Université de Montréal

Méthodologies pour la détection de diachronies sémantiques et leurs impacts

par

David Kletz

Département de mathématiques et de statistique
Faculté des arts et des sciences

Mémoire présenté en vue de l'obtention du grade de
Maître ès sciences (M.Sc.)
en Intelligence Artificielle

August 26, 2021

Université de Montréal

Faculté des arts et des sciences

Ce mémoire intitulé

Méthodologies pour la détection de diachronies sémantiques et leurs impacts

présenté par

David Kletz

a été évalué par un jury composé des personnes suivantes :

Nadia El-Mabrouk
(président-rapporteur)

Philippe Langlais
(directeur de recherche)

Patrick Drouin
(codirecteur)

Jian-Yun Nie
(membre du jury)

Résumé

Le sens d'un mot est sujet à des variations au cours du temps. Nombre de phénomènes motivent ces modifications comme l'apparition de nouveaux objets ou les changements d'habitudes. Ainsi, un même mot peut se voir assigner un nouveau sens, retirer un sens, ou encore rester stable entre deux dates.

L'étude de la diachronie sémantique est un domaine s'intéressant à ces changements de sens. Les récents travaux sur la diachronie sémantique proposent des méthodologies pour le repérage de diachronies. Pour ce faire, ils s'appuient sur des textes issus de plusieurs périodes temporelles différentes, et grâce auxquels sont entrainés des modèles de langue. Un alignement des représentations obtenues, et une comparaison de celles de mots-cibles leur permet de conclure quant à leur changement de sens. Néanmoins, l'absence de jeu de données (*dataset*) de référence pour la validation de ces méthodes conduit au développement de méthodes de validation alternatives, suggérant notamment de s'appuyer sur les changements de sens recensés dans les dictionnaires traditionnels.

Le travail réalisé au cours de ma maitrise s'attache à exposer une réflexion sur les méthodes existantes de repérage des diachronies. En nous appuyant sur un corpus journalistique couvrant l'ensemble du **XX**^{ème} siècle, nous proposons des méthodes complémentaires grâce auxquelles nous démontrons que les évaluations proposées font l'objet d'ambiguïtés. Celles-ci ne permettent dès lors pas de conclure quant à la qualité des méthodes. Nous nous sommes ensuite attachés à développer une méthodologie pour la construction d'un jeu de données de validation. Cette méthodologie tire parti d'un algorithme de désambiguïsation afin d'associer à tous les sens recensés d'un mot une date d'apparition au cours du temps. Nous proposons un jeu de données composé de 151 mots permettant d'évaluer le repérage de diachronies en français entre 1910 et 1990.

Mots clés : TAL, diachronie, évaluation, validation, jeu de données, désambiguïsation de sens

Abstract

The meaning of a word is subject to variations over time. Many phenomena motivate these modifications such as the appearance of new objects or changes in habits. Thus, the same word can be assigned a new meaning, or have a meaning withdrawn, or remain stable between two dates.

The study of semantic diachrony is a field that focuses on these changes in meaning. Recent work on semantic diachrony proposes methodologies for the detection of diachronies. In order to do so, they rely on texts from several different temporal periods, and through which language models are trained. An alignment of the obtained representations, and a comparison of those of target words enables one to infer the change of meaning. Nevertheless, the absence of a reference dataset for the validation of these methods leads to the development of alternative validation methods, suggesting in particular to rely on the changes of meaning identified in traditional dictionaries.

The work carried out during my master's degree aims at presenting a reflection on the existing methods of diachrony detection. Based on a corpus of newspapers covering the whole 20th century, we propose complementary methods thanks to which we demonstrate that the proposed evaluations are subject to ambiguities. These ambiguities do not allow us to ensure the quality of the methods. We then develop a methodology for the construction of a validation dataset. This methodology takes advantage of a disambiguation algorithm in order to associate a date of appearance in the course of time to all the senses of a word. We propose a dataset composed of 151 words allowing one to evaluate the identification of diachronies in French between 1910 and 1990.

Key words : NLP, diachrony, evaluation, validation, dataset, Word sens disambiguation (WSD)

Table des matières

Liste des tableaux

Liste des figures

Liste des sigles et des abréviations

IA	Intelligence Artificielle
ML	Machine Learning
OCR	Optical character recognition
PCA	Principal Component Analysis
PPMI	Positive Pointwise Mutual Information
SVC	Support Vector Classification
SVD	Single Value decomposition
TAL / TALN	Traitement automatique des langues / langages naturels
WSD	Word Sens Disambiguation

Remerciements

Il y a presque deux ans de cela je découvrais le traitement automatique du langage naturel dans un cours enseigné par Philippe Langlais, qui a joué un rôle déterminant dans l'orientation qu'ont depuis pris mes études tant ce domaine m'a captivé. Je ne peux qu'exprimer toute ma gratitude à Philippe, qui n'a depuis cessé de m'aiguiller. L'avoir comme directeur de recherche a été non seulement une félicité pour la qualité de son encadrement, mais également un plaisir d'échanges et d'apprentissage. La relation de confiance qui s'est immédiatement installée a grandement contribué à la qualité de direction de recherche.

Il va sans dire que l'excellence de mon encadrement a largement reposé sur les mérites de mes co-encadrants de l'Observatoire de Linguistique Sens-Texte : François Lareau et Patrick Drouin. Ce duo aussi original que savant, aussi complémentaire que chaleureux, a été un véritable mentor dans ma découverte des arcanes de la diachronie, et leur connaissance profonde des processus lexicographiques et sémantiques autant que nos discussions assidues ont guidé la progression de mes recherches. Mon appétence pour la linguistique, c'est à eux que je la dois.

Enfin, et en dépit du Covid, les occasions ont jalonné cette année au Rali de rencontrer et discuter avec tous les membres du laboratoire. Leurs conseils et encouragements ont été de précieux moments de convivialité et de réflexion.

Je conclue en remerciant l'UdeM dans son ensemble : cette université est un véritable paradis pour qui a soif d'étudier, tout y est dévotion pour les étudiants, l'apprentissage est un plaisir et la route qu'on y trace est pavée par la réussite. Je ne peux que lui souhaiter de rester ainsi.

Maintenant il vouloit decimer ses propres citoyens.

Amyot

(traduction de Vie de Camille de Plutarque)

Chapitre 1

Introduction

1.1. Contexte

L'origine de ce mémoire se situe dans le domaine de la lexicographie : l'Observatoire de linguistique Sens-Texte (OLST)[1] de l'UdeM cherche à développer un outil de veille permettant de notifier les lexicographes de l'existence du changement de sens d'un mot non repéré ni intégré aux dictionnaires.

Le terme de diachronie caractérise en linguistique un fait ayant connu une rupture au cours du temps. Notre sujet se place dans une étude du sens des mots et traite de ce fait de diachronie sémantique, dont l'axiome central s'énonce ainsi :

La signification associée à un mot (c'est-à-dire son sens) évolue au cours du temps.

Cette définition se déclinant également selon la formulation suivante : un mot donné revêt des significations différentes dépendamment de l'époque à laquelle il est employé. Ces changements de sens peuvent relever de l'apparition d'un nouveau sens, de la disparition ou encore d'un glissement de l'utilisation de sens.

Pour illustrer cette notion, amorçons une première étude diachronique avec la comparaison des sens du mot-forme "souris" (mot s'écrivant "souris") entre 1900 et 2021 : le premier sens répertorié de ce mot désigne un "*Petit mammifère rongeur omnivore de la famille des Muridés*" [2], sens existant déjà en 1900 et encore présent aujourd'hui.

A l'inverse, un second sens d'utilisation de ce mot-forme désigne un "*périphérique d'entrée relié à l'ordinateur et permettant, en guidant le déplacement du curseur sur l'écran, de sélectionner une commande ou une option, sans passer par le clavier de l'ordinateur*", sens qui pour sa part apparaît au cours du $\mathbf{XX}^{ème}$ siècle et n'a pas cours en 1900. Cette succincte

[1] http://www.olst.umontreal.ca

[2] Toutes les définitions lexicopgraphiques françaises citées dans ce mémoire sont issues du Trésor de la Langue Francaise Informatisé (TLFI) [**3**]

analyse caractérise ainsi un mot ayant subi une diachronie sémantique entre 1900 et 2021, prenant forme de l'apparition d'un nouveau sens.

Néanmoins, si cette analyse met en évidence un net changement de sens, la diachronie est un phénomène continu et par conséquent latent : le vocabulaire d'une langue vivante est au cœur d'un mouvement rebattant constamment sa signification, et c'est cette caractéristique qui rend laborieux le travail de lexicographie.

Intéressons nous à présent au mot-forme "producteur" : en 1900 celui-ci ne désigne qu'une "*Personne, société, firme qui engendre des biens, qui les commercialise ou qui assure certains services (p.oppos. à celui qui s'en sert, qui consomme).*", tandis que ce même mot-forme désigne également en 2021 une personne responsable du financement d'un spectacle, et en particulier d'un film : "*Personne, société qui assure le financement, la constitution de l'équipe de techniciens, le choix du metteur en scène d'un film*". Or ces deux significations désignent une activité de commercialisation et de production. Dès lors, une certaine ambiguïté s'immisce : le lexicographe doit-il ou non considérer ces sens comme définitions distinctes du mot ?

L'objectif du projet est de repérer ce type de changements survenus dans l'utilisation d'un mot au cours du temps et les notifier aux lexicographes. Et dès lors, les recherches en lien avec ce projet ont pour ambition d'étudier les méthodes de quantification des changements de sens de mots.

Si l'étude des mécanismes de diachronie relève davantage de la linguistique que du Traitement automatique du langage naturel (TALN), un champ d'étude de ce domaine s'intéresse au développement de techniques liées à la diachronie sémantique et par conséquent, leur utilisation dans le cadre du projet a pour objectif final une application à un vocabulaire large dans le but d'identifier des mots ayant subi de tels changements non recensés dans les dictionnaires.

Les recherches bibliographiques que nous avons effectuées nous ont permis de mettre en évidence des méthodes ambitieuses d'identification de diachronies autant qu'elles nous ont conduits à remarquer l'absence notable de jeu de données de référence permettant l'évaluation des méthodes existantes, et rendant difficile d'appréhender la qualité de celles-ci.

Or une analyse qualitative de ces travaux nous a permis de mettre en évidence un paradoxe dans les méthodologies employées : ces méthodes aspirent à repérer des diachronies non notifiées par les dictionnaires, et s'appuient sur ces mêmes dictionnaires comme outil de validation des méthodes.

Dans ce cadre, notre contribution au cours de ce mémoire est de mettre en exergue les ambiguïtés qu'entrainent tant l'absence de jeu de données que les méthodes alternatives de

validation, puis de proposer une méthodologie pour l'automatisation de la construction d'un jeu de données recensant des diachronies. Finalement, nous montrons que cette ambivalence est à l'origine d'une nécessité de réfléchir à la définition à adopter d'une diachronie.

En outre, si la littérature fait état de l'existence de quelques groupes de mots employés comme jeu de donnéess, le plus large d'entre eux ne comptait que 28 mots tous sélectionnés manuellement. Pour notre part, une première application de notre méthodologie nous a permis de construire un jeu de données de 151 exemples.

1.1.1. Définitions

Cette section propose une définition du vocabulaire technique et propre au domaine de la diachronie. Les définitions ci-dessous présentent en guise d'illustration d'autres définitions lexicographiques : celles-ci sont toutes issues du Trésor de la Langue française (TLFI) [**3**].

19xx-yy : Période de 10 ans s'étendant de 19xx à 19yy. La période 1960-70, désigne par exemple la période s'étendant de 1960 à 1970. Par extension, nous désignons la période s'étendant de 1990 à 2000 par "1990-00".

Diachronie : La diachronie se définit dans le cadre d'une étude temporelle. Elle caractérise en linguistique un fait connaissant une rupture au cours du temps. Un fait linguistique est alors dit diachronique si sa valeur durant une époque E_1 est non strictement égale à la sienne durant l'époque E_2.

Diachronie sémantique : Changement des sens utilisés d'un mot au cours du temps. Ce changement peut être de nature variée :

- apparition d'un nouveau sens, par exemple le mot "souris" qui se voit attribuer avec l'essor de l'informatique le sens de "*Périphérique d'entrée relié à l'ordinateur par un cordon et permettant, en guidant le déplacement du curseur sur l'écran, de sélectionner une commande ou une option, sans passer par le clavier de l'ordinateur*".
- disparition d'un sens en usage.
- glissement des usages des sens d'un mot : par exemple le mot "producteur" qui n'est utilisé en 1900 que pour désigner une personne responsable d'une production industrielle. "*Personne, société, firme qui engendre des biens, qui les commercialise ou qui assure certains services (p.oppos. à celui qui s'en sert, qui consomme).* ", et aujourd'hui désigne à part égal une personne responsable du financement d'un spectacle, et enparticulier d'un film : "*Personne, société qui assure le financement, la constitution de l'équipe de techniciens, le choix du metteur en scène d'un film*".

Époque : Terme désignant la période historique rattachée à l'intervalle de temps nommé *tranche* (voir définition) : l'époque caractérise donc au sens historique le temps sur lequel s'étale cette tranche.

Lexicographie : Domaine de la linguistique travaillant sur les mots et visant à les identifier et les définir. Les dictionnaires constituent dès lors des recueils lexicographiques.

Millésime (d'un dictionnaire) : Publication annuelle d'un dictionnaire. Chaque millésime enregistre l'entrée de nouveaux mots. De plus, il apporte des corrections dans les articles des mots présents dans le millésime précédent ; celles-ci peuvent relever de la modification de définition, ou de l'ajout et suppressions de définitions.

Monosémie/monosémique : Se dit d'un mot n'étant utilisé que dans un seul sens. Dans le cadre de ce travail, nous considérons comme "monosémique" un mot auquel n'est associée qu'une seule définition.

Mot graphique ou Mot-forme : Cette notion désigne un mot sans porter attention à son sens. Ainsi, un mot-forme ne se caractérise que par son écriture : deux mots dont l'orthographe est identique sont un même mot-forme.
Prenons comme exemple les occcurences dans un texte du mot-forme "aplanir" : celles-ci désignent toutes les apparitions de ce verbe, qu'il soit utilisé dans le sens de "*Rendre plan, uni, en faisant disparaître les inégalités*" ou "*S'atténuer, disparaître, s'uniformiser*".
Pour sa part, "mot" désigne une forme associée à un sens donné. D'après l'exemple précédent, nous pouvons donc dénombrer deux mots "applanir" (correspondant respectivement aux deux définitions proposées).

Plongement de mot (ou embedding): Représentation vectorielle d'un mot. Les vecteurs représentant les mots d'un corpus sont de même dimension. Leur valeur est attribuée par un modèle apprenant les représentations à partir de l'analyse des usages des mots du vocabulaire dans les textes composant un corpus d'entrainement. Ainsi, l'analyse du vecteur représentant un mot est porteur d'information sémantique sur celui-ci.

Rapport d'utilisation (d'un sens) : Le rapport d'utilisation d'un sens quantifie en pourcentage le taux d'utilisation d'un sens s d'un mot m parmi tous les sens recensés de ce mot.

Pour le calculer, nous dénombrons dans un corpus les occurrences du mot m (nombre noté *total*), et parmi celles-ci, nous recensons les n_s occurrences du sens s. En notant R_s ce rapport nous avons alors:

$$\boxed{R_s = \frac{n_s}{total}} \tag{1.1.1}$$

Stable (mot) :
Sur une période donnée séparant deux tranches, mot dont le sens ne connait pas de modifications majeures, c'est-à-dire ni ajout ni suppression de sens, ni encore de changement important (plus de 30%) de rapport d'utilisation d'un sens.
Exemple : le mot "excédent" conserve entre 1910 et 1990 pour seule définition : "Quantité de quelque chose qui dépasse une quantité donnée (longueur, volume); ce qui se trouve en surplus".

Tranche ou Strate: S'agissant de traiter de modifications temporelles, nous définissons le terme de tranche (ou strate) comme un corpus de texte issu d'une même période temporelle donnée, qui sera considérée comme l'unité de temps de notre travail. Ainsi, un corpus étendu sur une période de 100 ans pourra être découpé en 10 tranches contenant chacune respectivement tous les textes du corpus datant de 10 périodes de 10 ans chacune.

1.2. Etat de l'art : approche typique

1.2.1. Analyse diachronique

Les travaux en sémantique diachronique ont pour objectif de proposer une analyse du changement de sens d'un mot entre deux périodes sélectionnées.

Parmi les publications traitant du sujet, un article rédigé par Hamilton et al. [**8**] est en parfaite adéquation avec notre sujet, et constitue à ce titre la référence que nous utiliserons au cours de ce mémoire.

Cette publication présente pour objectif de proposer la définition de lois de changements sémantiques, c'est-à-dire prédisant les taux de changements de sens de mots et certaines caractéristiques de ceux-ci. Elle se découpe en deux parties : dans un premier temps Hamiton et al. proposent une approche typique d'identification de diachronies entre deux époques sélectionnées. Puis une seconde partie décrit l'application de ces méthodes à un large vocabulaire afin d'en déduire les lois de changements sémantiques.

Cet article nous permet d'amorcer cette étude en soumettant l'examen d'une figure dont elle est issue et décrivant parfaitement le résultat d'un processus d'analyse diachronique :

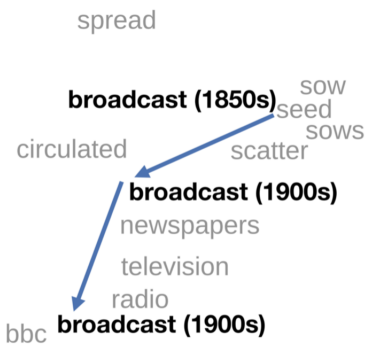


Fig. 1.1. Figure extraite de l'article d'Hamilton et al. [**8**] présentant le déplacement du vecteur du mot "broadcast" entre 1850 et 1990. Les mots en gris représentent l'environnement du mot "broadcast" aux différentes époques.

La figure 1.1 propose une visualisation en 2 dimensions des changements sémantiques du mot "broadcast" : elle affiche les modifications de l'utilisation de ce mot entre les années 1850 et 1990. Une analyse de cette figure indique par conséquent le parcours sémantique du mot "broadcast" : si ce mot désigne au milieu du XIX^{e} siècle l'action d'essaimer des graines, nous observons que l'invention des moyens de communication que sont la radio et la télévision au début du $\mathbf{XX}^{\text{ème}}$ siècle déplace ce mot dans un contexte de transmission de signal ("émettre" en français) au cours des périodes étudiées.

L'exemple de la figure 1.1 illustre ainsi le contexte dans lequel se place le développement de méthodes d'identification de diachronie : le suivi historique. Plus précisément, elle met en évidence les deux étapes de la comparaison des sens au cours du temps :

(1) Identification de l'utilisation d'un mot au cours des périodes étudiées, ici les trois périodes 1850, 1900 et 1990.

(2) Comparaison des sens relevés entre les différentes périodes étudiées, illustrée ici par les mots en gris, représentant les mots les plus similaires à "broadcast" à chaque époque, et variant du lexique de la plantation à celui des médias.

Afin de réaliser une identification statique du contexte d'utilisation, l'article préconise de s'appuyer sur :

- Un corpus composé de textes dont les dates respectives de rédaction couvrent la période à laquelle s'intéressent les travaux : c'est à partir de ces textes que les usages statiques seront déduits.
- Une représentation distributionnelle des mots : cette technique classique du TAL ambitionne d'associer un vecteur à chaque mot de vocabulaire ; la comparaison des vecteurs permet alors d'effectuer la comparaison des usages.

1.2.2. Corpus

La première étape consiste donc en une sélection du corpus de textes utilisés pour la connaissance des usages et du découpage du corpus. L'analyse de la diachronie d'un mot s'appuie sur ce corpus afin de comparer les différences d'usage de ce mot. Ainsi, il est nécessaire que ce corpus contienne des textes rédigés tout au long de la période historique étudiée. Les changements sémantiques survenus qui seront identifiés dépendent de la période sélectionnée : plus celle-ci est longue, plus le nombre et la visibilité des phénomènes seront importants. Néanmoins, Mair et al. [**11**] indiquent qu'une période de 30 ans est déjà suffisante pour repérer un changement syntaxique en anglais.

Plusieurs corpus standards respectent cette spécificité, parmi lesquels sont le Corpus of Historical American English (COHA) [**2**] ou encore le Google N-Gram corpus [**7**] tous deux utilisés par Hamilton et al. et couvrant une période de 200 ans s'étalant de 1800 à 1999.

La comparaison nécessite pour sa part un découpage : le corpus est séparé en sous-corpus contenant respectivement tous les textes issus d'une sous-période. La durée des sous-périodes doit être identique, et celles-ci doivent être contigues et non chevauchantes. Les sous-corpus obtenus sont nommés **strates** ou **tranches**. La période historique couverte par une strate constitue l'unité temporelle du travail et donc la précision maximale de l'analyse diachronique. Hamilton et al. découpent ainsi leurs corpus en 20 tranches de 10 ans.

1.2.3. Plongements de mots et contexte d'utilisation d'un mot

L'analyse du contexte d'utilisation d'un mot au sein d'une tranche passe par l'utilisation de divers algorithmes. Ceux-ci ne peuvent prendre en entrée que des données numériques. Afin de transformer les données dont nous disposons (c'est à dire les mots composant les textes des tranches) en données numériques, nous nous appuyons sur le développement d'un modèle de plongement de mots.

Cette technique classique du TAL propose d'entrainer un modèle afin d'attribuer une représentation numérique à chaque mot d'un vocabulaire. Ces représentations étant en plusieurs dimensions, il s'agit de vecteurs. Et chaque mot de vocabulaire se voit ainsi attribuer une représentation numérique unique.

Afin de parvenir à développer des représentations vectorielles qui soient porteuses d'information sémantique sur les mots qu'elles représentent respectivement, les valeurs de ces vecteurs sont calculées à partir d'une analyse des textes d'entrainement.

Parmi les méthodes utilisées pour le développement des représentations, Hamilton et al. présentent des méthodes classiques : positive point-wise mutual information (PPMI) et single value decomposition (SVD). En outre, ils ont également recourt aux méthodes neuronales au travers de Word2Vec ([**13**]). Le détail de ce processus d'obtention du contexte

d'un mot à partir de cette représentation est décrit dans le chapitre 4. C'est à partir de ce contexte qu'est déduit le sens d'utilisation d'un mot dans une tranche. La figure 1.2 expose une visualisation du contexte obtenu grâce à l'entrainement d'un modèle Word2Vec sur la tranche 1850-60 du corpus utilisé par Hamilton et al. Cette figure met en évidence un sens d'utilisation rattaché au domaine de l'agriculture et de l'élevage bovin.

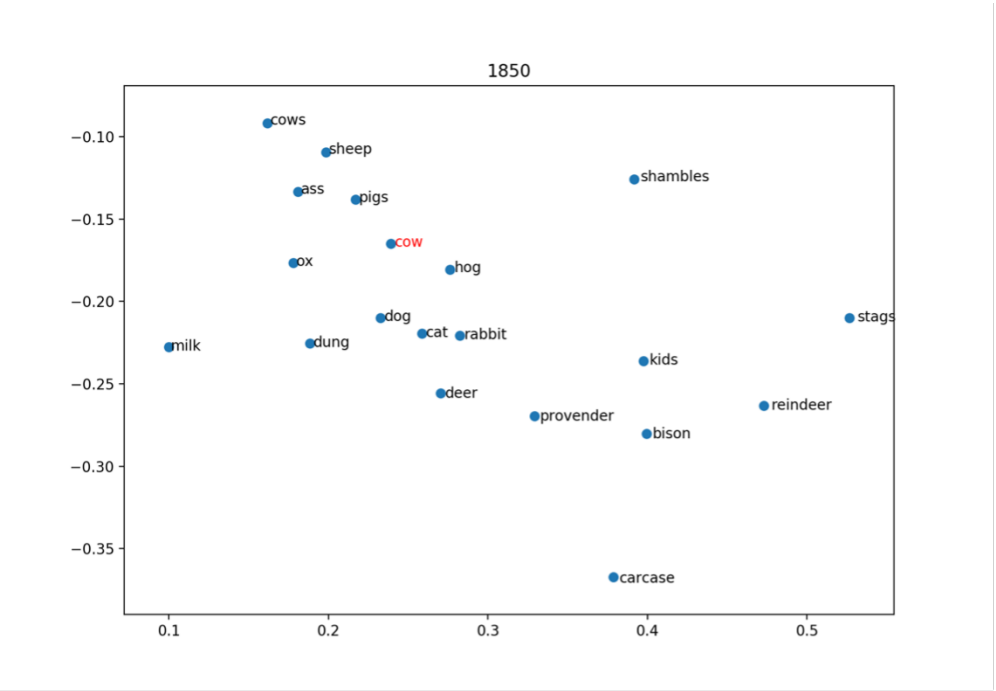


Fig. 1.2. Représentation du contexte du mot "cow" dans le modèle Word2Vec entrainé par Hamilton et al. dans la tranche 1850-60. Nous avons construit cette figure en calculant les coordonnées des vecteurs grâce à une projection dans un espace 2D (dont la construction est décrite au chapitre 2).

1.2.4. Série temporelle

La dernière étape de la méthodologie d'analyse diachronique constitue le suivi des mots étudiés au travers d'une série temporelle. Afin d'opérer un suivi des changements de sens d'un mot, des modèles de plongements de mot sont entrainés sur chacune des tranches du corpus. 20 modèles différents sont ainsi définis qui proposent chacun une représentation de chaque mot de vocabulaire et reflètent alors le sens de chaque mot dans la tranche d'entrainement. Pour conserver la valeur sémantique des zones de l'espace, les modèles appris sur chaque tranche sont alignés, c'est à dire qu'ils subissent une transformation mathématique permettant d'aligner l'origine et les axes de leurs espaces respectifs.

A l'issue de cet entrainement, Hamilton et al. définissent une série temporelle à partir de laquelle seront indiquées les diachronies. Cette série analyse deux points :

- Le déplacement du vecteur représentant un mot entre deux strates consécutives : un déplacement indique une modification de la valeur sémantique du mot considéré.
- Le changement du lexique des voisins les plus proches.

Ainsi, reprenons l'illustration (figure 1.3) du changement de sens d'un mot issu de l'article.

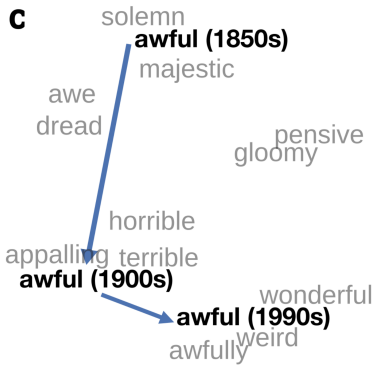


Fig. 1.3. Illustration tirée de l'article d'Hamilton et al. [**8**] représentant les changements sémantiques du mot "awful" entre les périodes 1850-60 et 1990-00.

Le suivi du mot "awful" montre bien que le suivi temporel s'appuyant sur les deux points présentés ci-dessus indique la diachronie subie par un mot :

- Les flèches bleues indiquent le parcours spatial subi par "awful" entre les périodes 1850-60 et 1990-00. Le changement d'utilisation du mot entre les textes de ces strates a provoqué une modification de la représentation apprise de ce mot.
- Les voisins les plus proches de "awful" représentés en gris varient de "solemn" et "majestic" en 1850-60 à "awfully" et "weird" en 1990-00, reflétant les nouveaux usages de ce mot.

1.2.5. Validation

La mise en pratique des méthodes développées est nécessairement précédée d'une étape de validation permettant de quantifier la performance des méthodes développées.

La validation de la diachronie s'appuie ainsi sur l'application des méthodes à un groupe de mots témoins dont les changements de sens sont connus sur les époques étudiées, et catégorisés selon ces changements. C'est enfin une comparaison entre les catégories connues et les résultats proposés par les méthodes qui permet de conclure quant à leur performance.

L'absence de jeu de données de référence amène les auteurs à proposer diverses approches de validation. Hamilton et al. proposent ainsi une première étape de validation afin de vérifier si leurs méthodes parviennent à capturer les diachronies connues. Pour ce faire, ils s'appuient sur un groupe de 28 mots dont le changement de sens est considéré comme

connu sur la période d'étude de leur recherche. Ces mots sont sélectionnés manuellement et proviennent de deux sources :

- les précédents travaux sur la diachronie proposent parmi leurs résultats des groupes de mots ayant subi une diachronie (tels que les articles de Wijaya et Yeniterzi [**20**], Jatowt et Duh [**9**])
- le dictionnaire classique Oxford English Dictionary (OED, [**16**]) : les définitions de cet ouvrage sont constellées d'indications lexicographiques parmi lesquelles le mot-clé "obsolete" qui indique la disparition de l'usage d'une définition, et ainsi considéré comme ayant subi une diachronie.

La validation mise en place par la suite propose d'appliquer les différentes approches distributionnelles à ces mots et vérifier si leur déplacement correspond au changement de sens attendu, c'est-à-dire si le plongement associé subi un déplacement spatial et si les environnements sémantiques au cours des périodes correspondent bien aux domaines d'utilisation prévus.

En sus, les auteurs proposent une seconde étape de validation visant cette fois-ci à vérifier si les méthodes parviennent à découvrir des changements de sens à partir des données. Ils proposent ainsi d'établir la liste des 10 mots ayant subi les changements de sens les plus importants selon leur méthodologie, puis d'en référer de nouveau à la littérature existant sur le sujet et en particulier à l'Oxford English Dictionary afin de confirmer ou infirmer le changement de sens proposé.

1.2.6. Evaluation des méthodes

La première étape de la validation de l'article d'Hamilton et al. s'intéresse donc à la concordance entre le changement de sens supposé d'une liste de 28 mots et le changement de sens analysé par leur méthode. Pour ce faire, les auteurs vérifient que chacun des mots étudiés se déplace dans l'espace sémantique vers les mots du lexique du nouveau sens. Par exemple, pour "broadcast" (voir figure 1.1), ils s'assurent que la représentation de "broadcast" se rapproche de celle de "radio" au cours du temps. De plus, ils proposent de vérifier si ce déplacement est statistiquement significatif. Cette étape est testée en faisant varier les modèles de plongement de mots (les trois évoqués en section 1.2.3) et les corpus utilisés (COHA ou Google N-Grams). Les résultats de cette étape rapportent un déplacement de 100% des mots évalués vers leur nouveau lexique dans tous les modèles sauf les PPMI entrainés sur Google N-Grams (96.9%). Néanmoins, nous remarquons des scores constamment inférieurs en vérifiant si les déplacements sont statistiquement significatifs.

La seconde étape est donc réalisée après avoir utilisé la méthodologie de séries temporelles selon les trois modèles sur tout le vocabulaire du corpus issu de Google N-Grams. Les 10 mots ayant subi le plus grand déplacement selon chaque modèle sont extraits et analysés pour juger de la qualité des méthodes.

Ces 10 mots extraits sont classés entre

- "vrais déplacements sémantiques" : les auteurs sont en accord avec la proposition d'un changement de sens.
- "cas ambigus" : mots dont le changement de sens n'est pas définitif, mais dont les contextes habituels d'utilisation ont évolué, par exemple à cause de changements sociétaux.
- "artefacts" : mots n'ayant pas changé de sens, mais dont le corpus et les méthodes sont responsables d'un déplacement.

Cette analyse des mots ayant subi les déplacements les plus importants signale des précisions bien inférieures à celles de la première étape. Ainsi, parmi les 10 mots évalués, un seul de ceux identifiés par la méthode PPMI est un "vrai déplacement sémantique", 4 de ceux identifiés par SVD et 7 par la méthode neuronale.

C'est cette différence entre les résultats des deux méthodes qui constitue le point de départ de notre réflexion.

1.3. Problématique

Etant donné que nous nous inscrivons dans un projet visant à utiliser des méthodes d'identification des diachronies, il nous est essentiel de pouvoir évaluer les capacités d'une méthode à proposer des résultats qualitatifs. Or, la lecture de l'étape de validation des méthodes nous a laissé entrevoir certaines lacunes, qui remettent ainsi en question l'évaluation des méthodes :

- Les méthodes de validation proposées font l'hypothèse que le déplacement du plongement d'un mot et son changement de voisinage sont une preuve de diachronie, or comment s'assurer du lien entre ces phénomènes ?
- L'identification de diachronies repose sur l'hypothèse que toutes les diachronies ne sont pas connues. Les mots repérés par les méthodes sont donc validés par des ressources dont nous ne pouvons garantir l'exactitude : certains mots notifiés comme stables peuvent avoir subi une diachronie non relevée.

Finalement, l'étude de ces méthodes nous a permis de mettre en exergue l'existence d'une dualité entre leurs aspirations et le choix des données de validation utilisées. Si les auteurs espèrent identifier de nouveaux changements de sens, l'utilisation comme référence des ressources existantes met en place une situation contradictoire : lors de la validation, l'apparition de termes non référencés comme ayant subi une diachronie est considérée comme une erreur.

Dès lors, nous devons nous concentrer sur une réflexion quant à la nature des phénomènes de diachronie sémantique que l'on doit viser. Et par conséquent, nous devons proposer la construction d'un jeu de données de référence qui permettrait d'évaluer rigoureusement la qualité d'une méthode d'identification de la diachronie.

Afin de proposer des réponses aux problématiques soulevées, nous proposons d'organiser notre mémoire selon le plan suivant :

Le chapitre introductif 1, décrit les enjeux liés à la diachronie et présente les travaux déjà réalisés dans la recherche de diachronies.

Le chapitre 2 est un chapitre technique au cours duquel sont détaillés les méthodologies et outils techniques utilisés au cours du mémoire.

Le chapitre 3 se propose de décrire le corpus utilisé et de détailler les choix effectués parmi les textes constituant celui-ci.

Au cours du chapitre 4 nous identifions les problèmes posés par l'évaluation des méthodes de repérage de diachronie. Nous décrivons ensuite des méthodes complémentaires d'identification de diachronies afin d'entreprendre des vérifications supplémentaires et nous permettant d'étayer nos interrogations. L'application de ces méthodes à notre corpus d'étude confirme les ambiguïtés relevées: de manière similaire aux résultats des articles étudiés, l'étude de mots révèle des termes considérés comme stables affichant des scores de diachronies élevés.

La mise en évidence de ces limites de la validation ne nous permettant pas de garantir la qualité des résultats, nous concluons en mettant en évidence le paradoxe que représente l'utilisation d'un dictionnaire pour labéliser les exemples et démontrer la nécessité de développer un jeu de données de référence.

Enfin, le chapitre 6 soumet une proposition pour remédier aux problèmes soulevés dans le chapitre 4. Nous y détaillons un processus d'élaboration de jeu de données recensant des mots ayant subi un changement de sens et avéré par l'usage de ces mots dans notre corpus d'étude. Cette méthode est automatique et repose sur une suite d'étapes toutes évaluées et validées. Étant donné la nature des problèmes soulevés précédemment, nous proposons que les mots contenus dans notre jeu de données soient identifiés eux-mêmes par une analyse des

usages afin d'éviter le recours à la datation issue de ressources dont nous faisons l'hypothèse qu'elles présentent des lacunes. Une analyse de désambiguïsation sur chaque tranche du corpus nous permet d'accéder aux sens utilisés dans chacune des périodes et en conséquent à cartographier les usages des mots par époque. Une confrontation des sens utilisés entre périodes nous permet de proposer une datation.

Enfin, nous obtenons une cartographie des sens utilisés des mots de vocabulaire. Un mot dont les mêmes sens sont utilisés au cours du temps est stable. A l'inverse un mot dont un sens est apparu ou a disparu a subi une diachronie.

Cette étape détaillant la construction d'un jeu de données s'accompagne d'une mise en pratique à partir de l'analyse de notre corpus d'étude, et nous permet de proposer un jeu de données en français de 151 mots labélisés "stables" ou "ayant subi une diachronie" entre les périodes 1910-20 et 1990-00, non issus d'un dictionnaire.

Chapitre 2

Outils

2.1. Outils de TAL

Nous présentons dans cette section les outils utilisés pour le développement des mesures présentées dans ce mémoire au cours du chapitre 4.

2.1.1. Plongements de mots

Afin d'obtenir des représentations des mots de vocabulaire des corpus, nous entrainons des modèles associant à chaque mot de vocabulaire un vecteur numérique. Les représentations ainsi obtenues nous permettent par la suite de proposer les méthodes de quantification de diachronie.

Les modèles utilisés sont entrainés par tranche : ils proposent une représentation des mots de vocabulaire pour chaque tranche de 10 ans du corpus. Nous nous appuyons sur un corpus [1] s'étendant de 1910 à 2000, ce qui nous permet d'entrainer 9 modèles différents, sur les 9 tranches de 10 ans du corpus.

Nous utilisons des modèles Word2Vec. Ceux-ci sont entrainés grâce à la bibliothèque python Gensim [**15**]. Les caractéristiques d'entrainement de nos modèles sont les suivantes :

- Taille des vecteurs : 300
- Échantillonnage négatif : 20
- Taux d'apprentissage initial alpha : 0.025
- Algorithme d'entrainement : skip-gram

2.1.2. Techniques de TAL

Les représentations obtenues sont ensuite employées par certaines techniques classiques de TAL que nous définissons dans cette section.

[1] Le corpus utilisé est présenté en détail dans le chapitre 3

La distance entre deux mots m_1 et m_2 est essentielle à la définition d'un voisinage sémantique. Celle-ci est calculée en appliquant la similarité cosinus entre eux : avec A, B vecteurs de dimension n, nous avons :

$$\boxed{\text{Similarité-Cosinus}(A,B) = \frac{A \cdot B}{\|A\|\|B\|}} \tag{2.1.1}$$

Dès lors, nous pouvons définir le voisinage d'un mot m dans un modèle comme l'environnement sémantique de ce mot m. Ce voisinage est constitué des plus proches voisins de m ; le nombre de voisin, noté $n_{voisins}$ est fixé. Pour l'obtenir, il faut calculer la distance entre m et chaque mot de vocabulaire du modèle. Un tri croissant des distances permet de déduire le voisinage.

Le pseudo-code de l'algorithme utilisé pour obtenir le voisinage d'un mot est proposé en annexe A.1.

A l'issue d'un tel algorithme, nous obtenons une liste de $n_{voisins}$ termes pour chaque mot de vocabulaire. Le tableau 2.1 affiche quelques exemples de plus proches voisins selon la valeur sélectionnée de $n_{voisins}$.

voisins	m	année	voisins
3	numérique	1930	extrême, accroître, faiblesse
5	ampoule	1960	tube, lampe, bouilloire, pylône, chandelier
10	bouteille	1980	bière, canette, tasse, cidre, farine, cuillère, flacon, riz, confiture, biscuit

Tableau 2.1. Liste des n plus proches voisins des mots "numérique", "ampoule" et "bouteille", avec n prenant respectivement les valeurs 3, 5, 10

La lecture de ces exemples apporte un complément à la notion de plus proches voisins : nous y retrouvons des termes du même lexique ("chandelier" et "ampoule"), des termes fréquemment associés ("faiblesse" et "numérique") ou encore des synonymes ("flacon" et "bouteille").

En nous intéressant dès lors au paysage constitué par les n mots les plus proches (ayant la distance la plus faible) du mot m nous définissons l'environnement sémantique d'un mot m: un environnement sémantique correspond à une partie délimitée de l'espace sémantique défini par un modèle de langue. Il regroupe l'ensemble des plus proches voisins situés dans cette partie de l'espace sémantique.

Enfin, la visualisation graphique de ces environnements nécessite une réduction de la dimension des vecteurs (de dimension 300 dans nos modèles). La méthode Principal Component Analysis (PCA) permet de réduire la dimension d'un ensemble de données multivariées de dimension N. Par analyse de ces variables, elle extrait les z (avec $z < N$) vecteurs principaux permettant de définir une nouvelle base B_z de dimension z. Les variables initiales sont alors projetées sur B_z. Très utilisée en apprentissage machine (ML), elle présente l'avantage de conserver les distances relatives entres les données.

2.2. Outils pour l'affichage d'un environnement sémantique

L'analyse graphique des environnements sémantiques du vocabulaire joue un rôle essentiel dans la compréhension des phénomènes sémantiques. Cette section présente les outils permettant l'affichage de ces environnements.

2.2.1. Projection

Nos mesures sont portées par une analyse graphique des environnements sémantiques de mots et de leur évolution, il est alors essentiel de pouvoir disposer d'une méthode de visualisation de ces environnements.

Tels que définis en section 2.1.1, les vecteurs obtenus après entrainement des modèles sont de taille 300. La visualisation de cet hyperespace étant impossible, nous nous proposons d'utiliser une méthode de PCA. En appliquant le PCA aux vecteurs d'un modèle et les réduisant à deux dimensions, nous obtenons une base correspondant à un plan graphique 2D. Dès lors une projection des mots de vocabulaire dévoile une représentation graphique de l'environnement sémantique d'un mot, représentant ses plus proches voisins et respectant les distances relatives entre les mots (voir exemple sur la figure 2.1).

Cette projection nous permet de confirmer la définition de l'environnement sémantique : ainsi, nous retrouvons parmi les plus proches voisins du mot "france", des mots issus du lexique de la géographie, avec majoritairement des noms de pays ("pologne", "grande-bretagne"...) et le mot "occident" caractérisant pour sa part le mot "france".

2.2.2. Voisinages à l'échelle du corpus

La projection des vecteurs sur un plan 2D permet l'affichage d'un état statique : elle représente le voisinage d'un mot m pendant une période P. Dans cette sous-partie nous présentons un outil proposant une visualisation synoptique et unifiée des voisinages entre les différentes tranches étudiées.

Afin de proposer une vision globale sur les T tranches du corpus, nous utilisons un outil de visualisation développé par l'OLST [**5**]. Cet outil construit une visualisation globale du mot m en deux temps :

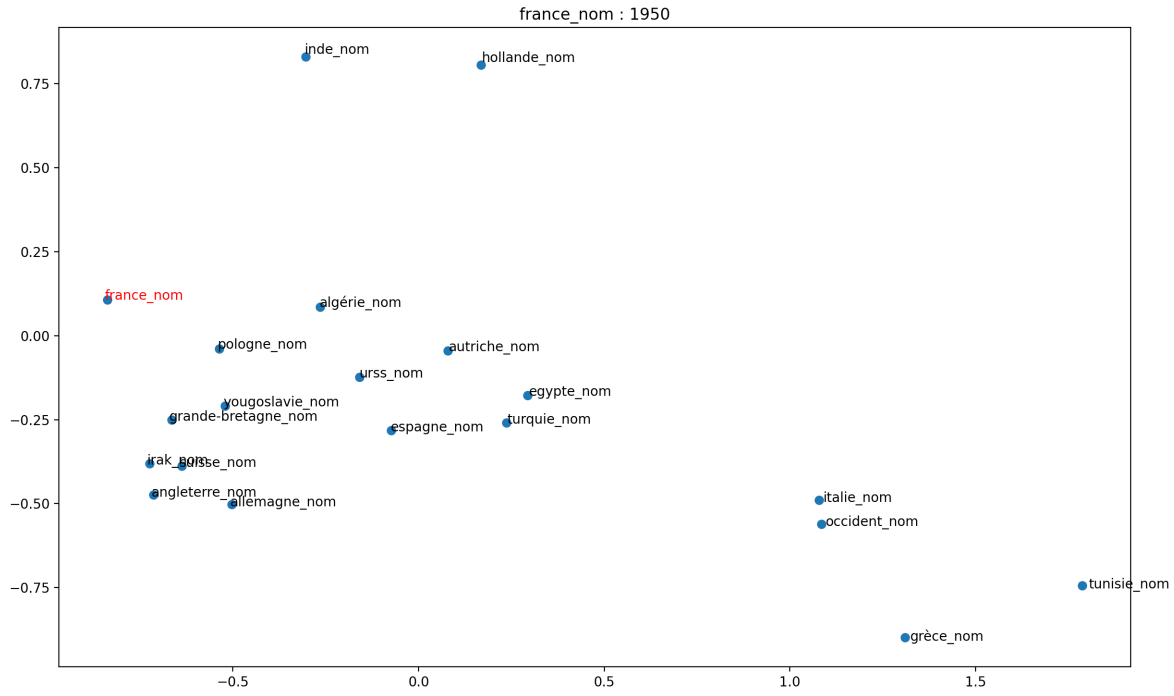


Fig. 2.1. Exemple d'une projection du voisinage du mot "france". Nous avons généré cette figure selon le protocole décrit dans la section 2.2 afin de présenter les 20 voisins les plus proches. L'espace affiché est un espace de projection 2D de l'espace sémantique du modèle obtenu. Les coordonnées des mots affichés correspondent ainsi aux nouvelles coordonnées des vecteurs représentant ces mots après projection dans le nouvel espace.

- En premier lieu il collecte les n plus proches voisins de m pour chacune des T tranches.
- Puis, les $n \times T$ voisins sont unifiés : une liste des voisins disponibles est créée, dans laquelle un voisin n'apparait qu'une seule fois (quelque soit le nombre de tranches où il apparait). A chaque voisin est alors associée une liste de tranches d'apparitions. Ainsi un même voisin présent dans le voisinage dans les tranches t_1 et t_4 n'est présent qu'une seule fois dans la liste, mais est marqué comme voisin dans plusieurs tranches.
- Enfin, une visualisation est construite dans laquelle m est décliné en T points notés $m_1, \ldots\ m_T$. représentants les T vecteurs de m respectivement issus des T modèles. Chaque m_i est relié à l'ensemble de ses plus proches voisins. Un voisin pouvant être plus proche voisin à plusieurs époques, il peut être rattaché à plusieurs m_i.

Finalement, nous obtenons une figure présentant les liens entre les voisins au cours des périodes constituant le corpus.

La figure 2.2 présente ainsi l'exemple d'un synoptique du nom "france" entre 1910 et 2000. Le mot est alors décliné en 10 nœuds du graphique, de couleurs différentes, nommés

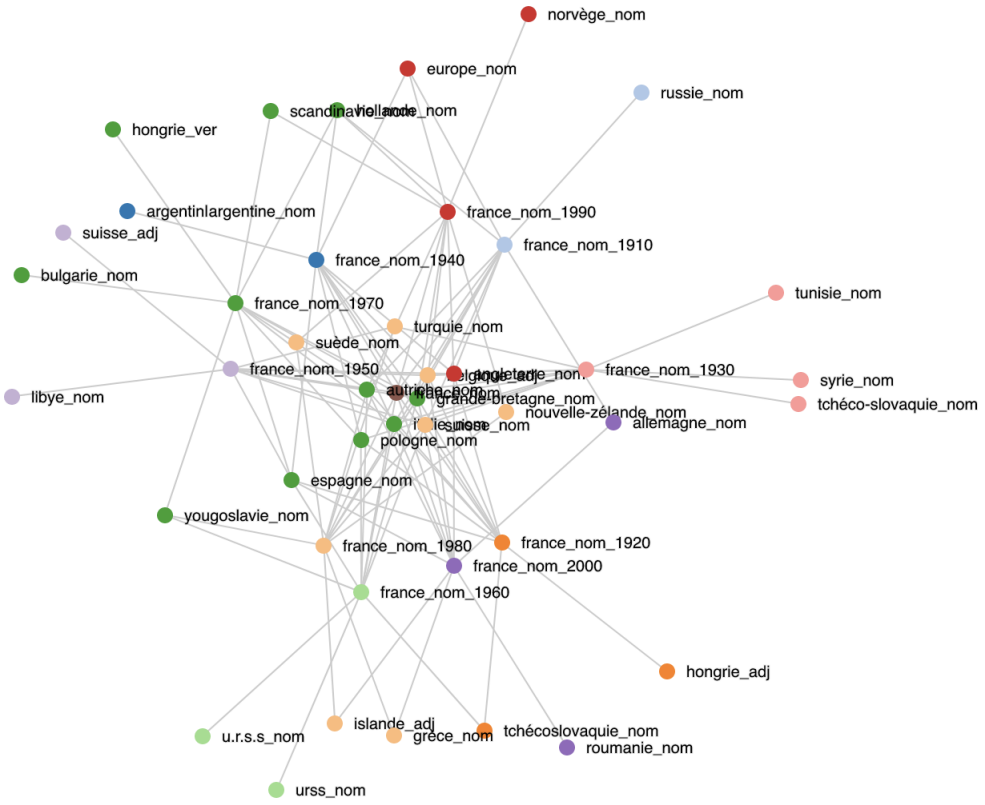


Fig. 2.2. Exemple de vision globale du voisinage du mot "france". Cette figure présente les 10 voisins les plus proches de ce mot entre 1910 et 2000. Chaque couleur correspond aux voisins d'une tranche différente. Les lignes grises représentent le lien de voisinage (plus proche voisin) entre un mot et la représentation du mot "france" dans une tranche.

france_nom_19xx, avec xx la date de début de période représentée. Ces 10 nœuds sont ensuite connectés à leurs 10 plus proches voisins. On remarque sur cette figure la constance des voisins : par exemple le mot "espagne" fait partie des plus proches voisins du mot "france" en 1920, 40, 50, 60, 70 et 2000. Tous les voisins affichés font référence à des noms de pays. Ceci s'explique en particulier par le fait que le sens du mot "france" désignant un lieu géographique est présent constamment au cours du $\mathbf{XX}^{ème}$ siècle, et donc situe sémantiquement ce mot dans un plus proche voisinage appartenant au lexique des pays.

2.3. Outils pour la désambiguisation

L'utilisation d'un mot ayant plusieurs sens est vecteur d'ambiguïté : à quel sens du mot fait référence cette utilisation? Dès lors, il existe des méthodes d'attribution de sens d'utilisation à l'occurence d'un mot dans un texte : il s'agit d'algorithmes de "désambiguïsation" ou *Word Sens Disambiguation* (WSD). Dans le contexte d'analyse des diachronies, il nous est utile d'être capables de réaliser une telle désambiguïsation. Pour ce faire, nous avons sélectionné un modèle de désambiguïsation. Nous présentons celui-ci dans cette section ainsi que les outils utilisés pour la désambiguïsation du sens d'un mot.

2.3.1. Wordnet et synset

L'algorithme de désambiguïsation que nous utilisons s'appuie sur WordNet [**6**] comme référence des sens d'usage existants d'un mot. Wordnet étant une ressource déclinée en de nombreuses langues (anglais, français, chinois...), son utilisation présente l'avantage de permettre d'envisager des applications en toutes ces langues par la suite.

Cette ressource associe à chaque mot du dictionnaire une liste de "*synsets*" (pour *synonyme set*) qui correspondent à une liste de sens admis d'utilisation du mot. Les synsets recensés sont accompagnés d'une définition en anglais précisant la nature de ce synset (voir figure 2.3).

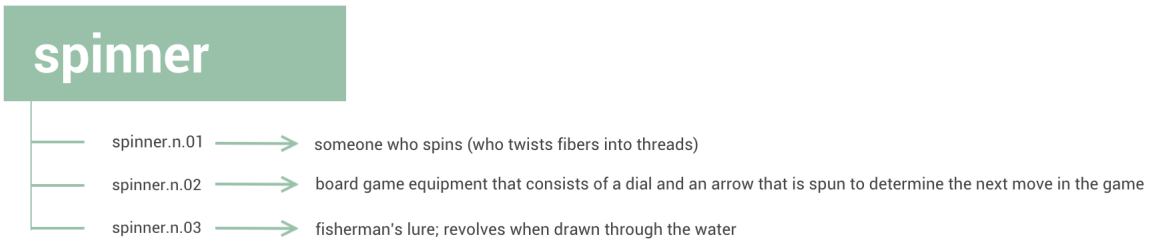


Fig. 2.3. Extrait de WordNet : synsets associés au mot "Spinner", accompagnés de leur définition.

2.3.2. Word Sens Disambiguation

Le modèle utilisé pour la désambiguïsation est issu d'un algorithme désambiguïsation de l'équipe GETALP [**17**]. Cet algorithme est proposé dans sa version française, et repose sur les éléments suivants :

- Un modèle de langue dérivé de BERT [**4**] ; travaillant sur un corpus français nous avons pu le tester avec FlauBERT [**10**] autant qu'avec CamemBERT [**12**], tous deux disponibles depuis la bibliothèque python Hugginface [**21**].
- Un modèle d'annotation unifié pour le français : UFSAC [**18**]
- Un corpus d'entrainement : FLUE [**10**]; celui-ci est également proposé par l'équipe GETALP et s'inspire de GLUE [**19**] pour proposer des *benchmarks* d'évaluation sur diverses tâches de TAL.

.

La tâche d'entrainement de l'algorithme est une tâche de classification.

La désambiguïsation procède selon le protocole en plusieurs étapes : l'algorithme reçoit d'abord en entrée une phrase et propose successivement une tokenisation de la phrase, puis il effectue une traduction token par token vers l'anglais avant d'associer une désambiguïsation à chacun des tokens. La désambiguïsation fournit en sortie des synsets issus de Wordnet. Nous présentons ci-dessous un exemple de sortie de l'algorithme utilisé. Le texte suivant est fourni en entrée :

"*Déjà aux prises avec des difficultés financières considérables les directions des universités*"

La sortie de l'algorithme reprend un par un chaque mot du texte et lui associe un synset issus de la liste disponible sur Wordnet. On obtient ainsi :

Déjà	|	not%4:02:00
aux	|	distressed%3:00:00:troubled:00
prises	|	distressed%3:00:00:troubled:00
avec	|	distressed%3:00:00:troubled:00
des	|	not%4:02:00
difficultés	|	asperity%1:07:01::
financières	|	fiscal%3:01:00::
considérables	|	considerable%3:00:00::
les	|	not%4:02:00
directions	|	management%1:14:00::
des	|	governement%1:09:00::
universités	|	university%1:14:00::

Le symbole "|" sépare ci-dessus les mots du texte original des indications correspondant à l'encodage du synset (tel que défini en section 2.3.1)

Le modèle utilisé s'appuie sur un modèle de langue afin d'obtenir des représentations vectorielles des tokens reçus en entrée. Puis, un réseau de neurones est entrainé à la sortie d'une pile de $transformers$. La couche de neurones linéaire attribue un score aux synsets associés au token traité et un $softmax$ permet de décider du résultat.

L'apprentissage de la désambiguïsation est effectuée après l'étape de traduction, aussi l'algorithme s'entraine sur cette étape en s'appuyant sur un corpus en anglais construit par la concaténation des corpus SemCor [**14**] et du Princeton WordNet Gloss Corpus (WNGC) [**6**]. Le corpus est annoté sous la forme UFSAC. L'évaluation est pour sa part effectuée sur la tâche 12 de SemEval 2013. Celle-ci propose 1445 instances labélisées, et évalue la correspondance entre le label (synset) proposé par l'algorithme et le label réel.

Chapitre 3

Corpus

Nos recherches bibliographiques en introduction ont permis d'exposer le fait qu'une démarche de recherche de diachronies s'appuie nécessairement sur un corpus historique, contenant des textes rédigés tout au long des périodes étudiées. Ce chapitre présente le corpus sur lequel nous nous appuyons, décrit les choix de découpages accomplis et en détaille les motivations.

3.1. Nature du corpus

Le corpus que nous utilisons est de nature journalistique : il est constitué de l'accumulation des parutions de journaux québecois datant de 1800 à 2000.

Le corpus a été mis à notre disposition dans le cadre du projet CO.SHS[1] (la cyberinfrastructure ouverte pour les sciences humaines et sociales), financé par la Fondation canadienne pour l'innovation (FCI) et soutenu par le consortium Érudit[2] [**1**].

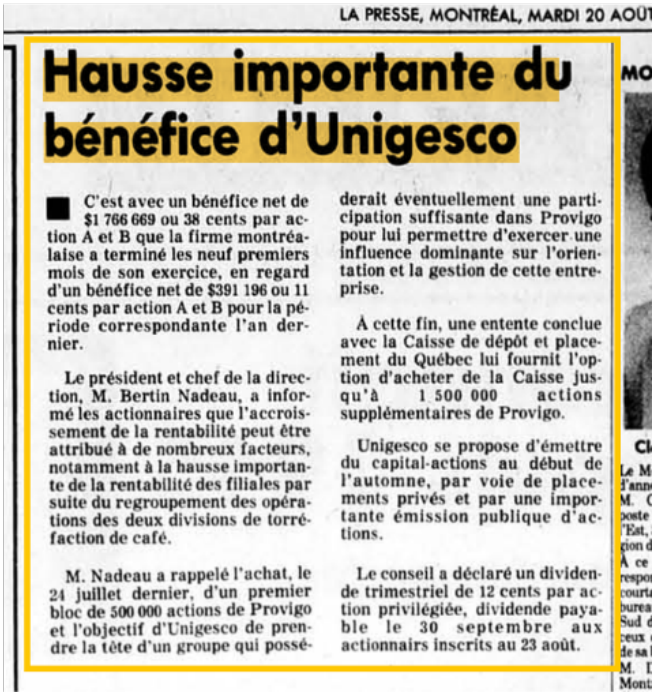
Au total, 196 titres composent ce corpus, et celui-ci réunit tous les types de presses ayant existé au Québec sur la période donnée et notamment :

- quotidiens généralistes (Le Devoir, La Presse. . .)
- presse religieuse (la vérité, Jeunesse et Héraut, Action Catholique, Prêtre aujourd'hui . . .)
- presse alimentaire
- presses locales (journal de Montréal)
- ou encore presse people : Télé-radio monde

Ces titres sont très majoritairement en français. Les outils de TAL que nous présenterons dans les chapitres 4 et 6 proposent des analyses s'appuyant sur les usages des termes tels qu'observés dans le corpus, aussi, afin d'éviter un mélange qui fausserait les résultats, nous avons d'office décidé de ne pas considérer les textes bilingues ou rédigés en anglais.

[1] https://co-shs.ca
[2] https://www.erudit.org

Les documents sont numérisés et fournis en format pdf. En outre, nous disposons des versions textes des titres, obtenues grâce à l'utilisation d'algorithmes d'OCR[3] (ceux-ci proposant une conversion d'un fichier pdf en fichier texte par reconnaissance caractère par caractère des textes imprimés, jusqu'à former un texte complet.[4])



LA PRESSE, MONTRÉAL, MARDI 20 AOÛT

Hausse importante du bénéfice d'Unigesco

C'est avec un bénéfice net de $1 766 669 ou 38 cents par action A et B que la firme montréalaise a terminé les neuf premiers mois de son exercice, en regard d'un bénéfice net de $391 196 ou 11 cents par action A et B pour la période correspondante l'an dernier.

Le président et chef de la direction, M. Bertin Nadeau, a informé les actionnaires que l'accroissement de la rentabilité peut être attribué à de nombreux facteurs, notamment à la hausse importante de la rentabilité des filiales par suite du regroupement des opérations des deux divisions de torréfaction de café.

M. Nadeau a rappelé l'achat, le 24 juillet dernier, d'un premier bloc de 500 000 actions de Provigo et l'objectif d'Unigesco de prendre la tête d'un groupe qui posséderait éventuellement une participation suffisante dans Provigo pour lui permettre d'exercer une influence dominante sur l'orientation et la gestion de cette entreprise.

À cette fin, une entente conclue avec la Caisse de dépôt et placement du Québec lui fournit l'option d'acheter de la Caisse jusqu'à 1 500 000 actions supplémentaires de Provigo.

Unigesco se propose d'émettre du capital-actions au début de l'automne, par voie de placements privés et par une importante émission publique d'actions.

Le conseil a déclaré un dividende trimestriel de 12 cents par action privilégiée, dividende payable le 30 septembre aux actionnairs inscrits au 23 août.

Fig. 3.1. Version numérisée d'un article de La Presse du *20/08/1985*.

```
Hausse importante du bénéfice d'Unigesco
i I C'est avec un bénéfice net de $1 766 669 ou 38 cents par action
A ci B que la firme montréalaise a terminé les neuf premiers mois de
son exercice, en regard d'un bénéfice net de $391 196 ou 11 cents
par action A et B pour la période correspondante l'an dernier.
Le président et chef de la direction, M. Bertin Nadeau, a informé
les actionnaires que l'accroissement de la rentabilité peut être
attribué à de nombreux facteurs, notamment à la hausse importante de
la rentabilité des filiales par suite du regroupement des opérations
des deux divisions de torréfaction de café.
M. Nadeau a rappelé l'achat, le 21 juillet dernier, d'un premier
bloc de 500 000 actions de Provigo et l'objectif d'Unigesco de
prendre la tête d'un groupe qui possé-
derait éventuellement une participation suffisante dans Provigo pour
lui permettre d'exercer une influence dominante sur l'orientation et
la gestion de cette entreprise.
À cette fin, une entente conclue avec la Caisse de dépôt et
placement du Québec lui fournit l'option d'acheter de la Caisse
jusqu'à 1 500 000 actions supplémentaires de Provigo.
Unigesco se propose d'émettre du capital-actions au début de
l'automne, par voie de place-monts privés et par une importante
émission publique d'actions.
Le conseil a déclaré un dividende trimestriel de 12 cents par action
privilégiée, dividende payable le 30 septembre aux actionnairs
inscrits au 23 août.
```

Fig. 3.2. Texte de l'article obtenu par algorithme d'OCR.

3.2. Description du vocabulaire, erreurs de transcriptions et conséquences

Les algorithmes d'OCR, bien que très utiles peuvent présenter des défauts d'interprétation. En effet, l'identification des caractères est sujette à divers aléas de l'environnement d'impression, par exemple le cas de textes non formellement isolés ou l'utilisation de paramètres inhabituels dans le choix de la police de caractère, l'encre utilisée ou encore le papier d'impression. Aussi, et afin de nous assurer de la qualité des textes utilisés, nous avons inspecté des extraits du corpus et avons comparé l'entrée (version numérique d'une page) et la sortie (version texte obtenue) produite par l'algorithme (voir par exemple les figures 3.1 et 3.2 montrant respectivement la version numérique et la version texte d'un même article).

[3]Ce processus a été réalisé en amont de notre travail. Nous n'avons pas accès à ces algorithmes mais uniquement aux sorties textuelles produites. Notre seule marge de manœuvre consiste donc en un travail sur les fichiers textuels

[4]Les mots sont donc conservés avec leurs particularités : majuscules, accentuations, pluriels... Les questions liées à la nature des mots utilisés seront davantage discutées au cours du chapitre 6

Ces analyses nous ont permis de mettre en évidence plusieurs défauts récurrents d'OCR. Ceux-ci correspondent à des usages fréquents dans le cadre d'écrits journalistiques, et impliquant des problèmes de reconnaissance des caractères. Nous en proposant une rapide anthologie compilant une description du défaut, et un exemple de textes imprimés de l'interprétation textuelles associée :

- Colonnes : les journaux sont en majorités rédigés en colonnes. Bien que les colonnes soient normalement reconnues par les algorithmes d'OCR, la présence d'autres défauts entraine une illisibilité de l'écriture en colonne, et les textes sont interprétés sans prendre en compte les passages à la ligne. Ainsi, les figures 3.3 et 3.4 mettent en évidence le fait que les problèmes de papier de journal entrainent une non reconnaissance des colonnes.

Fig. 3.3. Version numérisée d'un article de La Presse du *22/06/1928*.

RBcapitallsatlon probable de la Lake Superior Corp.
avait ivnîreprinee aux usines de bt, Ca- j faveur d#5 pori-llr.-» d'actioru? ordirlftlr#» therina»; il rn» fait plaisir de air# que 1 actueBé», soit au taux d'un* pour une. Ofis ligne», censista.nl surtout d# la- ; ^ tmU,^ s<*t 86,996 aettan» "A" «t biew'.x d# command# et inMaflatlon gé« j io,606 «otioni» "B", demeurera dans le nfirai# électrique, dont le marolifi s é-l trÀw .pour le» besoin» fulUîi» de la oom-ter:d oontlnuellement au Panada, ont été tri>« favoraidement accueülies pafi; si le plan est adopté, les dfirecte<9c>-» le» ollent». î ont Vin trillion de reoommandvr le pale-
<93>Vert directeur» fittidH-nt depuis quel- | nw-r.t d'«n dividende de *3 ipar action . qua iemt» un plan pour modifier !»i -A" et espèrent pouvoir «n dont muer raht-gfinfiraî, en sera ie pnésrident eijpn»! d< la oompagnl#: ce plan est, paiemaxit.
gérant-général, iinair, tenant terrotné et Son-, soumis aux! .....- ¦¦ .... actionnaire» S une aasemblfi» prochaine. I ~r"
ÎTn avl» de convocation ainsi qu'un# cd- ; pie du plan de rficapitAlieatk>n sont en- ; vpyéé' Immédiatement aux actionnaire» !
¦<95>Vos directeurs sont porteur» d'un! nombre considérable (l'action* de» deux ' catégories et approuvent unanimement l le plan comme étant dan* le meilleur [intfirft' rie tous le» actionnaire» et de la D'après le Wall Street Journal, le cobipsurnle. Si vtws ne pouvez P»" groupe canadien qui a acquis la contrôle lasolaier ft rassemblée, fafte» une propu-de la Lake Superior CoritoraÉlon est en [ration A l'ordre de vos directeurs, b, le train d'établir un plan pour modifier leiplan est adopté .le» action» «fWit plft-i-apltai iilnst que celui de quelque» ftilft-|eéev* sur une btu» de dividende Immé-le«. U «si iKisalbl» que la chart# actuelle I dlatement."
du New-Jersey soir remis# et que l

Fig. 3.4. Texte de l'article obtenu par algorithme d'OCR. Les zones surlignées du texte ont été encadrées sur l'image numérique de l'article, laissant apparaitre les problèmes de reconnaissance de colonnes : les deux colonnes de l'article sont transcrites à la suite comme ne constituant qu'une seule ligne.

- Papier journal ancien, abimé ou jauni : les problèmes liés à la qualité du papier journal abaisse les contrastes avec la police et brouille la lecture du texte.

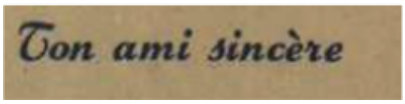


Fig. 3.5. Version numérisée d'un article de Jeunesse et Hérauts du *15/09/1947*.

V,
on ami âmcete

Fig. 3.6. Texte de l'article obtenu par algorithme d'OCR.

Ainsi, les figures 3.5 et 3.6 montrent un papier jauni, et dont la cumulation avec une police inhabituelle conduit à l'interprétation erronée du texte.

- Gros titres, changements de taille de police et polices inhabituelles : les variations de types de police au sein d'une même page s'opposent à une interprétation correcte du texte. Seule une police peut être correctement transcrite. Le reste du texte en police différente n'est soit pas identifié soit mal transcrit.

Fig. 3.7. Version numérisée d'un article de Télé RADIO MONDE de la semaine du *12* au *19/01/1980*.

```
###PAGE###1###
Mention incorrecte : Date

2 ao

.

A
VOLUME 42 NUM…RO 18 DU 12 AU 19 JANVIER 80 PRIX- 50e ó ….U. & N.B.: 5

¶-.Ô*1 i
I

N'AVAIT PAS
```

Fig. 3.8. Texte de l'article obtenu par algorithme d'OCR. Nous remarquons que seule la zone encadrée en rouge a été correctement interprétée.

- Les présentations hybrides : mêler texte et image aboutit de nouveaux à des défauts d'interprétation : les séparations de textes (par exemple entre les bulles de personnage de bandes dessinées) ne sont plus identifiées.

Fig. 3.9. Extraits numérique et textuel d'une même page de bande dessinée publiée de la parution du *15/09/1947* de Jeunesse et Hérauts. Les éléments du texte présentant des problèmes sont indiqués par des liens colorés. Nous remarquons ainsi qu'en orange est désignée une date qui a été interprétée comme un mot, et en vert un changement de colonnes qui a été omis.

- Les textes non alignés : quelques soient les polices utilisées, les algorithmes utilisés ne traitent les textes que ligne par ligne, parallèlement à l'orientation du papier. Le non-alignement du texte aboutit à une interprétation erronée.

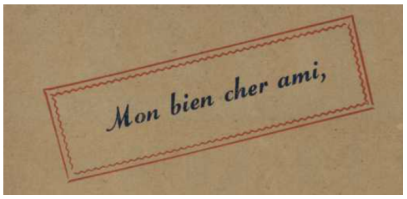


Fig. 3.10. Version numérisée d'un article de Jeunesse et Hérauts du *15/09/1947*.

N-
NWWxWiWUV
MM
mrük

Fig. 3.11. Texte de l'article obtenu par algorithme d'OCR.

Finalement, l'accumulation de ces erreurs aboutit à l'apparition de bruit, c'est-à-dire de termes dus à des erreurs, soit que des signes aient été mal interprétés et provoqué l'apparition d'un mot supplémentaire dans le texte, soit qu'un mot ait été mal transcrit.

Le parcours des textes composant le corpus nous permet empiriquement de proposer les deux tris suivants :

- Seuls les journaux généralistes quotidiens et hebdomadaires proposent majoritairement des pages avec une police identique, peu d'illustrations, et des colonnes clairement identifiables, aussi, par la suite nous ne considérerons plus les autres titres.
- Les journaux d'avant 1910 présentent un papier régulièrement abimé, ce qui nous pousse à ne considérer que les parutions ultérieures.

Ces considérations nous ont permis d'identifier 5 titres dont les parutions sont régulières, et couvrent des périodes au moins ultérieures à 1910 et longues d'au moins 10 ans. La figure 3.12 affiche les périodes de parution de chacune d'entre elles.

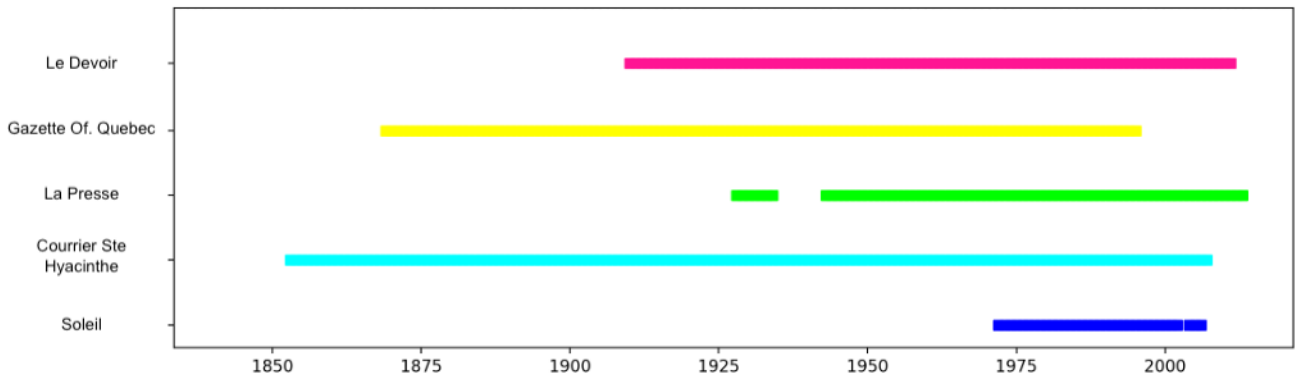


Fig. 3.12. Répartition chronologique des parutions disponibles de chaque titre disponible.

Dans le but d'estimer la qualité des textes, nous calculons le nombre d'occurrences de chaque mot utilisé, c'est-à-dire que nous identifions tous les mots différents utilisés dans le corpus, puis calculons le nombre total de leurs apparitions respectives dans chaque parution disponible. A l'issue de cette étape, nous repérons que seuls 20% de mots apparaissent plus de deux fois par texte. Ce faible taux est révélateur des problèmes remarqués ci-dessus.

Ainsi, les mots ayant peu d'occurrences sont en partie des mots mal identifiés. Le tableau 3.1 propose des exemples de mots apparus entre 5 et 10 fois dans une parution.

On remarque en conséquence que si certains de ces mots appartiennent au vocabulaire français-québécois, une partie non négligeable est encore le fait d'erreurs de transcription.

Une inspection supplémentaire des mots nous permet de décider de ne considérer que les mots apparaissant plus de 200 fois dans le corpus complet.

Mots	exemples
dûs à des erreurs	ut, Uu, V*,Vv, Vznd
corrects	Vanité, VENDU, Vend, vent, Véritable

Tableau 3.1. Exemple de mots apparus entre 5 et 10 fois dans la version texte de la parution du *22/06/1928* de La Presse. Les mots sont triés entre "corrects", c'est à dire mots disponibles dans les dictionnaires, leur conjugaison... et ceux ne correpondant à aucune forme, classés comme "dûs à des erreurs".

3.3. Configuration finale du corpus

Enfin, nous concluons cette partie en indiquant que comme recommandé par la littérature, notre corpus est divisé en tranches. La présence de textes rédigés entre 1910 et 2000 nous pousse à envisager un découpage en tranches de 10 ans.

Dans ce cadre, nous nous intéressons aux tailles du vocabulaire respectif de ces tranches envisagées ; en effet, malgré la qualité des mots recensés à l'issu du tri sur le nombre d'occurrences, nous désirons nous assurer d'une taille similaire de vocabulaire afin d'éviter qu'un déséquilibre influence les expériences menées dans la suite du mémoire.

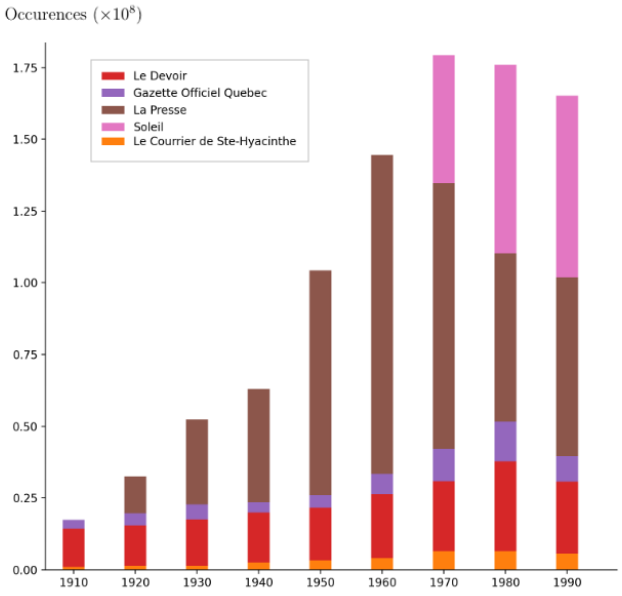


Fig. 3.13. Nombre total de mots dans chaque période de 10 ans couverte et leur répartition parmi les titres de journaux.

La figure 3.13 met en évidence l'existence d'un déséquilibre patent variant de 250 millions de mots en 1910-20 à plus de 1750 millions en 1970-80.

Néanmoins, cette figure révèle également que le journal Le Devoir affiche un nombre de mots tournant au cours de toutes les périodes envisagées autour de 200 millions de mots. Cette constance fait de ce titre un candidat naturel à devenir notre corpus d'étude.

Étant données toutes les contraintes décrites dans cette partie sur le vocabulaire, il nous a semblé dès lors judicieux d'envisager un travail ne s'appuyant que sur les textes issus de la numérisation des textes du Devoir. Celui-ci devient par la suite le seul titre constituant notre corpus.

Ainsi, les textes du Devoir sont répartis au sein de 9 sous-corpus non chevauchants recouvrant respectivement toutes les années des périodes entre 1910 et 1990.

Chapitre 4

Définition des mesures

L'objectif du projet dans lequel nous nous inscrivons est de proposer une analyse au terme de laquelle la diachronie d'un mot est quantifiable. Comme relevé en introduction, plusieurs méthodes ont été développées dans ce but, mais les étapes de validation associées laissent paraître certaines ambiguïtés. Nous proposons ainsi dans ce chapitre de détailler les interrogations que posent les méthodes issues de la littérature, puis proposer le développement de méthodes d'analyse du voisinage d'un mot évitant les ambiguïtés relevées et permettant de statuer quant aux résultats affichés par les auteurs. Pour ce faire, nous introduisons deux mesures de quantification : *Score* et *Link*, nous permettant de mettre en exergue les problèmes posés par les méthodes de validation des articles. Ces méthodes sont ensuite agrégées afin de nous permettre de déterminer une frontière de décision, c'est à dire à établir une séparation entre les vecteurs des mots ayant changé de sens et ceux de mots stables.

4.1. Le score d'Hamilton et ses limites

Telles que nous les avons présentées dans le chapitre 1, la lecture des méthodes proposées par Hamilton et al. nous a poussés à nous interroger.

Leur article propose une équivalence entre le déplacement spatial du plongement d'un mot et son changement sémantique. En outre, il affirme également que ce déplacement s'accompagne d'un changement d'environnement sémantique. Or ces affirmations nous semblent en contradiction avec certains résultats de leurs expériences:

- En premier lieu notons l'apparition de mots qui ne sont pas considérés comme "vrais déplacements sémantiques" parmi les 10 mots évalués par Hamilton et al. Cette apparition indique qu'un grand déplacement dans l'espace entre les différentes époques étudiées n'est pas seulement de l'apanage des mots ayant subi une diachronie. Le nombre réduit d'exemples permet mal d'appréhender précisément la qualité de la méthodologie proposée mais le taux de 20% d'erreurs parmi les mots ayant subi le plus grand déplacement dans le modèle Word2Vec, et 90% dans PPMI, interdit tout au

moins d'affirmer que le déplacement du plongement d'un mot implique absolument l'existence d'un changement de sens subi.

- Dès lors, la première évaluation, c'est-à-dire celle s'appuyant sur le jeu de données de 28 mots est remise en question : en effet, si certains mots censés être stables (par exemple "romance") subissent un déplacement, comment vérifier que le déplacement des vecteurs des 28 mots soit effectivement une preuve de la diachronie subie et non la conséquence d'autres processus (tels que l'imprécision des modèles ou des variation du taux d'utilisation des mots entre les tranches) entrainant certains déplacements ? L'absence d'implication entre déplacement et diachronie remet en cause leur interprétation.

- Si le déplacement des 28 mots du jeu de données vers leur lexique prévu respectif semble être un indicateur positif quant à la qualité de ces résultats, revenons sur le schéma 4.1 : en reprenant le mot "awful", nous lisons que les années 1850 l'approchent du mot "solemn", tandis que les années 90 le placent dans une zone sémantique proche de "weird". La réalisation de ce schéma utilise un procédé permettant d'obtenir un contexte (les voisins en gris) stable. Or le voisinage subit en réalité tout autant que "awful" un déplacement au cours des périodes. Comment vérifier que les anciens voisins ne sont pas également déplacés dans la même direction, au quel cas le déplacement dessiné marquerait davantage une modification globale de l'environnement, et non un changement de sens du mot étudié. Il serait ainsi intéressant de vérifier au-delà de la direction du déplacement comment évolue la relation entre un mot et ses anciens voisins et confirmer ainsi la qualité de la première évaluation.

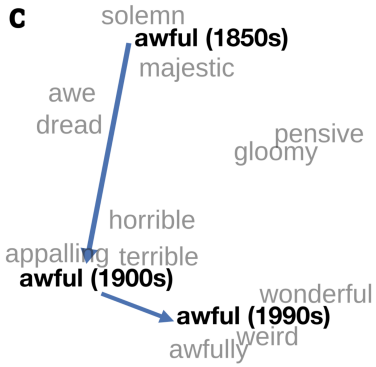


Fig. 4.1. Reprise de l'illustration tirée de l'article d'Hamilton et al. citée en figure 1.3. Nous remarquons dans cette représentation la stabilité du voisinage ("solemn", "wonderful"...) entre les différentes périodes.

Afin d'éviter les écueils que peuvent produire cette équivalence, nous proposons de poser comme fondements de nos méthodes les observations suivantes :

- Le changement de sens d'un mot s'accompagne d'une rupture temporelle de la composition de son environnement proche : ce changement de sens indique un changement de la composition de son environnement. Ainsi, repérer un changement passe par le repérage d'une rupture dans la composition des voisinages de la série temporelle.
- La stabilité du sens d'un mot s'accompagne d'une stabilité de l'emplacement du mot dans son environnement : ainsi, sans s'intéresser à la représentation du mot, il est possible de vérifier la constance de la proximité entre ce mot et ses voisins. Par exemple le mot "cow" restant utilisé dans un contexte bovin, il reste proche voisin du mot "milk" (lait) au cours des différentes périodes (figures 1.2 et 4.2).

Fig. 4.2. Représentation en 2D du contexte du mot "cow" d'après les vecteurs de plongement de mot issus du modèle Word2Vec entrainé par Hamilton et al. dans la tranche 1870-80. Nous remarquons en bas à gauche le mot milk, resté à proximité.

Nous avons essayé de nous inspirer des méthodes classiques d'étude diachronique pour quantifier le changement de sens d'un mot. Mais désormais, et afin de ne pas retomber dans les écueils notifiés, nous ne nous intéressons qu'aux changements survenus dans l'environnement proche d'un mot, en même temps que nous essayons de concilier les deux propositions précédentes. A l'aide de ces méthodes, nous essayons donc de qualifier la diachronie en n'observant que les variations de l'environnement d'un mot.

4.2. Mesures

Cette section présente la définition des mesures de la diachronie ne s'interessant qu'aux variations de l'environnement d'un mot. Nous les avons élaborées en complément de celles d'Hamilton et al. afin d'en confronter les résultats. Nous présentons ainsi deux mesures nommées respectivement Score et Link (L). Le Score est proposé sous deux formes : une forme classique (Score-base ou S_b), et une variante gaussienne (S_g)

4.2.1. Définition des Scores (S_b et S_g)

La première mesure développée est nommée *Score-base* (notée S_b).

Ce score (S_b) s'attache à mesurer l'évolution de l'environnement d'un mot m entre deux périodes. On calcule le score entre deux périodes P_1 et P_2[1] selon:

$$\boxed{S_b(m) = \sum_{v \in V} abs(d(m_1, v_1) - d(m_2, v_2))} \tag{4.2.1}$$

Où:

(1) V est l'ensemble des voisins sémantiques de m sur les deux périodes P_1 et P_2

(2) v_i est vecteur du voisin v dans le modèle entrainé sur la période P_i

(3) m_i est le vecteur du mot dont nous calculons le score dans le modèle entrainé sur la période P_i

(4) d(a, b) est la similarité cosinus entre a et b

Dans un second temps, intéressons nous au fait que la définition du contexte d'un mot m est davantage portée par les plus proches voisins que par ceux ayant une plus grande distance à m. Néanmoins, la mesure Score-base ne tient pas compte de l'importance accordée aux voisins les moins distants. Nous proposons ainsi une variante de la formule 4.2.1 offrant une mesure garante de ces effets.

Pour ce faire, nous nous inspirons de l'algorithme des "k plus proches voisins" appliqué avec une fenêtre souple (ou Parzen), et définissons une fenêtre souple constituée par tous les voisins situés à une distance inférieure ou égale à d (hyperparamètre). La contribution de chaque voisin à la distance totale est désormais modulée par leur distance à m. Les poids des voisins les plus proches sont proches de 1, ceux des voisins les plus éloignés sont proches de 0. La différence avec Score-base est due à l'introduction d'une loi normale dans le calcul du score. Une telle loi repose sur une fonction de type gaussienne (voir formule 4.2.2).

[1]Le détail de l'algorithme de calcul du Score-base est disponible en annexe A.2.1

$$\boxed{G_c(x) = \frac{1}{c\sqrt{2\pi}} e^{-(x-\mu)^2/2c^2}} \tag{4.2.2}$$

Où :

(1) x est la distance à moduler

(2) μ est le centre de la loi. Dans le cadre des mesures, nous utilisons une loi centrée, soit $\mu = 0$

(3) c est la covariance de la loi normale (un métaparamètre supplémentaire)

Afin d'appliquer une telle modulation des distances, nous proposons une mesure nommée Score-gauss (notée S_g) que nous définissons ainsi:

$$\boxed{S_g(m) = \sum_{v \in V} abs(G_c(d(m_1, v_1)) - G_c(d(m_2, v_2)))} \tag{4.2.3}$$

4.2.2. Definition du Link (L)

Cette section introduit une mesure supplémentaire de diachronie : le Link. Celle-ci s'attache en particulier à mesurer la constance des plus proches voisins.

Après nous être intéressés à la modification des distances entre m et les plus proches voisins, nous proposons d'analyser la constance de l'environnement.

Nos analyses qualitatives nous ont permis de proposer le postulat suivant : un mot v constamment présent dans la liste des plus proches voisins de m indique une constance sémantique de m [2]. Ceci présupposant à la fois que v soit resté assez proche pour définir l'environnement sémantique et qu'aucun nouveau sens ne soit entré assez fortement dans l'environnement pour en éliminer v. Dès lors, nous proposons la mesure suivante comme définition du Link [3]:

$$\boxed{\text{Link}(m) = \sum_{i}^{10} \sum_{\substack{j \\ j>i}}^{10} 10(j-i)|ppv_i(m) - ppv_j(m)|} \tag{4.2.4}$$

Où:

(1) ppv_i est l'ensemble des plus proches voisins de m à la période P_i

(2) $|\ .\ |$ la cardinalité d'un ensemble. (donc $|ppv_i(m) - ppv_j(m)|$ est le nombre de plus proches voisins de m communs à P_i et P_j)

La mesure calcule la constance des voisins les plus proches. Ainsi, elle considère les périodes les unes après les autres, et pour chacune d'entre elles compare les voisins les plus proches avec ceux des périodes ultérieures. Chaque voisin en commun prend pour valeur un

[2] Ces analyses qualitatives sont proposées en annexe B.2

[3] Le pseudo-code de calcul du link est proposé en annexe A.3.1

coefficient défini multiplié par l'écart entre les époques communes. Enfin, Link additionne les valeurs de tous les voisins.

4.3. Analyses des mesures

Cette section décrit les motivations et les intentions auxquelles répondent les mesures proposées dans la section 4.2.1. Les mesures sont ensuite appliquées à des exemples : cette application permet d'illuster la scission produite par ces mesures entre les mots ayant subi une diachronie et les mots stables.

Les mots qui sont utilisés comme exemples ont été sélectionnés selon une transposition en français de la méthode de Hamilton et al. [**8**], c'est à dire qu'ils sont issus d'articles ou de dictionnaires avérant leur stabilité ou changement de sens au cours du $\mathbf{XX}^{ème}$ siècle. Ils sont séparés entre mots supposés stables (représentés en orange) et supposés avoir subi une diachronie (représentés en bleu) selon ces sources. Par exemple, si le mot "vache" (en orange) conserve les mêmes définitions, le mot "disque" (en bleu) se voit ajouter une définition relevant du support de stockage informatique.

Les mesures affichées ont toutes été calculées entre les périodes 1910-20 et 1990-00.

4.3.1. Scores

Les analyses qualitatives réalisées nous ont poussés à nous intéresser à la distance parcourue par un proche voisin de m [4], et nous ont permis de déduire que la stabilité des sens d'usage d'un mot implique une stabilité du voisinage : le déplacement d'un plus proche voisin est un marqueur de changement de sens. Ce déplacement peut être un rapprochement de m ou un éloignement.

La définition de Score-base s'attache ainsi à additionner la valeur absolue du déplacement de tous les plus proches voisins des périodes étudiées, c'est-à-dire la différence entre la distance à m dans le modèle entrainé sur P_1 et celui entrainé sur P_2. Plus le score est faible plus un mot est stable entre les deux périodes évaluées et vice-versa.

Finalement, la formule du score proposée (4.2.1) nous offre des variations possibles sur plusieurs hyperparamètres :

- Le seuil : nombre d'apparitions d'un mot dans le corpus pour pouvoir être considéré comme un plus proche voisin
- Le nombre de voisins : combien de plus proches voisins sont sélectionnés dans chacune des périodes étudiées

Cette formule est ensuite testée sur les exemples décris ci-dessus, et les résultats sont présentés sur la figure 4.3.

[4] Ces analyses qualitatives sont proposées en annexe B.1

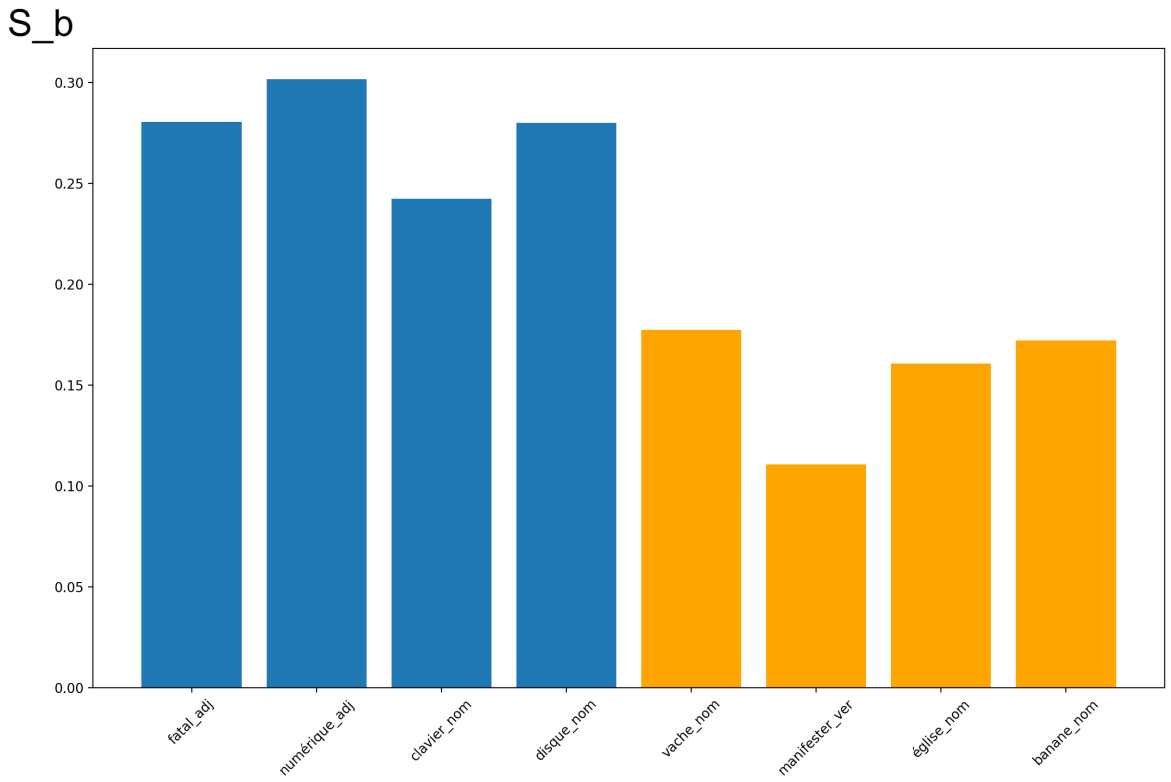


Fig. 4.3. Exemple des S_b entre 1910-20 et 1990-00 de quelques mots sélectionnés. Les mots notifiés comme ayant changé de sens entre les deux périodes sont en bleu, les mots notifiés stables sont en orange. La figure laisse apparaître un seuil aux alentours de 0.2.

Cette application du score montre une différence entre les scores de mots du groupe des mots supposés avoir subi une diachronie et ceux du groupe supposé stable ; le score de 0.2 pouvant être établi sur la base de ce graphique (à conditions identiques) comme frontière de décision.

Pour sa part Score-gauss est contrôlé par l'hyperparamètre c : une faible valeur de celui-ci insiste sur le poids des plus proches voisins, une valeur plus forte équilibre davantage l'influence entre tous les voisins.

La figure 4.4 montre un exemple de coefficients multiplicateurs appliqués au voisinage du mot "cow" (vu au chapitre 1). Le paramètre c de la fonction gaussienne contrôle l'applanissement de la fonction, et donc le diamètre des cercles représentés sur le voisinage.

En sus l'hyperparamètre controllant la valeur de la covariance c de la loi normale utilisé, les hyperparamètres de la formule de S_b sont applicables à S_g.

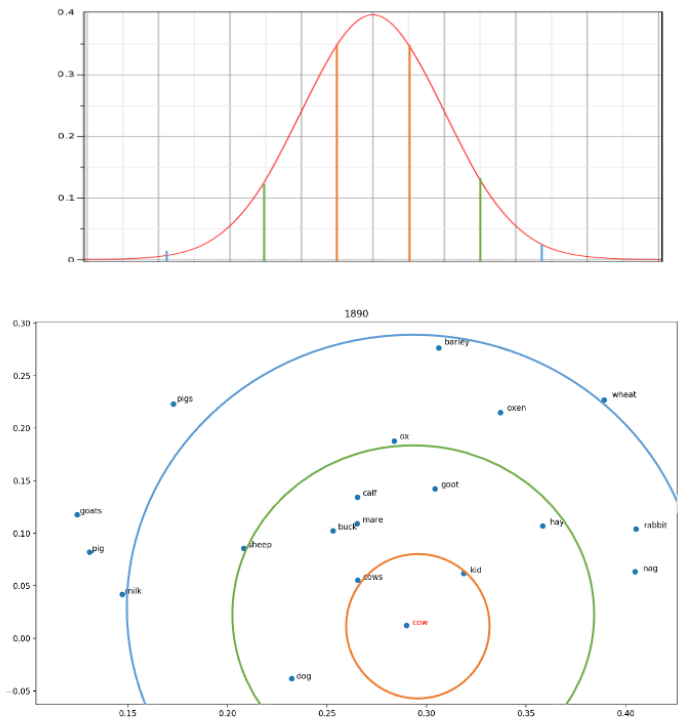


Fig. 4.4. Illustration de l'application d'une fonction gaussiennee au score. Les mots à l'intérieur du cercle orange voient leur score multiplié par 1, tandis que ceux à l'exterieur du cercle bleu sont multipliés par 0.5.

4.3.2. Link

La mesure Link permet à la fois d'attribuer un score croissant avec les liens établis entre les voisins les plus proches de différentes périodes et de favoriser les liens entre les périodes les plus éloignées :

- Un mot stable entre deux périodes établira des liens entre les périodes successives, ainsi qu'entre les périodes extrêmes, maximisant son Link.
- A l'inverse, un mot instable verra ses liens interrompus entre les périodes de diachronie, limitant la valeur finale du Link.

Cette définition offre la possibilité de choisir deux hyperparamètres:

- Le nombre de voisins considérés pour chaque période prise en compte.
- Le nombre de périodes intermédiaires : pour calculer Link entre deux périodes P_1 et P_2, quelles seront les périodes pour lesquelles les plus proches voisins seront analysés?

Nous appliquons Link à nos exemples sur la figure B.2 afin de l'illustrer. Nous avons choisi empiriquement avec les valeurs d'hyperparamètres suivantes :

- Nombre de voisins considérés : 10
- Nombre de périodes intermédiaires : les 9 époques entre 1910 et 1990

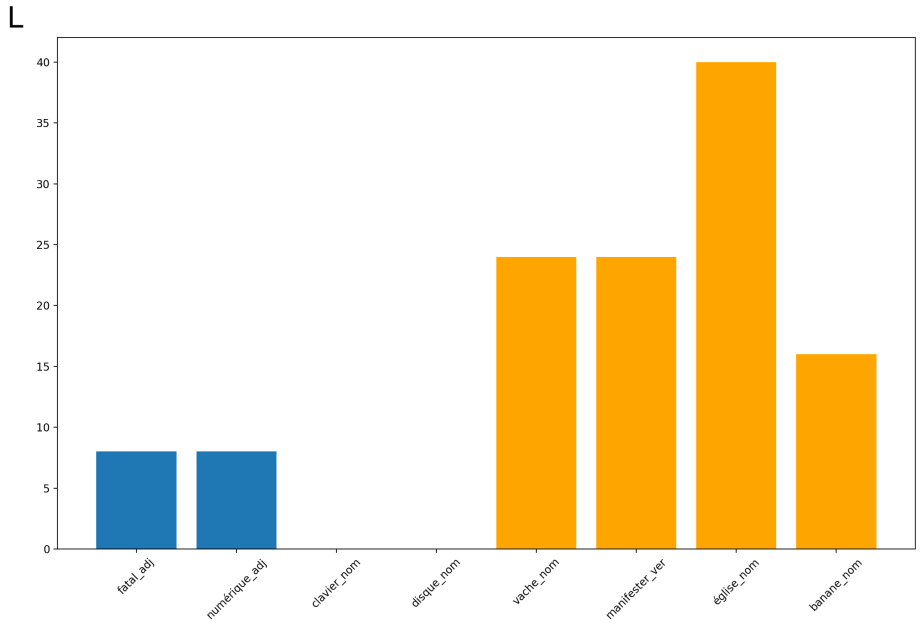


Fig. 4.5. Illustration du calcul de Link entre 1910 et 1990. Les mots en orange étaient notifiés stables, ceux en bleu avoir subi une diachronie. Les deux mots avec Link à 0 sont du groupe notifié avoir subi une diachronie.

Cette application du Link nous permet de mettre en évidence la gradation existant entre les mots supposés avoir changé de sens, et ceux supposés être restés stables : nous remarquons qu'une frontière de décision sur Link peut être tracée entre 10 et 12. En outre, les scores de 0 de deux mots supposés avoir subi une diachronie montrent une inconstance de leur proche voisinage, conséquence des changements de sens subis.

Chapitre 5

Application des mesures

Après avoir défini nos mesures, nous les utilisons conjointement pour la définition d'une nouvelle représentation spatiale des plongements de mots. Celle-ci nous permet notamment de soumettre un regard critique sur l'équivalence proposée entre déplacement du vecteur et déplacement sémantique ; c'est-à-dire vérifier si seuls les mots ayant subi une diachronie subissent un déplacement dans l'espace sémantique. Cette étape est l'occasion de questioner la pertinence du choix des exemples, et suggérer la nécessité du développement d'un jeu de données de validation complet, tout en en définissant les contraintes de sa construction.

5.1. Agrégation des scores

Revenons sur l'objectif des scores développés : avec leur aide, nous essayons de développer une méthode permettant de repérer les mots ayant subi une diachronie entre deux périodes.

Dans ce cadre, nos méthodes doivent permettre de déterminer une frontière de décision : les représentations vectorielles des mots étudiés évoluent dans un espace de 300 dimensions, et nous cherchons à calculer un hyper-plan établissant une séparation entre les vecteurs des mots ayant changé de sens et ceux de mots stables.

Les méthodes que nous avons proposées ramènent les points à un plan en 2 dimensions. La frontière peut donc être une droite s'il s'agit d'une limite linéaire ou de toute autre courbe (un cercle, une courbe logarithmique...) dans le cas contraire.

Les exemples que nous avons utilisés nous permettent de vérifier la qualité des méthodes. Intéressons-nous donc aux représentations de ces mots après application des mesures, en les affichant sous formes de vecteurs de coordonnées $(S_b(m), L(m))$.

Nous proposons ainsi de tracer ce plan et d'y placer les exemples du chapitre 4 dans la figure 5.1. Cette figure met en évidence l'existence d'une polarité entre les mots ayant changé de sens caractérisé par $S_b > 0.23$ et $L < 12$, et les mots stables caractérisés par $S_b < 0.18$ et $L > 15$. La polarisation de ce graphe indiquerait une qualité des méthodes.

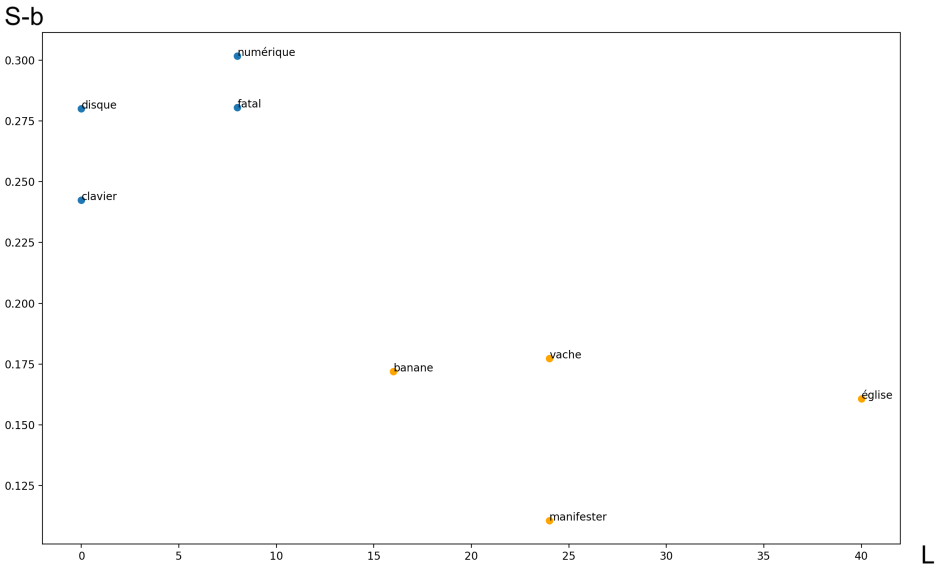


Fig. 5.1. Représentation des mots français sélectionnés en exemple selon des coordonnées (S_b, L) calculées entre les périodes 1900-10 et 1990-00. Les mots en bleu ont changé de sens entre ces deux périodes, les mots oranges sont considérés comme stables. Cette figure met en évidence la polarisation de la répartition des points.

Néanmoins, le point central du développement de ces méthodes était d'envisager une comparaison des voisinages des mots étudiés sans avoir recours au déplacement des plongements de mots, afin de proposer une analyse critique des méthodes proposées par Hamilton et al.

Réitérons donc cette même méthodologie en l'appliquant au vocabulaire décrit dans l'article, c'est à dire autant des exemples issus du jeu de 28 exemples, que des termes issus des top-10 des différentes méthodes et validés par les auteurs.

Dans le graphique 5.2, nous représentons ainsi en vert ces mots considérés comme ayant changé de sens. Nous avons en outre ajouté des mots que nous avons sélectionnés pour leur stabilité a priori et qui sont représentés en rouge. Enfin les mots issus des top-10 et considérés comme "artefacts" par Hamilton et al. sont affichés en noir.

Les scores affichés sont calculés à partir des représentations issues des modèles entrainés par Hamilton et al.

L'analyse de ce graphique met en exergue plusieurs points :

- En ne nous intéressant qu'aux mots présupposés stables (en rouge) et ceux ayant subi une diachronie (en vert), nous remarquons (ainsi que ce fut le cas lors des tests

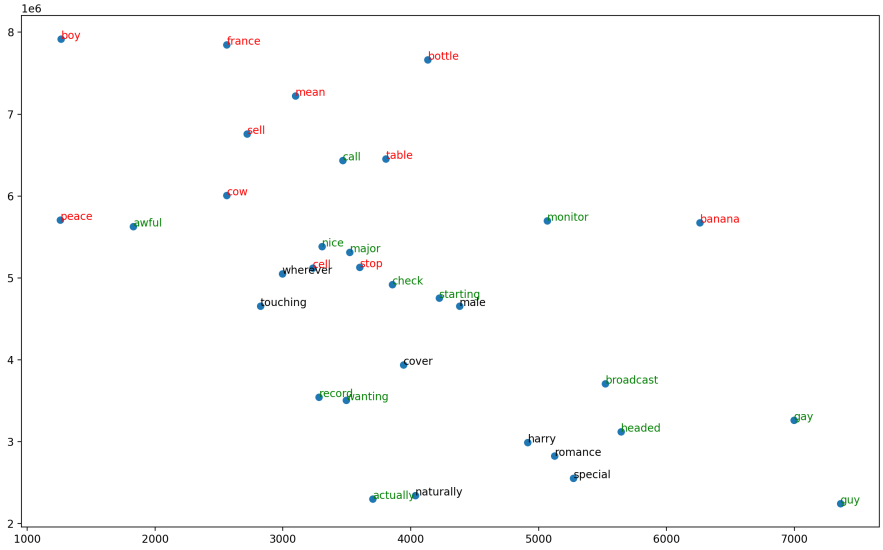


Fig. 5.2. Représentation des mots selon des coordonnées (S_b, L). Les mots utilisés sont issus des groupes définis en section 4.3. Ceux écrits en vert étaient ceux indiqués comme ayant changé de sens, les mots en rouge indiqués stables et les mots noirs ont été mal classés par les algorithmes de l'article d'Hamilton et al. La différence d'ordre de grandeur des valeurs avec la figure 5.1 est due au fait que ces deux figures ont été obtenues avec des modèles entrainés sur des corpus distincts.

précédents en français) l'existence d'une polarisation du graphique entre ces deux groupes.

Cette répartition permet de proposer la définition d'une limite linéaire déterminant le changement de sens. Celle-ci est calculée grâce un algorithme de Machine Learning : le *SVC linear*. Elle est tracée sur la figure 5.3.

Remarquons néanmoins le fait que l'augmentation du nombre de mots a dévoilé une frontière moins nette : certains mots comme "cell" ou "stop" sont situés du côté des mots ayant subi une diachronie, et "call" et "awful" se placent de l'autre côté.

- En nous concentrant d'autre part sur les mots noirs, nous remarquons immédiatement que leur position les place tous les huit dans la zone regroupant les mots ayant subi un changement de sens.

  Cette position sur le plan est particulièrement intéressante en ce qu'elle montre qu'en plus des hypothèses proposées par Hamilton et al. et décrites auparavant, ces termes connaissent des variations de voisinage aussi importantes que celles subies par les autres mots considérés comme ayant subi une diachronie par Hamilton et al.

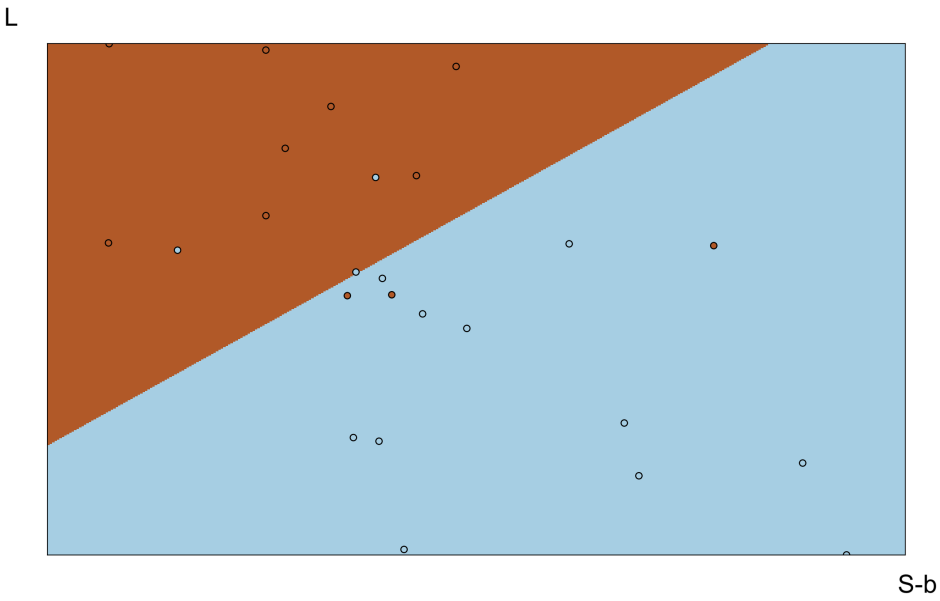


Fig. 5.3. Profil des zones de décisions dessinées par SVC linéaire d'après les mots issus de la figure 5.2. Les points en marron indiquent les coordonnées de mots supposés avoir changé de sens, les mots bleus de mots supposés stables. Les zones en marron et en bleu correspondent aux zones de classification.

5.2. Le paradoxe de la validation

Les figures obtenues sont ainsi l'occasion de revenir sur la méthode de validation qui a été proposée. En effet, tant les analyses proposées par Hamilton et al. que la figure 5.4 associent plusieurs termes considérés stables des caractéristiques identiques à celles de mots ayant changé de sens. Si les méthodes d'identification peuvent évidemment présenter des défauts à l'origine de ces résultats, il nous apparaît essentiel de questionner la nature de la méthode de validation. Comme nous l'avions fait remarquer, cette méthode s'appuie sur une recherche bibliographique pour confirmer ou infirmer une proposition de diachronie subie. La tension entre nos attentes et la méthode de validation peut se décliner selon les deux points suivants :

- D'un côté les labels des mots du jeu de données sont tirés des dictionnaires. Ainsi, nous utilisons les définitions lexicographiques, leur évolution et la connaissance des modifications associées comme validation de nos résultats.
- D'un autre côté nous cherchons à développer un outil capable d'identifier des diachronies non notifiées par les dictionnaires.

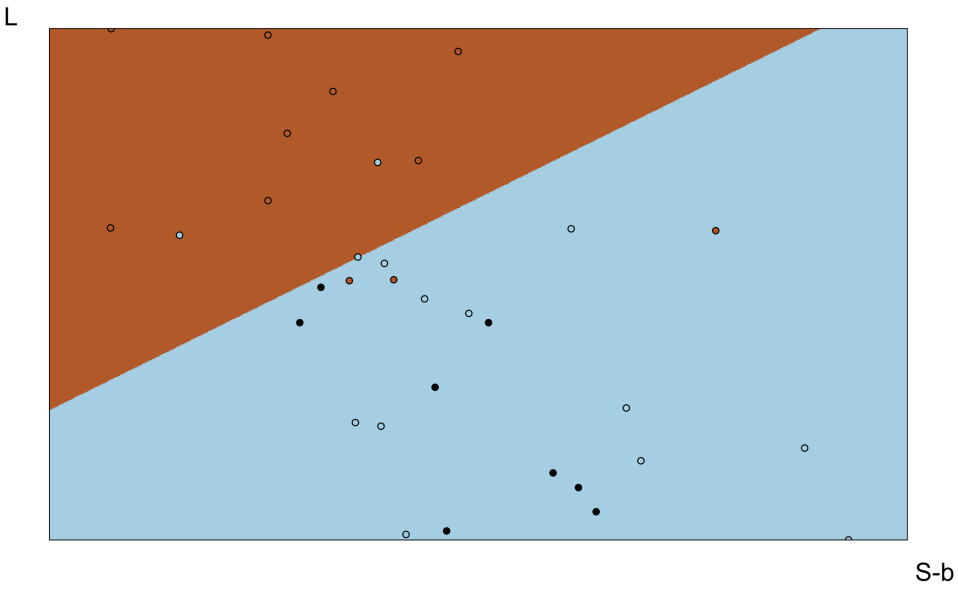


Fig. 5.4. Reprise de la figure 5.3 à laquelle sont ajoutés en noir les points des mots "artefacts" représentés dans la figure 5.2. Nous remarquons qu'ils sont tous situés dans la zone de décision correspondant aux mots ayant changé de sens.

A l'aune de ces considérations, il apparaît que la méthode de validation utilisée est paradoxale : les termes étudiés sont datés par les dictionnaires qui sont précisément les media auxquels nous souhaitons apporter de l'information et donc pour lesquels l'hypothèse est fait que les informations contenues sont partiellement incomplètes.

C'est donc cette différence entre la valeur des données du jeu de données et les aspirations des méthodes qui constitue le paradoxe central de notre mémoire : comment peut-on analyser les lacunes des dictionnaires si ceux-ci sont utilisés pour valider les résultats de nos méthodes?

Revenons sur un exemple issu de l'article d'Hamilton et al. Le mot "romance" (anglais) est classé par les méthodes d'Hamilton et al. comme ayant subi une diachronie, résultat que les auteurs jugent comme un artefact et en désaccord avec l'évolution réelle de ce mot en conséquence. Cette classification les conduit donc à classer "romance" comme une erreur d'annotation.

Nous affichons sur les figures 5.5 et 5.6 la représentation de ce mot et de son voisinage dans le Google N-Gram corpus aux périodes 1900-10 et 1990-00. Or, la considération de ces figures fait apparaitre une différence considérable de voisinage entre ces deux périodes : si la période la plus récente rattache le mot à un contexte traitant du sentimental, des histoires à l'eau de rose, la plus ancienne le place pour sa part dans un contexte d'enjeux

artistiques typiques du $XIX^{\text{è}}$ siècle. En reprenant la définition utilisée par Hamilton de diachronie, un mot subit une diachronie entre deux époques si son vecteur se déplace dans l'espace sémantique et qu'on voit une modification du contexte d'utilisation (voisinage) de celui-ci. L'exemple de "romance" vérifie ces deux critères.

Notons de plus, que nos mesures (figure 5.4) qui ne s'appuyent pour leur part que sur les changements de voisinage du mot, classaient également "romance" comme mot ayant subi une diachronie.

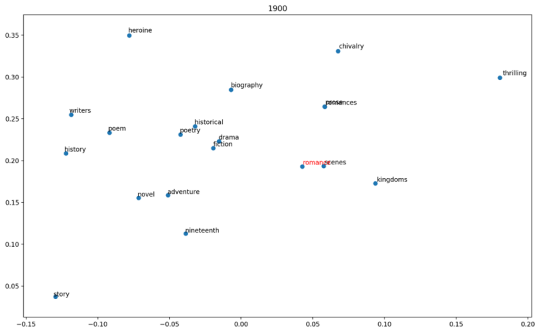


Fig. 5.5. Voisinage du mot "*romance*" (en anglais) selon la représentation entrainée par Hamilton et al. sur la tranche 1900-10 du Google N-Gram corpus

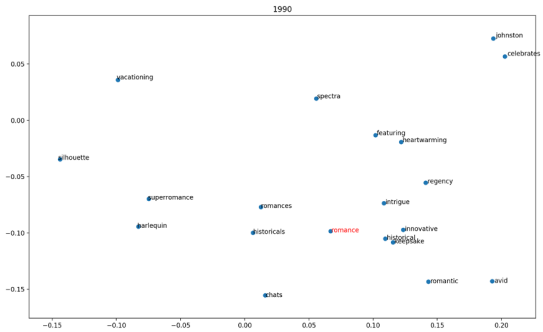


Fig. 5.6. Voisinage du mot "*romance*" (en anglais) selon la représentation entrainée par Hamilton et al. sur la tranche 1990-00 du Google N-Gram corpus

Ces résultats nous replongent donc dans les réflexions que nous avons présentées précedemment : si nous avons développé des méthodes ne s'appuyant pas sur le déplacement des vecteurs, les résultats obtenus présentent les mêmes ambiguïtés que celles relevées à la lecture des articles : les mots considérés comme n'ayant pas changé de sens présentent toujours des caractéristiques les assimilant à des mots ayant changé de sens.

Les exemples mettent ainsi en exergue la difficulté principale rencontrée pour l'évaluation : quelle définition de la diachronie sémantique devons-nous adopter ? Doit-on chercher des mots subissant un changement d'utilisation, c'est-à-dire ce qu'essaient de quantifier les méthodes présentées : une variation de l'environnement sémantique entre différentes périodes? Ou bien s'agit-il à l'inverse de variations de la définition d'un mot entre plusieurs dictionnaires, tels que sont identifiées les diachronies présentées dans le jeu de données de validation et dans la méthodologie proposés par Hamilton et al. [**8**] ?

Finalement les erreurs affichées par les différentes méthodes ne sont que des éléments supplémentaires soulignant la difficulté majeure que représente ce paradoxe. Avec la méthode de validation utilisée, aucune identification de nouveaux mots n'est possible : si les méthodes mettent à jour une diachronie non connue, le dictionnaire ne l'indiquera pas (par définition

de la diachronie non connue), ce qui conduirait à classer cette diachronie comme "erreur" de classification.

Et par conséquent les enjeux de notre sujet doivent nous conduire à réfléchir à l'élaboration d'une méthode de validation.

Dans ce cadre, nous remarquons une limite supplémentaire de la validation proposée par Hamilton et al. : si une vérification statique a été réalisée pour les mots stables, le jeu de données évaluant le déplacement diachronique ne contient pas d'exemples négatifs, c'est-à-dire que les méthodes ne sont jamais évaluées sur des mots dont il est avéré qu'ils n'ont pas changé de sens. La confirmation passera nécessairement par l'établissement d'un jeu de données plus complet contenant également des exemples négatifs. Et c'est précisément en travaillant sur ce point que nous pourrons nous assurer de l'immobilité des vecteurs de mots stables.

5.3. Vers le développement d'un jeu de données

Ces considérations et les résultats affichés sur la figure 5.4 nous ont ainsi ammenés à remarquer la nécessité de développer un nouveau jeu de données permettant de réaliser une évaluation des méthodes, et dont les datations sont obtenues en se passant de dictionnaires. Nous devons en particulier nous passer de la confrontation de plusieurs millésimes de dictionnaires pour en déduire la diachronie subie par un mot. A la place, il nous faut développer un jeu de données de subsitution corrigeant les lacunes du jeu de données précédent :

- Le jeu de données devra contenir autant des mots ayant subi une diachronie que des mots dont la stabilité est avérée.
- La construction du jeu de données devra se passer des datations proposées par les dictionnaires afin d'éviter l'introduction des biais notifiés précédemment.

Un tel jeu de données permettra enfin d'être une référence pour l'évaluation des méthodes d'identification des diachronies.

Chapitre 6

Création du jeu de données de validation

Dans le prolongement direct des conclusions établies précédemment, il nous est apparu nécessaire de créer un jeu de données de validation, c'est-à-dire un jeu de données recensant des mots auxquels sont associés des sens avec leur date d'apparition ou de disparition.

L'objectif de cette partie est donc de proposer une méthodologie pour créer ce jeu de données. Pour ce faire, nous nous appuyons sur des mots pour lesquels le dictionnaire recense plusieurs sens.

Parmi les pistes explorées, il nous a paru intéressant d'utiliser les différents millésimes d'un même dictionnaire. En effet, ces millésimes sont l'occasion de mettre à jour les définitions d'une entrée, en particulier en y supprimant les définitions obsolètes, et ajoutant les nouvelles définitions en vigueur. Néanmoins, en dépit de nos recherches, aucune ressource ne correspondait à nos besoins :

(1) Notons en premier lieu que si les millésimes les plus récents sont bien publiés simultanément sous formes numériques et physiques, ce n'est le cas que depuis récemment, et les millésimes les plus anciens n'existent que sous forme papier. Les périodes disponibles étaient trop resserrées et donc contraignantes pour pouvoir être utilisées.

(2) D'autre part, nous avons remarqué l'existence de projets visant à détailler les évolutions des dictionnaires en indiquant les changements de définitions ainsi que les apparitions et disparitions d'entrées de dictionnaires [1]. Néanmoins ces projets étant encore en cours, ils ne mettaient pas assez de données à disposition pour pouvoir s'appuyer dessus.

[1] Le projet le plus abouti que nous ayons consulté est celui de l'association *Club d'orthographe de Grenoble* qui propose les différentiels des dictionnaires du Petit Larousse et du Petit Robert. En mai 2021, seules les comparaisons entre les millésimes des périodes 1906-25 et 1998-2021 avaient été mises en ligne. Ce travail est toujours en cours et la liste des années disponibles s'agrandit. Leurs travaux sont disponibles sur le site `https://orthogrenoble.net/mots-nouveaux-dictionnaires/`

(3) Enfin, si nous avons songé à employer Wiktionary[2], l'absence d'une structuration générique entre les articles, et la présence assez aléatoire d'une rubrique étymologique indiquant la date d'apparition d'une définition nous a suffi à nous apercevoir que cette plateforme ne correspondait pas à nos besoins.

Finalement, avec l'ambition de rendre notre méthodologie translingue, facile d'utilisation et ne s'appuyant pas sur des données rares ou contraignantes, l'utilisation d'un algorithme de WSD nous a semblé être la plus adéquate. En effet, l'algorithme sélectionné s'appuie sur la ressource WordNet (voir section 2.3), ressource souple d'utilisation et largement répandue. Bien que l'utilisation de cette ressource nous limite aux sens de mots enregitrés dans sa base de donnée, elle devrait nous permettre d'établir une datation de ces sens enregistrés et donc de déterminer l'apparition d'une diachronie.

Ainsi, avec cet algorithme de WSD nous identifions la période d'apparition des sens des mots au cours du temps. L'intégration de ces périodes dans une chronologie permet de conclure quant aux diachronies. Nous appliquons cette méthode afin de construire le jeu de données requis en français.

6.1. Méthodologie

L'objectif étant de créer un jeu de données, nous proposons ici une méthodologie afin de récolter les mots qui constitueront ce jeu de données. Le jeu de données doit être constitué de mots sur lesquels nous disposons d'informations précises quant à leur diachronie sémantique entre deux périodes P1 et P2 définies.[3] Ainsi, nous nous intéressons à deux catégories de mots :

- Les mots stables entre les périodes P1 et P2 choisies : ces mots ne connaissent pas de diachronie sémantique ; la liste des sens d'utilisation pendant la période P1 est strictement identique à celle des sens de la période P2, et ces sens sont utilisés dans une proportion similaire.

Sens	1910-20	1990-00
Cat	92%	89%
Kitten	3%	9%
Autres	5%	2%

Tableau 6.1. Exemple de répartition des sens du mot "chat" stable entre les deux périodes

[2]Site disponible à l'adresse `https://www.wiktionary.org/`

[3]Notons que bien qu'il existe et qu'il soit fréquent, le phénomène de l'apparition d'un nouveau mot ne nous intéresse pas. Dans le cadre de notre étude nous ne nous intéressons qu'à l'acquisition ou la disparition du sens d'un mot déjà existant.

- Les mots présentant une apparition ou une disparition de sens entre les périodes P1 et P2 choisies : la liste des sens d'utilisation du mot pendant la période P1 est différente de la liste de la période P2. La diachronie se définit alors comme un ajout ou une suppression de sens.

Sens	Description	1910-20	1990-00
Canal	canaux drainant de l'eau	98%	33%
Canal Electrique	voie de circulation d'un signal éléctrique	∅	50%
Channel	canaux de communication du signal télévisé	∅	12%
Autres	∅	2%	5%

Tableau 6.2. Exemple de répartition des sens du mot "canaux" présentant l'apparition des sens *canalélectrique* et *channel* (canaux des chaînes télévisées) entre les deux périodes.

Grâce à l'identification de ces deux catégories de mots, nous construisons un jeu de données de validation pour les méthodes d'identification de diachronies.

Dans cette section, nous nous appuyons sur les textes de notre corpus issus du quotidien Le Devoir (décrit au chapitre 3).

La précision de ces processus étant conditionnée par certaines étapes, et nécessitant davantage de temps, nous avons décidé de nous concentrer sur les tranches 1910-20 et 1990-00, afin de vérifier l'apparition de sens de certains mots, sans chercher davantage de précision dans la datation. Celle-ci pourra faire l'objet de futurs travaux.

6.1.1. Détail du processus d'identification

La question étudiée présentement est donc celle de la datation de l'apparition ou disparition d'un sens S d'un mot m. Pour ce faire, nous proposons d'identifier pour les périodes P1 et P2 les sens d'utilisation du mot m, puis une comparaison entre ces sens permettra de conclure sur la diachronie.

Afin de déterminer la date d'apparition du sens S d'un mot m, nous employons une méthode de désambiguïsation qui a pour rôle d'attribuer un sens contextuel à chaque mot d'un texte. Ainsi, le traitement d'un texte par cet algorithme permet d'attribuer un sens d'utilisation à chaque mot de ce texte.

Nous employons cet algorithme sur les textes d'une tranche du corpus. Nous obtenons alors un fichier pour chaque tranche. Ce fichier contient dès lors tous les textes de cette tranche dans une forme désambiguïsée, c'est à dire que chaque mot présent dans le texte est associé à son sens d'utilisation (cette forme désambiguïsée a été illustrée dans la section figure 2.3.2). Insistons sur le fait qu'un mot utilisé à plusieurs reprises dans le texte se verra associé un sens pour chacune de ses occurrences.

Une fois ces fichiers générés, les informations peuvent être regroupées. Une liste de vocabulaire est d'abord établie : chaque mot différent utilisé dans la tranche est identifié. Puis, les occurrences de chacun de ces mots de vocabulaire sont recherchées et les sens utilisés enregistrés (voir figure 6.1).

Fig. 6.1. Exemple de réunion des sens des mots "déjà", "financière" et "des" utilisés dans le texte désambiguisé de l'exemple de la section 2.3.2

Le résultat de cette étape laisse apparaitre une cartographie de la tranche ; nous connaissons tous les mots de vocabulaire disponibles et pour chacun d'entre eux les sens d'utilisation. Enfin, nous calculons une répartition : pour chaque sens S1, S2, ..., Sn d'un mot donné, nous établissons son pourcentage d'utilisation par rapport au nombre d'occurrences total du mot m dans la tranche (n_m), et le nombre d'utilisation du sens S_x (n_{Sx}). Ces dénombrements permettent d'établir une cartographie statique globale : dans chaque tranche est calculée la répartition des sens du mot.

6.1.2. Construction du jeu de données par concaténation

Une dernière étape est nécessaire à l'établissement de la chronologie : En comparant les répartitions entre plusieurs tranches et dans un ordre chronologique, apparait un panorama "glissant", c'est-à-dire que nous pouvons désormais confronter les modifications de l'usage d'un sens au cours du temps et l'apparition ou la disparition de certains d'entre eux.

Finalement, une telle chronologie autorise la déduction de deux types d'informations :

- La modification d'usages : sur la figure 6.2, nous lisons qu'en 1900-10, le mot **direct** est employé dans le sens "*Direct*" (sens lié à la direction, l'orientation...) dans 88% des cas, tandis qu'en 1980-90, ce sens "*Direct*" n'est plus utilisé que dans 57% des cas. Nous assistons ainsi à un recul de l'utilisation de ce sens d'utilisation du mot **direct**.

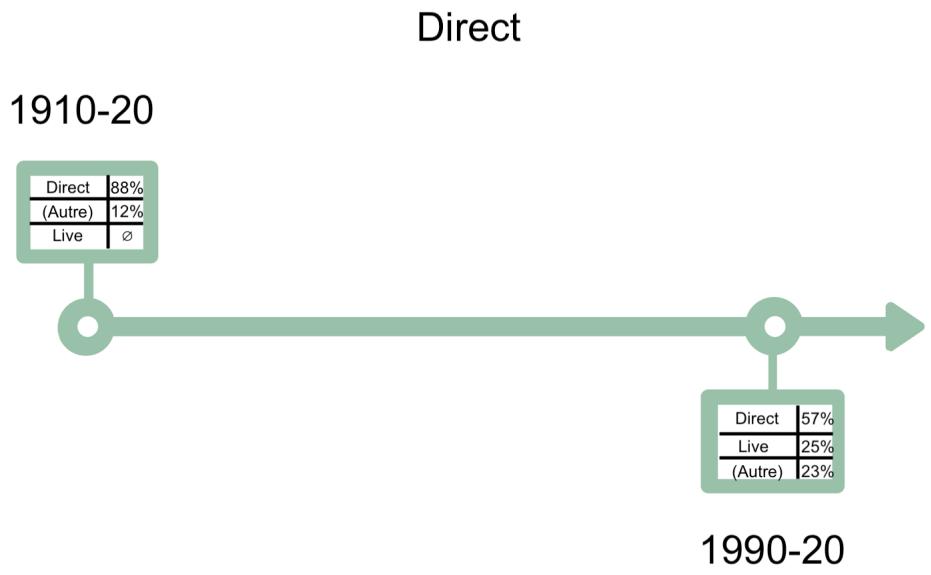


Fig. 6.2. Exemple de répartition des sens du mot "direct" au cours du $\mathbf{XX}^{ème}$ siècle : le sens "direct" désigne l'utilisation directionnel du mot (directement, chemin le plus direct...), tandis que le sens "live" fait référence à la diffusion radio-télévisée (être en direct). Nous remarquons que le sens "direct" apparaît entre les périodes affichées. on remarque l'apparition du sens "live" autour de 1940.

- En 1900-10, le sens "*Live*" du mot **direct** (sens désignant les emissions télévisées ou radiophoniques émises sans latence) n'était pas en usage, et le devient en 1990-00 ; nous en concluons que ce sens est apparu entre ces deux périodes : un mot tel que "direct" est alors labélisé **"diachronie subie entre les périodes** $1900-10$ **et** $1990-00$**".**

A l'inverse, dans le cas où les sens d'utilisation étaient restés identiques, et les pourcentages avaient peu varié, nous aurions labélisé le mot M comme **"stable entre les périodes** $1900-10$ **et** $1990-00$**"**.

6.2. Désambiguïsation

Afin d'obtenir une cartographie de l'utilisation d'un mot dans une tranche, nous utilisons un algorithme de désambiguisation qui associe à chaque mot d'un corpus un sens d'utilisation (voir section 2.3.2).

Le modèle de désambiguisation utilisé est un modèle développé par l'équipe GETALP du laboratoire LIG [**17**] présenté en section 2.3.2. Celui-ci utilise la taxonomie de WordNet [**6**] afin de déterminer les sens possibles d'usage d'un mot. C'est-à-dire qu'il utilise le contexte d'utilisation d'un mot afin de déterminer à quel sens elle fait référence parmi les sens possibles de ce mot.

6.2.1. Evaluation du modèle de désambiguïsation

La phase d'entrainement du modèle permet un *tuning* du réseau utilisé pour la classification, c'est à dire pour la sélection des synsets.

Pendant cette première phase d'entrainement, nous utilisons le modèle présenté en 2.3.2 et développé par Vial et al. [**17**] avec les hyperparamètres suivants :

- taille de batch : 25
- 24 couches de transformers
- encoder de type transformer avec une taille de couche cachée de 3072
- 16 têtes par encoder

Les hyperparamètres que nous modifions au cours des expériences sont : le modèle BERT utilisé (CamemBERT ou FlauBERT) et le nombre de périodes d'entrainement.

Nous présentons dans le tableau 6.3 les résultats de l'évaluation des modèles sur la tâche 12 de SemEval 2013 [**14**]. Celle-ci s'appuie sur un corpus constitué d'une liste de phrases en français dont chaque mot a été manuellement associé à un synset ; la tâche consiste dès lors pour le modèle à retrouver le même synset que celui associé à chaque mot du corpus.

Modèle BERT	*Taille*	*époques*	*Précision validation*
Flaubert	large	2	51.1%
Flaubert	large	20	52.4%
Camembert	large	2	25.6%
Camembert	large	20	**52.9%**

Tableau 6.3. Pourcentage de désambiguïsation et précision des modèles entrainés avec la version modifiée de l'algorithme

Les résultats obtenus en utilisant directement l'algorithme tel que proposé présentaient des résultats assez décevants (moins de 55%) et surtout trop faibles pour être utilisés pour notre tâche de datation.

L'étape de désambiguïsation n'aspire pas à fournir une désambiguïsation totale des textes utilisés. En effet, notre objectif étant d'obtenir une cartographie des utilisations d'un mot, il nous est davantage utile de garantir un niveau de confiance haut sur les sens associés aux occurrences plutôt que désambiguïser un texte complet ; c'est-à-dire qu'on préfèrera ne désambiguïser qu'une portion des mots pour lesquels un certain niveau de précision est garanti et non un maximum de mots.

Nous modifions ainsi le modèle en conséquence afin qu'il ne désambiguïse plus tous les mots d'un texte, mais uniquement ceux considérés comme fiables[4], c'est à dire pour lesquels le niveau de confiance du modèle est supérieur au seuil fixé (softmax supérieur à 0.5)..

Nous présentons dans le tableau 6.4 les résultats de l'évaluation des modèles modifiés sur la tâche 12 de SemEval 2013 [**14**]. La précision affichée n'est désormais calculée que sur les mots considérés comme fiables. Nous n'affichons en outre que le pourcentage de mots désambiguïsés.

Modèle BERT	*Taille*	*Époques*	*Précision validation*	*% désambiguïsation*
Flaubert	large	2	75.3%	42.9%
Flaubert	large	20	66.8%	**66.2%**
Camembert	large	2	**82.0%**	7.7%
Camembert	large	20	70.9%	59.9%

Tableau 6.4. Pourcentage de désambiguïsation et précision des modèles entrainés avec la version modifiée de l'algorithme.

Les deux caractéritiques proposées dans le tableau 6.4 ont un impact direct sur la qualité de la désambiguïsation:

- La précision sur le jeu de données de validation met en exergue la qualité des résultats proposés. Plus celle-ci est élevée, plus l'on pourra accorder de confiance aux résultats de la désambiguïsation, et ainsi proposer une finesse dans l'analyse. Notre objectif étant de proposer une identification la plus précise possible, nous cherchons à réduire les incertitudes dues à la qualité de l'algorithme de désambiguïsation.
- Le pourcentage de désambiguisation représente la part d'occurences d'un mot qui sera désambiguisée. De nouveau, la maximisation de cette part de désambigïsation permet d'envisager une confiance plus grande : statistiquement les incertitudes sont inversement proportionnelles à la taille de l'échantillon. En augmentant ce pourcentage, nous augmentons la qualité des résultats.

Finalement, nous devons viser un compromis acceptable, c'est à dire un équilibre entre la justesse des résultats et le nombre de résultats proposés, c'est pourquoi nous selectionnons le modèle reposant sur un CamemBERT$_{large}$ entrainé durant 20 époques.

Un tel modèle nous permet d'obtenir une liste de mots fiablement désambiguïsés parmi lesquels seront repérés les mots ayant subi une diachronie. Ainsi, au cours de cette étape, nous réduisons les occurences candidates afin de garantir un certain niveau de confiance des résultats proposés.

[4]La présentation de la modification de ce modèle et du détail de son fonctionnement sont proposés en annexe C.1

Remarquons que la littérature [**17**] laisse apparaitre des résultats très élevés (plus de 80% de précision) sans utilisation de seuils de confiance avec certains modèles anglophones de BERT, ce qui suggère que des résultats plus précis pourraient être obtenus en travaillant sur de la datation de vocabulaire anglais.

6.2.2. Désambiguïsation du corpus d'étude

Après avoir sélectionné et modifié notre algorithme de désambiguïsation, nous l'avons mis en œuvre afin de désambiguïser des textes issus de notre corpus.

Dans le cadre de cette création de jeu de données, deux éléments sont à noter :

(1) La comparaison entre des tranches temporellement éloignées permet de mettre à jour davantage de diachronies ; en effet, l'écart temporel des tranches implique une durée plus longue du glissement des sens des mots. Cela permet également d'observer l'apparition de nouveaux sens d'un mot, ou de dénominations à l'aide de mots déjà existants de concepts apparus dans l'intervalle.

(2) A l'inverse, la comparaison entre des tranches temporellement rapprochées permet d'obtenir une datation plus fine de la diachronie ; en effet, plus les tranches sont proches, plus les variations sont précises. Ainsi, les différences de répartition d'un sens entre deux périodes contiguës permet de dater un changement de sens à une précision égale au double de la taille des tranches (c'est à dire entre le début de la première et la fin de la seconde).

Ces informations connues, il nous faut ajouter les problèmes liés aux incertitudes : la section 6.2 a mis en évidence le fait que les algorithmes de désambiguïsations en français, même modifiés obtiennent des précisions maximales de 82%. Or le modèle sélectionné présente une précision de 70% ne permettant pas de conclure quant à la nature d'un changement de proportion de l'utilisation des sens d'un mot que si celui-ci est d'au moins 30%. Dès lors, nous nous retrouvons confrontés à la nécessité de faire apparaître les variations les plus fortes possibles, afin qu'elles excèdent les taux d'incertitudes et permettent de conclure définitivement quant à une variation du sens d'un mot.

Ces considérations nous ont poussés à proposer l'approche suivante : nous nous cantonnons à une désambiguïsation des tranches les plus éloignées, c'est-à-dire sur une période maximale notée P_{max}, afin d'en tirer les changements les plus marquants. Cette étape permettra de récupérer une liste de mots pour lesquels une diachronie sera connue durant la période P_{max}. Dans un second temps, de futurs travaux seront l'occasion de resserrer les mailles de la comparaison en comparant les tranches deux à deux. La comparaison des mots

de la liste obtenue nous permet d'obtenir le fameux panorama glissant (cf figure 6.2), et donc de proposer une datation plus précise.

6.2.3. Quantification des résultats

Comme décrit auparavant, nous utilisons les textes issus du Devoir. Les textes de ce quotidien s'étendent entre 1910 et 2000, permettant de former 9 tranches de 10 ans. Nous obtenons donc une période maximale $P_{max} = 90$ ans, et une datation précise à $2 \times P$ près, c'est-à-dire 20 ans.

La désambiguïsation des fichiers contenant les textes des tranches 1910-20 et 1990-00 aboutit aux résultats suivants :

Tranche	*# occurences du corpus*	*# occurences désambiguïsées (%)*	*Taille du vocabulaire*	*# mots vocabulaire désambiguïsés (%)*
1910-20	97.9M	27.9M (28%)	3.6M	0.80M (22%)
1990-00	193.2M	51.5M (27%)	3.3M	0.73M (22%)

Tableau 6.5. Quantification de la désambiguisation en occurences et en vocabulaire.

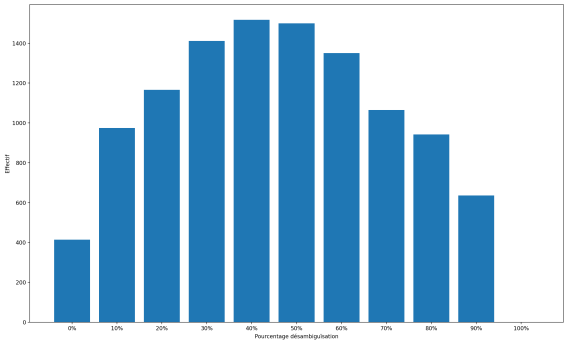


Fig. 6.3. 1910-20 : pourcentage désambïguation mots ayant été désambiguisés plus de 200 fois

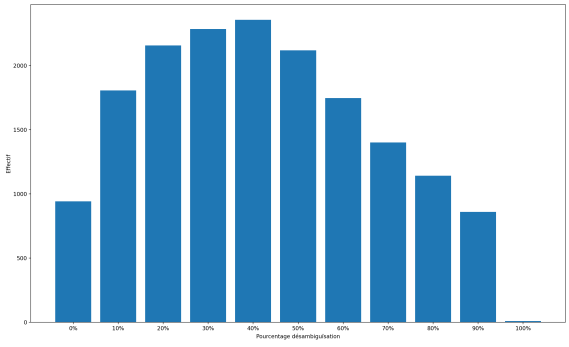


Fig. 6.4. 1990-00 : pourcentage désambïguation mots ayant été désambiguisés plus de 200 fois

Des analyses complémentaires montrent néanmoins une prédominance des mots de vocabulaire désambiguïsés une seule fois[5]. Afin de garantir la précision des résultats proposés par la suite, nous décidons de ne considérer que les mots de vocabulaire désambiguïsés au moins 200 fois.

[5]Ces analyses peuvent être consultées en annexe C.2

Les figures 6.3 et 6.4 présentent les pourcentages de désambiguïsation totales (donc par rapport à toutes les occurences de ces mots) des mots désambiguïsés au moins 200 fois. Nous remarquons un lissage avec un sommet aux environs de 50% en accord avec les taux de désambiguïsation obtenus lors de l'évaluation de l'algorithme de désambiguïsation utilisé.

6.3. Identification des mots

6.3.1. Méthodologie de sélection des mots

Afin d'établir une comparaison des sens utilisés entre les différentes périodes, nous unifions dans chaque tranche les sens utilisés de chaque mot (voir figure 6.2). Pour chaque tranche du corpus considérée, nous récupérons tout le vocabulaire sur cette tranche, puis pour chaque mot de vocabulaire, nous calculons la répartition des synsets associés au cours de la tranche. Le pseudo-code de l'algorithme utilisé est proposé en annexe A.4.

Cette unification nous permet d'obtenir une structure stockant les pourcentages d'utilisation de chaque sens de chaque mot. La comparaison de deux de ces structures permet de récupérer une liste mots de vocabulaire a priori utilisables dans le jeu de données.

Seuls les mots ayant changé de sens nous intéressent, nous ne gardons donc que les mots présents pendant les deux périodes. Puis, nous appliquons un tri selon les critères suivants :

- Nombre minimal de désambiguïsations par tranche : afin de garantir une fiabilité des résultats, nous appliquons un tri sur le nombre de désambiguisations effectuées du mot considéré ; un mot trop peu désambiguïsé présente des incertitudes trop grandes pour garantir une fiabilité de la comparaison. Ce nombre de désambiguïsation doit être atteint dans chacune des deux périodes considérées.

- Pourcentage minimal d'apparition d'un synset : pour être considéré dans l'analyse, un synset doit être apparu dans un pourcentage assez élevé dans au moins l'une des deux périodes considérées. Ceci permettant notamment de ne pas considérer l'association avec un synset due à une erreur de déambiguïsation. En particulier, la précision de l'algorithme utilisé présente une incertitude qui laisse envisager de telles erreurs.

- Différence de répartition minimale d'un synset : il s'agit de la définition d'un seuil de différence ; un synset dont la différence absolue de répartition entre les deux tranches excède ce seuil est considéré comme ayant subi une diachronie. Si la différence naturelle entre l'époque supérieure et l'époque inférieure est positive, il s'agit d'une augmentation de l'utilisation de ce sens ; une différence négative est la marque du recul de l'utilisation d'un sens.

- Différence de répartition maximale d'un synset : à l'inverse du précédent, il s'agit d'un seuil en deçà duquel un synset est considéré comme stable. Les variations légères d'utilisation d'un synset sont naturelles, notamment du fait des incertitudes. Nous définissons ainsi un taux de variation normal, indiquant les variations acceptables du pourcentage sans impliquer de variation d'utilisation d'un synset.

Afin de sélectionner les mots ayant subi une diachronie entre les deux périodes 1910-20 et 1990-00, nous recherchons des mots dont un sens est apparu ou a disparu entre ces deux périodes. Pour ce faire, nous appliquons le tri décrit ci-dessus. Afin d'éviter les ecueils liés aux incertitudes causées par les étapes d'OCR et de désambiguisation, nous choisissons des seuils dépassant celles-ci :

- Nombre minimal de désambiguïsations par tranche : 200 apparitions
- Pourcentage minimal d'apparition d'un synset : 10%
- Différence de répartition minimale d'un synset : 70%

Une deuxième méthode complète celle-ci : il s'agit de chercher les mots dont un synset était absent dans l'une des deux périodes, et présent dans l'autre période dans une proportion supérieure à 10%.

Finalement, nous obtenons une liste de mots dont au moins un synset est apparu ou a disparu entre les deux périodes. Le tableau 6.6 présente les répartitions du mot "émissions", considéré comme tel.[6]

Sens	1910-20	1990-00
issue	92%	10%
broadcast	<10%	42%
television program	<10%	37%

Tableau 6.6. Exemple de répartition des sens du mot "émissions" entre les périodes 1910-20 et 1990-00. Le sens *issue* fait référence aux fait de mettre en circulation; *broadcast* à toute transmission réalisée par le moyen d'ondes et *television program* aux programmes retransmis à la télévision.

[6]Cet exemple est l'occasion d'insister sur un point du traitement du vocabulaire. Il est fréquent en TAL de faire appel à des algorithmes de *lemmatization*, c'est à dire d'unification des occurences dérivées d'un mot. Ainsi, un tel algorithme proposera d'unifier les mots "emission" et "emissions" sous la banière "EMISSION", ou encore "vendu" et "vendit" sous la banière de "VENDRE". Or, nous avons fait le choix de ne pas faire appel à ce type de méthodes. En effet, certains termes ne sont utilisés dans certains contextes que sous une forme dérivée : c'est le cas du mot "vacances", qui n'est utilisé qu'au pluriel afin de désigner des congés, quand la forme "vacance" n'indique que "[l']état de ce qui est vacant" . Afin d'éviter de noyer l'utilisation spécifique d'un dérivé dans les occurences des autres formes, nous conservons tous les termes sous la forme utilisée dans les textes.

Pour les mots stables, nous devons nous intéresser à l'ensemble des synsets associés à ce mot, et nous assurer que tous ont une répartition stable au cours du temps. Les incertitudes d'OCR impliquent l'apparition de mots mal identifiés au sein du vocabulaire, et les erreurs de désambiguïsations rattachent jusqu'à 30% de sens mal identifiés. Afin de dépasser ces erreurs, nous proposons de ne considérer les éléments dépassant les seuils suivants :

- Nombre minimal de désambiguïsations par tranche : 200 apparitions
- Pourcentage minimal d'apparition d'un synset : 5%
- Différence de répartition maximale d'un synset : 10%

Ainsi, les mots dont tous les synsets respectent ces contraintes entre les deux périodes sont considérés comme stables. Afin de faciliter la phase d'évaluation (6.4), nous nous contentons des mots auxquels ne sont associés qu'un ou deux synsets. Le tableau 6.7 présente les synsets du mot "monarque", considéré comme stable.

Sens	1910-20	1990-00
sovereign	98%	99%

Tableau 6.7. Exemple de répartition des sens du mot "monarque" entre les périodes 1910-20 et 1990-00

6.3.2. Résultats de la comparaison

Nous appliquons successivement chaque critère décrit en 6.3.1 aux fichiers désambiguïsés des tranches 1910-20 et 1990-00. Nous trouvons ainsi 6424 mots présents durant les deux périodes et désambiguïsés plus de 200 fois durant les deux périodes, soit une réduction drastique du nombre de mots de vocabulaire considérés.

Afin d'identifier les mots ayant subi une diachronie, nous repérons parmi ceux-ci, 5514 mots ayant en outre soit :

- au moins un synset représentant 10% ou plus des occurences désambiguïsés en 1910-20 et disparu en 1990-00.
- au moins un synset absent en 1910-20 et représentant 10% ou plus des occurences désambiguïsés en 1990-00

Enfin, nous dénombrons 119 mots respectant toutes les contraintes fixées précédemment.

Dans le but d'identifier les synsets stables, nous reprenons les 6424 mots désambiguïsés plus de 200 fois durant les deux périodes. Parmi ceux-ci, nous en dénombrons 4652 dont aucun synset n'apparaît entre les deux périodes (selon les critères fixés précédemment).

Enfin, 3587 mots n'ont aucun synset connaissant une variation supérieure à 10%.

Finalement, nous obtenons:

- 122 synsets considérés comme ayant apparu ou disparu entre les deux périodes [7]
- 3587 mots stables ayant 2 sens ou moins, dont 2463 sont monosémiques

Le nombre d'exemples obtenus illustre quantitativement les problèmes liés au WSD et à l'OCR : si nos tranches comptaient plusieurs centaines de milliers de mots désambiguïsés, seuls 6424 parmi eux étaient désambiguïsés plus de 200 fois dans les deux tranches à la fois, réduisant drastiquement la taille du vocabulaire considéré. D'autre part, ces résultats présentent près de 50% de mots de stables, en dépit des contraintes drastiques appliquées à leur sélection, ainsi que 2% de mots ayant changé de sens, résultat en accord avec la littérature qui fait état d'un phénomène épisodique.

Qualitativement, ces résultats sont intéressants à plus d'un titre. En premier lieu, la lecture avec un regard non lexicographique suffit souvent à valider la diachronie ou la stabilité proposée. Mais il est tout aussi passionnant de s'intéresser aux champs d'utilisation de tels mots : Ainsi, du côté des nouveaux sens de mots, nous trouvons nombre d'exemples issus du lexique des nouvelles technologies, en accord avec les évolutions technologiques du **XX**^{ème} siècle. Une part non négligeable de ces exemples relèvent également de considérations géographiques voire géopolitiques. Enfin, ces changements de sens sont également très révélateurs de certains changements sociétaux survenus au cours du **XX**^{ème} siècle.

Nous proposons dans les tableaux 6.8 et 6.9 un échantillon de quelques exemples de mots récupérés grâce à la comparaison précédente.

Mot	**Ex. ancien sens**	**Nouveau sens**	domaine nouveau sens
tissu	Matériau	Framework (structure sous-jacente)	industrie
émissions	Mettre en circulation	Programme télévisé	technologie
direct	Direction spatiale	Réalisé en direct	technologie
union	Réunion, combinaison	Union Soviétique	géopolitique
nations	Etat	Organisation des Nations Unies	géopolitique

Tableau 6.8. Exemple de mots ayant subi une diachronie sémantique entre les périodes 1910-20 et 1990-00. Les anciens sens restent utilisé, mais les mots se voient ajouter les sens notés "Nouveaux".

6.4. Évaluation

Pour clore ce chapitre, nous terminons par une étape de validation. Le jeu de données qui est construit est soumis à l'évaluation d'annotateurs afin de confirmer la pertinence des résultats proposés.

[7] Parmi les 119 mots recensés, certains ont vu l'apparition ou la disparition de plus d'un synset, conduisant à un nombre de synsets recensés supérieur au nombre de mots.

Mot	Sens_1	Sens_2 (cas échéant)
quatre	cardinal	$\emptyset$
donc	therefore	consequently
symphonie	symphony	$\emptyset$
privilège	privilege	prerogative
couteau	knife	$\emptyset$
témoigner	testify	express

Tableau 6.9. Exemple de mots sémantiquement stables entre les périodes 1910-20 et 1990-00

6.4.1. Création du jeu de données à évaluer

Le jeu de données évalué est constitué de 272 associations "mot-sens" (notés respectivement M et S) candidates. Ces candidats sont labélisés selon deux catégories :

- Apparition : le sens S du mot M est absent du corpus en 1910-1920 et présent en 1990-00. 122 candidats sont compris dans le jeu de données.
- Stable : le sens S du mot M est présent du corpus en 1910-1920 et en 1990-00, dans des proportions différant de moins de 10%. 150 candidats sont compris dans le jeu de données. Il s'agit uniquement de candidats monosémiques échantillonés aléatoirement parmi les 2463 mots stables récupérés selon la méthodologie décrite en section 6.3.2.

6.4.2. Création de la tâche d'évaluation

Pour évaluer les résultats un outil d'évaluation a été créé (illustrée en figure 6.5). Celui-ci propose à des annotateurs une liste de candidats. Un candidat est constitué des informations suivantes :

- Un mot : mot dont nous cherchons à évaluer la présence ou l'absence du sens associé
- Un sens : utilisation possible du mot précédent
- La définition du sens considéré : issue de WordNet, définit en anglais le sens considéré
- Une liste de contextes : exemples d'utilisation du mot candidat dans le sens candidat. Ces contextes sont issus du corpus Le Devoir sur la période 1990-2000 et proposent un contexte constitué de 30 mots avant et 30 mots après le mot candidat

L'outil demande pour chaque sens candidat de répondre à la question suivante : "Le mot M peut-il avoir été employé dans le sens S en 1910-1920" ? Pour y répondre, l'évaluateur dispose des caractéristiques du candidat, et peut faire défiler la liste des contextes. Enfin, l'outil propose des boutons permettant à l'évaluateur de répondre à la question d'évaluation par oui ou par non. Les réponses sont enregistrées pour chaque candidat et centralisées.

L'évaluation consiste dès lors à vérifier si les sens labélisés comme "apparition" recueillent bien des réponses "non" de la part des évaluateurs, et les sens "stables" recueillent des "oui".

Enfin les évaluateurs se voient proposer une option "ne s'applique pas" afin d'indiquer que la proposition présente des obstacles qui rendrait obsolète et non adéquate une évaluation.

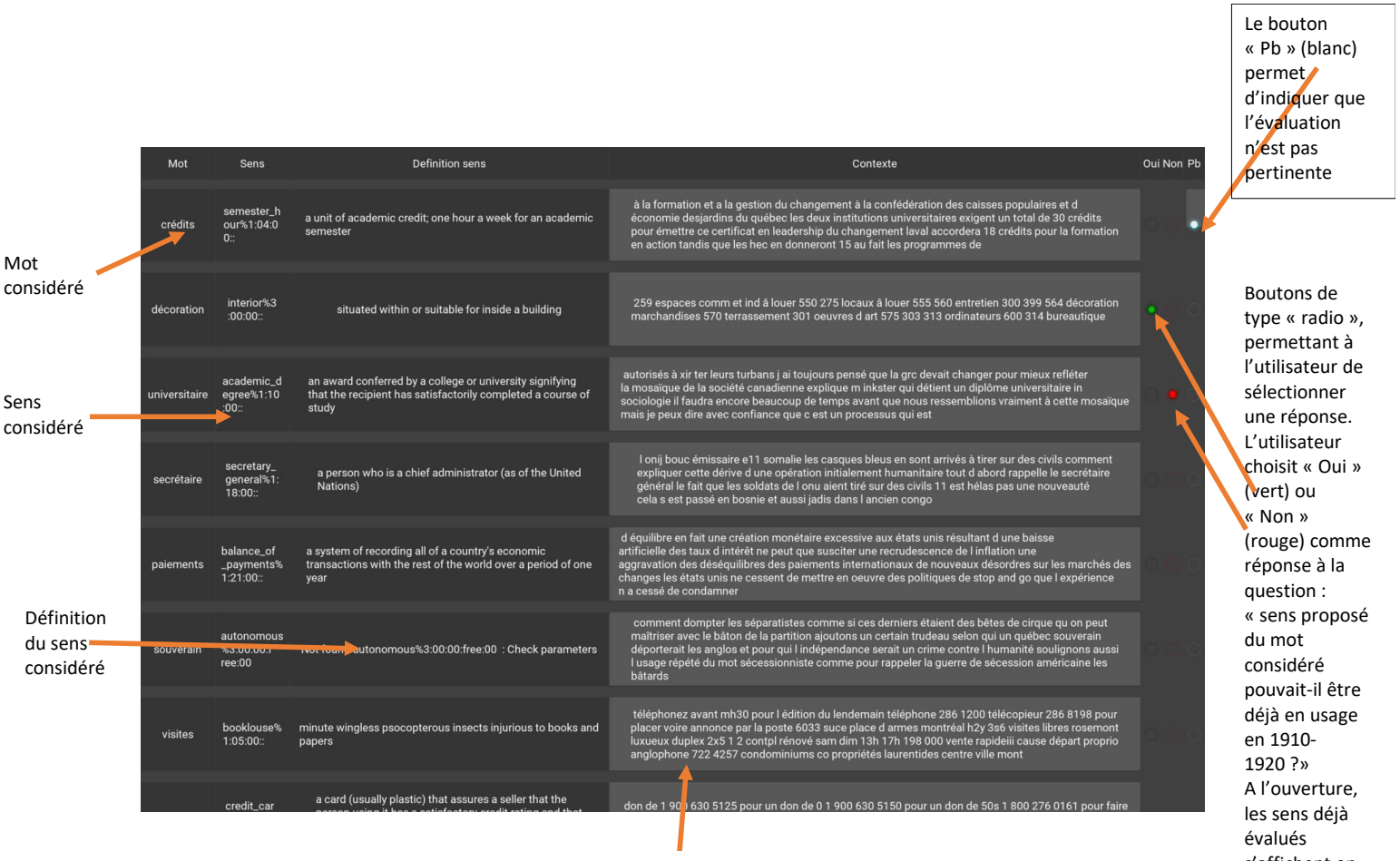


Fig. 6.5. Présentation de l'interface d'évaluation fournie aux évaluateurs.

6.4.3. Distribution de l'outil

Cet outil d'évaluation a été distribué à des membres de l'équipe enseignants-chercheurs au département de linguistique de l'UdeM.

Les résultats proposés ont été consolidés ainsi :

(1) Proposition examinée par un seul évaluateur est directement considérée comme valide si la proposition est identique à celle issue de notre algorithme.

(2) Proposition examinée par deux évaluateurs : considérée comme valide si et seulement si les deux propositions sont identiques à celle issue de notre algorithme. Ainsi dans les cas où les deux évaluateurs répondent différemment, la proposition la proposition est rejetée (erreur d'identification).

Durant la phase d'évaluation, les évaluateurs ont été confrontés à la même liste de mots, et à l'issue de cette phase, ils avaient eu connaissance de chaque mot proposé et pu soumettre une annotation (y compris l'annotation "problématique") pour chacun d'entre eux.

La première remarque des évaluateurs a concerné la qualité de la désambiguïsation, et ce en dépit des efforts accomplis pour l'amélioration des performances de l'algorithme de WSD. En effet, nombre de contextes identifiés ne correspondaient nullement à des phrases correctes en français. Ces problèmes concernaient particulièrement des séries de nombres entrecoupées de lettres et interprétés comme des contextes. Chaque mot est interprété par l'algorithme de WSD comme un terme à désambiguïser, et ces contextes ont donc conduit à nombre d'erreurs de désambiguïsations.

Ce cumul d'erreurs d'OCR et de désambiguïsation ont ainsi conduit à de fortes proportions de sens en erreur au sein de certaines tranches.

Ainsi, les évaluateurs n'ont désigné que 73% de mots comme ne présentant pas de problèmes. En détail, 82% des mots ayant changé de sens et 67% des mots stables proposés à l'évaluation ont été considérés comme issus de contextes non problématiques. Ces mots ont donc été annotés par au moins un évaluateur. Nous notons ainsi une proportion importante de mots n'ayant reçu aucune annotation (respectivement 18% et 33%).

En parallèle, et en ne considérant que les propositions non problématiques, les évaluateurs ont pu ainsi proposer de confirmer ou infirmer l'existence de certains sens en 1910-20. Ils ont toutefois indiqué avoir tendance à hésiter à refuser l'existence d'un sens dans cette période, et à l'inverse à valider aisément une existence. Ils ont également indiqué une nécessité de se contraindre à prendre la décision de cliquer sur le bouton de refus. De plus, ils ont indiqué une tendance générale à ressentir les sens comme existants, cela étant dû au fait que nous nous intéressons qu'à des sens aujourd'hui existants, et donc paraissant naturels.

Selon ce mode d'évaluation, nous obtenons les résultats suivants :

(1) Validation de 90.0% des mots stables proposés.

(2) Validation de 55.1% des diachronies proposées.

6.4.4. Description du jeu de données

A l'issue de cette étape d'évaluation, nous obtenons un jeu de données de 151 mots, dont 59 ayant subi une diachronie et 92 stables. La liste complète des mots est proposée en annexe D. De plus, tous les exemples proposés le long de ce mémoire proviennent des listes obtenues avec cette méthodologie.

Une première visualisation des exemples ayant subi une diachronie permet de confirmer les thématiques principales des mots concernés par ce phénomène : la technologie (16 exemples), la géopolitique (7 exemples), ou encore l'économie (9 exemples).

Une lecture davantage historique permet également au lecteur attentif d'obtenir une cartographie des transformations et évennements marquants du $\mathbf{XX}^{ème}$ siècle : les 16 exemples qui relèvent de sens associés à la technologie sont l'illustration des grands changements sociétaux survenus au $\mathbf{XX}^{ème}$ siècle, avec l'arrivée et la démocratisation de la radio et de la télévision (*station* au sens de chaînes de radio), le développement de l'analogique et du numérique (*enregistrer* au sens de stocker une information en format numérique ou analogique), le développement du secteur aérien (*vols* pour vols dans l'espace). De même, les termes relevant de la géopolotique sont associés à de nouveaux sens liés à des évenements tels que l'indépendance d'Etats et à leur nouvelle appellation (le nom de *Congo* qui désignait un fleuve, peut désormais faire référence à deux pays ayant obtenu leur indépendance en 1958 et 1960), ou aux conflits majeurs (*froide*, en référence à la guerre froide).

Néanmoins, il nous semble tout aussi essentiel d'examiner les mots stables. En effet, ils constituent une nouveauté par rapport aux groupes utilisés pour l'évaluation dans la littérature.

Parmi les mots stables obtenus, nous dénombrons ainsi diverses catégories grammaticales de mots, avec notamment 15 verbes dont 4 à l'infinitif, 18 adjectifs et 51 noms (dont 2 noms propres).

D'autre part, en nous intéressant aux thèmes de ces mots, nous constatons un hétéroclitisme rendant difficile toute tentative de dégagement de tendances ou d'établissement d'un classement. Arrêtons nous toutefois sur les mots relevant des thématiques majoritaires de la liste des mots ayant subi une diachronie. En effet, il est intéressant de constater par contraste quels mots relevant de la géopolitique, de l'économie ou de la technologie sont restés stables au cour du $\mathbf{XX}^{ème}$ siècle:

- *allemande* : l'empire d'Allemagne existait au début du $\mathbf{XX}^{ème}$ siècle et le mot désignant la nationalité associée (ainsi que les personnes de cette nationalité) est resté stable au cours du siècle.
- *payant* : mot du domaine économique.

- *explosifs* : mot désignant une technologie développée à partir du **XIX**^{ème} siècle, et resté stable depuis.

6.4.5. Discussion

Si la construction de ce jeu de données a mis en évidence les problèmes engendrés par les défauts de précisions des algorithmes successivement utilisés, elle a néanmoins abouti à l'identification d'un groupe de mots, structuré, annoté et validé.

Commençons par nous intéresser à l'aspect statistique de cette partie. Il est évident que le nombre restreint d'exemples obtenus constitue en apparence une déception. Néanmoins il nous semble important de prendre en compte les éléments suivants. D'une part, le jeu de données a été proposé en français, ce qui constitue en soi une contribution. D'autre part, remarquons que même en anglais, le jeu constitué par Hamilton et al. ne contenait que 28 termes ayant tous subi une diachronie, tandis que le dataset que nous proposons contient 59 mots ayant subi une diachronie auxquels s'ajoutent 92 mots stables.

Enfin, revenons sur la constitution du groupe de mots à valider : si tous les mots ayant apparemment subi une diachronie ont été annotés, nous n'avions sélectionné aléatoirement que 150 mots parmi ceux apparemment stables, alors que nous en avions plus de 2000 à notre disposition. Or, si la validation de ces mots a mis en évidence des problèmes exogènes (n'étant pas du ressort des méthodes que nous développons), les mots correctement identifiés, ont obtenu un taux de validation de 90%. Nous pouvons donc légitimement espérer que (une fois éliminés les 30% de contextes abusifs), près de 1300 termes pourront être ajoutés à notre jeu de données en tant que mots stables.

Quant à l'aspect méthodologique, le repérage de mots ayant changé de sens au travers des textes, est un travail particulièrement laborieux à réaliser à la main. Aussi, malgré le taux d'erreurs élevé relevé par les évaluateurs, il faut garder en tête que l'évaluation d'un jeu de données, c'est-à-dire un tri, nécessite moins d'une minute par exemple, abaissant considérablement le temps nécessaire à la construction d'un tel jeu de données.

Enfin, et nous insisterons dessus en conclusions (chapitre 7), si ce mémoire a permis de construire un premier jeu de données, il met en lumière le sujet de l'analyse de diachronie et affiche toutes les possibilités nouvelles d'analyse des sens des mots et de leurs changements au cours du temps, sans avoir d'a priori dessus. C'est donc également un rôle de preuve de concept que joue ce mémoire.

Chapitre 7

Conclusions

Ce mémoire a été l'occasion de présenter les méthodologies essentielles aux études de diachronies. Nous avons pu, dans un premier temps, montrer les étapes classiques d'une telle étude, tout en mettant en évidence leurs limites. Prolongeant les méthodes déjà existantes, nous avons proposé, dans un deuxième temps, de développer des formules aidant à la quantification des diachronies dans l'objectif de les confronter aux résultats d'autres méthodes issues de publications ultérieures. Cette étape décisive nous a permis de mettre en exergue les incertitudes qui demeuraient sur la qualité de celles-ci et de conclure quant au manque que faisait peser l'absence de jeu de données recensant des mots avérés stables ou ayant subi diachronies. Nous avons donc proposé de développer un tel jeu de données dans une dernière partie, ayant abouti à l'identification de 151 mots en français classés selon leur stabilité ou diachronie au cours du $\mathbf{XX}^{ème}$ siècle.

De plus, et comme cela fut notifié tout au long de ce document, nombre de méthodes développées s'appuient sur d'autres algorithmes relevant eux-mêmes de domaines différents du TAL. Les méthodologies que nous présentons resteront pertinentes, néanmoins les équipes s'attelant dans le futur à des tâches dans le domaine de la diachronie devront garder à l'esprit qu'ils pourront substituer les algorithmes que nous avons employés à chaque étape intermédiaire par les nouvelles versions développées dont les performances seront plus élevées. Ceci leur permettrait, grâce à une augmentation de la précision, d'obtenir davantage de données dans les jeux générés.

7.1. Travaux futurs

Les travaux présentés dans ce mémoire constituent une ouverture sur les possibilités et les besoins futurs du domaine de l'identification des diachronies et s'établit dès lors comme le point de départ de plusieurs travaux qui pourront être menés dans l'avenir.

D'autre part, il serait essentiel au cours de futurs travaux de proposer une évaluation précise des méthodes développées par Hamilton et al. Bien que le jeu de données que nous

avons constitué soit en français les méthodologies développées sont translingues (comme indiqué tout au long de ce mémoire). Ces méthodologies pourront donc être déclinées en diverses langues à condition de disposer d'un corpus historique rédigé dans la langue désirée. En particulier, des travaux relançant cette méthodologie afin d'obtenir un jeu de données en anglais rendraient possible l'application des méthodes d'Hamilton et al. à ce jeu et donc d'obtenir un score de précision sur la tâche d'identification des mots ayant subi une diachronie.

Nous espérons en outre que notre méthodologie permettra à d'autres langues d'être l'objet d'études de diachronie.

Au sein du projet d'identification des diachronies non repérées, ces méthodes joueront un rôle moteur. Avec la création de notre jeu de données nous ouvrons la porte à l'évaluation précise des méthodes déjà existantes dans ce domaine. Dans un second temps, leur application à tous les mots de vocabulaire des périodes étudiées permettra de mettre en évidence l'existence de changements de sens non identifiés par les dictionnaires et qui seront ainsi transmis aux lexicographes pour évaluations ultérieures.

Plus largement, bien que nos travaux aient participé au développement d'un jeu de données, celui-ci ne concerne que des évolutions du vocabulaire francophone entre les périodes 1910-20 et 1990-00. En conséquence, cette méthodologie ouvre la voie au développement de toute une série d'autres jeu de donnéess potentiels : en faisant varier les périodes de début et de fin, de nouvelles diachronies pourront être identifiées sur des périodes plus anciennes. Une diminution du délai séparant les périodes étudiées pourra également être envisagé. Cela permettrait de dater avec davantage de précision les phénomènes observés.

Enfin, nous concluons notre document en remarquant qu'une tâche supplémentaire a été traitée par ricochet. En effet, bien qu'il ne s'agisse pas de l'objet de nos recherches, nous avons conjointement développé une méthodologie permettant la datation d'une définition.

Dans le chapitre 6, nous proposions le développement d'un jeu de données, et pour ce faire, nous avons identifié des mots dont des sens apparaissaient entre deux périodes données. Cette méthode de parcours de textes issus de corpus diachroniques afin d'en déduire des usages s'inscrit dans une démarche linguistique classique (études de textes anciens, comparaison. . .) utile à bien des égards au-delà des études diachroniques.

En particulier une telle méthode offre la possibilité de réaliser une datation en utilisant exactement les mêmes algorithmes que nous, mais en menant le processus dans un sens différent de celui que nous avons suivi : Nous avons fixé deux périodes d'intérêt et cherché tous les sens apparus entre celles-ci. Dans le cadre d'une datation, il s'agirait de fixer une définition d'intérêt et chercher les périodes (les plus rapprochées) entre lesquelles apparaît

cette définition, et donc de réutiliser des textes désambiguïsés tels que nous l'avons réalisé, mais issus de nombreuses périodes concomitantes, et de calculer la répartition des sens du seul mot présentant un intérêt.

Références bibliographiques

[1] BANQ : Plateforme erudit. `https://www-erudit-org.res.banq.qc.ca/fr/`, consulté en juin 2020.

[2] Mark DAVIES : Corpus of Historical American English (COHA), 2015.

[3] ATILF / CNRS – Université de LORRAINE. : Trésor de la Langue Française informatisé (TLFi) [ressource en ligne]. `http://atilf.atilf.fr/tlf.htm`, consulté en mars 2021. Place: Nancy.

[4] 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.

[5] Patrick DROUIN : Les voisins du devoir. `http://olst.ling.umontreal.ca/~drouinp/devoir/`, consulté en octobre 2020.

[6] Christiane FELLBAUM, éditeur. *WordNet: An Electronic Lexical Database*. Language, Speech, and Communication. MIT Press, Cambridge, MA, 1998.

[7] GOOGLE : Google ngram viewer. http://books.google.com/ngrams/datasets, 2012.

[8] William L. "HAMILTON, Jure LESKOVEC et Dan" JURAFSKY : "diachronic word embeddings reveal statistical laws of semantic change". *In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1501, Berlin, Germany, août 2016. Association for Computational Linguistics.

[9] Adam JATOWT et Kevin DUH : A framework for analyzing semantic change of words across time. *In Proceedings of the 14th ACM/IEEE-CS Joint Conference on Digital Libraries*, JCDL '14, page 229–238. IEEE Press, 2014.

[10] 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, 2020.

[11] C. MAIR et Geoffrey LEECH : *Current change in English syntax.*, pages 318–342. Blackwell, 2006.

[12] 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. *In ACL 2020 - 58th Annual Meeting of the Association for Computational Linguistics*, Seattle / Virtual, United States, juillet 2020.

[13] Tomas MIKOLOV, Kai CHEN, Greg CORRADO et Jeffrey DEAN : Efficient estimation of word representations in vector space, 2013.

[14] George A. MILLER, Claudia LEACOCK, Randee TENGI et Ross T. BUNKER : A semantic concordance. *In Human Language Technology: Proceedings of a Workshop Held at Plainsboro, New Jersey, March 21-24, 1993*, 1993.

[15] Radim Řehůřek et Petr Sojka : Software Framework for Topic Modelling with Large Corpora. *In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, mai 2010. ELRA. `http://is.muni.cz/publication/884893/en`.

[16] John Andrew Simpson, Edmund SC Weiner et al. : The Oxford English Dictionary, volume 2, 1989.

[17] 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, 2019.

[18] Loïc Vial, Benjamin Lecouteux et Didier Schwab : UFSAC: Unification of Sense Annotated Corpora and Tools. *In* Nicoletta Calzolari (Conference chair), Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis et Takenobu Tokunaga, éditeurs : *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 7-12, 2018 2018. European Language Resources Association (ELRA).

[19] 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.

[20] Derry Tanti Wijaya et Reyyan Yeniterzi : Understanding semantic change of words over centuries. *In Proceedings of the 2011 International Workshop on DETecting and Exploiting Cultural DiversiTy on the Social Web*, DETECT '11, page 35–40, New York, NY, USA, 2011. Association for Computing Machinery.

[21] 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.

Annexe A

Algorithmes

A.1. Algorithme de récupération des plus proches voisins

Algorithme A.1.1. Pseudo-code décrivant la méthode de récupération des plus proches voisins d'un mot dans un modèle Word2Vec

---

fonction *get_dictionnaire_distances*(**modèle**, **nombre_voisins**, **mot**)**:**

"'

: *param* : *modèle* : *modèle Word2Vec utilisé*

: *param* : *nombre_voisins* : *nombre de voisins plus proches* à *rechercher*

: *param* : *mot* : *mot dont on calcule le score*

"'

dictionnaire_distances ← *{} # taille finale nombre_voisins*

vecteur_mot ← *modèle [mot]*

POUR chaque mot de vocabulaire *mot_compar* **du modèle :**

vecteur_compar ← *modèle* [*mot_compar*]

distance ← *cosine_similarité*(*vecteur_compar*, *vecteur_mot*)

Ajouter "*mot_compar*:distance" à **dictionnaire_distances**

Supprimer argmin(**dictionnaire_distances**)

A.2. Détail de l'algorithme de calcul du score (S_b)

Algorithme A.2.1. Pseudo-code décrivant le calcul du Score-base d'un mot entre deux périodes

---

mot $\#$ *choix utilisateur*
vecteur_mot_1 $\leftarrow$ **modèle_1**[**mot**]
vecteur_mot_2 $\leftarrow$ **modèle_2**[**mot**]
modèle_1 $\leftarrow$ $modèle[période\_1]$
modèle_2 $\leftarrow$ $modèle[période\_2]$
dictionnaire_distances_1 $\leftarrow$
    get_dictionnaire_distances($modèle\_1, nombre\_voisins, mot$)
dictionnaire_distances_2 $\leftarrow$
    get_dictionnaire_distances($modèle\_2, nombre\_voisins, mot$)
distance_totale $\leftarrow 0$
POUR chaque voisin v *dans* **dictionnaire_distances_1:**
    vecteur_voisin_2 $\leftarrow$ **modèle_2**[v]
    distance_1 $\leftarrow$ **dictionnaire_distances_1**[v]
    distance_2 $\leftarrow$ $cosine\_similarité$(**vecteur_mot_2**, **vecteur_voisin_2**)
    difference_distance $\leftarrow$ abs(**dictionnaire_distances_2** $-$ **dictionnaire_distances_1**)
    distance_totale$+=$ **difference_distance**
POUR chaque voisin v *dans* **dictionnaire_distances_2:**
    SI v $n'est$ **PAS** *dans* **dictionnaire_distances_1** :
        vecteur_voisin_1 $\leftarrow$ **modèle_1**[v]
        distance_2 $\leftarrow$ **dictionnaire_distances_2**[v]
        distance_1 $\leftarrow$ $cosine\_similarité$(**vecteur_mot_1**, **vecteur_voisin_1**)
        difference_distance $\leftarrow$
            abs(**dictionnaire_distances_2** $-$ **dictionnaire_distances_1**)
        distance_totale $\leftarrow$ **distance_totale** $+$ **difference_distance**

A.3. Détail de l'algorithme de calcul de Link (L)

Algorithme A.3.1. Pseudo-code décrivant le calcul du Link d'un mot entre deux périodes

---

mot $\#$ *choix utilisateur*
dictionnaire_modèles
 # choix utilisateur associe une date au modèle Word2Vec correspondant
$\textbf{link}_{total} \leftarrow 0$
POUR *chaque* **période p** *dans* **dictionnaire_modèles:**
    modèle_période $=\leftarrow$ **dictionnaire_modèles**$[p]$
    voisins $\leftarrow$ **get_dictionnaire_distances**$(modèle\_période, nombre\_voisins, mot)$
    POUR *chaque* **période** $\textbf{p}_n > \textbf{p}$ *dans* **dictionnaire_modèles:**
        ecart $\leftarrow \textbf{p}_n - \textbf{p}$
        $\textbf{link}_n \leftarrow \textbf{0}$
        modèle_période $\leftarrow$ **dictionnaire_modèles**$[p]$
        $\textbf{voisins}_n \leftarrow$ **get_dictionnaire_distances**$(modèle\_période, nombre\_voisins, mot)$
        POUR *chaque* **voisin** $\textbf{v}_n$ *dans* $\textbf{voisins}_n$:
            SI $\textbf{voisins}_n$ *dans***voisins** :
                $\textbf{link}_n \leftarrow \textbf{link}_n + 1 \times$ **ecart**
        $\textbf{link}_{total} \leftarrow \textbf{link}_{total} + \textbf{link}_n$

A.4. Algorithme d'unification des sens d'un mot dans un corpus

Algorithme A.4.1. Pseudo-code décrivant la méthode d'unification des sens d'un mot dans un corpus

---

dictionnaire_listes_mots = {}
Pour chaque tranche du corpus :
    liste_mots = [ ]
    Pour chaque mot du corpus :
        Si mot n'est pas dans *liste_mots* **:**
            Ajouter mot à *liste_mots*
    dictionnaire_listes_mots[**tranche**] = *liste_mots*

dictionnaire_dictionnaires_sens_mot = {}
Pour chaque tranche du corpus :
    dictionnaire_sens = { }
    liste_mots =*dictionnaire_listes_mots*[**tranche**]
    Pour chaque mot du corpus :
        Si mot n'est pas dans *dictionnaire_sens* **:**
            Ajouter clé:mot, valeur : { } à *dictionnaire_sens*
        Sens = disamb(mot)
        Si sens n'est pas dans *dictionnaire_sens*[**mot**]**:**
            Ajouter clé:sens, valeur : 1 à *dictionnaire_sens*[**mot**]
        Sinon:
            dictionnaire_sens[**mot**][**sens**] += **1**
    dictionnaire_dictionnaires_sens_mot [**tranche**] = *dictionnaire_sens*

Annexe B

Constats préliminaires d'introduction des mesures

B.1. Analyse de l'évolution de voisinage d'un mot

Nous nous intéressons dans cette section aux enjeux dus aux modifications du voisinage d'un mot.

Les articles évoqués en introduction (chapitre 1) nous permettent de proposer la définition suivante: "les changements de sens d'un mot s'accompagnent de variations de son voisinage".

Afin d'etayer empiriquement cette hypothèse, nous avons parcouru le voisinage de voisinage de nombreux exemples.

Nous nous proposons ici d'en considérer deux, en nous concentrant sur le rapport entre d'une part la distance séparant le plongement de mot de m de celle de ses plus proches voisins et d'autre part les changements de sens de m Pour ce faire, nous analysons les mouvements subis par le voisinage d'un mot m entre deux périodes : nous observons donc les voisins de la période p_1 et ceux de la période p_2.

Cette analyse de deux mots issus du lexique animalier met en exergue la différence de l'évolution du voisinage entre un mot ayant subi un changement de sens ("souris") et un autre stable ("vache") : Les 5 voisins les plus proches de "vache" en 1990 appartiennent tous au registre de l'agriculture : dinde, foie... Les 5 étaient déjà inclus dans le vocabulaire, et leur distance à "vache" est restée dans le même ordre de grandeur. A l'inverse, parmi les 5 plus proches voisins du mot "souris", 4 relèvent du domaine de la technologie, et 4 n'étaient pas même inclus dans le vocabulaire utilisé en 1920. Enfin le seul mot à apparaitre en 1920 (clavier), désignait un objet sans lien avec la souris, et se retrouve donc relégué plus loin de souris qu'en 1990.

Cette comparaison des voisinages nous a permis de vérifier l'hypothèse de départ. Deux caractéristiques sont mises en évidence :

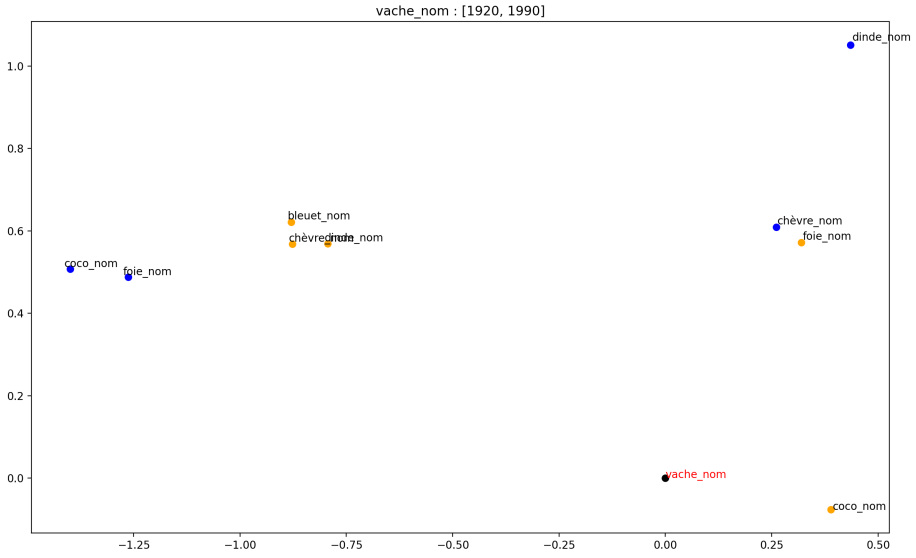


Fig. B.1. Exemple de l'évolution des distances entre le mot "vache" et ses 5 plus proches voisins en 1990 : en bleu sont représentées leur position en 1920 et en orange en 1990. Les mots n'apparaissant qu'une seule fois étaient absents en 1920. La projection est centrée et réduite sur la distance au mot "vache".

- Comme annoncé concernant la variation des n voisins les plus proches entre p_1 et p_2, nous remarquons une constance des voisins les plus proches implique une stablilité sémantique.
- De plus, nous constatons que la distance parcourue par les voisins non communs entre p_1 et p_2 est porteuse d'information : plus les anciens voisins sont éloignés de la position en p_2 et les nouveaux de la position en p_1 plus les contextes d'utilisation du mot m ont évolué.

B.1.1. Analyse de voisinage

Ces analyses qualitatives nous a permis de remarquer que le changement de sens d'un mot s'accompagne non seulement d'un changement de l'environnement proche, mais également d'une modification sensible des distances aux anciens voisins. En considérant l'espace construit par le modèle de langue comme un espace sémantique, nous pouvons traduire ce phénomène de déplacement du mot m dans cet espace, par un éloignement du lexique propre à l'ancien emploi du mot et donc de la zone de l'espace associée. D'autre part, les modifications de sens du mot se traduisent par un déplacement vers une nouvelle zone sémantique correspondant pour sa part au nouveau sens d'utilisation de m. Ces remarques nous ont

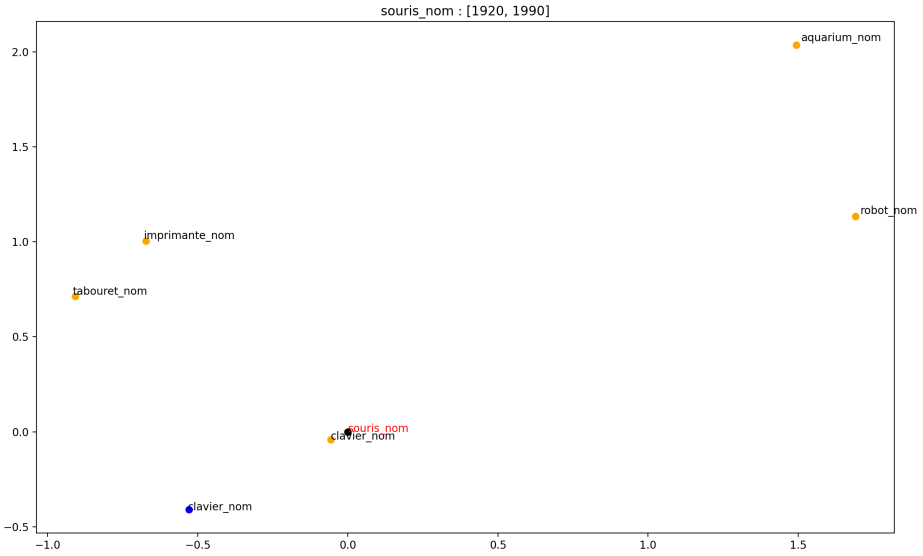


Fig. B.2. Exemple de l'évolution des distances entre le mot souris et ses 5 plus proches voisins en 1990 : en bleu sont représentées leur position en 1920 et en orange en 1990. La projection est centrée sur le mot souris, et les positions mutipliées par la distance dans l'espace initial.

Fig. B.3. Exemple de la représentation des 6 plus proches voisins du mot m durant la première période étudiée

Fig. B.4. Exemple de la représentation des 6 plus proches voisins du mot m durant la seconde période étudiée : les voisins en vert n'étaient pas présents durant la première période. Nous remarquons qu'en conséquence certains anciens voisins plus proches sont sortis du voisinage.

amenés à définir la mesure nommée "Score" qui proposer de quantifier le changement de sens d'un mot en s'appuyant sur les différentiels de distances des voisins les plus proches.

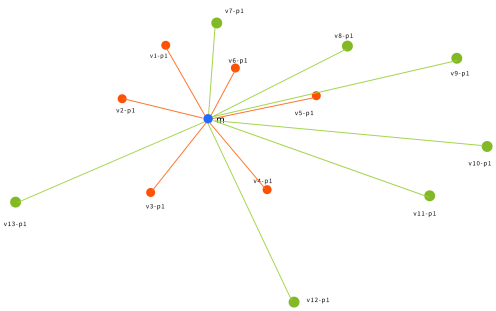


Fig. B.5. Exemple des 12 plus proches voisins d'un mot m durant une première période. Les 6 plus proches sont en orange, les 6 suivants sont en verts. Les lignes représentent la distance à m.

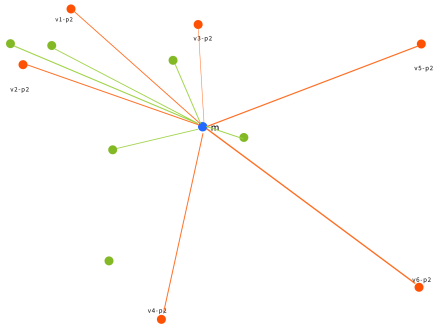


Fig. B.6. Exemple des 12 voisins de la figure B.5 durant la période. Nous remarquons que les 6 plus proches voisins de la première période s'éloignent de m et les 6 autres s'approchentde m.

B.2. Analyse de graphes panoptiques

Nous nous proposons ici d'analyser les voisins de deux mots : "souris" (voir figure B.8) mot subissant une diachronie dans les années 1960 avec l'arrivée du vocabulaire informatique et "vache" ou "france" (voir figure 2.2) mots stables entre 1920 et 2000.

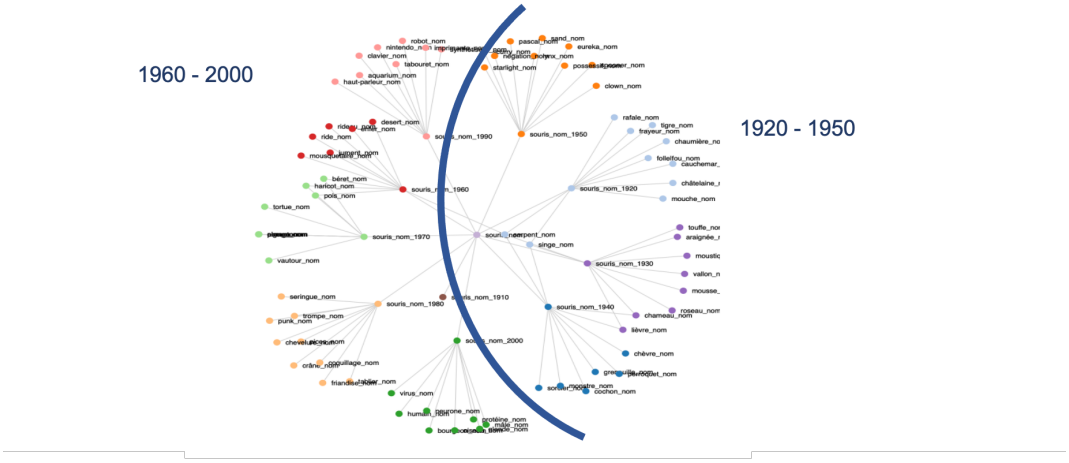


Fig. B.7. Analyse des 10 voisins les plus proches du mot "souris" entre 1920 et 2000. L'arc de cercle bleu marque la rupture entre les périodes 1920-1950 et 1960-2000 qui ne présentent aucun voisins en commun.

La mise en perspective des exemples précédents permet de confirmer les observations précédentes :

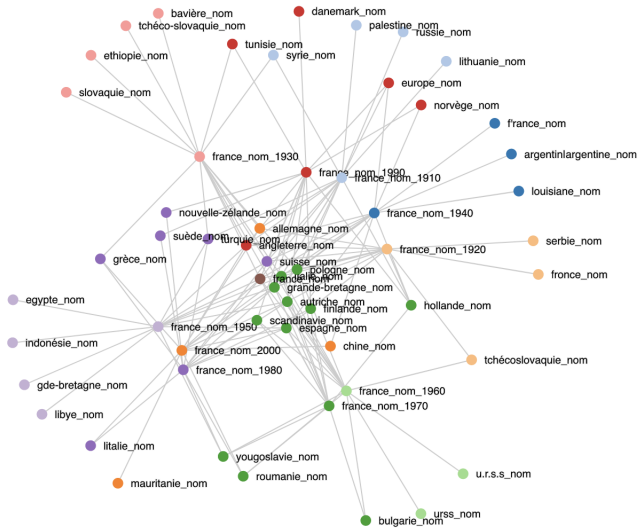


Fig. B.8. Analyse des 10 voisins les plus proches du mot "France" entre 1920 et 2000. L'arc de cercle bleu marque la rupture entre les périodes 1920-1950 et 1960-2000 qui ne présentent aucun voisins en commun.

- Le graphique du mot "vache" présente des liens entre toutes les périodes du corpus; ainsi le mot "animal" est à la fois plus proche voisins en 1920 et en 2000. A l'inverse, le graphique du mot "souris" fait apparaitre une rupture entre 1950 et 1960 : aucun voisin n'est commun entre les périodes antérieures et ultérieures.
- Les mots présents de façon concomitante entre deux périodes indiquent une stabilité sémantique, ainsi le voisin "singe" commun à "souris" entre 1920 et 1940 indique bien la stabilité de la définition "mammifère" de ce mot, et il en va de même pour "bouvillon" commun à "vache" entre 1910 et 1990.

Annexe C

Détails de la désambiguïsation

C.1. Entrainement de l'algorithme de désambiguïsation

Une analyse a été menée pour vérifier quels indicateurs de confiance pouvaient être= envisagés afin d'établir un tri en amont entre les mots. Nous cherchons alors à ne plus désambiguiser tous les mots d'un texte, mais uniquement ceux pour lesquels nous obtenons un niveau de confiance assez élevés.

Nous proposons donc une méthode de tri en sortie de la couche linéaire, qui décidera si la confiance accordée à une désambiguïsation est assez grande pour être acceptable ; dans le cas inverse, cette désambiguïsation sera omise. Pour ce faire, nous nous appuierons sur la précision, en n'évaluant que les évaluations réalisées, c'est-à-dire ayant un niveau de confiance assez élevé.

La figure propose une analyse des softmax calculés sur la couche linéaire en sortie du classifieur : pour chaque mot évalué l'algorithme est appliqué, et propose pour chaque synset un score ; le softmax est calculé grâce au rapport suivant (C.1.1):

$$\text{Softmax}(x_i) = \frac{\exp(x_i)}{\sum_j \exp(x_j)} \tag{C.1.1}$$

Avec:

- x_i le score le plus important des synsets proposés en sortie, c'est-à-dire celui du synset proposé par l'algorithme comme sens associé au token considéré.
- x_j les scores de tous les synsets proposés en sortie

On peut voir en orange les synsets mal associés, et en bleu les synsets bien associés. Il apparait clairement que les softmax les plus faibles sont source majeure d'erreur.

Étant donnée l'incertitude des désambiguïsations proposées par les softmax les plus bas, nous proposons d'instaurer un cut-off : seuls les mots pour lesquels le softmax est supérieur à un certain seuil est désambiguisé, aux autres est associé le token "tooLow". D'après les résultats précédents, nous fixons ce seuil à un softmax supérieur ou égal à 0.5.

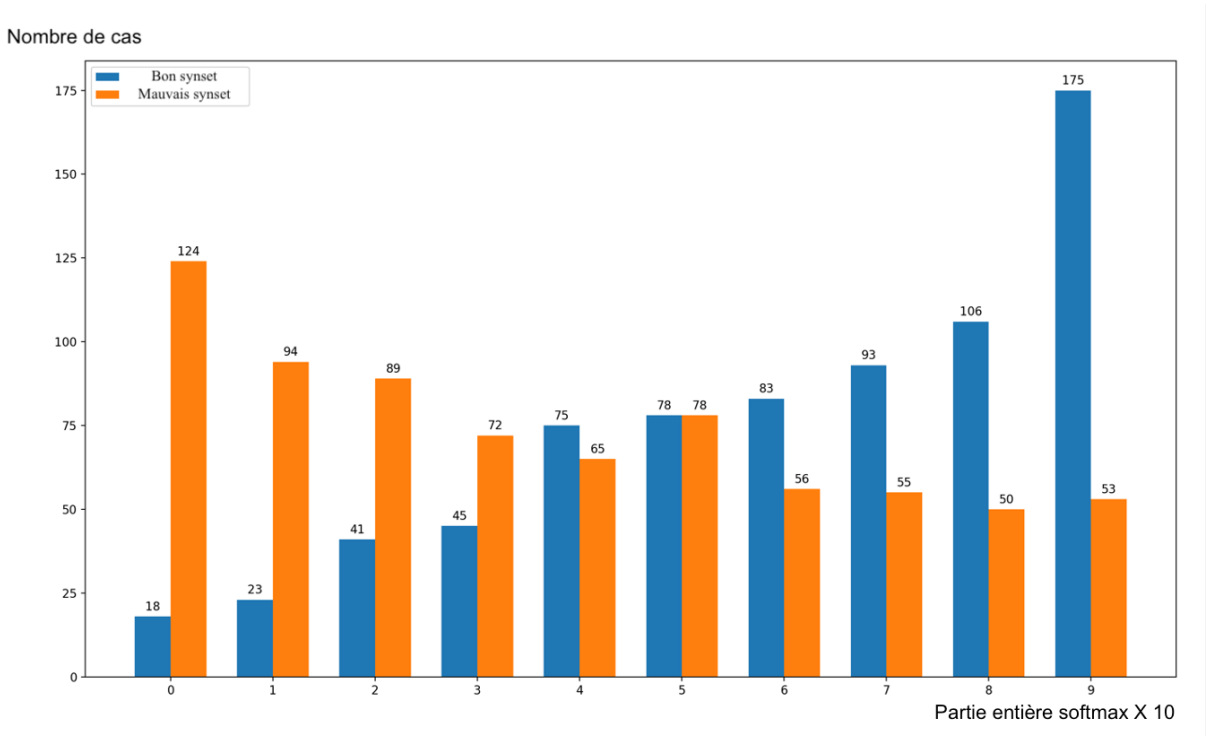


Fig. C.1. Comparaison des valeurs des softmax (multipliés par 10) des synsets proposés par l'algorithme. Les taux des synsets bien associés sont en bleu, et ceux des synsets mal associés en rouge.

Reprenons le texte présenté et déjà désambiguisé en section 2.3.2 et appliquons lui la nouvelle version de l'algorithme. On obtient la sortie suivante :

Déjà	already%4:02:00
aux	tooLow
prises	tooLow
avec	tooLow
des	not%4:02:00
difficultés	tooLow
financières	fiscal%3:01:00::
considérables	considerable%3:00:00::
les	tooLow
directions	tooLow
des	tooLow
universités	university%1:14:00::

Nous remarquons que sur les 12 tokens du texte, seuls 5 conservent leur synset. Les autres ne présentent pas un niveau de confiance assez élevé (softmax inférieur à 0.5) pour que l'algorithme leur propose une association avec un synset.

C.2. Tri du vocabulaire désambiguïsé

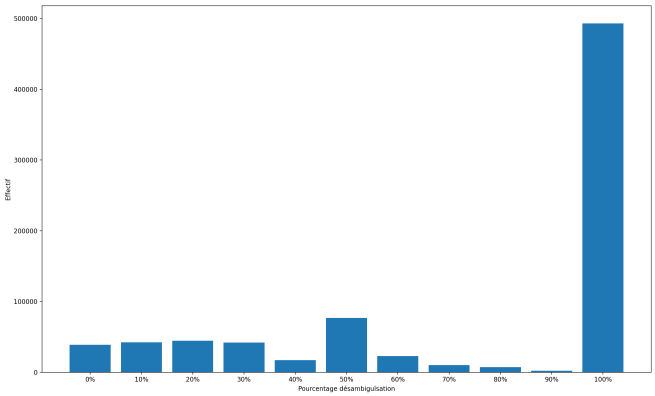


Fig. C.2. 1910-20 : pourcentage désambïguation

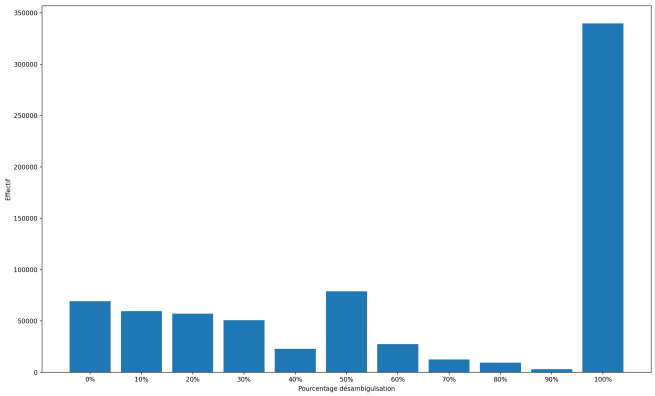


Fig. C.3. 1990-00 : pourcentage désambïguation

Les figures C.2 et C.3 montrent une proportion très élevée de mots ayant 100% d'occurences désambiguïsées. Nous proposons donc d'opposer ce pourcentage au nombre d'occurences des mots.

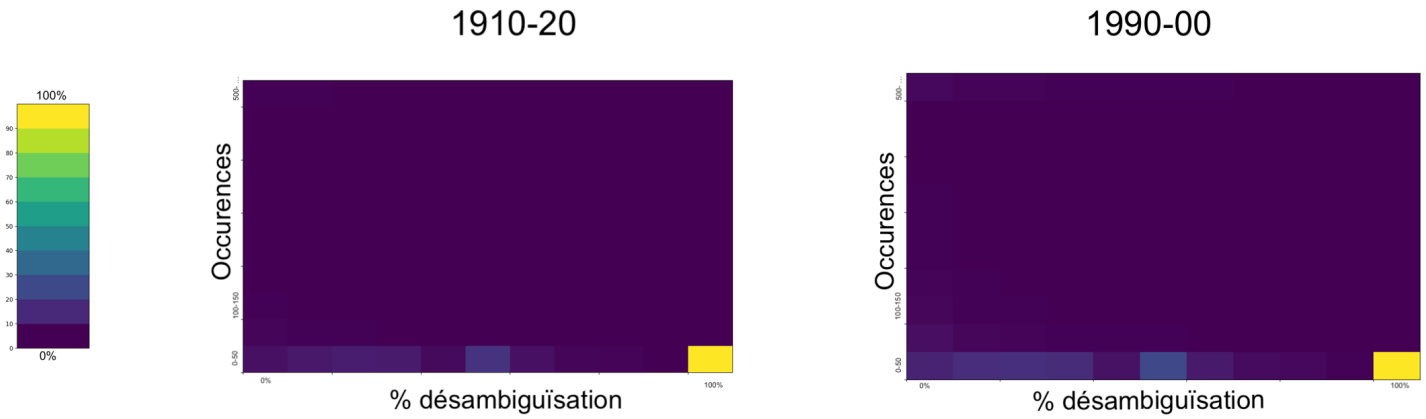


Fig. C.4. Répartition des désambiguisations du vocabulaire. Les cases représentent l'ensemble des mots dont le nombre d'occurences est inclus dans la tranche représentée sur l'axe des ordonnées, et le pourcentage de désambiguïsation sur l'axe des abscisses. La couleur représente le poucentage de l'effectif total contenu dans une case selon l'échelle affichée à gauche. Les deux périodes représentées affichent une case jaune dans la case correspondant à une désambiguïsation de 100% et un nombre d'occurences compris entre 0 et 50.

Les figures C.4 sont l'occasion de vérifier les caractéristiques des mots de vocabulaires désambiguïsés : nous avons ainsi décidé de séparer en catégories selon le nombre d'occurences

d'un mot dans la tranche et le pourcentage de désambiguïsation associé. Puis, nous affichons le pourcentage de mots correpondants à chaque catégorie, par rapport au nombre total de mots de vocabulaires disponibles.

Ces figures montrent toutes deux une case colorée en jaune regroupant les mots ayant 50 ou moins occurences et un pourcentage de désambiguïsation de 100%. Ceci indique que la grande majorité des mots désambiguïsés ont de telles caractéristiques. Or les problèmes liés à l'OCR ne permettent pas de considérer ces mots comme fiables. Afin de ne plus considérer les mots mal identifiés, nous ne considérons désormais plus que les mots ayant été désambiguisés au moins 200 fois.

Annexe D

Détail des mots du jeu de données

Les tableaux D.1 et D.2 listent respectivement les mots ayant subi une diachronie et stables entre les périodes 1910-20 et 1990-00 obtenus après l'étape de validation.

n°	Mot	Sens apparu
1	stations	radio station
2	souverain	autonomous
3	carré	square meter
4	crédits	semester hour
5	enregistrer	record
6	séparation	secede
7	charme	fashionable
8	allée	go
9	suites	suite
10	tombée	down
11	ordonne	authorize
12	échéance	run out
13	direct	live
14	carte	credit card
15	sommet	eighth
16	consultations	negotiation
17	couverture	coverage
18	manches	tournament
19	crime	organized crime
20	occupés	district
21	secteur	sector
22	producteur	producer
23	humaines	working
24	union	soviet union
25	pêches	fishing
26	variétés	television
27	garde	childcare
28	armes	nuclear weapon
29	émission	television program
30	station	radio station
31	placement	pass
32	haute	high technology
33	climat	manner
34	silencieux	sound
35	bande	gaza strip
36	textile	fabric
37	courrier	electronic mail
38	canaux	channel
39	émissions	television program
40	tournant	turn of the century
41	froide	cold war
42	vols	spaceflight
43	tissu	framework
44	enlevés	kidnap
45	discussions	discussion
46	l'union	soviet union
47	organisme	agency
48	lanceurs	pitcher
49	subvention	grant
50	avantageux	advantageous
51	fournisseurs	supplier
52	congo	congo
53	turquie	turkey
54	remercié	thank
55	volumes	bulk
56	occidentales	western
57	sûrs	straight
58	crédit	credit card
59	répondait	react

Tableau D.1. Liste des mots ayant subi une diachronie sémantique entre les périodes 1910-20 et 1990-00

n°	Mot
1	allemande
2	persécutions
3	bleue
4	fixant
5	urgents
6	conversation
7	musiciens
8	affectent
9	centre
10	proclamée
11	efficacement
12	payant
13	dangereuses
14	tactique
15	munich
16	extrémité
17	volés
18	forteresse
19	crimes
20	appel
21	recevront
22	échantillons
23	construire
24	l'été
25	veut
26	efforts
27	avec
28	adultes
29	debout
30	existé
31	travers
32	lions
33	sites
34	régiment
35	elle
36	regardant
37	soirée
38	commission
39	adoption
40	combattant
41	savez
42	flammes
43	tsar
44	d'achat
45	exigent
46	londres

n°	Mot
47	raconte
48	fonda
49	prétendent
50	couvertures
51	endommagé
52	vins
53	tante
54	mécontentement
55	intitulé
56	bruns
57	pages
58	comptabilité
59	solidité
60	explosifs
61	patience
62	s'attendre
63	inconnus
64	solennellement
65	pèlerinage
66	pauvreté
67	phase
68	principal
69	reçoivent
70	établie
71	rein
72	principale
73	révélations
74	merveilleux
75	requête
76	comtés
77	passager
78	soie
79	mélancolie
80	centaines
81	doublé
82	noire
83	révolte
84	artisans
85	maintenir
86	met
87	lue
88	annoncer
89	énergique
90	quarts
91	proprement
92	sénateurs

Tableau D.2. Liste des mots sémantiquement stables entre les périodes 1910-20 et 1990-00