverture correspond à environ 20% des occurrences de noms et de verbes en anglais.

2. Le *SemCor* (Miller, 1995) est un sous-ensemble du Corpus de Brown (Francis et Kučera, 1964). Sur les 700 000 mots de ce dernier, environ 230 000 sont annotés avec des *synsets* du *Princeton WordNet*. L'annotation porte au total sur 352 textes. Pour 186 d'entre eux, 192 639 mots (soit l'ensemble des noms, verbes, adjectifs et adverbes) sont annotés. Sur les 166 autres, seulement 41 497 verbes sont annotés.

3. Les corpus issus des campagnes d'évaluation. Depuis 1998, il y a eu plusieurs campagnes (SemEval-SensEval) destinées à évaluer la désambiguïsation lexicale. La plupart ont concerné l'anglais mais également le japonais, l'espagnol, le chinois ou le français. La taille de ces corpus est de l'ordre d'une centaine de fois plus petite que celle des deux précédents corpus, soit quelques milliers de mots.

Le tableau 5.1 présente un tableau des corpus existants pour l'anglais et annotés dans une des versions de *Princeton WordNet*, en septembre 2021 (Vial *et al.*, 2018b; Vial, 2020).

Corpus	Année de sortie	Inventaire de sens originel	Nombre de phrases	Nombre de mots	Nombre de mots annotés en sens
SemCor	1993	WordNet 1.6	37 176	778 587	229 533
DSO	1997	WordNet 1.5	101 004	2 705 190	176 197
WNGC	2007	WordNet 3.0	117 659	1 634 691	496 776
OMSTI	2015	WordNet 3.0	863 648	36 636 675	1 109 147
MASC	2008	WordNet 3.0	31 759	585 353	113 518
Ontonotes	2013	Ontonotes	124 852	2 475 987	233 616
Train-O-Matic	2017	WordNet 3.0	788 888	31 708 188	834 455
SensEval 2 (t.m.)	2001	WordNet 1.7	252	5 766	2 282
SensEval 2 (é.l., entr.)	2001	WordNet 1.7	8 611	251 772	8 451
SensEval 2 (é.l., éval.)	2001	WordNet 1.7	4 328	125 477	4 239
SensEval 3 (t.m.)	2004	WordNet 1.7.1	352	5 541	1 850
SensEval 3 (é.l., entr.)	2004	WordNet 1.7.1	7 860	241 565	7 819
SensEval 3 (é.l., éval.)	2004	WordNet 1.7.1	3 944	122 131	3 849
SemEval 2007 - 07	2007	WordNet 2.1	245	5 637	2 261
SemEval 2007 - 17	2007	WordNet 2.1	135	3 201	455
SemEval 2013	2013	BabelNet 1.1.1	306	8 391	1 644
SemEval 2015	2015	BabelNet 2.5	138	2 604	1 022

Table 5.1 – Statistiques générales des corpus annotés en sens. « é.l. » correspond aux tâches d'« échantillon lexical », « t.m. » correspond aux tâches de désambiguïsation lexicale « tous mots », « entr. » correspond aux corpus d'« entraînement » et « éval. » aux corpus d'« évaluation ». (Très légèrement) adapté de (Vial, 2020).

5.5.2.2 Difficultés liées à la construction d'un corpus annoté

Il existe peu de données manuellement annotées. La *Global WordNet Association* dresse la liste des 26 corpus annotés avec un wordnet [9]. Ces corpus concernent 17 langues. Seules trois d'entre elles (l'anglais, le néerlandais et le bulgare) atteignent les 100 000 annotations. À notre connaissance, il existait en 2015 très peu de données annotées en sens pour le français (environ 3600 mots annotés avec le dictionnaire Larousse pour la campagne Romanceval 1998 et 1656 mots annotés avec des sens de BabelNet pour la tâche 12 de la campagne SemEval 2013) et pour l'arabe (32000 du AQMAR 1.0 Arabic Wikipedia Supersense Corpus développé par Schneider *et al.* (2012) qui est un corpus multidomaines qui contient 65 000 mots issus de 28 articles de Wikipédia arabe, annotés à la main par des supersens propres à ce corpus. La version n°5 de l'OntoNotes propose également des annotations sur environ 13 000 mots parmi 300 000 issus de news. Ces deux langues qui nous ont intéressé plus particulièrement ces dernières années étaient ainsi peu dotées en ce domaine.

En effet, la construction d'un corpus manuellement annoté en sens est une tâche d'annotation particulièrement difficile. Ainsi, s'il n'y avait que 45 annotations possibles pour le *Penn Treebank* (Marcus *et al.*, 1993), un corpus annoté en parties du discours, il y en a autant que de synsets (117 000) pour une annotation en sens issus du *Princeton WordNet*. Pour l'annotation du corpus de la *Defense Science Organisation*, alors que les conditions étaient plus favorables que celles du *SemCor* (uniquement 191 mots différents pour seulement 1 800 annotations possibles), le taux d'annotation était seulement de 150 à 250 mots par heure (1 personne-année pour les 192 800 occurrences de mots) tandis que les annotateurs du *Penn Treebank* réalisaient 6 000 annotations par heure.

Qui plus est, l'annotation de corpus de sens doit être répétée pour chaque langue, pour chaque domaine et pour chaque inventaire de sens.

Des recherches ont visé à faciliter cette annotation. Par exemple, pour le néerlandais, Vossen *et al.* (2011) utilisent un algorithme de désambiguïsation automatique dont les annotations les moins sûres sont vérifiées/modifiées manuellement par les annotateurs, et Mihalcea et Chklovski (2003) utilisent une méthode de production participative (*crowdsourcing*) pour augmenter le nombre d'annotateurs.

Le même principe est utilisé par Taghipour et Ng (2015a), qui annotent une grande quantité de textes provenant de corpus parallèles anglais-chinois grâce à

9. http://globalwordnet.org/wordnet-annotated-corpora/ consultée le 6 septembre 2021. Il existe d'autres corpus annotés avec des synsets de wordnets comme ces corpus de domaines annotés avec des *synsets* de l'*Hindi WordNet* http://www.cfilt.iitb.ac.in/wsd/annotated_corpus/

un système de désambiguïsation lexicale utilisant les mots de la traduction alignés comme source principale (projection interlingue d'annotations). La qualité du corpus ainsi généré est ensuite démontrée via la création d'un système de désambiguïsation lexicale supervisée entraîné sur ce même corpus et obtenant des résultats légèrement en retrait pas rapport à son utilisation sur des corpus manuellement créés.

De notre côté, nous n'avons pas encore travaillé sur la récolte de corpus annotés en sens mais nous nous sommes concentrés sur l'unification du format des corpus anglais (UFSAC – voir section 3.2.1.1) ainsi que sur la génération de corpus synthétiques, à savoir les corpus UFSAC traduits depuis l'anglais vers une autre langue, puis sur le portage des annotations (UFSAC-Mult – voir section 3.2.2.2). Toutefois, dans le cadre des projets Propicto et InterAACtionBox (voir section 2.4), nous envisageons de récolter des corpus voix-écrit-pictogrammes qui pourront être facilement convertis en corpus annotés en sens *Princeton WordNet* via les pictogrammes.

5.6 Méthodes de désambiguïsation lexicale

Il existe dans la littérature scientifique différentes typologies de méthodes de désambiguïsation lexicale. Souvent, les chercheurs font la distinction entre les méthodes supervisées et les méthodes non supervisées. Les premières sont basées sur des techniques d'apprentissage machine qui utilisent des corpus annotés en sens, alors que les secondes ne les utilisent pas.

Si tout le monde semble d'accord sur la définition de la désambiguïsation lexicale supervisée, il n'en va pas de même pour la désambiguïsation lexicale non supervisée, comme le souligne (McCarthy, 2009). La désambiguïsation lexicale uniquement basée sur WordNet, par exemple, est-elle une sorte d'algorithme non supervisé (Navigli et Lapata, 2010) ou s'agit-il d'un troisième type de désambiguïsation lexicale (Agirre et Edmonds, 2007) ? Comme nous l'avons présenté précédemment, deux types de ressources nécessaires pour créer un système de désambiguïsation lexicale sont à la fois difficiles et coûteuses à construire : les corpus annotés en sens et les bases de données lexicales. Alors que les bases de données lexicales peuvent être utilisées directement dans de nombreuses applications ou par l'humain, les corpus annotés en sens ont moins d'applications et sont ainsi indirectement plus coûteux à construire. Bien entendu, ces deux dimensions ne sont pas indépendantes : par exemple, les bases de données lexicales telles que BabelNet (Navigli et Ponzetto, 2012b) sont étendues à l'aide de corpus annotés.

En 2013, j'ai proposé la typologie suivante lors du dépôt d'un projet ANR

Jeune Chercheur-Jeune Chercheuse. J'utilise également cette typologie dans les cours que j'ai donnés ou que je donne en Master MIASHS* et au Master MOSIG* de l'ENSIMAG/IM2AG, tous les deux à l'Université Grenoble Alpes. Cette typologie a ensuite été suivie dans un certain nombre des écrits de notre équipe. Je profite de l'écriture de ce mémoire pour la présenter de façon aussi complète que possible.

La figure 5.2 présente les différentes méthodes de désambiguïsation lexicale en fonction de leur utilisation de ressources lexicales et de corpus annotés en sens, telles qu'elles étaient schématiquement en 2015.

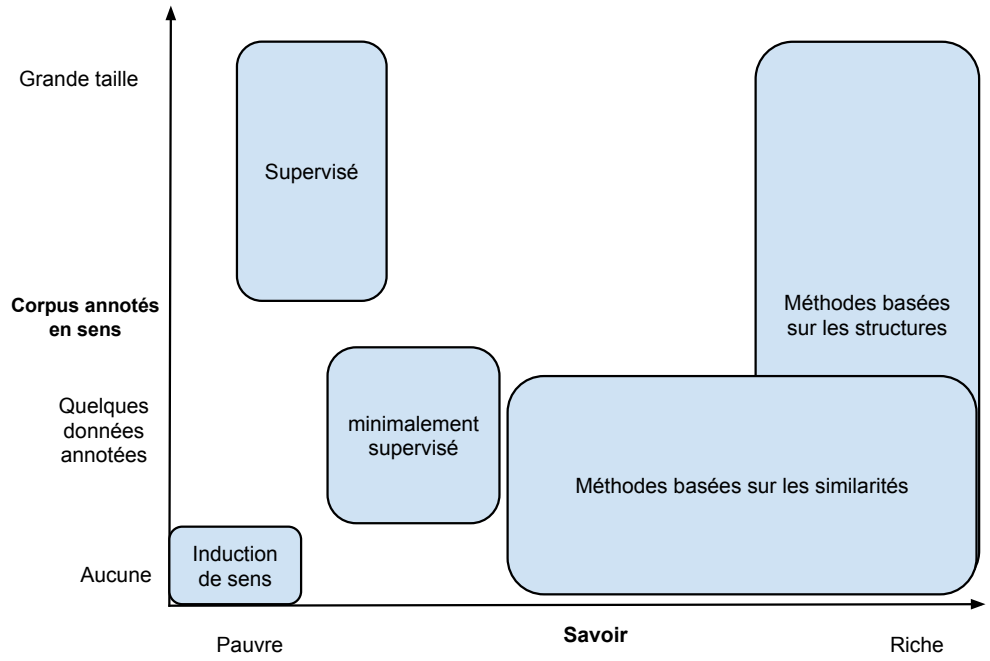

FIGURE 5.2 – Les différentes méthodes de désambiguïsation lexicale

5.6.1 Processus de mise en œuvre de la désambiguïsation lexicale

Quelle que soit la méthode de désambiguïsation lexicale, trois étapes sont nécessaires pour la mettre en place :

1. *Constitution d'une ressource générique* : plusieurs ressources non dédiées à la désambiguïsation lexicale sont possibles : dictionnaires, encyclopédies, corpus non annotés, corpus annotés, bases lexicales... Cette étape, optionnelle, est souvent réalisée par des équipes spécialisées. Certaines de ces res-

sources sont constituées manuellement (*Princeton WordNet*, *SemCor*) tandis que d'autres le sont largement automatiquement (*BabelNet*, *Princeton WordNet Gloss Corpus–WNGC*). Leur mode de constitution a évidemment des conséquences importantes sur leur qualité.

2. *Constitution d'une ressource dédiée à la désambiguïsation lexicale* – utilisation d'une ou plusieurs ressources génériques pour donner une représentation informatique à chacun des sens d'un mot. Il s'agit ainsi de constituer une ressource dédiée à la tâche. Ces sens sont définis par l'expertise humaine ou induits à partir des contextes d'utilisation dans les textes. Mathématiquement, ces représentations peuvent être des graphes (des réseaux lexicaux), des sacs de mots, des n-grammes ou encore des représentations vectorielles.
3. *Utilisation de la ressource dédiée pour désambiguïser des textes* – il s'agit de l'algorithme de désambiguïsation proprement dit. Il est plus ou moins complexe et dépend de la ressource dédiée. Il peut s'agir d'un algorithme de plus proches voisins, de réseaux de neurones, de recherche de concentrateurs (*hub*) dans un graphe, de pageRank, d'algorithmes génétiques ou encore d'algorithmes à colonie(s) de fourmis. Plusieurs paramètres peuvent entrer en compte. Certains sont communs à chaque algorithme comme la taille du contexte considéré pour le mot à désambiguïser (par exemple quelques mots avant ou après celui-ci, la phrase qui le contient, voire le texte si les ressources computationnelles et la combinatoire le permettent), tandis que d'autres dépendent du type d'algorithme : par exemple la limite à considérer pour la profondeur de la recherche dans un graphe, les paramètres à considérer pour des algorithmes stochastiques ou encore le type ou le nombre de couches et leurs connexions dans un réseau de neurones.

Ainsi, selon cette méthode, Schwab *et al.* (2013b) utilisent WordNet comme ressource générique ; des représentations en sac de mots issues des définitions des sens et de leurs liens comme ressource dédiée ; un algorithme à colonies de fourmis et une mesure de proximité entre les sacs de mots comme algorithme de désambiguïsation lexicale.

De même, Navigli et Ponzetto (2012b) utilisent de nombreux corpus annotés en sens, plusieurs wordnets et Wikipedia pour constituer une ressource générique, *BabelNet*. Il s'agit d'une base lexicale à grande échelle construite par alignement automatique des *synsets*, issus de *Princeton WordNet* et de pages Wikipedia correspondantes. BabelNet introduit la notion de *Babel Synset*, qui contient tout le contenu du *synset* correspondant dans le *Princeton WordNet*, ainsi qu'un ensemble de pages Wikipedia similaires. À partir de cette ressource, les auteurs construisent un grand réseau lexical dédié, dont ils exploitent la structure pour désambiguïser des textes.

Enfin, Vial *et al.* (2019c) utilise le *Princeton WordNet*, les corpus annotés *SemCor* et *WNGC* ainsi que BERT [10](Devlin *et al.*, 2019) comme ressource générique ; un réseau de neurones exploitant les poids issus de BERT et appris à partir des corpus annotés et du réseau lexical du *Princeton WordNet* comme ressource dédiée ; enfin ce réseau de neurones entraîné est utilisé comme algorithme de désambiguïsation lexicale.

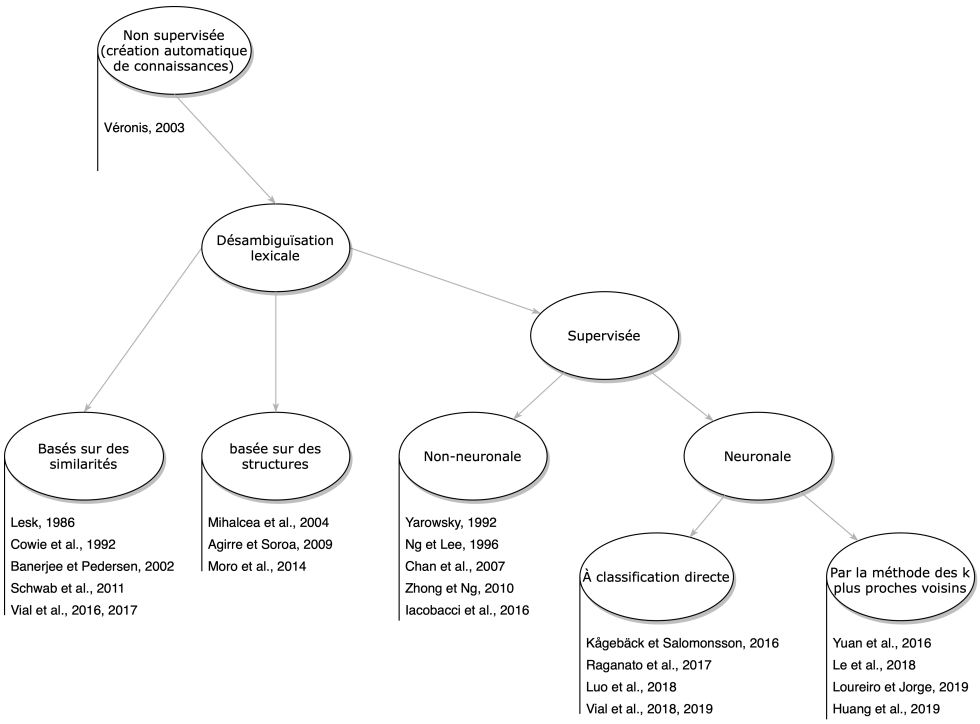

FIGURE 5.3 – Quelques-unes des différentes approches de la désambiguïsation lexicales. Nous nous restreignons ici à celles que nous avons plus particulièrement étudiées au cours de nos recherches. Les approches non supervisées sont situées un peu au dessus, car la plupart des méthodes de désambiguïsation lexicale sont utilisables une fois la ressource générique constituée.

10. Plus précisément, le modèle *bert-large-cased*.

5.6.2 Méthodes non supervisés (induction de sens)

Les méthodes non supervisées constituent leur ressource générique uniquement à partir de corpus non annotés en sens. Cette étape, appelée également induction des sens des mots utilise des techniques d'apprentissage par machine sur les corpus non annotés en sens sans avoir de connaissance *a priori* sur la tâche. Ces algorithmes induisent les sens des mots en considérant les co-occurrences suivant l'hypothèse distributionnelle (Harris *et al.*, 1989) pour laquelle deux mots sont considérés comme sémantiquement proches (similaires) s'ils sont utilisés dans les mêmes contextes. Les principales techniques de regroupement comprennent : l'identification de structures particulières dans les graphes de co-occurrence (Véronis, 2003) et le regroupement de vecteurs construits à partir du contexte (voisins ou voisins des vecteurs voisins).

Une fois l'étape de regroupement réalisée, il s'agit d'une désambiguïsation lexicale classique fonction de la représentation calculée obtenue. Ainsi, avec un graphe, on aura, par exemple, des méthodes basées sur les concentrateurs ; avec des vecteurs des méthodes de distance et de similarité, à partir de corpus annotés, les méthodes supervisées deviennent accessibles ...

Ces méthodes sont particulièrement utilisées pour les langues les moins dotées. Nous les avons, par exemple, utilisées dans les travaux de Mohammad Nasiruddin avec Hervé Blanchon (GETALP, LIG, UGA) sur le bengali (Nasiruddin *et al.*, 2014).

5.6.3 Méthodes supervisées

Les méthodes supervisées utilisent des techniques d'apprentissage automatique. Elles utilisent des classifieurs classiques entraînés sur des corpus annotés en sens : séparateurs à vaste marge (SVM) – NUS-PT (Chan *et al.*, 2007) – , classifieurs naïfs bayésiens – NUS-ML (Cai *et al.*, 2007) – , combinaison de séparateurs à vaste marge, entropie maximale – LCC-WSD (Novischi *et al.*, 2007). On ne peut pas vraiment affirmer que tel ou tel classifieur est meilleur qu'un autre et ce qui différencie les performances des systèmes est principalement directement lié à la taille des données annotées. Bien que LCC-WSD et NUS-ML utilisent uniquement SemCor, NUS-PT utilise à la fois SemCor et le DSO.

Pendant une dizaine d'années, la recherche sur les méthodes supervisées a stagné et il a fallu attendre l'arrivée des méthodes neuronales pour que l'état de l'art sur l'anglais évolue de façon significative. Nous avons particulièrement contribué à ces méthodes en incluant d'autres langues comme l'arabe ou le français et en proposant une méthode originale permettant de joindre les informations issues des

textes et les informations issues des réseaux lexicaux grâce aux travaux menés lors des thèses de Loïc Vial et Marwa Hadj Salah (voir section 2.3.1.2 et section 5.7.4).

5.6.4 Méthodes basées sur les similarités

Nous avons intensément travaillé sur ce sujet entre 2010 et 2018 parce qu'elles étaient les seules possibles sur un certain nombre de langues à cause du manque de corpus annotés en sens disponibles à l'époque. Ma première contribution a été de proposer une définition de ces méthodes basées sur un double algorithme, un algorithme local et un algorithme global.

5.6.4.1 Algorithmes locaux : similarités entre sens

Les algorithmes locaux correspondent aux mesures de similarité sémantique et permettent d'estimer la proximité entre deux sens de mots du texte. Ainsi, ‹*policier*› et ‹*commissariat*› devraient être considérés comme plus proches que ‹*hélicoptère*› et ‹*éléphant*›. Ces méthodes consistent ainsi à donner un score censé refléter la proximité des objets linguistiques (généralement des mots ou des sens de mots) comparés. Un très grand nombre d'études ont été consacrées à la similarité entre mots et nous avons déjà évoqué ce sujet dans le cadre des représentations vectorielles dans le chapitre précédent. Il existe ainsi des dizaines de mesures peut-être même des centaines si on considère leurs variantes. On peut les distinguer en fonction de leurs domaines d'arrivée :

— $[0, 1]$: ce sont des similarités pour lesquelles une valeur avoisinant 1 indique des sens proches alors qu'une valeur avoisinant 0 indique des sens éloignés. C'est le cas, par exemple des mesures vectorielles comme dans LSA (Deerwester *et al.*, 1990) ou Word2Vec (Mikolov *et al.*, 2013) ;

— $[0, \pi/2]$ ou $[0, 90]$: un angle mesuré en radians ou en degrés comme c'est le cas, par exemple, pour les vecteurs d'idées (Schwab, 2005). Un angle proche de $0°$ (0 radian) correspond alors à des sens voisins et un angle proche de $90°$ ($\pi/2$ radians) correspond à des sens très éloignés ;

— $\mathbb{N}^+$, un nombre entier positif, comme c'est le cas pour les mesures comme celle de Lesk (1986).

Parmi ces mesures, on peut également citer celle de Hirst et St-Onge (1998), basée sur la distance entre deux sens dans un réseau lexical ; (Rada *et al.*, 1989) et *Leacock and Chodorow* similaires à la précédente mais ne considérant que les liens de type hyperonymie ; les mesures ou distances entre vecteurs (voir chapter 4) ou encore celle de Lesk (1986) qui mesure le nombre de mots entre les définitions des sens correspondants.

Le lecteur pourra consulter (Pedersen *et al.*, 2005; Cramer *et al.*, 2010) ou (Navigli, 2009) et surtout (Harispe *et al.*, 2015) pour un panorama plus complet. En désambiguïsation lexicale basée sur les similarités, ces méthodes sont utilisées de façon locale entre deux sens de mots, et sont ensuite appliquées à un niveau global.

5.6.4.2 Algorithmes globaux : cohérence globale de l'énoncé

L'algorithme global, de son côté, propage les mesures locales aux niveaux supérieurs (syntagmes, phrases, paragraphes, voire tout le texte, selon l'algorithme choisi) afin de désambiguïser l'ensemble de l'énoncé.

Ces algorithmes globaux ont besoin d'une mesure pour permettre une évaluation pertinente d'une configuration donnée. Rappelons qu'une configuration C du problème est représentée par un vecteur d'entiers tel que $j = C[i]$ est le sens $w_{i,j}$. Le score du sens sélectionné pour un mot donné peut être exprimé par la somme des scores locaux entre ce sens et les sens sélectionnés de tous les autres mots du contexte considéré. Ainsi, pour évaluer une configuration du problème donné, on peut considérer comme *mesure globale*, comme fonction de coût, la somme des scores de tous les sens sélectionnés des mots du texte :

$$Score(C) = \sum_{i=1}^{m} \sum_{j=i}^{m} M_{local}(w_{i,C[i]}, w_{j,C[j]}) \tag{5.4}$$

Le calcul exhaustif de l'ensemble des possibilités est impossible pour des raisons d'explosion combinatoire comme nous l'avons montré dans (Schwab *et al.*, 2013a) et il faut se tourner vers des méthodes approchées pour résoudre ce problème. L'objectif des algorithmes globaux est ainsi de trouver la ou les solutions qui maximisent la cohérence sémantique du texte selon le ou les algorithmes locaux utilisés.

Parmi les algorithmes globaux, on trouve ainsi, entre autres, des algorithmes génétiques (Gelbukh *et al.*, 2003), de recuit simulé (Cowie *et al.*, 1992), ainsi que toute une batterie de méthodes bio-inspirées qui ont longtemps arpenté les couloirs du LIG – colonies de fourmis (Schwab *et al.*, 2011; Schwab *et al.*), colonies d'abeilles (Abualhaija et Zimmermann, 2016), de coucous (Vial *et al.*, 2017a), ou de chauves-souris (Vial *et al.*, 2017e).

Pour en savoir plus sur les algorithme globaux, le lecteur pourra se reporter en particulier aux articles où nous développons plus particulièrement cette thématique, à savoir (Schwab *et al.*; Tchechmedjiev *et al.*; Schwab *et al.*, 2013b) et (Vial *et al.*, 2017e).

5.6.5 Algorithmes locaux et algorithmes globaux

Nous montrons dans ces articles que la qualité de la désambiguïsation est essentiellement déterminée par les algorithmes locaux. En effet, l'algorithme global permet de parcourir plus ou moins efficacement l'espace afin de maximiser le score global calculé grâce aux algorithmes locaux. Ainsi, une mesure de proximité sémantique parfaite parviendrait à refléter la proximité qui existe entre les sens possibles pour les mots du texte et aiderait au mieux l'algorithme global à trouver le ou les sens les plus appropriés pour chaque mot d'un texte. Autrement dit, en un temps infini, tous les algorithmes globaux trouveraient la ou les solutions qui maximisent la cohérence sémantique du texte selon le ou les algorithmes locaux utilisés et, c'est donc de la performance de ces derniers que vient principalement la qualité de la désambiguïsation lexicale.

Dans (Schwab *et al.*, 2014), nous proposons une méthode qui analyse la corrélation entre score local et performance de la désambiguïsation lexicale dans le cadre d'une évaluation *in vivo*. L'hypothèse est que, si la corrélation est importante, l'apprentissage trouve plus facilement une configuration obtenant une bonne performance. Ces travaux sont restés préliminaires et il serait intéressant de les reprendre et de les affiner. On peut également regretter que les ressources n'aient pas été publiées à l'époque ; je me propose de prendre le temps de le faire faire. Peut-être le cadre de Propicto m'en offrira-t-il la possibilité.

5.6.6 Algorithmes basés sur les structures

Ces algorithmes reposent sur la topologie, la structure de grands graphes lexicaux. Des exemples typiques de cette catégorie sont les travaux de Roberto Navigli exploitant *BabelNet*. La ressource dédiée à la désambiguïsation est fabriquée à partir des liens issus de cette grande base lexicale, elle-même constituée de très nombreuses ressources (autres bases lexicales, corpus annotés en sens). Pour des raisons de complexité calculatoire, la désambiguïsation se fait souvent dans le contexte de la phrase et l'idée de base est de construire un nouveau graphe à partir de ses mots. Ces systèmes (Navigli et Ponzetto, 2012b) sont assez bons, quand on a une quantité importante de ressources, mais ces méthodes semblent limitées pour les langues qui ont moins de ressources et en particulier les langues pauvres en ressources annotées et la désambiguïsation ne peut se faire que par l'intermédiaire des autres langues. À notre connaissance, il existe assez peu de travaux utilisant BabelNet pour d'autres langues que l'anglais et aucun, par exemple, sur l'arabe ou le bengali, deux des langues qui nous ont plus particulièrement intéressés.

5.6.7 De l'anglais comme exemple de ce qu'il est possible d'obtenir ?

Le tableau 5.2 récapitule les types de méthodes qui existent et les ressources qu'elles utilisent classiquement. Il est ainsi possible d'analyser une langue en fonction des ressources dont elle dispose pour arriver à établir quel type de désambiguïsation lexicale est envisageable.

L'anglais dispose aujourd'hui sans nul doute des ressources ayant la meilleure qualité et la plus large couverture que ce soit pour les bases lexicales ou pour les corpus annotés en sens. C'est évidemment dû aux moyens humains et financiers dont cette langue bénéficie depuis des décennies.

Nous voyons l'étude des travaux sur cette langue comme un moyen d'estimer quelles ressources sont utiles et à quel point pour désambiguïser des textes dans une langue donnée. Ainsi, on peut par exemple, n'utiliser que des corpus bruts afin d'estimer quelle qualité il est possible d'obtenir de manière totalement automatique ou n'utiliser que des corpus annotés pour estimer l'intérêt de constituer une telle ressource pour une autre langue. Enfin, on pourrait n'utiliser qu'une partie des corpus pour mieux estimer la quantité de données nécessaires.

Cette vision ne doit pas conduire en revanche à un cercle vicieux qui continuerait à augmenter toujours plus les ressources en anglais sans travailler sur les ressources pour d'autres langues. C'est d'autant plus vrai pour les chercheurs dont l'anglais n'est pas la langue maternelle ce qui est le cas de la plupart des chercheurs de l'équipe. Les langues existent par elles-mêmes et il est important de travailler sur autant de langues que possible afin de ne pas perdre la richesse culturelle que chacune véhicule et garantir aux locuteurs l'accès à l'information.

Ce n'est pas uniquement vrai pour des raisons éthiques, ça l'est également car les langues ont des caractéristiques différentes et il n'est possible de vraiment comprendre un phénomène que si on l'a analysé dans plusieurs langues, chacune pouvant apporter à la compréhension globale du phénomène.

Ainsi, un équilibre doit être trouvé entre travail sur l'anglais pour essayer d'analyser ce qu'il est possible de réaliser, et travail sur la constitution de ressources pour les autres langues. Nous avons développé ces idées dans la section 3.1. Dans notre équipe, nous avons essayé de répondre à cet équilibre en travaillant sur des applications particulières visant un public spécifique grâce à des collaboration avec des entreprises et/ou des équipes de recherche pour travailler sur d'autres langues.

Nous avons travaillé sur le *bengali* avec Mohammad Nasiruddin, étudiant bangladais, Hervé Blanchon (GETALP, LIG, UGA) et Andon Tchechmedjiev doctorant dans l'équipe entre 2012 et 2015. Le bengali, également appelé bangla, est la

septième langue la plus parlée au monde avec plus de 250 millions de locuteurs et la plus orientale des langues indo-européennes. Elle est essentiellement parlée au Bangladesh (75% de la population) et la deuxième la plus parlée en Inde où elle est langue officielle de trois des vingt-trois états (Garry et Rubino, 2001). C'est pour cette langue que nous avons travaillé sur la création automatique de bases lexicales (induction de sens) puis mis au point la première traduction du SemCor vers une autre langue.

Nous avons travaillé sur l'*arabe* avec Marwa Hadj Salah et Hervé Blanchon (GETALP, LIG, UGA) que nous avons déjà évoqué à plusieurs reprises ici (voir section 2.3.1.2 et section 3.2.2.2). Le travail a consisté d'un côté à créer un corpus d'évaluation pour l'arabe et d'un autre d'un gros travail sur le portage des corpus d'UFSAC en anglais vers l'arabe ou tout autre langue du moment qu'elle possède une traduction automatique depuis l'anglais.

Nous avons travaillé sur le *français* avec les travaux avec Loïc Vial et Benjamin Lecouteux (GETALP, LIG, UGA), intégrés dans FLUE (voir section 3.3.3) et dans ceux menés avec Céline Vaschalde (alors étudiante en master 2 à l'Université d'Orléans). Ces travaux nous ont permis d'étudier les verrous liés à la transcription automatique de la parole sous forme pictographique pour participer à la communication avec les personnes en situation de handicap cognitif. Ces travaux sont précurseurs de ceux du projet ProPicto (voir section 2.4.3).

Enfin, nous avons travaillé sur le *multilingue* avec les travaux d'Andon tchechmedjiev entre 2011 et 2016 (de son M1 à son doctorat) avec Jérôme Goulian (GETALP, LIG, UGA) et Gilles Sérasset (GETALP, LIG, UGA) dans lesquels nous utilisons les ponts entre les langues grâce à BabelNet. Les limites constatées à l'époque sur cette base fabriquée de manière très automatique, et l'explosion des méthodes neuronales profondes, nous ont ensuite éloignés de cette piste. Les travaux autour des pictogrammes et leurs liens des bases lexicales comme WordNet devraient nous ramener vers ces possibilités, en particulier à travers Dbnary de Gilles Sérasset.

Concluons ce chapitre en abordant la manière dont nous avons investi la désambiguïsation lexicale neuronale et proposé une méthode qui exploitait conjointement les informations issues de corpus annotés en sens et celles issues de réseaux lexicaux.

Type de méthodes	Ressources génériques			Ressource dédiée	Algorithme de désambiguïsation	Langues concernées
	Bases lexicales	Corpus annotés en sens	Autres Corpus			
Non supervisées – induction de sens	-	-	+++	Créée par induction des sens en fonction du cotexte	Fonction de la structure de données de la ressources dédiée	Langues disposant uniquement de corpus de textes
Faiblement supervisées	++	+	+++	Itérations création système supervisé, anno--tation corpus non annoté	*Idem* supervisés	Langues disposant de corpus de textes et d'un inventaire de sens
Supervisées	++	+++	-	Méthodes d'apprentissage automatique à partir des textes annotés en sens	Application des modèles entraînés sur les textes	Essentiellement l'anglais. D'autres langues (jap, nld) dans une moindre mesure
Basées sur les structures	+++	+++	-	Graphe intégrant un maximum de ressources génériques	Algorithmes classiques de graphes (*PageRank*, degrés des nœuds, poids des chemins…)	Anglais et langues fortement liées dans des bases lexicales
Basées sur les similarités	+++	-	-	Basées sur les définitions et/ou les réseaux lexicaux issus des bases lexicales	Algorithmes locaux (similarités entre sens) ; Algorithmes globaux (similarité globale)	Langues disposant d'une base lexicale (dictionnaire, réseaux lexicaux))

TABLE 5.2 – Analyse des méthodes de désambiguïsation lexicale en fonction du processus de désambiguïsation lexicale qu'elles suivent et des langues auxquelles elles peuvent s'appliquer.

5.7 Désambiguïsation lexicale neuronale

Les méthodes supervisées [11] sont de loin les plus représentées car elles donnent généralement les meilleurs résultats dans les campagnes d'évaluation – par exemple (Navigli *et al.*, 2007). Les classifieurs à l'état de l'art combinaient jusqu'à récemment des caractéristiques précises telles que les parties du discours et les lemmes des mots voisins, (Zhong et Ng, 2010), mais ils sont maintenant remplacés par des réseaux de neurones récurrents qui apprennent leur propre représentation des mots (Raganato *et al.*, 2017; Le *et al.*, 2018; Vial *et al.*, 2019a).

Plusieurs avancées ont été réalisées dans la création de nouvelles architectures neuronales pour les systèmes supervisés de désambiguïsation lexicale. Ces systèmes atteignent des performances à l'état de l'art et certains peuvent intégrer des sources de connaissances externes. Dans cette section, nous donnons un aperçu de ces travaux.

5.7.1 Approches basées sur un modèle de langue

Ces approches reposent sur les approches neuronales contextualisées dont nous avons parlé dans le chapitre précédant (voir section 4.5). Dans ce type d'approches, initié par Yuan *et al.* (2016) et réimplémenté par Le *et al.* (2018), le composant principal est un modèle de langue neuronal capable de prédire un mot en tenant compte des mots qui l'entourent, grâce à un réseau neuronal entraîné sur une quantité massive de données non annotées (100 milliards de mots pour Yuan *et al.* (2016) et 1,8 milliards pour Le *et al.* (2018)).

Une fois le modèle de langue entraîné, il est utilisé pour produire des vecteurs de sens en moyennant les vecteurs de mots prédits par le modèle à l'endroit où ces mots sont annotés avec un sens particulier.

Au moment du test, le modèle de langue est utilisé pour prédire un vecteur en fonction du contexte environnant, et le sens le plus proche du vecteur prédit est attribué à chaque mot.

Ces systèmes ont l'avantage de contourner le problème de l'absence de données annotées en sens en concentrant le pouvoir d'abstraction offert par les réseaux neuronaux récurrents sur un modèle de langue de bonne qualité et entraîné de manière non supervisée. Cependant, ces méthodes souffrent toujours du manque de corpus annotés en sens, étant donné qu'ils restent indispensables pour la création des vecteurs de sens.

11. Cette section est principalement issue de (Vial *et al.*, 2019b).

5.7.2 Approches basées sur un classifieur linéaire et la fonction *softmax*

Dans ces systèmes, le réseau neuronal principal classifie et attribue directement un sens à chaque mot donné en entrée à l'aide d'une distribution de probabilité calculée par la fonction *softmax*. Les annotations en sens sont simplement considérées comme des balises placées sur chaque mot, à la manière d'une tâche d'étiquetage en parties du discours par exemple.

On peut distinguer deux branches distinctes de ces types de réseaux neuronaux :

1. Ceux dans lesquels il y a un réseau neuronal (ou classifieur) distinct et spécifique à chaque lemme du dictionnaire (Iacobacci *et al.*, 2016; Kågebäck et Salomonsson, 2016). Chaque classifieur est capable de gérer un lemme particulier avec ses sens. Par exemple, l'un des classifieurs est spécialisé dans le choix entre les quatre sens possibles du nom « souris ». Ce type d'approche est particulièrement adapté aux tâches d'échantillon lexical (*lexical sample*), où seul un petit nombre de mots distincts et très ambigus doivent être annotés dans plusieurs contextes. Mais cette approche nécessiterait plusieurs milliers de réseaux différents [12] pour pouvoir être utilisés dans les tâches de désambiguïsation lexicale totale (*all words*), dans lesquelles tous les mots d'un document doivent être annotés en sens.
2. Ceux dans lesquels il y a un seul réseau neuronal, plus grand et capable de gérer tous les lemmes du lexique, qui attribuent à un mot un sens issu de l'ensemble de tous les sens de l'inventaire de sens utilisé (Raganato *et al.*, 2017; Vial *et al.*, 2019a).

L'avantage de la première branche est que pour désambiguïser un mot, il est beaucoup plus facile de limiter notre choix à l'un de ses sens possibles que de chercher parmi tous les sens de tous les mots du lexique. Pour se donner une idée, le nombre moyen de sens des mots polysémiques dans WordNet est d'environ 3, alors que le nombre total de sens en considérant tous les mots est 206 941. [13]

La seconde approche a cependant une propriété intéressante : tous les sens résident dans le même espace vectoriel et partagent donc des caractéristiques dans les couches cachées du réseau. Cela permet au modèle non seulement de prédire un sens identique pour deux mots différents (synonymes), mais aussi de prédire un sens pour un mot non présent dans le dictionnaire (néologisme, faute d'orthographe, etc.).

12. L'ensemble de WordNet contient par exemple 26 896 mots polysémiques (https://wordnet.princeton.edu/documentation/wnstats7wn)
13. https://wordnet.princeton.edu/documentation/wnstats7wn

Enfin, Luo *et al.* (2018b,a) ont introduit une amélioration de ce type d'architectures, en calculant une attention entre le contexte d'un mot cible et les définitions de ses différents sens. Ainsi, leur travail est le premier à incorporer les connaissances de WordNet dans un système de désambiguïsation neuronal.

5.7.3 Les corpus annotés en sens, une limite des approches neuronales pour la désambiguïsation lexicale

Cependant, une des limitations majeures des systèmes supervisés est la quantité limitée de corpus manuellement annotés en sens. En effet, le SemCor (Miller *et al.*, 1993), qui est le plus grand corpus manuellement annoté en sens disponible, contient 33 760 labels de sens différents, ce qui correspond à seulement environ 16% de l'inventaire de sens de WordNet [14] (Miller *et al.*, 1990), la base de données lexicales de référence largement utilisée en désambiguïsation lexicale. De nombreux travaux tentent de résoudre ce problème via la création de nouveaux corpus annotés en sens, générés soit automatiquement (Pasini et Navigli, 2017), semi-automatiquement (Taghipour et Ng, 2015b), ou bien par *crowdsourcing* (Yuan *et al.*, 2016), nous en avons déjà parlé (voir section 5.5.2.2).

Une méthode complémentaire, que nous explorons ici, consiste à tirer parti des relations sémantiques présentes entre les sens de WordNet comme l'hyperonymie, l'hyponymie, l'antonymie, la méronymie, etc.

Notre méthode est basée sur les observations suivantes.

1. Un sens et ses sens voisins dans le graphe des relations sémantiques de WordNet véhiculent tous une même idée ou concept, à des niveaux d'abstraction différents.
2. Dans certains cas, un mot peut être désambiguïsé en utilisant seulement les sens voisins de ses sens, et pas nécessairement ses sens propres.
3. Par conséquent, nous n'avons pas besoin de connaître tous les sens de WordNet pour désambiguïser tous les mots de WordNet.

Par exemple, considérons le mot « souris » et deux de ses sens : la souris **d'ordinateur** et la souris **l'animal**. Les notions plus générales comme « être vivant » (hyperonyme de souris/animal) et « appareil électronique » (hyperonyme de souris/ordinateur), permettent déjà de distinguer les deux sens, et toutes les notions plus spécialisées telles que « rongeur » ou « mammifère » sont, elles, superflues. En regroupant ces étiquettes de sens, on peut bénéficier de tous les autres exemples

14. https://wordnet.princeton.edu/documentation/wnstats7wn

mentionnant un appareil électronique ou un être vivant dans un corpus d'entraînement, même si le mot « souris » n'est pas mentionné spécifiquement, pour désambiguïser le mot « souris ».

L'hypothèse de ce travail est ainsi qu'un sous-ensemble des sens du *Princeton WordNet* peut suffire pour désambiguïser tous les mots de la base lexicale. Nous proposons deux méthodes différentes pour construire un tel sous-ensemble, que nous appelons méthodes de compression de vocabulaire de sens. En utilisant ces techniques, nous sommes en mesure d'améliorer considérablement la couverture des systèmes de désambiguïsation lexicale supervisé, en éliminant quasiment le besoin d'une stratégie de repli habituellement employée pour les mots jamais observés pendant l'entraînement. Nous présentons des résultats qui surpassent l'état de l'art de façon significative sur toutes les tâches d'évaluation de la désambiguïsation lexicale ; nous fournissons à la communauté notre outil ainsi que nos meilleurs modèles préentraînés, sur un dépôt GitHub dédié [15].

5.7.4 Compression de vocabulaire de sens

Les systèmes supervisés neuronaux à l'état de l'art tels que (Yuan *et al.*, 2016; Raganato *et al.*, 2017; Le *et al.*, 2018; Luo *et al.*, 2018b,a; Vial *et al.*, 2019a) sont tous confrontés aux mêmes limitations.

1. La quantité de données annotées manuellement en sens étant très limitée, il se peut qu'un mot cible ne soit jamais observé pendant l'entraînement. Dans ce cas, le système ne peut pas être en mesure de l'annoter, et une stratégie de repli est généralement adoptée (par exemple utiliser le premier sens du mot dans WordNet).
2. Pour la même raison, un mot peut être observé, mais pas tous ses sens. Dans ce cas, le système est capable d'annoter ce mot, mais si le sens attendu n'a jamais été observé, le résultat sera faux, quelle que soit l'architecture sous-jacente du système supervisé.
3. L'empreinte mémoire des modèles neuronaux et leurs temps d'entraînement et d'exécution augmentent avec la quantité de données d'apprentissage et le nombre d'étiquettes de sens différentes prises en compte, nombre qui monte jusqu'à 206 941 si l'on considère toutes les étiquettes de sens de WordNet.

Afin de résoudre ces problèmes, nous proposons deux nouvelles méthodes permettant de regrouper des étiquettes de sens qui se réfèrent à des concepts similaires, tout en nous assurant que ces groupes de sens permettent toujours de discriminer

15. https://github.com/getalp/disambiguate

les différents sens de tous les mots du lexique, afin de retrouver l'étiquette de sens originale pour un mot au moment de le désambiguïser. En conséquence, le vocabulaire de sens, c'est à dire le nombre total d'étiquettes de sens dans notre inventaire de sens, diminue, et le système est capable de mieux généraliser, et sa couverture augmente.

5.7.5 Des sens aux *synsets* : une première compression de vocabulaire de sens à travers la synonymie

Dans le *Princeton WordNet* (Miller *et al.*, 1990), les sens sont organisés en ensembles de synonymes appelés *synsets*. Un *synset* est concrètement un groupe d'un ou plusieurs sens qui ont la même définition. Par exemple, les premiers sens des mots ‹*eye*›, ‹*optic*› et ‹*oculus*› appartiennent tous au même *synset* dont la définition est **l'organe de la vue**.

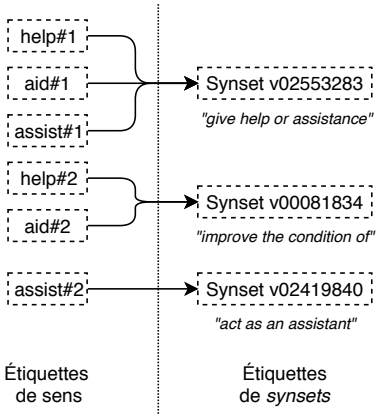

Figure 5.4 – Conversion des étiquettes de sens vers des étiquettes de synsets, appliqué aux deux premiers sens des mots ‹help›, ‹aid› *et* ‹assist›. *Le nombre de sens différents dans notre vocabulaire passe ainsi de six à trois (issue de (Vial* et al.*, 2019b)).*

La conversion des étiquettes de sens (« *Xème sens du mot N* ») aux étiquettes de *synsets* (« *synset numéro Y* »), illustrée dans la figure 5.4, est ainsi une façon de compresser le vocabulaire qui est déjà appliquée dans plusieurs travaux (Yuan *et al.*, 2016; Le *et al.*, 2018; Vial *et al.*, 2019a) sans être toujours explicitement précisée. Cette méthode contribue pourtant clairement à améliorer la couverture des systèmes supervisés. En effet, si le verbe ‹*aid*› annoté avec son premier sens est observé dans les données d'apprentissage, le contexte autour du mot cible peut être aussi utile pour annoter ultérieurement les verbes ‹*assist*› ou ‹*help*› avec la même étiquette de *synset*.

En allant plus loin, on peut trouver d'autres informations dans le *Princeton WordNet* qui peuvent aider à mieux généraliser. La première nouvelle méthode que nous proposons repose ainsi sur ce même principe de regroupement de sens, mais en exploitant les relations d'hyperonymie et d'hyponymie entre les sens.

5.7.5.1 Compression de vocabulaire de sens à travers les relations d'hyperonymie, d'hyponymie et d'instance

Selon Polguère (2003), l'hyperonymie et l'hyponymie sont deux relations sémantiques qui correspondent à un cas particulier d'inclusion de sens : l'hyponyme d'un terme est une spécialisation de ce terme, alors que son hyperonyme est une généralisation. Par exemple, une ‹*souris*› est un type de ‹*rongeur*› qui est à son tour un type d'‹*animal*›. Dans le *Princeton WordNet*, ces relations lient presque tous les noms allant de la racine générique, le nœud « entité » aux feuilles les plus spécifiques, par exemple « souris à pattes blanches ». Si l'on prend aussi en compte la relation d'instance, qui fonctionne de la même manière mais qui lie les entitées nommées aux noms courants (par exemple, ‹*Einstein*› est une instance de ‹*physicien*›), tous les noms du *Princeton WordNet* font partie de cette même hiérarchie.

Ces relations sont également présentes sur plusieurs verbes : ainsi, par exemple, « additionner » est une manière de « calculer » qui est à son tour une manière de « raisonner ».

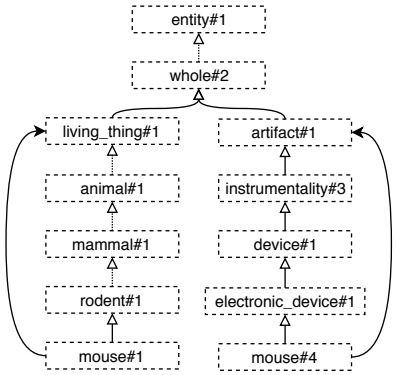

FIGURE 5.5 – Compression de vocabulaire utilisant la hiérarchie d'hyperonymie, appliquée au premier et quatrième sens du mot ‹mouse›. *Les lignes en pointillés indiquent que des nœuds ont été omis pour la clarté (issue de (Vial et al., 2019b)).*

Pour la désambiguïsation lexicale, tout comme le regroupement des synonymes en *synsets* aide à mieux généraliser, nous faisons l'hypothèse que le regroupement des sens faisant partie d'une même hiérarchie d'hyperonymie peut aussi aider à mieux généraliser, et que les concepts les plus spécialisés de WordNet sont souvent superflus. En effet, si l'on considère un sous-ensemble de WordNet qui ne comprend que le mot ‹*souris*›, avec son premier sens (le petit rongeur), son quatrième sens (le dispositif électronique), et tous leurs hyperonymes, tel qu'illustré dans la figure 5.5, on voit que les concepts **artefact** et **être vivant** suffisent à différencier les deux sens, et toutes les étiquettes plus spécialisées pourrait être ramenées à ces deux concepts. Ainsi, non seulement le vocabulaire de sens, c'est à dire le nombre d'étiquettes de sens dans notre inventaire, sera réduit, mais en plus tous

les autres **êtres vivants** donneront des exemples qui pourront ensuite permettre de différencier les deux sens de souris.

En considérant maintenant tout le vocabulaire de WordNet, l'objectif de notre méthode est ainsi de faire correspondre chaque sens à son ancêtre le plus haut dans sa hiérarchie d'hyperonymie, avec les contraintes suivantes : Premièrement, cet ancêtre doit permettre de discriminer tous les différents sens du mot cible. Deuxièmement, nous devons conserver les hyperonymes qui sont indispensables pour discriminer les sens des autres mots du dictionnaire. Par exemple, en prenant tout WordNet en considération, nous ne pouvons pas faire correspondre « souris#1 » à « être vivant#1 », parce qu'une étiquette plus spécifique, « animal#1 » est nécessaire pour distinguer les deux sens du mot « proie » (un sens décrit une personne et l'autre un animal).

Notre méthode fonctionne en deux étapes.

1. Nous marquons comme « nécessaires » les enfants du premier ancêtre commun de chaque paire de sens de chaque mot de WordNet.
2. Nous faisons correspondre chaque sens à son ancêtre le plus bas dans sa hiérarchie d'hyperonymie ayant été précédemment marqué comme « nécessaire ».

En conséquence, les sens les plus spécifiques de l'arbre qui ne sont pas indispensables pour distinguer un mot de l'inventaire lexical seront automatiquement supprimés du vocabulaire. En d'autres termes, l'ensemble de sens qui reste dans le vocabulaire est le plus petit sous-ensemble de tous les *synsets* qui sont nécessaires pour distinguer chaque sens de chaque mot de WordNet, en considérant seulement les liens d'hyperonymie, d'hyponymie et d'instance.

5.7.5.2 Compression de vocabulaire de sens à travers l'ensemble des relations sémantiques de WordNet

En plus de l'hyperonymie, de l'hyponymie et de la relation d'instance, WordNet contient plusieurs autres relations entre *synsets*, telles que la méronymie (X fait partie de Y, ou X est un membre de Y) et son opposé l'holonymie, l'antonymie (X est le contraire de Y) et son opposé la similarité, etc.

Nous proposons ainsi une deuxième nouvelle méthode de compression du vocabulaire de sens, qui prend en compte toutes les relations sémantiques offertes par WordNet, afin de former des groupes de *synsets* proches.

Par exemple, en utilisant toutes les relations sémantiques disponibles, nous pourrions former un groupe contenant « physicien », « physique » (domaine), « Einstein » (instance), « astronome » (hyponyme), mais aussi d'autres sens connexes

tels que « photon », car c'est un méronyme de « rayonnement », qui est un hyponyme de « énergie », qui appartient au même domaine de « physique », etc.

Notre méthode fonctionne en construisant ces groupes de manière itérative. Soit S l'ensemble des *synsets* de WordNet et C l'ensemble des groupes de *synsets* que l'on cherche à construire, on initialise d'abord C comme des singletons contenant chacun un *synset* différent.

$$S = \{s_0, s_1, ..., s_n\} \qquad C = \{c_0, c_1, ..., c_n\} \qquad C = \{\{s_0\}, \{s_1\}, ..., \{s_n\}\}$$

Ensuite, à chaque étape, on trie C par taille de groupes, et on sélectionne le plus petit groupe c_x ainsi que le plus petit groupe relié à c_x, c_y. On considère qu'un groupe c_a est relié à un groupe c_b si un *synset* $s_a \in c_a$ est relié à un *synset* $s_b \in c_b$ par n'importe quel lien sémantique. On fusionne c_x et c_y, si et seulement si le résultat de l'opération permet de discriminer les différents sens de tous les mots de la base lexicale. Si c'est le cas, on valide la fusion et on passe à l'étape suivante. Si ce n'est pas le cas, on annule la fusion et on essaye avec un autre groupe relié à c_x. S'il est impossible de fusionner un groupe avec c_x, alors on essaye avec le plus petit groupe suivant, et si aucune fusion n'est possible pour aucun des groupes, l'algorithme s'arrête.

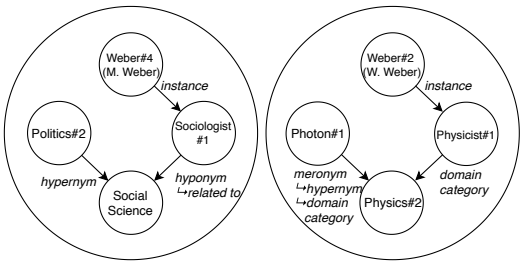

FIGURE 5.6 – Exemple de groupes de sens pouvant résulter de notre méthode, si on ne considère que deux sens du nom « Weber » et seulement certaines relations (issue de (Vial et al., 2019b)).

Dans la figure 5.6, nous montrons un ensemble possible de groupes qui pourraient résulter de notre méthode.

Cette méthode produit des groupes significativement plus grands que celle s'appuyant sur les hyperonymes. En effet, en moyenne, un groupe contient 5 *synsets* avec cette dernière, alors qu'il en contient 17 avec cette méthode. De plus, cette méthode, contrairement à la précédente, est également stochastique, parce qu'à chaque fois qu'on ordonne les groupes par taille, l'algorithme de tri place les groupes de même taille dans un ordre aléatoire. Cependant, comme nous réordonnons les groupes après chaque fusion, les groupes sont de taille assez équilibrés, et nous avons observé que la taille finale du vocabulaire (c.-à-d. le nombre de groupes) se situe toujours entre 11 000 et 13 000.

Dans la suite, on considère un ensemble C généré après que l'algorithme se soit arrêté après 105 775 étapes de fusion (générant ainsi 11 885 groupes de sens).

Méthode de compression	Taille du vocabulaire	Taux de compression	Couverture du SemCor
Référence	206 941	référence	16%
Synonymes	117 659	43%	22%
Hyperonymes	39 147	81%	32%
Toutes relations	11 885	94%	39%

Table 5.3 – Résultats de nos deux méthodes de compression de vocabulaire sur la taille du vocabulaire et la couverture du SemCor. La couverture correspond au rapport entre le nombre d'étiquettes de sens différentes observables dans le corpus et le nombre total d'étiquettes (taille du vocabulaire) – issue de (Vial et al., 2019b).

La tableau 5.3 montre l'effet de la compression de vocabulaire via les synonymes (sens vers *synsets*), de notre première nouvelle méthode utilisant les hyperonymes, ainsi que de notre deuxième nouvelle méthode utilisant toutes les relations de WordNet, sur la taille du vocabulaire de sens de WordNet, et sur la couverture du SemCor. Comme nous pouvons le constater, la taille du vocabulaire diminue considérablement grâce à nos méthodes, et la couverture d'un même corpus est nettement améliorée.

5.7.6 Protocole expérimental

Afin d'évaluer nos méthodes de compression de vocabulaire de sens, nous les avons appliquées à un système neuronal de désambiguïsation lexicale à l'état de l'art similaire à celui de Vial *et al.* (2019a) (voir la section 5.7.2). Notre réseau de neurones prend ainsi en entrée directement les mots sous forme vectorielle, à partir d'un modèle de vecteurs de mots préentraîné. Il a une ou plusieurs couches cachées, et une couche de sortie, qui associe à chaque mot une distribution de probabilité sur tous les sens du vocabulaire utilisé, à l'aide de la fonction *softmax*.

5.7.6.1 Détails de l'architecture

En entrée de notre réseau, nous utilisons les vecteurs contextualisés BERT (Devlin *et al.*, 2019). Nous avons utilisé le modèle pour l'anglais « bert-large-cased » qui est préentraîné sur BookCorpus (Zhu *et al.*, 2015) et Wikipedia, et qui produit des vecteurs de dimension 1 024.

Pour les couches cachées, nous avons 6 couches d'encodeurs Transformeur (Vaswani *et al.*, 2017a), avec les mêmes paramètres que le modèle « base » de

l'article original (8 têtes d'attention, dimension cachée de 2 048, et régularisation *dropout* à 0,1). Ces couches *Transformeur* sont basés sur le mécanisme d'auto-attention. Nous les avons utilisées à la place des cellules récurrentes plus classiques comme des *LSTM* ou des *GRU*, parce que plusieurs travaux récents ont montré leur plus grande efficacité dans une multitude de tâches, par exemple en traduction automatique (Vaswani *et al.*, 2017a; Ott *et al.*, 2018) et en modélisation de la langue (Devlin *et al.*, 2019).

De plus, étant donné que les vecteurs retournés par BERT encodent directement les positions des mots, il n'est pas nécessaire d'avoir une récurrence au niveau des couches cachées. Ainsi, nous n'ajoutons pas de vecteurs de positions supplémentaires en entrée de notre encodeur.

Pour tous les autres paramètres du modèle, comme le nombre de phrases par mini-lot, et la méthode d'optimisation, nous avons utilisé les mêmes paramètres que Vial *et al.* (2019a).

5.7.6.2 Entraînement du modèle

Nous avons comparé nos méthodes sur deux ensembles de corpus d'entraînement : le SemCor (Miller *et al.*, 1993), le plus grand corpus annoté en sens utilisé pour l'apprentissage de la plupart des systèmes supervisés de désambiguïsation lexicale, et la concaténation du SemCor et du WordNet Gloss Tagged [16]. Ce dernier est un corpus distribué dans WordNet depuis sa version 3.0, et il contient les définitions de tous les sens de WordNet, annotées manuellement ou semi-automatiquement en sens. Nous avons utilisé les versions de ces corpus fournies avec la ressource UFSAC 2.1 [17] (Vial *et al.*, 2017c) présentée dans la section 3.2.1.1.

Nous avons choisi d'ajouter uniquement le WNGT en plus du SemCor à nos données d'entraînement, et pas tous les corpus de la ressource UFSAC 2.1, parce que (1) c'est le seul, avec le SemCor, dans lequel tous les mots sont annotés en sens ; (2) l'inventaire de sens utilisé par les annotateurs est WordNet (3) la qualité est plutôt bonne car les annotations ne sont pas entièrement automatiques (4) la ressource est libre ce qui facilite la reproductibilité (voir section 3.1.3). Nous avons ainsi cherché à utiliser seulement des données de la meilleure qualité possible, pour éviter d'ajouter du bruit et/ou de rallonger le temps d'entraînement de nos modèles.

Nous avons entraîné chaque modèle sur 20 passes de nos données d'entraînement. Au début de chaque passe, nous avons mélangé toutes les phrases aléatoirement, et à la fin de chaque passe, nous avons évalué notre modèle sur un jeu

16. http://wordnetcode.princeton.edu/glosstag-files/glosstag.shtml
17. https://github.com/getalp/UFSAC

Système	SE2	SE3	SE07 17	SE13	SE15	ALL	SE07 07
SemCor, référence	91,15	96,76	97,58	91,06	94,78	93,23	92,84
SemCor, hyperonymes	98,03	99,19	99,78	99,15	98,39	98,75	98,85
SemCor, toutes relations	99,56	99,84	100	100	98,92	99,67	99,69
SemCor+WNGT, référence	97,81	98,92	99,34	97,63	99,34	98,26	98,45
SemCor+WNGT, hyperonymes	99,74	99,95	100	99,76	99,91	99,83	99,91
SemCor+WNGT, toutes relations	100	100	100	100	99,91	99,99	100
Nombre de mots à annoter	2282	1850	455	1644	1022	7253	2261

TABLE 5.4 – Couverture (en %) des corpus d'évaluation en fonction du corpus d'apprentissage et de l'utilisation de notre méthode de compression de vocabulaire – issue de (Vial et al., 2019b).

de développement, et nous avons conservé celui qui a obtenu le meilleur score F1 de désambiguïsation lexicale. Le corpus de développement est constitué de 4 000 phrases prises aléatoirement du WNGT pour le système entraîné sur le SemCor seul, et de 4 000 phrases extraites aléatoirement de nos données d'entraînement pour les autres.

Nous avons ainsi entraîné trois systèmes :

1. un système « référence » dont le vocabulaire de sens est celui de tous les *synsets* vus pendant l'entraînement (utilisant ainsi la compression classique via les synonymes) ;
2. un système « hyperonymes » entraîné dans les mêmes conditions, mais avec notre première méthode de compression du vocabulaire via les hyperonymes, les hyponymes et les instances appliquée sur le corpus d'entraînement ;
3. un système « toutes relations » qui applique cette fois-ci sur le corpus d'entraînement notre deuxième méthode de compression de vocabulaire via toutes les relations sémantiques de WordNet.

	Nombre de paramètres	
Système	SemCor	SemCor+WNGT
Référence	77,15M	120,85M
Hyperonymes	63,44M	79,85M
Toutes relations	55,16M	60,27M

TABLE 5.5 – Nombre de paramètres d'un modèle en fonction du corpus d'apprentissage et de notre méthode de compression de vocabulaire – issue de (Vial et al., 2019b).

Tous les entraînements ont été effectués sur un seul GPU Titan X de Nvidia. Dans le tableau 5.5, nous montrons le nombre de paramètres des différents mo-

dèles, en fonction du corpus d'entraînement et de notre méthode de compression du vocabulaire. Comme nous pouvons le voir, ce nombre est réduit par un facteur de 1,2 à 2 grâce à nos méthodes de compression.

5.7.7 Résultats

Nous avons évalué nos modèles sur tous les corpus d'évaluation de la désambiguïsation lexicale de l'anglais des campagnes d'évaluation SensEval/SemEval, c'est-à-dire les corpus d'évaluation « *à grain fin* » de SensEval 2 (Edmonds et Cotton, 2001), SensEval 3 (Snyder et Palmer, 2004), SemEval 2007 (tâche 17) (Pradhan *et al.*, 2007), SemEval 2013 (Navigli *et al.*, 2013) et SemEval 2015 (Moro et Navigli, 2015), ainsi que le corpus « *ALL* » constitué de leur concaténation. Nous avons également comparé nos résultats sur la tâche « *à grain fin* » de SemEval 2007 (tâche 7) (Navigli *et al.*, 2007).

Pour chaque évaluation, nous avons entraîné 8 modèles indépendants, et nous donnons le score obtenu par un système « ensemble » qui moyenne leurs prédictions à l'aide d'une moyenne géométrique.

Les scores obtenus par nos systèmes en comparaison avec les meilleurs systèmes de l'état de l'art et l'étalon du premier sens sont présentés dans le tableau 5.6, et le tableau 5.4 montre la couverture de nos systèmes sur les tâches d'évaluation.

Concernant les résultats présentés dans le tableau 5.6, nous observons que nos systèmes qui utilisent nos méthodes de compression de vocabulaire, que ce soit à travers les relations d'hyperonymie ou à travers toutes les relations obtiennent des scores qui sont globalement équivalents ou légèrement supérieurs aux systèmes « référence » qui n'utilisent pas nos méthodes.

Nos méthodes de compression améliorent cependant grandement la couverture de nos systèmes. En effet, comme nous pouvons le voir dans le tableau 5.4, sur un total de 7 253 mots à annoter pour le corpus « *ALL* », le système de référence entraîné sur le SemCor n'est pas capable d'annoter 491 d'entre eux, alors qu'avec la compression du vocabulaire à travers les hyperonymes, ce nombre descend à 91, et 24 avec la compression à travers toutes les relations.

Lors de l'ajout du WordNet Gloss Tagged aux données d'entraînement, seulement 12 mots ne peuvent pas être annotés avec le système « hyperonymes », et avec le système « toutes relations », et il n'y a plus qu'un seul mot (l'adjectif monosémique « cytotoxic ») ne peut pas être annoté parce que son sens n'a pas été vu pendant l'entraînement. Si nous prenons en compte uniquement les mots polysémiques, le système basé sur la compression à travers toutes les relations et entraîné

Système	SE2	SE3	SE07 17	SE13	SE15	ALL (concat. tâches précédentes) noms	verbes	adj.	adv.	total	SE07 07
Étalon du premier sens	65,6	66,0	54,5	63,8	67,1	67,7	49,8	73,1	80,5	65,5	78,9
UFSAC+1M (Vial *et al.*, 2019a)	74,6	69,4	60,7	69,8	74,2	-	-	-	-	†71,1	85,0
HCAN (Luo *et al.*, 2018b)	72,8	70,3	-	68,5	72,8	72,7	58,2	77,4	84,1	71,1	-
LSTMLP (Yuan *et al.*, 2016)	73,8	71,8	63,5	69,5	72,6	†73,9	-	-	-	†71,5	83,6
SemCor, référence	77,2	76,5	70,1	74,7	77,4	78,7	65,2	79,1	85,5	76,0	87,7
SemCor, hyperonymes	77,5	77,4	69,5	76,0	78,3	79,6	65,9	79,5	85,5	76,7	87,6
SemCor, toutes relations	76,6	76,9	69,0	73,8	75,4	77,2	66,0	80,1	85,0	75,4	86,7
SemCor+WNGT, référence	**79,7**	76,1	**74,1**	78,6	80,4	80,6	68,1	82,4	86,1	78,3	90,4
SemCor+WNGT, hyperonymes	**79,7**	77,8	73,4	**78,7**	**82,6**	81,4	68,7	83,7	85,5	**79,0**	90,4
SemCor+WNGT, toutes relations	79,4	**78,1**	71,4	77,8	81,4	80,7	68,6	82,8	85,5	78,5	**90,6**

*TABLE 5.6 – Scores F1 (%) sur les tâches de désambiguïsation lexicale de l'anglais des campagnes d'évaluation SensEval/SemEval. La tâche « ALL » est la concaténation de SE2, SE3, SE07 17, SE13 et SE15. La stratégie de repli est appliquée sur les mots dont aucun sens n'a été observé pendant l'entraînement. Les scores en **gras** sont à notre connaissance les meilleurs résultats obtenus sur la tâche. Les scores prefixés par une obélisque (†) ne sont pas fournis par les auteurs mais sont déduits de leurs autres scores.*

Corpus d'entraînement	Vecteurs de mots préentraînés	Ensemble	Scores F1 sur la tâche "ALL" (%)					
			Référence		Hyperonymes		Toutes relations	
			$\bar{x}$	σ	$\bar{x}$	σ	$\bar{x}$	σ
SemCor+WNGT	BERT	Oui	78,27	-	79,00	-	78,48	-
SemCor+WNGT	BERT	Non	76,97	±0,38	77,08	±0,17	76,52	±0,36
SemCor+WNGT	ELMo	Oui	75,16	-	74,65	-	70,58	-
SemCor+WNGT	ELMo	Non	74,56	±0,27	74.36	±0,27	68,77	±0.30
SemCor+WNGT	GloVe	Oui	72,23	-	72,74	-	71,42	-
SemCor+WNGT	GloVe	Non	71,93	±0,35	71,79	±0,29	69,60	±0,32
SemCor	BERT	Oui	76,02	-	76,73	-	75,40	-
SemCor	BERT	Non	75,06	±0,26	75,59	±0,16	73,91	±0,33
SemCor	ELMo	Oui	72,55	-	73,09	-	69,43	-
SemCor	ELMo	Non	72,21	±0,13	72,83	±0,24	68,74	±0,29
SemCor	GloVe	Oui	70,77	-	71,18	-	68,44	-
SemCor	GloVe	Non	70,51	±0,16	70,77	±0,21	67,48	±0,55
Système « élève » (Vial *et al.*, 2019a)								
SemCor+UFSAC+1M News 2016	GloVe	Oui	71,1					
HCAN (Luo *et al.*, 2018b)								
SemCor+WordNet glosses	GloVe	Non	71,1					
LSTMLP (Yuan *et al.*, 2016)								
SemCor+1K (private)	private	Non	71,5					

Table 5.7 – Étude des hyperparamètres sur la tâche "ALL" (concaténation des corpus de toutes les tâches de désambiguïsation lexicale à granularité fine de SensEval/SemEval). Pour les systèmes qui n'utilisent pas l'ensemble, nous montrons la moyenne des scores ($\bar{x}$) de huit modèles entraînés séparément, avec l'écart type (σ).

sur le SemCor n'est pas capable d'annoter un seul mot (l'adverbe « eloquently »). Avec le WNGT en plus, il a une couverture de 100%.

Par rapport aux autres travaux, nous obtenons des résultats surpassant significativement l'état de l'art dans toutes les tâches, notamment grâce à l'ajout du WordNet Gloss Tagged aux données d'entraînement, et des vecteurs BERT en entrée de notre système.

5.7.8 Étude des hyperparamètres

Afin de mieux comprendre l'origine de nos scores, nous étudions l'impact de nos principaux paramètres sur les résultats. En plus du corpus d'entraînement et de la méthode de compression du vocabulaire, nous avons choisi deux paramètres qui nous différencient de l'état de l'art : le modèle de vecteurs de mots préentraînés, et la méthode d'ensemble, et nous les avons fait varier.

Pour le modèle de vecteurs de mots, nous avons expérimenté non seulement avec BERT (Devlin *et al.*, 2019) comme pour nos résultats principaux, mais aussi avec ELMo (Peters *et al.*, 2018a) et GloVe (Pennington *et al.*, 2014). Pour ELMo, nous avons utilisé le modèle entraîné sur Wikipedia et les données monolingues de WMT 2008-2012. [18] Pour GloVe, nous avons utilisé le même modèle que Luo *et al.* (2018b) et Vial *et al.* (2019a) entraîné sur Wikipedia 2014 et Gigaword 5. [19] Comme les représentations vectorielles de GloVe n'encodent pas la position des mots (un mot a la même représentation quelle que soit sa position ou son contexte), nous avons réutilisé une couche de cellules *LSTM* bidirectionnelles de taille 1 000 par direction pour les couches cachées, comme nous avions fait dans Vial *et al.* (2019a).

Pour la méthode d'ensemble, nous avons expérimenté soit en l'utilisant, comme dans nos résultats principaux, c'est-à-dire en moyennant les prédictions de 8 modèles entraînés séparément, ou bien en donnant la moyenne et l'écart type des scores des 8 modèles évalués individuellement.

Comme nous pouvons le voir dans la tableau 5.7, le corpus d'entraînement supplémentaire (WNGT) et encore plus l'utilisation de BERT en tant que vecteurs de mots ont tous les deux un impact majeur sur nos résultats et conduisent à des scores supérieurs à l'état de l'art. L'utilisation de BERT au lieu de ELMo ou GloVe améliore respectivement le score d'environ 3 et 5 points dans chaque expérience, et l'ajout du WNGT aux données d'entraînement l'améliore encore d'environ 2 points. Enfin, l'utilisation d'ensembles ajoute environ 1 point au score F1 final.

Enfin, à travers les scores obtenus par les modèles individuels (sans ensemble), nous pouvons observer sur les écarts-types que la méthode de compression du vocabulaire par les hyperonymes n'a jamais d'impact significatif sur le score final. Cependant, la méthode de compression via toutes les relations semble avoir un impact négatif sur les résultats dans certains cas (en utilisant GloVe et ELMo particulièrement, et en utilisant le SemCor seul comme corpus d'entraînement).

18. https://allennlp.org/elmo
19. https://nlp.stanford.edu/projects/glove/

5.8 Conclusion

Dans ce chapitre, nous avons présenté les recherches en désambiguïsation lexicale que nous avons menées à Grenoble depuis une dizaine d'année. Nous avons parlé de la situation en terme de ressources en 2015 et expliqué les limites que cela posait en terme de performance dans la plupart des langues. Nous avons proposé UFSAC-Mult, une méthode qui consiste à traduire et à porter des annotations de corpus en anglais annotés en sens issus du *Princeton WordNet* par des méthodes neuronales.

Nous avons également exposé les différentes méthodes que nous avons explorées jusqu'aux dernières méthodes neuronales. Nous avons ainsi présenté deux nouvelles méthodes que nous avons appelées « *compression de vocabulaire de sens* » qui améliorent la couverture et la capacité de généralisation des systèmes de désambiguïsation lexicale supervisées, en réduisant le nombre d'étiquettes de sens différentes dans le *Princeton WordNet* afin de ne conserver que celles qui sont essentielles pour différencier les sens de tous les mots présents dans la base lexicale.

En considérant qu'une même étiquette de sens peut être appliquée à plusieurs mots différents, cette méthode permet effectivement une meilleure couverture. À l'échelle de l'ensemble de la base de données lexicales, nous avons montré que ces méthodes permettaient de réduire le nombre total d'étiquettes de sens différentes dans WordNet à seulement 6% de sa taille originale, et que la couverture d'un même corpus d'entraînement est ensuite plus que doublée.

Nous avons entraîné un système de désambiguïsation lexicale neuronal à l'état de l'art et nous avons montré que nos méthodes permettaient de réduire la taille des modèles par un facteur de 1,2 à 2 et de largement augmenter leur couverture, sans dégrader leurs performances. Au total, nous obtenons une couverture de 99,99% sur l'ensemble des tâches d'évaluation (soit un seul mot manquant sur les 7 253) lorsque l'on entraîne notre système sur le SemCor uniquement, et 100% lorsque l'on ajoute le WordNet Gloss Tagged aux données d'entraînement. On élimine ainsi quasiment le besoin d'une méthode de repli pour désambiguïser n'importe quel mot du vocabulaire de WordNet.

Notre méthode, combinée avec les récentes avancées en terme de vecteurs de mots préentraînés, a permis à notre système de surpasser nettement l'état de l'art dans toutes les tâches d'évaluation de la désambiguïsation lexicale de l'anglais, avec une bien meilleure couverture.

Les thèses de Loïc Vial et Marwa Hadj Salah ont montré l'apport de la désambiguïsation lexicale pour la Traduction Automatique. Les méthodes présentées ici

seront affinées et améliorées, en particulier dans le cadre de Propicto et de la Communication Alternative et Augmentée dans les années qui viennent.

Chapitre 6

Conclusions et perspectives

« *Donner la parole à ceux qui ne l'ont pas* ». C'est en quelques mots sans doute trop laconiques ce qu'est la Communication Alternative et Augmentée. Le Traitement Automatique des Langues et de la Parole est, pour faire simple, « *le domaine d'étude des techniques d'analyse (compréhension) et de génération (production) automatiques d'énoncées oraux ou écrits* ». Les liens entre les deux sont relativement naturels. Ainsi, le Graal théorique et technologique du multilinguisme pour lequel le GETALP se veut être un acteur de terrain particulièrement pertinent et crédible, et qui consisterait à parler dans sa propre langue et que l'autre entende dans la sienne, peut facilement être revisité. Le locuteur communique de la manière qui lui est la plus pratique et le ou les destinataires reçoivent le message de la manière qui leur est la plus pratique. La machine se charge de la médiation. Ce n'est ainsi plus seulement la parole ou l'écrit qui véhiculent le message mais bien tout canal de communication langagière voire même non-langagière qu'il s'agit d'interpréter. En d'autres termes, il s'agit d'estimer la compréhension, l'entendement de l'autre.

Je suis loin d'être le premier à mener des travaux à l'interface entre TALP et CAA. L'exercice de l'Habilitation à Diriger les Recherches et le temps m'ont conduit a considérer essentiellement ma propre perspective, passant sous silence les travaux d'autres chercheurs dans le domaine. Il conviendrait ainsi de revenir en détail sur d'autres recherches comme, par exemple, celles qui suivent.

— Bianco *et al.* (2006) proposent une « *plateforme de communication alternative* » portée par une société commerciale aujourd'hui disparue. Le logiciel bénéficiait de plusieurs modes d'accessibilité (clavier, souris, capteurs de mouvements) et permettait de reformuler une séquence de pictogrammes constituant le message en une phrase en langage naturel syntaxiquement et

sémantiquement correcte.

— Les travaux que mène Laurianne Sitbon à la *Queensland University of Technologie* à Brisbane en Australie. Je pense en particulier à TalkingBox (Bircanin *et al.*, 2020b) proche de certains jeux de GazePlay dans le monde du tangible ou de (Roberston *et al.*, 2021), un outil de communication à l'interface entre la grille de pictogrammes et le logiciel de scène visuelle. De plus, la démarche suivie semble similaire à certains aspects de la nôtre (Bircanin *et al.*, 2020a, 2021; Bayor *et al.*, 2021) puisque son équipe teste en particulier les différentes versions de ses prototypes dans plusieurs institutions locales et coconstruit les outils avec les personnes concernées et leur entourage.

— Les travaux qui se font, de manière plus générale, à l'étranger, comme ceux de l'association Rett italienne qui a une action similaire à celle de l'AFSR, et notre interAACtionBox [1] mais semble-t-il sans la dimension recherche.

— Les travaux de Jean-Yves Antoine et, en particulier, ceux autour de Sybille (Wandmacher *et al.*, 2008) repris depuis dans LifeCompanion [2], un projet libre et ouvert dont la sortie est annoncée pour l'automne 2021.

Nous connaissons ces travaux et nous espérons que les structures que nous mettons petit à petit en place nous permettront de collaborer avec certains d'entre-eux. L'ANR *AAC4All* devrait en particulier nous aider à trouver des synergies entre les travaux présentés dans cette habilitation et *LifeCompanion*.

Dans ce projet, financé par l'ANR (2022 - 2025), je serai responsable de la partie UGA. Ce projet regroupe un certain nombre d'acteurs français de la CAA, dont deux hôpitaux importants pour le handicap, le Centre Mutualiste de Kerpape (56) et l'Hôpital Raymond-Poincaré de Garches (92). Ce projet concerne [3] la Communication Alternative et Augmentée pour les personnes atteintes de handicaps sévères (infirmité motrice cérébrale, locked-in syndrome, maladies neurodégénératives, tétraplégie sévère...). Comme nous l'avons déjà souligné, quel que soit le mode d'interaction (manuel, regard...) ou de communication (texte, pictogrammes...), la saisie est lente et fatigante. Ce projet inclut donc une prédiction de mots ou de pictogrammes afin d'accélérer la communication.

AAC4All rassemble un consortium de partenaires (IRIT à Toulouse, LIFAT à Bois, Modyco à Nanterre, LIG à Grenoble) qui proposeront tous les composants logiciels. Ces modules intégreront des outils d'évaluation et de suivi qui faciliteront l'analyse par les thérapeutes du soutien apporté par un système/dispositif de

1. https://www.airett.it/amelielamica-delle-bimbe-che-da-voce-ai-loro-occhi/
2. https://esante.gouv.fr/projets-structures-3.0/life-companion
3. le texte qui suit est essentiellement une traduction de la proposition déposé auprès de l'ANR.

CAA. Plus précisément, le projet a les objectifs suivants.

— La définition d'une architecture ouverte pour l'interopérabilité des composants de la plateforme *AAC4All* (dispositif de saisie de l'utilisateur, clavier souple, prédiction de mots, support de saisie pictographique ou d'images, synthèse vocale), intégrant non seulement des modules d'évaluation et de suivi, mais aussi des modules de soutien à l'utilisation des systèmes (jeux sérieux pour favoriser la formation et l'appropriation du système par les utilisateurs). Notre expérience avec GazePlay sera essentielle ici (Voir section 2.4.2).

— Des avancées scientifiques et technologiques sur certains modules de CAA, afin non seulement de fournir des composants de pointe aux entreprises et aux chercheurs, mais aussi de relever les défis scientifiques qui restent un frein à la mise en place d'un système de CAA optimal. En particulier, nous étudierons des interfaces innovantes optimisant l'utilisation de la prédiction de mots, le couplage de la prédiction de mots avec la vérification orthographique à la volée, ainsi qu'une compréhension plus profonde des processus cognitifs impliqués dans la communication pictographique et/ou par images. Dans ce cadre, les travaux autour des modèles basés sur des vecteurs (voir chapter 4) seront fondamentaux pour représenter les usages en langue. Les travaux que j'ai menés avec Michael Zock (Zock et Schwab, 2011, 2013) autour du « *problème du mot sur le bout de la langue* » pourront également être intéressants pour combiner pictogrammes, idées et mots.

— La définition de lignes directrices ergonomiques aidant à la recommandation des systèmes de CAA par les ergothérapeutes. Cet objectif crucial sera abordé par l'évaluation expérimentale de différentes configurations des systèmes. Deux types d'évaluation sont prévus : (1) des évaluations technocentriques analysant les performances de chaque composant logiciel ; (2) des évaluations globales centrées sur l'utilisateur insistant sur le couplage entre la prédiction et le clavier virtuel. Des évaluations par des experts seront menées dans la même perspective.

  Les composants logiciels de AAC4ALL et de la plateforme seront en open-source afin d'assurer une large visibilité et reproductibilité. Le projet sera soutenu et partagé par une société émergente conçue pour développer et soutenir des projets de santé et de soins liés aux technologies d'assistance.

Dans les mois qui viennent, je devrais devenir co-directeur de plusieurs thèses dont le financement est d'ores et déjà assuré. À plusieurs reprises dans ce mémoire, j'ai parfois ouvert les perspectives sur certains de ces travaux. Commençons par présenter ces travaux avant de voir certains des problèmes de recherche auxquels ils nous permettront d'être confrontés.

— *les trajectoires de patients* (« *Le patient sera-t-il hospitalisé à la prochaine visite ?* » ; « *Aurait-il besoin d'un nouveau rendez-vous ?* »)[4] où nous explorons l'apport des données textuelles rédigées par les médecins aux données objectives (température, tension,...).

— *des travaux sur l'analyse sémantique*[5] qui consistent, entre autres, à identifier les rôles sémantiques joués par les éléments d'une phrase. Il s'agira ainsi de s'attaquer aux différentes ambiguïtés à traiter pour la clarification du sens (voir section 2.3.1.1) à savoir (1) l'ambiguïté lexicale (2) les références (3) les rattachements prépositionnels (4) les Fonctions lexicales (5) les chemins interprétatifs.

— *d'autres travaux autour de la Communication Alternative et Augmentée*[6] directement liés au TALP comme les deux projets suivants. (1) Le projet franco-suisse Propicto qui concerne la transcription automatique de la parole française sous forme pictographique dans les deux contextes que sont, d'un côté la communication avec des personnes ayant des déficits cognitifs et, de l'autre, la communication avec des patients n'ayant pas la même langue que le praticien. (2) L'organisation automatique des grilles de pictogrammes dans les logiciels de communication et en particulier AugCom (voir section 2.4.2). À ce jour, il n'existe aucune norme, ni aucune évaluation systématique permettant de mesurer objectivement l'adéquation de ces outils à une langue donnée ou un handicap donné. Nous avons, dans un premier temps, proposé des mesures objectives pour calculer le coût de production d'un texte donné pour une grille donnée Chasseur *et al.* (2020). Nos travaux actuels exploitent ces mesures pour trouver automatiquement des organisations optimales de grilles à partir de corpus oraux transcrits.

Je vais également continuer la supervision de deux doctorants qui ont commencé en 2020 avec les travaux sur la *Traduction Automatique de l'oral et de l'écrit massivement multilingue* dans le cadre de projets avec les entreprises FaceBook et Naver Labs. Il s'agit de produire des traductions vers plusieurs langues, y compris les paires de langues ne bénéficiant pas (*zero shot learning*) ou de peu de données parallèles.

4. Avec Lorraine Goeuriot, maîtresse de conférences en informatique à l'UGA et Thierry Chevalier, chef de clinique des universités à l'UGA et médecin généraliste à SOS Medecin Grenoble

5. Avec Maximin Coavoux, chargé de recherche CNRS dans l'équipe GETALP et Cédric Lopez, Directeur de recherche dans la société Emvista que nous devrions mener dans le cadre d'une thèse CIFRE avec cette société (Montpellier).

6. Avec Benjamin Lecouteux, maître de conférences à l'UGA, Pierrette Bouillon, Professeur à l'Université de Genève, Hervé Spechbach, médecin aux Hôpitaux Universitaires de Genève, Amélie Rochet-Capellan, chargée de recherche en sciences du langage au CNRS du Gipsa-Lab et Marion Dohen, maîtresse de conférences en sciences du langage à Grenoble INP/UGA et au Gipsa-Lab.

De ces contextes, plusieurs problèmes de recherche étroitement liés seront ainsi abordés.

— *Le transfert de connaissances à travers les langues* : il s'agit en particulier de se demander à quel point on peut considérer avoir des représentations interlingues dans les réseaux profonds en étudiant, par exemple, quelles connaissances passent d'une langue à l'autre pour une même tâche ou entre paires de langues, en particulier pour celles pour lesquelles on ne bénéficie pas de données parallèles.

— *Le transfert de connaissances à travers les tâches* : il s'agit en particulier de se demander à quel point on peut considérer avoir des représentations inter-tâches dans les réseaux profonds. Parmi les pistes susceptibles de permettre de répondre à cette question, identifier des tâches complémentaires, identifier des tâches aux profils similaires qui bénéficieraient à être apprises conjointement… (Caruana, 1997)

— *À quel point peut-on réduire la taille des représentations sans (trop) de perte de connaissances* : il s'agit d'un élément essentiel dans un monde où les problèmes liés au changement climatique sont prégnants ou simplement pour être mis à disposition sur des systèmes embarqués ou dans un contexte hospitalier pour limiter les échanges et les pertes énergétiques. L'extension des modèles doit ainsi se penser avec un ajout aussi minimal que possible de paramètres. Une autre piste est de chercher à remplacer les coûteux mécanismes d'auto-attention en les remplaçant par exemple par des transformées de fourier (Lee-Thorp *et al.*, 2021).

— *Qu'est-ce qui caractérise les aspects du sens* que les réseaux profonds sont capables de capturer et ceux qui leur sont moins accessibles, voire qu'ils sont incapables de capturer et pourquoi (impossibilités intrinsèques, manque de données…).

Ce dernier point nous amène vers des travaux entamés autour du dialogue et de la clarification du sens. J'avais déjà effectué une tentative avec Jean-Pierre Chevallet (MRIM, LIG, UGA) et Jibril Frej pour faire de la conduite de dialogue entre la machine et l'utilisateur pour élargir l'usage et l'efficacité des systèmes de recherche d'informations. En 2018, le manque de données nous avait convaincus de laisser de côté la partie dialogue. Avec les récentes avancées dans les systèmes de dialogues (Roller *et al.*, 2021), certaines limites seraient aujourd'hui levées. J'ai entamé de nouveaux travaux depuis début 2021 en particulier avec François Portet (GETALP, LIG, UGA) autour de la mise au point de deux applications : une première de type art science pour créer des systèmes de dialogue dans le cadre d'une pièce de théâtre et une seconde dans le cadre d'une application de santé féminine.

Je pense que ces recherches peuvent s'inscrire dans une recherche à bien plus

long terme autour du sens et de l'entendement, la faculté de compréhension des personnes. Les systèmes de dialogue peuvent ainsi être revisités dans le cadre de la Communication Alternative et Augmentée où le dialogue se réalise généralement entre une personne et une personne privée de la parole. On peut, par exemple, imaginer un jour la machine capable d'augmenter le dialogue en aidant l'autre à analyser les états mentaux de son interlocuteur. Cette augmentation pourrait se faire à l'aide de synthèse vocale, de pictogrammes, de texte ou également d'émojis.

Il va sans dire que de telles possibilités posent des questions éthiques fondamentales et que la notion de déconnexion de tels outils par une des deux personnes est essentielle.

Avec cet exemple, on touche ce qu'est, pour moi, la CAA. C'est un domaine quelque part à l'interface entre les neurosciences, l'intelligence artificielle, et le Traitement Automatique des Langues et de la Parole.

Dans mes recherches, je reste, pour le moment, au niveau langagier. Je participe à la conception de systèmes intégrant de l'analyse de la parole, de texte, de pictogrammes, et à la conception de systèmes produisant des séquences de pictogrammes, du texte certainement demain des émojis et, à une moindre mesure, de la parole puisque j'utilise des outils de synthèse vocale développés par d'autres.

Si l'interaction avec les systèmes peut se faire par le regard, par exemple, pour sélectionner des pictogrammes, le regard n'est pas directement analysé et interprété. Il s'agit peut-être de la clé à très long terme. Étudier les données communicationnelles et pas seulement les données langagières. En d'autres termes, utiliser l'ensemble du contexte, des informations informations pour la clarification du sens, de l'entendement, de la compréhension des autres : étudier ainsi **la représentation, l'acquisition et l'exploitation du sens pour et par la clarification des données communicationnelles.**

Annexe A

Glossaire

Note : Les définitions suivantes ont été rassemblées originellement pour ma thèse Schwab (2005) et régulièrement mises à jour jusqu'à ce document.

Adaptateur : Les adaptateurs sont de petits modules rajoutés entre les couches transformer d'un modèle neuronal. Lors de l'apprentissage, seuls les adaptateurs sont mis à jour. Ainsi, peu de paramètres sont rajoutés au modèle et le nombre de paramètres à mettre à jour est nettement plus petit. En revanche, le modèle résultat comporte plus de paramètres et en phase d'exploitation, l'inférence est nettement plus longue.

AFSA : Association Française du syndrome d'Angelman.

Aidants : Personne de l'entourage d'une personne en situation de handicap.

Allophone : Personne dont la ou les langues maternelles ne sont pas les langues officielles du territoire où elle se trouve.

Apprentissage automatique : Acquisition et exploitation par des machines de lois générales établies à partir de l'observation de cas particuliers. Ces cas particuliers sont des données (dîtes d'entraînement) qui peuvent être annotées (apprentissage supervisé) ou non-annoté (apprentissage non-supervisé). Les lois générales sont constituées des structures et des opérations réalisables sur ces structures. Ces lois sont appelées modèles, l'acquisition est appelée induction, l'exploitation est appelée inférence.

Apprentissage auto-supervisé : Ces approches utilisent des méthodes d'appren-

tissage qui sont appelées, depuis quelques années, les méthodes auto-supervisées. L'idée est soit (1) de dégrader les données originelles, soit (2) d'exploiter une partie des données pour en deviner une autre. On constitue ainsi des paires dont on peut se servir avec des algorithmes supervisés.

Dans le premier cas, on constitue des paires (données originelles - données dégradées), on met alors en entrée les données dégradées, et on essaye de retrouver les données originelles en sortie. Par exemple :

— en masquant un ou plusieurs mots dans une phrase et en cherchant à les retrouver ;
— en prédisant la phrase suivante ;
— en supprimant des espaces et en cherchant à retrouver le segment textuel initial ;
— en supprimant des caractères et en cherchant à retrouver le segment textuel initial ;
— en modifiant (suppression, ajout, échange) aléatoirement des mots et en cherchant à retrouver le segment textuel.

Dans le second cas, une autre possibilité est de constituer des paires exploitant le corpus originel. Il s'agit alors de :

— prédire le mot w précédant une séquence de mots ;
— prédire le mot w suivant une séquence de mots ;
— prédire la phrase Y précédant la phrase X ;
— prédire la phrase Y suivant la phrase X ;

Apprentissage supervisé : Apprentissage automatique qui utilise des données d'entraînement annotés pour créer un modèle. En phase d'exploitation, le modèle prédit à partir de données non étiquetées, les étiquettes les plus pertinentes.

Apprentissage non-supervisé : Apprentissage automatique qui cherche à découvrir des motifs récurrents dans des données non annotées pour en induire des étiquettes. En phase d'exploitation, le modèle prédit à partir d'autres données non étiquetées, les étiquettes les plus pertinentes.

AugCom : AugCom est un outil de grilles de Communication Alternative et Augmenté libre et en sources ouvertes, basé sur des études comparatives des différents

logiciels existants, sur les besoins exprimés par les aidants et thérapeuthes. Le fondement principal de ces outils repose sur la génération vocale réalisée à partir de pictogrammes : elle permet de composer un message à partir d'un ensemble de pictogrammes afin de les associer entre eux pour qu'une synthèse vocale énonce le message au destinataire. Ce type d'outil permet de créer et d'exploiter des grilles de communication. AugCom a pour vocation de lire tout les types de fichier de grilles de communication et permet d'enregistrer dans le format OPFG (*Open Pictogram Grid Format*), un format ouvert basé sur json*.

Acception : Une acception est un sens particulier d'un mot, admis et reconnu par l'usage. Il s'agit d'une unité sémantique propre à une langue donnée (Sérasset et Mangeot, 2001). Par exemple, le terme ‹*botte*› possède au moins trois acceptions, la ‹*chaussure*›, l'‹*amas de paille*› et le ‹*coup porté en escrime*›. Les acceptions sont donc monosémiques.

Affixe : cf. *morphème*

Aire sémantique : L'aire sémantique d'un item lexical est l'ensemble des significations qu'il est susceptible d'avoir.

Antidictionnaire : Un antidictionnaire est une liste de mots qui doivent être ignorés car considérés comme non pertinents dans le cadre d'une certaine application. Ainsi, une telle liste, dans le cadre d'une application visant la sémantique, comprendra les mots vides de sens comme les mots outils (pronoms, articles, . . .), dans un cadre distributionnel, les mots trop fréquents dans le corpus. (anglais : *stop-list*)

Antonymie : Deux items lexicaux sont en relation d'antonymie si on peut exhiber une symétrie de leurs traits sémantiques par rapport à un axe.

Anaphore : Un mot à valeur anaphorique ne peut être interprété que lorsqu'il est mis en relation avec un autre élément de l'énoncé. Par exemple, dans « *En ce moment, le second attira de nouveau l'attention du capitaine. Celui-ci suspendit sa promenade et dirigea sa lunette vers le point indiqué.* », ‹‹celui-ci›› est une anaphore de ‹‹capitaine››.

Apraxie : Difficulté de planification et d'exécution des gestes en l'absence de perturbation des mécanismes qui président à la motricité élémentaire (définiton de (Fuchs, 2017)).

Archisémène : Dans une analyse sémique, l'archisémène est l'objet lexicalisé ou non dont les traits sémantiques sont l'intersection mathématique des traits des sémènes étudiés (Nyckees, 1998, p. 211).

Bain langagier : Ensemble des productions de communication auxquelles est confrontée une personne quel que soit leur mode. Il s'agit classiquement de la parole mais également des gestes, des signes ou des pictogrammes.

Base de la collocations : cf. *collocation*.

Base lexicale (ou base de données lexicale) : Une base lexicale est une source de connaissances structurées qui contient des informations sur les objets lexicaux d'une ou plusieurs langues et qui est accessible par logiciel. On retrouve ainsi parmi les bases lexicales des dictionnaires électroniques, des thésaurus, des ontologies, des graphes de connaissances. Le *Princeton WordNet* , DBNary ou encore l'ensemble des ressources du *Linguistic Linked Open Data cloud* sont considérés comme des bases lexicales.

BATX : Acronyme désignant les 4 entreprises géantes du Web chinois : Baidu, Alibaba, Tencent et Xiaomi.

CAA : Voir Communication Alternative et Augmentée.

Catégorie grammaticale : cf. *nature*

Coévolution : Terme provenant de la théorie de l'évolution dans laquelle deux espèces évoluent de manière conjointes. Une première s'adaptant à la seconde qui dans le même temps s'adapte en retour à la première. De manière plus générale et métaphorique car les processus de mise en œuvre sont bien différents, se dit de deux entités ou organisations qui s'adaptent l'une à l'autre. On peut citer la coévolution ressources-modèles ou la coévolution humain-machine dans laquelle l'humain s'adapte à la machine et la machine s'adapte à l'humain.

Collocatif : cf. *collocation*.

Collocation : L'énoncé AB (ou BA) formé des items lexicaux A et B est une collocation si, pour produire cette expression, le locuteur sélectionne A librement d'après son sens alors qu'il sélectionne B pour exprimer un autre sens en fonction de A (Polguère, 2003). On appelle A *base de la collocation* et B *collocatif*. On peut

citer comme exemples de collocations en français : ‹*tir*›$_{[=A]}$ ‹*nourrit*›$_{[=B]}$, ‹*peur*›$_{[=A]}$ ‹*bleue*›$_{[=B]}$, ‹*forte*›$_{[=B]}$ ‹*fièvre*›$_{[=A]}$,‹*dormir*›$_{[=A]}$ ‹*profondément*›$_{[=B]}$.

Comités d'éthique : comité de réflexion et d'analyse, donnant des avis éclairés sur des recherches ayant lien avec la personne humaine, tant au niveau des moyens mis en œuvre pour mener celles-ci, qu'à leurs conséquences potentielles à court ou long terme, pour qu'elles se conforment aux recommandations légales et aux consensus moraux.

Communication Alternative et Augmentée (CAA) : Selon l'ASHA (*American Speech-Language-Hearing Association* [1]), la Communication Alternative et Augmentée répond aux besoins des individus avec des troubles de communication importants et complexes caractérisés par des déficiences de la parole, que ce soit en production ou en compréhension.

On dit que la CAA est :

— *augmentée* lorsqu'elle est utilisée pour compléter un langage préexistant ;

— *alternative* lorsqu'elle est utilisée en remplacement d'un langage non existant ou dysfonctionnel ;

— *temporaire* lorsqu'elle est utilisée par les patients en postopératoire ou en soins intensifs ;

— *permanente* lorsqu'elle est utilisée par une personne qui aura besoin d'une forme quelconque de CAA tout au long de sa vie.

La CAA utilise une variété de techniques et d'outils :

— objets tangibles (boutons, claviers, images plastifiées…) ;

— gestes comme ceux d'une langue des signes, et aussi des gestes associés au discours ou à des images comme pour le Makaton ;

— pictogrammes, images correspondant à un ou plusieurs mots ;

— logiciels de synthèse vocale qui permettent de synthétiser la parole à partir de texte ou de pictogrammes ;

— oculomètres (eye-trackers) qui permettent aux personnes qui ne peuvent pas utiliser leurs membres d'interagir avec un ordinateur.

Componentielle (linguistique) : La linguistique componentielle suppose l'existence d'une atomisation de la signification, c'est-à-dire que le sens d'un terme n'est plus considéré comme primitif, mais peut être décomposé en éléments de

1. https://www.asha.org/practice-portal/professional-issues/augmentative-and-alternative-communication/

sens plus petits appelés suivant les diverses écoles : sèmes, traits sémantiques, atomes de sens, primitives, ... Par exemple, ‹*Ferrari*› peut être construit à partir des idées VOITURE, ROUGE, RAPIDE.

Compositionnalité sémantique (principe de) : D'après le principe de compositionnalité sémantique, « *le tout est calculable à partir du sens de ses parties* ». Ainsi, un énoncé est directement calculable (dans sa composition lexicale et sa structure syntaxique) à partir de la combinaison du sens de chacun des ses constituants (Polguère, 2003). Par exemple, le sens d'une phrase comme « *L'enfant voit la mer.* » est calculable à partir :

— des items lexicaux ‹*le*›, ‹*enfant*›, ‹*voir*›, ‹*la*›, ‹*mer*› ;

— des règles syntaxiques et morphologiques du français utilisées dans la phrase.

Constituants (d'une phrase) : En syntaxe, les constituants de la phrases sont les unités linguistiques qui composent la phrase : les *mots* et les *syntagmes*.

Co-texte : Le co-texte d'un mot est l'ensemble des mots qui constituent son entourage qu'ils apparaissent avant ou après dans l'énoncé. Par exemple, dans la phrase « *Une légère pente aboutissait à un fond accidenté.* », le co-texte de ‹‹pente›› est constitué des mots ‹‹une››, ‹‹légère››, ‹‹aboutissait››, ‹‹à››, ‹‹un››, ‹‹fond››, ‹‹accidenté››. Le co-texte est parfois appelé *contexte linguistique*.

Contexte : Au niveau pragmatique, la situation dans laquelle se déroule l'énoncé.

Contexte linguistique : cf. *co-texte*

Corpus : Un corpus est une collection de données langagières qui sont sélectionnées et organisées selon des critères linguistiques et extra-linguistiques explicites pour servir d'échantillon d'emplois déterminés d'une langue (définiton de (Habert *et al.*, 1998))

Délégué à la protection des données (DPO) : En droit européen, le Délégué à la protection des données (DPO, pour *Data Protection Officer*) est la personne chargée de la protection des données personnelles au sein d'une organisation [2].

Désambiguïsation sémantique : Opération qui consiste à résoudre l'ensemble des ambiguïtés posées par le sens dans un texte : ambiguïté lexicale, résolution d'anaphore, d'ellipse, ...

2. Définition inspirée de celle de Wikipedia – https://fr.wikipedia.org/wiki/Délégué_à_la_protection_des_données consultée le 26 septembre 2021

Désambiguïsation lexicale : Opération qui consiste à trouver un sens préférentiel ou une combinaison de sens préférentiel pour les mots d'un énoncé. Ainsi, dans la phrase « *L'avocat est véreux.* » deux combinaisons peuvent être considérées comme préférentielles : d'un côté *avocat*/**personne** et *véreux*/**crapuleux** et de l'autre *avocat*/**fruit** et *véreux*/**ver**.

Diachronie : cf. *synchronie*

Distance angulaire : Angle entre deux vecteur. Elle est généralement considérée comme une distance thématique dans le cas où les vecteurs correspondent à des objets textuels. Elle est calculée à partir de la similarité entre deux vecteurs.

$$\vartheta^2 \rightarrow [0, \frac{\pi}{2}] : \quad D_A(X, Y) = \arccos(Sim(X, Y)$$

Distributionnelle (linguistique) : La linguistique distributionnelle considère que le sens d'un terme peut être donné par l'ensemble de ses contextes. Par exemple, la sémantique de l'item ‹*lait*› peut être décrite grâce à la liste {‹*vache*› ‹*bouteille*› ‹*fromage*› ‹*yaourt*›, ...}

Données d'entraînement : En apprentissage automatique, les données d'entraînement sont les cas particuliers qui sont observés par les machines pour établir des lois générales.

Données langagières naturelles : Par opposition aux données synthétiques, données langagières produites par des humains

Données langagières synthétiques : Par opposition aux données naturelles, données langagières produites par des machines

Écologique (condition) : Condition réelle, la plus proche possible de celle dans laquelle une personne évolue.

Efficacité : Capacité de produire un résultat.

Efficience : Rapport entre les résultats obtenus et les ressources utilisées.

Ellipse : Omission volontaire d'une partie de phrase non nécessaire à la compréhension de l'ensemble. Exemple : « *Elle marche vite, moi (je marche) lentement.* ».

Énoncé : Séquence de termes et de phrases en langue naturelle prononcée (appelée alors paroles ou énoncé oral) ou écrite (texte ou énoncé écrit) constituant un tout.

Forme canonique : La *forme canonique* d'un mot est la forme de ce mot telle qu'on peut la trouver comme entrée d'un dictionnaire par opposition à la forme fléchie. Par définition, un item lexical est donc toujours dans une forme canonique. Traditionnellement, suivant la nature de l'item, une forme particulière est choisie :

— *verbe* : à l'infinitif
— *nom* : au singulier (s'il existe)
— *adjectif* : au masculin singulier

Exemples : ‹*sourire*›, ‹*souris*›, ‹*orgue*›, ‹*orgues*›, ‹*petit*›, ...

Notons que pour les mots invariables, formes fléchie et canonique sont identiques.

FAIR : *Facebook Artificial Intelligence Research* : Centre de Recherche de Facebook. Il est implanté sur plusieurs sites dont New-York, Paris ou Menlo Park (Californie).

FAIR data : Le terme Fair data regroupe tout un ensemble de bonnes pratiques pour la construction, le stockage, la présentation et la publication des données afin de faire en sorte qu'elles soient Facile à trouver, Accessible, Interopérable et Réutilisable.

FairSeq : Fairseq est une boîte à outils de modélisation de séquences pour le TALP comme par exemple, la traduction automatique de la langue et de la parole, le résumé automatique, la modélisation du langage.

Fonctions lexicales pour l'analyse : Permettent de modéliser les connaissances du monde et celles qui permettent de modéliser les connaissances lexicales (Schwab, 2005, p. 196). Il existe deux types de fonctions lexicales, les fonctions lexicales paradigmatiques qui formalisent les relations sémantiques qui existent entre les items lexicaux (synonymie, antonymie, méronymie, hyperonymie, ...) et les fonctions lexicales syntagmatiques c'est-à-dire les fonctions lexicales qui modélisent les combinaisons d'items lexicaux qui prévalent sur d'autres sans qu'il ne semble n'y avoir de motif logique (« *dormir profondément* » et non *« *dormir intensément* »). Pour les connaissances du monde, les FL permettent par exemple de

lier les classes aux instances (Inst(‹*Homme politique*›) = ‹*François Hollande*› et Class(‹*François Hollande*›) = ‹*Homme politique*›.

Forme fléchie : Les mots sous forme *fléchie* comportent un radical et une ou plusieurs désinences. Les désinences sont les morphèmes porteurs des indications de nombre et de genre pour les noms, adjectifs et déterminants, de personnes, de temps et de mode pour les verbes. Ainsi, ‹‹lisions›› est constitué du radical *lis-* issu de l'item ‹*lire*›, de la désinence temporelle *-i-* et de la désinence personnelle *-ons* tandis que ‹‹rattes›› est lui formé par *rat* (radical) + *te* (féminin) + s (pluriel). En aucun cas, la flexion ne modifie donc la catégorie syntaxique.

Fréquence d'un terme : On appelle fréquence d'un terme (*term frequency*) le nombre de fois où ce terme apparaît, on parle aussi du *nombre d'occurrences* ou de la *fréquence d'occurrence*.

Fréquence d'occurrences : cf. *fréquence d'un terme*.

Fréquence inverse en documents : Évaluation de l'importance d'un terme dans un corpus. Plus le terme est préesent, moindre sera l'idf. Une formule souvent utilisée $log(\frac{N}{n})$ où N est le nombre total de documents du corpus et n le nombre de documents du corpus où le terme apparaît au moins une fois.

GAFAM : Acronyme désignant les 5 entreprises géantes du Web : Google, Apple, Facebook, Amazon et Microsoft.

Gazeplay : Logiciel gratuit et open-source qui regroupe plusieurs mini-jeux sérieux intégrant plusieurs interactions (gestes, regard). Cinq grands types de compétences pouvant être développées grâce à GazePlay : les compétences d'action-réaction, de sélection, de mémorisation, de littératie et de logique mathématiques.

Généralisation : En apprentissage automatique, la généralisation est la capacité à établir des lois générales à partir de l'observation de cas particuliers.

GIPSA-lab : Grenoble Images Parole Signal Automatique est une unité mixte du CNRS, de Grenoble-INP et de l'Université de Grenoble-Alpes. Il mène des recherches théoriques et appliquées sur les signaux et les systèmes.

Gloses : Informations que l'on trouve dans certains dictionnaires (en particulier de traduction ou de synonymie) pour préciser le sens d'un terme.

HCERES : Haut Conseil de l'évaluation de la recherche et de l'enseignement supérieur – autorité administrative indépendante française, chargée de l'évaluation de l'enseignement supérieur et de la recherche publique.

Holonymie : cf. *méronymie*

Homonymie : cf. *polysémie*

Human-beatbox : Le *human-beatbox* consiste à produire des percussions vocales ainsi que des imitations d'instruments de musique, comme la trompette ou la guitare.

Hyperonymie : cf. *hyponymie*.

Hyponymie : La relation d'hyponymie est la relation hiérarchique qui lie un hyponyme à un item plus général l'hyperonyme. La relation d'hyperonymie est la relation inverse. Exemples : ‹*chat*›\‹*animal*›, ‹*voilier*›\‹*bateau*›, ‹*bateau*›\‹*véhicule*›, ‹*rose*›\‹*fleur*›.

Informatique du quotidien : Informatique utilisée par la plupart des gens dans leur vie quotidienne par opposition à celle des informaticiens.

Infixe : cf. *morphème*

Induction : Création de lois générales à partir de l'observation de cas particuliers. En apprentissage automatique, il s'agit de créer des modèles à partir des données exemples.

Inférence : Exploitation de lois générales dans le but de résoudre une tâche particulière. Opération qui applique à un modèle des données et en calcule le résultat.

***In vitro* (méthode d'évaluation)** : Méthode d'évaluation qui consiste à évaluer la tâche sur un jeu de données directement destiné à l'évaluation de cette tâche.

***In vivo* (méthode d'évaluation)** : Méthode d'évaluation qui consiste à déléguer à une autre tâche l'évaluation d'un système conçu pour résoudre une tâche. On peut ainsi évaluer un système de traduction automatique dans une tâche de recherche d'information translingue ou un système de désambiguïsation lexicale dans une tâche de traduction automatique. Les méthodes d'évaluation *in vivo* reposent sur

l'hypothèse que toutes choses sont égales par ailleurs et donc l'évaluation de l'application constitue un indice de la performance. Cette hypothèse est généralement raisonnable mais la manipulation de données qui peut être entraînée par l'intégration dans l'autre tâche peut réduire artificiellement la performance d'un système.

Item lexical : Un item lexical est une suite de caractères formant une unité sémantique et pouvant constituer une entrée de dictionnaire. Par exemple, ‹*voiture*› tout comme ‹*pomme de terre*›, ‹*moulin à vent*› et même des termes techniques comme ‹*pompe bivalve à échappement central*› sont des items lexicaux.

Interaaction (Groupe) : Groupe de recherche et de formation sur la Communication Augmentée et Alternatives inter-disciplinaire de l'université de Grenoble Alpes constitué autour de Marion Dohen (Grenoble INP-UGA, GipsaLab), Benjamin Lecouteux, Amélie Rochet Capellan (CNRS-GipsaLab) et Didier Schwab (UGA, GETALP) - voir `http://interaaction.com`

InteraactionScene : Logiciel interactif et configurable de scènes visuelles pour apprendre le vocabulaire de base aux enfants, tout en faisant un premier pas vers la Communication Alternative et Augmentée.

Interopérabilité : L'interopérabilité est la capacité d'éléments matériels ou logiciels à échanger des données et à utiliser mutuellement les données échangées par le recours à des standards ouverts de communication Duponchelle (2015). Cette définition implique que les spécifications des interfaces de ces éléments sont intégralement connues et peuvent ainsi fonctionner avec d'autres sans restriction d'accès ou de mise en œuvre.

JSON (*JavaScript Object Notation*) : Format de fichier textuel ouvert, très léger et facilement lisible par un humain originellement créé pour le langage de programmation JavaScript. Il est devenu *de facto* un format d'échange standard en informatique en remplaçant XML (*Extensible Markup Language*) bien plus verbeux.

Kaldi : Boite à outils de reconnaissance de la parole.

Langage naturel (ou langue naturelle) : langage tel qu'il est parlé quotidiennement par les êtres humains et qu'ils ont créé de façon émergente (comme le français, l'anglais, le chinois ou le malais) par opposition aux langages artificiels construits de façon consciente par l'être humain et utilisés en logique, mathéma-

tiques ou informatique.

Lexie : Objet lexical correspondant à un sens d'un item lexical dans un certain dictionnaire. Ainsi, dans Larousse (2022) pour ‹*canard*›, on peut trouver 7 lexies (**oiseau**, **fausse note**, **sucre**, **aviation**, **fausse nouvelle**, **journal**) tandis que dans Robert (2020) on en trouve 6 (**oiseau**, **marche**, **sucre**, **fausse note**, **fausse nouvelle**, **terme d'affection**)

LIRMM : Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier.

Locuteur natif : Personne qui parle la langue dont il est question comme langue maternelle.

Locution : Les locutions sont les items lexicaux constitués d'un groupe de mots figé. On distingue :

- les locutions *adjectivales* : ‹*mezza voce*›,‹*quel que*›, ...
- les locutions *adverbiales* : ‹*grosso modo*›, ‹*en partance*›, ...
- les locutions *prépositionnelles* : ‹*en ce qui concerne*›, ‹*au profit de*›, ...
- les locutions *nominales* : ‹*pomme de terre*›, ‹*moulin à vent*›, ...
- les locutions *verbales* : ‹*aller à l'encontre de*›, ‹*se faire marcher sur les pieds*›, ‹*retirer une épine du pied*›, ...

Makaton : Méthode de communication qui permet un encodage de l'information à la fois par la voix, le signe et le pictogramme. L'individu a donc accès à un input multimodal et va pouvoir s'appuyer sur plusieurs médias pour construire du sens.

Méronymie : La relation de méronymie est la relation hiérarchique qui lie la partie au tout. Un des éléments de la relation est une partie de l'autre élément. Les deux relations sont symétriques c'est-à-dire que le tout est l'holonyme de la partie tandis que la partie est le méronyme du tout.

MIASHS : Master Mathématiques et Informatique Appliquées aux Sciences Humaines et Sociales de l'UFR Sciences Humaines et Sociales de l'Université Grenoble Alpes.

Mode langagier : façon de produire du langage naturel : texte manuscrit, texte imprimé, texte informatisé, parole, gestes, signes, pictogrammes...

Modèle : En science, un modèle permet d'expliquer, de reproduire ou encore d'estimer un phénomène particulier (la gravité par exemple). En TALP, un modèle permet de reproduire un phénomène particulier (comme la résolution d'une tâche). Parfois, il s'agit d'expliquer le phénomène tel qu'il est naturellement réalisé mais c'est un cas d'application particulièrement rare en TALP où les chercheurs ne prétendent généralement pas utiliser les mêmes processus que ceux mis en œuvre par l'humain. Un modèle repose sur des données langagières, des structures de données et des opérations appliquées à ces structures.

Modèles pré-entraînés : Modèles entraînés à résoudre une tâche et destinés à être utilisés par une autre tâche (transfert d'apprentissage). Ces modèles peuvent être mis à disposition de la communauté et réduire le coût d'entrée. Ces modèles peuvent être généralistes ou s'attaquer à une tâche particulière (traduction automatique, désambiguïsation lexicale, reconnaissance de la parole)

Monosème : cf. *monosémie*

Monosémie : Caractéristique des items qui n'ont qu'un seul sens. Ainsi, des termes comme ‹*calame*›, ‹*cajou*›, ‹*neuroleptique*›, ‹*polyamide*› semblent n'avoir qu'une seule signification. On parle alors d'item *monosémique* ou *monosème*

Monosémique : cf. *monosémie*

Morphologie : Partie de la linguistique et du TALN qui s'intéresse aux morphèmes des mots.

Morphologie d'un item lexical : La morphologie d'un item lexical regroupe les informations concernant sa nature et son genre. Ainsi, ‹*courir*› est un *verbe*, ‹*souris*› est un *nom féminin* , ‹*orgues*› est un *nom masculin pluriel*…

Morphème : Les morphèmes sont les unités minimales significatives qui constituent les mots. Par exemple, le mot ‹‹fleurs›› est constitué de deux morphèmes : le radical (ou base) correspondant à l'item ‹*fleur*› et du suffixe marquant le pluriel *s*. Il existe deux types de morphèmes :

— les *morphèmes lexicaux* qui correspondent aux items lexicaux ou à une légère variante ;
— les *morphèmes grammaticaux*, autrement appelés *affixes*. Situé avant le radical, un affixe est dit *préfixe*, après le radical, il est dit *suffixe* et dans le radical, *infixe*.

MOSIG : *Master of Science in Informatics in Grenoble.*.

Mot : Un mot est la forme fléchie d'un item lexical.

NATU : Les 4 sociétés entreprises dont le taux de croissante depuis les années 2010 laissent penser qu'elles pourraient challenger les GAFAM : Netflix, Airbnb, Tesla et Uber

Nature (mot) : Les termes de même nature se caractérisent par la possibilité de les substituer syntaxiquement. Elle est constituée, entre autres, des *verbes*, des *adjectifs*, des *noms*, des *adverbes*. La nature est aussi appelée parfois *catégorie grammaticale* ou *partie du discours*.

Nombre d'occurrences : cf. *fréquence d'un terme*

Oculométrie : Ensemble de techniques permettant d'analyser le regard.

Objet lexical : La classe des *objets lexicaux* regroupe l'ensemble des objets du lexique : *item lexical*, *lexie*, *acception*.

Objet linguistique : La classe des *objets linguistiques* regroupe l'ensemble des objets de la langue : les *objets lexicaux* mais aussi les *segments textuels*.

OPFG (*Open Pictogram Grid Format*) : Format de fichier textuel ouvert, très léger et facilement lisible par un humain utilisé dans AugCom pour l'export des grilles de communication. Il permet d'exporter l'ensemble des informations des grilles (organisation, interaction, images, prononciations…).

Ontonote : Corpus annoté libre de droit en anglais, arabe et chinois.Les annotations concernent les informations structurelles (syntaxe et structures prédicat-arguments) et sémantiques superfi- cielles (sens du mot lié à une ontologie et coréférence)

Oxymore : On appelle *oxymore* ou *oxymoron* le rapprochement de termes qui semblent contradictoires comme c'est le cas pour ‹*mort-vivant*› ou ‹*clair-obscur*›.

Paradigmatique (plan) : Le plan paradigmatique, ou plan du sens, est le plan dans lequel les termes sont unis par leur sens à l'intérieur du lexique. Ces liens sont des relations sémantiques comme la synonymie, l'antonymie ou l'hyperonymie.

Paradigmatiquement, ‹*peur*› est, par exemple, relié à ‹*peureux*›, ‹*frayeur*›, ‹*calme*›, ...Le plan paradigmatique est le plan orthogonal au plan syntagmatique.

Partie du discours : cf. *nature*. En Anglais, part of speech (POS)

Préfixe : cf. *morphème*

Polysème : cf. *polysémie*

Polysémie : Caractéristique des items qui ont plusieurs sens entre lesquels il existe un lien. On parle alors d'item *polysémique* ou *polysème*. Habituellement, on distingue de la polysémie, l'*homonymie* qui est la caractéristique des items qui ont plusieurs sens entre lesquels il n'existe pas de lien. Dans cette thèse nous utilisons le qualificatif de polysémique pour un terme sans chercher à distinguer s'il s'agit d'un vrai cas de polysémie ou d'un cas d'homonymie.

Polysémique : cf. *polysémie*

Phonèmes : Segments phoniques minimaux dont la fonction est de constituer les signifiants et de les distinguer entre eux dans une langue parlée donnée. les sons interchangeables dans une langue sans changer le sens d'un énoncé ne forme qu'un seul phonème. Le français, par exemple, comprend 36 phonèmes (16 voyelles et 20 consonnes). Les phonèmes sont notés habituellement par des lettres placées entre des barres obliques : a, â, an, b, ch, d. Les items ‹*pou*› [pu] et ‹*cou*› [ku] diffèrent par le phonème p et k

Pragmatique : La pragmatique est l'étude du sens des énoncés en contextes c'est-à-dire l'ensemble des significations que peut lui donner un être humain. Le niveau pragmatique de la compréhension de textes consiste ainsi à découvrir le bon sens d'un énoncé en fonction des conditions situationnelles et contextuelles dans lesquelles il apparaît. La pragmatique s'occupe en particulier des problèmes d'anaphore, de subjectivité.

Pré-neuronal : Datant d'avant la mode du tout neuronal.

Produit terme à terme normalisé : Soient X et Y deux vecteurs, leur *produit terme à terme normalisé* V est défini par : $\vartheta^2 \rightarrow \vartheta \ : V = X \otimes Y \quad | \quad v_i = \sqrt{x_i y_i}$ L'opérateur $\otimes$ peut être interprété comme un opérateur d'intersection entre vecteurs. Si l'intersection entre deux vecteurs est le vecteur nul, alors ils n'ont rien

en commun. Du point de vue des vecteurs conceptuels, cette opération permet donc de sélectionner les idées communes à un ensemble de termes.

Relations sémantiques externes : Les relations sémantiques externes (ou relations sémantiques lexicales) sont les liens sémantiques qui relient les items lexicaux entre eux. Les principales relations sémantiques externes sont la synonymie, l'antonymie, l'hyperonymie/hyponymie, l'holonymie/méronymie.

Relations sémantiques internes : Les relations sémantiques internes sont les liens sémantiques qui relient les différentes acceptions d'un même item lexical.

Relations sémantiques lexicales : cf. *relations sémantiques externes*

Ressources langagières : On distingue les ressources langagières statiques* (ou données langagières) et les ressources langagières dynamiques* (ou outils du traitement automatique des langues et de la parole).

Ressources langagières basées sur des données multimodales : données langagières qui peuvent combiner un ou plusieurs modes langagiers* ; aussi appelé corpus.

Ressources langagières basées sur des items : données langagières centrées entrées quel que soit leur mode langagier*. On y trouve les dictionnaires, les encyclopédies, les lexiques…

Ressources langagières dynamiques : Outils du traitement automatique des langues et de la parole permettant de traiter les ressources langagières statiques. On y trouve des analyseurs syntaxiques, sémantiques, de discours, d'opinion…

Ressources langagières statiques : Ensemble de données langagières pouvant aller de leur forme brute (texte brut, écriture manuscrite, signal audio) à leur forme annotée de manière plus ou moins approfondie. On peut y distinguer les ressources langagières basé sur des items* et les ressources langagières basé sur des données multimodales*.

Réutilisabilité : la réutilisabilité permet d'exploiter les données dans d'autres contextes que pour celui il avait été originellement conçu. Il s'agit ainsi d'utiliser des formats standards mais également de préciser les conditions légales du partage (licences), la manière dont elles ont été collectées ou si elles ont été générées ou produites manuellement.

Science participative : Science qui cherche à associer les citoyens intéressés directement ou indirectement par le sujet en bénéficiant de leur point de vue formel ou informel sur les fonctionnalités, les modes d'évaluation, d'acquisition et/ou l'éthique

Segment textuel : La classe des *segments textuels* regroupe les portions d'un texte ayant une unité sémantique *mots*, *syntagme*, *phrase*, *paragraphe*, *texte*, ...

Sémantique : La sémantique est l'étude du sens des énoncés.

Sens : Le sens d'une expression linguistique est la propriété qu'elle partage avec toutes ses paraphrases.

Similarité : Soit $Sim(X, Y)$ une des mesures de *similarité* entre deux vecteurs X et Y, utilisée habituellement en recherche d'informations (Salton et McGill, 1983, p. 121). Cette valeur est le cosinus de l'angle entre les deux vecteurs.

$$\vartheta^2 \to [0,1] : \quad Sim(X,Y) = \cos(\widehat{X,Y}) = \frac{X \cdot Y}{\|X\| \times \|Y\|}$$

Somme vectorielle normée : S oient X et Y deux vecteurs, leur *somme vectorielle normée* V est définie par :

$$\vartheta^2 \to \vartheta \ : V = X \oplus Y \quad | \quad V_i = \frac{X_i + Y_i}{\|X + Y\|}$$

où ϑ est l'ensemble des vecteurs conceptuels, V_i (resp X_i, Y_i) représente la i-ème composante du vecteur V (resp. X, Y).

La somme vectorielle normée de deux vecteurs donne un vecteur équidistant en termes d'angle des deux premiers vecteurs. Il s'agit en fait d'une moyenne des vecteurs sommés. En tant qu'opération sur les vecteurs conceptuels, on peut donc voir la somme vectorielle normée comme l'union des idées contenues dans les termes.

Suffixe : cf. *morphème*

synchronie : L'étude synchronique (ou descriptive) de la langue s'intéresse à décrire la langue sans tenir compte des facteurs temps contrairement à l'étude dia-

chronique (ou historique) qui elle porte sur l'évolution de la langue (étymologie).

Synonymie : La synonymie est la relation sémantique qu'il existe entre deux items lexicaux qui diffèrent sur leur forme mais expriment le même sens ou un sens très proche.

Syndrome de Rett : Le syndrome de Rett est une maladie génétique rare qui se développe chez le très jeune enfant, principalement la fille, et provoque un handicap mental et des atteintes motrices sévères. C'est la première cause de polyhandicap d'origine génétique en France chez les filles [3].

Syndrome d'Angelman : Trouble sévère du développement neurologique dont l'origine est génétique. La maladie est caractérisée par une déficience mentale plus ou moins sévère, et une apparence et un comportement caractéristiques. La prévalence du SA est estimée entre 1/10 000 et 1/20 000. Des personnes tant de sexe masculin que féminin peuvent être touchées. [4].

Syntagmatique (plan) : Le plan syntagmatique est le plan dans lequel les termes sont unis à l'intérieur de la phrase. Il s'agit du plan sur lequel s'exercent les phénomènes de collocations. Ainsi, ‹*peur*› est relié à ‹*grande*›, ‹*énorme*›, ‹*avoir la peur de sa vie*›, ...Le plan syntagmatique est le plan orthogonal au plan paradigmatique.

Syntagme : cf. *syntaxe*

Syntaxe : La syntaxe étudie la manière dont les mots se combinent pour former des syntagmes et les syntagmes se combinent pour former des phrases.

TALN (Traitement Automatique du Langage naturel (ou des langues naturelles) : domaine d'étude des techniques d'analyse (compréhension) et de génération (production) automatiques d'énoncés oraux ou écrits

Taxinomie (ou taxonomie) : Le terme taxinomie signifie littéralement, la « *loi du rangement* ». Il désigne une classification systématique d'un ensemble d'éléments dans un domaine précis (taxinomie des êtres vivants, taxinomie de Flynn sur les

3. Définition issue du site https://afsr.fr/le-syndrome-de-rett/ consulté le 30 septembre 2021.

4. Définition issue du site https://www.angelman-afsa.org/le-syndrome/description/definition/definition-du-syndrome-dangelman consulté le 30 septembre 2021.

architectures informatiques, ...) ou général. Ce terme désigne aussi la science qui vise à établir de telles classifications.

Théorie de l'esprit : Faculté de se représenter les états mentaux d'un autre individu afin de comprendre son comportement.

Traduction automatique multilingue : Traduction automatique d'une langue source vers plusieurs langues cibles.

Annexe B

Quelques projets non présentés directement dans le document

B.1 La détection Automatique du plagiat

La société Compilatio [1] commercialise des solutions anti-plagiat qui reposent sur des techniques de comparaison textuelles. Dans le cadre d'une collaboration avec cette société (2014-2017), la thèse cifre de Jérémy Ferrero (Ferrero, 2017) a concerné les similarités textuelles sémantiques translingues avec pour objectif, la détection automatique du plagiat par traduction. Nous avons commencé par chercher comment évaluer les méthodes de l'état de l'art et proposé un corpus multilingue, multi-genre et multi-granularité (voir section 3.3.2). Nous avons également proposé plusieurs méthodes à l'époque état de l'art, certaines inspirées de mesures mises aux point pendant ma thèse de doctorat exploitant en particulier les parties du discours et des représentations vectorielles. Nous obtenions ainsi de meilleurs résultats que l'état de l'art dans la totalité des sous-corpus étudiés.

1. https://www.compilatio.net

Détection automatique de plagiat par traduction

Rôle Co-encadrant de thèse
Période Jérémy Ferrero (doctorant), Laurent Besacier (UGA, LIG, GETALP) et Frédéric Agnès (Société Compilatio)
Encadrement/Collaborations 2014-2017
Financement(s) principal(aux) Thèse Cifre avec la société Compilatio
Valorisation :

Ressources langagières Corpus multilingue, multi-genre et multi-granularité pour l'évaluation de la détection du plagiat trans-lingue [a]

Modèles Internes à l'entreprise

Références Ferrero (2017); Ferrero *et al.* (2016, 2017a,c); Ferrero (2017); Billah Nagoudi *et al.* (2017b,a,c)

Notes Première place à la campagne de similarité textuelle sémantique (STS 2017)

a. https ://github.com/FerreroJeremy/Cross-Language-Dataset

B.2 La Recherche d'Information par apprentissage profond

Ce travail s'inscrit dans le cadre d'une collaboration avec l'équipe MRIM [2] du LIG autour de la recherche d'information textuelle. Depuis ma thèse, je défends l'idée que les représentation vectorielles basées sur le corpus ne sont pas suffisantes dans le cadre de l'analyse sémantique [3] et que l'ajout de connaissances externes est indispensable. Dans ce projet, nous avons étudié les moyens d'intégrer de telles ressources à des réseaux de neurones afin de palier outre cette représentation du texte basée uniquement sur l'analyse statistique, (1) la nécessité de disposer de grandes quantités de données étiquetées et (2) le manque d'efficacité des modèles. Pour ce faire, la production de ressources s'est révélée indispensable et un accent particulier a été apporté comme l'utilisation de Wikipedia pour construire des collections de recherche d'information (Wikir - voir section 3.2.1.3).

2. Modélisation et Recherche d'Information Multimédia – http://lig-mrim.imag.fr/
3. Ce sont d'ailleurs des points qu'il a fallu rappeler ces derniers temps à la partie de la communauté scientifique et industrielle qui faisait preuve d'un enthousiasme réel ou feint pour les méthodes uniquement basées sur les corpus (Bender et Koller, 2020; Bender *et al.*, 2021). Le lecteur y apprendra, si ce n'est déjà fait que la lecture d'une expérience ne remplace pas toujours l'expérience ou que répondre de manière cohérente n'assure pas que la question a bien été comprise.

Incorporation de Connaissances a priori pour la Recherche d'Information Textuelle Neuronale

Rôle Co-encadrant de thèse
Période 2017-2021
Encadrement/Collaborations Jibril Frej (doctorant), Jean-Pierre Chevallet (UGA, LIG, MRIM), Philippe Mulhem (CNRS, LIG, MRIM)
Financement(s) principal(aux) Allocation de recherche
Valorisation :

Ressources langagières WikIR78k, WikIRS78k, MLWikir [a]

Modèles Intégrés dans des librairies internes. Leur mises à disposition sera visée avec la thèse d'Ilyes Bentebib

Références Frej (2021), FREJ *et al.* (2017), FREJ *et al.* (2020), FREJ *et al.* (2018) , Frej *et al.* (2020a), Frej *et al.* (2020b)

Notes ∅

a. https://github.com/getalp/wikIR

Annexe C

Bibliographie personnelle

La stratégie de publication est axée sur les **conférences/workshops majeur(e)s** lié(e)s aux domaines énoncés :

— **Traitement Automatique des Langues et de la Parole** : TALN (conférence francophone réputée), Coling, *ACL, Interspeech (conférences anglophones très réputées)

— **Corpus/logiciels** : LREC (diffusion des ressources, fort impact dans le domaine TALN)

— **Communication Alternative et Augmentée** : ICCHP, SIGACCESS (conférences internationales réputées dans le domaine de l'accessibilité)

Sur l'ensemble des papiers présentés, j'ai contribué aux idées, encadrements, développement des systèmes ainsi qu'aux expérimentations. Les publications sont ordonnées suivant la perspective adoptée dans ce mémoire selon qu'elles relèvent du Traitement Automatique des Langues et de la Parole, de la Communication Alternative et Augmentée ou des deux. Les cinq publications considérées comme prototypique de mes recherches.

L'ensemble des publications sont consultables sur le CV hal : https://cv.archives-ouvertes.fr/didier-schwab.
La liste suivante est arrêtée au 30 septembre 2021 et ne comprends pas les soumissions.

Articles liés au Traitement Automatique des Langues et de la parole

[1] Didier Schwab, Mathieu Lafourcade, Violaine Prince. Vers l'Apprentissage Automatique pour et par les Vecteurs Conceptuels de Fonctions Lexicales – L'exemple de l'Antonymie. TALN : Traitement Automatique des Langues Naturelles, 2002, Nancy, France.

[2] Mathieu Lafourcade, Violaine Prince, Didier Schwab. Vecteurs Conceptuels et Structuration Emergent de Terminologies. Revue TAL, ATALA (Association pour le Traitement Automatique des Langues), 2002, 43 (1), pp.43–72.

[3] Mathieu Lafourcade, Violaine Prince, Didier Schwab. Vecteurs conceptuels et structuration émergente de terminologies. Revue TAL, ATALA (Association pour le Traitement Automatique des Langues), 2002, 43 (1), pp.43–72.

[4] Didier Schwab, Mathieu Lafourcade, Violaine Prince. Hypothèses pour la Construction et l'Exploitation Conjointe d'une Base Lexicale Sémantique Basée sur les Vecteurs Conceptuels. JADT'04 : Journées Internationales d'Analyse Statistique des Données Textuelles, Mar 2004, Louvain–la–Neuve, pp.1008–1018.

[5] Mathieu Lafourcade, F. Rodrigo, Didier Schwab. Low Cost Automated Conceptual Vector Generation From Mono and Bilingual Resources. PAPILLON'04, Aug 2004, Grenoble (France), pp.10.

[6] Mathieu Lafourcade, Didier Schwab. Estimation Automatique de la Distribution des Sens de Termes. INFORSID'05 : INFormatique des Organisations et Systèmes d'Information et de Décision, May 2005, Grenoble (France)

[7] Didier Schwab, Mathieu Lafourcade, Violaine Prince. Extraction Semi–Supervisée de Couples d'Antonymes grâce à leur Morphologie. TALN : Traitement Automatique des Langues Naturelles, Jun 2005, Dourdan, France.

[8] Didier Schwab. Approche hybride – lexicale et thématique – pour la modélisation, la détection et l'exploitation des fonctions lexicales en vue de l'analyse sémantique de texte. Interface homme–machine [cs.HC]. Université Montpellier II – Sciences et Techniques du Languedoc, 2005. Français.

[9] Alain Joubert, Mathieu Lafourcade, Didier Schwab. Approche évolutive des notions de base pour une représentation thématique des connaissances générales. TALN : Traitement Automatique des Langues Naturelles, Apr 2006, Louvain, Belgique. pp.512–521.

[10] Didier Schwab, Mathieu Lafourcade. Modelling, Detection and Exploita-

tion of Lexical Functions for Analysis. ECTI–CIT Transactions on Computer and Information Technology, ECTI, 2006, 2 (2), pp.97–108.

[11] Didier Schwab, Lim Lian Tze, Mathieu Lafourcade. Les vecteurs conceptuels, un outil complémentaire aux réseaux lexicaux. TALN : Traitement Automatique des Langues Naturelles, Jun 2007, Toulouse, France. pp.293–302.

[12] Didier Schwab, Lim Lian Tze, Mathieu Lafourcade. Conceptual Vectors, a complementary tool to Lexical Networks. NLPCS'07 : 4th International Workshop on Natural Language Processing and Cognitive Science, Jun 2007, pp.10.

[13] Didier Schwab, Mathieu Lafourcade. Lexical Functions for Ants Based Semantic Analysis. ICAI'07 : International Conference on Artificial Intelligence, Jun 2007, pp.7.

[14] Michael Zock, Didier Schwab. Lexical access based on underspecified input. Coling 2008 : Proceedings of the Workshop on Cognitive Aspects of the Lexicon (COGALEX 2008), 2008, Manchester, United Kingdom. pp.9—17.

[15] Didier Schwab, Mathieu Lafourcade. Hardening of Acceptions Links Through Vectorized Lexical Functions. PAPILLON'02, Tokyo, Japan

[16] Didier Schwab, Mathieu Lafourcade, Violaine Prince. Antonymy and Conceptual Vectors. COLING'02 : 19th Conference on Computational Linguistics, Taipei, Japan, pp.904–910.

[17] Didier Schwab, Mathieu Lafourcade, Violaine Prince. Amélioration de la Représentation Sémantique Lexicale par les Vecteurs Conceptuels : Le Rôle de l'Antonymie. JADT'02 : Journées Internationales d'Analyse Statistique des Données Textuelles, St Malo, France, pp.701–712.

[18] Didier Schwab, Mathieu Lafourcade, Violaine Prince. Amélioration de Liens entre Acceptions par Fonctions Lexicales Vectorielles Symétriques. TALN'03 : 10ème Conférence Internationale sur le Traitement Automatique du Langage Naturel, Batz–sur–Mer (France), pp. 235–244.

[19] Didier Schwab. Société d'Agents Apprenants et Sémantique Lexicale : Comment Construire des Vecteurs Conceptuels à l'Aide de la Double Boucle. RECITAL'03, Batz–sur–Mer (France), pp. 478–489.

[20] Lian–Tze Lim, Didier Schwab. Limits of Lexical Semantic Relatedness with Ontology–based Conceptual Vectors. 5th International Workshop on Natural Language Processing and Cognitive Science (NLPCS 2008), Jun 2008, Barcelone, Spain.

[21] Lian Lim, Didier Schwab. Limits of Lexical Semantic Relatedness with

Ontology–based Conceptual Vectors. NLPCS 2008 : Natural Language Processing and Cognitive Science, Jun 2008, Barcelone, Spain.

[22] Didier Schwab, Lian Lim. Blexisma2 : a Distributed Agent Framework for Constructing a Semantic Lexical Database based on Conceptual Vectors. DFMA 2008 International Conference on Distributed Framework and Applications, Oct 2008, Penang, Malaysia.

[23] Didier Schwab. Introduction aux vecteurs conceptuels. Séminaire du LIRIS, école centrale de Lyon, 2009, Ecully, France.

[24] Tang Enya Kong, Lim Lian Tze, Ye Hong Hoe, Didier Schwab. Grid–enabled Blexisma2. Belaton, Bahari and Lian Tze, Lim. Grid Computing Cluster : The Development and Integration of Grid Applications and Services, Platform for Information Communication Technology Research, Universiti Sains Malaysia, pp.23—26, 2009, 9789833986583.

[25] Didier Schwab. Modélisation, Détection Et Exploitation de Fonctions Lexicales : Approches lexicales et thématique pour la modélisation, la détection et l'exploitation des fonctions lexicales en vue de l'analyse sémantique de texte. Éditions Universitaires Européennes, 2010.

[26] Michael Zock, Didier Schwab, Nirina Rakotonanahary. Lexical Access, a Search–Problem. 2nd SIGLEX endorsed COLING Workshop on Cognitive Aspects of the Lexicon, Enhancing the Structure and Look–up Mechanisms of Electronic Dictionaries, 2010, Beijing, China. pp.75–84.

[27] Achille Falaise, David Rouquet, Didier Schwab, Hervé Blanchon, Christian Boitet. Ontology driven content extraction using interlingual annotation of texts in the OMNIA project. CLIA workshop, COLING 2010, 2010, Beijing, China.

[28] Michael Zock, Olivier Ferret, Didier Schwab. Deliberate word access : an intuition, a roadmap and some preliminary empirical results. International Journal of Speech Technology, Springer Verlag, 2010, 13 (4), pp.107—117.

[29] David Rouquet, Achille Fallaise, Didier Schwab, Hervé Blanchon, Valérie Belinck, et al.. Classification multilingue et multimédia pour la recherche d'images dans le projet OMNIA. RISE'2010 : second atelier Recherche d'Information SEmantique associé à la conférence (INFORSID), May 2010, Marseille, France.

[30] Didier Schwab, Jérôme Goulian, Nathan Guillaume. Désambiguısation lexicale par propagation de mesures sémantiques locales par algorithmes à colonies de fourmis. TALN'2011 : Traitement Automatique des Langues Naturelles, 2011, Montpellier, France.

[31] Michael Zock, Didier Schwab. Storage does not Guarantee Access : The

Problem of Organizing and Accessing Words in a Speaker's Lexicon. Journal of Cognitive Science, Institute for Cognitive Science, Seoul National University, 2011, 12, pp.233–258.

[32] Michael Zock, Didier Schwab. Le problème du " mot sur le bout de la langue " et comment y remédier. Cognisciences : Journal des sciences cognitives, École Nationale Supérieure de Cognitique de Bordeaux, 2011.

[33] Didier Schwab, Nathan Guillaume. A Global Ant Colony Algorithm for Word Sense Disambiguation based on Semantic Relatedness. PAAMS 2011 : 9th Conference on Practical Applications of Agents and Multi–Agent Systems, 2011, Salamanca, Spain.

[34] Michael Zock, Didier Schwab. Si tous les chemins mènent à Rome, ils ne se valent pas tous. Le problème d'accès lexical.. Ataa Allah Fadoual. Invited paper to the 4th international workshop on Amazighe and new Technologies, IRCAM, Institut Royal, pp.1–9, 2011.

[35] Michael Zock, Didier Schwab. Et si Boileau s'était trompé ? Le problème du mot sur le bout de la langue et comment y remédier.. Cognisciences : Journal des sciences cognitives, École Nationale Supérieure de Cognitique de Bordeaux, 2011, Revue CogniScience, 7 (5—7), pp.5–7.

[36] Sandra Skaff, David Rouquet, Emmanuel Dellandrea, Achille Falaise, Valérie Bellynck, et al.. Multimodal Search for Graphic Designers. IVAPP'2011 : International Conference on Information Visualization Theory and Applications, 2011, Algarve, Portugal.

[37] Alain Joubert, Mathieu Lafourcade, Didier Schwab, Michael Zock. Évaluation et consolidation d'un réseau lexical grâce à un assistant ludique pour le " mot sur le bout de la langue ". TALN : Traitement Automatique des Langues Naturelles, Jun 2011, Montpellier, France. pp.295–306.

[38] Alain Joubert, Mathieu Lafourcade, Didier Schwab, Michael Zock. Évaluation et consolidation d'un réseau lexical via un outil pour retrouver le mot sur le bout de la langue. TALN : Traitement Automatique des Langues Naturelles, Jun 2011, Montpellier, France. pp.295–306.

[39] Jorge Mauricio Molina Mejia, Didier Schwab, Gilles Sérasset. Actes de la conférence conjointe JEP–TALN–RECITAL 2012, volume 3 : RECITAL. Molina Mejia, Jorge Mauricio and Schwab, Didier and Sérasset, Gilles. ATALA/AFCP, pp.x–x, 2012.

[40] Didier Schwab, Jérôme Goulian, Andon Tchechmedjiev, Hervé Blanchon. Ant Colony Algorithm for the Unsupervised Word Sense Disambiguation of Texts :

Comparison and Evaluation. Proceedings of the 25th International Conference on Computational Linguistics (COLING 2012), 2012, Mumbai, India.

[41] Andon Tchechmedjiev, Didier Schwab, Jérôme Goulian, Gilles Sérasset. Parameter estimation under uncertainty with Simulated Annealing applied to an ant colony based probabilistic WSD algorithm. Proceedings of the 1st International Workshop on Optimization Techniques for Human Language Technology, on behalf of (COLING 2012), 2012, Mumbai, India.

[42] Jorge Mauricio Molina Mejia, Didier Schwab, Gilles Serasset. Actes de la conférence conjointe JEP–TALN–RECITAL 2012. Volume 3 : RECITAL. Association Francophone pour la Communication Parlée (AFCP) et Association pour le Traitement Automatique des Langues (ATALA). 2012.

[43] Didier Schwab, Jérôme Goulian, Andon Tchechmedjiev. Désambiguïsation lexicale de textes : efficacité qualitative et temporelle d'un algorithme à colonies de fourmis. Revue TAL, ATALA (Association pour le Traitement Automatique des Langues), 2013, 54 (1), pp.99–138.

[44] Didier Schwab, Andon Tchechmedjiev, Jérôme Goulian, Mohammad Nasiruddin, Gilles Sérasset, et al.. GETALP System : Propagation of a Lesk Measure through an Ant Colony Algorithm. Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2 : Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013), 2013, Atlanta, Georgia, United States. pp.232—240. ;hal–00953724x27E9 ; [45] Andon Tchechmedjiev, Jérôme Goulian, Didier Schwab. Fusion strategies applied to multilingual features for an knowledge–based Word Sense Disambiguation algorithm : evaluation and comparison. Proceedings of the 14th International Conference on Intelligent Text Processing and Computational Linguistics (CiCling 2013), 2013, Samos, Greece.

[46] Michael Zock, Didier Schwab. Chercher un mot dans un dictionnaire sans bon index est un peu comme s'orienter sur une île déserte sans carte convenable. Lingvisticae Investigationes, Philadelphia ; Amsterdam : John Benjamins, 2013.

[47] Didier Schwab, Jérôme Goulian, Andon Tchechmedjiev. Worst–case Complexity and Empirical Evaluation of Artificial Intelligence Methods for Unsupervised Word Sense Disambiguation. International Journal of Web Engineering and Technology, Inderscience, 2013, 8, pp.124–153.

[48] Andon Tchechmedjiev, Jérôme Goulian, Didier Schwab. Evaluation and Comparison of multilingual fusion strategies for similarity–based Word Sense Disambiguation. Research in Computing Science, National Polytechnic Institute, 2013, pp.69–80. ;hal–00953648x27E9 ; [49] Michael Zock, Didier Schwab. L'index, une ressource vitale pour guider les auteurs à trouver le mot bloqué sur le bout de la

langue.. Nuria Gala and Michael Zock. Ressources lexicales : construction et utilisation, 10, Lingvisticae Investigationes, John Benjamins, Amsterdam, The Netherlands, pp.313—354, 2013.

[50] Michael Zock, Didier Schwab. Le fait d'avoir stocké des mots garantit nullement leur accès.. Ressources Lexicales et Traitement de la Langue, atelier TALN–2014, 2014, Marseille, France. pp.221—230.

[51] Andon Tchechmedjiev, Gilles Sérasset, Jérôme Goulian, Didier Schwab. Attaching Translations to Proper Lexical Senses in DBnary. 3rd Workshop on Linked Data in Linguistics : Multilingual Knowledge Resources and Natural Language Processing, May 2014, Reykjavik, Iceland.

[52] Benjamin Lecouteux, Didier Schwab. Décodage de graphe à l'aide de colonies de fourmis. 30èmes Journées d'étude de la parole, Jun 2014, Le mans, France. pp.6.

[53] Mohammad Nasiruddin, Didier Schwab, Andon Tchechmedjiev, Gilles Sérasset, Hervé Blanchon. Induction de sens pour enrichir des ressources lexicales. 21ème conférence sur le Traitement Automatique des Langues Naturelles (TALN 2014), Jul 2014, Marseille, France. pp.6.

[54] Didier Schwab, Andon Tchechmedjiev, Jerome Goulian, Gilles Sérasset. Comparisons of Relatedness Measures through a Word Sense Disambiguation Task. Advances in Language Production, Cognition and the Lexicon,, Springer, pp.23, 2014.

[55] Benjamin Lecouteux, Didier Schwab. Ant Colony Algorithm Applied to Automatic speech Recognition Graph Decoding. Interspeech, 2015, Dresde, Germany.

[56] Mohammad Nasiruddin, Andon Tchechmedjiev, Hervé Blanchon, Didier Schwab. Création rapide et efficace d'un système de désambiguïsation lexicale pour une langue peu dotée. 22ème conférence sur le Traitement Automatique des Langues Naturelles, Jun 2015, Caen, France.

[57] Benjamin Lecouteux, Didier Schwab. AntT Colony Algorithm applied to Automatic Speech recognition Graph Decoding. Interspeech 2015, Sep 2015, Dresden, Germany.

[58] Bakhouche Abdelaali, Yamina Tlili–Guiassa, Didier Schwab, Andon Tchechmedjiev. Ant Colony Algorithm for Arabic Word Sense Disambiguation through English lexical information. International Journal of Metadata, Semantics and Ontologies, Inderscience, 2015, 10 (3), pp.202–211.

[59] Michael Zock, Didier Schwab. WordNet and beyond.. Christiane Fellbaum

and Piek Vossen and Verginica Mititelu and Corina Forăscu. 8th International Global WordNet conference (http ://gwc2016.racai.ro), pp.436–444, 2016.

[60] Jérémy Ferrero, Frédéric Agnès, Laurent Besacier, Didier Schwab. A Multilingual, Multi–Style and Multi–Granularity Dataset for Cross–Language Textual Similarity Detection. 10th edition of the Language Resources and Evaluation Conference, May 2016, Portorož, Slovenia.

[61] Marwa Hadj Salah, Hervé Blanchon, Mounir Zrigui, Didier Schwab. Amélioration de la traduction automatique d'un corpus annoté. JEP–TALN–RECITAL 2016, Jul 2016, Paris, France.

[62] Loïc Vial, Andon Tchechmedjiev, Didier Schwab. Extension lexicale de définitions grâce à des corpus annotés en sens. 23ème Conférence sur le Traitement Automatique des Langues Naturelles, Jul 2016, Paris, France.

[63] Loïc Vial, Benjamin Lecouteux, Didier Schwab. Représentation vectorielle de sens pour la désambiguïsation lexicale à base de connaissances. TALN, 2017, Orléans, France.

[64] Loïc Vial, Benjamin Lecouteux, Didier Schwab. Sense Embeddings in Knowledge–Based Word Sense Disambiguation. IWCS, 2017, Montpellier, France.

[65] El Moatez Billah Nagoudi, Didier Schwab. Semantic Similarity of Arabic Sentences with Word Embeddings. Third Arabic Natural Language Processing Workshop, Apr 2017, Valencia, France. pp.18 – 24.

[66] Jérémy Ferrero, Frédéric Agnès, Laurent Besacier, Didier Schwab. Using Word Embedding for Cross–Language Plagiarism Detection. EACL 2017, Apr 2017, Valence, Spain. pp.415 – 421.

[67] Loïc Vial, Benjamin Lecouteux, Didier Schwab. Uniformisation de corpus anglais annotés en sens. 24ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN), Jun 2017, Orléans, France.

[68] Loïc Vial, Benjamin Lecouteux, Didier Schwab. Représentation vectorielle de sens pour la désambiguïsation lexicale à base de connaissances. 24ème Conférence sur le Traitement Automatique des Langues Naturelles, Jun 2017, Orléans, France.

[69] El Moatez Billah Nagoudi, Jérémy Ferrero, Didier Schwab. Amélioration de la similarité sémantique vectorielle par méthodes non–supervisées. 24e conférence sur le Traitement Automatique des Langues Naturelles (TALN 2017), Jun 2017, Orléans, France.

[70] Loïc Vial, Benjamin Lecouteux, Didier Schwab. Uniformisation de corpus

anglais annotés en sens. 24ème Conférence sur le Traitement Automatique des Langues Naturelles, Jun 2017, Orléans, France.

[71] Jibril Frej, Jean–Pierre Chevallet, Didier Schwab. Enhancing Translation Language Models with Word Embedding for Information Retrieval. 9ème Atelier Recherche d'Information SEmantique, Jul 2017, Caen, France.

[72] El Moatez Billah Nagoudi, Jérémy Ferrero, Didier Schwab. LIM–LIG at SemEval–2017 Task1 : Enhancing the Semantic Similarity for Arabic Sentences with Vectors Weighting. International Workshop on Semantic Evaluations (SemEval–2017), Aug 2017, Vancouver, Canada. pp.125 – 129.

[73] Jérémy Ferrero, Laurent Besacier, Didier Schwab, Frédéric Agnès. CompiLIG at SemEval–2017 Task 1 : Cross–Language Plagiarism Detection Methods for Semantic Textual Similarity. Proceedings of the 11th International Workshop on Semantic Evaluations (SemEval–2017),, Aug 2017, Vancouver, Canada.

[74] Jérémy Ferrero, Laurent Besacier, Didier Schwab, Frédéric Agnès. Deep Investigation of Cross–Language Plagiarism Detection Methods. BUCC, 10th Workshop on Building and Using Comparable Corpora, Aug 2017, Vancouver, Canada.

[75] Loïc Vial, Benjamin Lecouteux, Didier Schwab. Sense Embeddings in Knowledge–Based Word Sense Disambiguation. 12th International Conference on Computational Semantics, Sep 2017, Montpellier, France.

[76] El Moatez Billah Nagoudi, Jérémy Ferrero, Didier Schwab, Hadda Cherroun. Word Embedding–Based Approaches for Measuring Semantic Similarity of Arabic–English Sentences. 6th International Conference on Arabic Language Processing, Oct 2017, Fez, Morocco.

[77] Amel Ziani, Nabiha Azizi, Didier Schwab, Monther Aldwairi, Nassira Chekkai, et al.. Recommender System Through Sentiment Analysis. 2nd International Conference on Automatic Control, Telecommunications and Signals, Dec 2017, Annaba, Algeria.

[78] Loïc Vial, Benjamin Lecouteux, Didier Schwab. UFSAC : Unification of Sense Annotated Corpora and Tools. LREC, 2018, Miyazaki, Japan.

[79] El Moatez Billah Nagoudi, Ahmed Khorsi, Hadda Cherroun, Didier Schwab. A Two–Level Plagiarism Detection System for Arabic Documents. Cybernetics and Information Technologies, In press, 18 (1).

[80] Loïc Vial, Benjamin Lecouteux, Didier Schwab. UFSAC : Unification of Sense Annotated Corpora and Tools. Language Resources and Evaluation Conference (LREC), May 2018, Miyazaki, Japan.

[81] Marwa Hadj Salah, Hervé Blanchon, Mounir Zrigui, Didier Schwab. Un corpus en arabe annoté manuellement avec des sens WordNet. 25e conférence sur le Traitement Automatique des Langues Naturelles, May 2018, Rennes, France.

[82] Didier Schwab. Enjeux et perspectives de la recherche d'information sémantique – Ressources lexicales. Recherche d'Information Semantique 2018, May 2018, Rennes, France.

[83] Jibril Frej, Philippe Mulhem, Didier Schwab, Jean–Pierre Chevallet. Combining Subword information and Language model for Information Retrieval. 15e Conférence en Recherche d'Information et Applications, May 2018, Rennes, France.

[84] Loïc Vial, Benjamin Lecouteux, Didier Schwab. Approche supervisée à base de cellules LSTM bidirectionnelles pour la désambiguïsation lexicale. 25e conférence sur le Traitement Automatique des Langues Naturelles, May 2018, Rennes, France.

[85] Marwa Hadj Salah, Loïc Vial, Hervé Blanchon, Mounir Zrigui, Benjamin Lecouteux, et al.. La désambiguïsation lexicale d'une langue moins bien dotée, l'exemple de l'arabe. 25e conférence sur le Traitement Automatique des Langues Naturelles, May 2018, Rennes, France.

[86] Didier Schwab. Cognitive Ability Estimation and Reinforcement with Eye–tracking Games for Children with Multiple Disabilities. Grenoble Workshop on Models and Analysis of Eye Movements, Jun 2018, Grenoble, France.

[87] Didier Schwab, Pierre Zweigenbaum. Actes de TALIA 2018 – Journée « Traitement Automatique des Langues Intelligence Artificielle ». 2018.

[88] Loïc Vial, Benjamin Lecouteux, Didier Schwab. Approche supervisée à base de cellules LSTM bidirectionnelles pour la désambiguïsation lexicale. Revue TAL, ATALA (Association pour le Traitement Automatique des Langues), 2019.

[89] Loïc Vial, Benjamin Lecouteux, Didier Schwab, Hang Le, Laurent Besacier. The LIG system for the English–Czech Text Translation Task of IWSLT 2019. IWSLT (16th International Workshop on Spoken Language Translation), 2019, Hong–Kong, China.

[90] Loïc Vial, Benjamin Lecouteux, Didier Schwab. Sense Vocabulary Compression through the Semantic Knowledge of WordNet for Neural Word Sense Disambiguation. Global Wordnet Conference, 2019, Wroclaw, Poland.

[91] Loïc Vial, Benjamin Lecouteux, Didier Schwab. Compression de vocabulaire de sens grâce aux relations sémantiques pour la désambiguïsation lexicale. TALN 2019 (Conférence sur le Traitement Automatique des Langues Naturelles), Jul 2019, Toulouse, France.

[92] Raki Lachraf, El Moatez Billah Nagoudi, Youcef Ayachi, Ahmed Abdelali, Didier Schwab. ArbEngVec : Arabic–English Cross–Lingual Word Embedding Model. The Fourth Arabic Natural Language Processing Workshop, Jul 2019, Florence, Italy.

[93] Solène Evain, Adrien Contesse, Antoine Pinchaud, Didier Schwab, B Lecouteux, et al.. Beatbox sounds recognition using a speech–dedicated HMM–GMM based system. MAVEBA 2019 – 11th International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications, Dec 2019, Florence, Italy.

[94] Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, et al.. FlauBERT : Unsupervised Language Model Pre–training for French. LREC, 2020, Marseille, France.

[95] Jibril Frej, Didier Schwab, Jean–Pierre Chevallet. Modèle Transformer à base de Connaissances pour la Recherche d'Information dans des Domaines Spécialisés. Extraction et Gestion des Connaissances (EGC), 2020.

[96] Solène Evain, Adrien Contesse, Antoine Pinchaud, Didier Schwab, Benjamin Lecouteux, et al.. Reconnaissance de parole beatboxée à l'aide d'un système HMM–GMM inspiré de la reconnaissance automatique de la parole. JEP–TALN–RECITAL 2020 – 6e conférence conjointe 33e Journées d'Études sur la Parole, 27e Traitement Automatique des Langues Naturelles, 22e Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues, Jun 2020, Nancy, France. pp.208–216.

[97] Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, et al.. FlauBERT : des modèles de langue contextualisés pré–entraînés pour le français. 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Volume 2 : Traitement Automatique des Langues Naturelles, Jun 2020, Nancy, France. pp.268–278.

[98] Jibril Frej, Didier Schwab, Jean–Pierre Chevallet. MLWIKIR : A Python toolkit for building large–scale Wikipedia–based Information Retrieval Datasets in Chinese, English, French, Italian, Japanese, Spanish and more. Joint Conference of the Information Retrieval Communities in Europe, Jul 2020, Toulouse, France.

[99] Jibril Frej, Jean–Pierre Chevallet, Didier Schwab. Knowledge Based Transformer Model for Information Retrieval. Joint Conference of the Information Retrieval Communities in Europe (CIRCLE 2020), Jul 2020, Samatan, France.

[100] Jibril Frej, Philippe Mulhem, Didier Schwab, Jean–Pierre Chevallet. Learning Term Discrimination. 43rd International ACM SIGIR Conference on Re-

search and Development in Information Retrieval (SIGIR 2020), Jul 2020, Xi'an, China. pp.1993–1996.

[101] Philippe Mulhem, Gabriela Gonzalez Saez, Aidan Mannion, Didier Schwab, Jibril Frej. LIG–Health at Adhoc and Spoken IR Consumer Health Search : expanding queries using UMLS and FastText. CLEF 2020, Sep 2020, Thessaloniki (on line), Greece.

[102] Hang Le, Juan Pino, Changhan Wang, Jiatao Gu, Didier Schwab, et al.. Dual–decoder Transformer for Joint Automatic Speech Recognition and Multilingual Speech Translation. COLING 2020 (long paper), Dec 2020, Virtual, Spain.

[103] Amel Ziani, Nabiha Azizi, Didier Schwab, Djamel Zenakhra, Monther Aldwairi, et al.. Deceptive Opinions Detection Using New Proposed Arabic Semantic Features. Procedia Computer Science, Elsevier, 2021, 189, pp.29 – 36.

[104] Solène Evain, Benjamin Lecouteux, Didier Schwab, Adrien Contesse, Antoine Pinchaud, et al.. Human Beatbox Sound Recognition using an Automatic Speech Recognition Toolkit. Biomedical Signal Processing and Control, Elsevier, 2021, 67.

[105] Hang Le, Juan Pino, Changhan Wang, Jiatao Gu, Didier Schwab, et al.. Lightweight Adapter Tuning for Multilingual Speech Translation. The Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL–IJCNLP 2021), Aug 2021, Bangkok (Virtual), Thailand.

[106] Zae Kim, Laurent Besacier, Vassilina Nikoulina, Didier Schwab. Do Multilingual Neural Machine Translation Models Contain Language Pair Specific Attention Heads ?. Findings of ACL 2021, Aug 2021, Bangkok (virtual), Thailand.

[107] Hang Le, Florentin Barbier, Ha Nguyen, Natalia Tomashenko, Salima Mdhaffar, et al.. ON–TRAC' systems for the IWSLT 2021 low–resource speech translation and multilingual speech translation shared tasks. International Conference on Spoken Language Translation (IWSLT), Aug 2021, Bangkok (virtual), Thailand.

[108] Solène Evain, Ha Nguyen, Hang Le, Marcely Zanon Boito, Salima Mdhaffar, et al.. LeBenchmark : A Reproducible Framework for Assessing Self–Supervised Representation Learning from Speech. INTERSPEECH 2021 : Conference of the International Speech Communication Association, Aug 2021, Brno, Czech Republic.

Articles liés à la Communication Alternative et Augmentée

[109] Didier Schwab. GazePlay : Creation of a community to help the development of a Free and Open–source plateform to make eye–tracker Video Games accessible to everyone. 5ème EUROPEAN RETT–SYNDROME CONGRESS, Nov 2017, Berlin, Germany.

[110] Didier Schwab, Amela Fejza, Loïc Vial, Yann Robert. The GazePlay Project : Overview in February 2018. [Research Report] LIG lab. 2018, pp.1–5.

[111] Didier Schwab, Amela Fejza, Loïc Vial, Yann Robert. The GazePlay Project : Open and Free Eye–trackers Games and a Community for People with Multiple Disabilities. ICCHP 2018 – 16th International Conference on Computers Helping People with Special Needs, Jul 2018, Linz, Austria. pp.254–261.

[112] Sébastien Riou, Didier Schwab, François Bérard. Interactions par franchissement grâce a un système de suivi du regard. [Research Report] Groupe d'Étude en Traduction Automatique/Traitement Automatisé des Langues et de la Parole et équipe Ingénierie de l'Interaction Humain-Machine ; LIG (Laboratoire informatique de Grenoble) ; Université Grenoble Alpes. 2019.

[113] Didier Schwab, Sébastien Riou, Amela Fejza, Loïc Vial, Johana Marku, et al.. Le projet GazePlay : des jeux ouverts, gratuits et une communauté pour les personnes en situation de polyhandicap. 1024 – Bulletin de la Société informatique de France, 2020.

[114] Nairit Bandyopadhyay, Sébastien Riou, Didier Schwab. Webcam as Alternate Option for Eye–Trackers in Gaze Gaming Software : GazePlay. [Research Report] Groupe d'Étude en Traduction Automatique/Traitement Automatisé des Langues et de la Parole ; LIG (Laboratoire informatique de Grenoble) ; Université Grenoble Alpes. 2021.

[115] Nairit Bandyopadhyay, Sébastien Riou, Didier Schwab. Effect Of Personalized Calibration On Gaze Estimation Using Deep-Learning. [Research Report] Groupe d'Étude en Traduction Automatique/Traitement Automatisé des Langues et de la Parole ; LIG (Laboratoire informatique de Grenoble) ; Université Grenoble Alpes. 2021.

Articles liés à la fois au Traitement Automatique des Langues et de la Parole et de la Communication Alternative et Augmentée

[116] Céline Vaschalde, Pauline Trial, Emmanuelle Esperança–Rodier, Didier Schwab, Benjamin Lecouteux. Automatic pictogram generation from speech to help the implementation of a mediated communication. Conference on Barrier–free Communication, Nov 2018, Geneva, Switzerland.

[117] Céline Vaschalde, Benjamin Lecouteux, Didier Schwab. Génération de pictogrammes à partir de la parole spontanée pour la mise en place d'une communication médiée. 50 ans de linguistique sur corpus oraux : Apports à l'étude de la variation, Nov 2018, Orléans, France.

[118] Didier Schwab, Pauline Trial, Céline Vaschalde, Loïc Vial, Benjamin Lecouteux. Apporter des connaissances sémantiques à un jeu de pictogrammes destiné à des personnes en situation de handicap : Un ensemble de liens entre Wordnet et Arasaac, Arasaac–WN. TALN 2019, 2019, Toulouse, France.

[119] Lucie Chasseur, Marion Dohen, Benjamin Lecouteux, Sébastien Riou, Amélie Rochet–Capellan, et al.. Evaluation of the acceptability and usability of augmentative and alternative communication (AAC) tools : the example of pictogram grid communication systems with voice output. ACM SIGACCESS 2020 – Conference on Computers and Accessibility, Oct 2020, Athènes, Greece.

Bibliographie

Slimane Abdellaoui, Valérie Bellynck, Mathieu Mangeot et Christian Boitet : Outillage de l'accès aux textes par la lecture active étymologique multilingue pour apprenants berbérophones et arabophones. *In Traitement Automatique des Langues Africaines TALAf 2018*, Grenoble, France, septembre 2018. URL https://hal.archives-ouvertes.fr/hal-02054881.

Lahsen Abouenour, Karim Bouzoubaa et Paolo Rosso : On the evaluation and improvement of Arabic wordnet coverage and usability. *Language Resources and Evaluation*, 47(3):891–917, 2013.

Sallam Abualhaija et Karl-Heinz Zimmermann : D-bees : A novel method inspired by bee colony optimization for solving word sense disambiguation. *Swarm and Evolutionary Computation*, pages –, 2016. ISSN 2210-6502. URL http://www.sciencedirect.com/science/article/pii/S221065021500098X.

Eneko Agirre et Philip Edmonds : *Word Sense Disambiguation : Algorithms and Applications*. Springer Publishing Company, Incorporated, 1st édition, 2007. ISBN 1402068700.

Roee Aharoni, Melvin Johnson et Orhan Firat : Massively multilingual neural machine translation. *In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3874–3884, Minneapolis, Minnesota, juin 2019. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/N19-1388.

Cindy Aloui, Carlos Ramisch, Alexis Nasr et Lucie Barque : SLICE : Supersense-based lightweight interpretable contextual embeddings. *In Proceedings of the 28th International Conference on Computational Linguistics*, pages 3357–3370, Barcelona, Spain (Online), décembre 2020. International Committee on Computational Linguistics. URL https://aclanthology.org/2020.coling-main.298.

Mikel Artetxe et Holger Schwenk : Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610, 2019.

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma et Isabelle Augenstein : A diagnostic study of explainability techniques for text classification. *In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3256–3274, Online, novembre 2020. Association for Computational Linguistics. URL https://aclanthology.org/2020.emnlp-main.263.

Marco Baroni, Georgiana Dinu et Kruszewski : Don't count, predict ! a systematic comparison of context-counting vs. context-predicting semantic vectors. *In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, pages 238–247, Baltimore, Maryland, June 2014. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/P14-1023.

Andrew A. Bayor, Margot Brereton, Laurianne Sitbon, Bernd Ploderer, Filip Bircanin, Benoit Favre et Stewart Koplick : Toward a competency-based approach to co-designing technologies with people with intellectual disability. *ACM Trans. Access. Comput.*, 14(2), juillet 2021. ISSN 1936-7228. URL https://doi.org/10.1145/3450355.

Emily M. Bender : On achieving and evaluating language-independence in nlp. *Linguistic Issues in Language Technology*, 6, 2011.

Emily M. Bender et Batya Friedman : Data statements for natural language processing : Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604, 2018. URL https://www.aclweb.org/anthology/Q18-1041.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major et Shmargaret Shmitchell : On the dangers of stochastic parrots : Can language models be too big ? . *In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383097. URL https://doi.org/10.1145/3442188.3445922.

Emily M. Bender et Alexander Koller : Climbing towards NLU : On meaning, form, and understanding in the age of data. *In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198,

Online, juillet 2020. Association for Computational Linguistics. URL https://aclanthology.org/2020.acl-main.463.

F. Vanden Berghen et H. Bersini : Condor, a new parallel, constrained extension of powell's uobyqa algorithm : Experimental results and comparison with the dfo algorithm. *Journal of Computational and Applied Mathematics*, 181(1): 157–175, septembre 2005. URL <GotoISI>://000229805500011.

David R Beukelman et Janice C. Light : *Augmentative & alternative communication : Supporting children and adults with complex communication needs*. Brookes Publishing Co., 2020.

David R Beukelman et Pat Mirenda : *Communication alternative et améliorée : Aider les enfants et les adultes avec des difficultés de communication*. De Boeck Supérieur, 2017.

Anick Bianco, Philippe Blache, Julie Marty et Stéphane Rauzy : La Plateforme de Communication Alternative : un outil de communication et de rééducation. *In Congrès Scientifique International des Orthophonistes*, pages 127–140, France, 2006. Fédération Nationale des Orthophonistes. URL https://hal.archives-ouvertes.fr/hal-00142431.

El Moatez Billah Nagoudi, Jérémy Ferrero et Didier Schwab : Amélioration de la similarité sémantique vectorielle par méthodes non-supervisées. *In 24e conférence sur le Traitement Automatique des Langues Naturelles (TALN 2017)*, Orléans, France, juin 2017a. URL https://hal.archives-ouvertes.fr/hal-01531886.

El Moatez Billah Nagoudi, Jérémy Ferrero et Didier Schwab : LIM-LIG at SemEval-2017 Task1 : Enhancing the Semantic Similarity for Arabic Sentences with Vectors Weighting. *In International Workshop on Semantic Evaluations (SemEval-2017)*, Proceedings of the 11th International Workshop on Semantic Evaluations (SemEval-2017), pages 125 – 129, Vancouver, Canada, août 2017b. URL https://hal.archives-ouvertes.fr/hal-01531255.

El Moatez Billah Nagoudi, Jérémy Ferrero, Didier Schwab et Hadda Cherroun : Word Embedding-Based Approaches for Measuring Semantic Similarity of Arabic-English Sentences. *In 6th International Conference on Arabic Language Processing*, Fez, Morocco, octobre 2017c. URL https://hal.archives-ouvertes.fr/hal-01683494.

Filip Ircanin, Margot Brereton, Laurianne Sitbon, Bernd Ploderer, Andrew Azaabanye Bayor et Stewart Koplick : Including adults with severe in-

tellectual disabilities in co-design through active support. *In* Yoshifumi Kitamura, Aaron Quigley, Katherine Isbister, Takeo Igarashi, Pernille Bjørn et Steven Mark Drucker, éditeurs : *CHI '21 : CHI Conference on Human Factors in Computing Systems, Virtual Event / Yokohama, Japan, May 8-13, 2021*, pages 486 :1–486 :12. ACM, 2021. URL https://doi.org/10.1145/3411764.3445057.

Filip Bircanin, Laurianne Sitbon, Benoit Favre et Margot Brereton : Designing an IIR Research Apparatus with Users with Severe Intellectual Disability. *In ACM SIGIR Conference on Human Information Interaction and Retrieval (CHIIR)*, volume 5, pages 412–416, Vancouver, Canada, mars 2020a. ACM. URL https://hal-amu.archives-ouvertes.fr/hal-02470797.

Filip Bircanin, Laurianne Sitbon, Bernd Ploderer, Andrew Azaabanye Bayor, Michael Esteban, Stewart Koplick et Margot Brereton : The talkingbox. : Revealing strengths of adults with severe cognitive disabilities. *In The 22nd International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '20, New York, NY, USA, 2020b. Association for Computing Machinery. ISBN 9781450371032. URL https://doi.org/10.1145/3373625.3417025.

Piotr Bojanowski, Edouard Grave, Armand Joulin et Tomas Mikolov : Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017. URL https://aclanthology.org/Q17-1010.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Kohd, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Naraya-

nan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou et Percy Liang : On the opportunities and risks of foundation models. https://arxiv.org/abs/2108.07258, 2021.

Tomáš Brychcín et Lukáš Svoboda : UWB at SemEval-2016 task 1 : Semantic textual similarity using lexical, syntactic, and semantic information. *In Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 588–594, San Diego, California, juin 2016. Association for Computational Linguistics. URL https://aclanthology.org/S16-1089.

Lou Burnard : The british national corpus, 1998.

Jun Fu Cai, Wee Sun Lee et Yee Whye Teh : NUS-ML :improving word sense disambiguation using topic features. *In Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 249–252, Prague, Czech Republic, juin 2007. Association for Computational Linguistics. URL https://aclanthology.org/S07-1053.

Rich Caruana : Multitask learning. *Mach. Learn.*, 28(1):41–75, juillet 1997. ISSN 0885-6125. URL https://doi.org/10.1023/A:1007379606734.

Yee Seng Chan, Hwee Tou Ng et Zhi Zhong : NUS-PT : Exploiting parallel texts for word sense disambiguation in the English all-words tasks. *In Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 253–256, Prague, Czech Republic, juin 2007. Association for Computational Linguistics. URL https://aclanthology.org/S07-1054.

Lucie Chasseur, Marion Dohen, Benjamin Lecouteux, Sébastien Riou, Amélie Rochet-Capellan et Didier Schwab : Evaluation of the acceptability and usability of augmentative and alternative communication (AAC) tools : the example of pictogram grid communication systems with voice output. *In ACM SIGACCESS 2020 - Conference on Computers and Accessibility*, ASSETS '20. The 22nd International ACM SIGACCESS Conference on Compu-

ters and Accessibility - Virtual Event, Athenes, Greece, octobre 2020. URL https://hal.univ-grenoble-alpes.fr/hal-02896668.

Jacques Chauché : Un outil multidimensionnel de l'analyse du discours. *In COLING'1984 : 10th International Conference on Computational Linguistics*, pages 11–15, Stanford University, California, 1984.

Jacques Chauché, Violaine Prince, Simon Jaillet et Maguelonne Teisseire : Classification automatique de textes à partir de leur analyse syntaxico-sémantique. *In TALN 2003*, volume 1, pages 55–64, Batz-Sur-Mer, France, Juin 2003.

Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Si Wei, Hui Jiang et Diana Inkpen : Enhanced lstm for natural language inference. *In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, pages 1657–1668, 2017.

Christian Chiarcos, John P. McCrae, Philipp Cimiano et Christiane Fellbaum : Towards open data for linguistics : Linguistic linked data. *In New Trends of Research in Ontologies and Lexical Resources*, 2013.

Narcisa Chirvasiu et Elena Simion-Blândă : Alternative and augmentative communication in support of persons with language development retardation. *Revista Romaneasca pentru Educatie Multidimensionala*, 10:28, 07 2018.

Vincent Claveau : Detecting fake news in tweets from text and propagation graph : IRISA's participation to the FakeNews task at MediaEval 2020. *In MediaEval 2020 - MediaEval Benchmarking Initiative for Multimedia Evaluation*, pages 1–3, online, United States, décembre 2020. URL https://hal.archives-ouvertes.fr/hal-03116027.

Vincent Claveau, Antoine Chaffin et Ewa Kijak : La génération de textes artificiels en substitution ou en complément de données d'apprentissage. *In Traitement Automatique des Langues Naturelles (TALN)*, Lille, France, 2021.

K. Bretonnel Cohen, Jingbo Xia, Pierre Zweigenbaum, Tiffany Callahan, Orin Hargraves, Foster Goss, Nancy Ide, Aurélie Névéol, Cyril Grouin et Lawrence E. Hunter : Three dimensions of reproducibility in natural language processing. *In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, mai 2018. European Language Resources Association (ELRA). URL https://www.aclweb.org/anthology/L18-1025.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer et Veselin Stoyanov : Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*, 2019.

Matthieu Constant, Marie Candito et Djamé Seddah : The ligm-alpage architecture for the spmrl 2013 shared task : Multiword expression analysis and dependency parsing. *In Proceedings of the EMNLP Workshop on Statistical Parsing of Morphologically Rich Languages (SPMRL 2013)*, 2013.

Jim Cowie, Joe Guthrie et Louise Guthrie : Lexical disambiguation using simulated annealing. *In Proceedings of the workshop on Speech and Natural Language (HLT '91)*, 1992.

Irene Cramer, Tonio Wandmacher et Uli Waltinger : *WordNet : An electronic lexical database*, chapitre Modeling, Learning and Processing of Text Technological Data Structures. Springer, 2010.

Andrew M Dai et Quoc V Le : Semi-supervised sequence learning. *In Advances in neural information processing systems*, pages 3079–3087, 2015.

Scott C. Deerwester, Susan T. Dumais, Thomas K. Landauer, George W. Furnas et Richard A. Harshman : Indexing by latent semantic analysis. *Journal of the American Society of Information Science*, 41(6), 1990. URL http://citeseer.nj.nec.com/deerwester90indexing.html.

Mathieu Dehouck et Pascal Denis : Delexicalized word embeddings for cross-lingual dependency parsing. *In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics : Volume 1, Long Papers*, pages 241–250, Valencia, Spain, avril 2017. Association for Computational Linguistics. URL https://aclanthology.org/E17-1023.

Catalina Devandas-Aguilar : Rapport de la rapporteuse spéciale sur les droits des personnes handicapées. en visite en france. *New York : ONU*, 2019. URL https://inshea.fr/sites/default/files/www/sites/default/files/medias/ONU%20Rapport.pdf.

Jacob Devlin, Ming-Wei Chang, Kenton Lee et Kristina Toutanova : BERT : Pre-training of deep bidirectional transformers for language understanding. *In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, juin 2019. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/N19-1423.

Mattia A. Di Gangi, Roldano Cattoni, Luisa Bentivogli, Matteo Negri et Marco Turchi : MuST-C : a Multilingual Speech Translation Corpus. *In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2012–2017, Minneapolis, Minnesota, juin 2019. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/N19-1202.

Marie Duponchelle : *Le droit à l'interopérabilité : études de droit de la consommation*. Thèse, Université Panthéon-Sorbonne - Paris I, avril 2015. URL https://tel.archives-ouvertes.fr/tel-01618804.

Angela Duvivier-Senis : *John Rupert Firth Historien de la linguistique et fondateur de la "London School"*. Thèse de doctorat, Université Sorbonne Paris Cité, Université Paris Diderot, 11 2016.

Moussa Kamal Eddine, Antoine J. P. Tixier et Michalis Vazirgiannis : Barthez : a skilled pretrained french sequence-to-sequence model, 2021.

Philip Edmonds et Scott Cotton : Senseval-2 : Overview. *In The Proceedings of the Second International Workshop on Evaluating Word Sense Disambiguation Systems*, SENSEVAL '01, pages 1–5, Stroudsburg, PA, USA, 2001. Association for Computational Linguistics. URL http://dl.acm.org/citation.cfm?id=2387364.2387365.

Julian Eisenschlos, Sebastian Ruder, Piotr Czapla, Marcin Kardas, Sylvain Gugger et Jeremy Howard : Multifit : Efficient multi-lingual language model fine-tuning. *In Proceedings of the 2019 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2019.

William Black Horacio Rodríguez-Musa Alkhalifa Piek Vossen Adam Pease Elkateb, Sabri et Christiane Fellbaum : Building a wordnet for Arabic. *In In Proceedings of The fifth international conference on Language Resources and Evaluation (LREC 2006)*, 2006.

Solène Evain, Adrien Contesse, Antoine Pinchaud, Didier Schwab, B Lecouteux et Nathalie Henrich Bernardoni : Beatbox sounds recognition using a speech-dedicated HMM-GMM based system. *In Models and Analysis of Vocal Emissions for Biomedical Applications : 11th International Workshop*, Florence, Italy, décembre 2019. URL https://hal.archives-ouvertes.fr/hal-02429730.

Solène Evain, Adrien Contesse, Antoine Pinchaud, Didier Schwab, Benjamin Lecouteux et Nathalie Henrich Bernardoni : Reconnaissance de parole beatboxée à l'aide d'un système HMM-GMM inspiré de la reconnaissance automatique de la parole. *In* Christophe Benzitoun, Chloé Braud, Laurine Huber, David Langlois, Slim Ouni, Sylvain Pogodalla et Stéphane Schneider, éditeurs : *JEP-TALN-RECITAL 2020 - 6e conférence conjointe 33e Journées d'Études sur la Parole, 27e Traitement Automatique des Langues Naturelles, 22e Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues*, pages 208–216, Nancy, France, juin 2020. ATALA. URL https://hal.archives-ouvertes.fr/hal-02798538.

Solène Evain, Benjamin Lecouteux, Didier Schwab, Adrien Contesse, Antoine Pinchaud et Nathalie Henrich Bernardoni : Human Beatbox Sound Recognition using an Automatic Speech Recognition Toolkit. *Biomedical Signal Processing and Control*, 67:102468, mai 2021a. URL https://hal.archives-ouvertes.fr/hal-02896690.

Solène Evain, Ha Nguyen, Hang Le, Marcely Zanon Boito, Salima Mdhaffar, Sina Alisamir, Ziyi Tong, Natalia Tomashenko, Marco Dinarelli, Titouan Parcollet, Alexandre Allauzen, Yannick Estève, Benjamin Lecouteux, François Portet, Solange Rossato, Fabien Ringeval, Didier Schwab et Laurent Besacier : LeBenchmark : A Reproducible Framework for Assessing Self-Supervised Representation Learning from Speech. *In INTERSPEECH 2021 : Conference of the International Speech Communication Association*, Brno, Czech Republic, août 2021b. URL https://hal.archives-ouvertes.fr/hal-03317730.

Christiane Fellbaum : *WordNet : An Electronic Lexical Database (Language, Speech, and Communication)*. The MIT Press, mai 1998. ISBN 026206197X. URL http://www.amazon.ca/exec/obidos/redirect?tag=citeulike09-20&path=ASIN/026206197X.

Jacques Ferber, Olivier Gutknecht et Fabien Michel : From agents to organizations : An organizational view of multi-agent systems. *In* Paolo Giorgini, Jörg P. Müller et James Odell, éditeurs : *Agent-Oriented Software Engineering IV*, pages 214–230, Berlin, Heidelberg, 2004. Springer Berlin Heidelberg. ISBN 978-3-540-24620-6.

Jérémy Ferrero : *Similarités Textuelles Sémantiques Translingues : vers la Détection Automatique du Plagiat par Traduction*. thèse, LA COMMUNAUTÉ UNIVERSITÉ GRENOBLE ALPES, décembre 2017. URL https://tel.archives-ouvertes.fr/tel-01721390.

Jérémy Ferrero, Frédéric Agnès, Laurent Besacier et Didier Schwab : A Multilingual, Multi-Style and Multi-Granularity Dataset for Cross-Language Textual Similarity Detection. *In 10th edition of the Language Resources and Evaluation Conference*, Portorož, Slovenia, mai 2016. URL https://hal.archives-ouvertes.fr/hal-01303135.

Jérémy Ferrero, Frédéric Agnès, Laurent Besacier et Didier Schwab : Using Word Embedding for Cross-Language Plagiarism Detection. *In EACL 2017*, volume 2, pages 415 – 421, Valence, Spain, avril 2017a. URL https://hal.archives-ouvertes.fr/hal-01502146.

Jérémy Ferrero, Laurent Besacier, Didier Schwab et Frédéric Agnès : CompiLIG at SemEval-2017 Task 1 : Cross-Language Plagiarism Detection Methods for Semantic Textual Similarity. *In Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval 2017)*, Vancouver, Canada, August 2017b.

Jérémy Ferrero, Laurent Besacier, Didier Schwab et Frédéric Agnès : CompiLIG at SemEval-2017 Task 1 : Cross-Language Plagiarism Detection Methods for Semantic Textual Similarity. *In Proceedings of the 11th International Workshop on Semantic Evaluations (SemEval-2017),*, Vancouver, Canada, août 2017c. URL https://hal.archives-ouvertes.fr/hal-01531330.

J. R. Firth : A synopsis of linguistic theory 1930-55. 1952-59:1–32, 1957.

W. N. Francis et H. Kučera : A Standard Corpus of Present-Day Edited American English, for use with Digital Computers (Brown). Rapport technique, Brown University, Providence, Rhode Island, 1964.

W. N. Francis et H. Kucera : Brown corpus manual. Rapport technique, Department of Linguistics, Brown University, Providence, Rhode Island, US, 1979. URL http://icame.uib.no/brown/bcm.html.

Peter Frankopan : *Les routes de la soie : l'histoire du cœur du monde*. Flammarion, Paris, 2019. ISBN 978-2081480407.

Jibril Frej : *Incorporation de connaissances a priori pour la Recherche d'Information Textuelle Neuronale*. thèse, Université Grenoble Alpes [2020-....], février 2021.

Jibril FREJ, Jean-Pierre Chevallet et Didier Schwab : Enhancing Translation Language Models with Word Embedding for Information Retrieval. *In 9ème*

Atelier Recherche d'Information SEmantique, Caen, France, juillet 2017. URL https://hal.archives-ouvertes.fr/hal-01681311.

Jibril Frej, Jean-Pierre Chevallet et Didier Schwab : Knowledge Based Transformer Model for Information Retrieval. *In* Iván Cantador, Max Chevalier, Massimo Melucci et Josiane Mothe, éditeurs : *Joint Conference of the Information Retrieval Communities in Europe (CIRCLE 2020)*, volume 2621 de *CIRCLE 2020 Information Retrieval Communities in Europe 2020*, Samatan, France, juillet 2020a. URL https://hal.archives-ouvertes.fr/hal-03263784.

Jibril FREJ, Philippe Mulhem, Didier Schwab et Jean-Pierre Chevallet : Combining Subword information and Language model for Information Retrieval. *In 15e Conférence en Recherche d'Information et Applications*, Rennes, France, mai 2018. URL https://hal.archives-ouvertes.fr/hal-01781181.

Jibril Frej, Didier Schwab et Jean-Pierre Chevallet : MLWIKIR : A python toolkit for building large-scale wikipedia-based information retrieval datasets in chinese, english, french, italian, japanese, spanish and more. *In* Iván Cantador, Max Chevalier, Massimo Melucci et Josiane Mothe, éditeurs : *Proceedings of the Joint Conference of the Information Retrieval Communities in Europe (CIRCLE 2020), Samatan, Gers, France, July 6-9, 2020*, volume 2621 de *CEUR Workshop Proceedings*. CEUR-WS.org, 2020b. URL http://ceur-ws.org/Vol-2621/CIRCLE20_22.pdf.

Jibril FREJ, Didier Schwab et Jean-Pierre Chevallet : Modèle Transformer à base de Connaissances pour la Recherche d'Information dans des Domaines Spécialisés. *Extraction et Gestion des Connaissances (EGC)*, janvier 2020. URL https://hal.archives-ouvertes.fr/hal-02474706.

Jibril Frej, Didier Schwab et Jean-Pierre Chevallet : WIKIR : A python toolkit for building a large-scale wikipedia-based english information retrieval dataset. *In Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*, pages 1926–1933, 2020a. URL https://www.aclweb.org/anthology/2020.lrec-1.237/.

Jibril Frej, Didier Schwab et Jean-Pierre Chevallet : WIKIR : A python toolkit for building a large-scale Wikipedia-based English information retrieval dataset. *In Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1926–1933, Marseille, France, mai 2020b. European Language Resources

Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.237.

C. Fuchs : *La linguistique cognitive*. Cogniprisme. Éditions de la Maison des sciences de l'homme, 2017. ISBN 9782735119233. URL https://books.google.fr/books?id=Y6o2DwAAQBAJ.

Souhir Gahbiche-Braham, Hélène Bonneau-Maynard, Thomas Lavergne et François Yvon : Joint segmentation and POS tagging for Arabic using a CRF-based classifier. *In Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2107–2113, Istanbul, Turkey, mai 2012. European Language Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2012/pdf/855_Paper.pdf.

Jane Garry et Carl Rubino : Facts about the world's languages. *HW Wilson*, 2001.

Alexander Gelbukh, Grigori Sidorov et Sang Yong Han : Evolutionary approach to natural language wsd through global coherence optimization. *WSEAS Transactions on Communications*, 2(1):11–19, 2003.

Joke J. Geytenbeek, Lidwine B. Mokkink, Dirk L. Knol, R. Jeroen Vermeulen et Kim J. Oostrom : Reliability and validity of the c-billt : A new instrument to assess comprehension of spoken language in young children with cerebral palsy and complex communication needs. *Augmentative and Alternative Communication*, 30(3):252–266, 2014. URL https://doi.org/10.3109/07434618.2014.924992.

Sahar Ghannay, Christophe Servan et Sophie Rosset : Neural Networks approaches focused on French Spoken Language Understanding : application to the MEDIA Evaluation Task. *In In Proceedings of The 28th International Conference on Computational Linguistics (COLING'2020), 2020*, Barcelona (online), Spain, décembre 2020. URL https://hal.archives-ouvertes.fr/hal-03007482.

E. Smeets R. Van de Berg M. Van den Berg Leopold M.G. Curfs Gillian S. Townend, Peter B. Marschik : Eye gaze technology as a form of augmentative and alternative communication for individuals with rett syndrome : Experiences of families in the netherlands. 28(1):101–112, 2016. ISSN 1056-263X, 1573-3580. URL http://link.springer.com/10.1007/s10882-015-9455-z.

Palash Goyal et Emilio Ferrara : Graph embedding techniques, applications, and performance : A survey. *Knowledge-Based Systems*, 151:78–94, 2018. ISSN 0950-7051. URL https://www.sciencedirect.com/science/article/pii/S0950705118301540.

Algirdas Julien Greimas : *Structural Semantics : An Attempt at a Method*. University of Nebraska Press, 1984.

Benoit Habert, Cécile Fabre et Fabrice Issac : *DE L'ECRIT AU NUMERIQUE. Constituer, normaliser et exploiter les corpus électroniques*. Numéro ISBN : 2-225-82953-5. ELSEVIER MASSON, 1998.

Marwa Hadj Salah : *Arabic word sense disambiguation for and by machine translation*. thèse, Université Grenoble Alpes ; Université de Sfax (Tunisie). Faculté des Sciences économiques et de gestion, décembre 2018. URL https://tel.archives-ouvertes.fr/tel-02139438.

Marwa Hadj Salah, Hervé Blanchon, Mounir Zrigui et Didier Schwab : Amélioration de la traduction automatique d'un corpus annoté. *In JEP-TALN-RECITAL 2016*, Paris, France, juillet 2016. URL https://hal.archives-ouvertes.fr/hal-01680553.

Marwa Hadj Salah, Hervé Blanchon, Mounir Zrigui et Didier Schwab : Un corpus en arabe annoté manuellement avec des sens WordNet. *In 25e conférence sur le Traitement Automatique des Langues Naturelles*, Rennes, France, mai 2018a. URL https://hal.archives-ouvertes.fr/hal-01781188.

Marwa Hadj Salah, Loïc Vial, Hervé Blanchon, Mounir Zrigui, Benjamin Lecouteux et Didier Schwab : La désambiguïsation lexicale d'une langue moins bien dotée, l'exemple de l'arabe. *In 25e conférence sur le Traitement Automatique des Langues Naturelles*, Rennes, France, mai 2018b. URL https://hal.archives-ouvertes.fr/hal-01781185.

Coleman Haley : This is a BERT. now there are several of them. can they generalize to novel words ? *In Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 333–341, Online, novembre 2020. Association for Computational Linguistics. URL https://aclanthology.org/2020.blackboxnlp-1.31.

Sébastien Harispe, Sylvie Ranwez, Stefan Janaqi et Jacky Montmain : Semantic similarity from natural language and ontology analysis. *Synthesis Lectures on Human Language Technologies*, 8(1):1–254, May 2015. ISSN 1947-4059. URL http://dx.doi.org/10.2200/S00639ED1V01Y201504HLT027.

Zellig S. Harris, Michael Gottfried, Thomas Ryckman, Paul Mattick Jr., Anne Daladier, T.N. Harris et S. Harris : *The Form of Information in Science*, volume 104 de *Boston Studies in the Philosophy and History of Science*. Springer Netherlands, 1989.

Graeme Hirst et David St-Onge : Lexical chains as representations of context for the detection and correction of malapropisms. *In* Christiane Fellbaum, éditeur : *WordNet : An Electronic Lexical Database*, pages 305–332. MIT Press, 1998.

Thomas Hofmann : Probabilistic latent semantic indexing. *In* Fredric Gey, Marti Hearst et Richard Tong, éditeurs : *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '99), August 15-19, 1999, Berkeley, CA, USA*, pages 50–57. ACM Press, New York, NY, USA, 1999.

Francois Houllier et Jean-Baptiste Merilhou-Goudard : Les sciences participatives en France. Rapport technique. URL https://hal.inrae.fr/hal-02801940.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan et Sylvain Gelly : Parameter-efficient transfer learning for NLP. *In* Kamalika Chaudhuri et Ruslan Salakhutdinov, éditeurs : *Proceedings of the 36th International Conference on Machine Learning*, volume 97 de *Proceedings of Machine Learning Research*, pages 2790–2799. PMLR, 09–15 Jun 2019. URL http://proceedings.mlr.press/v97/houlsby19a.html.

Jeremy Howard et Sebastian Ruder : Universal language model fine-tuning for text classification. *In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, pages 328–339, 2018.

Ignacio Iacobacci, Mohammad Taher Pilehvar et Roberto Navigli : Embeddings for word sense disambiguation : An evaluation study. *In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, pages 897–907, Berlin, Germany, August 2016. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/P16-1085.

Nancy Ide et Catherine Macleod : The american national corpus : A standardized resource of american english. *In Proceedings of Corpus Linguistics 2001*, volume 3, 2001.

Lionel Jean : Les emprunts et la langue française. le phénomène des échanges linguistiques. https://www.axl.cefan.ulaval.ca/francophon

ie/HIST_FR_s92_Emprunts.htm, 2020. [en ligne ; dernière mise à jour le 25 juin 2020 accédé le 30 juin 2021].

Yangfeng Ji et Jacob Eisenstein : Discriminative improvements to distributional sentence similarity. *In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 891–896, Seattle, Washington, USA, octobre 2013. Association for Computational Linguistics. URL https://aclanthology.org/D13-1090.

Russell Jones, éditeur. *Loan-Words in Indonesian and Malay*. KITLV Press, Leiden, 2007. ISBN 9789067183048.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali et Monojit Choudhury : The state and fate of linguistic diversity and inclusion in the NLP world. *In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online, juillet 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.acl-main.560.

Mikael Kågebäck et Hans Salomonsson : Word sense disambiguation using a bidirectional lstm. *In 5th Workshop on Cognitive Aspects of the Lexicon (CogALex)*. Association for Computational Linguistics, 2016.

Zae Myung Kim, Laurent Besacier, Vassilina Nikoulina et Didier Schwab : Do Multilingual Neural Machine Translation Models Contain Language Pair Specific Attention Heads ? *In Findings of ACL 2021*, Bangkok (virtual), France, août 2021. URL https://hal.archives-ouvertes.fr/hal-03299010.

Diederik P Kingma et Jimmy Ba : Adam : A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Nikita Kitaev, Steven Cao et Dan Klein : Multilingual constituency parsing with self-attention and pre-training. *In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3499–3505, Florence, Italy, juillet 2019. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/P19-1340.

H. Kucera et W. N. Francis : *Computational analysis of present-day American English*. Brown University Press, Providence, RI, 1967.

Sawan Kumar, Sharmistha Jat, Karan Saxena et Partha Talukdar : Zero-shot word sense disambiguation using sense definition embeddings. *In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*,

pages 5670–5681, Florence, Italy, juillet 2019. Association for Computational Linguistics.

Mathieu Lafourcade : Conceptual vector learning - comparing bootstrapping from a thesaurus or induction by emergence. *In Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy, mai 2006. European Language Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2006/pdf/618_pdf.pdf.

Mathieu Lafourcade : Lexique et analyse sémantique de textes - structures, acquisitions, calculs, et jeux de mots. HDR de l'Université Montpellier II, décembre 2011.

Mathieu Lafourcade et Violaine Prince : Synonymies et vecteurs conceptuels. *In TALN'2001*, Tours, France, Juillet 2001a.

Mathieu Lafourcade et Violaine Prince : Synonymy and conceptual vectors. *In NLPRS'2001*, pages 127–134, Tokyo, Japon, Novembre 2001b.

Mathieu Lafourcade et Eugène Sandford : Analyse et désambiguïsation lexicale par vecteurs sémantiques. *In TALN'99*, pages 351–356, Cargèse, France, Juillet 1999.

Guillaume Lample et Alexis : Cross-lingual language model pretraining. *In Advances in neural information processing systems*, 2019.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma et Radu Soricut : Albert : A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*, 2019.

Thomas K. Landauer et Susan T. Dumais : A solution to Plato's problem : The latent semantic analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review*, 104:211–240, 1997. URL http://www.welchco.com/02/14/01/60/96/02/2901.HTM.

Larousse, éditeur. *Thésaurus Larousse - des idées aux mots, des mots aux idées*. Larousse, 1992.

Larousse, éditeur. *Le Petit Larousse Illustré 2022*. Larousse, 2022.

Hang Le, Florentin Barbier, Ha Nguyen, Natalia Tomashenko, Salima Mdhaffar, Souhir Gahbiche, Bougares Fethi, Benjamin Lecouteux, Didier Schwab et Yannick Estève : ON-TRAC' systems for the IWSLT 2021 low-resource

speech translation and multilingual speech translation shared tasks. *In International Conference on Spoken Language Translation (IWSLT)*, Bangkok (virtual), Thailand, août 2021a. URL https://hal.archives-ouvertes.fr/hal-03298854.

Hang Le, Juan Pino, Changhan Wang, Jiatao Gu, Didier Schwab et Laurent Besacier : Dual-decoder Transformer for Joint Automatic Speech Recognition and Multilingual Speech Translation. *In COLING 2020 (long paper)*, Virtual, Spain, décembre 2020a. URL https://hal.archives-ouvertes.fr/hal-02991564.

Hang Le, Juan Pino, Changhan Wang, Jiatao Gu, Didier Schwab et Laurent Besacier : Dual-decoder transformer for joint automatic speech recognition and multilingual speech translation. *In Proceedings of the 28th International Conference on Computational Linguistics*, pages 3520–3533, Barcelona, Spain (Online), décembre 2020b. International Committee on Computational Linguistics. URL https://www.aclweb.org/anthology/2020.coling-main.314.

Hang Le, Juan Pino, Changhan Wang, Jiatao Gu, Didier Schwab et Laurent Besacier : Lightweight Adapter Tuning for Multilingual Speech Translation. *In The Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP 2021)*, Bangkok (Virtual), Thailand, août 2021b. URL https://hal.archives-ouvertes.fr/hal-03294912.

Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoît Crabbé, Laurent Besacier et Didier Schwab : FlauBERT : des modèles de langue contextualisés pré-entraînés pour le français. *In* Christophe Benzitoun, Chloé Braud, Laurine Huber, David Langlois, Slim Ouni, Sylvain Pogodalla et Stéphane Schneider, éditeurs : *6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Volume 2 : Traitement Automatique des Langues Naturelles*, pages 268–278, Nancy, France, juin 2020c. ATALA. URL https://hal.archives-ouvertes.fr/hal-02784776.

Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoit Crabbe, Laurent Besacier et Didier Schwab : FlauBERT : Unsupervised Language Model Pre-training for French.

In LREC, Marseille, France, 2020d. URL https://hal.archives-ouvertes.fr/hal-02890258.

Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoit Crabbé, Laurent Besacier et Didier Schwab : Flaubert : des modèles de langue contextualisés pré-entraînés pour le français. *In 27ème conférence sur le Traitement Automatique des Langues Naturelles (TALN 2020)*, 2020e.

Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoit Crabbé, Laurent Besacier et Didier Schwab : Flaubert : Unsupervised language model pre-training for french. *In Proceedings of the Twelfth International Conference on Language Resources and Evaluation (LREC 2020)*, 2020f.

Minh Le, Marten Postma, Jacopo Urbani et Piek Vossen : A deep dive into word sense disambiguation with lstm. *In Proceedings of the 27th International Conference on Computational Linguistics*, pages 354–365. Association for Computational Linguistics, 2018. URL http://aclweb.org/anthology/C18-1030.

Yann Le Cun : *Quand la machine apprend : la révolution des neurones artificiels et de l'apprentissage profond / Yann Le Cun ; avec la collaboration de Caroline Brizard*. Odile Jacob, Paris, 2019. ISBN 978-2-7381-4931-2.

Christophe Lecerf : *Une leçon de piano ou la double boucle de l'apprentissage cognitif.*, volume 3-1997. Université Paris 8 - Vincenne-Saint-Denis, Université Paris 8, Vincennes Saint-denis, Mars 1997. revue Travaux et Documents.

Jacques Leclerc : Page sur la Malaisie sur le site Web « *l'aménagement linguistique dans le monde* ». http://www.axl.cefan.ulaval.ca/asie/malaysia.htm, 2016. [en ligne ; dernière mise à jour le 25 novembre 2016 accédé le 30 juin 2021].

Benjamin Lecouteux et Didier Schwab : Décodage de graphe à l'aide de colonies de fourmis. *In 30èmes Journées d'étude de la parole*, page 6, Le Mans, France, juin 2014. URL https://hal.archives-ouvertes.fr/hal-01003001.

Benjamin Lecouteux et Didier Schwab : Ant Colony Algorithm applied to Automatic Speech Recognition Graph Decoding. *In Interspeech 2015*, Dresden, Germany, septembre 2015. URL https://hal.archives-ouvertes.fr/hal-01170535.

James Lee-Thorp, Joshua Ainslie, Ilya Eckstein et Santiago Ontanon : Fnet : Mixing tokens with fourier transforms, 2021.

Jacqueline Léon : Le CNRS et les débuts de la traduction automatique en France. *La revue pour l'histoire du CNRS*, 6:6–24, 2002. URL https://halshs.archives-ouvertes.fr/halshs-01145021.

Michael Lesk : Automatic sense disambiguation using mrd : how to tell a pine cone from an ice cream cone. *In Proceedings of SIGDOC '86*, pages 24–26, New York, NY, USA, 1986. ACM. ISBN 0-89791-224-1.

Omer Levy et Yoav Goldberg : Dependency-based word embeddings. *In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Volume 2 : Short Papers*, pages 302–308, 2014. URL http://aclweb.org/anthology/P/P14/P14-2050.pdf.

Quentin Lhoest, Albert Villanova del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, Joe Davison, Mario Šaško, Gunjan Chhablani, Bhavitvya Malik, Simon Brandeis, Teven Le Scao, Victor Sanh, Canwen Xu, Nicolas Patry, Angelina McMillan-Major, Philipp Schmid, Sylvain Gugger, Clément Delangue, Théo Matussière, Lysandre Debut, Stas Bekman, Pierric Cistac, Thibault Goehringer, Victor Mustar, François Lagunas, Alexander M. Rush et Thomas Wolf : Datasets : A community library for natural language processing, 2021.

Xian Li, Paul Michel, Antonios Anastasopoulos, Yonatan Belinkov, Nadir Durrani, Orhan Firat, Philipp Koehn, Graham Neubig, Juan Pino et Hassan Sajjad : Findings of the first shared task on machine translation robustness. *Fourth Conference on Machine Translation (WMT19)*, pages 91–102, 2019.

Lian Lim et Didier Schwab : Limits of Lexical Semantic Relatedness with Ontology-based Conceptual Vectors. *In NLPCS 2008 : Natural Language Processing and Cognitive Science*, Barcelone, Spain, juin 2008. URL https://hal.archives-ouvertes.fr/hal-03319015.

Tal Linzen : Issues in evaluating semantic spaces using word analogies. *In Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 13–18, Berlin, Germany, août 2016. Association for Computational Linguistics. URL https://aclanthology.org/W16-2503.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer et Veselin Stoyanov : Roberta : A robustly optimized bert pretraining approach. *arXiv preprint arXiv :1907.11692*, 2019.

Anthony Lodge : *Le français, histoire d'un dialecte devenu langue*. Fayard, Paris, 1997. ISBN 978-2213598628.

Fuli Luo, Tianyu Liu, Zexue He, Qiaolin Xia, Zhifang Sui et Baobao Chang : Leveraging gloss knowledge in neural word sense disambiguation by hierarchical co-attention. *In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1402–1411. Association for Computational Linguistics, 2018a. URL http://aclweb.org/anthology/D18-1170.

Fuli Luo, Tianyu Liu, Qiaolin Xia, Baobao Chang et Zhifang Sui : Incorporating glosses into neural word sense disambiguation. *In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, pages 2473–2482. Association for Computational Linguistics, 2018b. URL http://aclweb.org/anthology/P18-1230.

Ingo Lütkebohle : Page sur la malaisie sur le site web « *l'ethnologue* ». https://www.ethnologue.com/country/my, 2021. [en ligne ; accédé le 30 juin 2021].

Sean MacAvaney, Andrew Yates, Sergey Feldman, Doug Downey, Arman Cohan et Nazli Goharian : Simplified data wrangling with ir$_{datasets}$, 2021.

Saber Mansour et Jacques Ferber : Un modèle organisationnel pour les systèmes multi-agents ouverts. *In Journées Francophones des Systèmes Multi-Agents*, France, octobre 2007. URL https://hal.archives-ouvertes.fr/hal-00390421.

Gary Marcus et Ernest Davis : Has ai found a new foundation ? *The Gradient*, 2021.

Mitchell P Marcus, Mary Ann Marcinkiewicz et Beatrice Santorini : Building a large annotated corpus of english : The penn treebank. *Computational linguistics*, 19(2):313–330, 1993.

d'Hoffschmidt Martin, Vidal Maxime, Belblidia Wacim et Brendlé Tom : FQuAD : French Question Answering Dataset. *arXiv e-prints*, page arXiv :2002.06071, Feb 2020.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric Villemonte de la Clergerie, Djamé Seddah et Benoît Sagot : CamemBERT : a Tasty French Language Model. *arXiv preprint arXiv:1911.03894*, Nov 2019.

Diana McCarthy : Lexical substitution as a task for WSD evaluation. *In Proceedings of the ACL-02 Workshop on Word Sense Disambiguation : Recent Successes and Future Directions*, pages 089–115. Association for Computational Linguistics, juillet 2002. URL https://www.aclweb.org/anthology/W02-0816.

Diana McCarthy : Word sense disambiguation : An overview. *Language and Linguistics Compass*, 3(2):537–558, 2009. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1749-818X.2009.00131.x.

John Philip McCrae, Christian Chiarcos, Francis Bond, Philipp Cimiano, Thierry Declerck, Gerard de Melo, Jorge Gracia, Sebastian Hellmann, Bettina Klimek, Steven Moran, Petya Osenova, Antonio Pareja-Lora et Jonathan Pool : The open linguistics working group : Developing the linguistic linked open data cloud. *In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2435–2441, Portorož, Slovenia, mai 2016. European Language Resources Association (ELRA). URL https://www.aclweb.org/anthology/L16-1386.

Margot Mieskes, Karën Fort, Aurélie Névéol, Cyril Grouin et Kevin B Cohen : NLP Community Perspectives on Replicability. *In Recent Advances in Natural Language Processing*, Varna, Bulgaria, septembre 2019. URL https://hal.archives-ouvertes.fr/hal-02282794.

Rada Mihalcea et Timothy Chklovski : *Building sense tagged corpora with volunteer contributions over the Web*, pages 357–402. John Benjamin Publishing Compagny, 2003.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado et Jeff Dean : Distributed representations of words and phrases and their compositionality. *In* C.J.C. Burges, L. Bottou, M. Welling, Z. Ghahramani et K.Q. Weinberger, éditeurs : *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc., 2013. URL http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf.

George A Miller : Wordnet : a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross et Katherine Miller : Wordnet : An on-line lexical database. *International Journal of Lexicography*, 3:235–244, 1990.

George A. Miller, Claudia Leacock, Randee Tengi et Ross T. Bunker : A semantic concordance. *In Proceedings of the workshop on Human Language Technology*, HLT '93, pages 303–308, Stroudsburg, PA, USA, 1993. Association for Computational Linguistics. ISBN 1-55860-324-7. URL http://dx.doi.org/10.3115/1075671.1075742.

Jacques Moeschler : *Pourquoi le langage ? : des Inuits à Google*. La lettre et l'idée. Armand Colin, Paris, 2020. ISBN 2200628552.

Andrea Moro et Roberto Navigli : Semeval-2015 task 13 : Multilingual all-words sense disambiguation and entity linking. *In Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 288–297, Denver, Colorado, June 2015. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/S15-2049.

Mohammad Nasiruddin, Didier Schwab, Andon Tchechmedjiev, Gilles Sérasset et Hervé Blanchon : Induction de sens pour enrichir des ressources lexicales. *In 21ème conférence sur le Traitement Automatique des Langues Naturelles (TALN 2014)*, page 6, Marseille, France, juillet 2014. URL https://hal.archives-ouvertes.fr/hal-01003002.

Mohammad Nasiruddin, Andon Tchechmedjiev, Hervé Blanchon et Didier Schwab : Création rapide et efficace d'un système de désambiguïsation lexicale pour une langue peu dotée. *In 22ème conférence sur le Traitement Automatique des Langues Naturelles*, Caen, France, juin 2015. URL https://hal.archives-ouvertes.fr/hal-01856098.

Roberto Navigli : Word sense disambiguation : A survey. *ACM Computing Surveys*, 41(2):10 :1–10 :69, feb. 2009. ISSN 0360-0300. URL http://doi.acm.org/10.1145/1459352.1459355.

Roberto Navigli, David Jurgens et Daniele Vannella : SemEval-2013 Task 12 : Multilingual Word Sense Disambiguation. *In Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2 : Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 222–231, 2013. URL http://www.aclweb.org/anthology/S13-2040.

Roberto Navigli et Mirella Lapata : An experimental study of graph connectivity for unsupervised word sense disambiguation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 32:678–692, April 2010. ISSN 0162-8828.

Roberto NAVIGLI, Kenneth C. LITKOWSKI et Orin HARGRAVES : Semeval-2007 task 07 : Coarse-grained english all-words task. *In SemEval-2007*, pages 30–35, Prague, Czech Republic, June 2007.

Roberto NAVIGLI et Simone Paolo PONZETTO : Babelnet : The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193(0):217 – 250, 2012a. ISSN 0004-3702. URL http://www.sciencedirect.com/science/article/pii/S0004370212000793.

Roberto NAVIGLI et Simone Paolo PONZETTO : BabelNet : The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250, 2012b.

Hwee Tou NG : Exemplar-based word sense disambiguation" some recent improvements. *In Second Conference on Empirical Methods in Natural Language Processing*, 1997. URL https://www.aclweb.org/anthology/W97-0323.

Hwee Tou NG et Hian Beng LEE : Integrating multiple knowledge sources to disambiguate word sense : an exemplar-based approach. *In Proceedings of the 34th annual meeting on Association for Computational Linguistics*, ACL '96, pages 40–47, Stroudsburg, PA, USA, 1996. Association for Computational Linguistics. URL http://dx.doi.org/10.3115/981863.981869.

Joakim NIVRE : *Inductive Dependency Parsing*, volume 34 de *Text, speech and language technology*. Springer, 2006. ISBN 978-1-4020-4888-3. URL http://www.springer.com/education+%26+language/linguistics/book/978-1-4020-4888-3.

Adrian NOVISCHI, Muirathnam SRIKANTH et Andrew BENNETT : LCC-WSD : System description for English coarse grained all words task at SemEval 2007. *In Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 223–226, Prague, Czech Republic, juin 2007. Association for Computational Linguistics. URL https://aclanthology.org/S07-1047.

Vincent NYCKEES : *La sémantique*. Belin, 1998.

Myle OTT, Sergey EDUNOV, David GRANGIER et Michael AULI : Scaling neural machine translation. *In Proceedings of the Third Conference on Machine Translation : Research Papers*, pages 1–9, Belgium, Brussels, octobre 2018. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W18-6301.

Sungjoon Park, Jihyung Moon, Sungdong Kim, Won Ik Cho, Jiyoon Han, Jangwon Park, Chisung Song, Junseong Kim, Yongsook Song, Taehwan Oh, Joohong Lee, Juhyun Oh, Sungwon Lyu, Younghoon Jeong, Inkwon Lee, Sangwoo Seo, Dongjun Lee, Hyunwoo Kim, Myeonghwa Lee, Seongbo Jang, Seungwon Do, Sunkyoung Kim, Kyungtae Lim, Jongwon Lee, Kyumin Park, Jamin Shin, Seonghyun Kim, Lucy Park, Alice Oh, Jung-Woo Ha et Kyunghyun Cho : Klue : Korean language understanding evaluation, 2021.

Tommaso Pasini et Roberto Navigli : Train-o-matic : Large-scale supervised word sense disambiguation in multiple languages without manual training data. *In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 78–88. Association for Computational Linguistics, 2017. URL http://aclweb.org/anthology/D17-1008.

Tommaso Pasini, Alessandro Raganato et Roberto Navigli : XL-WSD : An extra-large and cross-lingual evaluation framework for word sense disambiguation. *In Proc. of AAAI*, 2021.

T. Pedersen, S. Banerjee et S. Patwardhan : Maximizing Semantic Relatedness to Perform WSD. Research report, University of Minnesota Supercomputing Institute, March 2005.

Jeffrey Pennington, Richard Socher et Christopher D. Manning : Glove : Global vectors for word representation. *In Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014. URL http://www.aclweb.org/anthology/D14-1162.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee et Luke Zettlemoyer : Deep contextualized word representations. *In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana, juin 2018a. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/N18-1202.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee et Luke Zettlemoyer : Deep contextualized word representations. *In Proceedings of NAACL-HLT*, pages 2227–2237, 2018b.

Geneviève Petitpierre et Koviljka Barisnikov : *Le Vygotskij que nous (ne) connaissons (pas). Les principaux travaux de Vygotskij et la chronologie de leur composition*. Delachaux et Niestlé, 1 édition, 1 1994. ISBN 2-603-00944-3.

Jerin Philip, Alexandre Berard, Matthias Gallé et Laurent Besacier : Monolingual adapters for zero-shot neural machine translation. *In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4465–4470, Online, novembre 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.emnlp-main.361.

Pascal Picq : *La Plus Belle Histoire du Langage*, chapitre Aux sources du langage, pages 17–74. Seuil, 2008. ISBN 978-2020406673.

Mohammad Taher Pilehvar et Jose Camacho-Collados : WiC : the word-in-context dataset for evaluating context-sensitive meaning representations. *In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1267–1273, Minneapolis, Minnesota, juin 2019. Association for Computational Linguistics.

Alain Polguère : *Lexicologie et sémantique lexicale*. Les Presses de l'Université de Montréal, 2003.

Sameer S. Pradhan, Edward Loper, Dmitriy Dligach et Martha Palmer : Semeval-2007 task 17 : English lexical sample, srl and all words. *In Proceedings of the 4th International Workshop on Semantic Evaluations*, SemEval '07, pages 87–92, Stroudsburg, PA, USA, 2007. Association for Computational Linguistics. URL http://dl.acm.org.gate6.inist.fr/citation.cfm?id=1621474.1621490.

Vineel Pratap, Qiantong Xu, Anuroop Sriram, Gabriel Synnaeve et Ronan Collobert : MLS : A Large-Scale Multilingual Dataset for Speech Research. *In Proc. Interspeech 2020*, pages 2757–2761, 2020. URL http://dx.doi.org/10.21437/Interspeech.2020-2826.

R. Rada, H. Mili, E. Bicknell et M. Blettner : Development and application of a metric on semantic nets. *IEEE Transactions on Systems, Man, and Cybernetics*, 19(1):17–30, 1989.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li et Peter J Liu : Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*, 2019.

Alessandro Raganato, Claudio Delli Bovi et Roberto Navigli : Neural sequence learning models for word sense disambiguation. *In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1167–1178.

Association for Computational Linguistics, 2017. URL http://www.aclweb.org/anthology/D17-1121.

Prajit Ramachandran, Peter Liu et Quoc Le : Unsupervised pretraining for sequence to sequence learning. *In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 383–391, 2017.

François Rastier : *Sémantique et Recherche Cognitive*. Presses Universitaires de France, 1991.

Nicholas Roberston, Filip Bircadin et Laurianne Sitbon : Designing a pictorial communication web application with people with intellectual disabilities. *ASSET 2021 – The 23rd International ACM SIGACCESS Conference on Computers and Accessibility*, 2021.

Le Robert, éditeur. *Le Robert Illustré*. Éditions Le Robert, 2020.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau et Jason Weston : Recipes for building an open-domain chatbot. *In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics : Main Volume*, pages 300–325, Online, avril 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.eacl-main.24.

MaryAnn Romski et Rose A Sevcik : Augmentative communication and early intervention : Myths and realities. *Infants & Young Children*, 18(3):174–185, 2005. URL https://depts.washington.edu/isei/iyc/romski_18_3.pdf.

Andreas Rücklé, Gregor Geigle, Max Glockner, Tilman Beck, Jonas Pfeiffer, Nils Reimers et Iryna Gurevych : AdapterDrop : On the efficiency of adapters in transformers. *arXiv*, 2020. URL https://arxiv.org/abs/2010.11918.

N. Bahi-Buisson M. Voisin Université Pierre et Marie Curie (Paris) UFR de médecine Pierre et Marie Curie S. Patachon, M. Duclos : Syndrome de rett : instauration d'une communication alternative par commande oculaire : étude de cas unique. 2015. OCLC : 927105157.

Laurent Sagart : *La plus belle histoire du langage*, chapitre Mystérieuse langue mère. Seuil, 2008. ISBN 978-2020406673.

Gerard. Salton : *Automatic Information Organization and Retrieval.* McGraw Hill Text, 1968. ISBN 0070544859.

Gerard Salton : The smart retrieval system – experiments in automatic document processing, 1971.

Gerard Salton : The Smart document retrieval project. *In Proc. of the 14th Annual Int'l ACM/SIGIR Conf. on Research and Development in Information Retrieval*, Chicago, 1991.

Gerard Salton et Michael McGill : *Introduction to Modern Information Retrieval*. McGrawHill, New York, 1983.

E. Santus, Qin Lu, A. Lenci et Chu ren Huang : Unsupervised antonym-synonym discrimination in vector space. pages 328–333, 2014a. Italian Conference on Computational Linguistics [CLiC-it] ; Conference date : 01-01-2014.

Enrico Santus : *Making sense : from word distribution to meaning*. Thèse de doctorat, Hong Kong Polytechnic University, septembre 2016. URL https://theses.lib.polyu.edu.hk/handle/200/8805.

Enrico Santus, Qin Lu, Alessandro Lenci et Chu-Ren Huang : Taking antonymy mask off in vector space. *In Proceedings of the 28th Pacific Asia Conference on Language, Information and Computing*, pages 135–144, Phuket,Thailand, décembre 2014b. Department of Linguistics, Chulalongkorn University. URL https://aclanthology.org/Y14-1018.

Nathan Schneider, Behrang Mohit, Kemal Oflazer et Noah A. Smith : Coarse lexical semantic annotation with supersenses : An Arabic case study. *In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2 : Short Papers)*, pages 253–258, Jeju Island, Korea, juillet 2012. Association for Computational Linguistics. URL https://aclanthology.org/P12-2050.

Didier Schwab : Vecteurs conceptuels et fonctions lexicales :application à l'antonymie. Mémoire de dea/master, Université Montpellier 2, juin 2001. URL https://hal.archives-ouvertes.fr/hal-03328266.

Didier Schwab : *Approche hybride - lexicale et thématique - pour la modélisation, la détection et l'exploitation des fonctions lexicales en vue de l'analyse sémantique de texte*. Thèse, Université Montpellier II - Sciences et Techniques du Languedoc, décembre 2005. URL https://tel.archives-ouvertes.fr/tel-00333334.

Didier Schwab : Cognitive Ability Estimation and Reinforcement with Eye-tracking Games for Children with Multiple Disabilities. *In Grenoble Workshop on Models*

and Analysis of Eye Movements, Grenoble, France, juin 2018. URL https://hal.archives-ouvertes.fr/hal-01806290.

Didier Schwab, Amela Fejza, Loïc Vial et Yann Robert : The GazePlay Project : Open and Free Eye-trackers Games and a Community for People with Multiple Disabilities. *In ICCHP 2018 - 16th International Conference on Computers Helping People with Special Needs*, volume 10896 de *LNCS*, pages 254–261, Linz, Austria, juillet 2018. Springer. URL https://hal.archives-ouvertes.fr/hal-01804271.

Didier Schwab, Jérôme Goulian et Nathan Guillaume : Désambiguïsation lexicale par propagation de mesures sémantiques locales par algorithmes à colonies de fourmis. *In TALN'2011 : Traitement Automatique des Langues Naturelles*, pages x–x, Montpellier, France, 2011. URL https://hal.archives-ouvertes.fr/hal-00959146.

Didier Schwab, Jérôme Goulian et Andon Tchechmedjiev : Désambiguïsation lexicale de textes : efficacité qualitative et temporelle d'un algorithme à colonies de fourmis. *Traitement Automatique des Langues*, 54(1):99–138, 2013a. URL https://hal.archives-ouvertes.fr/hal-00953647.

Didier Schwab, Jérôme Goulian et Andon Tchechmedjiev : Worst-case complexity and empirical evaluation of artificial intelligence methods for unsupervised word sense disambiguation. *International Journal of Web Engineering and Technology*, 8(2):124–153, 2013b.

Didier Schwab, Jérôme Goulian, Andon Tchechmedjiev et Hervé Blanchon : Ant Colony Algorithm for the Unsupervised Word Sense Disambiguation of Texts : Comparison and Evaluation. *In Proceedings of the 25th International Conference on Computational Linguistics (COLING 2012)*, Mumbai, India. URL https://hal.archives-ouvertes.fr/hal-00953818.

Didier Schwab, Mathieu Lafourcade et Violaine Prince : Amélioration de la Représentation Sémantique Lexicale par les Vecteurs Conceptuels : Le Rôle de l'Antonymie. *In JADT'02 : Journées Internationales d'Analyse Statistique des Données Textuelles*, pages 701–712, St Malo, France, avril 2002a. URL https://hal-lirmm.ccsd.cnrs.fr/lirmm-00268623.

Didier Schwab, Mathieu Lafourcade et Violaine Prince : Antonymy and Conceptual Vectors. *In COLING'02 : 19th Conference on Computational Linguistics*, numéro Vol. 2/2, pages 904–910, Taipei, Japan, avril 2002b. URL https://hal-lirmm.ccsd.cnrs.fr/lirmm-00268558.

Didier SCHWAB, Mathieu LAFOURCADE et Violaine PRINCE : Vers l'Apprentissage Automatique pour et par les Vecteurs Conceptuels de Fonctions Lexicales - L'exemple de l'Antonymie. *In TALN : Traitement Automatique des Langues Naturelles*, Nancy, France, 2002c. URL https://hal-lirmm.ccsd.cnrs.fr/lirmm-00268559.

Didier SCHWAB, Mathieu LAFOURCADE et Violaine PRINCE : Extraction Semi-Supervisée de Couples d'Antonymes grâce à leur Morphologie. *In TALN : Traitement Automatique des Langues Naturelles*, Dourdan, France, juin 2005. URL https://hal-lirmm.ccsd.cnrs.fr/lirmm-00105981.

Didier SCHWAB et Lian Tze LIM : Blexisma2 : a Distributed Agent Framework for Constructing a Semantic Lexical Database based on Conceptual Vectors. *In DFMA 2008 International Conference on Distributed Framework and Applications*, Penang, Malaysia, octobre 2008. URL https://hal.archives-ouvertes.fr/hal-03319027.

Didier SCHWAB, Lian Tze LIM et Mathieu LAFOURCADE : Conceptual vectors, a complementary tool to lexical networks. *NLPCS 2007 : The 4th International Workshop on Natural Language Processing and Cognitive Science*, 2007.

Didier SCHWAB, Sébastien RIOU, Amela FEJZA, Loïc VIAL, Johana MARKU, Wafaa El HUSSEINI, E K SANNARA, Miles BARDON et Yann ROBERT : Le projet GazePlay : des jeux ouverts, gratuits et une communauté pour les personnes en situation de polyhandicap. *In 1024 – Bulletin de la Société informatique de France*. avril 2020a. URL https://hal.archives-ouvertes.fr/hal-03004915.

Didier SCHWAB, Andon TCHECHMEDJIEV, Jérôme GOULIAN et Gilles SÉRASSET : Comparisons of Relatedness Measures through a Word Sense Disambiguation Task. *In Advances in Language Production, Cognition and the Lexicon,*, page 23. Springer, juillet 2014. URL https://hal.archives-ouvertes.fr/hal-01003000.

Didier SCHWAB, Pauline TRIAL, Céline VASCHALDE, Loïc VIAL et Benjamin LECOUTEUX : Apporter des connaissances sémantiques à un jeu de pictogrammes destiné à des personnes en situation de handicap : Un ensemble de liens entre Wordnet et Arasaac, Arasaac-WN. *In TALN 2019*, Toulouse, France, 2019. URL https://hal.archives-ouvertes.fr/hal-02127258.

Didier SCHWAB, Pauline TRIAL, Céline VASCHALDE, L. VIAL, Emmanuelle ESPERANÇA-RODIER et Benjamin LECOUTEUX : Providing semantic knowledge to a set of pictograms for people with disabilities : a set of links between WordNet and Arasaac : Arasaac-WN. *In LREC*, Marseille, France, 2020b. URL https://hal.archives-ouvertes.fr/hal-02888279.

Vincent Segonne, Marie Candito et Benoit Crabbé : Using wiktionary as a resource for wsd : the case of french verbs. *In Proceedings of the 13th International Conference on Computational Semantics-Long Papers*, pages 259–270, 2019.

Rico Sennrich, Barry Haddow et Alexandra Birch : Neural machine translation of rare words with subword units. *In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, pages 1715–1725, 2016.

Gilles Sérasset : DBnary : Wiktionary as a Lemon-Based Multilingual Lexical Resource in RDF. *Semantic Web – Interoperability, Usability, Applicability*, 6 (4):355–361, 2015a. URL https://hal.archives-ouvertes.fr/hal-00953638.

Gilles Sérasset : Dbnary : Wiktionary as a lemon-based multilingual lexical resource in rdf. *Semantic Web*, 6(4):355–361, 2015b.

Gilles Sérasset et Mathieu Mangeot : Papillon lexical databases project : monolingual dictionaries and interlingual links. *In NLPRS 2001*, pages 119–125, 2001.

Jeff Sigafoos, Larah van der Meer, Ralf Schlosser, Giulio Lancioni, Mark O'Reilly et Vanessa Green : *Augmentative and Alternative Communication (AAC) in Intellectual and Developmental Disabilities*, pages 255–285. 12 2016. ISBN 9780128020753.

Benjamin Snyder et Martha Palmer : The english all-words task. *In Proceedings of SENSEVAL-3, the Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, 2004. URL http://www.aclweb.org/anthology/W04-0811.

Kaveh Taghipour et Hwee Tou Ng : One million sense-tagged instances for word sense disambiguation and induction. *In Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, pages 338–344, Beijing, China, July 2015a. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/K15-1037.

Kaveh Taghipour et Hwee Tou Ng : One Million Sense-Tagged Instances for Word Sense Disambiguation and Induction. *In Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, pages 338–344, Beijing, China, July 2015b. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/K15-1037.

Choon Hoe Tan : *Penang Hokkien Dialect*. Tan Choon Hoe, Penang, 2001. ISBN 983-40774-0-8.

Andon Tchechmedjiev : état de l'art : Mesures de similarité sémantique et algorithmes globaux pour la désambiguïsation lexicale a base de connaissances. *In RECITAL 2012*, Grenoble, June 2012. ATALA.

Andon Tchechmedjiev : *Interopérabilité Sémantique Multi-lingue des Ressources Lexicales en Données Liées Ouvertes*. Thèse de doctorat, Université Grenoble Alpes, 2016.

Andon Tchechmedjiev, Didier Schwab, Jérôme Goulian et Gilles Sérasset : Parameter estimation under uncertainty with Simulated Annealing applied to an ant colony based probabilistic WSD algorithm. *In Proceedings of the 1st International Workshop on Optimization Techniques for Human Language Technology, on behalf of (COLING 2012)*, Mumbai, India. URL https://hal.archives-ouvertes.fr/hal-00953822.

Andon Tchechmedjiev, Gilles Sérasset, Jérôme Goulian et Didier Schwab : Attaching Translations to Proper Lexical Senses in DBnary. *In 3rd Workshop on Linked Data in Linguistics : Multilingual Knowledge Resources and Natural Language Processing*, page to appear, Reykjavik, Iceland, mai 2014. URL https://hal.archives-ouvertes.fr/hal-00990870.

Jörg Tiedemann : Parallel data, tools and interfaces in OPUS. *In Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey, mai 2012a. European Language Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2012/pdf/463_Paper.pdf.

Jörg Tiedemann : Parallel data, tools and interfaces in opus. *In* Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Ugur Dogan, Bente Maegaard, Joseph Mariani, Jan Odijk et Stelios Piperidis, éditeurs : *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may 2012b. European Language Resources Association (ELRA). ISBN 978-2-9517408-7-7.

Tobii : Tobii annual report 2020, 2020. URL https://www.tobii.com/contentassets/b2fa5728796f413a852eb00f76ce2e7f/wkr0006.pdf/?v=1.0.

Céline Vaschalde, Benjamin Lecouteux et Didier Schwab : Génération de pictogrammes à partir de la parole spontanée pour la mise en place d'une communication médiée. *In 50 ans de linguistique sur corpus oraux : Apports à l'étude de la variation*, Orléans, France, novembre 2018a. URL https://hal.archives-ouvertes.fr/hal-01876781.

Céline Vaschalde, Pauline Trial, Emmanuelle Esperança-Rodier, Didier Schwab et Benjamin Lecouteux : Automatic pictogram generation from speech to help the implementation of a mediated communication. *In Conference on Barrier-free Communication*, Geneva, Switzerland, novembre 2018b. URL https://hal.archives-ouvertes.fr/hal-01880744.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser et Illia Polosukhin : Attention is all you need. *In* I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan et R. Garnett, éditeurs : *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc., 2017a. URL http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser et Illia Polosukhin : Attention is all you need. *In Advances in neural information processing systems*, pages 5998–6008, 2017b.

Jean Véronis : Cartographie lexicale pour la recherche d"information. *In Actes de la 10ème conférence sur le Traitement Automatique des Langues Naturelles. Articles longs*, pages 265–274, Batz-sur-Mer, France, juin 2003. ATALA. URL https://aclanthology.org/2003.jeptalnrecital-long.25.

Loïc Vial, Benjamin Lecouteux et Didier Schwab : Représentation vectorielle de sens pour la désambiguïsation lexicale à base de connaissances. *In 24ème Conférence sur le Traitement Automatique des Langues Naturelles*, Orléans, France, juin 2017a. URL https://hal.archives-ouvertes.fr/hal-01599572.

Loïc Vial, Benjamin Lecouteux et Didier Schwab : Sense Embeddings in Knowledge-Based Word Sense Disambiguation. *In 12th International Conference on Computational Semantics*, Montpellier, France, septembre 2017b. URL https://hal.archives-ouvertes.fr/hal-01599685.

Loïc Vial, Benjamin Lecouteux et Didier Schwab : UFSAC : Unification of Sense Annotated Corpora and Tools. Research report, UGA - Université Grenoble Alpes, décembre 2017c. URL https://hal.archives-ouvertes.fr/hal-01680739.

Loïc Vial, Benjamin Lecouteux et Didier Schwab : Uniformisation de corpus anglais annotés en sens. *In 24ème Conférence sur le Traitement Automatique des Langues Naturelles*, Orléans, France, juin 2017d. URL https://hal.archives-ouvertes.fr/hal-01599578.

Loïc VIAL, Benjamin LECOUTEUX et Didier SCHWAB : Approche supervisée à base de cellules LSTM bidirectionnelles pour la désambiguïsation lexicale. *In 25e conférence sur le Traitement Automatique des Langues Naturelles*, Rennes, France, mai 2018a. URL https://hal.archives-ouvertes.fr/hal-01781183.

Loïc VIAL, Benjamin LECOUTEUX et Didier SCHWAB : UFSAC : Unification of Sense Annotated Corpora and Tools. *In Language Resources and Evaluation Conference (LREC)*, Miyazaki, Japan, mai 2018b. URL https://hal.archives-ouvertes.fr/hal-01718237.

Loïc VIAL, Benjamin LECOUTEUX et Didier SCHWAB : Approche supervisée à base de cellules LSTM bidirectionnelles pour la désambiguïsation lexicale. *journal Traitement Automatique des Langues*, 2019a. URL https://hal.archives-ouvertes.fr/hal-02010901.

Loïc VIAL, Benjamin LECOUTEUX et Didier SCHWAB : Compression de vocabulaire de sens grâce aux relations sémantiques pour la désambiguïsation lexicale. *In TALN 2019 (Conférence sur le Traitement Automatique des Langues Naturelles)*, Toulouse, France, juillet 2019b. URL https://hal.archives-ouvertes.fr/hal-02127237.

Loïc VIAL, Benjamin LECOUTEUX et Didier SCHWAB : Sense Vocabulary Compression through the Semantic Knowledge of WordNet for Neural Word Sense Disambiguation. *In Global Wordnet Conference*, Wroclaw, Poland, 2019c. URL https://hal.archives-ouvertes.fr/hal-02131872.

Loïc VIAL, Andon TCHECHMEDJIEV et Didier SCHWAB : Extension lexicale de définitions grâce à des corpus annotés en sens. *In 23ème Conférence sur le Traitement Automatique des Langues Naturelles*, Paris, France, juillet 2016. URL https://hal.archives-ouvertes.fr/hal-01332850.

Loïc VIAL : Word sense disambiguation : improvement and integration in machine translation. Mémoire de D.E.A., Master of Science in Informatics at Grenoble, 2016.

Loïc VIAL : *Modèles neuronaux joints de désambiguïsation lexicale et de traduction automatique*. thèse, Université Grenoble Alpes [2020-....], juillet 2020. URL https://tel.archives-ouvertes.fr/tel-03033118.

Loïc VIAL, Andon TCHECHMEDJIEV et Didier SCHWAB : Comparison of global algorithms in word sense disambiguation, 2017e.

David Vickrey, Luke Biewald, Marc Teyssier et Daphne Koller : Word-sense disambiguation for machine translation. *In Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, pages 771–778, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics. URL http://dx.doi.org/10.3115/1220575.1220672.

Piek Vossen, Attila Görög, Fons Laan, Maarten Van Gompel, Rubén Izquierdo-Bevia et Antal Van Den Bosch : Dutchsemco : building a semantically annotated corpus for dutch. *In Electronic lexicography in the 21st century : New Applications for New Users : Proceedings of eLex 2011, Bled, 10-12 November 2011*, pages 286–296, 2011.

Ivan Vulić : Cross-lingual syntactically informed distributed word representations. *In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics : Volume 2, Short Papers*, pages 408–414, Valencia, Spain, avril 2017. Association for Computational Linguistics. URL https://aclanthology.org/E17-2065.

Henriette Walter : *L'aventure des mots français venus d'ailleurs*. R. Laffont, Paris, 1997. ISBN 2-221-08275-3.

Henriette Walter : Va-et-vient lexicaux : l'exemple de l'anglais et du français. *Dilbilim*, pages 103 – 112, 2012. ISSN 0255-674X.

Tonio Wandmacher, Jean-Yves Antoine, J.-P. Departe et Franck Poirier : SIBYLLE, an Assistive Communication System Adapting to the Context and its User. *ACM - Transactions on Speech and Language Processing*, 1(1):1–30, 2008. URL https://hal.archives-ouvertes.fr/hal-00516757.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy et Samuel R Bowman : Superglue : A stickier benchmark for general-purpose language understanding systems. *arXiv preprint arXiv :1905.00537*, 2019.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy et Samuel Bowman : GLUE : A multi-task benchmark and analysis platform for natural language understanding. *In Proceedings of the 2018 EMNLP Workshop BlackboxNLP : Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium, novembre 2018. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W18-5446.

Changhan Wang, Morgane Rivière, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino et Emmanuel Dupoux : Voxpopuli : A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation, 2021.

Anna Wierzbicka : *Semantics : Primes and Universals*. Oxford University Press, 1996.

Mark D Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E Bourne *et al.* : The fair guiding principles for scientific data management and stewardship. *Scientific data*, 3, 2016.

Andreas Witt, Ulrich Heid, Felix Sasaki et Gilles Sérasset : Multilingual language resources and interoperability. *Language Resources and Evaluation*, 43(1):1–14, 2009. URL https://hal.archives-ouvertes.fr/hal-00953703.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest et Alexander Rush : Transformers : State-of-the-art natural language processing. *In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing : System Demonstrations*, pages 38–45, Online, octobre 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.emnlp-demos.6.

Liang Xu, Hai Hu, Xuanwei Zhang, Lu Li, Chenjie Cao, Yudong Li, Yechen Xu, Kai Sun, Dian Yu, Cong Yu, Yin Tian, Qianqian Dong, Weitang Liu, Bo Shi, Yiming Cui, Junyi Li, Jun Zeng, Rongzhao Wang, Weijian Xie, Yanting Li, Yina Patterson, Zuoyu Tian, Yiwen Zhang, He Zhou, Shaoweihua Liu, Zhe Zhao, Qipeng Zhao, Cong Yue, Xinrui Zhang, Zhengliang Yang, Kyle Richardson et Zhenzhong Lan : CLUE : A Chinese language understanding evaluation benchmark. *In Proceedings of the 28th International Conference on Computational Linguistics*, pages 4762–4772, Barcelona, Spain (Online), décembre 2020. International Committee on Computational Linguistics. URL https://aclanthology.org/2020.coling-main.419.

Mengjia Xu : Understanding graph embedding methods and their applications. *CoRR*, abs/2012.08019, 2020. URL https://arxiv.org/abs/2012.08019.

Zhilin YANG, Zihang DAI, Yiming YANG, Jaime CARBONELL, Ruslan SALAKHUTDINOV et Quoc V. LE : Xlnet : Generalized autoregressive pretraining for language understanding. *arXiv preprint arXiv :1906.08237*, 2019.

Jason YOSINSKI, Jeff CLUNE, Yoshua BENGIO et Hod LIPSON : How transferable are features in deep neural networks ? *In* Z. GHAHRAMANI, M. WELLING, C. CORTES, N. LAWRENCE et K. Q. WEINBERGER, éditeurs : *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. URL https://proceedings.neurips.cc/paper/2014/file/375c71349b295fbe2dcdca9206f20a06-Paper.pdf.

Dayu YUAN, Julian RICHARDSON, Ryan DOHERTY, Colin EVANS et Eric ALTENDORF : Semi-supervised word sense disambiguation with neural models. *In COLING 2016*, 2016.

Marcely ZANON BOITO, William N HAVARD, Mahault GARNERIN, Éric LE FERRAND et Laurent BESACIER : MaSS : A Large and Clean Multilingual Corpus of Sentence-aligned Spoken Utterances Extracted from the Bible. *In Proceedings of The 12th Language Resources and Evaluation Conference*, pages 6486 – 6493, Marseille, France, mai 2020. URL https://hal.archives-ouvertes.fr/hal-02611059.

Zhi ZHONG et Hwee Tou NG : It makes sense : A wide-coverage word sense disambiguation system for free text. *In Proceedings of the ACL 2010 System Demonstrations*, ACLDemos '10, pages 78–83, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics. URL http://dl.acm.org/citation.cfm?id=1858933.1858947.

Yukun ZHU, Ryan KIROS, Rich ZEMEL, Ruslan SALAKHUTDINOV, Raquel URTASUN, Antonio TORRALBA et Sanja FIDLER : Aligning books and movies : Towards story-like visual explanations by watching movies and reading books. *2015 IEEE International Conference on Computer Vision (ICCV)*, Dec 2015. URL http://dx.doi.org/10.1109/ICCV.2015.11.

Michael ZOCK : Sorry, what was your name again, or how to overcome the tip-of-the tongue with the help of a computer ? *In SemaNet'02 : Building and Using Semantic Networks*, Taipei, Taiwan, 2002.

Michael ZOCK et Didier SCHWAB : Storage does not Guarantee Access : The Problem of Organizing and Accessing Words in a Speaker's Lexicon. *Journal of Cognitive Science*, 12:233–258, 2011. URL https://hal.archives-ouvertes.fr/hal-00953672. (Impact-F 3.52 estim. in 2012).

Michael Zock et Didier Schwab : Chercher un mot dans un dictionnaire sans bon index est un peu comme s'orienter sur une île déserte sans carte convenable. *Linguisticae Investigationes*, pages –, 2013. URL https://hal.archives-ouvertes.fr/hal-00953650.

Michael Zock et Didier Schwab : Le fait d'avoir stocké des mots garantit nullement leur accès. *In* Michael Zock, Gemma Bel-Enguix et Reinhard Rapp, éditeurs : *Ressources Lexicales et Traitement de la Langue, atelier TALN-2014*, pages 221–230, Marseille, France, 2014. TALN. URL https://hal.archives-ouvertes.fr/hal-01480407.

Ozan İrsoy, Adrian Benton et Karl Stratos : kōan : A corrected cbow implementation, 2020.