

THÈSE DE DOCTORAT

pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ DE LILLE

dans la spécialité « INFORMATIQUE ET APPLICATIONS »

par

Bastien Liétard

Concept-Aware Computational Models of Lexical Semantic Change

Modèles computationnels du changement lexico-sémantique par
reconnaissance des concepts

Thèse soutenue le 15 janvier 2026 devant le jury composé de :

Rapporteurs

<i>M.</i>	MATHIEU CONSTANT	Professeur	ATILF, Université de Lorraine
<i>M.</i>	YANG XU	Associate Professor	University of Toronto

Examineur·ices

<i>Mme.</i>	KATRIN ERK	Professor	University of Massachusetts Amherst
<i>M.</i>	KRIS HEYLEN	Researcher	Dutch Language Institute

Directeur·ices de thèse

<i>Mme.</i>	ANNE CARLIER	Professeure	STIH, Sorbonne-Université
<i>M.</i>	MARC TOMMASI	Professeur	CRIStAL, Université de Lille

Encadrant·e de thèse

<i>M.</i>	PASCAL DENIS	Chargé de Recherche	Inria, Université de Lille
<i>Mme.</i>	MIKAELA KELLER	Maîtresse de Conférences	CRIStAL, Université de Lille

Président du jury : MATHIEU CONSTANT

Centre Inria de l'Université de Lille,

Centre de Recherche en Informatique, Signal et Automatique de Lille (CRIStAL)

UMR 9189 Équipe Magnet, 59650, Villeneuve d'Ascq, France

Abstract

Lexical semantics, the study of words and their meanings, can be approached from a diachronic perspective by analyzing how conventional word-meaning associations emerge or disappear. These changes over time are referred to as Lexical Semantic Change, about which can be examined two complementary perspectives. The semasiological approach starts with a given word and describes its various meanings over time. Conversely, the onomasiological approach begins with a specific concept and examines the words language users employ to express it. Research in Natural Language Processing (NLP) has developed models of Lexical Semantic Change and has proposed standard evaluations for these methods. However, this line of research has focused almost exclusively on detecting and quantifying semasiological changes, with little attention to the characterization of change and the study of its onomasiological aspects.

The central objective of this thesis is to propose NLP models to characterize onomasiological as well as semasiological changes. Specifically, we aim to leverage word embeddings to extract semantic concepts and their associated words, and to study the evolution of word-meaning mappings in historical textual data. Starting from a synchronic perspective, we investigate how semantic concepts can be extracted from Contextualized Language Models representations. We develop a framework, called Concept Induction, which evaluates a system’s ability to represent word-meaning associations by learning a soft-clustering of words into synonym groups that instantiate a common concept directly from the data. We propose a method based on a bilevel clustering of the word occurrences to extract these concept-clusters from representations of a Contextualized Language Models. Subsequently we explore strategies to improve the encoding of lexical semantic information in these representations. Concept Induction is reformulated as a binary classification task over pairs of word usages, and contrastive learning is used to fine-tune a series of multilingual representation models to perform this task effectively. Evaluation on a variety of lexical semantic tasks demonstrates that the obtained models, called Concept-Aligned Embeddings (CALE), better encode lexical meaning and conceptual information than pre-trained models and state-of-the-art word representations. Moreover, the CALE models exhibit interesting spatial properties for the study of semantic concepts in representations.

In the diachronic setting, we present an adaptation of our bilevel clustering method to learn concepts from historical data. The original synchronic approach is modified to learn concepts across several time periods and without concept-annotated data. We apply this Diachronic Concept Induction method to a large historical corpus of French using representations from a CALE model and conduct a qualitative evaluation of the obtained clusters and their evolution over time. This methodology identifies both onomasiological and semasiological changes, enabling the study of lexical semantic change from both perspectives. Next, we examine the lexical semantic change of synonyms as an example of onomasiological changes. Building on prior work investigating two historical linguistic laws that predict different evolution for synonyms, we define an evaluation framework for computational approaches to these changes. We show that the proposed model, close to classical NLP models of lexical semantic change, is inherently limited in characterizing change in synonyms because it is restricted to a semasiological description of the problem. Finally, we demonstrate that our Diachronic Concept Induction method, when applied to historical data, allows for a more comprehensive description of the semantic change of synonyms and reveals that the seemingly contradictory laws could actually operate simultaneously within synonym groups.

Résumé

La sémantique lexicale, qui étudie les mots et leur sens, peut être abordée sous un angle diachronique en analysant comment les associations conventionnelles entre les mots et leur signification apparaissent ou disparaissent. Ces changements sont appelés « changements lexico-sémantiques », et peuvent être étudiés dans deux perspectives complémentaires. L’approche sémasiologique part d’un mot donné et décrit ses différentes significations au fil du temps. À l’inverse, l’approche onomasiologique part d’un concept spécifique et examine les expressions lexicales que les locuteurs utilisent pour l’exprimer. La recherche en Traitement Automatique des Langues (TAL) a développé des modèles du changement lexico-sémantique et a proposé des méthodes d’évaluation standards pour ces modèles. Cependant, cette ligne de recherche s’est concentrée presque exclusivement sur la détection et la quantification des changements sémasiologiques, sans prêter beaucoup d’attention à la caractérisation du changement et à l’étude de ses aspects onomasiologiques.

L’objectif principal de cette thèse est de proposer des modèles de TAL pour caractériser les changements onomasiologiques et sémasiologiques. Plus précisément, nous utilisons les plongements lexicaux pour extraire des concepts sémantiques et les mots qui leur sont associés, et pour étudier l’évolution de ces associations dans des textes historiques. Dans un cadre synchronique d’abord, nous étudions l’extraction de concepts sémantiques à partir de représentations de modèles de langue contextualisés. Nous développons un cadre d’évaluation, appelé « Concept Induction », pour mesurer la capacité d’un système à représenter les associations mots-sens en apprenant un regroupement des mots en groupes de synonymes instanciant un concept commun dans les données. Nous proposons une méthode basée sur un partitionnement à deux niveaux des occurrences de mots afin d’extraire ces groupes. Nous explorons ensuite des moyens de mieux représenter les informations lexico-sémantiques avec des modèles de langue. Nous affinons une série de modèles de représentation multilingues pour un nouvel objectif de sémantique lexicale. L’évaluation démontre que les modèles obtenus, appelés « Concept-Aligned Embeddings » (CALE), encodent mieux les informations conceptuelles que les modèles pré-entraînés et d’autres modèles de pointe, et qu’ils présentent des propriétés spatiales intéressantes pour l’étude des concepts.

Dans un cadre diachronique, nous adaptons notre regroupement à deux niveaux pour apprendre des concepts dans des données historiques. L'approche synchronique initiale est modifiée afin d'apprendre des concepts sur plusieurs périodes et sans données annotées. Nous appliquons cette méthode d'induction diachronique de concepts à un vaste corpus historique français et procédons à une évaluation qualitative des clusters obtenus et de leur évolution dans le temps. Nous démontrons que cette méthode permet d'identifier à la fois les changements onomasiologiques et sémasiologiques et de décrire les changements sémantiques lexicaux sous ces deux angles. Nous examinons ensuite les changements lexico-sémantiques des synonymes comme exemple de changements onomasiologiques. En nous appuyant sur une étude computationnelle de deux lois linguistiques qui prédisent une évolution différente pour les synonymes, nous définissons un cadre d'évaluation pour les approches computationnelles de ces changements. Nous montrons que le modèle proposé est intrinsèquement limité dans la caractérisation du changement des synonymes, car il se restreint à une description sémasiologique du problème. Enfin, nous démontrons que notre méthode d'induction diachronique des concepts, lorsqu'elle est appliquée à des données historiques, permet de décrire plus précisément le changement sémantique des synonymes et révèle que les lois apparemment contradictoires peuvent en fait s'appliquer simultanément à des groupes de synonymes.

Acknowledgements

Remerciements

Before we jump into the heart of the work, there are several people to whom I would like to express my gratitude.

To my supervisors. First, I want to thank Anne, Marc, Mikaela and Pascal, for their advice during my PhD. Pascal and Mikaela, thank you for the day-to-day supervision. You helped me grow as a researcher, and with your support I gained in autonomy, in scientific rigor and other skills that are essential on the path to become an accomplished researcher. I really appreciated our many discussions and your help in developing my ideas and pushing myself to my best. Thank you also for your kindness and your support over the past three years, especially given that self-confidence is not my forte. Anne, I am grateful for these enlightening discussions we had. You brought a highly valuable perspective to my work, and I can't thank you enough for your thorough feedback while I was writing this thesis. Marc, I am thankful for the your wise advice. Although you tend to minimize your impact on this PhD. due your more distant supervisory role, you played an important part in my academic growth and you earn my most sincere gratitude for that.

To my jury. Many thanks to Katrin Erk, Kris Heylen, Mathieu Constant and Yang Xu for agreeing to be on my jury and for dedicating their time to this mission; reading my thesis, attending my defense and providing these insightful comments about my work. You are all truly inspiring researchers and it has been a real pleasure to exchange with you on the day of the defense. Special thanks to Mathieu and Yang, for agreeing to be my *rapporteurs* and especially since you had to review my manuscript right before the holiday; to Kris, for proofreading my manuscript so thoroughly that you spotted the few remaining typos, which, as you pointed out on the day of the defense, are almost a good sign in the 'Generative AI' era; and to Katrin, for following my research journey since my first PhD. year and for providing valuable feedback over the years.

To my colleagues. A healthy work environment is a key ingredient to a successful and enjoyable PhD experience. Therefore I am grateful for my colleagues

in the Magnet team. I feel fortunate to have come to the lab every week knowing that our discussions and time spent together, at lunch or during coffee breaks, would bring me joy, motivation and laughter. I would particularly like to thank all of the volunteer bakers who made cakes to share with the team for the *Jeudi Gâteau* over these three years. Besides this weekly meeting, thank you for game nights, drinks in bars, conferences we organized, and other moments we shared. I would like to express my gratitude again to Marc for being such a dedicated and benevolent team leader. A huge thank to Michaël, our discussion on clustering and other topics significantly impacted my research and I benefited greatly from your expertise. Thank you to all of those who shared my office for an extended period – most notably Gaurav, Batiste, Viet-Toan and Gabriel – as well as those who shared it for a shorter period.

To people I met through academia. I am grateful for the people I met at seminars, conferences, and other scientific events. Particularly, thank you to my fellow members of the local organizing team for the 2025 LLcD conference; as well as and to people I met in *ACL conferences, especially to people involved in the LChange workshop. Thank you Aina Garí Soler, although now you would best fall into the next category.

To my friends. Part of the success of a PhD actually comes from outside of the PhD itself and that's where friends and loved ones play a crucial role. I feel so lucky to have the friends I have today, who listen to me no matter the topic and to support me in my endeavors. Whether you have been there for me in Lille, or other cities in France (and sometimes outside of France), or mostly online; whether I have known you since high school, during my bachelor's or master's degree or more recently, I thank you. My most sincere gratitude to those of you who attended my defense – and those who couldn't be there that day – despite not knowing what I was working on, at least for most of you. However I won't take the risk of trying to name everyone, I would likely forget someone important and feel guilty for the rest of my life. But you will probably recognize yourself in one of the previous lines. Your unconditional support over past three years has been valuable beyond words, thank you for it.

And now let me switch to French for the remainder of the acknowledgements, or more exactly: *les remerciements*.

À ma famille. Merci à ma famille d'avoir été mes premier·e·s supportrices et supporters dès le moment où j'ai commencé ma thèse, et même bien avant pour être honnête. Merci de m'avoir toujours offert une parenthèse bienveillante,

à l'écart de mon quotidien lillois, d'avoir toujours cru en mes capacités même lorsque (et surtout lorsque) j'en doutais, et de vous être réjouis avec moi à chaque étape. Je m'estime chanceux de vous avoir.

À mes chats. Il peut paraître étrange de remercier quelqu'un en sachant pertinemment que cette personne ne pourra pas lire, mais je ne peux pas sincèrement rédiger ces paragraphes sans un mot pour Hylia et Percy (*les meufs* comme je les appelle familièrement). Vous êtes toutes deux des formidables compagnons de vie qui ont sans cesse rempli mon cœur de joie, d'émerveillement et de sérénité au travers des multiples étapes qui composent le parcours d'une thèse.

À Honorine. Merci de partager ma vie; de m'avoir toujours écouté que ce soit pour me plaindre, me réjouir, m'émerveiller d'un détail ridiculement insignifiant ou déblatérer pendant trop longtemps sur un sujet trop niche. Merci de m'avoir toujours encouragé et soutenu, notamment lors de la rédaction de cette thèse. Plus généralement, merci de me rendre chaque journée un peu plus facile.

Contents

Abstract	iii
Résumé	v
Acknowledgements	vii
Contents	xi
List of Figures	xvi
List of Tables	xix
List of Abbreviations	xxi
List of Notations	xxiii
1 Introduction	1
1.1 Contributions	3
1.2 Outline	7
1.3 List of publications	8
I Background	11
2 The Study of Lexical Meaning and Its Evolution	13
2.1 Theories of Lexical Meaning	13
2.2 Notions of Lexical Semantics	16
2.2.1 Fundamental Definitions	16
2.2.2 Meanings and Concepts	20
2.2.3 The Mapping Between Words and Concepts	22
2.2.4 Semasiology and Onomasiology	24
2.2.5 Semasiological Structures	26
2.2.6 Onomasiological Structures and Relations	27
2.3 Lexical Semantic Change	31
2.3.1 Lexical Semantic Change in the Word-Meaning Mapping	31
2.3.2 Development and Loss of Word-Meaning Associations	33
2.3.3 Classification of Semasiological Changes	35
2.3.4 Case Studies and Regularities	37
2.3.5 Resources and Corpora	41
2.4 Summary	43

3	Computational Modelling of Lexical Semantic Change	45
3.1	NLP Tasks in the Word-Concept Mapping	45
3.1.1	Between-word tasks	46
3.1.2	Within-word tasks	48
3.1.3	Lexical Semantic Change Detection	50
3.1.4	Novel Sense Detection and Definition Generation	53
3.2	Computational Models of Lexical Semantic Change	54
3.2.1	Static vectors	55
3.2.2	Contextualized representations	61
3.3	Beyond Change Detection	67
3.4	Summary and Limitations	71
II	Synchronic Contributions: Representing and Extracting Concepts From Data	75
4	Concept Induction: Clustering Data to Learn Latent Concepts	77
4.1	Definition of Concept Induction	77
4.1.1	Task definition	78
4.1.2	Formal framework	78
4.2	Methodology	80
4.2.1	Proposed Bi-level method	80
4.2.2	Local-only and Global-only	81
4.3	Experiments	82
4.3.1	Comparing clusterings with BCubed	82
4.3.2	Settings	84
4.3.3	Systems and baselines	85
4.3.4	Concept Induction in SemCor	86
4.3.5	Qualitative analysis of concepts clusters	87
4.3.6	Benefits at the local level	88
4.4	Extrinsic Evaluation with Concept-aware Embeddings	89
4.5	Summary	91
5	CALE: Fine-tuning a Model for Concept-Aligned Embeddings	95
5.1	Related Work on Semantic Representations	96
5.2	Concept Differentiation	97
5.2.1	Task Definition	97
5.2.2	Dataset for Concept Differentiation	98
5.3	Concept-Aligned Embeddings	99
5.4	Model Evaluation	101

5.4.1	Concept Differentiation Evaluation	102
5.4.2	Lexical Semantic Change Detection.	103
5.4.3	In-Context Lexical Similarity.	105
5.5	Spatial Organization	106
5.6	Conclusions and Limitations	109
5.7	Summary of Part II	110

III Diachronic Contributions: Studying Semasiological and Onomasiological Changes 111

6	Diachronic Concept Induction: Learning Concepts in Historical Data	113
6.1	Learning Concepts in Diachronic Data	114
6.1.1	‘All-Periods’ Concept Clustering	114
6.1.2	Formalism	117
6.1.3	Types of Semasiological Changes	118
6.1.4	Types of Onomasiological Changes	119
6.1.5	Interactions Between Semasiological and Onomasiological Changes	121
6.2	Methodology for Diachronic Concept Induction	122
6.2.1	Bi-level clustering in diachronic corpora with time-agnostic concepts.	122
6.2.2	Tuning Hyperparameters Without Concept-Annotations . .	124
6.2.3	Filtering obtained clusters	126
6.2.4	Summary of the Adaptations	127
6.3	Evaluating Diachronic Concept Induction.	127
6.4	Experiments	129
6.4.1	The PRESTO Corpus	129
6.4.2	Experimental Settings	132
6.4.3	Cluster Filtering Results	132
6.4.4	Synchronic Onomasiological Results	133
6.4.5	Onomasiological Changes	135
6.4.6	Semasiological Changes	141
6.5	Summary and Limitations	145
7	Case Study: Lexical Semantic Change in Synonyms	149
7.1	Laws of Semantic Change in Synonymy	150
7.2	Computational Studies	153
7.3	Evaluation and Predictions from Static Representations	155
7.3.1	The task of tracking change in synonymy	155

7.3.2	Evaluation Dataset	157
7.3.3	Models	160
7.3.4	Experiments	164
7.3.5	Interim Conclusion	168
7.4	Changes in Synonymy with Diachronic Concept Induction	170
7.4.1	Reallocation and Parallel Change in Concept Clusters . . .	170
7.4.2	Experiments	172
7.5	Summary	176

IV Conclusions and Appendices 179

8	Conclusions and Perspectives	181
8.1	Summary of the Thesis	181
8.2	Future Work and Perspectives	183
8.2.1	Clustering Improvements	183
8.2.2	Evaluation and Interpretability	184

Bibliography 187

A	Appendix	223
A.1	Appendices for Chapter 4	223
A.1.1	Splits and dataset statistics	223
A.1.2	Used hyperparameters and layers	223
A.1.3	Concept Clusters Size Distribution	224
A.2	Appendices for Chapter 5	225
A.2.1	SPCD Dataset	225
A.2.2	Fine-Tuning Hyperparameters	225
A.2.3	Test Metrics in SPCD	226
A.2.4	APD and PRT results on DWUGs	226
A.3	Appendices for Chapter 6	227
A.3.1	HDBSCAN cluster selection	227
A.4	Appendices for Chapter 7	230
A.4.1	Target words selection	230
A.4.2	Frequency groups	230
A.4.3	Unsupervised models	231
A.4.4	Components of selected models	231
A.4.5	Full classification results	232
A.4.6	Predictive variables in our model	232

List of Figures

2.1	Dictionary entry for <i>show</i>	19
2.2	The three layers of lexical semantics	21
2.3	Mapping Between Lexemes and Concepts	23
2.4	Mapping Between Lexemes, Occurrences and Concepts	24
2.5	Prototypicality effects in the senses of <i>fruit</i>	26
2.6	Radial network of the senses of <i>fruit</i>	27
2.7	Example of a taxonomy	30
4.1	Illustration of Concept Induction’s framework. The word <i>drum</i> is polysemous and has two senses corresponding to two different concepts, and is synonym with <i>barrel</i> for this second meaning. . .	78
5.1	Distribution of cosine distance between embeddings from the candidate models’ spaces for SPCD <i>test</i> pairs, according to their category. The dashed line is the decision threshold for Concept Differentiation learned from the other splits.	107
6.1	Distribution of PRESTO’s texts over time.	130
6.2	Distribution of obtained cluster naming sizes per period.	134
7.1	Illustration of the Law of Reallocation	151
7.2	Illustration of Parallel Change	152
7.3	Distribution of shortest distances in WordNet’s hypernymy graphs	159
7.4	Distribution of cosine distances in the SGNS space	161
7.5	Pairs of word embeddings for words w and v at two time periods T_1 and T_P and associated diachronic and synchronic distances. . .	161
7.6	Proportions of predictions of the models w.r.t. the actual distance d in WordNet of <i>noun</i> pairs. Pairs with $d = 0$ are synonymous pairs in WordNet.	167
A.1	Distribution of cluster size (in number of lemmas) obtained by the Bi-level Agglo system.	224
A.2	HDBSCAN’s hierarchy	228
A.3	HDBSCAN’s condensed tree	229
A.4	HDBSCAN’s selected clusters	230

A.5	Histograms of the value of divergence Δ of synonymous pairs, depending whether they differentiated (orange) or stayed synonyms (blue).	231
-----	---	-----

List of Tables

2.1	The different types of semasiology and onomasiology	25
3.1	DWUG datasets	52
4.1	Concept Induction BCubed Precision (P), Recall (R) and F_1 on the SemCor data averaged over 5 runs.	86
4.2	Qualitative manual evaluation of obtained word clusters of size ≥ 2	87
4.3	WSI BCubed F_1 and sense number correlation coefficient ρ on SemCor full data. Not computed for K-means because the number of cluster is constant.	89
4.4	Accuracy scores on the nouns of the WiC test dataset (Pilehvar and Camacho-Collados, 2019).	90
5.1	Balanced Accuracies on test pairs from SPCD. “SL” (resp. “DL”) refers to pairs of occurrences of the <u>s</u> ame <u>l</u> exeme (resp. of <u>d</u> ifferent <u>l</u> exemes).	102
5.2	Spearman Correlation between model Average Pairwise Divergence (APD) and gold change scores in DWUG datasets. The highest score for each dataset is in bold, and other scores that are not significantly different from it are marked with *.	104
5.3	Pearson Correlation of CoSimLex Subtask 1 (upper table) and Spearman Correlation of CoSimLex Subtask 2 (lower table). Values marked with $\dagger$ are non-significant correlation values. The highest score for each dataset is in bold, and other scores that are not significantly different from it are marked with *.	105
5.4	Silhouette scores with synsets (S) or with Unique Beginners (UB) as cluster labels ; and Spearman Correlation of cosine similarities with Wu-Palmer’s similarities in WordNet.	106
6.1	Summary statistics of our extraction from the PRESTO corpus.	131
6.2	Summary statistics of the filtering made on the PRESTO corpus from clustering results.	133
6.3	Qualitative evaluation of concept-clusters in the 1800-1949 period.	134
6.4	Evolution of the concept-clusters between Middle and Modern French.	136

6.5	Evolution of a sample of 50 <i>new</i> concept-clusters between the Middle & Preclassical French and Modern French.	137
6.6	Evolution of the 97 <i>disappearing</i> concept-clusters between the Middle & Preclassical French and Modern French.	138
6.7	Evolution of a sample of 50 <i>expanding</i> concept-clusters between the Middle & Preclassical French and Modern French.	139
6.8	Evolution of a sample of 50 <i>shrinking</i> concept-clusters between the Middle & Preclassical French and Modern French.	140
6.9	Evolution of lexemes between the Middle & Preclassical French and Modern French.	142
6.10	Evolution of a sample of 50 lexemes <i>gaining meanings</i> between the Middle & Preclassical French and Modern French.	143
6.11	Evolution of a sample of 50 lexemes <i>loosing meanings</i> between the Middle & Preclassical French and Modern French.	144
7.1	Numbers of synonymous pairs extracted from Fernald (1896) (T_1) displayed by Part-of-Speech, and numbers of those that are also considered as synonyms or hypernyms/hyponyms in WordNet (T_P). For hypernyms, we detail the proportions of hypernym/hyponym pairs that are separated by 1, 2 or 3 nodes in the WordNet graph.	158
7.2	Random samples of Xu and Kemp’s control pairs	163
7.3	Performances of the different approaches. Results are averaged over 20 random splits.	165
7.4	Cluster counts of synonym changes from the PRESTO data.	173
7.5	Average number of lexemes per concept-cluster at time T_1 w.r.t. the type of synonym change from the PRESTO data.	175
A.1	Statistics on the different data splits in annotated SemCor. The split “Synon” only contains occurrences of concepts instantiated with multiple lemmas (cases of synonymy). d_{Lex} is the average number of unique lemmas per concept, d_{polysemy} is the average number of distinct concepts per lemma.	223
A.2	Best hyperparameters on the <i>Dev</i> split.	223
A.3	CALE models hyper-parameters.	226
A.4	F1 scores on test pairs from SemCor. “SL” (resp. “DL”) refers to pairs of occurrences of the same lemma (resp. of different lemmas) and “SC” (resp. “DC”) to pairs of occurrences referring to the same concept (resp. to different concepts).	226

A.5	SPCD dataset description in number of pairs. The train split used 49,601 unique occurrences, 6,504 for validation and 14,274 for test. “SL” (resp. “DL”) refers to pairs of occurrences of the same lemma (resp. of different lemmas) and “SC” (resp. “DC”) to pairs of occurrences referring to the same concept (resp. to different concepts).	227
A.6	Accuracies on test pairs from SemCor (Balanced Accuracies are indicated with †). The higher the better. “All” refers to all pairs, “SL” (resp. “DL”) to pairs of occurrences of the same lemma (resp. of different lemmas) and “SC” (resp. “DC”) to pairs of occurrences referring to the same concept (resp. to different concepts).	227
A.7	Spearman correlations between gold change scores and predicted change scores for LSCD on the DWUG datasets, for APD and PRT measures.	228
A.8	Performances of the different approaches. Results are averaged over 20 random splits.	232
A.9	Average values of some variables for data subset based on the prediction of our best-performing LR model. TS,FS,TD,FD stand for True/False Syn/Diff. $FG_w^{(T)}$ stands for Frequency Groups of word w at time T . Significant difference within a pair of columns are in bold.	233

Abbreviations

Abbreviation	Meaning (related section)
CALE	Concept-Aligned Embeddings (Section 5.3)
CALE	Concept Differentiation (Section 5.2)
CI	Concept Induction (Section 4.1)
DCI	Diachronic Concept Induction (Section 6.2)
DWUG	Diachronic Word Usage Graph (Section 3.1.3)
LD	Law of Differentiation (Section 7.1)
LPC	Law of Parallel Change (Section 7.1)
LSC	Lexical Semantic Change (Section 2.3)
LSCD	Lexical Semantic Change Detection (Section 3.1.3)
NLP	Natural Language Processing
WiC	Word-in-Context (Section 3.1.2)
WSD	Word Sense Disambiguation (Section 3.1.2)
WSI	Word Sense Induction (Section 3.1.2)

Notations

Notation	Superset	Definition
COMMON TO ALL SETTINGS		
$\mathcal{C}$		The set of semantic concepts.
W		The lexicon, i.e. the set of all considered lexemes.
S		The set of word-senses of all the considered lexemes.
O		The corpus, the set of all occurrences of all considered lexemes.
IN THE SYNCHRONIC SETTING		
O_w	O	The set of occurrences of lexeme w .
$o_{w,i}$	O_w	The i -th occurrence of lexeme w .
M_w	$\mathcal{C}$	The set of concepts a lexeme w is associated to (its potential meanings).
N_k	W	The set of lexemes associated to a concept c_k (its naming).
S_w	S	The set of senses of a lexeme w .
$s_{w,j}$	S_w	The j -th sense of lexeme w .
IN THE DIACHRONIC SETTING		
T		The set of P (disjoint) considered time periods.
T_p	T	The p -th considered time period.
O_w	O	The set of occurrences of lexeme w from all time periods.
$O_w^{(p)}$	O_w	The set of occurrences of lexeme w at time T_p .
$o_{w,i}^{(p)}$	$O_w^{(p)}$	The i -th occurrence of lexeme w at time T_p .
$M_w^{(p)}$	$\mathcal{C}$	The set of concepts a lexeme w is associated to (its potential meanings) at time T_p .
$N_k^{(p)}$	W	The set of lexemes associated to a concept c_k (its naming) at time T_p .
S_w	S	The set of senses available to a lexeme w across all time periods.
$S_w^{(p)}$	S_w	The set of available senses of a lexeme w at time T_p .
$s_{w,j}^{(p)}$	$S_w^{(p)}$	The j -th sense of lexeme w at time T_p .

1

Introduction

Languages are dynamic systems, shaped by their users and evolving over time, across generations of speakers. Phonetical change is perhaps the most thoroughly studied type of language change in the literature (Campbell 2013, Chapter 2), with precise descriptions of the mechanisms at work in sound change and an attempt to formulate phonetic laws describing regular sound change. In contrast, semantic change is less extensively studied in linguistics, and has been the subject of debates on whether it is a regular phenomenon or a completely arbitrary one (De Saussure, 1916).

This thesis focuses particularly on *Lexical Semantic Change* (sometimes abbreviated as ‘LSC’), traditionally defined as shifts in the meaning and in the use of words (Tahmasebi et al., 2021). The study of words and their meanings can be seen as the study of conventionalized word-meaning associations: the word *mouse* is conventionally used to refer to small rodents or to ‘computer devices made to interact with a graphical interface’, while a cylindrical container for liquids is conventionally given the name of *drum* or *barrel*. The mapping between words and meanings is not one-to-one, but rather many-to-many. The fact that a word like *mouse* can be associated to multiple, different meanings is called *polysemy*, and the fact that distinct words such as *drum* and *barrel* are associated to the same meaning is referred to as *synonymy*. Moreover, word meanings are conventional by nature: ideas are associated to words insofar as the relationship between a word’s form and its meaning is not natural or inherent, but based on shared social agreement within a language community. In other words, words mean what they do because speakers collectively use and understand them in that way. From these observations, a more precise definition of Lexical Semantic Change can be formulated: lexical semantic change is the evolution of the mapping between words and their meanings. This evolution takes the form of additions to or removals from the list of conventionally recognized associations between words and meanings in a community of language users.

The word-meaning mapping, and thus lexical semantic change, can be approached from two complementary perspectives (Koch, 2008; Geeraerts et al., 2023). On the one hand, the *semasiological* approach starts from words and studies their meanings. Semasiological change, or meaning change, is the development or loss of senses for a given word. On the other hand, an *onomasiological* study starts from concepts and investigates the variety of words used by speakers to evoke this concept (whether it is an object, an action, an event or an idea), which is referred to as the ‘naming’ or ‘designation’ of the concept. Onomasiological change refers to a type of lexical semantic change that focuses on how the *expression* of a particular *concept* evolves over time. The semasiological approach corresponds to the word-to-concepts view of the mapping and naturally models polysemy, while the onomasiological approach corresponds to the concept-to-words view of the mapping and primarily models instances of synonymy. Since the first half of the 20th century and particularly after the impulse of structuralist linguistics, historical linguistics has used the onomasiological perspective in complement of the semasiological one to analyse lexical semantic change and describe its regularities and mechanisms. For instance, many new word senses are conventionalized through the systematic metaphorical mapping of one conceptual domain onto another – a process by which abstract or unfamiliar experiences are understood in terms of more concrete, familiar ones, as described by (Lakoff and Johnson, 1980). For instance, spatial relationships are conceived in terms of bodyparts (e.g. the *foot* of a mountain), time is conceived in terms of spatial relationships (e.g. *before* and *after*), human qualities may be expressed through animals such as a *lion* for being brave, or a *snake* for being treacherous (Lehrer, 1985). The onomasiological perspective also appears in the on-going debate on whether language users tend to naturally erase synonymy over time and create specializations of meaning between words designating the same concept (Clark, 1993; Blythe and Croft, 2012). It is clear that a full understanding of lexical semantic change requires examining both semasiological and onomasiological dimensions. Only by considering the changes in meaning and the ways concepts are lexicalized can we capture the regularities and patterns that underlie the evolution of the lexicon.

The last two decades, research in Natural Language Processing (NLP) has developed computational approaches of lexical semantic change, mostly focusing on detecting semantic changes in words, thus undertaking the semasiological approach. Most models are based on word embeddings i.e. dense vector representations of word usages (Tahmasebi et al., 2021; Periti and Montanelli, 2024). The modelization of lexical semantic change is indeed usually broken down to three components: representing words and their meaning in multiple time peri-

ods, aligning these representations if needed to enable diachronic comparisons, and finally measuring a degree of semantic change. The quality of these computational models is usually evaluated on semasiological benchmarks (Periti and Tahmasebi, 2024a), in which change measures produced by systems for a given list of target words are compared to gold-standard measurements of semantic change from human annotations.

Current computational approaches face two crucial limitations. First, because detecting semantic change is often reduced to the prediction of the degree of change, models rarely allow for the *characterization* of lexical semantic change. Characteristics of change to be studied include the nature of the observed change (e.g. a sense gain or a sense loss), the linguistic process involved (e.g. metaphorization, ellipsis), its causes and its consequences (e.g. correlated changes of other words, reorganization of a semantic field, etc.). Second, almost all of the computational studies of lexical semantic change in NLP adopt a semasiological approach, focusing on how the polysemic nature of a word is subject to change over time, while the onomasiological perspective, investigating how the lexical expression of a particular concept evolves over time, is lacking.

This thesis aims to provide a comprehensive characterization of both the semasiological and onomasiological dimensions of Lexical Semantic Change through computational models, with a particular focus on embedding-based approaches. Our research is guided by two central questions: how can embeddings be leveraged to model both semasiological and onomasiological change, and to what extent can such models facilitate a broader understanding of Lexical Semantic Change? The study is organized in two main stages: first, we develop models of lexical meaning in a synchronic perspective; then, we extend these models to a diachronic perspective to capture the evolution of the mappings between words and concepts over time. Starting in the synchronic setting facilitates the development of models of lexical meaning and their evaluation, as more datasets are available; after that we can adapt these to the diachronic setting and apply them to address concrete questions about lexical semantic change.

1.1 Contributions

The contributions presented in this thesis can be summarized as follows:

Concept Induction: clustering data to learn latent concepts. Our first contribution is set in a synchronic setting and focuses on contextualized embeddings, vector representations computed for each occurrence of a word in a cor-

pus, and conceptual information they can reflect. In this work, we address two research questions: (1) to what extent can we extract information about concepts from contextualized embeddings ? and (2) how to evaluate a system’s ability to learn the mapping between words and concepts from data ?

We first introduce Concept Induction, the task of clustering words into groups referring to shared concepts based on their usages in data. Contrary to existing lexical semantic tasks like Word Sense Induction, this framework simultaneously addresses both polysemy and synonymy: polysemy is captured when a word is present in multiple clusters, and synonymy when several words are grouped in the same cluster. Concept Induction therefore models both the semasiological and onomasiological sides of the words-concepts mapping. Another advantage of this framework is that a concept-inducing system does not need a pre-defined set of concepts; instead the clustering uncovers latent concepts, clusters that are evaluated through the words they contain.

Then we propose a methodology based on a bilevel clustering: a first one, applied to occurrences of each word individually to learn its senses, and a second one applied to all of the uncovered word sense clusters, to compare usages of different words and find similarities in meaning. This method is compared to traditional approaches to word sense induction, as well as more direct global clustering methods. All of these methods are applied to contextualized embeddings extracted from the intermediate layers of a pre-trained language models.

Our first finding is that contextualized language models indeed convey information about semantic concepts in their representation. Our results also indicate that the uncovered clusters of word occurrence are likely non-convex and may be best defined with a hierarchical approach. Furthermore, we observe that the bi-level approach yields better performance than direct clustering or word sense induction techniques, indicating that it is a promising method for extracting conceptual information from contextualized embeddings. This work has been published in Liétard et al. (2024a).

CALE: fine-tuning a model for concept-aligned embeddings. With Concept Induction, we have shown that pre-trained language models convey semantic information about concepts, and proposed ways to leverage that information to learn concepts. But neural language models can also be fine-tuned with a specific objective in order to exhibit useful properties for a certain task, and diminish the weight of irrelevant information from the signal. This second contribution, also set in the synchronic setting, brings an answer to the question of whether Contextualized Language Models can successfully be fine-tuned to represent lexical

meaning at the concept level and not only at the word sense level, and the extent to which this objective is beneficial in lexical semantic tasks. Inspired by recent state-of-the-art studies on fine-tuning models for lexical semantic tasks, we propose a series of models trained specifically to make contextualized embeddings reflect semantic concepts more accurately, so that words referring to the same concept end up close in the representational space.

To train these models (referred to as the CALE models, for ‘Concept-Aligned Embeddings’), we introduce a new task called “Concept Differentiation” which reformulates Concept Induction as a binary classification task among pairs of word occurrences. We also curate a dataset of hundreds of thousands of usage pairs in English. We fine-tune a multilingual Contextualized Language Model for this objective using siamese networks and contrastive learning.

The obtained models are compared to their pre-trained counterparts, as well as a state-of-the-art fine-tuned model trained only to differentiate word senses (and not to compare different words). Our experiments, conducted on a variety of lexical semantic tasks that compare word usages both in isolation and to other words, show that CALE models indeed provides the most accurate representations of lexical meaning, and closer analyses show that the obtained representational spaces exhibit interesting properties, such as being more easily clustered into concepts than those of pre-trained models.

This work has been peer-reviewed and will be submitted to EACL in the upcoming months, and a preprint version is available (Liétard and Loiseau, 2025). The dataset and the models have been publicly released on the HuggingFace platform.

Diachronic Concept Induction: learning concepts in historical corpora. Compared to the synchronic setting, learning concepts in diachronic data offers additional challenges and requires adapting the clustering-based method of Concept Induction to this new setting. In this third contribution (and first diachronic one) we address two questions: how can concepts be learned from contextualized representation in historical data, and how can we evaluate whether the proposed method accurately describes changes in the words-concepts mapping.

We propose an adaptation of the clustering approach of our first contribution to the diachronic setting, enabling the learning concepts across multiple time periods, without relying on semantic annotations on the data. We apply the proposed method, now *Diachronic* Concept Induction, to the PRESTO corpus, composed of French historical texts from the 16th century to the 20th century. The overall system is a bi-level clustering on the occurrence embeddings (produced with a

CALE model from the previous contribution) from all time periods at once, with hyperparameters selected to maximize the retrieval of synonymy among clusters.

To evaluate our method, we conduct a thorough manual evaluation comparing the onomasiological and semasiological changes uncovered by our method to changes actually attested in a reference historical dictionary, namely the *Trésor de la Langue Française informatisé*. Our evaluation shows that most of the detected changes correspond to attested historical shifts, across both the semasiological and onomasiological perspectives; and we find that our method more accurately captures changes ‘by gains’ (e.g. a word gaining a new sense) than changes ‘by loss’ (e.g. a word no longer associated with a concept). We can compare our results to previously proposed statistical laws of semantic change, showcasing the versatility of the proposed method to characterize lexical semantic change on both sides of the mapping between words and concepts.

The methodological part of this work has been published in Liétard et al. (2024b), while experimental results and discussions are more recent material.

A case study on the semantic evolution of synonymy over time. Our fourth contribution focuses on two different mechanisms of change among synonyms posited by historical linguists that lead to apparently contradictory generalizations. On the one hand the so-called law of ‘reallocation’ states that synonyms cannot stay synonyms for long: either they differentiate in meaning, or they undergo a selection process; suggesting that synonyms compete for a given meaning. On the other hand, historical studies have also documented cases of ‘parallel change’ where synonyms evolve in a similar way and maintain their semantic association through shifts of meaning, thereby extending their synonymy to other concepts. These two apparently contradicting phenomena set the basis for this case study, in which we investigate the extent to which computational models of semantic change can predict and characterize complex cases of change, and how integrating onomasiological considerations can enhance a computational approach. Our contribution to the study of these changes is twofold.

First, building upon a prior work of Xu and Kemp (2015), we provide a framework assessing a model’s ability to predict whether pairs of synonyms have differentiated in meaning over time or not, and we evaluate a series of computational models of semantic change based on static embeddings. To build an evaluation dataset, we compare two lists of synonyms from different historical periods: if pairs of the earlier list are not found in the more recent list, the synonyms are assumed to have differentiated in meaning. Our experiments show that changes in our static representations are not fully predictive of the observed semantic

changes, as these representations struggle to capture a consistent notion of synonymy and to account for polysemy. Interestingly, both the curated data and tested models display a prevalence of a differentiation of synonyms, contrary to the findings of Xu and Kemp.

Second, we apply the Diachronic Concept Induction method along with the results obtained from French historical data to investigate both reallocation and parallel change from a more onomasiological perspective, exploiting the structure of this method to provide a more comprehensive characterization of the targeted phenomena. Doing so, we provide a concrete use-case for the method that has been elaborated throughout our first three contributions. Ultimately, we find that the parallel changes are less common than the reallocation processes, but that most synonym groups undergo neither type of change and instead persist over time. Our results also suggest that reallocation and parallel change can both apply for a given concept and are not mutually exclusive. Contrarily to our first intuition, we further find that reallocation is not stronger in larger sets of synonyms than in smaller ones.

Comparing the different approaches, we show that different representations of synonymy (static embeddings + pairwise distances; or contextualized embeddings + clusters) lead to different characterization of the problem of lexical semantic change in synonyms and that the framework of Diachronic Concept Induction, at the expense of additional hypotheses, allows for a more comprehensive characterization of this change and bypasses the need for an external expert resource to identify synonyms.

The study based on crossed synonym lists and static embeddings has been published in Liétard et al. (2023), while the results obtained with the Diachronic Concept Induction method represent more recent work.

1.2 Outline

The first part of this thesis provides the necessary background for this study. Chapter 2 introduces the key notions and definitions, including a discussion on the mapping between words and meanings (or “concepts” in particular), as well as a concise overview on the different approaches to semasiology and onomasiology in lexical semantics. It subsequently reviews the main theories of lexical semantic change in diachronic linguistics, classifications of semantic change and lists several regularities that have been observed by linguists. The chapter concludes with an overview of major lexical resources for the study of lexical semantic change, and a presentation of the most widely employed diachronic text

corpora.

Chapter 3 provides an extensive review of the literature on Natural Language Processing approaches to lexical semantics on the one hand, and to lexical semantic change on the other hand. The chapter first discusses how NLP has framed these questions in terms of *tasks*, and then describes the range of models developed for the study of lexical semantic change. Particular attention is given to vector-based models (i.e. word embeddings), as they are the primary focus of this thesis and represent the most widely used techniques in recent years. The chapter concludes with a review of the various instances and regularities of lexical semantic change that have been studied and/or uncovered by computational studies.

Chapters 4 to 7 present the original contributions we have described above. Chapter 4 is dedicated to Concept Induction in the synchronic setting, its evaluation and the proposed models. Chapter 5 introduces the CALE models as well as Concept Differentiation and the related dataset. Chapter 6 presents the Diachronic Concept Induction method, its application to the PRESTO corpus as well as the evaluation comparing detected changes with attested changes. Chapter 7 is dedicated to the case study on the evolution of synonymy, encompassing both the work based on static embeddings and the application of Diachronic Concept Induction with contextualized embeddings.

Finally, Chapter 8 summarizes the outcomes of our contributions and explains how they address the formulated research questions. It also discusses the remaining limitations of the proposed approaches and outlines possible directions for future research.

1.3 List of publications

Bastien Liétard, Mikaela Keller, and Pascal Denis. A tale of two laws of semantic change: Predicting synonym changes with distributional semantic models. In *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023)*, July 2023. Association for Computational Linguistics. URL <https://aclanthology.org/2023.starsem-1.30/>

Bastien Liétard, Mikaela Keller, and Pascal Denis. Towards an onomasiological study of lexical semantic change through the induction of concepts. In *Proceedings of the 5th Workshop on Computational Approaches to Historical Language Change (L'Change 24)*, August 2024. Association for Computational Linguistics. URL <https://aclanthology.org/2024.lchange-1.15/>

Bastien Liétard, Pascal Denis and Mikaela Keller. To word senses and beyond: Inducing concepts with contextualized language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP 2024)*, November 2024. Association for Computational Linguistics. URL <https://aclanthology.org/2024.emnlp-main.156/>

Bastien Liétard and Gabriel Loiseau. CALE : Concept-Aligned Embeddings for both within-lemma and inter-lemma sense differentiation, 2025. URL <https://arxiv.org/abs/2508.04494>. arXiv preprint

Part I

Background

2

The Study of Lexical Meaning and Its Evolution

This chapter provides an overview of lexical semantics and lexical semantic change as they are studied in linguistics. The first section of this chapter situates our work among the various theories of lexical semantics: it discusses several notable axes along which the different approaches are contrasted, and explicitly states the theoretical framework of this thesis. The second section of this chapter is dedicated to some fundamental notions of lexical semantics that form the basis for the rest of the present work. This includes the definition of lexemes and concepts, as well as discussions about polysemy and synonymy, semasiology and onomasiology, taxonomies and categories.

The third section of this chapter reviews major advancements and proposals about the temporal evolution of lexical meaning in diachronic linguistics, as well as several widely-used corpora and resources for the study of lexical semantic change.

2.1 Theories of Lexical Meaning

Lexical semantics finds its roots in the lexicographic tradition, with the development of bilingual and then monolingual dictionaries during the 16th century (Leca-Tsiomis, 2006). Lexicographers paved the way for the scientific study of the relation between words and meanings, beginning from the late 19th century onward with *historical-philological* approaches to lexical semantics. These studies were inherently diachronic, focusing on how word meanings emerged from older meanings and older forms, and on the relations that exists between these senses in the first scientific studies of polysemy (Paul, 1880). They are often based on the idea that word meanings are subject to change because they are the product in *interpretation*, a combination of psychological processes and so-

cial conventions, constantly invoked, changed or revoked within the language community. Semanticists of the historical-philological tradition elaborated classifications of change patterns (Stern, 1931) and edicted laws about the emergences, disappearances and shifts of word meanings over time Bréal (1897). The historical-philological tradition was superseded by another approach to lexical semantics, the *structuralist* approach, which was itself superseded by a variety of other theoretical frameworks, including for instance *cognitive* semantics, *formal* semantics and *distributional* semantics.

Synchronic vs. diachronic approaches. The various frameworks under which the relation between words and meanings has been studied can be contrasted along various dimensions. As mentionned above, historical-philological semantics is inherently *diachronic*, as it studies word meanings by examining their changes over time. In contrast, structuralist linguistics generally rejects the diachronic aspect, in favor of a *synchronic* study of the language (De Saussure 1916). The cognitive approach to lexical meaning for instance is interested both the diachronic processes of change and the synchronic variations, alternatives and oppositions (Geeraerts, 1997, chapter 5). In this thesis, we study word meaning in both the synchronic and diachronic settings: first we study modelizations of the word-meaning mapping in a single time period first (chapters 4 and 5), before we extend the methodology to multiple time periods and compare the different stages of this mapping (chapters 6 and 7).

System-level vs. usage-level approaches. The analysis of structuralist semantics is fundamentally at the *system* level: word meanings are not studied in isolation from one another, instead the analysis describes the interactions and relations within a set of words/a set of meanings. The so-called system is considered in isolation from human interactions and mental states. In the framework of *generative grammars*, semanticist follow the Saussurean distinction between '*langue*' (the language as a system of items, rules and conditions) and '*parole*' (the production of language in context) and aim at describing only the former. Cognitive semantics on the other hand are focused on *usage* rather than a system: the study of word meaning primary consists of analyses of how speakers use words to convey meaning (Croft and Cruse, 2004). As pointed by Geeraerts (2010), most attempts at distinguishing linguistic meaning and encyclopedic meaning failed to clearly set the frontier between them. The present work aims at providing a more systemic description of word meanings, and the proposed methods are used to probe the overall distribution of words over meanings across a large portion of the lexicon. Yet, we do not make a distinction between a

so-called “purely-linguistic” meaning and encyclopedic knowledge; we integrate them both as parts of word meanings, as it is through their combination that language users determine the words they employ.

Decontextualized vs. contextualized approaches. Another difference between the various framework is the level of *contextualization*: structuralist semantics usually study word meanings outside of any precise linguistic context; on the other end of the spectrum, *distributional* semantics argues that the meaning of a word is found in the contexts in which it appears (Harris, 1954; Firth, 1957; Geeraerts et al., 2023). The distributional semantics framework is at the foundation of most models representing lexical meaning in Natural Language Processing, from the so-called Distributional Semantic Models (representations of word meanings using direct co-occurrence statistics) to the more recent Contextualized Language Models (models that learn representations of in-context word usages from high-order co-occurrence statistics). A longer discussion on how these models are used to represent lexical meaning in Chapter 3. The work presented in this thesis is fundamentally grounded in context: we examine the contexts in which a word appears in order to determine its meaning. From this investigation, we extract a view of the mapping between words and meanings.

Different conceptions of lexical meaning. Another point of debate is the very nature of word meanings. Both the pre-structuralist historical approach to lexical semantics and post-structuralist cognitive semantics consider lexical meaning to be a psychological/cognitive entity, the product of a particular mental state (Sweetser, 1984; Koch, 2008). Structuralist semantics and generative semantics emphasize on the distinction between *encyclopedic* (that relates to the world and individual experiences) knowledge and *purely linguistic* knowledge, arguing that only the latter should be the object of lexical semantics (Geeraerts, 2010, chapter 2). In the structuralist framework, word meanings can for instance be described their oppositions and their distribution over a conceptual field (Trier, 1934; Pottier, 1964). Lexical meaning can also be studied by examining the set of real-world referents a word is or can be applied to, as in formal semantics (Kripke et al., 1980; Green, 2007). The theoretical framework of this thesis is that of concepts, mental representations of categories within the language users mind. Studying the mapping between words and meanings, we study the conventionally established relations between lexical items and the concepts they can invoke. The meaning of a word in a given context is the concept to which it refers. More discussion is provided in Section 2.2.2.

Semasiology vs. onomasiology. Finally, the various approaches to lexical semantics can be distinguished by their perspective on the relation between words and meanings. The historical-philological tradition for instance investigates only the *semasiological* perspective¹, studying for instance the links that exist between the many senses of a word. The primary object of semasiology is polysemy, the fact that a single word has multiple meanings. The structuralist tradition on the contrary adopts an *onomasiological* approach to lexical meaning, as it studies relations between words and between word meanings within a lexical field for instance. Onomasiology is interested in a phenomenon like *synonymy* (the fact that multiple words can bear a common meaning), *hyponymy* (the fact that two words can be related by an inclusion-type relation between their meanings), *meronymy* (when there is a part-whole relation between the meanings of two words) or *antonymy* (the fact that two words refer to opposite ends on the spectrum of a given property). While most of the literature about lexical semantic change in Natural Language Processing adopts a semasiological-only approach, we believe these two perspectives are complementary and can be studied together. Thus the present thesis is an attempt to model onomasiological aspects alongside semasiological ones using computational approaches. More discussion about the definitions of semasiology and onomasiology in Section 2.2.4.

Now that we have situated our research among the various approaches to lexical semantics, let us define the theoretical apparatus and the fundamental notions we will be using throughout this thesis.

2.2 Notions of Lexical Semantics

This section defines the fundamental notions we will use throughout this work and discusses the model of a mapping between lexemes and concepts. We examine some of the theoretical implications of a concept-based view of meaning. Finally we discuss the difference between semasiology and onomasiology, highlighting some of the key features of the latter approach.

2.2.1 Fundamental Definitions

In this thesis, we use the terms ‘word’ and ‘lexical item’ as a synonym for **lexeme**. Studying word meaning, we will consider all forms of a lexeme, regardless of the impact of inflections on the meaning. For instance, in *She hit her foot on the table corner* / *I have large feet compared to the average human*, we will consider that

¹ There are some exceptions to this semasiological-only trend in the historical-philological tradition, such as the *Wörter und Sachen* movement.

both *foot* and *feet* share the same lexical meaning, ignoring the difference in number. We will refer to a lexeme using its **lemma**, i.e. the form conventionally chosen to represent the set of its forms. The notion of ‘lemma’ is particularly important in lexicography, as it is typically the form found in dictionary entries that readers query when looking for the meaning of a lexeme. We will consider that lexemes are bound to a given part-of-speech. For instance, *fly* as a noun (the insect) and *fly* as a verb (the action) are treated as two distinct lexemes. The former can take the forms *fly* and *flies*, while the latter can occur as *fly*, *flying*, *flew*, and so on.

The **lexicon** is the list of lexemes available to a speaker of a given language. As this thesis does not explore cross-linguistic studies, nor make distinction based on individual speakers, the lexicon is merely be the set of all lexemes in a given language. Because this computational work is necessarily limited to available data, which do not provide an exhaustive coverage of all lexemes of the language at a given time, the lexicon will often be restricted to those lexemes that occur frequently enough in the data. We will refer to these lexemes as **target words**.

Word meanings. In this work on the diachronic evolution of lexical meaning, we will study changes in *conventionalized* associations between words and meanings. Since Frege et al. (1892), it has been customary to distinguish between the *intension* of a word, i.e. the abstract idea, and its *extension* which encompasses all past, present, future, potential, or fictive entities that can be denoted by this word. For instance, the intension of *legislative chamber* can be represented defined by a specific Wikipedia entry (https://en.wikipedia.org/wiki/Legislative_chamber), while its extension would be the corresponding Wikipedia category (https://en.wikipedia.org/wiki/Category:Legislative_chambers). Of course, intension and extension are not independent: the more precise the definition, the more restricted is the class of referents; and the larger the set of denoted entities, the more general the definition becomes. Our work focuses on the intension-side of lexical meaning. In particular, we say that a word meaning is (intensionally) the **concept** it refers to in a given context. If no particular context is given, we say that its meaning is the set of concepts it is conventionally associated with. A concept is an idea, it is a non-linguistic, mental representation of a class of entities (Murphy, 2002). Concepts can be concrete or abstract, general or specific, etc. For the sake of simplicity, the set of concepts is considered discrete (but not necessarily finite). In order to study the mapping between words and meanings, we actually investigate the word-concept associations. Section 2.2.2 provides a more comprehensive discussion on the notion of concepts and its relation to word meanings.

Typographical conventions. Lexemes will be denoted with their lemma in light italics (e.g. *mouse*). Concepts will be denoted using the most relevant lexeme in small capital letters (e.g. LEGISLATIVE CHAMBER), and in figures we may make use of emojis to illustrate concepts and avoid ambiguity between linguistic forms and non-linguistic mental representations. If a concept is usually instantiated by a single lexeme that may be ambiguous, we will make use of definitions or contextual information in the concept’s reference if judged necessary. For instance, in the domain of computer peripherals, we will avoid to say that the word *mouse* refers to the concept of MOUSE as the lexeme is polysemous and it could lead to confusion with the rodent concept. Instead, we will say that it refers to the concept of MOUSE (COMPUTER DEVICE), defined as ‘a computer device used to point and click in a graphical interface’ (in opposition to the concept of MOUSE (RODENT)).

Occurrences. We call “occurrence of lexeme *w*” the use of any form of *w* in a specific context in a given corpus. Consider the following examples:

- (1) the jury found them guilty in a fair trial.
- (2) candidates competed in a trial of skill.
- (3) the hero underwent a test of strength.

Here, (1) and (2) are occurrences of the lexeme *trial*, while (3) is an occurrence of *test*. We would say that (1) and (2) realize two different senses of *trial*, as they contain two very different usages of the word and denote two distinct concepts. In particular, (1) refers to the concept of PROCESS OF LAW TO DETERMINE GUILT, while (2) refer to the concept of CHALLENGE TO EVALUATE A SKILL. The occurrence (3) of *test* also refers to the latter concept.

Concept namings and synonymy. Across the lexicon, there may be more than one word that can be used to refer to the same concept. For instance, in the sentence *the fire crackled as she lit the match*, the word *fire* could be replaced by *flame* without any change of meaning at the sentence level because they both refer to the concept of FLAME. We call **concept naming** the set of lexemes that can be used to refer to a given concept, and we call **synonyms** all the words belonging to the same naming. Namings can sometimes also called “designations”. In our example, both *fire* and *flame* belong to the naming of FLAME and therefore are synonyms. Note that we define synonymy at the *denotational* level (as opposed to the *connotational* level): in context in which they refer to the same concept, synonyms should be interchangeable without alteration of the literal meaning,

regardless of emotional or cultural value associated to them. For instance, *cop* and *policeman* are said to be synonyms, even though they should probably not be used in the same context from a connotational standpoint. This notion of synonymy is equivalent to cognitive synonymy in Stanojevic (2009). More discussion on the different notions of synonymy can be found later in Section 2.2.6.

Word senses and polysemy. Word senses (sometimes called “readings” as in Geeraerts (2010) for instance) are conventionally established ways to use a word to refer to a concept. They are called “word senses” in the NLP literature because they are defined as bound to a specific lexeme. Dictionaries typically provide this information: each entry corresponds to the lemma of a lexeme, which is associated with a list of senses, as illustrated in Figure 2.1.

1	: to cause or permit to be seen : EXHIBIT
	showed pictures of the baby
2	: to offer for sale
	stores were <i>showing</i> new spring suits
3	: to present as a public spectacle : PERFORM
4	: to reveal by one's condition, nature, or behavior
	<i>showed</i> themselves to be cowards
5	: to give indication or record of
	an anemometer <i>shows</i> wind speed
6 a	: to point out : direct attention to
	<i>showed</i> the view from the terrace
b	: CONDUCT, USHER
	<i>showed</i> me to an aisle seat

Figure 2.1: Entry for *show* in Merriam-Webster dictionary (2025). In black are text definitions, in gray are use-context examples and in blue are synonyms.

A word sense can be defined using:

- the associated concept or a text definition of the associated concept (e.g. *mouse* in the sense of ‘computer peripheral used to point and click on a graphical interface’).
- a conceptual category to which the associated concept belongs (e.g. *mouse* in the sense of COMPUTER PERIPHERALS)
- a prototypical use-context that lifts ambiguity (e.g. *mouse* as in *double left-click with the mouse to open a file.*)
- a synonym or near-synonym (e.g. *mouse* in the sense of *rat*).

Word senses can also be latently defined by grouping similar usages of the lexeme. Figure 2.1 provides an excerpt from Merriam-Webster dictionary (2025)’s dictionary entry for the lemma *show*, that displays use of different techniques to define the word’s senses.

Each sense of a lexeme is a pointer toward one and only one concept. We say that a word is **polysemous** / exhibits **polysemy** if it has strictly more than one conventionally established sense. The word *mouse* is polysemous, with a first sense for the concept of MOUSE (RODENT) and another sense for MOUSE (COMPUTER DEVICE).

We have seen in examples (1) and (2) that different occurrences of the same word can refer to different concepts, thus illustrating the polysemous nature of the lexeme. In that case, the context determines a selection of a particular word sense among the lexeme’s senses inventory. In this thesis, for the sake of simplicity, we assume that the linguistic context surrounding a word provides sufficient cues for disambiguation such that each occurrence realizes only one sense of a word and thus refers to one concept. Future work could study ambiguity, cases where several concepts could plausibly be associated to the same occurrence without a sufficient contextual information to select one in particular.

2.2.2 Meanings and Concepts

In many theories of lexical meaning, and particularly in cognitive semantics, the notion of **concepts** plays a central role. It is often associated with the notion of *categories*, as concepts are the mental tool that helps us classify and understand our experiences. In this thesis, we will use the notion of concepts and of their relation to categories given in the introduction of Murphy (2002):

“Concepts are the glue that holds our mental world together.

...

If we have formed a concept (a mental representation) corresponding to that category (the class of objects in the world), then the concept will help us understand and respond appropriately to a new entity in that category.

...

Our concepts embody much of our knowledge of the world, telling us what things there are and what properties they have.”

(*The Big Book of Concepts*, Murphy 2002).

In this quotation, the term “objects” is to be taken in its largest sense: not only physical objects but also entities, abstract notions, events, states. **Concepts** are therefore defined as *mental representations* of categories that we abstract from

our experiences in the world. Through our perceptions and other sources of interactions with the world, we construct an idea of that category.

Concepts, categories and word meanings. We have previously established that we consider concepts to be the intensional component of lexical meaning; categories are therefore related to its extensional component. To understand the relation between words, concepts and categories, let us examine Figure 2.2 from Geeraerts et al. (2023).

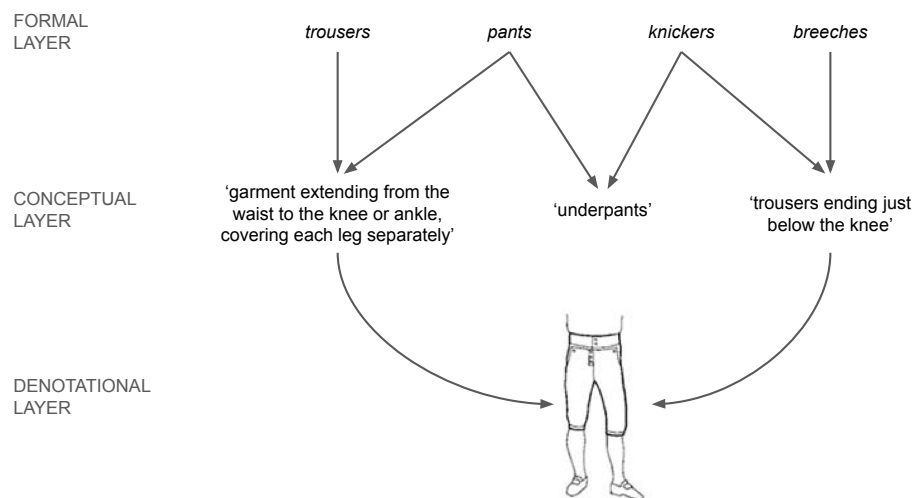


Figure 2.2: The three layers of lexical semantics, from Geeraerts et al. (2023).

Geeraerts et al. (2023) distinguish three layers to describe lexical meaning: the *formal* layer, that of lexical items; the *conceptual* layer, that of concepts; and the *denotational* layer², that of real-world or possible entities. Words are used to refer to concepts (their intension); concepts are mental representations that help us grasp our experiences and organize entities into categories. Traditionally, the extension of a word is defined as the set of entities it can be applied to; thus in this figure it is the set of referents (at the denotational layer) that are linked to the word (at the formal layer) through the conceptual layer.

Interaction between the conceptual and the denotational layers. On the one hand, the act of assigning an entity at the denotational layer to a concept is called *categorization*, as it determines the class within which the referent will be considered. On the other hand, psychological research has shown that human categorization is subject to so-called **prototypicality effects** (Rosch, 1973, 1975;

² Geeraerts et al. (2023) make a different use of the term “denotational” than the one we made earlier. In this instance, it is not opposed to “connotational”, it is rather a synonym of *referential*.

Rosch and Mervis, 1975): some members are better representatives of the category than other (they are more ‘salient’ or ‘typical’) and categories are therefore organized around them in a radial structure, where central members are more representative than peripheral ones (they are said to be more *prototypical*). In this model, category membership is determined by the degree of similarity to the most prototypical members, as opposed to a model of necessary and sufficient conditions. The concept of RED, for instance, is better understood as encompassing as ‘colours similar to that of hexadecimal code FF0000’ rather than a precise range of hexadecimal values.

As Figure 2.2 illustrates, two concepts (at the conceptual layer) can apply to categorize the same represented object (at the denotational layer). This ambiguity in categorization results in a linguistic ambiguity: should the word *trousers* or the word *breeches* be used to refer to this clothing item ? In a given sentence where the speaker refers to this object, interchanging these two terms would likely not affect the truth-value of the utterance (the item belongs to the extensions of both lexemes), yet two different concepts would be instantiated. The degree to which the item is typical with respect to each of the two categories may determine the preferred lexical choice; contexts may also favor the choice of one lexical item over the other.

Scope of our work. This thesis will focus only on intensional meaning, and thus we won’t consider the denotational layer. Therefore, when comparing the lexical meaning of two words in their respective occurrences, we do not address the question of whether they are referring to the same entity; instead, we address only the question of whether they are evoking the same concept. In the example illustrated in Figure 2.2, we are interested in the fact that *breeches* and *trousers* refer to two different concepts, while *pants* and *knickers* can refer to the same concept in some context. From now onwards, unless explicitly mentioned otherwise, the term ‘meaning’ will be used equivalent to ‘concept’. In general, we are interested in the mapping between words and concepts.

2.2.3 The Mapping Between Words and Concepts

In this thesis, we are particularly interested in the mapping between words that speaker use, and meanings they associate with them. In this perspective, Lexical Semantic Change can be seen as the process of altering this mapping by the conventional establishment or disappearance of some word-meaning associations. The mapping between lexemes and semantic concepts has been studied under different denominations in other fields: François (2022) refers to it as “patterns

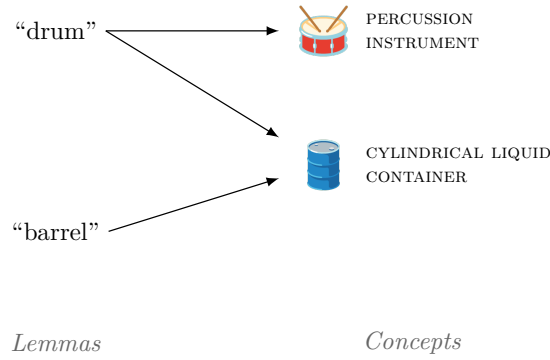


Figure 2.3: Illustration of the mapping between Lexemes and concepts.

of lexification”, while some psycholinguistics works discuss the notion of “mental lexicon” (Elman, 2004; Jarema and Libben, 2007; Paivio, 2010). In this work, the lexeme-concept mapping is a modelization tool to better describe and visualize conventionally established (among a community of speakers) associations between words and meanings, following recent approaches in linguistic typology François (2022) or Haspelmath (2023). We do not claim that a mapping between lexemes and concepts is an accurate depiction of language functions in the human mind. This thesis is not intended to engage in the on-going debate in psycholinguistics concerning the existence and structure of a mental lexicon in our cognition.

In Figure 2.3 we illustrates of this mapping using the two lexemes *barrel* and *drum*, and the concepts of CYLINDRICAL CONTAINER FOR LIQUIDS and PERCUSSION INSTRUMENT. Word senses are represented by arrows linking words to concepts. We would observe a case of **polysemy**, as *drum* can be used to refer to both PERCUSSION INSTRUMENT and CYLINDRICAL LIQUID CONTAINER. Conversely, the concept of CYLINDRICAL CONTAINER FOR LIQUIDS’s naming includes both *drum* and *barrel*, illustrating **synonymy**. Our research primarily focuses around polysemy and synonymy. Other lexical relations, such as hypernymy or antonymy, are left to future investigation.

In practice, there are two intermediary levels: occurrences and senses. When we use a word in a given context (an occurrence), it realizes a particular conventionalized way of using that word (a sense) in order to refer to some concept. This is illustrated in Figure 2.4, where $o_{w,i}$ is the i -th occurrence of lexeme w and $s_{w,j}$ the j -th sense of word w , as it could possibly be listed in the dictionary entry for w . Generally, we denote W the considered lexicon; O the set of all occurrences of all considered words (i.e. the corpus); and O_w the occurrence set of a word w . Similarly, we denote S_w the set of senses of a word w , and S is the

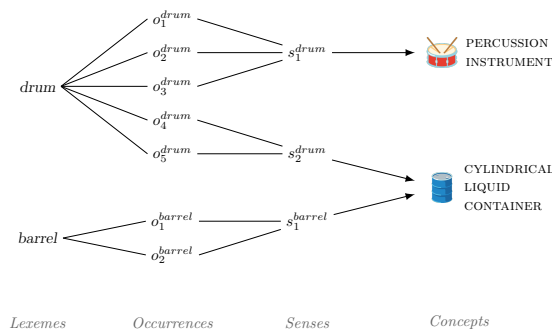


Figure 2.4: Illustration of the mapping between lexemes, their occurrences and the corresponding concepts.

set of all senses for all considered words. Finally, the set of concepts is denoted $\mathcal{C}$. We will use these notations throughout this work.

The mapping between occurrences and concepts is many-to-one: each concept in $\mathcal{C}$ can be evoked in many occurrences in O , but each occurrence evokes only one concept. The same holds for the mapping between senses and concepts: each concept can be realized through multiple senses (from different words) in S , but a word sense points towards only one concept. In contrast, the mapping between lexemes in W and concepts in $\mathcal{C}$ is many-to-many, illustrating polysemy and synonymy: some words evoke multiple concepts (polysemy), and some concepts can be designated by several words (synonymy). As noted Murphy (2002, chapter 11), at a given point in time in given language, certain concepts may not be instantiated by any words at all, concepts that nonetheless exist in human experience, or that may be lexicalized in other languages and/or in other periods of time.

In Chapter 4, we introduce Concept Induction, the task of modeling the many-to-many mapping between lexemes in W and concepts in $\mathcal{C}$ using a corpus O . In practice, we exploit many-to-one mappings between O and S in a first step, and between S and $\mathcal{C}$ in a second step.

2.2.4 Semasiology and Onomasiology

The mapping between forms and meanings can be approached from two complementary perspectives. On the one hand, **semasiology** starts out from lexemes and studies the different meanings that are associated to a given item; on the other hand, **onomasiology** takes the reverse perspective and studies the set of lexemes used to refer to a given concept. In Figure 2.3, an onomasiological approach would observe that the concept of CYLINDRICAL LIQUID CONTAINER can be instantiated by two lexemes: *drum* and *barrel*. A semasiological approach would

highlight that the word *drum* has two different meanings, that of CYLINDRICAL LIQUID CONTAINER and that of PERCUSSION INSTRUMENT.

Let us recall the three layer of lexical meaning according to Geeraerts et al. (2023) and illustrated in Figure 2.2. The *formal layer* is the layer of lexemes; the *conceptual layer* is that of concepts and categories; the *denotational layer* is that of referents, of objects and real-world instances. The interactions between these three layers are characterized by four features: *polysemy* (the multiplicity of concepts associated to a given form); *synonymy* (the multiplicity of lexemes to express a given concept); *prototypicality* (the extent to which a referent is representative of a given concept); and *categorization* (the potential conceptual categories that can be applied to a given referent).

Building on these four phenomena, Geeraerts et al. (2023) distinguish two kinds of semasiology. First, *between-sense* semasiology investigates polysemy – the multiple senses of a lexeme – from the formal layer to the conceptual layer. Second, *within-sense* semasiology describes the prototypicality levels of referents associated to a category, from the conceptual layer to the denotational layer. The authors also distinguishes two types of onomasiology: *formal* onomasiology, which studies synonymy and the lexemes associated to a given concept, from the conceptual layer to the formal layer; and *conceptual* onomasiology, which is focused on the categorization of a given referent, from the denotational layer to the conceptual layer. This is summarized in Table 2.1.

	SEMASIOLOGY	ONOMASIOLOGY
formal layer w.r.t. conceptual layer	Between-sense semasiology <i>polysemy</i> (lexemes to concepts)	Formal onomasiology <i>synonymy</i> (concepts to lexemes)
conceptual layer w.r.t. denotational layer	Within-sense semasiology <i>prototypicality</i> (concepts to referents)	Conceptual onomasiology <i>categorization</i> (referents to concepts)

Table 2.1: The different types of semasiology and onomasiology, according to Geeraerts et al. (2023)

Our research primarily focuses on the relation between the formal layer and the conceptual layer, which corresponds to the word-meaning mapping we have presented earlier. Consequently, our analysis is limited to *polysemy* and *synonymy*, and their evolution over time. Throughout this thesis, the term *semasiology* specifically refers to between-sense semasiology only, and the term *onomasiology* to formal onomasiology. The expression *semasiological changes* will

designate changes in sense inventories, i.e. in polysemies, while *onomasiological changes* will refer to changes in concept namings, i.e. in synonymy. Section 2.3 discusses the different semasiological and onomasiological aspects of Lexical Semantic Change; and Chapter 3 examines how existing Natural Language Processing studies addresses both perspectives.

2.2.5 Semasiological Structures

Prototypicality effects not only apply to the referents within a conceptual category, they also structure polysemies, in a semasiological perspective (more precisely, at the level of *between-sense semasiology* in Table 2.1).

According to Lakoff “Most words and morphemes have multiple meanings – meanings that are related to one another. These meanings can be seen as forming a radial category, in which there is a central meaning and a structure of related meanings which are motivated by the central meaning.” (Lakoff, 1987, page 291).

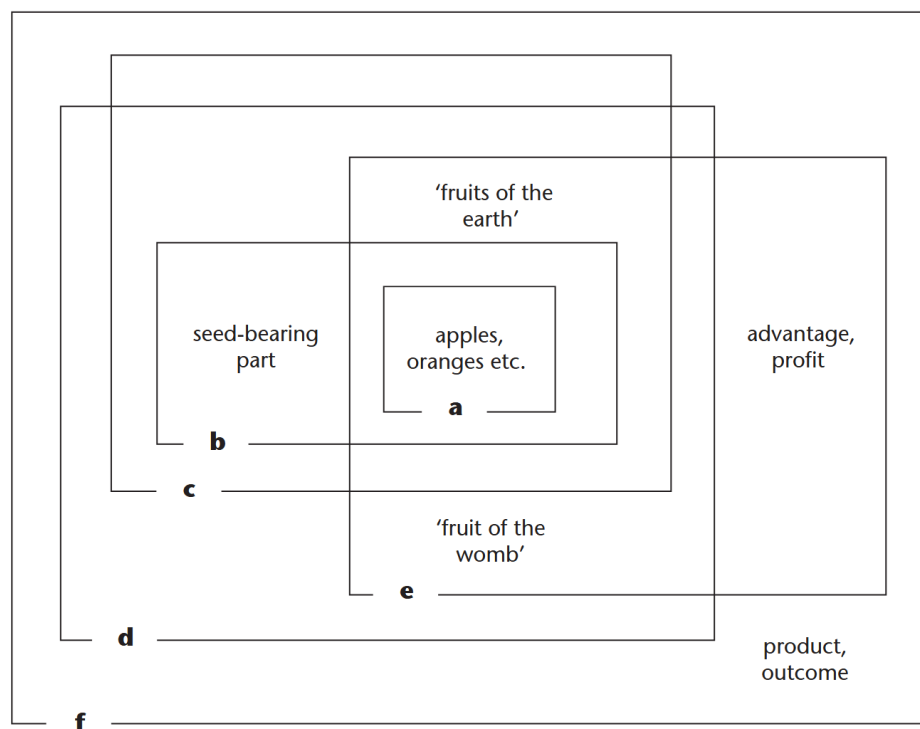


Figure 2.5: Prototypicality effects in the senses of *fruit*, from Geeraerts (2010)

Building on this idea, Geeraerts (2010) examines six senses of the word *fruit* and establishes the radial network presented in Figure 2.5. Decomposing each meaning as a combination of 6 features (e.g. ‘being sweet, juicy, commonly used as dessert’, ‘being the natural result of an organic process’, ‘being the outcome of

a process or activity', etc.), he shows that the central senses have many features in common, while the more peripheral senses exhibit only a selected few of these features; as such, some senses are more similar to one another, and some senses are very dissimilar. This network illustrates that the meaning pertaining to apples and oranges is more very typical while the sense of 'product, outcome' is much more peripheral. The network can be represented as a directed graph that relates senses to one another; edges can be labelled with the kind of semantic relations that underlies them and provide information on the historical development of senses. Example of such graph is provided in Figure 2.6. The view of polysemy as subject to prototypicality effects forms the basis of *diachronic prototype theory* of lexical semantic change that will be discussed in Section 2.3.2. Furthermore, Section 2.3.3 reviews several typologies of the semasiological developments that can take place in a word's sense inventory (e.g. *generalization* or *metaphor*).

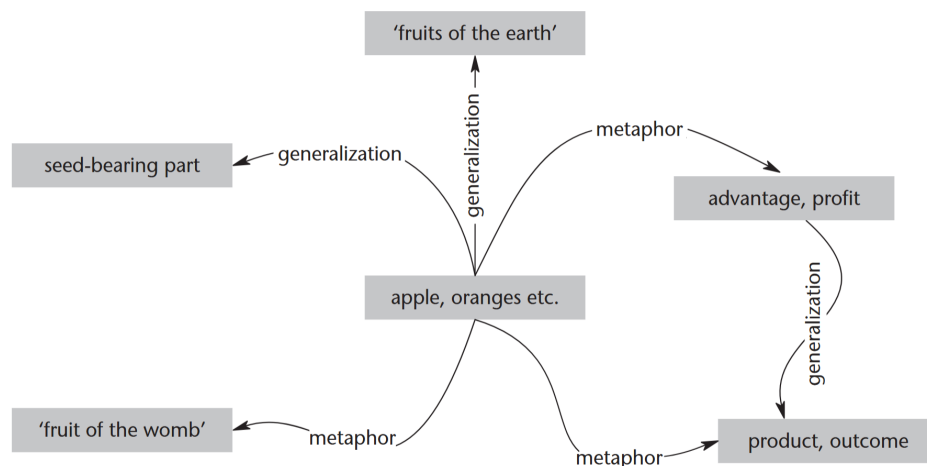


Figure 2.6: Radial network of the senses of *fruit*, from Geeraerts (2010)

2.2.6 Onomasiological Structures and Relations

This section discusses several relations and structures relevant for an onomasiological study, such as synonymy (central to the onomasiological approach), hyponymy and taxononmy.

Synonymy. The relation of synonymy, crucial for this thesis, is loosely defined as **identity in meaning**. More precisely, Cruse (1986) describes synonymy as a scale. Absolute synonymy is the complete and strict identity of all meanings of two or more terms; cognitive synonymy is established between terms that have a sense of common meaning; and near-synonymy between words having readings of highly similar but not identical meaning. Absolute synonyms must

be perfectly interchangeable in all contexts without any modification of meaning or truth-conditions of the utterance in which they occur. **Absolute synonymy is ideal** but hardly achieved in practice, and Cruse discusses many examples of candidate pairs for absolute synonymy and find for each pair a set of contexts in which both terms could not readily be exchanged.

According to Cruse, absolute synonymy is just a theoretical end of the spectrum of synonymy. What most semanticists consider to be synonymy is actually **cognitive synonymy**. There are two main differences with absolute synonyms. First, cognitive synonyms are identity in meaning in at least some of their readings but not necessarily in all of them. Second, the readings on which cognitive synonymy is evaluated must have identical *descriptive* (or “propositional”, in Cruse’s terminology) meaning while they may differ in *connotational* (or “expressive”) meaning. The idea is that substituting one for the other does not alter the truth-conditions of an utterance and both refer to the same category in this context, even though one may convey a different emotion or be may inappropriate in some social settings. In terms of intension, cognitive synonyms have senses with equivalent definition if we ignore emotional and cultural markers. In terms of extension, they have senses that can be applied to nearly-identical sets of entities. For instance, *cop* and *policeman* are cognitive synonyms (they have the same descriptive meaning) even though they cannot always be used interchangeably, as the former is much more informal than the latter. In a concept-based view of meaning, cognitive synonyms are words that can instantiate the same concept in some contexts as each possesses a conventionalized sense pointing towards this concept.

Finally, **near-synonyms** (which Cruse calls “plesionyms”) are terms whose substitution results in minor changes to an utterance’s truth conditions (for instance *fog/mist*, *murder/execute*). To identify near-synonyms, Cruse proposes to rely on the acceptability of using the discursive correction phrases *more exactly* and *not exactly*. For example, in *he was murdered — or, more exactly, he was executed*, the pair *murder/execute* qualifies as near-synonyms as they could be exchanged and the sentence would still be judged semantically valid. In contrast, a pair of hyper-/hyponyms would not yield a valid correction (e.g. in *he was killed — or, more exactly, he was murdered*, *kill* and *murder* could not be exchanged). In a concept-based framework, near-synonyms may refer to the same concept but emphasize different features of the corresponding category; or they may refer to different but highly similar concepts. It is generally acknowledged that, on the continuous scale of synonymy, the distinction between cognitive synonyms and near-synonyms is often clear, whereas the boundary between near-synonymy

and non-synonymy is blurry.

A more comprehensive discussion on cognitive synonymy, how they are defined and how they can differ, can be found in Stanojevic (2009). In the remainder of this thesis, unless explicitly stated otherwise, the use of “synonymy” alone should be understood as a shorthand for “cognitive synonymy”.

Hyponymy and taxonomy. The notion of hyponymy is normally conceived in terms of class inclusion. Hyponymy is the inclusion of a conceptual category within another one. Intensionally, if the definition of concept A is a specification of the definition of concept B, then we say that ‘A is an hyponym of B’. Extensionally, the category associated with concept A is included in the category associated with concept B. The converse relation is called ‘hypernymy’ (or sometimes ‘hyperonymy’). The general concept is the **hypernym**, and the specific, subordinate concept is the **hyponym**. This terminology also applies to lexemes referring to these categories. Of course, the relation is not exclusive, as the hypernym likely have multiple hyponyms. The relation is also transitive: if A is hyponym of concept B and B is hyponym of another concept C, then A is also an hyponym of C. Figure 2.7 provides two examples of the hyponymy/hypernymy. Both FINCH and CROW are hyponym of BIRD, which is itself hyponym of ANIMAL. As a result of transitivity, ANIMAL is also an hypernym of both CROW and FINCH.

With hyponymy, one can build a hierarchical structure of meanings called a **taxonomy**, such as the two presented in Figure 2.7. The taxonomical organization reveals that there exist multiple levels in this structure. Items at the same level and hyponym to the same superordinate item are said to be **co-hyponyms**. In our example, FINCH and CROW are co-hyponyms of BIRD, but the same could not be said for the pairs CROW and HOODED CROW for BIRD, or MINISKIRT and SKIRT for GARMENT, because the paired items are not at the same hierarchical level in the taxonomical structure.

Cruse (2002) establishes that taxonomies should be based on “is a kind of” hyponymy only (thus referred as “taxonomical hyponymy”, e.g. CROW is a kind of BIRD) and not on simple cases of “is a” (the “non-taxonomical hyponymy”, e.g. a KITTEN is a CAT). Thus in such a taxonomy, KITTEN would be reported under YOUNG MAMMAL for instance and not under CAT.

In lexicography, hyponymy has a central role in defining meanings; a word’s sense can be defined using the most direct hypernym term, followed by specific characteristics that distinguish it from other co-hyponyms. The clothing-related reading of *skirt* can be defined as being “a garment hanging from the waist, with a unique opening for both legs”, as the concept of GARMENT is the most direct

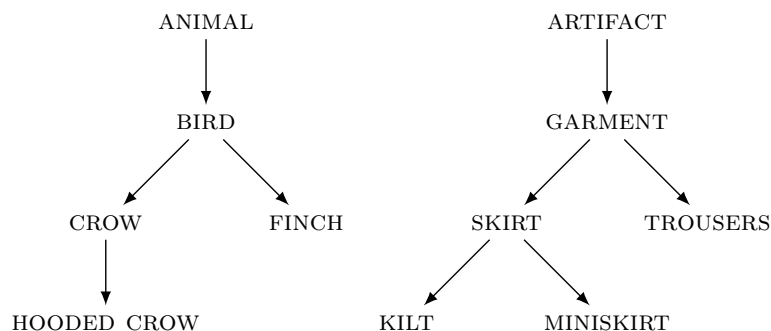


Figure 2.7: Example of a taxonomy of hyponyms (subordinates) and hypernyms (superordinates).

hypernym of this sense of *skirt*.

An Example of a Taxonomically Organized Lexical Database: WordNet. WordNet is a large-scale, taxonomically organized lexical database of Modern English developed at Princeton University under the direction of George A. Miller (Miller, 1995; Fellbaum, 1998) which models the organization of the lexicon by grouping words into sets of cognitive synonyms, called *synsets* (short for ‘synonym sets’). Every synset contains one or more lemmas that are interchangeable in at least some contexts without altering the propositional meaning of an utterance. Synsets are further described with a definition of the concept, and contextual examples. Synsets are organized in a taxonomy and linked to one another by the relation of hyponymy/hypernymy. Thus, WordNet describes a large tree of concepts, arranged from superordinate concepts to most subordinate ones. It can be seen as both a semasiological dictionary and an onomasiological dictionary, as WordNet can be queried in two fashions: a lemma as input, which will result in the list of synsets in which the lemma appears; or a synset as input, which will result in the list of lemmas associated to this concept. On the highest level of WordNet’s taxonomical tree are *Unique Beginners* (sometimes called “supersenses”), which are very large conceptual categories, such as ANIMAL, PLANT, ENTITY, etc. An abstract node, superordinate to all Unique Beginners, is often considered when comparing concepts in WordNet. WordNet distinguishes between four main lexical categories—nouns, verbs, adjectives, and adverbs—each forming its own relational network. Beyond hyponymy, WordNet models other lexical relations between lemmas, such as antonymy (oppositeness of meaning as in *ascend/descend*), or between synsets (such as holonymy, the part-whole relation as in ARM/BODY), but these relations fall outside the scope of the present thesis. Note that for adjectives and adverbs, WordNet does not represent hypernymy and only models other relations. WordNet is purely synchronic (i.e. it

does not model temporal variations of meaning) and is extensively used in NLP for its valuable descriptors of language and form-meaning associations. WordNet is accompanied with a synchronic corpus called *SemCor*³ ('Semantic Correspondances'), which consists of 352 texts from the Brown Corpus⁴ and annotated with WordNet synsets, as well as part-of-speech tags.

In this thesis, we make use of WordNet as a reference taxonomy for English. In Chapter 4, it allows us to evaluate a model's ability to describe the mapping between words and concepts learned from data (more precisely, from the *SemCor* corpus). In Chapter 5, it also us to curate a dataset of labeled occurrence pairs to supervise the fine-tuning of a Contextualized Language Models, in order to obtain better in-context representations of lexical meaning. In Sectin 7.3 of Chapter 7, we use WordNet to obtain a list of synonyms for modern-day English.

2.3 Lexical Semantic Change

This section adopts a diachronic perspective and focuses on the study of lexical semantic change over time. Three questions will be adressed:: *how* does word meaning change (mechanisms of change) ? *what kinds* of change are observed (nature of change) ? and *what regularities* does semantic change follow (patterns of change) ? Section 2.3.1 defines Lexical Semantic Change in terms of the mapping between lexemes and concepts. Section 2.3.2 examines several mechanisms that can create new meanings or alter existing ones. Section 2.3.3 reviews the exploration of several proposed classifications of lexical semantic changes. Then in Section 2.3.4 we discuss the various patterns and regularities that have been uncovered or hypothesized about semantic change. Finally, Section 2.3.5 presents available computational resources and diachronic corpora available for research in this area.

2.3.1 Lexical Semantic Change in the Word-Meaning Mapping

Over time, the mapping presented in Section 2.2.3 is subject to change: some lexeme-concept associations may be created, while others may be lost. Lexical semantic change corresponds precisely to such changes in the mapping between words and the concepts they denote.

Semasiological and onomasiological change. A **semasiological change** refers to an evolution of the set of meanings a given word has; that is, if the word gains

³<https://www.sketchengine.eu/semcor-annotated-corpus/>

⁴<https://www.sketchengine.eu/brown-corpus/>

or loses a sense over time. New senses can be the result of emerging concepts requiring a linguistic expression, and sense loss can be provoked by the obsolescence of a concept; both processes can also simply result from a change in the way words are allocated to persistent concepts. Conversely, **onomasiological change** refers to changes in the way concepts are named over time, rather than changes in the meanings of the words themselves. It is the study of how different words are used to refer to the same concept in a language, and how these naming patterns evolve historically. Onomasiological change notably include the appearance or the disappearance of associations between words and this concept. In the former case, the naming is expanded by one or several new words; and in the latter case, it is shrunk by one or more words. In addition, new concepts may appear and obsolete concepts may disappear, creating new associations or cutting existing associations in the word-concept mapping.

Notations. In the remainder of this chapter, we will have to provide examples of semasiological changes. We will denote *lexeme* X , $\text{CONCEPT } A > \text{CONCEPT } B$ the fact that the word X has shifted in meaning from $\text{CONCEPT } A$ to $\text{CONCEPT } B$. In other words, X has gained the meaning of $\text{CONCEPT } B$ but simultaneously it has lost the meaning of $\text{CONCEPT } A$. When the new meaning $\text{CONCEPT } B$ does not replace an existing meaning, we will denote the change with *lexeme* X , $+ > \text{CONCEPT } B$. Similarly, for onomasiological changes, we will denote $\text{CONCEPT } X$ *lexeme* $A > \text{lexeme } B$ the fact that the concept x used to be refer to with *lexeme* A and is later referred to with *lexeme* B . In other words, *lexeme* B entered the naming of $\text{CONCEPT } X$ while *lexeme* A left its naming. If *lexeme* B does not replace a word previously existing in the naming of $\text{CONCEPT } X$, we will use $\text{CONCEPT } X + > \text{lexeme } B$. On the contrary, if *lexeme* A is removed from the naming of $\text{CONCEPT } X$ without replacement, we denote it $\text{CONCEPT } X / > \text{lexeme } A$.

Concept drifts. While the notion of ‘lexical semantic change’ is traditionally employed to refer to changes in the interaction between the formal layer and the conceptual layer (that is, between words and concepts) as presented in Figure 2.2, there can also be diachronic changes as to the mapping between the conceptual layer and the denotational layer. Particularly, such changes may happen when a concept is perceived differently by speakers (e.g. in terms of emotional value) or because of changes at the level of intension (i.e. the very definition of a concept has changed). Such changes are called ‘concept drifts’ or ‘conceptual changes’, to distinguish them from lexical semantic change (Fokkens et al., 2016). Concept drifts fall outside the scope of this manuscript, but in terms of computational studies we can notably cite Sommerauer and Fokkens (2019) for a

computational approach to the change of the concept of RACISM, and Vylomova et al. (2019) for a study of the drifts in harm-related concepts in psychology (such as ADDICTION, HARASSMENT, TRAUMA, etc.).

2.3.2 Development and Loss of Word-Meaning Associations

Sense salience and peripheral changes. In Section 2.2.5 we mentioned that, in the semasiological perspective, word senses exhibit prototypicality effects. Investigating the nature of semasiological meanings within the framework of his *diachronic prototype theory*, Geeraerts (1997) showed that the prototypical effects governing senses have consequences in their diachronic evolution.

First, Geeraerts found that new meanings emerge as modulations of a core case; that is, if a lexeme has a prototypical meaning whose referents exhibit features ABCDE, then arising meanings will likely have referents with feature sets like ABCD, ABCE, ABDE, etc. before new meanings associated to features ABC can develop. Second, semasiological change involves a change in the prototypical structure of meanings: senses overlap and intertwine, and a new reading may emerge from multiple already conventionalized readings. Finally, new senses likely emerges at the periphery of the radial polysemy network, and core meanings are likely to remain very stable over time while peripheral meanings may disappear quickly.

Geeraerts (1997) also examined *semantic polygenesis*, a phenomenon whereby some meanings of a given lexeme may come into existence at several points in time, on an independent basis, and quickly disappear. Semantic polygenesis occurs with extremely peripheral meanings, that are semantically close enough to be spontaneously inferred by speakers yet too distant from the core meanings to become stable or long-lasting. This phenomenon illustrates the vagueness of the boundaries of a word meaning, which is an important property of prototypicality effects.

Conceptual Metaphor and Metonymy. Metaphor and metonymy are among the most productive sources of modulation of word meaning, and thus of lexical semantic change. Metaphor is usually described as based on *similarity in meaning* between two items, while metonymy is said to be based on *contiguity or association in meaning* between items (Ullmann, 1962).

Cognitive semantics research has shed new light on the study of metaphors, in particular through the framework of the *Conceptual Metaphor Theory* (Lakoff and Johnson, 1980). In this approach, metaphors are described as systematic

patterns of associations between a *source domain* and a *target domain*, creating a mapping between elements of the source domain *X* towards analogous elements in the target domain *Y*. For instance, in the metaphor ARGUMENT IS WAR, claims may be *attacked* are *defended* (like a strategic point); criticisms *strike true*; you will *defend* your thesis using an argumentative *strategy*; an eloquent opponent can *demolish* your arguments; and the debate can be *won*. The structure of the source domain is projected to the target domain – e.g. in the WAR domain, ‘*winning the war*’ is the result of ‘*attacking the opponent’s forces*’; in the ARGUMENT domain ‘*winning the argument*’ being the result of successfully countering of the opponent’s claims. Mappings are not exhaustive, some features of the source domain may not be applicable to the target. Also, metaphors are not unique with respect to the target domain. For instance, ARGUMENT IS BUILDING also conceptualizes argumentation as a constructed entity (e.g. claims need *support* otherwise they would *collapse*).

According to Lakoff and Johnson (1980), metonymies are similar to metaphors in many aspects but they involve only one conceptual domain instead of two (e.g. “crocodile handbag” where *crocodile* stays within the animal domain but the focus shifts to one of its salient parts or properties: its skin). Peirsman and Geeraerts (2006) rely on a prototypicality-based approach to classify and compare various recurrent types of metonymy: PART-WHOLE, CONTAINER-CONTAINED, PRODUCER-PRODUCT, ACTION-PARTICIPANT, etc.

Other models of semasiological change. The *Invited Inference Theory of Semantic Change* (Traugott and Dasher, 2002) has gained acceptance in the recent years as a pragmatics-based model to describe the emergence of new meanings. Conceptually, they distinguish three stages for a new meaning to emerge: (i) *utterance-token meaning*, innovative use of a lexeme that does not correspond to conventionalized meanings but can be understood by the audience in a specific context; (ii) *utterance-type meaning*, after a repeated use of the utterance-token meaning, and (iii) *coded meaning*, after the utterance-type meaning has been crystallized enough to gain autonomy and can be invoked out of context without ambiguity.

The primary goal of the Invited Inference Theory of Semantic Change was to explain a process called *grammaticalization* (Bybee, 2003; Lehmann, 2015; Hopper and Traugott, 2003), the gradual change in meaning of a content word (i.e. a lexeme of autonomous meaning) to a grammatical word (a lexeme whose semantic contribution lies in a grammatical function). Grammaticalization is a process through which a lexeme gains a new meaning (specifically, a functional

meaning) and often loses its previous meanings.

The Invited Inference Theory and Grammaticalization fall outside the scope of the present work, and have generally lead to few computational studies in the recent years. Their investigation, especially with recent NLP representation methods, is left to future work.

2.3.3 Classification of Semasiological Changes

Over the past centuries, research from historical-philological works to cognitive semantics studies has explored optimal ways to classify the various instances of lexical semantic change into categories that best capture regular patterns of development.

Our summary of the proposed typologies is based on the classifications proposed by Bréal (1897) (reprinted with updated definitions in Traugott (2017)), Stern (1931), Ullmann (1962) and Blank (1997). Common to most typologies of lexical semantic change are the following categories:

1. **Metaphorization:** A lexeme is innovatively used to express one thing in terms of another, based on the perceived similarity between the two things. For instance, the word *mouse* acquired the meaning of COMPUTER DEVICE TO POINT AND CLICK because the shape of these devices resembles that of the small rodents usually referred to as *mouse*.
2. **Metonymization:** A lexeme, originally used for a given category, is innovatively applied to a new category because of a conceptual relation between the two categories (e.g. *board*, TABLE > COUNCIL OF PEOPLE).
3. **Ellipsis:** A lexeme acquires a meaning that encompasses the whole semantic content of a multi-word expression of which it was a component. For instance, *private* developed the meaning of COMMON SOLDIER from the complete phrase ‘private soldier’. It was originally a simple adjective, and has become an autonomous noun for this meaning.
4. **Popular etymology** (or folk-etymology): A lexeme receives a new meaning from another lexeme due to their formal similarity. As an example, *hangnail* lost its meaning of ‘corn on the foot’ after being reinterpreted as ‘a bit of skin hanging loose at the side or root of a fingernail’ (based on its similarity with the combination of *hang* and *nail*, despite its true etymology actually being *agnail* in Middle English).
5. **Generalization** (or *broadening* or *extension of range*): A lexeme acquires a more general meaning than its original one. For instance, the meaning

of the French word *panier* ‘basket’ is an extension of its Latin ancestor *panarium* ‘bread basket’.

6. **Specialization** (or *narrowing* or *restriction of range*): A lexeme acquires a more specific meaning than its older meaning. For instance, the meaning of the French word *viande* changed from FOOD to MEAT around the 17th century.

Some typologies, like that of Bréal (1897) or of Ullmann (1962), account for **amelioration** and **pejoration**. They correspond respectively to cases in which a new meaning acquires a more positive or more negative emotional value than the original meaning. The word *disease* underwent pejoration (DISCOMFORT > ILLNESS), and *nice* underwent amelioration (LASCIVIOUS > KIND, PLEASANT). Many pejorations originate from *euphemisms*: speakers use a less-negatively connotated lexeme to refer to a negative concept; and over time the lexeme’s meaning ends up bearing the negative weight (which is the case of *disease* for instance). A number of ameliorations do not necessarily lead to a strictly positive meaning, sometimes amelioration of meaning merely cancels out the negative weight (e.g. French *regretter*, LAMENT OVER THE DEAD > MISS, REGRET).

Blank (1997, 1999)’s classification additionally accounts for other phenomena:

- **Co-hyponymous transfer**: A lexeme gains a new sense that is at the same taxonomical level than one of its existing sense, i.e. they share a common close hypernym (Occitan *bòrni*, ONE-EYED, can be used to mean BLIND).
- **Auto-antonymy**: A lexeme develops a new meaning that is opposite to one of its existing meaning (e.g. *awful*, AWE-INSPIRING > VERY BAD).
- **Antiphrasis** : A lexeme acquires a new meaning based on a contrastive aspect of the concepts, often through ironic or euphemistic usage (e.g. Old French *oste*, GUEST, gained the meaning of HOSTAGE).
- **Auto-converse**: A lexeme that used to lexicalize one extreme of a relationship comes to mean the opposite extreme too (e.g. *take* in the dialectal use as GIVE).

Traugott (2017) argues that pairs of opposite phenomena, such as amelioration/pejoration or generalization/specialization, tend to obfuscate the discussion on the regularities and patterns of lexical semantic change and may discourage the research on directionality of semantic change.

2.3.4 Case Studies and Regularities

In this section, we will review several patterns of Lexical Semantic Change that have been studied in the literature. While the previous section focused on typologies and *what kind* of change there might be, this section rather focuses on *when* change occurs and *which kind of meanings* are affected. In contrast with the last two sections that focused on more *semasiological* aspects of change, the present section enumerates several patterns that have been uncovered thanks to an *onomasiological* approach.

Change in semantic fields. Analyses of semantic changes through the lens of semantic fields (i.e. sets of words and concepts corresponding to related meanings) were very prolific in uncovering regular patterns of lexical semantic change. Trier (1934) describes how the lexical field of ‘knowledge’ was described by different lexical items over time in German. During medieval times, the field was divided between *Kunst* (knowledge and skills of the courtly knight) and *List* (practical knowledge) while the word *Wisheit* was a general term for all kinds of knowledge, encompassing the former two. Later, *Wisheit* underwent specialization and came to refer to religious knowledge. *List* changed in meaning and was replaced by *Wizzen* which, jointly with *Kunst*, came to refer to non-religious knowledge, without social class distinction. In modern German, *Wissen* means KNOWLEDGE of any kind, *Kunst* has the meaning of ARTS and *Weisheit* means WISDOM.

In the field of SPEED, Stern (1931) studies the development of English words for IMMEDIATELY, showing that most of them come from terms associated to RAPIDLY instead. This change, from the open-ended RAPIDLY to the half-limited IMMEDIATELY were also observed by Buck (1949) in a large cross-linguistic study. Buck argues that this change is unidirectional, i.e. that the change from IMMEDIATELY to RAPIDLY would have been highly unlikely.

The ‘concrete to abstract’ paradigm. Kronasser (1952) posits the general principle that terms associated to concrete meanings are likely to develop new, abstract meanings; and that this change from concrete to abstract is unidirectional. Numerous subsequent studies have verified the CONCRETE>ABSTRACT directionality. Derrig (1978) investigated metaphors based on color terms applied to human qualities, finding that terms associated to HUE often gain meanings of emotional or moral qualities (e.g. *blue*, +> MELANCHOLY, *white* +> INNOCENT) while lexemes associated to BRIGHTNESS are likely associated to intellectual qualities (e.g. *bright* +> CLEVER; *clear* +> OF KEEN MIND, ALERT, *dark*

+ > IGNORANT). Derrig (1978); Sweetser (1990) showed that PERCEPTION lexemes are often extended through metaphorization to the domain of INTELLECT (e.g. *see* to mean UNDERSTAND). Traugott (1978) highlighted that lexemes related to SPACE are regularly applied in the more abstract domain of TIME (e.g. *before*, IN FRONT OF > AT A TIME EARLIER). Clark and Clark (1978) showed that the SPACE domain is also a source of lexemes for the MENTAL ACTS domain (e.g. *follow* + > UNDERSTAND A DISCOURSE; the word *deduce* coming from Latin word *ducere*, LEAD). Klepanski (1990) studied a number of regular FOOD > HUMAN BEINGS WITH PEJORATIVE CONNOTATIONS patterns (e.g. Polish *grzyb*, MUSHROOM + > OLD, UNPLEASANT MAN). Lehrer (1985) provides many examples of ANIMAL nouns applied to human beings for negatively connotated behavior, such as *snake* to mean TREACHEROUS PERSON. Kahr (1975) discussed how body parts can be used for space-related concepts (as in ‘the *foot* of the mountain’), which is in line with Langacker (1987)’s idea that the human body and the sensory experiences are the most concrete concepts.

All of these examples illustrate that cross-domain lexical semantic change usually takes place in a specific direction, from concrete concepts to abstract concepts. For most of these changes, we have provided a semasiological example as it is the most straightforward way to conceive lexical semantic change. However, an onomasiological approach is required to reach a better understanding of the underlying regular patterns: comparing concrete and abstract conceptual domains, researchers have noted that the intersection between the namings of concepts within one concrete domain and the namings of concepts within a second, abstract domain grows larger over time.

While most of the examples discussed here are drawn from English, the vast majority of these studies were not limited to English and often included non Indo-European languages, strengthening the idea of a universal directionality for cross-domain lexical semantic change.

To study lexical semantic change in a given semantic field or regular transfers across conceptual domains, computational studies need expert knowledge to determine the set of words that relate to concepts in this field (e.g. Baes et al. (2023) for mental-health related concepts, Paccosi et al. (2023) for the olfactory field, Boholm et al. (2024) with political dogwhistles) and a categorization of concepts into domains. This requirement restricts the scale at which computational studies of semantic fields and domains can be conducted. In our research, we choose to adopt a large-scale approach, not focusing on a specific semantic field / pair of domains, and we do not make use of such external knowledge.

Allocation, conventionality, contrast. Bréal (1897) proposed several laws of lexical semantic change. The first one is the *loi de répartition* ('law of reallocation'), which we quote as follows:

Nous appelons « répartition » l'ordre intentionnel par suite duquel des mots qui devraient être synonymes, et qui l'étaient en effet, ont pris cependant des sens différents et ne peuvent plus s'employer l'un pour l'autre. (...) [Le peuple] admet l'existence d'une répartition : il ne croit pas qu'il y ait dans le langage des termes absolument identiques. (...) [S]on opinion qu'il n'y a pas de synonymes fait qu'en réalité les synonymes n'existent pas longtemps : ou bien ils se différencient, ou bien l'un des deux termes disparaît.

We call "reallocation" the deliberate order from which words that should be synonyms, and were actually, have adopted different meanings and can no longer be used interchangeably. (...) People acknowledge the existence of allocation: they do not believe that there are terms in the language that are absolutely identical. (...) Their opinion that there are no synonyms induces that synonyms do not exist for long: either they become differentiated, or one of the two terms disappears.

Essai de Sémantique, Bréal 1897

According to Bréal, when words happen to be synonymous, this synonymy cannot last: only one of the terms should remain or the terms must differentiate in meaning (which leads Xu and Kemp (2015) to refer to this principle as the *law of differentiation*). An example of Reallocation in French is GARDEN /> *courtil*: the word *courtil* ceased to be used after *jardin* entered the naming of GARDEN.

This principle concerning the evolution of synonyms was later supported by the works of Eve Clark in the field of pragmatics and language acquisition in children (Clark and Clark, 1978; Clark, 1993, 2007). Clark's work revolves around two fundamental pragmatic principles: *contrast* and *conventionality*. The Principle of Contrast is defined as follows: "*Speakers take every difference in form to mark a difference in meaning.*" (Clark, 1993). Thus, Bréal's law of reallocation and Clark's principle of contrast both dictates that two different lexemes should not have the same meaning, even if they share some semantic properties. A difference should always exist, and if such a divergence is temporarily absent it is expected to emerge over time. On the other hand, Clark (1993) also defines the principle of Conventionality, defined as a corollary to Contrast: "*For certain meanings, there is a form that speakers expect to be used in the language community*". Combining contrast and conventionality: two lexemes should have different meanings, and a meaning should be expressed by a only one word. From this perspective, there is no room for synonymy.

Synchronically, when synonymy still occurs, the Principle of Conventionality implies that "*speakers typically make use of only one term out of the synonymous set*"

(Clark, 1993). Diachronically, synonyms must differentiate according to Bréal's law of allocation, and to the Principle of Contrast. These views are complementary : Bréal's Differentiation ('forms of near-identical meaning diverge over time') is the process, encouraged by the Principle of Contrast ('two forms must make two meanings'), that results in the Principle of Conventionality ('one form for one meaning').

The Law of Reallocation is, by definition, an onomasiological change: the naming of a concept encompassing several distinct lexemes tend to shrink down to only one word.

Associations and change. Revisiting Stern (1921)'s work on *swiftly*, *quickly* and a number of related lexemes, Lehrer (1985) points out that Stern's findings – that these words *all* changed in meaning, from RAPIDLY to QUICKLY, and that their changed caused the semantic development of related words such as associated adjectives – support a different hypothesis than differentiation. Her claim is that “semantically related words may undergo parallel semantic changes”, or in other words, that semantic relationships tend to remain stable, even when the meaning of individual words changes. Building on Stern's analysis of *swiftly* and synonyms, Lehrer studies a variety of animal metaphors. She shows that the development of a metaphorical meaning for one lexeme consistently entails the emergence of the same meaning for its (near-)synonyms. For instance, *snake*'s more recent meaning of TREACHEROUS PERSON has been attested from 1590, after its near-synonym *viper* gained this new meaning in 1526. The word *viper* was introduced in English in 1526 by Tindale's translation of the Bible from the Old French, carrying both its literal meaning and its figurative meaning. Lehrer argues that it is the established association in *viper* between the animal and the human quality that enabled the development of this moral meaning for *snake*. Similarly, the meaning related to BRUTISHNESS first associated with *gorilla* later spreads to *baboon* and *ape*. By 1785, both lexemes *mule* and *donkey* had developed the metaphorical meaning of SILLY PERSON. In the case of *donkey*, this meaning was extended to OBSTINATE PERSON, which facilitated the same change for *mule*.

It remains unclear whether synonymous associations are doomed to disappear, as suggested by Bréal and Clark, or if they are to be preserved over time, as claimed by Lehrer. More precisely, both apparently opposite phenomena are at work in the whole lexicon, but the conditions in which each of them realizes remain to be determined. Furthermore, they need not be mutually exclusive: Stern (1921) observed that the parallel change RAPIDLY > IMMEDIATELY of *swiftly* and

all of its synonyms was likely caused by the emergence of *faste* (Old English for *fast*) as a concurrent term meaning RAPIDLY. Indeed, Stern also describes how *fast* gradually developed its new meaning from FIRM, initially through constructions like *run fast* ('to run steadily', which became 'to run rapidly'), opening the way for *fast* thereby acquiring the new coded meaning of RAPIDLY. In Chapter 7, we take this alleged contradiction as a case study and we conduct a data-driven investigation using various kinds of computational approaches to quantify these two phenomena and discuss the extent to which the proposed models can characterize them.

2.3.5 Resources and Corpora

In the previous section, we reviewed a wide range of observations and theories on Lexical Semantic Change developed within linguistic research. Before turning to the next chapter and discuss how Natural Language Processing addresses questions of lexical semantics and meaning evolution over time, we will first present several crucial resources at the basis of many computational and non-computational studies of Lexical Semantic Change.

Historical Thesaurus. The Historical Thesaurus (Kay et al., 2020) is a historical taxonomical database of English maintained and published by the University of Glasgow. Its structure is similar to that of WordNet and traditional thesaurus, with concepts being organized in a tree structure in which categories become increasingly specific at deeper levels of classifications, each concept is associated with a set of lexemes that are used or have been used to refer to this category. For each word belonging a given concept, the Thesaurus provides the time period in which this lexeme was associated to this meaning. The thesaurus covers word forms from Old English to present-day English. Some categories are supplemented with subcategories, which do not represent an additional level in the main taxonomical structure but are local subordinates to a specific categories. While the Historical Thesaurus is available online at <https://ht.ac.uk/category/> for manual inspection, its API is not publicly available and data download is not permitted, preventing a systematic use of the Historical Thesaurus in large-scale computational studies.

CLICS³. The *Database of Cross-Linguistic Colexifications*, abbreviated as CLICS (Rzymiski et al., 2020), is a multi-lingual database useful to investigate relations between concepts and regular cross-domain associations. It is available at <https://clics.clld.org/>. The base material of this resource are *colexification* data. The term 'colexification' as defined by Haspelmath (2023) refers

to the phenomenon whereby two meanings are expressed by the same lexeme in a given language (a polysemous lexeme). The database is usually explored through *semantic maps*, which are graphs in which nodes represent concepts and edges indicate the number of languages in which both concepts are colexified. Thus, two concepts are strongly associated if they are encoded by the same lexeme in many languages. CLICS contains colexification data from a wide variety of modern day languages, as well as from older languages such as Old English, Old French, Gothic, Akkadian, Old Testament Hebrew, etc. This temporal diversity enables diachronic investigations of regular patterns of meaning change, in particular in an onomasiological perspective. For instance, Georgakopoulos and Polis (2021) used the CLICS data base to study the change of time-related lexemes in Ancient Greek and Ancient Egyptian.

Database of Semantic Shifts. In an attempt to build the largest collection of attested semasiological changes, the Database of Semantic Shifts (Zalizniak et al., 2012a) compiles 9,870 cross-domain shifts found in 44,667 realizations, across 2,183 language varieties. For instance, the change from LISTEN to OBEY, COMPLY has been attested in 241 realizations (e.g. in English I told you to do it but you don't listen to me anymore). Many entries specify the direction of change, with a left-to-right arrow if the change is only found in one direction, or a bidirectional arrow if the reverse change is also realized. Note that the database, unlike the Historical Thesaurus for instance, only list semasiological changes, i.e. development of new polysemies. The database is available at <https://datsemshift.ru/>.

Diachronic corpora. Computational studies of Lexical Semantic Change require textual data from different time periods. Corpora like the *Google Book Ngrams Corpus* (Michel et al., 2011; Lin et al., 2012)⁵, the *Corpus of Historical American English* (COHA, Davies 2012; Alatrash et al. 2020)⁶ or the *Penn-BFM Parsed Corpus of Historical French* (Martineau et al., 2021) and the *MCVF Corpus* Kroch and Santorini (2021)⁷, provide collections of texts and/or cooccurrence information for different time periods in a variety of languages. The Google Book corpus and the COHA are the most widely used corpora in English computational studies of lexical semantic change, but these data can be biased if the genre of texts is not balanced ; they can also be noisy because of OCR-induced error or sparsity for older time periods (Pechenick et al., 2015).

⁵ <https://storage.googleapis.com/books/ngrams/books/datasetv3.html>

⁶ <https://www.english-corpora.org/coha/>

⁷ <https://github.com/beatrice57/mcvf-plus-ppchf>

Of particular interest for the diachronic analysis conducted in this thesis is the PRESTO corpus (Blumenthal et al., 2017), available at <http://presto.ens-lyon.fr/>. This French diachronic corpus comprises 315 texts, spanning from 1501 to 1549. PRESTO has both syntactic annotations (Part-of-Speech tags) and morphological annotations, as word forms are annotated with the corresponding lemma. These morphological annotations, not common in large-scale diachronic corpora, are extremely valuable for tracking the semantic evolution of a lexeme and avoid the difficulties caused by orthographic variation over such an extended period. We will make use of PRESTO in experimental studies of the Chapter 6 and 7.

2.4 Summary

In this chapter, we have established a theoretical framework for introducing the contributions of this thesis. We have presented the notions of lexeme, occurrence and concept, word senses and concept namings, as well as polysemy and synonymy. Particular attention has been paid to the difference between semasiology and onomasiology in the mapping between lexemes and concepts. By examining synonymy and hyponymy, we have introduced taxonomies such as WordNet that provide a modelization of the word-meaning mapping and can be used as external resources in computational studies. Then, we have reviewed several studies on Lexical Semantic Change in historical linguistics. We have seen that Lexical Semantic Change can be seen as the evolution of word-meaning associations, with words gaining or losing senses (on the semasiological side) and concept namings expanding or shrinking (on the onomasiological side). We have questioned *how* meanings emerge or disappear, *what kind* of changes may occur and *what regularities* govern these changes. Finally, we have introduced several resources of interest for computational approaches to lexical semantic change.

Studies in historical linguistics are based on careful examination of examples of word usages over time, allowing researchers to precisely describe how associations between words and meanings evolve. The drawback of such methodology is its large cost in time and human resources, which limits the scale of such studies to a relatively small scale and requires a strong supervision of the analytical process. In some cases, proposed theories of semantic change are supported by a small number of examples and would benefit from a more systematic evaluation in large historical corpora. In the next chapter, we will see how Natural Language Processing addresses these issues and we review the most widely used models of lexical semantic change. Computational studies allow for large-scale,

automatic exploration of semantic change, by comparing vast amount of data over time. The computational approach is complementary to the linguistic approach, operating at a different level on the corpus data and allowing to uncover new regularities in lexical semantic change.

3

Computational Modelling of Lexical Semantic Change

In the previous chapters, we have outlined the dominant approaches to lexical semantics in general, and to lexical semantic change in particular, and we have argued that lexical semantic change can be seen as an evolution of the mapping between lexemes and semantic concepts. Natural Language Processing (NLP) has interacted with Linguistics since its earliest days and has proven a valuable tool for large-scale, systematic studies of language phenomena. In this chapter, we discuss the various challenges that NLP research has framed to describe the mapping between words and meanings, including the specific tasks that have been formulated to evaluate computational models of lexical semantic change. Then, we give an overview of the various approaches in the NLP literature to detect and model lexical semantic change. We provide a review of several computational studies that explored and uncovered patterns and regularities of lexical semantic change. We conclude the chapter by outlining the major limitations of current NLP approaches to lexical semantic change.

3.1 NLP Tasks in the Word-Concept Mapping

Research in Natural Language Processing has translated several challenges related to lexical semantics into different tasks. These tasks are usually set within two different perspectives. On the one hand, there are tasks that involve comparisons between the various usages of a target word. On the other hand, other tasks emphasize on comparisons between the meanings of different words. In this section, we describe some of the most widely studied tasks dealing with lexical meaning, and particularly polysemy and synonymy in the mapping between lexemes and concepts. We will notably distinguish two categories of tasks: *between-word* tasks focus on comparisons between words and on the degree to which speakers perceives them as (dis)similar in meaning (with synonymy be-

ing the maximal degree of similarity); and *within-word* tasks are dealing with polysemy and how the multiple senses of a word are reflected in its usages.

3.1.1 Between-word tasks

Among between-word tasks, we first describe **Lexical Substitution**. First introduced as a means to evaluate Word Sense Disambiguation system (see Section 3.1.2) by McCarthy (2002), Lexical Substitution aims to find and evaluate suitable substitutes for a target word in a given context. In the example below, a Lexical Substitution system should judge that *workers* could be a good substitute for *employees* in this context.

- (4) The factory employees have been on strike for days.

McCarthy (2002) discussed several evaluation settings for Lexical Substitution. McCarthy first proposed that candidate systems should pick a valid substitute – or a valid set of substitutes – among a pre-defined set of options, and the evaluation would assess whether the right answer was chosen. Alternatively, candidate systems can be tasked with providing a substitute on their own. In the former setting, the task is easier to evaluate because the set of possible answers is constrained, simplifying the comparison to ground truth; but it creates a bias towards systems that rely on the same lexical resource than the one used for training (e.g. a system using WordNet’s synsets on a benchmark that relied on WordNet to determine the possible answers.). The substitute-prediction setting is less unfair towards candidate systems that may be using a different lexical resources or no lexical resource at all (such as methods purely based on word embeddings), but is less trivial to obtain an exhaustive ground-truth from human annotators and to compute inter-annotators agreement.

Notable datasets for Lexical Substitution in English include the *SemEval-2007 dataset* (McCarthy and Navigli, 2007) and *CoInCo* (“Concepts in Context”, Kremer et al. 2014). Cholakov et al. (2014) introduced a lexical substitution dataset for German.

Another major between-word tasks is **Lexical Similarity**, the task of predicting how similar the meanings of two different words are. It is important to distinguish between *similarity* and *relatedness*. Similarity refers to the fact that two words can be exchanged because their meanings are close enough and that they may occupy the same role in the sentence. In contrast, relatedness indicates that the meaning of two words are related enough to co-occur.¹ Similarity is obviously

¹ Chronis and Erk (2020) makes the distinction as follows: “Similarity is paradigmatic: highly

not a binary feature between two words, no two words are perfectly similar nor entirely dissimilar. Thus, similarity is measured on a continuous scale. Lexical Similarity as a task is usually realised in two settings. On the one hand, we have *Out-of-context* (OOC) Lexical Similarity, where only pairs of words are presented and candidate systems must propose a similarity score to each word pair. For instance, the pair (*cat*, *dog*) would typically receive a higher similarity score than the pair (*paper*, *airplane*). On the other hand, we have *In-Context* (IC) Lexical Similarity, where systems are asked to judge the similarity of two words in provided contexts. The two words may be presented in the same sentence, like in example 5, or presented in separate sentences or contexts, like in example 6.

(5) **Shared context:**

The prisoner's cell was large enough that several pieces of furniture could fit in the room.

(6) **Distinct contexts:**

- a. The treatment to cure infected cells was working perfectly.
- b. They must make enough room in the kitchen to place a fridge, an oven and a counter.

In both the OOC and IC settings, evaluation is usually achieved by computing the correlation between gold-standard judgements of similarity in the pairs and predictions of the candidate system.

In-Context Lexical Similarity, despite being a between-word task by definition, allows within-lexeme considerations. Its advantage over its Out-Of-Context counterpart is that it requires systems to disambiguate word senses before making the similarity judgement between target words. Indeed, both *cell* and *room* exhibit polysemy in (5) and (6), with meanings SINGLE ROOM and SMALLEST AUTONOMOUS UNIT OF LIVING MATTER for *cell* and PART OF A BUILDING and UN-OCCUPIED SPACE for *room*. Thus, their meanings are very similar in (5) but very dissimilar in 6, and systems for the Lexical Similarity task should account for these variations of meaning.

Among popular Lexical Similarity datasets in the Out-Of-Context setting, we find the *WordSim353-Sim* dataset (Agirre et al. 2009), *SimLex-999* (Hill et al. 2015) and the *CAmbridge Rare word Dataset* (CARD-660, Pilehvar et al. 2018) for English, and *Multi-SimLex* for 12 diverse languages (Vulić et al. 2020). The

similar words are more likely to occupy the same 'slot' in a sentence (i.e., *The bug is on the rug/carpet*). Semantic relatedness, on the other hand is syntagmatic: two highly related words are likely to appear in succession, as in *She filled the car with gasoline.*"

most widely-used In-Context datasets include Stanford Contextual Word Similarity (SCWS, Huang et al. 2012) in English and CoSimLex (Armendariz et al., 2020) in 4 different languages. In Chapter 5, we use CoSimLex to evaluate a model’s ability to represent the lexical meaning of different words in context.

3.1.2 Within-word tasks

Word Sense Disambiguation (WSD) is a widely studied task (Navigli, 2009; Bevilacqua et al., 2021). From a machine-learning perspective, it is a supervised problem where candidate systems are tasked to assign the correct sense label to each occurrence of a word, chosen among a pre-defined set of available senses for this word. The particularity of this classification problem is that the set of classes is distinct for each word.

For instance, consider the following occurrences of *room*:

- (7) a. They must make enough room in the kitchen to place a frigde, an oven and a counter.
- b. Her conclusions leave no room for doubt.
- c. He wanted to decorate the room with pastel colors.
- d. The evaluation showed there was still room for improvement.

Knowing that the word *room* has meanings (A) UNOCCUPIED SPACE, (B) PART OF A BUILDING’S INSIDE and (C) SUITABLE OCCASION OR OPPORTUNITY, the system should label the occurrences presented in Example (7) as pertaining to senses A, C, B, and C respectively.

Raganato et al. (2017) proposed a unified framework that brings together most major evaluation datasets for English WSD: Senseval-2 (Edmonds and Cotton, 2001), Senseval-3 English All-Words (Snyder and Palmer, 2004), SemEval-2007 task 17 (Pradhan et al., 2007), SemEval-20013 task 12 (Navigli et al., 2013), SemEval-2015 task 13 (Moro and Navigli, 2015). Under this framework, all of these benchmarks use WordNet’s senses as ground-truth. More recently, Blevins et al. (2021) proposed an English dataset based on Wiktionary’s senses. Pasini et al. (2021) introduced XL-WSD that extends Raganato et al. (2017)’s framework to 18 other languages, Basque, Bulgarian, Catalan, Chinese, Croatian, Danish, Dutch, English, Estonian, French, Galician, German, Hungarian, Italian, Japanese, Korean, Slovenian, and Spanish from six families. XL-WSD’s sense inventories are drawn from BabelNet’s synsets, restricted to those containing at least one English WordNet sense. Most of the non-English dataset they gathered were annotated using English WordNet’s synsets too. This reliance on English

WordNet can lead to discrepancies in cases where a given language lexicalizes concepts that are not lexicalized in English or not found in WordNet.

Word Sense Disambiguation is a supervised task, meaning that candidate systems are provided a sense-annotated training corpus from which they learn how to map word occurrences to the appropriate senses. Derived from WSD, **Word Sense Induction** is the unsupervised task of clustering word occurrences into senses (Van de Cruys and Apidianaki, 2011; Navigli, 2012). All occurrences in the same cluster must be instances of the same sense, corresponding to one reading, without requiring the system to know the reading’s definition in advance. WSI systems are agnostic to sense inventories, they simply group occurrences of similar meaning together for each word.

In English, SemEval-2010 task 14 (Manandhar et al., 2010) and SemEval-2013 task 13 (Jurgens and Klapaftis, 2013) are the most widely used benchmarks for WSI. The former contains 8,915 occurrences for 100 distinct target words (50 nouns and 50 verbs), while the latter contains 4,664 occurrences for 50 target words (20 nouns, 20 verbs, 10 adjectives). To perform and evaluate multilingual WSI, in the absence of widely-known multilingual WSI benchmark, studies like Kokosinskii and Arefyev (2024) have relied on WSD datasets like XL-WSD. WSI is evaluated using clustering comparison metrics like Adjusted Rand Index (Hubert and Arabie, 1985) or BCubed F1 (Bagga and Baldwin, 1998) to compare gold-standard clusters to predicted clusters. Amigó et al. (2009) compared a large variety of clustering metrics and their statistical properties, showing that no one metric fits all constraints, yet that BCubed metrics satisfy the largest number of them; and proposed an extension of BCubed computation to account for “fuzzy” clusters (i.e. overlapping clusters where instances may be assigned to multiple clusters, with a membership probability for each). However, these metrics do not penalize over-specification of small senses (as long as most occurrences are in valid, large clusters) so Amrami and Goldberg (2019) proposed to also compute the correlation between the number of clusters found for each word with the gold numbers of senses.

The major limitations of WSD and WSI is that their evaluation rely on a annotated data and therefore on a pre-defined set of senses given by a reference lexical resource. As such their evaluation favors any systems that leverage this resources and discourages other attempts. Discussing this limitation in the context of WSD, McCarthy (2002) proposed shifting the evaluation of WSD systems to external applications, and suggested Lexical Substitution as a suitable extrinsic evaluation of WSD systems.

The **Word-in-Context** (WiC) task has more recently been defined by Pilehvar and

Camacho-Collados (2019) as a binary classification task where candidates must, given a pair of occurrences of a target word, predict whether these occurrences reflect the same word sense or not. Authors relate this to In-Context Lexical Similarity with distinct contexts (like the SCWS dataset), but emphasize that pairs in WiC contain occurrences from the same word only, thus shifting the focus from similarity between different lexemes to an individual word’s polysemy degree. WiC can be related to WSI: in WiC, pairs of occurrences are assigned to label 1 (“same sense”) if they would have been in the same cluster in WSI, and label 0 if they would have been in different clusters. However, it is easier to gather annotations for WiC than for WSI; one does not need large sense-annotated corpora with a pre-defined set of senses, but rather only judgements on whether two instances evoke the same meaning or not. Sense annotation can be a difficult task for human annotators, and either requires a pre-defined set of senses (i.e. previously discussed limitation), or requires annotators to be linguist experts. A pair judgement is easier to achieve for non-expert human annotators, and a large WiC dataset can be more readily derived from available resources like the Wiktionary despite their small amount of tagged occurrences per sense because the task is not on occurrences alone but on pairs. Another crucial difference between WSI and WiC is that the former is intended to be fully unsupervised, while the latter is supervised and provides by training data.

As a binary classification task, WiC is traditionally evaluated with *Accuracy* (or *Balanced Accuracy*, the average of Recalls for classes 0 and 1). The dataset provided in Pilehvar and Camacho-Collados (2019), called the *WiC dataset* (or sometimes *WiC-original*), contains 1,400 pairs of occurrences for 1,184 unique target lexemes in its *test* split, with 5,428 pairs of occurrences for 1,256 target lexemes in its *train* split. Only 36% of the target lexemes in the test set appear in the train set, which makes the task challenging for systems relying on naive pattern matching. Several datasets have been proposed subsequently in the multilingual setting.

3.1.3 Lexical Semantic Change Detection

The **Lexical Semantic Change Detection** (LSCD) task was introduced for the first time as SemEval-2020 Task 1 in Schlechtweg et al. (2020). It has become the most common form of evaluation for NLP models of Lexical Semantic Change since 2020, with several dedicated benchmarks in different languages. We first describe the task in itself, and then describe the standard procedure to create the benchmarks and the available benchmarks.

Definition of the LSCD Task. The most common setting for LSCD is *Graded Change Detection*. Given two sets of occurrences of a given target word extracted from two text corpora of different time periods, candidate systems must compare occurrences from the early period to occurrences of the later period and compute a change score that quantifies how much the word has changed in meaning between the two periods. The system’s predictions are evaluated by computing the Spearman correlation coefficient between predicted change scores and gold change scores, provided by the evaluation dataset. The other setting for LSCD is *Binary Change Detection*. In that case, system must predict whether a word has changed in meaning between the two periods, using provided sets of occurrences, as a binary signal: 1 (“the word underwent lexical semantic change”) or 0 (“the word’s meaning remained unchanged”). In a follow-up shared task, Zamora-Reina et al. (2022) decomposed the Binary Change Detection task into two subtasks: *Sense Gain Detection* and *Sense Loss Detection*, where systems must predict whether a word has, respectively, gained a new sense or lost a sense from period 1 to period 2.

Most datasets for LSCD are created using *Diachronic Word Usage Graphs* (DWUG, Schlechtweg et al. 2021), in which occurrences of a given word from different time periods are compared by human annotators to provide a gold-standard change score. In Table 3.1 are summarized the different DWUG datasets that have been created to evaluate unsupervised LSCD. **SemEval-2020** (Schlechtweg et al., 2020) covers 4 different languages, English, Swedish, Latin and German; **LSCDiscovery** (Zamora-Reina et al., 2022) is in Spanish; **RuSemShift** (Rodina and Kutuzov, 2020) is a russian dataset actually using three time periods and therefore offering three pairwise time comparisons; **NorDiaChange** (Kutuzov et al., 2022) similarly offers 2 benchmarks in Norwegian, one for changes on large periods and one for short-term changes; **ChiWUG** (Chen et al., 2023) brings DWUGs to Chinese, and the more recent **DWUGs-IT** (Cassotti et al., 2024a) introduced a smaller dataset for Italian. All of these datasets are available at <https://www.ims.uni-stuttgart.de/en/research/resources/experiment-data/wugs/>.

Lexical Semantic Change Detection, as formulated in the SemEval-2020 shared task, did not represent a new research problem: it was rather a standardisation of common research frameworks that existed in NLP research since the late 2000’s. Sagi et al. (2009) investigated the semantic change of three lexemes (*do*, *dog* and *deer*) from Middle English to Modern English, finding that early occurrences of *dog* and *do* exhibit more variations of contexts in late periods than in earlier periods (due to the generalization process that *dog* underwent

Name	Language	# targets	Periods	Source corpora
SemEval-2020	EN	46	1810-1860 ; 1960-2010	CCOHA
	SV	44	1790-1830 ; 1895-1903	Kubhist
	LA	40	200 BC. - 0 ; 0-2000	LatinISE
	DE	50	1800-1899 ; 1946-1990	T_1 : DTA, T_2 : BZ+ND
LSCDiscovery	ES	46	1810-1860 ; 1960-2010	CCOHA
RuSemShift	RU	111	1700-1916 ; 1918-1990	RNC
		111	1700-1916 ; 1992-2016	RNC
		111	1918-1990 ; 1992-2016	RNC
NorDiaChange	NO	40	1929-1965 ; 1970-2013	NBdigital
		40	1980-1990 ; 2012-2019	T_1 : NBdigital, T_2 : NAK
ChiWUG	ZH	40	1954-1978 ; 1979-2003	People’s Daily
DWUGs-IT	IT	26	1948-1970 ; 1990-2014	L’Unità

Table 3.1: Overview of the DWUG datasets in the different languages. SemEval-2020 (Schlechtweg et al., 2020) introduced 4 datasets in different languages. RuSemShift (Rodina and Kutuzov, 2020) offer three benchmarks for the same 111 target words across 3 time periods, with one benchmark for each pair of periods. NorDiaChange (Kutuzov et al., 2022) proposes two datasets with different periods, one with small time framed and the other with larger periods.

from a specific breed of dog to the whole category of DOG, and the grammaticalization process that the verb *do* underwent), while occurrences of *deer* were more varied in early periods than in later periods (*deer*’s meaning has shifted from SMALL WILD ANIMAL to the ANIMAL OF THE FAMILY CERVIDAE.) (Kim et al., 2014) used computational models to established a list of 10 words that changed the most in meaning between 1900 and 2010, and another list of 10 words that remained the most stable in meaning. Hamilton et al. (2016b) evaluated their computational model using a list of 10 words for which they knew, from previous literature, that underwent semantic change. As such, the LSCD task can be seen as an attempt to unify these research initiatives under a common framework. Note however that LSCD focuses on only two, disjointed and often not consecutive, time periods, while most of the NLP works before 2020 considered a higher number of time periods, and even continuous time indexing (Rosenfeld and Erk, 2018).

Most crucially, the nature of LSCD is *purely semasiological*, as it only evaluates a system’s ability to detect if a target word gained or lost senses. To the best of our knowledge, the existing NLP literature has not yet define a task that would evaluate the a system’s ability to model onomasiological lexical semantic change. In this thesis, we aim to build models of lexical meaning that account for both semasiological and onomasiological changes, and that go beyond the mere detec-

tion of these changes. Such model should provide additional information about the type of change that was observed. The approach proposed in Chapter 6 for instance can distinguish between sense gains and sense losses on the semasiological side, and naming expansions and naming shrinkages on the onomasiological side.

3.1.4 Novel Sense Detection and Definition Generation

Fedorova et al. (2024b) introduced the AXOLOTL’24 shared task, which features two subtasks related to Lexical Semantic Change. In Subtask 1, participants are provided with sets of occurrences of a target word for two different time periods (‘old’ and ‘new’), as well as a gold-standard sense inventory for the old time period. Participating systems are tasked to map occurrences in the ‘new’ set to the corresponding senses, and detect occurrences that belong to a newly gained sense, i.e. a sense that does not appear in the provided inventory. The task is similar to Word Sense Disambiguation in that occurrences are to be mapped to pre-defined senses, except that systems should be able to detect whenever an occurrence does not correspond to any of the provided senses and correspond to a novel one.

The task is evaluated with two metrics applied to the ‘new’ set: *Adjusted Rand Index*, which compares the candidate clustering of occurrences into senses with the gold sense partition; and the F_1 score on whether occurrences are labeled with the right sense or not. The latter metric is computed only for occurrences tagged with one of the provided senses and does not account for occurrences of novel senses.

In subtask 2, participants must generate the most accurate definition for the novel senses of each target word in the dataset. Importantly, temporal information plays no role in Subtask 2 as the goal is simply to generate a definition from occurrences of novel sense, regardless of other occurrences nor the time period these occurrence are drawn from.

The benchmark covers three languages (Finnish, Russian and German) with a *Train*, *Dev* and *Test* splits available for Finnish and Russian and only a *Test* set for German. The absence of *Train* and *Dev* splits for German is motivated by the task organizers’ desire to evaluate the multilingual transferability of candidate models on a surprise language.

While dealing with historical data and novel sense gain, one could argue that the task is hardly diachronic in nature. Subtask 1 can be seen as a variant of the

WSD/WSI tasks with a subtle difference². It is so subtle that the best performing model in the shared task is a straightforward WSD model with no novel-sense detection capabilities.

3.2 Computational Models of Lexical Semantic Change

In Section 3.1.3 we introduced the Lexical Semantic Change Detection task, the most widely used evaluation framework for computational models investigating Lexical Semantic Change. In this section, we elaborate on these models. Most of the methods presented below are rooted in the literature tackling the larger lexical semantic challenges presented in Section 3.1.

Computational models of lexical semantic change can be decomposed into three components. The first component is representations of lexical meaning, the basis of every computational study of LSC. The second component covers alignment techniques. Indeed, because studies of LSC span across different time periods, and thus across several sets of data, resulting representations may require to be aligned before any comparison can be made. The final component corresponds to how Lexical Semantic Change is measured and observed in diachronic data, given a representation of meaning. The choice of representations determines which alignment techniques are applicable, and how temporal comparisons can be made. This section is organized according to the kind of representations of meaning.

The core question of any computational approach to lexical semantic change is how should represent lexical meaning be represented for a given time period, given a corresponding text corpus? We will notably distinguish two families of approaches. The first family comprises *type-level* models (also called *static* models), where each word w in the vocabulary is associated to a single representation $\mathbf{r}^w$. The second family of approaches regroups *occurrence-level* models (or *contextualized* models), which assign a representation $\mathbf{r}_i^w$ for each occurrence $o_{w,i}$ of the lexeme w in the dataset, denoting $o_{w,i}$ the i -th occurrence of word w .

In the diachronic setting, the common practice is to compute representations for each time period; as such we will distinguish $\mathbf{r}_{(k)}^w$ the type-level representation of w at time T_k , and $\mathbf{r}_{(k),i}^w$ the representation of occurrence $o_{w,i}^{(k)}$, which is the i -th occurrence of w at time T_k . Most of these approaches are based on the so-called Distributional Hypothesis (Harris, 1954), grounded in the idea that “a word is characterized by the company it keeps” (Firth, 1957). Most popular models

² Fedorova et al. recognize “Taking classification of new examples with old senses into account made Subtask 1 more similar to pure WSD than LSCD”

of word meaning produce vector representations, whether they are probability distributions or dense vectors (*embeddings*).

Non-vector models include bayesian topic models (Lau et al., 2012; Frermann and Lapata, 2016; Cook et al., 2014), models based on comparisons of generated definitions (Giulianelli et al., 2023; Fedorova et al., 2024a) and ensemble-based methods such as *grammatical profiles* (Kutuzov et al., 2021; Giulianelli et al., 2022). Our research focuses on the use of (static or contextualized) word embeddings, thus these other approaches fall outside the scope of this section.

In the remainder of this section, we will first describe type-level vector methods, then occurrence-level vector methods.

3.2.1 Static vectors

With type-level representations, the standard procedure is to learn representations of a separate set of the target words for each time period, and compare them afterwards.

3.2.1.1 Representations

Co-occurrence matrices. Co-occurrence models, which were most popular in the 2000s and the early 2010s, are based on a straightforward application of the distributional hypothesis: the meaning of a word is represented by the distribution of other words it appears with. By comparing co-occurrence matrices from each time period, researchers can track how these distribution evolve over time for each word individually. In such matrices, rows correspond to target words and columns to context words. The key parameter of this method is the **window size**, which defines how many words are observed around a target word to get co-occurrence statistics.

However, raw co-occurrence matrices are rarely used as is, because they suffer from two major limitations: (i) they are very sparse, with a limited amount of useful information compared to their size and (ii) they are often normalized such that each line sums to one (to obtain a probability distribution of contexts for each target word), but this normalization erases the importance of more rare contexts that should still be accounted.

Co-occurrence matrices are often reweighted using *Pointwise Mutual Information* (PMI) or its truncated variant, *Positive Pointwise Mutual Information* (Pantel and Lin, 2002; Bullinaria and Levy, 2007). This reweighted mitigates the negative effect of normalization against rare contexts. The obtained sparse matrices are

often transformed into dense representations using Singular Value Decomposition (SVD) to reduce the dimensionality of the matrices to work with. (Turney and Pantel, 2010; Bullinaria and Levy, 2012).

In their study of temporal variations of meaning, Basile et al. (2014) and Caputo et al. (2015) relied on a different approach to obtain dense vector representation from co-occurrence matrices, using a method called **Random Indexing**. Each word of the vocabulary V is assigned to a random vector of d dimensions containing only zeros except at two randomly-chosen positions, which values are set to -1 and 1 respectively, obtaining a random matrix R of dimensions $|V| \times d$. The co-occurrence matrix $M_{(k)}$ is then multiplied to R to obtain dense representation matrix $M_{(k)}^{RI} = M_{(k)}R$. In this matrix, the representation of a target lexeme (row of the matrix) will be highly influenced by the random indexing of context words it often co-occurs with.

While popular in the late 2000s and the early 2010s, these matrix representations directly based on co-occurrence counts have been progressively abandoned in favor of neural vector models.

Neural vector models. With advancements in neural network, Word2Vec’s *Continuous Bag of Words* (CBOW, Mikolov et al. 2013a) and *Skip-Grams with Negative Sampling* (SGNS, Mikolov et al. 2013b) have been widely adopted in NLP lexical semantic studies to represent the meaning of a lexeme. In both cases, the model is a simple feed-forward neural network with an input layer, a single hidden layer and an output layer. In the case of CBOW, the model is optimized to maximize the probability distribution of a target word given its context. Skip-Grams models, in their original definition (Mikolov et al., 2013a), were optimized to maximize the probability distribution of context words given a target lexeme. In Mikolov et al. (2013b), authors further refined the model into SGNS by adding so-called “negative sampling” so that the model also has to distinguish between the target word and random noise examples.

In the study of Lexical Semantic Change, Gulordava and Baroni (2011) used PMI matrices; Sagi et al. (2009) used SVD reduction on co-occurrence matrices; Rodda et al. (2017) used SVD reduction on PPMI matrices. Neural embedding models based on Word2Vec have been extensively used in Kim et al. (2014); Kulkarni et al. (2015); Kutuzov et al. (2017); Rodina et al. (2019); Pražák et al. (2020); Zhou and Li (2020).

Hamilton et al. (2016b); Schlechtweg et al. (2019) compared results obtained using PPMI, SVD of PPMI matrices, and SGNS vectors. Schlechtweg et al. (2019) also tested the Random Indexing method. Overall, experimental comparisons

show that neural SGNS provide better representations of lexical meaning for the investigation of semantic change.

3.2.1.2 Temporal Alignment

Co-occurrence matrices and (P)PMI matrices from different time periods can be directly compared across time periods as long as the considered vocabulary for contexts is shared across periods (i.e. all matrices have the same amount of columns). However, dense representations produced by SVD, RI or neural models produce non-comparable spaces if they are run on different corpora, because their dimensions can be arbitrarily permuted with prevents the use of dimension-wise comparisons (like spatial distance, e.g. cosine distance). To enable comparisons of these representations over time, researchers rely either on second-order comparisons, or on alignment techniques that makes the spaces comparable.

For Random Indexing in the temporal setting, Caputo et al. (2015) show that using the same random indexing matrix R for the different time periods makes the transformation of co-occurrence matrices comparable, provided that the context vocabulary is shared. That is, if A is the count matrix for T_1 and B the matrix for T_2 , then $A^{RI} = AR$ and $B^{RI} = BR$ are comparable as they use the same R to project count matrices into a shared dense space.

For other dense models (SVD, CBOW and SGNS), Hamilton et al. (2016b) proposed using *Orthogonal Procrustes* to align a source space (from a later period) to a target space (an earlier period). Given two $n \times d$ representation matrices A and B which corresponds to two time periods, the Orthogonal Procrustes problem seeks a transition matrix $Q \in \mathcal{M}_{n,n}(\mathbb{R})$ such that

$$Q^* = \underset{Q \in \mathcal{M}_{n,n}(\mathbb{R})}{\operatorname{argmin}} \|QB - A\|_F \quad \text{with } Q^T Q = I_n.$$

Seeing Q^* as a transformation operator, this is a linear optimisation problem in which one want to minimize the distance between A and the transformed B . Placing the solution to that problem under the orthogonality constraint allows the preservation of distance between vectors of B . A general solution to this problem can be obtained through the SVD of AB^T , given by $AB^T = U\Sigma V^T$, with Σ diagonal with non-negative values. Then, the optimal solution for the Orthogonal Procrustes is $Q^* = UV^T$.

An alternative alignment technique, known as *incremental training*, was proposed by Kim et al. (2014). In this method, the vector model of period T_k is used as initialization of the model for period T_{k+1} , thereby maintaining temporal

consistency without explicit alignment.

Dynamic type-level models. A subsequent line of research has tried to adapt static representation methods to the specifics of a diachronic setting, integrating temporal information to the process of learning representations, thereby eliminating the need for alignment. Bamler and Mandt (2017) proposed a Bayesian version of the skip-gram model, in which embeddings evolve as a latent time series, constrained by the temporal ordering of documents (e.g. information about past documents is available but not about future documents). Yao et al. (2018) created an optimization to learn period-specific PPMI matrices all at once, jointly instead of separately. Rosenfeld and Erk (2018) proposed a model that integrates a SGNS model and a separate feed-forward neural network that integrates time as a variable, and output of both submodels are combined into a later integration unit that modulate the SGNS embedding with the time embedding. The system is therefore able to create period-specific word embeddings for every lexeme of the vocabulary and every period in which they occur. Dubossarsky et al. (2019) proposed a method to learn type-level representations for all periods at once, simply by indexing target word w with the period they occur in (for instance, in an occurrence from 1790, *cat* would be replaced by *cat_1790*) and fitting the representation model on all corpora from all time periods at once. A target word in a period T_1 is treated independently from the same target word in period T_2 as if they were two different lexemes. Because the context words are unchanged and shared across periods, the resulting representations are naturally aligned. They experimentally evaluated their method, called *Temporal Referencing*, with using PPMI matrices and SGNS as representation models. Dynamic type-level models inherently provide temporal alignment, as the same model produces representations for all time periods.

3.2.1.3 Measurements

Cosine distance. Once the representation spaces have been made comparable across time periods, the next step is to measure semantic change. The usual approach is to compute the cosine distance between (aligned) representation of the same word (e.g. Sagi et al. (2009); Gulordava and Baroni (2011); Kim et al. (2014); Kulkarni et al. (2015); Kutuzov et al. (2017); Zhou and Li (2020)), that is:

$$\text{change}_{\cos}(w, T_1, T_2) = d_{\cos}(\mathbf{r}_{(1)}^w, \mathbf{r}_{(2)}^w) = 1 - \text{cos-sim}(\mathbf{r}_{(1)}^w, \mathbf{r}_{(2)}^w) = 1 - \frac{\mathbf{r}_{(1)}^w \cdot \mathbf{r}_{(2)}^w}{\|\mathbf{r}_{(1)}^w\| \|\mathbf{r}_{(2)}^w\|} \quad (3.1)$$

where $\mathbf{r}_{(i)}^w$ is the vector representation of w at time T_i .

Distance between distributions. Co-occurrence matrices and (P)PMI matrices can also be row-normalized to be considered probability distributions over contexts, which allows to use probability distribution distances such as Jensen-Shannon Distance (JSD), as in Xu and Kemp (2015). JSD is derived from the Jensen Shannon Divergence (JSDiv), which is defined between two distributions P and Q as:

$$\text{JSDiv}(P, Q) = \frac{1}{2} \left(KL(P \parallel \text{avg}[P, Q]) + KL(Q \parallel \text{avg}[P, Q]) \right). \quad (3.2)$$

with $KL(p \parallel p')$ the Kullback-Leibler divergence between two distributions p and p' , and avg the averaging operator. JSD is defined as the square root of JSDiv, as the latter does not respect the triangular inequality as is and therefore is not a properly defined distance, contrarily to the former. If $\mathbf{r}_{(i)}^w$ the vector representation of w at time T_i is a probability distribution, semantic change score from JSD is:

$$\text{change}_{JSD}(w, T_1, T_2) = \text{JSD}(\mathbf{r}_{(1)}^w, \mathbf{r}_{(2)}^w) = \sqrt{\text{JSDiv}(\mathbf{r}_{(1)}^w, \mathbf{r}_{(2)}^w)} \quad (3.3)$$

Neighborhood-based distance. Xu and Kemp also computed a second-order distance based on the neighborhood. Let us denote $\mathcal{n}_{(1)}^{w,k}$ and $\mathcal{n}_{(2)}^{w,k}$ the sets of k nearest neighbors of w at time T_1 and T_2 respectively. To compare w at two time periods, they compute 1 minus the proportion of shared neighbors in the old period and in the later period; that is:

$$\text{change}_{\text{neigh}}(w, T_1, T_2, k) = 1 - \frac{|\mathcal{n}_{(1)}^{w,k} \cap \mathcal{n}_{(2)}^{w,k}|}{|\mathcal{n}_{(1)}^{w,k}|} \quad (3.4)$$

This is close to the formulation of the Jaccard distance as used in Xu et al. (2021), which is defined as the ratio between the size of intersection and the size of union between two sets. In our case, this amounts to:

$$\text{change}_{\text{jaccard}}(w, T_1, T_2, k) = 1 - \frac{|\mathcal{n}_{(1)}^{w,k} \cap \mathcal{n}_{(2)}^{w,k}|}{|\mathcal{n}_{(1)}^{w,k} \cup \mathcal{n}_{(2)}^{w,k}|} \quad (3.5)$$

Hamilton et al. (2016a) proposed a *Local Neighborhood Measure*. This local neighborhood measure for a word w between time T_1 and T_2 is based on vectors of the cosine distances between the word's representation at T_i and the those of words in the union of neighborhoods at both periods $\mathcal{n}_{(1)}^{w,k} \cup \mathcal{n}_{(2)}^{w,k}$. Given $\mathbf{r}_{(i)}^w$ the representation of the lexeme w at time T_i and $d_{\cos}$ the cosine distance operator,

they compute for T_1 the vector

$$\mathbf{s}_{(1)}^k = \left\{ d_{\cos}(\mathbf{r}_{(i)}^w, \mathbf{r}_{(i)}^{w'}) \right\}_{w' \in n_{(i)}^{w,k} \cup n_{(i+1)}^{w,k}}. \quad (3.6)$$

and compute $\mathbf{s}_{(2)}^k$ similarly. Each $\mathbf{s}_{(i)}^k$ measures how close is w to other words in the union of neighborhoods. Then, given these two vectors, the Local Neighborhood Measure predicts a value of semantic change as

$$\text{change}_{LND}(w, T_1, T_2, k) = d_{\cos}(\mathbf{s}_{(1)}^k, \mathbf{s}_{(2)}^k). \quad (3.7)$$

Empirical comparisons. Schlechtweg et al. (2019) compared many of the methods described in this section on German diachronic data and confirmed the findings of Hamilton et al. (2016b): SGNS representations aligned with Orthogonal Procrustes and compared with Cosine Distance provide the most reliable results for Lexical Semantic Change Detection. Models using SGNS and Orthogonal Procrustes achieved higher performances on the SemEval-2020 shared task (Schlechtweg et al., 2020) and have subsequently been adopted as a baseline for LSCDiscovery-2022 shared task (Zamora-Reina et al., 2022).

Change-point detection. Kulkarni et al. (2015) used type-level representations – namely SGNS aligned with Orthogonal Procrustes, as well as raw frequency counts and ppart-of-speech tags distributions – to measure lexical semantic change across multiple time periods. They apply a *change point detection* algorithm based on the *Mean Shift* model (Taylor, 2000) to identify whether a word has changed in meaning, as well as the date around which the shift has occurred, given enough diachronic data over multiple consecutive periods. However, the large majority of studies have been focusing on only two periods, leaving no room for change point detection.

3.2.1.4 Limitations of Type-level Vectors

Static models do not account for word order: co-occurrences counts and Word2Vec models indeed rely on the **bag-of-words** approach. Type-level approaches also struggle to handle a word’s polysemy, as they create a single representation per word and must therefore conflate its various senses. In Chapter 7, we make use of SGNS representations aligned with Orthogonal Procrustes to model semantic change in synonym pairs. Our experimental results confirm that these representations struggle to maintain accurate representation of lexical meaning for words with high polysemy degrees.

Having multiple context-specific representations can help accounting for the multiplicity of senses. For these reasons, recent research has preferred to explore Lexical Semantic Change with contextualized representations of meaning.

3.2.2 Contextualized representations

With the exponential development of Language Models over the last years, a lot of the research work in the study of Lexical Semantic Change in NLP has shifted from type-level representations to occurrence-level representations, or *contextualized representations*. The key idea is that such model creates one representation for each use of a word, each representation being modulated by the context around the target word in this occurrence.

3.2.2.1 Representations

Contextualized Language Models. To obtain these representations, one uses a *Contextualized Language Model* (CLM). Contextualized Language Models are large neural networks that are trained to output probabilities of missing words in this input sequence. Through the Language Model's *Tokenizer*, the input sequence is usually decomposed into smaller pieces, *tokens* (which may equate to words in some cases but are usually smaller pieces, like frequent character combinations).

Two main training scenarios exist. On the one hand, *Auto-regressive* or *Causal* Language Models are trained to perform *Next Token Prediction*, i.e. predicting the next token in the sequence from left to right, and can recursively be called to generate full sentences and text documents. On the other hand, the objective of *Masked* Language Models is to predict the actual value of a masked token in the input sequence, using both left and right contexts. NLP studies of Lexical Semantic Change with contextualized representations almost exclusively use Masked Language Models, most often to extract their latent states to obtain a dense vector representation of an occurrence. Indeed, Contextualized Language Models are trained on large amounts of data, and learn to build latent representations that contain semantic information about a token in its context.

The most widely used CLMs for the study of Lexical Semantic Change are ELMo (Peters et al., 2018), BERT and mBERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), and XLM-R (Conneau et al., 2020).

- *ELMo* is a bidirectional 2-layers LSTM network trained on a English data, and produces embeddings in 1024 dimensions.

- *BERT Base* is based on the Transformer architecture (Vaswani et al., 2017) with 12 Transformer layers with hidden size of 768; BERT is trained on English data, is able to take two separate sentences as input using a special delimiter token, which allows for a secondary pre-training objective of Next Sentence Prediction. BERT is trained on English data only.
- *BERT Large*, similar to BERT Base with 24 layers and a hidden size of 1024.
- *mBERT* is the multilingual counterpart of BERT, pre-trained on a corpus of 100 languages. mBERT exists in both *Base* and *Large* variants.
- *RoBERTa* is an improved variant of BERT that removes its secondary pre-training objective and leverage some adjustments in the training procedure to obtain better downstream performances. It is also released in a *Base* and a *Large* version with the same characteristics as BERT.
- *XLM-R* is a RoBERTa-Large model pre-trained on cross-lingual data, making it suitable for multilingual non-English studies of lexical semantic change.

Extracting embeddings from Language Models. To study lexical semantic change with CLMs, we want to obtain a single embedding for our target word in the context of its occurrence. However, when passing the occurrence through the model’s tokenizer, the target word may be split into multiple tokens (the *subwords*). The standard method in such cases is to average together the vector representations of the subwords into a single embedding.

Representational embeddings are obtained from the hidden layers of these neural Language Models. Research studies on the properties of CLMs like Ethayarajh (2019) showed that CLMs tend to disseminate semantic information non-uniformly across layers, as the later layers put more focus on context while early layer focus more on morpho-syntactic and form-based information. As a result, research work usually extract embeddings from the last layers of the model, and often uses embeddings from several layers before aggregating them (sum or average) into a single vector (Martinc et al., 2020a; Beck, 2020; Rosin and Radinsky, 2022). Using latent vectors from several layers allows for more varied information to be captured in the final vector.

Among studies using CLMs to investigate Lexical Semantic Change, most of them have used BERT or mBERT (Martinc et al., 2020a; Kutuzov and Giulianelli, 2020; Beck, 2020; Montariol et al., 2021; Periti et al., 2022; Rosin and Radinsky, 2022; Lendvai et al., 2023; Periti et al., 2024). several multilingual or non-English studies of LSC have also relied on XLM-R (Cuba Gyllensten et al., 2020; Giulianelli et al., 2022; Homskiy and Arefyev, 2022; Rachinskiy and Arefyev, 2022;

Kashleva et al., 2022). We refer to Periti and Montanelli (2024) for a more comprehensive survey that details which CLMs were used and how they were used (layers, aggregation method, etc.) for Lexical Semantic Change Detection.

XL-LEXEME. Contextualized Language Models (CLMs) are pre-trained models that can further be trained on new data and/or with a new objective (*fine-tuning*) to obtain a representational vector space with desired properties. In order to study Lexical Semantic Change Detection, Cassotti et al. (2023) leveraged the multilingual pretraining of XLM-R and fine-tuned it on several large multilingual Word-in-Context datasets, to increase the amount of lexical semantic information captured by the model. Indeed, the Masked Language Modeling objective of XLM-R still requires a significant amount of morpho-syntactic information that may not be necessary for the study of Lexical Semantic Change. Fine-tuning the model on a task specifically focused on lexical meaning may diminish the importance of less useful information in favor of more accurate lexical meaning representations. The resulting fine-tuned model, called **XL-LEXEME**, achieves state-of-the-art performances in the LSCD task (Periti and Tahmasebi, 2024a). Indeed, XL-LEXEME is fine-tuned to distinguish whether two occurrences are of the same sense, and as such it can naturally be applied to compare old occurrences to new occurrences in the LSCD benchmarks discussed in Section 3.1.3.

Aggregating occurrences. We may use bags of occurrence vectors as is, but they can also be aggregated in different ways. In their *word prototypes* (PRT) method, Kutuzov and Giulianelli (2020) averaged the occurrence vectors of the same target word to obtain a single vector representation for each lexeme for each period, which they could compare over time to measure how much the word has changed in meaning. This amounts to create a type-level representation from occurrence-level ones. Another line of research performs clustering of occurrence representations to identify the word’s senses. We refer to these approaches as **Word Sense Induction-based** (WSI-based) methods. Kutuzov and Giulianelli (2020) clustered ELMo and BERT embeddings using the *Affinity Propagation* clustering algorithm. Giulianelli et al. (2020) used *K-means* clustering and *Silhouette score* optimization to choose the amount of clusters, but Martinc et al. (2020b) showed that clustering silhouette scores correlate only poorly with performances on the English LSCD task benchmark. Arefyev and Zhikov (2020) used *Hierarchical Agglomerative Clustering* on bag-of-substitutes vectors. After clustering into senses, one can now represent a word’s meaning with the vector corresponding to the senses distribution among its occurrences, for each time period.

3.2.2.2 Accounting for Temporal Variations

Since representations are computed using the same CLM, most studies bypass the alignment step, deeming that representations from different periods are naturally in the same space. However, available CLMs are often pre-trained on modern data only (wikipedia data, recent books, etc.) and therefore may not accurately capture subtle semantic differences from earlier periods. Several studies have explored different techniques to account for temporal variations, without drastically changing the model’s representational space so that the alignment between representations of different periods is maintained.

Periti and Montanelli (2024) surveyed several studies using Contextualized Language Models for Lexical Semantic Change Detection, and found two different approaches to temporal adaptation of CLM representations within the surveyed literature. First is *domain-adaptation*, where the CLM is further trained for few epochs on the diachronic data for a given period of time, using the same objective as its pre-training (Montariol et al., 2021). This allows to capture the specificities of each time period. Kutuzov and Giulianelli (2020) even proposed an incremental variant, where a model adapted to period T_i serves as base model to be fine-tuned on data for T_{i+1} .

The second approach, which Periti and Montanelli (2024) refer to as *fine-tuning*, consists on training the pre-trained model with a different objective than the one of its pre-training. Pömsl and Lyapin (2020) fine-tuned BERT models to predict, given an occurrence, the time period from which it comes from. Rosin et al. (2022a) appended a temporal token to each occurrence (encoding its time period), and fine-tuned the model to predict this token using the mask token.

With XL-LEXEME, Cassotti et al. (2023) showed that using a model trained on a carefully chosen lexical semantic task (namely Word-in-Context) to represent occurrences from all time periods leads to better results on the Lexical Semantic Change Detection (LSCD) benchmarks than fine-tuning a language model on the data of each time period (Kutuzov and Giulianelli, 2020) or fine-tuning it to incorporate the date of an occurrence in the input (Rosin et al., 2022b) or in the model’s attention mechanism (Rosin and Radinsky, 2022), in all four languages in the SemEval2020-Task1 benchmark (Schlechtweg et al., 2020). Inspired by this idea, in Chapter 5 we propose an extension of Word-in-Context to cover a more general aspect of lexical meaning and we fine-tune a Contextualized Language Model for this objective, reaching equivalent or better performances than XL-LEXEME in the LSCD benchmarks.

In WSI-based approaches, clusters are usually learned among all occurrence vec-

tors of all time periods together. To better account for temporal variations, (Periti et al., 2022) proposed an incremental approach modifying the Affinity Propagation clustering algorithm to integrate cluster centroids of time T_k into the clustering of occurrences of time T_{k+1} , such that the clusters are incrementally updated. (Periti and Tahmasebi, 2024b) discussed several other approaches to clustering, such as using a sliding time window instead of learning clustering on all occurrences at once, but they did not test this setting experimentally.

3.2.2.3 Measurements

Measuring semantic changes with contextualized representations depends on whether occurrence vector have been aggregated or not, and how they would have been.

Distance between prototypes. We denote $\mathbf{p}_{(k)}^w$ the word prototype of word w at time T_k . Kutuzov and Giulianelli (2020)’s *PRT* method computes the inverted cosine similarity between word prototypes of two different periods:

$$\text{change}_{PRT}(w, T_1, T_2) = \frac{1}{\text{cos-sim}(\mathbf{p}_{(1)}^w, \mathbf{p}_{(2)}^w)} \quad (3.8)$$

Distance between usages. Let us consider $\Phi_{(k)}^w = \{\mathbf{r}_{(k),i}^w\}_{i \in [1, n_k]}$ the set of occurrence vector for the n_k occurrences of lexeme w at time T_k . Giulianelli et al. (2020) and Kutuzov and Giulianelli (2020) also introduced *Average Pairwise Distance*, commonly referred to as *APD*. APD is the average of cosine distances among any pairs of occurrences of T_1 with occurrences of T_2 .

$$\text{change}_{APD}(w, T_1, T_2) = \frac{1}{n_1 \cdot n_2} \sum_{\mathbf{r}_1 \in \Phi_{(1)}^w} \sum_{\mathbf{r}_2 \in \Phi_{(2)}^w} d_{\cos}(\mathbf{r}_1, \mathbf{r}_2) \quad (3.9)$$

APD is the most widely used measure of semantic change in the recent NLP literature (Cassotti et al., 2023). Periti and Montanelli (2024)’s survey highlighted that APD and PRT outperform in the LSCD task the other variants proposed in the NLP literature.

Distance between clusters. We denote $\Sigma_{(k),j}^w \subset \Phi_{(k)}^w$ the j -th cluster of word w at time T_k , and $\mathbf{s}_{(k),j}^w$ the average vector representation of occurrences in the cluster. We denote $\mathbf{q}_{(k)}^w$ the distribution of senses obtained by weighting clusters with the number of occurrences, for lexeme w at time T_k .

Jensen Shannon Divergence/Distance are the most widely used metric to compare

sense cluster distributions over time (Rodina et al., 2020; Kanjirang et al., 2020; Arefyev and Zhikov, 2020; Rother et al., 2020; Montariol et al., 2021; Kashleva et al., 2022). The formulation of this measure of change is similar to that of Equation 3.3:

$$\text{change}_{JSD-senses}(w, T_1, T_2) = \text{JSD}(\mathbf{q}_{(1)}^w, \mathbf{q}_{(2)}^w) = \sqrt{\text{JSDiv}(\mathbf{q}_{(1)}^w, \mathbf{q}_{(2)}^w)} \quad (3.10)$$

Maximum square is a different measure proposed in Hu et al. (2019), which only focuses on the maximum change observed in the sense distributions. That is,

$$\text{change}_{MS}(w, T_1, T_2) = \max_i ([\mathbf{q}_{(1)}^w]_i - [\mathbf{q}_{(2)}^w]_i)^2 \quad (3.11)$$

The intuition is that for semantic change to be detected, only a change in one of the senses is required rather than a shift across the entire distribution.

Montariol et al. (2021) introduced the use of *Optimal Transport* to measure the cost of changing from one sense distribution to another, an idea that later inspired Pranjić et al. (2024); Kishino et al. (2025). They used the *Wasserstein Distance*, which solves the problem of finding a transport plan Γ , a matrix that describes the change from cluster distribution $\mathbf{q}_{(1)}^w$ to distribution $\mathbf{q}_{(2)}^w$.

$$\text{change}_{WD}(w, T_1, T_2) = \text{WD}(\mathbf{q}_{(1)}^w, \mathbf{q}_{(2)}^w) \quad (3.12)$$

$$= \min_{\gamma} \sum_i^{|\Sigma_{(1)}^w|} \sum_j^{|\Sigma_{(2)}^w|} d_{\cos}(\Sigma_{(i),1}^w, \Sigma_{(j),2}^w) \Gamma_{\Sigma_{(i),1}^w \rightarrow \Sigma_{(j),2}^w} \quad (3.13)$$

$$\begin{aligned} \text{such that: } & \Gamma_{\Sigma_{(i),1}^w \rightarrow \Sigma_{(j),2}^w} \geq 0 \\ & \sum_i \Gamma_{\Sigma_{(i),1}^w \rightarrow \Sigma_{(j),2}^w} = \mathbf{q}_{(1)}^w \\ & \sum_j \Gamma_{\Sigma_{(i),1}^w \rightarrow \Sigma_{(j),2}^w} = \mathbf{q}_{(2)}^w \end{aligned}$$

Periti and Montanelli (2024) describe other measure with sense distribution, but none of them seem to have been largely accepted with the research community of NLP for Lexical Semantic Change.

Empirical comparisons. Systematically comparing the Graded LSCD performances reached by many of the approaches described in this section over DWUG datasets for a large variety of languages, Periti and Tahmasebi (2024a) showed that representations from **XL-LEXEME** compared over time using **APD** was the top-performing system across most of the LSCD DWUG benchmarks. For in-

stance, in the English benchmark from SemEval2020 Task1, XL-LEXEME+APD reaches a correlation score of .886; for comparison, BERT+APD reaches a score of .563, XL-LEXEME+Clustering+JSD achieves a score of .493. The evolutionary clustering approach proposed by Periti et al. (2022) reaches a correlation with gold scores of .652. For reference, SGNS aligned with Orthogonal Procrustes achieved a correlation score of .422 on this benchmark. Across the 11 settings tested by Periti and Tahmasebi (2024a), XL-LEXEME+APD is the best-performing model in 9 of them. One could expect clustering-based approaches to perform better due to a finer discrimination of polysemy and therefore more explicit information about sense change. Yet, clustering-based approaches usually perform worse than more direct methods (APD, PRT), indicating that there is a difference between what appears interesting and interpretable at the human-level and what is actually effective for machine-level performances.

3.3 Beyond Change Detection

Detecting lexical semantic change is not the end goal of the current research in NLP. Rather, the LSCD task and datasets were made to thoroughly evaluate these models. Several studies have subsequently applied them to investigate regularities in lexical semantic change and to collaborate with historical linguistics in order to better identify and understand semantic change phenomena. In this section, we will review the major insights that computational studies have uncovered about the evolution of word meanings.

The law of prototypicality. Dubossarsky et al. (2015) applied Word2Vec vectors, aligned by incremental training and compared using cosine distance, using the ‘fiction’ fraction of the GoogleBooks corpus, focusing on the span between 1850 and 2009. They also performed a clustering step using the K-means algorithm on type-level representations, to obtain clusters of words. Considering word clusters as *categories*, and they computed the average vector – the centroid – of each category. They found that higher measures of semantic change over time (i.e. larger cosine distances between embeddings of the same word across time) correlated with weaker prototypicality (i.e. greater cosine distance to the centroid of the category the word belongs to). The observed phenomenon is later called the **law of prototypicality**: less prototypical words are more subject to change than very typical ones. This finding supports Geeraerts (1997)’s theory of the diachronic evolution of semasiological prototypes, according to which sense emergence and loss occur at the edges of a word’s polysemic network rather than at its center.

Xu et al. (2021) investigated the relationship between semantic change and prototypicality in specific semantic domains related to emotions. Using SGNS representations provided by Hamilton et al. (2016b) in the GoogleBooks corpus and measuring semantic change using the Jaccard Distance between neighborhoods, they found that semantic change highly correlates with prototypicality in emotion terms, but not in terms related to birds. This study raises attention on the challenge of drawing general conclusions about laws of semantic change, since regularities may only be valid at local scales (e.g. lexical fields) rather than universally.

Laws of innovation and conformity. The study of Hamilton et al. (2016b) investigated the same corpus for the period 1800-2000, binned into decades. Using SVD representations³ aligned with Orthogonal Procrustes and cosine distance as a change measure, they uncovered two regular patterns: (i) word frequency negatively correlates with the semantic change, and (ii) contextual variability in word usages correlate positively with semantic change. To measure contextual variability, they compute a co-occurrence graph: for each target word, the local clustering coefficient of its neighbors is measured, which indicates whether the word is used in varied, disjointed contexts or in similar contexts. They use contextual variations as a proxy for polysemy, arguing that polysemous words are expected to be used in varied contexts. They summarize their conclusions in two statistical laws of semantic change: the **law of conformity**, which posits that frequent words are more stable in meaning and only words of lower frequency undergo semantic change; and the **law of innovation**, that posits that polysemous words are more likely to undergo semantic change.

Revisiting statistical laws of semantic change. Dubossarsky et al. (2017) revisited these three laws by conducting control experiments. In particular, they shuffled occurrences from the different periods of the original corpus (thus destroying the temporal structure), such that each bin of the data contained an equal share of all periods. In this shuffled conditions, they were able to observe the same correlations that those observed in Hamilton et al. (2016b), with only an 8% difference in effect between the original and the control conditions. They also highlight that frequent word representations are less distorted by the SVD transformation than infrequent words, and argue that this may explain that change score highly correlates with frequency even in the shuffled conditions. Regarding the law of prototypicality, they observed that the previously described effect could also be observed in shuffled conditions, to a lower degree. They

³ They mentioned having observed the same tendencies with SGNS.

ultimately conclude that, while the effects of these three laws should not to be entirely disregarded, they may be less significant than originally postulated.

Uban et al. (2019) extended the investigation cross-linguistically, through inspections of cognate sets. Cognates are words in different languages that share a common origin (for instance, *arhitect* in Romanian, *architecte* in French, *arquitecto* in Spanish, all originating from *architectus* in Latin). The authors aligned type-level representations between languages via a linear regression problem and they measured semantic change using a second-order distance between words in cognate sets that measure ‘falseness’, i.e. the likelihood that cognates have become false friends. They found that the semantic divergence between cognates was correlated with their degree of polysemy in WordNet (cf. law of innovation). However, they also found that frequency correlated positively with semantic divergence in cognates, in contradiction with the law of conformity. Kawasaki et al. (2022) revisited this study in romance languages, addressing some methodological shortcomings of the previous study. To measure polysemy, they referred to the number of senses in Latin rather than in the modern languages, ensuring that polysemy is measured *before* the potential semantic shift. To avoid the shortcomings addressed in Dubossarsky et al. (2017), they control for correlation between frequency and change in random word pairs. Ultimately, their results confirmed that both the law of conformity and the law of innovation hold in Romance cognates.

Factors of semantic change. The work of Fugikawa et al. (2023) also fuels the debate regarding the different factors that may correlate with semantic change. Training a model on quantitative judgements of concreteness and valence (the emotional charge of a word) as well as frequency counts to predict the directionality of source→target changes described in the Database of Semantic Shifts (Zalizniak et al., 2012b), they showed that concreteness is the most accurate predictor of the direction of lexical semantic change. Using contextualized representations from a BERT-derived model, and attempting to infer the target of changes in the database, their results show that semantic *similarity* and *analogy* are good predictors of the target sense, the predictors being related to metaphor and metonymy respectively.

Semantic change in synonyms. Xu and Kemp (2015) conducted a computational evaluation of two opposite hypotheses in historical linguistics. On the one hand, they consider the *Law of Differentiation*, abbreviated ‘LD’ (from Bréal (1897)’s law of reallocation discussed in Section 2.3.4), that posits that synonymous words cannot share (near-)identical meaning for long, and as such must

differentiate in meaning. On the other hand, they study the **Law of Parallel Change**, or ‘LPC’, following the observations of Stern (1921) and Lehrer (1985) describing a trend in groups of (near-)synonyms to undergo parallel evolutions when one of the group’s member changed in meaning. Using co-occurrence matrices from the GoogleBooks corpus in English between 1890 and 1990, they measured semantic proximity in pairs of synonyms described in synonym dictionaries from the 1890’s and compared these to semantic proximity in random control pairs. They found that the semantic proximity in synonym pairs tend to be more stable over time than those in control pairs, and ultimately conclude in favor of LPC rather than LD. Yet, like many studies before 2020, this work critically lacks an evaluation procedure to assess the quality of both representations and measures of change. In Chapter 7, we will revisit this study, addressing some of its shortcomings and highlighting the challenges of studying lexical semantic change.

Onomasiological study of DESTROY. Franco et al. (2022) use second-order PPMI matrices to study the diachronic evolution of two Dutch words referring to the concept of DESTROY, namely *vernietigen* and *vernielen*. The second-order representations are constructed by first computing PPMI representations for all considered context words in their corpus, and then using these context representations to build a unique representation for each occurrence of *vernietigen* or *vernielen* depending on its contexts. They run Hierarchical Agglomerative Clustering to identify sense clusters of the two target words. Examining the dominant context in the clusters, they are able to identify the different senses that the two target words had across the studied period (16th to 20th century). They observe that the two words were initially interchangeable but gradually specialized into distinct core meanings, each of them referring to a specific nuance of the concept of DESTROY.

Grammaticalization in deverbal prepositions. The recent work of Nagata et al. (2024) investigated the formation of deverbal prepositions, that is prepositions that originated from verbs (e.g. *during*, *regarding*, *according to*, etc.). Using both type-level (SGNS) and occurrence-level (BERT) representations in the Corpus Of Historical American English, and comparing target word vectors to other vectors corresponding to reference prepositions and verbs, they were able to measure the degree of grammaticalization of a target word by comparing whether it was closer to reference verbs or to reference prepositions. They compared their results to two previous linguistic studies that made ratings of “how prepositional” were the investigated items, and used their empirical evidences

to conclude in favor of one or the other hypothesis for each of the investigated items. For example, they concluded that *following* is much closer to being a preposition than a verb, with the 3rd highest prepositionality degree among their list of 48 target words.

Characterizing semasiological changes with sense relations. Finally, Cas-sotti et al. (2024b) introduced **LSC-CTD**, a dataset of instances of semasiological semantic change annotated with their type of change as reported in the typology of Blank (1997). In particular, they focus on 4 types of change: *generalization*, *specialization*, *co-hyponymous transfer* and *auto-antonymy*. In the dataset, each entry is a pair of sense definition, and annotated with the type of change that relates the senses in the pair. The authors used senses definitions from Word-Net to train a classification model that distinguishes between hyponymy, hypernymy, antonymy, co-hyponymy, and homonymy (no lexical relation) in pairs of sense definitions and further applied this classification model to the LSC-CTD dataset. The model – trained on synchronic sense-relation identification, which are mapped to the corresponding change type – was able to reach more than more than 60% accuracy for specialization, generalization and auto-antonymy, but struggled to identify cases of auto-antonymy.

Addressing other types of semasiological changes, Yu and Xu (2023) use the predictive ability of contextualized language models to study *word sense extension*, defined as the modulation of an existing sense to a new kind of use-contexts. As such, this framework encompasses both change by metaphorization and change by metonymization. Training a series of Contextualized Language Models from scratch on wikipedia data simulating word sense extensions in polysemous words, they propose chaining-based models to predict these extensions from the output probabilities of these language models.

This research direction is highly interesting, as it shifts the focus from *detecting* semantic change (which is the only focus of most of studies surveyed in this section) to *characterizing* it.

3.4 Summary and Limitations

In this chapter, we provided an overview of the current state of the art in Natural Language Processing concerning lexical semantics, and in particular lexical semantic change. We first described the various tasks and benchmarks that have been formulated and curated in NLP to tackle problems related to the mapping between words and concepts. We reviewed two categories of tasks, namely

between-word tasks which address relations between lexemes and similarity in meaning between them, and *within-word tasks* which deal with polysemy and contextual variations of word usage. Then we described the Lexical Semantic Change Detection task (LSCD), the reference evaluation for computational models of semasiological lexical semantic change. Alongside LSCD, we briefly introduced the alternative tasks proposed to evaluate diachronic NLP models, namely Novel Sense Detection and Sense Definition Generation. However, we noted that the proposed tasks fail to properly evaluate the diachronic validity of candidate models. Next, we conducted a thorough review of the existing NLP literature about computational models of Lexical Semantic Change, focusing on vector-based models and showing that they can usually be decomposed into three steps – representation, temporal alignment, measurement – and classified them according to the type of representations they use, that is static vectors or contextualized vectors. Finally, we summarized the findings of computational studies about regularities in semantic change and the proposed laws grounded in empirical studies on large-scale corpora. In particular, we mentioned the ongoing discussion on the relation between semantic change and several factors, such as polysemy, frequency, prototypicality, valence and concreteness. We also reviewed studies dealing with change in pairs of synonyms, grammaticalization of verbs into prepositions and the relationship between synchronic sense relations and types of lexical semantic change.

Yet, the existing literature suffers from several major limitations we will now discuss. Most of the studies reviewed in this section focus on detecting lexical semantic change rather than describing this change and uncover potential regularities. This is especially true for studies published after 2019, i.e. after the release of the SemEval-2020 shared task on Unsupervised LSCD. Before that, the community was in need of robust evaluation methods, and most studies conducted an evaluation of their models on very small samples – if the methods were evaluated at all. The LSCD benchmarks derived from Diachronic Word Usage Graphs have allowed systematic comparisons across a large variety of methodologies and a variety of different languages. Many of the subsequent research has proposed innovative methodologies to achieve higher performances on these benchmarks, at the expense of the research interest in investigating laws and mechanisms of semantic change. However, this tendency seems to become attenuated, as an increasing number of studies in the last two years focus on characterizing semantic change types and applying NLP semantic change models to specific case studies.

The models that perform the best in the LSCD datasets are not necessarily the

most informative about Lexical Semantic Change. Indeed, a simple numerical score gives little to no information about how a word has changed in meaning. More informative systems, such as those using WSI methods to discriminate word senses in data are more interpretable, but they are often outperformed by more direct approaches that compute change score through Average Pairwise Divergence or distance between type-level representations. It appears there must be a tradeoff between interpretability and efficiency. Zamora-Reina et al. (2022) also showed that most computational models had a high Recall but only a moderate Precision, indicating that these methods tend to produce a high number of false positives in detecting semantic change.

Yet, we argue that the most important limitation to existing studies of lexical semantic change in Natural Language Processing resides in the dominance of semasiological studies compared to the rare, onomasiological ones. To the best of our knowledge, the only studies that can be considered as onomasiological are Franco et al. (2022) on the concept of DESTROY in Dutch, and Xu and Kemp (2015) in pairs of synonyms, although the latter primarily investigates whether words are still synonyms and not on the concept they (used to) refer to, which is closer to a semasiological approach than an onomasiological one. The lack of onomasiological investigation in the NLP community contrasts with research on lexical semantic change in diachronic linguistics, which often adopts an onomasiological perspective to describe regularities and patterns of change. The scarcity of onomasiological studies in NLP can be explained by the fact that most of the proposed models are inherently semasiological, as they create representation for words and not for concepts. Indeed, words are explicitly present in data, while meanings must be inferred. It is therefore more straightforward then to build word representation rather than concept representations. Studying the evolution of concept namings over time would require to have a pre-determined set of words and the associated concepts across multiple time periods, such as the *Historical Thesaurus* (Kay et al., 2020). Unfortunately, the latter's API is not publicly available, and exists for English, which notably limits the possibility of its application in large-scale, systematic and multilingual computational studies.

The ability to determine word associations by concepts across multiple time periods – i.e. lists of (near-)synonyms – would provide a valuable basis for investigating Lexical Semantic Change. As we will discuss in Chapter 7, combining resources from different origins and from different periods may raises issues of comparability and this misalignment would diminish the strength of a computational study. In this thesis, we explore ways to extract groups of (near-)synonyms corresponding to latent concepts directly from the data, in order to obtain these

lists across multiple time periods. The overall objective is to model semantic evolutions within groups of words that relate to the same concept. As such, the very first step is the identification of groups of synonyms directly from the data, without external knowledge. We begin our study in the synchronic setting, to later extend it to diachronic data.

Part II

Synchronic Contributions: Representing and Extracting Concepts From Data

4

Concept Induction: Clustering Data to Learn Latent Concepts

Current models used to represent meaning in Natural Language Processing limit the scope to which we can understand Lexical Semantic Change with computational models by often enforcing a semasiological approach. In this first part of our contributions, we focus on representing and extracting concepts from data in the synchronic setting. This will be our first step towards a more complete computational model of lexical semantic change.

In this chapter in particular, we elaborate a novel lexical semantic task, **Concept Induction**, that addresses both the semasiological side and the onomasiological side of the word-concept mapping, by clustering words into clusters of shared concepts. As a result, one can access both the sense inventories of target lexemes, and the namings of found concepts, thus allowing a larger investigation of lexical meaning. The learned concepts are latent concepts (i.e. unlabeled), thus they do not require external supervision or a pre-defined set of concepts. To address Concept Induction, we propose a **bi-level methodology based on dual clusterings** that leverage *local* differences between senses of a given word, and *global* similarities in meaning between usages of different lexemes.

4.1 Definition of Concept Induction

Our main motivation behind Concept Induction is to present a view of the mapping between words and their meaning(s), as defined in Section 2.2.3. This view is systemic, meaning that it should not be defined for individual lexemes neither for individual concepts, but rather acknowledging these as a whole with interactions and relations. This extends beyond the primary objective of Word Sense Induction (WSI), which defines word senses as pertaining to individual words only and does not explore relations between lexemes or concepts.

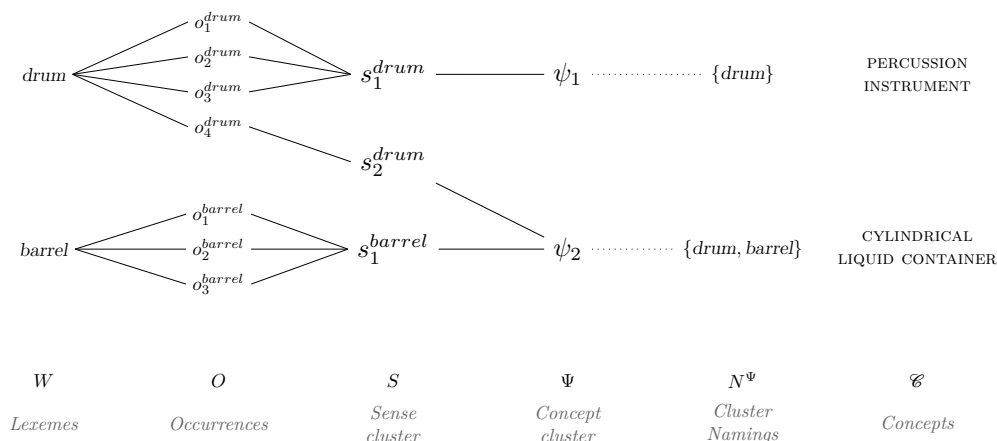


Figure 4.1: Illustration of Concept Induction’s framework. The word *drum* is polysemous and has two senses corresponding to two different concepts, and is synonym with *barrel* for this second meaning.

4.1.1 Task definition

The goal of Concept Induction (CI) is to automatically learn a set of latent concepts directly from the data, i.e. learning a soft clustering N_Ψ in the set of target words W that should correspond to the multiple concepts instantiated by occurrences of the corpus. Being sets of words associated to concepts, clusters in N_Ψ should correspond to the **namings** of concepts found in the data. N_Ψ is a *soft* clustering because a word can be assigned to several clusters (when it is polysemous). Using a different perspective than Word Sense Induction, the framework of Concept Induction provides a more complete view on meaning across the lexicon. Both WSI and CI capture *polysemy*, but CI also reveals *synonymy* across the lexicon. One advantage of WSI over Word Sense Disambiguation is that the former does not require explicit set of senses unlike the latter; similarly, Concept Induction does not require a pre-defined set of concepts, as they will be latently defined and inferred from the data.

4.1.2 Formal framework

Our goal is to model the mapping between lexemes and concepts as it appears in the occurrences from a given corpus. Word Sense Induction, in contrast, models only the mapping between lexemes and senses. The relation between lexemes, occurrences, word senses and concepts has already been discussed in Section 2.2.3 of Chapter 2; and in particular illustrated in Figure 2.4. In this figure, senses and concepts determine different ways to divide the set of occurrences. In this section, we provide the formalism necessary to describe our clustering task.

Let W be the *lexicon*. For all word w in W , we denote $o_{w,i}$ the i -th occurrence of w in the corpus. We define $O_w = \{o_{w,i}\}_{1 \leq i \leq m_w}$ the set of m_w occurrences of w . The corpus, denoted O , is the union of all O_w .

For a given word $w \in W$, the set O_w can be partitioned according to its different senses. We denote $s_{w,j}$ the part of occurrences of w in the corpus corresponding to the j -th sense of w . We refer to these groups of occurrences as the *sense clusters* of w . The set $S_w = \{s_{w,j}\}_{1 \leq j \leq n_w}$ forms a partition of O_w , and we call S the set of all sense clusters of all words, i.e. $S = \bigcup_{w \in W} S_w$. S is a “local” (lexeme-centric) partition of the whole O . The task of Word Sense Induction aims at learning the partition S given a corpus O .

In this chapter, we aim at dividing the corpus into *concepts* instead of senses. Assume that there is a set of semantic concepts, $\mathcal{C}$, and the k -th concept in $\mathcal{C}$ is denoted c_k . We denote ψ_k the group of occurrences of words corresponding to c_k and we refer to it as a *concept cluster*. We denote $\Psi = \{\psi_k\}_{1 \leq k \leq p}$ the partition of O in p concept clusters. Unlike sense clusters of S , a concept cluster $\psi_k \in \Psi$ can gather occurrences of different words: Ψ is a “global” partition. Each occurrence $o_{w,i}$ of a word $w \in W$ is associated to a sense cluster $s_{w,j}$ and a concept cluster $\psi_k \in \Psi$. We can say that a concept c_k corresponding to ψ_k is instantiated by occurrence $o_{w,i}$ through the sense corresponding to $s_{w,j}$, or conversely that $o_{w,i}$ uses the sense reflected in $s_{w,j}$ to mean the concept c_k described by concept cluster ψ_k . All occurrences of sense cluster $s_{w,j} \in S$ appear in the same concept cluster $c_k \in \mathcal{C}$.

In summary, S and Ψ are partitions of O and are naturally constrained as follows:

1. By definition, a sense cluster in S contains occurrences of one and only one word $w \in W$.
2. An occurrence $o_{w,i}$ is assigned to one and only one $s_{w,j} \in S$ (an occurrence realizes only one sense).
3. An occurrence $o_{w,i}$ is assigned to one concept cluster $\psi_k \in \Psi$ (an occurrence evokes only one concept).
4. In a given sense cluster $s_{w,j} \in S$, all occurrences are assigned to the same concept cluster $\psi_k \in \Psi$.
5. All the $s_{w,j} \in S_w$ are included into *distinct* concept clusters (senses of a given word points towards different concepts).

Word-level clustering and namings. From the partition Ψ on occurrences, one can derive N_Ψ , a clustering of the set of words W into concepts. To each concept

cluster $\psi_k \in \Psi$ we associate a word cluster N_k in N_Ψ that contains all lexemes of W whose occurrences were assigned to ψ_k . We say that N_k is the **naming** of concept cluster ψ_k , because words in N_k should be the designation of concept c_k . In N_Ψ , a polysemous word with n senses appears in n distinct clusters (one per sense), and synonyms appear in at least one common cluster (one per shared concept).

In Figure 4.1 we illustrate this framework, using a corpus of occurrences of the words *barrel* and *drum*. In this scenario, $W = \{\textit{barrel}, \textit{drum}\}$ and two concepts are instantiated: PERCUSSION INSTRUMENT and CYLINDRICAL LIQUID CONTAINER. The lexeme *drum* exhibits two senses as it has occurrences corresponding to both concepts: *drum* is polysemous. The second concept is also instantiated by occurrences of *barrel*, therefore *drum* and *barrel* show synonymy in this case. This toy example also follows all constraints formulated above.

4.2 Methodology

In this section we describe the methods we propose and evaluate for Concept Induction. We learn a clustering N_Ψ drawing inspiration from the relations between O , S , $\mathcal{C}$ described in Section 2.2.3. In particular, the overall objective of our methodology consist in finding Ψ (i.e. partition occurrences into concept clusters) to derive N_Ψ . Section 4.1.2 highlighted that there are two levels of partitions: a local level (senses) and a global one (concepts). The proposed approaches rely on both levels and the use of a Contextualized Language Model (CLM) to gather representations of occurrences influenced by the context.

4.2.1 Proposed Bi-level method

Local (lexeme-centric) clustering Firstly, we propose to learn a word-sense partition for each target words individually. Using the CLM hidden layers, we extract a vector representation (the occurrence embedding) of every occurrence $o_{w,i}$. We then learn a partition $\hat{S}_w$ of each O_w using a clustering algorithm on the embeddings. Each $\hat{S}_w$ describes the *locally estimated* sense clusters of word w . Jointly considering these partitions for all $w \in W$, we obtain a partition $\hat{S}$ of the whole set of occurrences O . This partition is *local* in the sense that each word has its occurrences clustered independently from other words.

In the context of our bi-level method, we call these locally estimated clusters **proto-sense clusters**, as they do not necessarily correspond to “finalized” sense clusters for a given word and may undergo merging operations if the next step

of our method judges that the correspond to the same concept.

Global (cross-lexicon) clustering Once we have a local clustering $\hat{S}$, we turn from considering words independently to consider all words together. In this step, we learn a *global* clustering by merging local clusters of occurrences. To do so, we average embeddings of all occurrences in the same proto-sense cluster to get a single embedding representing each proto-sense cluster. Then we run a second clustering algorithm, this time using the averaged embeddings of proto-sense clusters as training data. This global clustering defines a new partition Ψ of the corpus O : when two local proto-sense clusters $\hat{s}_{w_1,j}$ and $\hat{s}_{w_2,j'}$ are merged into the same global cluster ψ_k (because their embeddings were clustered together), all their occurrences are assigned to global cluster ψ_k . The global clusters ψ_k are the learned **concept clusters**. From this global occurrence partition Ψ we can easily extract N_Ψ , a word-level soft-clustering of lexemes whose occurrences appear in the same ψ_k .

This Bi-level method directly implements the system of constraints described in Section 4.1.2. Only constraint 5 is not enforced by design. Indeed, our proto-sense clusters being learned and not informed by an expert, the local clustering step may make errors, especially if the data for a given word are sparse. Allowing the global clustering to merge local clusters enables the correction of local clustering’s recall errors using information from the global level. This is why we refer to local clusters as ‘proto-sense cluster’ and not simply ‘sense clusters’, because occurrence of different senses normally evoke distinct concepts.

We also want to highlight that the proposed methodology is generic, in the sense that it is not tied to a specific choice of clustering algorithm.

4.2.2 Local-only and Global-only

Sense-inducing systems (WSI approaches) that create only local clusters of occurrences for each word are said to be Local-only systems. We use them as baseline models that only produce word-level clusters of size 1 and do not reflect synonymy, but still learn polysemy.

On the other hand, consider a system does not perform a local step and instead directly divides occurrences into global concept clusters. We refer to this kind of system as Global-only approaches. They allow to evaluate the contribution of the local clustering step in the overall process: we hypothesize that the local step in Bi-level will reduce potential variance among occurrences by aggregating them, thereby increasing Precision compared to Global-only.

4.3 Experiments

In this section, we evaluate the abilities of the proposed methods to induce concepts and compare the proposed bi-level approach to other methods. We investigate the advantages of the bi-level approach not only for the global viewpoint but also in the local setting.

4.3.1 Comparing clusterings with BCubed

Evaluation of Concept Induction We compare the learned word clustering to the reference one. We choose to use the *BCubed metrics* (Bagga and Baldwin, 1998), obtaining *Precision* and *Recall* for the evaluated clustering compared to the reference, as well as an F_1 score. To account for overlapping clusters (as we are in the case of soft clustering), we use the *Extended BCubed metrics* proposed by Amigó et al. (2009), which has already been used as evaluation in SemEval2013 WSI task (Jurgens and Klapaftis, 2013).

The extension of BCubed for overlapping clusters rely on two quantities, *Multiplicity Precision* (MP) and *Multiplicity Recall* (MR).

Given two clustering functions, r (the reference one) and f (the one we want to evaluate), and two lexemes w_1 and w_2 , let us denote $F_{w_1, w_2} = |f(w_1) \cap f(w_2)|$ the number of common clusters between w_1 and w_2 in the clustering induced by f ; and $R_{w_1, w_2} = |r(w_1) \cap r(w_2)|$ the set of common clusters in the clustering induced by r . MP and MR between two lexemes are defined as follows:

$$\text{MP}(w_1, w_2) = \frac{\min(F_{w_1, w_2}, R_{w_1, w_2})}{F_{w_1, w_2}}$$

$$\text{MR}(w_1, w_2) = \frac{\min(F_{w_1, w_2}, R_{w_1, w_2})}{R_{w_1, w_2}}$$

For a given pair of lexemes w_1 and w_2 , if candidate f does not find enough common concepts between w_1 and w_2 compared to the expected number of shared concepts ($F_{w_1, w_2} < R_{w_1, w_2}$), then the precision criterion $\text{MP}(w_1, w_2)$ remains to 1, but the recall criterion $\text{MR}(w_1, w_2)$ decreases. On the contrary, if f finds too many common concepts ($F_{w_1, w_2} > R_{w_1, w_2}$), then $\text{MP}(w_1, w_2)$ decreases while $\text{MR}(w_1, w_2) = 1$.

It is worth mentioning that MP and MR can be (and will be) computed between a lexeme w_i and itself (w_i, w_i). In that case, MP indicates whether we retrieved the right amount or too many senses of w_i , while MR indicates whether we retrieved

all senses of w_i or too few.

MP (resp. MR) can be computed for every lexeme w_1 with every other lexeme w_2 sharing at least one cluster with w_1 in f (resp. in r) (including w_1 itself). We denote $\overline{\text{MP}}(w_1)$ and $\overline{\text{MR}}(w_1)$ the obtained averages. In the case of non-overlapping clusters, this formulation gives the same result as the original (non-extended) BCubed. To evaluate WSI, the formulation is the same but we do not evaluate at the word-level but at the occurrence-level.

Precision, Recall and F-score are obtained as follows:

$$\begin{aligned}\text{Precision} &= \frac{1}{|W|} \sum_{w \in W} \overline{\text{MP}}(w) \\ \text{Recall} &= \frac{1}{|W|} \sum_{w \in W} \overline{\text{MR}}(w) \\ \text{F}_\beta &= (1 + \beta^2) \frac{\text{Recall} \times \text{Precision}}{\beta^2 \times \text{Precision} + \text{Recall}}.\end{aligned}$$

By default we fix $\beta = 1$, as we compare the learned clustering and the reference clustering as equals and therefore do not find that Precision and Recall should be weighted differently.

Amigó et al. (2009) showed that the benefits of BCubed over other clustering scores. For instance, *Rand Index* does not handle well the case of many small clusters, which is likely to be the case for Concept Induction. We also prefer *Extended BCubed* over *Overlapping Normalized Mutual Information* (McDaid et al., 2011) as the latter is matching-based. That is, the repetition (or non-repetition) of identical clusters will have no impact on the measure. However, identical clusters of words may correspond to distinct concepts. In *Extended BCubed*, repeated clusters are taken in account as we measure the number of times two lexemes are clustered together. The denominator of MP ensures that over-estimating the number of common clusters is also penalized, and that of MR ensures that under-estimating is penalized. *Min* operators are there to prevent both quantities to exceed 1.

Using BCubed metrics, for a given evaluated clustering, low precision would mean that lexemes have been incorrectly clustered together because none of their occurrence annotations maps to a shared concept according to annotations. Low recall means that the evaluated system fails to capture clusters of lexemes whose occurrences share a concept according to annotations. The number of common clusters between two words also impacts BCubed metrics: if two lexemes appear together in too many clusters compared to the reference clustering,

precision is decreased; if the number of common clusters is too low, recall is decreased.

4.3.2 Settings

Data. We choose to use the annotated part of the *SemCor 3.0*¹ corpus. This dataset contains occurrences for a wide number of words, and morphosyntactic annotations providing each lexeme with a part-of-speech tag. Among all lexemes having at least 10 annotated occurrences, we keep only nouns (excluding proper nouns)² composed only of alphabetical characters with a minimum length of 3 letters. The resulting lexicon W contains 1,560 different lexemes, for which we gather a corpus O containing a total of 52,997 occurrences³. *SemCor* is also semantically annotated, with each occurrence of a target lexeme assigned to a synset in WordNet, that we consider to be the concept it refers to. We derive a reference partition of the occurrences Ψ and a reference soft-clustering of the words N_Ψ from annotations, for a total of 3,855 different concepts (WordNet’s synsets) covered in O . This set of concepts is the subset of WordNet corresponding to the textual data.

Development. To learn the clustering, candidate systems have access to the full set of occurrences-in-context but not their annotations. To choose the appropriate set of hyperparameters, we create a *Dev* split of the annotations by randomly sampling 10% of concepts and revealing semantic annotations of the corresponding occurrences. We use them to evaluate Concept Induction for this small set of concepts, and choose the set of hyperparameters that scores best in BCubed F_1 .

Evaluation splits In the final evaluation phase, we compute scores on all concepts / all occurrences, including the *Dev* split, as concepts in it are part of the whole subset of WordNet described by *SemCor*’s annotations. In the full data, we found that 88% of the concepts were instantiated using only a single lexeme. To better evaluate cases of synonymy, we also evaluate systems on a subset of the corpus, denoted “*Synon*”, that contains only occurrences of concepts show-

¹ <http://web.eecs.umich.edu/~mihalcea/downloads.html#semcor>

² For the sake of simplicity and clarity, this study is focused only on nouns. Indeed, other parts-of-speech induce extra difficulties. Verbs for instance required extra preprocessing steps and decisions (e.g. include or exclude gerundive uses, past participle employed like adjectives, etc.). Extension of experiments to other PoS is left to future work.

³ Sentences in which the lexeme appears, paired with its position within them. If the lexeme appears multiple times in the same sentence, we create several distinct occurrences, where only the position varies.

ing synonymy (the remaining 12% of concepts, instantiated through at least 2 distinct lexemes). Statistics are provided in Table A.1 in Appendix A.1.1. Note that it only changes the set of concepts/lexemes for which the system is being evaluated, not the clustering’s training data.

4.3.3 Systems and baselines

Clustering Algorithms. We try two different clustering algorithms relying on different paradigms: *K-means* (used in Chronis and Erk (2020)), a centroid-based algorithm with a fixed number of clusters, and *Agglomerative clustering* (used in Saidi and Jarray (2023); Velasco et al. (2023); dubbed “Agglo” for short), a deterministic hierarchical approach using a distance threshold to create a dynamic number of clusters instead of using a fixed one. Another difference between K-means and Agglo is that the K-means assumes that clusters are of nearly-spherical shape and balanced in number of points, while Agglo does not make assumptions on the shape of data. Details of hyperparameters and tested values are provided in Appendix A.1.2.

Representations. Following Chronis and Erk (2020) and Eyal et al. (2022), we use *BERT Large* (Devlin et al., 2019), a *masked* language model with 24 layers and 345M parameters. This choice allows for direct comparisons with these approaches. Also, BERT Large was found by Haber and Poesio (2021) to yield better grouping of sense interpretations than other LLMs.⁴ We average subwords’ embeddings if needed. It is a common practice in previous work on semantic-related tasks to use the average of the last 4 layers to get embeddings; we decided to adopt the same “4 layers average pooling” strategy, but trying with different possible sets of layers (see Appendix A.1.2). Therefore, for a set of four layers, we average hidden states across the selected layers to get a single 1024-dimensional vector. We found that layers 14 to 17 obtained the best results on *Dev* for all methods (global/local-only and bilevel).

Sense-inducing systems. Comparison to Local-only systems will give a (strong) baseline just by inducing senses without aiming at concepts. We used the same clustering algorithms. We also implement the WSI method proposed by Eyal et al. (2022). It relies on a different paradigm, using the Language Model for *substitution* instead of word embeddings. From lists of substitutes, they build a graph of substitutes in which they find communities and then assign each occurrence to a community of substitutes to find the word senses. Because Local-only

⁴ We leave to further work the use of autoregressive and/or newer Language Models.

Concept Induction	Full data			Synon.		
	P	R	F_1	P	R	F_1
Baselines						
lexemes	1.0	.43	.61	1.0	.61	.50
Oracle WSI	1.0	.75	.86	1.0	.39	.56
Local-only Systems						
K-means Local	.73	.70	.71	.67	.38	.49
Agglo Local	.92	.53	.67	.92	.35	.50
Eyal et al. (2022)	.31	.75	.44	.37	.39	.38
CI Systems						
K-means Global	.48	.65	.56	.68	.54	.60
K-means Bi-level	.70	.59	.64	.82	.47	.59
Agglo Global	.61	.60	.60	.82	.50	.62
Agglo Bi-level	.75	.60	.66	.86	.49	.62

Table 4.1: Concept Induction BCubed Precision (P), Recall (R) and F_1 on the SemCor data averaged over 5 runs.

methods only induce senses, their hyperparameters are chosen to maximize a WSI objective on polysemous words of the *Dev* split.

Baselines We construct a candidate clustering N_Ψ where each lexeme has its own cluster. This baseline model is referred to as the “lexemes” baseline. This is to evaluate the extent to which the information contained by the lexeme alone can be used to induce concepts without any knowledge on word senses neither on context. As a second baseline, we create for each lexeme as many singleton clusters as the number of different concepts its occurrences are annotated with. All created clusters are of size 1: we account perfectly for polysemy but not at all for synonymy. This second baseline is dubbed “Oracle WSI”.

4.3.4 Concept Induction in SemCor

In Table 4.1 we display the Concept Induction scores (F_1) of proposed baselines and systems on the full SemCor data and on the *Synon.* split. On the full data, both the lexemes and Oracle WSI baselines achieve very good performance because they have, by design, a perfect precision (they do not cluster lexemes at all and do not overestimate the number of clusters) and because 88% of concepts are instantiated with only a single lexeme (thus their recall is still good). However, their performance is poor on the *Synon.* split of the data, where concepts are instantiated with multiple lexemes, showing their inability to model synonymy.

The proposed Concept Induction systems reach scores ranging from .56 to .66 on

	Cluster size		
	2	3	4+
Nb. of annotated clusters	50	50	23
Category (% of annotated clusters)			
Synonyms	42	38	17
Near-synonyms	24	24	35
Related	26	36	48
Invalid	08	02	0

Table 4.2: Qualitative manual evaluation of obtained word clusters of size ≥ 2 .

the full data, half of them outperforming the lexemes baseline, and from .59 to .62 on the *Synon.* split, outperforming all other systems. While still challenging, it exhibits that it is indeed possible to induce WordNet-based *concepts* in a corpus using LMs hidden layers vectors.

We also see that K-means-based approaches are consistently outperformed by Agglomerative methods. This indicates that the representational spaces in LM hidden layers are not organized in a nearly-spherical fashion as K-means algorithm assumes, but rather are populated less uniformly. This is reflected in precision and recall: Agglomerative systems reach a higher precision than K-means with similar recall.

Overall, results are in favor of *Bi-level* approaches over *Global-only* systems, with substantial improvements in F_1 on the full data while obtaining (nearly) identical performance on concepts of multiple lexemes, and large increases in precision while the loss in recall is minimal. This demonstrates that considering the local (lexeme-centric) perspective is beneficial to a global (cross-lexicon) view when inducing concepts. The local clustering, with the subsequent representation averaging, helps reducing variance in occurrences and therefore allow to reach higher levels of precision in the global clustering compared to Global-only. We would also like to emphasize that, while Global-only systems are more simple in design, their computational cost is usually higher than Bi-level ones, especially when the clustering algorithm’s time complexity is quadratic with respect to the number of occurrences.

4.3.5 Qualitative analysis of concepts clusters

We manually annotate word clusters (obtained from our best-performing approach, the Agglo Bi-level system) containing at least 2 lexemes according to the semantic similarity between lexemes. Distribution of cluster sizes (in number of lexemes) can be found in Appendix A.1.3. We distinguish four categories:

synonyms when lexemes are cognitive synonyms (e.g. *necessity* and *need*), *near-synonyms* for lexemes close to be synonyms but showing slight difference in meaning (e.g. *duty* and *task*, the former being stronger than the latter),⁵ *related* when lexemes show a topical (e.g. *dirt*, *sand* and *mud*) or lexical relations (e.g. antonyms like *man* and *woman*) and *invalid* clusters when lexemes show no semantic relation (e.g. *child* and *idea*).

Proportions of these annotations are displayed in Table 4.2 with respect to the cluster size, the number of lexemes in the cluster. For a given cluster size, if the number of clusters exceeds 50, we randomly sample 50 clusters to be annotated. Overall, the proportion of synonyms and near-synonyms is generally above 50% and less than 10% of clusters are invalid, indicating that most learned concepts are reliable and meaningful. We argue that the remaining *related* term clusters, while not synonyms, may still be interesting in less fine-grained studies. The portion of *related* clusters is in line with findings from previous work showing that BERT can reflect other lexical relations, such as hyponymy (Hanna and Mareček, 2021).

4.3.6 Benefits at the local level

We now turn back to the local level and assess whether the information brought at the global level helps distinguishing senses of individual words. Here we do not evaluate the word-level soft clustering, but the occurrence-level division of SemCor’s data, considering each word independently. In other words, we evaluate WSI in SemCor using annotations as the reference sense clustering.

Evaluation of induced senses For each word $w \in W$, we compare how its set of occurrences O_w is divided in Ψ to how it is divided in the reference $\mathcal{C}$ provided by annotations using BCubed metrics, and we average scores obtained across W . We display the WSI BCubed F_1 , as in previous WSI tasks like Jurgens and Klapaftis (2013). Following Amrami and Goldberg (2019), we report ρ the Spearman correlation coefficient between the number of clusters a lexeme is assigned to and its number of senses according to annotations, to ensure that the number of created senses actually scales with the actual degree of polysemy.

Note that, for CI systems, we evaluate the division of occurrences provided by the *final* clustering Ψ (i.e. how occurrences are clustered after the global step and its potential merge operations). The quality of sense clusters induced by the local-step only is actually evaluated with Local-only systems.

⁵ Notions of *cognitive synonymy* and *near-synonymy* are discussed by Stanojevic (2009).

	WSI F_1	ρ
Local-only Systems		
K-means Local	.61	NA
Agglo Local	.77	.04
Eyal et al. (2022)	.46	.51
CI Systems		
K-means Global	.76	.51
K-means Bi-level	.78	.30
Agglo Global	.80	.53
Agglo Bi-level	.80	.46

Table 4.3: WSI BCubed F_1 and sense number correlation coefficient ρ on SemCor full data. Not computed for K-means because the number of cluster is constant.

Local results. Results of this local evaluation are displayed in Table 4.3. Let us recall that Local-only systems’ hyperparameters are chosen to maximize the WSI F_1 on the *Dev* split, while those of CI systems maximize the Concept Induction F_1 . Nonetheless, one can observe that all CI systems outperform their Local-only counterparts, achieving higher WSI F_1 and ρ even though their hyperparameters are not chosen to match the WSI itself. This indicates that the information brought at the global level by considering cross-lexicon relations may indeed help improving WSI, and benefits between local and global levels go both ways.

We explain the relatively poor performance of state-of-the-art WSI system by the fact that we are in a particular setting, where the number of occurrences per lexeme is relatively low in SemCor (30 per lexeme on average) and so is the average number of occurrences per concept. Data sparsity is a favorable ground for word senses to be misrepresented. As such, methods meant to be applied on larger datasets like the one of Eyal et al. (2022) may not work as well as expected. Our results offer two key insights: (i) WSI underperforms in low-resource or sparse-data settings. This raises a problem for the study of Lexical Semantic Change, since diachronic corpora are often small and unbalanced (Tahmasebi et al., 2021); (ii) focusing on concepts rather than on senses can improve performance, even with limited data, including for studying semantic evolution over time.

4.4 Extrinsic Evaluation with Concept-aware Embeddings

In their work, Eyal et al. (2022) derive *sense-aware* static embeddings from their WSI method, training them on the *Wikipedia* dataset and used them for the

Model	Acc.
Eyal et al. (2022) (CBOW)	59.3
Eyal et al. (2022) (Skip-Grams)	61.9
Ours (Agglo global)	58.8
Ours (Agglo bi-level)	59.7

Table 4.4: Accuracy scores on the nouns of the WiC test dataset (Pilehvar and Camacho-Collados, 2019).

Word-in-Context (WiC) task. They achieve nearly-SotA results on the dataset proposed by Pilehvar and Camacho-Collados (2019), and report to be outperformed only by methods using external lexical knowledge and resources.

We proceed to the same extrinsic evaluation of our work, constructing *concept-aware* embeddings using concept clusters of Concept Induction systems (Global-only and Bi-level Agglo). To obtain such embeddings, we average all vectors representating occurrences in SemCor contained each global cluster to get one vector per concept cluster.

The WiC task consists of determining whether two occurrences of a target lexeme w correspond to the same sense. The WiC dataset’s target words are nouns and verbs, but like in the rest of this paper, we restrict our scope to nouns.

To solve the task, we use BERT Large to create representations of the two target occurrences. Each of them is assigned to a concept by finding the closest concept-aware using cosine distance. The decision depends on whether the two occurrences are mapped to the same concept (*true*) or to distinct ones (*false*). Results are displayed in Table 4.4. Our concept-aware embeddings obtain very similar results to those of Eyal, Sadde, Taub-Tabib, and Goldberg’s sense-aware embeddings, with ours derived from our bi-level approach even outperforming their CBOW method⁶. Interestingly, our embeddings were trained with far fewer resources than theirs, as we used 52 997 occurrences from the SemCor dataset while they used a dump of Wikipedia, gathering millions of occurrences. This highlights the value of concept-aware embeddings: the use of cross-lexicon information allows competitive results with fewer resources.

⁶ After using their Word Sense Induction method to label all words of a corpus with a sense tag, they train a CBOW model that considers each sense as an independent type: therefore they obtain a set of sense-aware embeddings for each lexeme. To assign a sense embedding to a target word in a given context from the Word-in-Context dataset, they compute the average embedding of words in the context (the context embedding) and take the sense-aware embedding of the target word that is the closest to this context embedding.

4.5 Summary

In this chapter, we elaborated on the mapping existing between lexemes and concepts. This mapping is both many-to-one, as multiple words can be used to refer to the same concept (i.e. synonymy), and sometimes also one-to-many, as a single word can have multiple meanings (i.e. polysemy). We showed that current lexical semantic tasks in Natural Language Processing are not designed to face both sides of the mapping at once, thus refraining associated systems to be used to investigate both the semasiological and the onomasiological aspects of Lexical Semantic Change.

Then we introduced Concept Induction (CI), the task of clustering lexemes in groups of synonyms that share a common concept, using a corpus of their occurrences. We explicated how Concept Induction tackles both the semasiological and the onomasiological sides of the mapping between words and meanings, and compare several methods to address the task. Relying on vector representations of occurrences from Contextualized Language Models, we made use of global clustering models, that directly compare usages from different words; local clustering models, that only compare occurrences from each lexeme independently ; and proposed a bi-level method that rely on a dual clustering, to leverage the precision of local models and the inter-words comparisons made by global models. Overall, we show that the bi-level method outperforms other approaches, with BCubed F_1 above 0.6, and also that some clustering algorithms that put hard assumptions on cluster shapes such as K-means underperform in the Concept Induction task. We empirically observed that in our realistic dataset with variable number of occurrences across target lexemes, bi-level concept inducing systems were providing a better sense-differentiation than some Word Sense Induction (local-only) models, and we hypothesize that it may be caused by the relatively low number of occurrences of some lexemes, making it difficult for WSI models to significantly distinguish word senses, while CI models could make use of occurrences from other lexemes.

Limitations Evaluating Concept Induction is mainly limited by the low amount of suitable annotated corpora. Not only the data need to be annotated in concepts, but these annotations must cover a wide variety of lexemes for synonymy to be sufficiently represented in the corpus. Future work may find or create datasets meeting these requirements to evaluate Concept Induction outside of SemCor, especially for other part-of-speech than Nouns.

Our Bi-level method allows the global clustering to *merge* local clusters, lever-

aging lexicon-level information to be used to correct Word Sense Induction errors at the lemma-level. By its sequential nature, our method does not allow to *split* local clusters using global-level information, which could lead to better results. Further research directions include creating an iterative version of our methodology (alternating local and global clustering), or attempting to tackle both clustering objectives simultaneously with bi-level constrained clustering.

Our results about sense-induction at the local level showed that usual WSI methods may not be robust in our setting where there are few occurrences for some lexemes. We demonstrated that, in this setting, concept-inducing methods provided a better division in word senses. In many fields of linguistics, corpora are not very large and do not contain hundreds of occurrences for each word. Nonetheless, it is still uncertain if this observed advantage of CI systems would still hold on bigger datasets with many occurrences per word, a setting better-suited for usual WSI methods.

To choose the optimal parameters for each combination of clustering algorithms, we revealed a portion of the data’s annotations to tune our hyperparameters to maximize the BCubed F1 on this section of the data. This makes the overall approach ‘less unsupervised’, and in practice most available datasets are not concept-annotated, thus future work may investigate different approaches to tune hyperparameters or using algorithms with fewer and/or more interpretable hyperparameters. In particular, the HDBSCAN method is equipped with a simple and efficient algorithm to extract flat clusters from a hierarchy. Agglomerative Clustering usually requires to set an arbitrary distance threshold to obtain flat clusters from the hierarchy; in contrast, HDBSCAN’s method uses a single parameter that is the minimum number of samples considered in a cluster. More details on this method in Appendix A.3.1. In Chapter 6, we will be using this cluster-selection method on top of an Agglomerative Clustering hierarchy to perform Concept Induction in a context where we do not have concept-annotated data.

Besides limitations related to data and clustering operations, the main limitation of our approach is that it is based on the representational power of the Contextualized Language Model. Improving the clustering and having more data is only beneficial to the extent that the representation model can accurately capture semantic information and represent the similarities and differences in meaning between occurrences. However, pre-trained language models are not trained to perform only lexical semantic tasks, they also encode syntactical and morphological information (Ethayarajh, 2019). Therefore their representations may be sub-optimal with respect to how accurately and explicitly they reflect lexical

meaning.

Following up on this limitation, the next chapter is dedicated to the possible advantages of fine-tuning a Contextualized Language Model to obtain better representations of lexical meaning in occurrences.

5

CALE: Fine-tuning a Model for Concept-Aligned Embeddings

The Concept Induction models presented in the previous chapter are clustering systems, built on top of vector representations from a pretrained Contextualized Language Model. As such, this range of approaches is inherently limited by the semantic information conveyed within these representation models and the extent to which embeddings reflect similarities and differences in meaning between occurrences. A limited performance in the Concept Induction task could be explained by an inadequate clustering method, but it could also be the result of sub-optimal semantic representations. In general, any NLP approach to lexical semantic change based on contextualized representations suffer from the same limitation. Vector representations of the lexical meaning should reflect that occurrences instantiating similar concepts have close representations in the vector space, and occurrences of identical meaning should have the smallest distance between their representation vectors. In the pursuit of the extraction of semantic information from contextualized representations, the previous chapter has explored differences in clustering systems; the present chapter focuses on representations, at a more basic level. We have seen in Chapter 3 that the XL-LEXEME model (Cassotti et al., 2023) has improved representations with a sense-differentiation objective and reached state-of-the-art results in the task of comparing usages to predict a word's change in meaning, but two questions remain. First, would it be interesting to fine-tune a representation model not only in the same-lexeme setting, as it is the case for XL-LEXEME, but also in the cross-lexeme perspective ? And second, do these models encode lexical meaning more accurately and in a more explicit fashion than pre-trained models ? Pre-trained Contextualized Language Models indeed not only encode semantic information, the representations they produce also convey syntactic and morphological aspects, which are likely not relevant to our research objectives.

This chapter aims at using of Contextualized Language Models to obtain multi-

purpose (as opposed to “task-specific”) vector representation of word meaning. Word occurrences that instantiate the same concept in respective contexts should have similar representations, while occurrences of words that refer to different concepts should have dissimilar embeddings. A model with this desired property would be useful in studies of the mapping between word forms and concepts, like Haspelmath (2023) for instance, or in any computational lexical semantic applications mentioned in Chapter 3. As such, we expect the representations not to be constrained to a specific task but rather versatile for various downstream usages. As in Chapter 4 we focus on synonymy and polysemy, leaving to future work the integration of other relations such as hyponymy.

We start this chapter by a short literature review on existing contextualized models and their benefits for representation of lexical meaning. Then we introduce a new task, *Concept Differentiation*, related to *Word-in-Context* and *Concept Induction*. We create a dataset of occurrence pairs for this task, accounting for both same-lexeme and different-lexemes scenarios. This task serves as our fine-tuning objective to obtain better models of lexical meaning representation. We describe how we fine-tune Contextualized Language Models for this objective, using the framework of Contrastive Learning. We conclude the chapter with two series of experimental comparisons: first, we compare the obtained models (dubbed ‘CALE’ for Concept-Aligned Embeddings) to other models on several benchmarks of different lexical semantic tasks (see Section 3.1 of Chapter 3 for an overview of these tasks); second, we compare the spatial properties of the different models to discuss the general organization of their occurrence representations.

5.1 Related Work on Semantic Representations

Contextualized word embeddings from models like ELMo (Peters et al., 2018) and BERT (Devlin et al., 2019) generate dynamic representations based on usage in context, enabling better semantic modeling than earlier static or sparse methods. However, pooling token-level embeddings into meaningful sentence-level vectors remains challenging due to issues like anisotropy in the embedding space (Reimers and Gurevych, 2019). To improve semantic similarity and retrieval, Sentence-BERT (SBERT) (Reimers and Gurevych, 2019) introduced a Siamese training architecture, followed by improvements using contrastive learning (Gao et al., 2021) and multi-task objectives (Zhang et al., 2022) to enhance cross-task generalization. Such models have been adapted for lexical semantics. Mosolova et al. (2024) fine-tune language models using Wiktionary supervision to improve contextualized token embeddings for tasks like Word-

in-Context. Similarly, the XL-LEXEME model (Cassotti et al., 2023), based on XLM-RoBERTa, extends SBERT for multilingual word meaning tasks, showing robustness on LSCD benchmarks and supporting cross-lingual, context-sensitive embedding spaces.

Other efforts fine-tune BERT for word-level classification tasks (Garí Soler and Apidianaki, 2020), though these cross-encoder models often lack flexibility for single-occurrence representation, which represents a limitation for tasks requiring generalized concept modeling. Embedding-based models have also proven effective in capturing non-literal or stylistic nuances, as shown by applications in authorship verification (Rivera-Soto et al., 2021). Additionally, SenseBERT (Levine et al., 2020) enhances contextual embeddings by incorporating sense information from WordNet during pretraining, improving performance on sense-sensitive tasks and highlighting the benefit of explicitly modeling lexical semantics in transformer-based representations. Collectively, these developments demonstrate the adaptability of sentence-transformer architectures for fine-grained semantic representations.

5.2 Concept Differentiation

Our primary goal is to build a model able to accurately represent semantic information in order to distinguish concepts in context, from which we can obtain concept-aligned embeddings. To train this model, we rely on a supervised task that captures both polysemy and synonymy, which we label *Concept Differentiation*, a supervised alternative to Concept Induction.

5.2.1 Task Definition

We define *Concept Differentiation* as a binary classification task where the goal is to determine whether two word usages represent the same underlying concept, regardless of whether they originate from the same lexeme or different lexemes. For illustration, consider the following three examples:

1. the boy could easily distinguish the different note values
2. the patient’s ability to recognize forms and shapes
3. the government had refused to recognize their autonomy and existence as a state

The verbs *distinguish* in (1) and *recognize* in (2) both express the concept of DIS-CERN — to tell things apart — so the pair (1-2) is labeled 1 (same concept). In

contrast, while (2) and (3) both use *recognize*, only (3) refers to ACKNOWLEDGE, so that pair (2-3) is labeled 0 (different concepts). This task is broader than traditional Word-in-Context, which is limited to same-lexeme comparisons¹. Here, we extend the comparison to any pair of word usages — same lexeme or not — capturing finer and more flexible distinctions of meaning. It can also be seen as a binary reformulation of the Concept Induction task presented in the previous chapter, where instead of clustering occurrences into concept clusters, we ask if two usages share a common concept. This setup allows us to train models that learn to compare usages in context, effectively capturing subtle semantic distinctions or consistencies between different instances of target words across the corpus.

5.2.2 Dataset for Concept Differentiation

In the absence of a dedicated dataset for this task, we construct a new dataset using the SemCor corpus, using its WordNet synsets annotations (Miller, 1995; Fellbaum, 1998) as concept labels. For each lexeme with enough labeled instances, we retrieve its occurrences following the procedure detailed in Appendix A.2.1. We gather 70.3k occurrences for 1902 target lexemes, covering 5899 concepts. To create labeled training pairs, we sample pairs of word occurrences (o_j, o_k) and assign a label of 1 if both instances are annotated with the same concept, and 0 otherwise. Let us underline that, contrarily to the previous chapter, we use the notation o_i for occurrences instead of $o_{w,i}$ and this new notation does not specify the involved lexeme. Indeed, while for Concept Induction we needed to know the lexemes in order to derive a word-concept mapping from the clustering of occurrences, for Concept Differentiation the fact that the two occurrences come from the same lexeme or from two different words does not matter, the only important aspect is whether they refer to the same concept or not. For the sake of clarity and to avoid unnecessary ambiguity, we will use notation o_i for occurrences in this chapter.

To rigorously assess generalization, we enforce a strict separation between training, validation, and test sets by first partitioning the set of concepts and the set of target words into disjoint subsets for each split. Specifically, we randomly sample subsets of concepts and lexemes to be assigned exclusively to the validation or test split (5% of concepts / lexemes in validation, 10% in test). Once these partitions are defined, we extract all word occurrences from the SemCor corpus that are annotated with the selected concepts and lexemes, and include

¹ This task formulation is also similar to the annotation guidelines for the ‘different words for the same meaning’ case in Baldissin et al. (2022)

them in the corresponding data split. This ensures a more rigorous evaluation of a model’s ability to generalize to entirely unseen meanings and lexical items, preventing models from relying on memorized associations between words and concepts. In each split separately, we create pairs by attempting to match each occurrence to 4 others, one for each of the following categories if possible: same concept and same lexeme (SC/SL), same concept but different lexeme (SC/DL, only if we can find an occurrence of another lexeme for the concept), different concept but same lexeme (DC/SL, only if the word is polysemous) and different concept and different lexeme (DC/DL, unrelated pairs). The dataset contains 44k pairs (from the 14.3k reserved occurrences) in the *test* split, 20k (from the 6.5k reserved occurrences) in the *validation* split, and 156k (from the 49.6k remaining occurrences) in the *train* split. In all splits, the resulting proportion of label ‘1’ (same concept) is around 40%.

The resulting dataset, later referred as *SemCor Pairs for Concept Differentiation* (SPCD), supports research in semantic representation and word sense disambiguation and is made publicly available to facilitate future benchmarking and exploration². Further description as well as more details about the extraction process can be found in Appendix A.2.1.

5.3 Concept-Aligned Embeddings

We use our resulting pair classification dataset to train a representation model. The goal is to learn a function f mapping an occurrence o_i to a vector representation e_i in a vector space in which the embeddings of two occurrences of the same concept have higher cosine similarity than embeddings of two occurrences of a different concept. To fit f , our architecture consists of a pre-trained contextualized language model followed by a pooling layer to obtain fixed-dimensional representations. We fine-tune this entire model using SentenceBERT-style siamese training and contrastive learning, which proceeds as follows.

Each training step considers a batch of occurrence pairs (o_i, o_j) with an associated label $y_{ij} \in \{0, 1\}$, where $y_{ij} = 1$ indicates that both occurrences share the same semantic concept, and $y_{ij} = 0$ otherwise. Occurrences of the pair are encoded into vector representations (e_i, e_j) by the Siamese models (two instances of the same model with shared weights during training). We then compute the cosine similarity between the two embeddings $\text{sim}_{ij} = \cos(e_i, e_j)$. The training objective is to maximize similarity for positive pairs ($y_{ij} = 1$) and minimize it for negative pairs ($y_{ij} = 0$). We fit the model to minimize a contrastive loss (Hadsell et al.,

²<https://hf.co/datasets/gabrielloiseau/CALE-SPCD/>

2006) defined as:

$$\frac{1}{2} [y \cdot \text{sim}_{ij}^2 + (1 - y) \cdot \max(0, m - \text{sim}_{ij})^2]$$

where m defines the minimum distance that should exist between dissimilar (negative) pairs in the embedding space, known as the margin.

This contrastive training encourages the model to organize the embedding space such that occurrences sharing the same concept are pulled closer together, while unrelated occurrences are pushed apart. Fine-tuning under this objective enables the model to produce embeddings that better reflect conceptual similarity across varied lexical and contextual realizations.

In practice, an occurrence o_i is a sequence of words $(t_1, \dots, t_n)$ with a target word w in position k ($t_k = w$). To be used as input for the model, the occurrence is transformed as follows:

$$\underbrace{t_1, \dots, t_{k-1}}_{\text{left context}}, <\text{t}>, w, </\text{t}>, \underbrace{t_{k+1}, \dots, t_n}_{\text{right context}},$$

using special tags $<\text{t}></\text{t}>$ to delimit the target word in the sentence. During training, the model will learn that the contrast must be made based on the meaning of words between the delimiters. In doing so, we ensure the embedding corresponding to the full sequence is the representation of o_i and distinct from the embedding of another target word occurring in the same sentence, without needing to account for potential subwords that may result from the model’s tokenizer. This method, also used by Cassotti et al. (2023), makes the model easy to use in practice at inference time, as we only have to add/move the delimiters to shift the model’s focus from one word to another in the sentence.

We refer to this fine-tuning approach and the resulting models as **CALE** (*Concept Aligned Embeddings*) and adopt this terminology throughout the paper. Our models are made publicly available through HuggingFace³. The scope of our study is limited to single-word targets, as it is the primary focus of this thesis; but the framework we just presented could be applied to capture the meaning of Multi-Word Expressions in future work.

³<https://huggingface.co/collections/gabrielloiseau/cale>

5.4 Model Evaluation

In this section, we report the performances of different models, including models with CALE fine-tuning, for the Concept Differentiation task on the test split of the *Semcor Pairs for Concept Differentiation* (SPCD) dataset, as well as other external lexical semantic tasks, namely Lexical Semantic Change Detection and In-context Lexical Similarity. It should be noted that the Lexical Semantic Change Detection task is only relevant to this chapter insofar as it is a complex within-lexeme task, enabling us to evaluate whether CALE models actually represent differences in meaning between occurrences of a given lexeme. The discussions in this chapter pay minimal attention to the task’s diachronic dimension. The extent to which CALE models can be used for a complete diachronic study, and the benefit that CALE representations may have for onomasiological and semasiological studies of change will be discussed in more depth in Chapter 6 and 7.

Compared Models. XLM-R (Conneau et al., 2020) is based on the RoBERTa Large architecture and pretrained on cross-lingual data. We consider three sets of layers to extract embeddings from : layers 1 to 4 (dubbed *first*), layers 14 to 17 (dubbed *mid*⁴) and the last four layers (21 to 24, dubbed *last*). Embeddings of word occurrences are obtained by averaging first over selected layers and then over subword tokens selection. XL-LEXEME (Cassotti et al., 2023) is a XLM-R model fine-tuned for multilingual WiC. It is a SentenceBERT model, meaning that a single embedding is produced per occurrence (using a final pooling layer). Compared to our models, XL-LEXEME’s fine-tuning was rather data-hungry, as three Word-in-Context datasets were used. XL-LEXEME established a new State-of-the-Art in some semantic tasks and is therefore our main reference point. Regrettably, models from Garí Soler and Apidianaki (2020) and Mosolova et al. (2024) are not publicly available, preventing us to include them in our experiments without re-implementing the whole dataset curation process and fine-tuning method. Our CALE models generate embeddings in a similar fashion than XL-LEXEME, but they are fine-tuned on Concept Differentiation, as described earlier. We experiment with several base models: XLM-R (to compare with XL-LEXEME) XL-LEXEME (to observe the impact of double fine-tuning and of XL-LEXEME’s fine-tuning multilinguality) and a monolingual English ModernBERT (Warner et al., 2024), a more-recent State-of-The-Art actualization of the BERT model⁵, evaluated only in English datasets in order to discuss the potential

⁴ Chronis and Erk (2020) showed that late-intermediate layers were best for tasks related to semantic similarity

⁵ ModernBERT was not presented in Chapter 3 because, to the best of our knowledge, it has not yet been used as part of a computational model of lexical semantic change.

	All	SL	DL
1L1C (baseline)	65.1	50.0	50.0
XLM-R (first)	60.8	54.2	50.0
XLM-R (mid)	70.9	59.1	68.4
XLM-R (last)	69.1	59.6	63.5
XL-LEXEME	76.7	62.4	82.7
CALE (XLM-R)	79.3	65.9	86.4
CALE (MBERT)	78.7	66.6	83.5
CALE (XL-LEX)	79.2	67.2	84.4

Table 5.1: Balanced Accuracies on test pairs from SPCD. “SL” (resp. “DL”) refers to pairs of occurrences of the same lexeme (resp. of different lexemes).

difference between language specialization and pre-training multilinguality.

5.4.1 Concept Differentiation Evaluation

In Table 5.1 we present the results for the Concept Differentiation task on the test split of the SPCD datasets. For each model, the classification process for a given pair of occurrences is the following: we compute the embeddings for each occurrence using the candidate model, then compare the cosine distance between them to a threshold to label the pair: 1 (same concept) if the cosine distance below the threshold, 0 (different concepts) otherwise. The threshold is chosen to maximize accuracy on the *train* and *validation* splits. We also provide a *baseline system*, where we label a pair as 1 if and only if the two occurrences are from the same lexeme. We refer to this baseline as 1L1C for “1 lexeme 1 Concept”. We use Balanced Accuracy (the average of the Recalls from both classes) as the classes are not exactly balanced. We also tried F1 scores and the observed tendencies are the same, thus we only report them in Appendix A.2.3, with also a more complete overview of accuracies per class.

Our first observation is that the 1L1C baseline reaches high accuracy, providing a challenging baseline. This is in line with prior observations made for Concept Induction in Section 4.3.4 of Chapter 4. This indicates that the mapping between words and concepts in SemCor may resemble a one-to-one mapping with some deviations, and as such a model for Concept Differentiation would benefit from not ignoring whether two occurrences are of the same lexeme or not. We also observe that XLM-R performs best using late-intermediate layers than other layers and outperforms the baseline verifying findings of Chronis and Erk (2020) that these layers are best for tasks related to semantic similarity.

We also find that XL-LEXEME largely outperforms XLM-R even when presented

pairs of Different lexemes, meaning that by learning to differentiate senses of the same lexeme, it captured more global information about relations between different words.

We find that CALE models outperforms XL-LEXEME. It is expected to some extent, as Concept Differentiation is the task CALE has been fine-tuned for. We also observe that Monolingual CALE (MBERT) is not better than the multilingual ones, underlining the advantage of multilingual pre-training.

The CALE model based on XL-LEXEME is globally on-par with the one based on XLM-R (79.3 and 79.2 balanced accuracies). The former is slightly better in the Same-lexeme setting (i.e. better at distinguishing senses) than the latter, but slightly worse in the Different-lexeme setting (i.e. worse at identifying synonyms), as expected because XL-LEXEME has been trained strictly in the Same-lexeme setting and therefore likely provides benefit in this context only.

5.4.2 Lexical Semantic Change Detection.

To assess models’ ability to capture meaning variation within lexemes across contexts, we evaluate on the Lexical Semantic Change Detection (LSCD) task, following “Subtask 2” of Schlechtweg et al. (2020). We use Diachronic Word Usage Graphs (DWUGs) datasets across six languages: English, German, Latin, Swedish, Spanish, and Italian (Schlechtweg et al., 2020; Zamora-Reina et al., 2022; Cassotti et al., 2024a)⁶. As we have seen in Section 3.1.3, each dataset contains target words with occurrences sampled from two time periods (T_1 , T_2), and the task is to assign a change score reflecting how much a word’s meaning has shifted. LSCD is a within-lexeme task: only usages of the same word are compared and semantic change is assessed by measuring differences in the word’s senses between time periods. Performance is measured by Spearman correlation between predicted and gold change scores. Cassotti et al. (2023) also evaluated their model XL-LEXEME on these benchmarks and established new State-of-the-Art scores in English, German and Swedish.

We use the Average Pairwise Divergence (APD) measure from Kutuzov and Giulianelli (2020) to compute change scores, relying on cosine distances between embeddings from T_1 and T_2 :

$$\text{APD}(w) = \frac{1}{n \times m} \sum_{i=1}^n \sum_{j=1}^m d(e_{w,i}^1, e_{w,j}^2)$$

with d the cosine distance operator, $e_{w,i}^k$ the embedding of the i -th occurrence of

⁶ <https://www.ims.uni-stuttgart.de/en/research/resources/experiment-data/wugs/>

	EN	DE	ES	IT	LA	SV
# targets	46	50	100	26	40	44
XLM-R (first)	.22	.14	.08	.18	.00*	.10
XLM-R (mid)	.57	.54	.49	.32	.24*	.58
XLM-R (last)	.56	.34	.53	.27	.00*	.52
XL-LEXEME	.79	.80*	.63*	.51*	.14*	.85*
CALE (XLM-R)	.71*	.78*	.65*	.75	.11*	.83*
CALE (MBERT)	.66*	-	-	-	-	-
CALE (XL-LEX)	.73*	.83	.68	.54	.20*	.86

Table 5.2: Spearman Correlation between model Average Pairwise Divergence (APD) and gold change scores in DWUG datasets. The highest score for each dataset is in bold, and other scores that are not significantly different from it are marked with *.

word w at time T_k , for a target word w with n occurrences at T_1 and m occurrences at T_2 . APD is generally accepted as standard for Lexical Semantic Change (see the survey study of Periti and Montanelli (2024)) and was used by Casotti et al. (2023) to evaluate XL-LEXEME on the same task on DWUGs datasets. In Appendix A.2.4 we tried PRT, another measure from Kutuzov and Giulianelli (2020).

We report APD-based results in Table 5.2. We use Steiger’s Z-test on dependent correlations to determine whether differences to the best-performing model are significant or not (Steiger, 1980).

Results show that CALE fine-tuning consistently improves over base XLM-R in all languages, even though it was trained on English data only, showcasing the cross-lingual transferability of XLM-R. CALE also matches or slightly outperforms XL-LEXEME in 5 out of 6 languages, despite XL-LEXEME being fine-tuned on larger multilingual datasets. As in prior work, Latin proves challenging across all models.

While XL-LEXEME is fine-tuned for Word-in-Context (a lexeme-centric task closely related to LSCD), CALE is optimized for Concept Differentiation, a broader task involving both within- and between-lexeme comparisons. One might expect this to hinder LSCD performance, yet results suggest otherwise: CALE (XL-LEXEME) slightly improves over XL-LEXEME alone, suggesting that Concept Differentiation enhances a model’s ability to detect semantic change. Due to the limited number of target words, differences are not statistically significant, but findings indicate that CALE fine-tuning does not degrade LSCD performance.

SUBTASK 1 : CONTEXTUAL CHANGES				
	EN	FI	HR	SL
# examples	340	24	112	111
XLM-R (first)	.18	.32 [†]	.20	.33
XLM-R (mid)	.69*	.36 [†]	.58	.52
XLM-R (last)	.63	.27 [†]	.51	.45
XL-LEXEME	.70	.47*	.67*	.66*
CALE (XLM-R)	.74*	.63	.72	.70
CALE (MBERT)	.68	-	-	-
CALE (XL-LEX)	.74	.60*	.72*	.67*

SUBTASK 2 : SIMILARITY RATINGS				
	EN	FI	HR	SL
# pairs	680	48	224	222
XLM-R (first)	.27	-.04 [†]	.25	.25
XLM-R (mid)	.61	.25 [†]	.53	.51
XLM-R (last)	.59	.16 [†]	.52	.46
XL-LEXEME	.63	.59*	.64	.67*
CALE (XLM-R)	.67	.64	.68*	.68
CALE (MBERT)	.67*	-	-	-
CALE (XL-LEX)	.67*	.57*	.73	.68*

Table 5.3: Pearson Correlation of CoSimLex Subtask 1 (upper table) and Spearman Correlation of CoSimLex Subtask 2 (lower table). Values marked with [†] are non-significant correlation values. The highest score for each dataset is in bold, and other scores that are not significantly different from it are marked with *.

5.4.3 In-Context Lexical Similarity.

We now evaluate candidate models in the cross-lexeme setting. To assess our models' ability to represent conceptual similarity of two distinct words in a given context, we use the CoSimLex dataset (Armendariz et al., 2020) for In-Context Lexical Similarity. This resource proposes two subtasks in 4 different languages, English, Finnish, Croatian and Slovene. Each entry of CoSimLex is a pair of target words, and two different contexts in which both words appear. For each context is provided a measure of lexical similarity between the two words. Subtask 1, Contextual Changes, aims at predicting the change of similarity between the two words, from context 1 to context 2. This is measured with Pearson Correlation coefficient between the differences of gold similarities and the differences of predicted similarities. Subtask 2, Similarity Ratings, adopts a more traditional setting, where candidate systems are simply tasked to predict a mea-

	Silh (S)	Silh (UB)	WuP ρ
XLM-R (first)	-0.166	-0.267	0.187
XLM-R (mid)	-0.064	-0.261	0.338
XLM-R (last)	-0.088	-0.260	0.311
XL-LEXEME	-0.124	-0.370	0.489
CALE (XLM-R)	-0.041	-0.192	0.509
CALE (MBERT)	0.042	-0.104	0.514
CALE (XL-LEX)	0.003	-0.228	0.532

Table 5.4: Silhouette scores with synsets (S) or with Unique Beginners (UB) as cluster labels ; and Spearman Correlation of cosine similarities with Wu-Palmer’s similarities in WordNet.

sure of lexical similarity between the two words in each context, and measuring the Spearman Correlation between gold measures of similarity and predicted ones. Subtask 2 uses independently each context provided in Subtask 1 instead of pairing them, and as such covers twice as many examples. To predict lexical similarity between the target lexemes in a given context for both subtasks, we compute the embeddings corresponding to each target and compute the cosine similarity between them. Results are displayed in Table 5.3.

CALE models reach the highest correlations values in all languages, ranging between 0.63 to 0.73 in subtask 1 and 0.57 to 0.68 in subtask 2, and consistently improve over XLM-R. Surprisingly, XL-LEXEME also performs very well and is most often not deemed significantly different from CALE results, despite being trained only on a same-lexeme task and not on between-words relations. CALE (XL-LEX) does not perform significantly better than the more simple CALE (XLM-R), indicating that CALE fine-tuning is usually sufficient to obtain accurate in-context similarities. Again, the good performances reached by CALE fine-tuning highlight the cross-lingual transferability of XLM-R models.

5.5 Spatial Organization

We investigate how CALE fine-tuning affects representation space structure. Figure 5.1 shows cosine distance distributions across the *test* split of SPCD by pair category, while Table 5.4 analyzes how well embedding spaces reflect WordNet’s conceptual organization using synsets annotations.

Distance distributions. In Figure 5.1 we plot the distributions of distance in pairs according to their categories: ‘SL’ and ‘DL’ refer to “same lexeme” and “different lexemes” respectively, ‘SC’ and ‘DC’ to “same concept” and “different concepts” respectively. We first observe that XLM-R suffers from major anisotropy

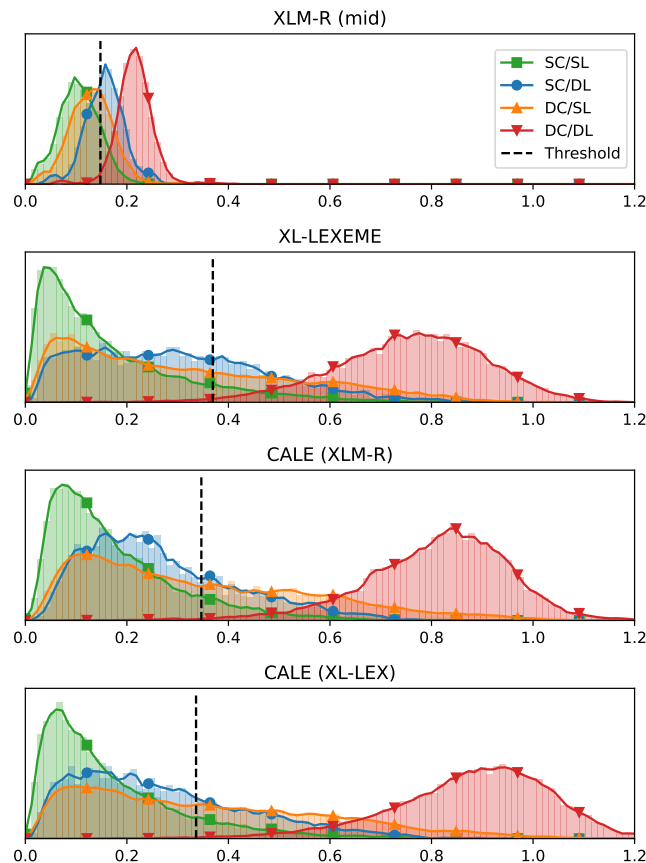


Figure 5.1: Distribution of cosine distance between embeddings from the candidate models’ spaces for SPCD *test* pairs, according to their category. The dashed line is the decision threshold for Concept Differentiation learned from the other splits.

(i.e. all cosine distances are low even in unrelated pairs). The fine-tuning of CALE and XL-LEXEME push occurrences in DC/DL pairs away from one another, a consequence of the margin parameter. Another observation is that CALE’s fine-tuning successfully shifts the focus from *lexeme*-centric to *concept*-centric: in XLM-R, DC/SL pairs (cases of polysemy) are generally closer than SC/DL pairs (cases of synonymy); this tendency is reversed in CALE (XLM-R) (average distance of .130 for DC/SL vs .157 for SC/DL before fine-tuning; .343 vs .269 after fine-tuning).

Similarity correlations. WordNet’s synsets are linked to one another through relations of hyper-/hyponymy, in a tree-like structure. A short path between two synsets indicate that the two concepts are similar. We compute Spearman correlations between model’s cosine similarities and Wu-Palmer similarities (Wu and Palmer, 1994) between synsets for each occurrence pair⁷. CALE models achieve

⁷ We also tried Lin similarity, with similar results.

the highest correlations (above 0.5), indicating a strong alignment between embedding similarity and semantic similarity. All correlations and pairwise differences are found statistically significant.

Cluster analysis. We evaluate whether occurrences of the same concept form tight, well-separated clusters using the Silhouette score, with WordNet synsets as labels. Given o_i the i -th occurrence found in corpus O , represented with e_i and assigned to cluster y_i , and denoting $d_{sc}(e_i)$ the average distance between e_i and all other members of the same cluster; and $d_{nn}(e_i)$ the average distance between e_i and all other members in the next nearest cluster, Silhouette score is defined as:

$$s = \sum_{i=1}^{|O|} \frac{d_{nn}(e_i) - d_{sc}(e_i)}{\max(d_{sc}(e_i), d_{nn}(e_i))}$$

The Silhouette score s lies between -1 and +1 and assess the general separability of clusters: if s is close to 1, it means the clusters are well-separated; values close to 0 indicate that there are some overlaps, while values close to -1 reflect strong overlaps between clusters and a lack of separability. Using true synset labels as cluster labels, we assess whether semantic concepts are well-separated in the representational space of a model, not only in terms of pairwise distance but in terms of clusters.

CALE models yield the highest scores, suggesting better conceptual clustering. However, all scores remain below 0.1, highlighting substantial overlap between clusters even for the best models, which is likely due to the fine granularity of WordNet synsets.

Broader clusters. We repeat the clustering analysis using WordNet’s Unique Beginners (UB) broad semantic categories like “entity”, “feeling”, or “communication” as labels, limited to nouns and verbs because such Unique Beginners do not exist for adjectives. Overall scores are negative, indicating poor alignment with these higher-level groupings. Nonetheless, CALE performs best among models, suggesting a modest ability to reflect coarse semantic distinctions.

Conclusions. These findings show that CALE embeddings align more closely with WordNet’s structure than other models, both in distances and clustering. Notably, ModernBERT-based CALE shows stronger structural alignment than multilingual models. XL-LEXEME, despite good pairwise similarity, shows weaker clustering, presumably because its fine-tuning was sense-level and did not require broad conceptual organization.

5.6 Conclusions and Limitations

We introduced *Concept Differentiation*, a supervised task for predicting whether two occurrences share the same semantic concept. We curated the SPCD (*Semcor Pairs for Concept Differentiation*) dataset from SemCor and developed CALE models (*Concept Aligned Embeddings*) using contrastive fine-tuning of SentenceBERT-style architectures.

Evaluation shows that CALE improves lexical semantic representations across tasks, including Concept Differentiation, Lexical Semantic Change Detection, and Lexical Similarity. These gains hold for both within-lexeme and between-lexeme tasks, with CALE matching or surpassing strong baselines like XL-LEXEME, even on multilingual benchmarks despite English-only training. This suggests that Concept Differentiation subsumes other lexical tasks dealing with the mapping between words and meanings, and that a broader view is beneficial for multi-purpose semantic representations.

Analysis of CALE’s embedding geometry reveals a shift from lexeme-centric to concept-centric representations and closer alignment with WordNet’s organization. These results position Concept Differentiation as a promising general-purpose training signal for lexical semantics.

Limitations While CALE models demonstrate cross-lingual transferability, fine-tuning was performed exclusively on English language data. Even if multilingual models (e.g., XLM-R) generalize well to other languages in tasks like LSCD and CoSimLex, the performance in lower-resource or typologically distant languages has not been thoroughly evaluated due to the lack of data sources. A more rigorous multilingual training and evaluation setup is needed to assess CALE’s robustness across languages.

Second, the binary nature of Concept Differentiation may oversimplify semantic similarity, which is often graded rather than categorical. This could limit the model’s applicability in tasks requiring nuanced semantic judgments.

Third, while we show improved alignment with WordNet’s conceptual structure, our models receive no direct supervision about ontological hierarchy or fine-grained relations between concepts (e.g., hyper-/hyponymy, holo-/meronymy, antonymy). Further work is needed to evaluate and potentially incorporate structured knowledge explicitly during training.

Lastly, we relied on small architectural modifications compared to XL-LEXEME for model comparison, a key aspect of our studies is model and data scaling.

However, other scaling axes, notably in terms of model parameters and embedding dimension size are left unexplored.

5.7 Summary of Part II

In Chapter 4, we have introduced Concept Induction, the task of learning from data a clustering of a set of target words into groups of synonymous lexemes referring to the same concept. This task provides a perspective on the mapping between words and concepts, and allows evaluating a system’s ability to model both polysemy and synonymy at the same time. We have proposed a methodology based on contextualized representations derived from a pretrained language model, performing a bi-level clustering of these occurrences to obtain concept clusters. By examining the word groupings that arise from concept clusters, we can obtain the soft-clustering of words into groups of synonyms corresponding to a common, latent concept. These concepts are latent in the sense that they are not explicitly labeled nor pre-defined, they are completely learned from data using similarities between contexts.

This bi-level clustering method extracts semantic knowledge about concepts from contextualized representations. In Chapter 5, we have elaborated methods to fine-tune a model in order to obtain representations of lexical meaning that are semantically rich and explicit. To this end, we have introduced a task that reframes Concept Induction into a binary classification problem, and used contrastive learning to fine-tune a series of models (CALE) for this objective.

Combining these two chapters, we have addressed the problem of extracting concepts from contextualized representations in the synchronic setting: semantic information can be efficiently conveyed by CALE representations, and can be further retrieved using bi-level clustering to obtain a view of the word-meaning mapping as it emerges in the data. As the present thesis aims to develop models for studying temporal changes in the mapping between words and concepts, this overall approach is promising but currently limited to the synchronic setting. The next part, and final round of our contributions, is dedicated to the diachronic setting and the application of our methodology for investigating both semasiological and onomasiological aspects of lexical semantic change. Chapter 6 in particular describes the adaptations made to the methodology to account for the specificities of the diachronic setting, and evaluates the changes identified by the proposed method.

Part III

Diachronic Contributions: Studying Semasiological and Onomasiological Changes

6

Diachronic Concept Induction: Learning Concepts in Historical Data

Chapter 3 has provided an overview of the recent computational approaches to the modeling of lexical semantic change relying on contextualized representations from Language Models. Yet, these approaches adopt a word-centric perspective, primarily focusing on detecting semasiological changes, as such they often overlook the onomasiological aspect of semantic change. In Chapter 4 we have investigated how pre-trained Language Models can effectively be used to extract latent concepts and the clusters of words related to them. In Chapter 5 we proposed a series of fine-tuned, multilingual Language Models specifically designed to generate concept-aware representations, thus improving over state-of-the-art lexical semantic representations. The approaches proposed in Chapter 4 and 5 both explore the ability of Contextualized Language Models to represent lexical meanings as concepts within specific contexts. Beyond mere representations, the clustering method proposed for Concept Induction can be leveraged to retrieve the concepts instantiated in the data and the sets of words associated to them.

By tracking the evolution of concept clusters over time – specifically, the changes in the lexemes that belong to these clusters – it becomes possible to detect both onomasiological and semasiological changes. This is a promising improvement over the current state-of-the-art computational approaches to lexical semantic change, usually limited to detecting semasiological changes and are therefore unable to fully characterize the nature of the observed change. However to achieve *Diachronic* Concept Induction in historical data, certain adaptations will be necessary to address the specific constraints of this setting. Specifically, the method must be able to learn concept clusters across multiple time periods, accounting for temporal variation among occurrences; and the method must be fully unsupervised, as there are usually no semantic annotations in historical

corpora. This latter requirement contrasts with the method presented in Chapter 4, where the clustering hyperparameters were chosen by revealing a subset of the concept annotations.

This chapter we will (1) discuss how we account for temporal variability in the data and define the semasiological and onomasiological changes we expect to identify with a concept-inducing approach; (2) discuss the chosen algorithms, hyperparameters and how they are tuned in unannotated historical data; (3) discuss how to evaluate the quality of the onomasiological and semasiological changes uncovered by a diachronic concept-inducing system; (4) apply the described method to a French historical corpus, namely the PRESTO corpus, and discuss how we can leverage it to investigate both semasiological and onomasiological changes.

6.1 Learning Concepts in Diachronic Data

Chapter 4 has introduced a *bi-level* methodology based on a dual clustering over Contextualized Language Models representations to extract sets of words instantiating a common concept in the data, but only in a synchronic setting. In the diachronic setting, word occurrences may come from various historical origins, which calls for a method to learn concepts with such a temporal variability. In this section, we propose to learn concept clusters across all time periods at once, therefore assuming that there is a set of stable concepts to which word occurrences can be mapped. After discussing this assumption, we define how semasiological and onomasiological changes could be uncovered in this setting.

6.1.1 ‘All-Periods’ Concept Clustering

Periti and Tahmasebi (2024b) discussed several possibilities to learn clusters over multiple time periods, including (a) learning a stable set of clusters in occurrences from all periods at once; (b) learning a set of clusters from occurrences of each period individually and aligning the clusters *a posteriori*; (c) learning clusters in a sliding window of several time periods, relying on shared occurrences between two windows to provide cluster evolutions; etc. Unfortunately, their study did not implement these different strategies and there exists no systematic comparison of these different approaches.

In our investigation, we choose to learn clusters across occurrences from all periods simultaneously. This approach offers a significant advantage: using the same cluster labels across all time periods makes it easier to track the change of

a given concept, or to compare word meanings over time. In contrast, a strategy consisting in learning clusters in each time period individually would require an additional alignment step to compare clusters from the multiple periods and identify, for each new cluster, whether one of the previous cluster might be its ancestor. To achieve this ‘*all-periods*’ clustering, the bilevel methodology presented in Section 4.2 of Chapter 4 learns proto-sense clusters from all of the occurrences of a word, regardless of their temporal origin. The global clustering step is then performed on all the obtained proto-sense clusters. The distinction between time periods is only made after the clusters are learnt, in the analyses of the clustering’s results. Clustering occurrences from all time periods simultaneously implies that we must make two hypotheses: that representations accurately reflect lexical meaning in all time periods and are comparable over time; and that concepts are drawn from a fixed set.

Synchronic similarities and differences. We expect an embedding model to accurately reflect similarities and differences in lexical meaning within a single time period: occurrences instantiating the same concept should be represented as close to one another, while occurrences referring to different concepts should be more distant in the embedding space. In Chapter 4 and 5, we demonstrated that Contextualized Language Models are able to reflect these (dis)similarities in concept and lexical meaning, especially if they were fine-tuned on lexical semantic tasks, such as XL-LEXEME with cross-lingual Word-in-Context tasks, or CALE models with Concept Differentiation.

Diachronic accuracy. We expect the representations from different time periods to be comparable: as we showed in Section 3.2 of Chapter 3, some representation models required an extra alignment step to allow fair comparison of their embeddings across different time periods. If the representational vector spaces for different periods are trained independently, the constructed dimensions may not be aligned and thus the spaces may not be comparable. This is particularly an issue for static embeddings, with solutions like Orthogonal Procrustes or incremental training. For contextualized representations, most approaches use the same pre-trained model (fine-tuned or not), ensuring that the representations are comparable dimension-wise. However, the question of whether a model that is primarily pre-trained only on modern-days texts can accurately represent lexical meaning in old periods remains. Several fine-tuning approaches have been proposed to handle this issue: fine-tuning a pretrained CLM on the historical data of each period, fine-tuning it to incorporate the date of an occurrence in the input or even integrating the occurrence’s date directly in the CLM’s attention mecha-

nism (see Chapter 3, Section 3.2). Yet with a time-agnostic model, Cassotti et al. (2023) obtain state-of-the-art performances on the Lexical Semantic Change Detection benchmarks (Schlechtweg et al., 2020; Periti and Tahmasebi, 2024a) by fine-tuning a CLM on a relevant lexical semantic task (in that case, cross-lingual Word-in-Context). Their approach outperforms time-aware approaches on the Lexical Semantic Change Detection tasks, suggesting that temporal accuracy may be obtained by focusing the model’s attention on abstract lexical meaning, rather than time-dependent morpho-syntactic features.

Time-agnostic concepts. The second implication of this ‘all-periods-at-once’ approach to concept learning is that it assumes a set of stable concepts to learn. It is hypothesized that the concepts themselves do not change over time, only their associations to words. Under this assumption, concepts may or may not be instantiated in a given time period, new concepts may emerge (when they are associated to words for the first time) and old concepts may disappear (when they lose all of their associations to words), but their definition remains stable; in other words, the concepts are *time-agnostic*.

This hypothesis is already present in linguistic and psycholinguistic studies that rely on concepts, especially those investigating the mapping between words and meanings. For instance, the model of gained/lost associations between words and a fixed set of concepts is similar to the lexicon model proposed by Murphy (2002). In linguistic typology, Semantic Maps also rely on a fixed set of concepts, using them as nodes of a graph where the edges represent the amount of languages that colexify two concepts (i.e. that use a single polysemous word to refer to both concepts). See Georgakopoulos and Polis (2018) for a general survey on semantic maps, and Haspelmath (2023) for updated definitions and terminological clarifications. In order to account for semantic shifts, it seems reasonable to consider that concepts are fixed over time: while Old English used the word *hound* to refer to all the domesticated canines, and only later has *dog* taken this role, the concept itself has not changed but only its designation. Moreover, several studies using semantic maps with fixed concepts that have successfully been applied in the historical perspective to study patterns of regular colexifications between certain meanings observed in the evolution of languages (Georgakopoulos and Polis, 2021).

The downside of considering concepts as a fixed set over time is that this view cannot account for *concept drifts*, which are changes induced by societal, cultural or scientific changes that alter the way speakers perceive or understand a concept. For instance, the concept of HOMOSEXUALITY used to be strongly

negatively-connotated and perceived as a mental illness in the past. The understanding of the concept of ATOM has changed in the last century, alongside research in chemistry and physics. As such, concept drifts do not concern the conceptual layer as such, but rather the mapping between the conceptual and denotational levels (see Figure 2.2 of Chapter 2). In the context of the proposed approach to Diachronic Concept Induction, concept drifts will either be represented as different clusters if the perceptual or emotional differences are strong enough (as if there were two different concepts), or as a monolithic cluster if the differences are too small (as if there was no difference).

6.1.2 Formalism

In this remainder of this section, we illustrate the different types of change we aim at identifying in diachronic data, both in the semasiological and the onomasiological perspectives using Diachronic Concept Induction. We extend the formalism introduced in Section 4.1.2 to account for historical data.

Let $T = [T_1, \dots, T_P]$ be the set of P disjoint time periods covered in the study. Let W be the target lexicon, i.e. the set of lexemes included in the study, and O be the diachronic corpus, i.e. the set of all occurrences from all lexemes in the lexicon, from any of the P periods. For a particular lexeme w in the lexicon W , we denote $O_w^{(p)}$ the set of the occurrences of w from time period T_p . In the diachronic setting, O_w refers to the set of all occurrences of w , across all time periods. Thus $O_w = \bigcup_{p=1}^P O_w^{(p)}$ and, consequently, $\forall p \in [1, P], O_w^{(p)} \subset O_w \subset O$.

Through Concept Induction, we aimed at identifying semantic concepts as they emerge in data, and the sets of words associated to them. The goal was to identify, for each concept c_k in the set of concepts $\mathcal{C}$, a cluster of occurrences $\psi_k \subset O$ that instantiate that concept, and derive the *concept cluster's naming* N_k , which is the set of words referring to this concept among these occurrences. In this chapter corresponding to the diachronic setting, assuming that there is an underlying set $\mathcal{C}$ of time-agnostic concepts, we want to apply a model of Diachronic Concept Induction to learn the set of time-agnostic concept clusters Ψ . Each concept cluster $\psi_k \in \Psi$ encompasses occurrences from the historical corpus O that refer to the time-agnostic concept c_k , regardless of the target lexemes and of the time period from which occurrences originate. Concept clusters are learned in data in a fully unsupervised fashion: in the present study, we lack the means to identify the precise concept corresponding to each cluster. Mapping concept clusters to clearly defined concepts, or automatically mapping concept clusters to more interpretable spaces is left to future work. In the absence of explicit concept labels, our observations of lexical semantic changes are simply made on the latent

concept clusters. In this section, we describe how changes in lexeme-concept associations are expected to be reflected in the assignments of word occurrences to concept clusters.

In each time period T_p , a **naming** $N_k^{(p)}$ is defined for each concept cluster ψ_k . This concept cluster naming is defined as the set of words whose occurrences from time T_p are found in the concept cluster ψ_k , as in the following equation.

$$N_k^{(p)} = \left\{ w \in W \mid \exists o_{w,i}^{(p)} \in O_w^{(p)}, o_{w,i}^{(p)} \in \psi_k^{(p)} \right\} \quad (6.1)$$

Conversely, at each period T_p each lexeme w is associated to a **sense inventory** $M_w^{(p)}$, i.e. the set of concept clusters in which occurrences of w at T_p are found.

$$M_w^{(p)} = \left\{ \psi_k \in \Psi \mid \exists o_{w,i}^{(p)} \in O_w^{(p)}, o_{w,i}^{(p)} \in \psi_k \right\} \quad (6.2)$$

Using cluster namings and sense inventories, we now describe semasiological and onomasiological changes.

6.1.3 Types of Semasiological Changes

Semasiological changes are evolutions in a word's set of senses. Therefore, observing such changes between T_1 and T_P for a lexeme w relies on comparing $M_w^{(1)}$ and $M_w^{(P)}$.

Assume that a concept-inducing systems finds that w has two meanings at time T_1 , i.e. we find two concept clusters ψ_1 and ψ_2 containing occurrences of w from time period T_1 , yielding $M_w^{(1)} = \{\psi_1, \psi_2\}$. These concept clusters are meant to correspond to latent concepts c_1 and c_2 . Let us now examine the possible changes this word's sense inventory may undergo, by analyzing the difference between $M_w^{(1)}$ and $M_w^{(P)}$.

Sense gain. If w undergoes a *sense gain* between time T_1 and time T_P , this means that at time T_P it may be used to refer to a new concept c_3 . For instance, the word *broadcast* gained a new sense in the 20th century, to refer to the concept of TRANSMIT BY MEANS OF RADIO OR TELEVISION. In $M_w^{(P)}$ we observe a new concept cluster that was not part of $M_w^{(1)}$, e.g. $M_w^{(P)} = \{c_1, c_2, c_3\}$.

Sense loss. If w undergoes a *sense loss* between time T_1 and time T_P , this means that one of the concepts w used to refer to at time T_1 no longer appears in the set of concepts w refers to at T_P . For instance, the word *abstract* has lost its

meaning of EXTRACTED since the 16th century. One of the concept clusters in $M_w^{(1)}$ no longer appears in $M_w^{(P)}$, e.g. $M_w^{(P)} = \{\psi_1\}$.

Combined changes. A word may undergo *both* changes simultaneously. For instance, *gay* gained the meaning of PERTAINING TO HOMOSEXUALITY during the 20th century, while its older sense of COLORFUL, BRIGHT is now considered *dated*. If a word undergoes both a sense gain and a sense loss, some concept clusters from $M_w^{(1)}$ disappear from $M_w^{(P)}$, while new concept clusters appear in it ($M_w^{(P)} = \{c_1, c_3\}$).

There can a continuity between the sense that was lost and the one that was gained (we may say refer to this phenomenon as a *shift* in meaning). Such continuity may arise from some of the change types discussed in Section 2.3.3, such as *generalization*, *specialization*, etc. (e.g. *hound* used to refer to DOG as a whole class and now commonly refers to HUNTING DOG only). However, the proposed method currently lacks means to characterize this continuity.

6.1.4 Types of Onomasiological Changes

Onomasiological changes are evolutions in a concept's naming, i.e. the set of words that refer to this concept. Thus, observing such changes between T_1 and T_P for a concept c_k requires comparing between the namings of the corresponding clusters $\psi_k^{(1)}$ and $\psi_k^{(P)}$, i.e. comparing $N_k^{(1)}$ and $N_k^{(P)}$.

Assume that both words w and v can be used to refer to c_k at time T_1 , such that occurrences of both w and v appear in ψ_k . Thus $N_k^{(1)} = \{w, v\}$.

Naming expansion. If ψ_k undergoes *naming expansion* between time T_1 and time T_P , then a new word appears in its naming at time T_P . $N_w^{(P)}$ should contain a word u absent from $N_k^{(1)}$, i.e. $N_w^{(P)} = \{w, v, u\}$. For instance, around the late 18th / early 19th century, the concept of CYLINDRICAL CONTAINER FOR LIQUIDS previously containing only *barrel* has been expanded with *drum*.

Naming shrinkage. If ψ_k undergoes *naming shrinkage* between time T_1 and time T_P , then one of the words that appeared in its naming at time T_1 is no longer part of it at time T_P . One word from $N_k^{(1)}$ should be missing from $N_w^{(P)}$, e.g. $N_w^{(P)} = \{w, v, u\}$. In French for instance, the word *courtil* used to mean GARDEN, alongside *jardin*. But the former word has been completely replaced by the latter in modern French, and as such *courtil* has been removed from the naming of GARDEN.

Combined changes. A concept may undergo *both* changes simultaneously. In the naming of concept cluster ψ_k , we should have $N_k^{(1)} = \{w, v\}$ but $N_k^{(1)} = \{u\}$ for instance. As an example, the concept of RAPIDLY was referred to with *quickly* and *swiftly* in Old English. Both words no longer instantiate this meaning (as they now mean of IMMEDIATELY), while *fast* has become the preferred word for this concept.

Studying the set of concepts and the change in their designations, one may encounter the appearance of a new concept or the disappearance of an obsolete concept.

Emerging and disappearing concepts. Across centuries, multiple factors drive the emergence of new concepts: technological progress, scientific research, cultural changes, historical events, etc. For instance, with the invention of new communication technologies and of computers, many new concepts have emerged, such as COMPUTER PERIPHERAL TO POINT AND CLICK (a *mouse*), TRANSMIT A PROGRAM (*broadcast*), CONTINUOUSLY DELIVERED DIGITAL DATA (*stream*). Eighteenth and nineteenth-century scholars have elaborated new concepts to describe economic and political theories, such as CAPITALISM, SOCIALISM, ANARCHISM, COMMUNISM, etc. These same reasons may render concepts obsolete in modern societies. While not disappearing completely from language, these concepts may become so rarely referred to that they may no longer appear from large text corpora in more recent periods. With Christianity's decreased importance in Western Europe since the end of the Middle Ages, a lot of concepts related to religious practices, items and symbols have become less significant daily life and are more rarely discussed by language users. French for instance had a word (*patenôtrier*) for the concept of CRAFTSMAN OF CHRISTIAN PRAYER BEADS, an old job title that no longer exists in modern days.

In the Diachronic Concept Induction framework, **emerging concepts** manifest as concept clusters that contain no occurrence from early time periods but gather occurrences of late periods only. A concept-inducing system may fail to capture emerging concepts if the new concept is too similar to an existing one (in which case they may be grouped together in the same cluster). Alternatively, if there are not enough occurrences referring to this new concept, it may not be identified at all. Conversely, a concept may be mistakenly identified as 'new' if it fails to capture old occurrences referring to this concept within the corresponding concept cluster.

Disappearing concepts manifest as concept clusters containing occurrences from early time periods only. A concept-inducing system may fail to identify disap-

pearing concepts if the concept in question is too similar to one that exists in the later time periods (in which case they may be grouped together in the same cluster). It may falsely identify a concept as ‘disappearing’ if it fails to capture later occurrences that refer to the concept, and assigns them to a different cluster instead.

6.1.5 Interactions Between Semasiological and Onomasiological Changes

As introduced in Chapter 2 when discussing the Word-Meaning mapping, semasiology and onomasiology are two complementary views on the same object, the mapping between words and concepts. As such, semasiological changes (i.e. shifts in sense inventory), are reflected in some onomasiological changes (i.e. shifts in concept naming), and vice-versa.

When a word w gains a new sense referring to concept c_k , it means that either c_k is a new concept, or the designation of c_k now includes w . Conversely, if the naming of c_k expands with a new word w , it means that word w is either a new term, or it gained the sense referring to c_k . In concept clusters, we will observe a correspondance between the semasiological change from $M_w^{(1)} = \{\psi_1, \psi_2\}$ to $M_w^{(P)} = \{\psi_1, \psi_2, \psi_k\}$ and the onomasiological change from $N_k^{(1)} = \{v\}$ to $N_k^{(P)} = \{w, v\}$. If w is a new term, $M_w^{(1)}$ will simply be empty, and if c_k is a new concept, then $N_k^{(1)}$ will be empty.

Similarly, when a word w loses a sense referring to concept c_k , it means that either c_k has disappeared, or the designation of c_k no longer includes w . Conversely, if the naming of c_k shrinks and loses word w , it means that word w is either is an obsolete term, or it lost the sense referring to c_k . In concept clusters, we will observe a correspondance between the semasiological change from $M_w^{(1)} = \{\psi_1, \psi_k\}$ to $M_w^{(P)} = \{\psi_1\}$ and the onomasiological change from $N_k^{(1)} = \{w, v\}$ to $N_k^{(P)} = \{v\}$. If the word w becomes obsolete, $M_w^{(P)}$ will be empty, and if the c_k is disappears, then $N_k^{(P)}$ will be empty.

This correspondance between semasiological and onomasiological changes is fundamental in this chapter. Indeed, evaluating a method’s ability to detect and characterize lexical semantic changes and particularly onomasiological changes requires a reference resource that indicates whether the detected changes correspond to actual changes in the history of the language, or if they are mistakes made by the system. However, expert resources for onomasiological changes are hardly available: to our knowledge the only resource is the *Historical Thesaurus of English* (Kay et al., 2020), but its API is currently closed. Thus, to

evaluate a model’s detection capacity, we will rely on semasiological resources, dictionaries that include a word’s history, that are more easily available and for more languages. To evaluate the quality of detected onomasiological changes, we will look for attested semasiological changes in the reference resource, and see whether we can find an attested semasiological counterpart for a given onomasiological change identified by the proposed model.

6.2 Methodology for Diachronic Concept Induction

6.2.1 Bi-level clustering in diachronic corpora with time-agnostic concepts.

Reformulation. Let us redefine the bi-level approach introduced in Section 4.2 in the diachronic setting, given the considerations described above. For the first, local sense clustering step, we consider for each lexeme $w \in W$ the set of its occurrences across all time periods O_w . Clustering these occurrences provides S_w , i.e. the set of all **proto-sense clusters** in all periods for a word w . Each proto-sense cluster $s_{w,j} \in S_w$ (the j -th proto-sense cluster of word w) is time-agnostic: it may contain occurrence of w from any time periods. The set of all proto-sense clusters from all lexemes across all periods is denoted $S = \bigcup_{w \in W} S_w$. S draws a partition of the whole historical corpus that remains at the lexeme-level. In the second step, proto-sense clusters of S are once again represented with the *average vector* of occurrences in the cluster.

Then the representations of all proto-sense clusters are fed to a global, cross-lexicon clustering run. All the occurrences within proto-sense clusters that are grouped together form a single **concept cluster**. A concept cluster denoted ψ_k may contain occurrences from different time periods, from different lexemes, and even from different proto-sense clusters related to the same lexeme.¹ The set of all time-agnostic concept clusters is denoted Ψ .

Certain words may be present in certain periods and absent from other, because they were brought to life in the language more recently or conversely because they fell in disuse, or simply because the historical corpus lacks occurrences for these words in a given time period. We define $W^{(p)} = \{w \in W \mid O_w^{(p)} \neq \emptyset\}$ the subset of words in the target lexicon that have occurrences in time periods T_p . New words are those that are present in $W^{(P)}$ despite being absent from $W^{(1)}$. Obsolete words are no longer in $W^{(P)}$ despite being present in $W^{(1)}$.

¹ Indeed, proto-sense clusters are not considered finalized senses, they may correspond to different contextual variations of the same sense, as in Chapter 4.

Similarly, certain concepts may be present in certain time periods, but not instantiated in other periods. We define $\Psi^{(p)}$ the set of concept clusters corresponding to concepts that were instantiated at least once during time period T_p as follows:

$$\Psi^{(p)} = \left\{ \psi_k \in \Psi \mid \exists w \in W, \exists o_{w,i}^{(p)} \in O_w^{(p)}, o_{w,i}^{(p)} \in \psi_k \right\}.$$

Emerging concepts are those that appear in $\Psi^{(P)}$ but are absent from $\Psi^{(1)}$. Disappearing concepts are those that were present in $\Psi^{(1)}$ but no longer in $\Psi^{(P)}$.

Representations and Algorithms. To represent occurrences, we use the CALE model based on XL-LEXEME introduced in Chapter 5, to benefit from both XL-LEXEME’s multilingual fine-tuning and CALE’s broader lexical semantic objective. On the Lexical Semantic Change detection benchmark, this model has obtained the highest performances compared to other tested models or was on-par with the best model, in almost all languages. The clustering algorithm used in both the local and the global steps is Hierarchical Agglomerative Clustering, as this combination reached the best performances in the evaluation performed in Chapter 4.

For the global step, to extract clusters from the constructed hierarchy, we set an arbitrary distance threshold τ and we keep clusters at the corresponding level in the hierarchy: cluster merges above this threshold are ignored, whereas merge operations below this threshold are considered. For the local step, however, we use HDBSCAN’s stability-based cluster selection method. This method, presented in details in Appendix A.3.1, computes a measure of stability for each cluster in the hierarchy and compares the stability values of the different clusters in the hierarchy to keep the most important ones. This method has three advantages over the simple threshold-based selection.

First, it allows to select clusters in different levels of the hierarchy: clusters originating from very early merge operations can be selected along clusters from later merge operations, if both are deemed stable enough compared to their children and parent clusters. Enabling the selection of clusters at different levels reduces the risks of over-fragmentation (being too restrictive with merge operations which results in a lot of very small clusters, if the threshold is too low), and over-regrouping (being too permissive with merge operations and thus ending up with a few clusters containing all occurrences, if the threshold is too high). Let us recall that the time complexity of constructing a hierarchy with Agglomerative Clustering scales quadratically with the number of instances: thus, the computation time required for clustering at the global step would become too important if we over-fragment proto-sense clusters. Using HDBSCAN’s cluster

selection allows to select a reasonable amount of proto-sense clusters for each word while limiting the risk of over-regrouping.

Second, only clusters with more than m occurrences within them are taken into account, m being the only hyperparameter of HDBSCAN’s selection method. This hyperparameter is simple, interpretable and therefore more convenient to tune than an arbitrary distance threshold.

Finally, this algorithm is able to identify noise points: occurrences that do not belong to any selected cluster. Large diachronic corpora can contain a significant portion of noisy data, induced by errors in the various automatic systems involved (e.g. Optical Character Recognition, automatic sentence parsing or automatic lemmatization.) or due to other reasons such as document formatting (e.g. chapter titles or publisher/copyright notices). Identifying these noisy points in the first step of our bilevel clustering ensures they do not affect the resulting concept clusters.

6.2.2 Tuning Hyperparameters Without Concept-Annotations

In the synchronic setting, a few concept-annotated corpora are available, like SemCor that we used in Chapter 4 and 5, which allowed us to exploit these annotations to find the best set of hyperparameters for each clustering model. In contrast, in the diachronic setting, no comparable concept-annotated corpora spanning over multiple time periods exists. The proposed bi-level clustering approach relies on the following hyperparameters: the linkage definition for the local step, m the minimum number of occurrences in proto-sense clusters, the linkage definition used in the global step and τ the distance threshold used to obtain concept clusters. Each of these parameters can take a range of different values could be used, and we need a method to select a set of hyperparameters that fits our problem setting without relying on concept annotations.

We use the following procedure for selecting our hyperparameters: given a set of possible hyperparameter combinations Θ , we try the different combinations in the set and select the combination $\hat{\theta}$ that results in the largest number of concept clusters containing between 2 and n different lexemes. The upper bound n is arbitrarily set and its values must correspond to an expected maximum number of synonyms for the same concept in a given language. It seems reasonable to assume that synonymy is not a large-scale phenomenon and that there should not be a high number of synonyms for the same concept. The value should be set such that concept clusters containing occurrences of more than n different lexemes are likely not groups of actual synonyms, as The method is summarized

in Algorithm 1.

Algorithm 1: Our hyperparameter selection algorithm.

Data: Θ the set of possible parameter combinations; O the historical corpora.
Result: $\hat{\theta}$ the selected set of hyperparameters.
 $D \leftarrow -1$; /* Initialize counter. */
for $\theta \in \Theta$ **do**
 $\Psi^{(*)} \leftarrow \text{Bilevel}(\theta, O)$;
 $d_\theta \leftarrow |\{\psi_k \in \Psi^{(*)} \mid 2 \leq N_k \leq n\}|$; /* Count clusters with 2 to n lexemes. */
 if $d \geq D$ **then**
 $D \leftarrow d$;
 $\hat{\theta} \leftarrow \theta$; /* Becomes the new selected parameters. */
 end
end

This method encourages higher rates of discovered synonymies, while it penalizes clusterings that emphasize only on single-word concepts. This method leverages the hierarchical nature of the data and of the algorithms used, and most directly impacts the choice of threshold at the global, cross-lexicon level, where synonym groups are created. The lower bound of 2 considers only clusters of strictly more than one lexeme and promotes merge operations across different lexemes with a sufficiently high threshold value, in order to retrieve synonymy. Indeed the onomasiological aspect of change is crucial for this thesis, and the retrieval of synonymy is fundamental to such approach. The upper bound n ensures that the threshold is not too high, which would result in the creation of too large clusters containing occurrences of words that do not refer to the same concept. With this policy we aim at preserving a balance between being conservative with merge operations while uncovering a maximum number of synonym groups.

Importantly, this method does not impact the learning process nor does it prevent the creation of concept clusters containing only one lexeme; it is an approach to select *a posteriori* one clustering obtained with hyperparameters $\hat{\theta}$ among a set of candidates clusterings with different hyperparameter values. As we will show in Section 6.4, the final selected clustering still contains a large number of concept clusters containing occurrences of only one word. The extent to which this method impacts the retrieval of polysemy is left to future work.

Other approaches could have been adopted to select a set of hyperparameters (and the corresponding clustering) without using concept annotations. To cluster

occurrences in historical data into word senses, many studies of the existing NLP literature rely on the Affinity Propagation for which no hyperparameter must be tuned (Martinc et al., 2020c; Montariol et al., 2021; Periti et al., 2022). Other approaches to historical word senses in the NLP community use the Silhouette score as an indicator of the clustering quality, and choose the set of hyperparameters that maximize the *Silhouette coefficient* of the clustering (Kanjirangat et al., 2020). Martinc et al. (2020b) however showed that Silhouette score did not necessarily correlate with performances on the Lexical Semantic Change Detection task, casting doubt on the relevance of such hyperparameters tuning in our setting. Future work may explore and compare different hyperparameter tuning methods and identify the most well-suited approach in the context of diachronic corpora.

6.2.3 Filtering obtained clusters

In practice (cf. Section 6.4.3), despite our initial filtering on the target words and the ability of HDBSCAN’s stability-based cluster selection used at the local step to identify noise points, we still observe that many of the discovered concept clusters exhibit unexpected properties (e.g. very few occurrences but containing many different lexemes). This irregularity may be explained by the noisy nature of the data², but also by weaknesses in the representations, in the choice of clustering algorithms and/or in the hyperparameter selection method. In this study, we focus more on evaluating the strength of discovered changes and less on uncovering a maximum number of changes, mostly because we have no means to evaluate the method’s Recall. Thus,; we have decided to filter out a number of the occurrences and clusters, based on the following two criteria.

First, for a lexeme to truly belong a cluster at a given point in time, we retain its occurrences for a half-century only if there are more than 5 of them. This decreases the risk of detecting the presence of a lexeme in the concept namings while it could be only due to some unrelated statistical artifacts. Then, we keep clusters and the occurrences they contain only if a cluster contains more than 50 occurrences in total. Preliminary inspection of the data showed that there was many very small clusters corresponding to statistical artifacts (like occurrences drawn from an extremely unusual/distinctive context) rather than actual meanings.

² The PRESTO corpus is rather large compared to many other diachronic corpora, especially in its *extended* version. As a consequence, text formatting, character recognition and automatic lemmatization to modern forms have not been manually checked for the whole corpus.

6.2.4 Summary of the Adaptations

We have identified that the diachronic setting imposes two fundamental changes to the approach we have elaborated in the previous chapters.

To achieve **Diachronic Concept Induction**, there is now several different time periods to consider instead of one, raising the question on how to compare occurrences to learn concepts. Learning concepts on all the data at once implies that there is a stable set of concepts over time.

In practice, we run the bilevel clustering model in all time periods at once, first learning time-agnostic proto-senses and their clustering their centroids into global clusters to obtain concept clusters. We use the CALE-XL-LEXEME model to represent occurrences in the different time periods to ensure that occurrence meanings are well represented in all periods and comparable over time.

Second, there are no diachronic concept annotations on the data, impacting the way we select our model's parameters. To select the most suitable set of hyperparameters, we take the set of values that maximizes the number of clusters containing occurrences of more than 2 distinct lexemes but no more than n lexemes, n being chosen arbitrarily depending on the language. This criterion does not make assumptions on the cluster shapes and is based on the linguistic intuition that there is only a small set of valid substitutes (synonyms or near-synonyms) for a word in a given context.

Third, we elaborate a filtering strategy to improve the quality of our analyses by removing small clusters and occurrences of words that are not numerous enough in a concept cluster.

6.3 Evaluating Diachronic Concept Induction.

Diachronic Concept Induction (DCI) aims to identify semasiological and onomasiological changes as shown in Sections 6.1.3 and 6.1.4. Its evaluation can thus be decomposed into a notion of *Precision* (do retrieved changes correspond to actual changes ?) and a notion of *Recall* (did the method captured all the changes that were present in the data ?). In Chapter 4 we have used concept-annotation on the data to derive a reference word-level clustering to which we compared the learned word clustering. In the absence of semantic annotations, we do not know the concepts instantiated in the historical data, nor the synonym groups that are to be uncovered; thus we cannot assess the Recall of a proposed method. But we may evaluate Precision by comparing identified changes to known lexical semantic changes.

Evaluating semasiological changes. Despite projects of historical WordNets (de Melo, 2014; Khan et al., 2023), we are not aware of the existence of an available resource from which we could systematically extract concept lists over time, nor groups of synonyms over time. There are, however, semasiological dictionaries that provide the history of a word’s meanings: the order in which they developed, the first time each meaning was attested, etc. Such a dictionary could naturally help us evaluate semasiological changes. Assuming that a Diachronic Concept Induction system uncovers a sense gain between periods T_1 and T_P , we can verify whether the corresponding sense in the historical dictionary is first attested between these two periods. For sense loss, the sense should be first attested before T_1 and should now be labeled as old, archaic or obsolete in the dictionary (assuming that T_P is close enough to the dictionary’s edition date).

Evaluating onomasiological changes. In Section 6.1.5, we have shown that onomasiological changes can be observed through their semasiological counterparts. We propose to use the historical semasiological dictionary to verify whether onomasiological changes correspond to attested changes.

- **Naming expansions:** if w appears in the naming of a concept cluster ψ_k between T_1 and T_P , then the dictionary should indicate a sense of w first attested between these two periods.
- **Naming shrinkages:** if w leaves the naming of a concept cluster ψ_k between T_1 and T_P , then the corresponding sense of w should appear as old, archaic or obsolete in the dictionary.
- **Emerging concepts:** if a concept cluster emerges from T_1 to T_P and contains only occurrences posterior to T_1 , all words in its naming should have a corresponding sense that is first attested after T_1 in the historical dictionary.
- **Disappearing concepts:** if a cluster does not contain any occurrence from T_P , all words in its naming at T_1 should have a corresponding sense labeled as old, archaic or obsolete in the dictionary.

Interpreting concept clusters. Of course, such evaluation presupposes to map concept clusters to sense definitions in the reference dictionary. To help in doing such correspondance and interpret the semantic content of concept clusters, we propose to use the most frequent context words of occurrences in the cluster, as in Montariol et al. (2021). For each cluster, we retrieve the top-15 most frequent words that appear on the left and the right side of the target word for every

occurrences in the cluster, excluding some stopwords (grammatical morphemes, verb auxiliaries, ...). This contextual information, used alongside the naming of a concept cluster, should help mapping concept clusters to the corresponding semasiological senses in the dictionary.

The proposed evaluation process is nonautomated, it requires human intervention to (1) interpret correspondances between concept clusters and semasiological senses in the dictionary and (2) check that the date and nature of the observed changes correspond to semasiological changes attested in the expert resource. Automation of this evaluation is an important direction for future research, especially through means of Word Sense Disambiguation techniques to establish cluster-sense mappings.

Having detailed the adjustments required for achieve Diachronic Concept Induction, how it should reveal onomasiological and semasiological changes in diachronic data as well as means to evaluate such method, we move to an experimental evaluation of the proposed approach.

6.4 Experiments

In this section, we apply Diachronic Concept Induction to a large historical corpus, and assess whether changes in sense inventories and concept cluster namings described by the method actually reflect attested changes. The goal of this experimental section is to show that this methodology can indeed characterize onomasiological changes, while still allowing a more traditional semasiological study of semantic change.

6.4.1 The PRESTO Corpus

We choose the PRESTO corpus (Blumenthal et al., 2017), an historical collection of texts spanning from 1500 to 1949. The corpus is composed of 315 texts, for a total of 28 millions of tokens³. It comprises equal proportions of four discursive genres: narrative literature (novels, fairytales, short novels, etc.), poetry, dramatic literature and ‘generic’ genre (comprising essays and treatises). The temporal distribution of texts is displayed in Figure 6.1.

This corpus was chosen for two reasons. first, since the evaluation is purely qualitative, we wanted the corpus to be in French, our native language, to ensure the reliability of the evaluation and avoid interpretative ambiguities; and

³ A subset of the corpus is available online at <http://presto.ens-lyon.fr/>. We thank the PRESTO project coordinators for giving us access to the full version.

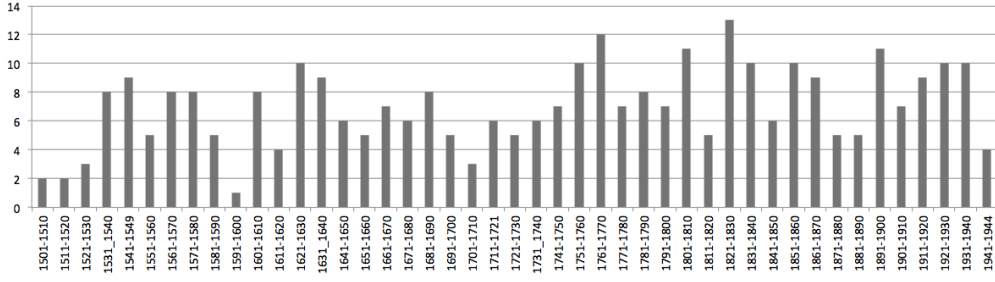


Figure 6.1: Distribution of PRESTO's texts over time.

second because it is one of the largest historical corpora of French that has part-of-speech and lemmatization annotations, both of which are crucial components to our method. Part-of-speech tags are useful to distinguish between the different grammatical categories and allow to cluster nouns, verbs and adjectives separately (as synonymy only exists between words of the same category). Lemmatization is used to track word occurrences over time and to cope with orthographic variations, very common in French between the 16th and the 18th century.

Target words selection. To select target words and extract their occurrences, we regroup the corpus' texts into half-centuries. Given that the corpus spans across four centuries, we estimate that a half century granularity provides an appropriate length to judge distributional continuity across the corpus: decades would be very small divisions of the total span and likely lead to a lot of irregularities; centuries would represent overly large divisions and imbalances would arise, especially given that only the first half of the 20th century is present in the corpus. In this study, we focus on adjectives, nouns and verbs. Target words are chosen according to a minimum frequency per half century, denoted f . Crucial to the target word selection is the idea of distributional continuity over time. We aim to include newly emerging words, words that fall out of use and words that remain consistently frequent over time, while avoiding words with too few occurrences or 'blinking' words that abruptly and frequently appear and disappear (due to the corpus' sampling rather than a linguistic reason). We therefore retain only words that fit one of the following three distributional patterns:

- **Stable words:** words that have a minimum of f occurrence per half-century across the corpus.
- **Appearing words:** words that have less than f occurrence in every half-century until a given point, and at least f occurrences in every subsequent half-century.

- **Disappearing words:** words that have more than f occurrence for every half-century and fewer than f occurrences per half-century thereafter.

In our experiments, trying different values for f , we find that a value of 10 effectively filters out most irregular distributional patterns without removing an excessive portion of the data. We further exclude words that have less than 50 occurrences in total, words that appear in less than two half-centuries and words composed of less than 3 letters. Ensuring a sufficient number of occurrences for each target word – per half-centuries or in total – is crucial for the bi-level clustering approach, as since the local step clusters word occurrences per lexeme and a proper clustering requires a sufficient number of instances.

Occurrence extraction. Occurrence for each of the target words are extracted as follows. We sample a maximum m occurrences per half-century for every target words, or we take all available occurrences if it has less than m occurrences in a given half-century. The context associated to a target word in an occurrence is limited to a context window of half-size h words centered around the target word. In our experiments, we set h to 20 words, and the sample maximum criteria m to 200. While the minimum frequency per half-century and in total guarantees that we select only target words for which we have enough occurrences for clustering, this maximum sample limit ensures that the clustering computational costs remain reasonable while preserving sense distributions. To mitigate the impact of orthographic variations (French spelling stabilized only after the 18th century), we replace nouns, verbs, adverbs and adjectives in the occurrence with their modern lemmatized form. This may degrade the quality of the obtained representations (as the contextualized model was not trained to work on lemmatized sentences) but it ensures that the representations are comparable, especially between the oldest and the most recent time periods, and avoid tokenization errors on old orthographic forms (as the contextualized model’s tokenizer has seen data only for modern French).

Summary statistics for the obtained corpus extraction are provided in Table 6.1.

Part-of-speech	# of targets	# of occurrences	average # of occs. per target and half-century
Nouns	4,195	2.1M	64
Verbs	1,382	903k	78
Adjectives	1,408	603k	55

Table 6.1: Summary statistics of our extraction from the PRESTO corpus.

6.4.2 Experimental Settings

Bilevel clustering system. To represent occurrences, we use the CALE (XL-LEXEME) model presented in Chapter 5. The local, intra-lexeme clustering step is performed using Agglomerative Clustering with single-linkage (i.e. the minimum distance between clusters) and clusters are selected using HDBSCAN’s stability-based selection. After the occurrences in pseudo-sense clusters are averaged, they are fed to the global, cross-lexicon clustering step which employs Agglomerative Clustering with average linkage and a fixed distance threshold.

Hyperparameter values. For our hyperparameter selection method, we set the upper limit n to 5 for French. To limit the impact of noisy or insignificant clusters, only clusters of more than 20 occurrences were considered when counting clusters grouping between 2 and 5 distinct lexemes. Preliminary experiments have revealed that very small clusters (containing very few occurrences) may appear and regroup several lexemes due to the noisy nature of some of the data points. For the local clustering step, we tested minimum cluster sizes of 3, 5 and 10. The value of 3 was systematically preferred by our criterion, whereas a value of 10 lead to an excessive number of occurrences labeled as noise. For the global, cross-lexicon clustering step, the threshold was set to a percentile of the distribution of distances between pseudo-sense cluster representatives, exploring percentile values ranging from 0.001% to 15% of the distribution.

Semasiological reference. To assist our evaluation of the obtained clusters and observed changes, the semasiological historical dictionary that we use is the *Trésor de la Langue Française informatisé* (TLFi)⁴, the online version of a dictionary of French of the 19th- and 20th-century, which corresponds to the later centuries of our corpus. The dictionary provides, for each entry, a historical description of the meaning(s), including the earliest attestation of each meaning found in the FRANTEXT database. Since PRESTO texts are also extracted from FRANTEXT, this ensures fair comparability between changes observed in PRESTO and attestation dates in the TLFi.

6.4.3 Cluster Filtering Results

The impact of the filtering on the obtained cluster is displayed in Table 6.2. As we can see, this selection drastically reduces the number of considered clusters, e.g. from 18.2k to 2,950 for nouns. It demonstrates that the method, whether

⁴The TLFi is not an historical dictionary as such, but it provides an historical section for each entry.

Part-of-speech	# of targets	# of occurrences	# of concept-clusters
BEFORE FILTERING			
Nouns	4,195	2.1M	18.2k
Verbs	1,382	903k	6.6k
Adjectives	1,408	603k	1.9k
AFTER FILTERING			
Nouns	3,819	1.4M	2,950
Verbs	1,306	605k	1,274
Adjectives	1,334	424k	794
<i>Total</i>	6,459	2.4M	5,018

Table 6.2: Summary statistics of the filtering made on the PRESTO corpus from clustering results.

due to noise in the data, the representations or to the clustering, has produced a lot of small, irregular clusters that would have negatively impacted the quality of the analyses. As a result, we obtain a total of 5018 concept clusters, for 6,459 words.

6.4.4 Synchronic Onomasiological Results

Let us now have a first look at our resulting 5,018 clusters. To guide our analysis, we will divide the corpus span in three time periods, corresponding to different stages of the development of French (Marchello-Nizia et al., 2020):

- 1500-1649, Middle & Preclassical French (PFr);
- 1650-1799, Classical French (CFr);
- 1800-1949, Modern French (MFr).

Observations. The distribution of the number of words in cluster namings with respect to temporal periods is shown in Figure 6.2. Clusters with ‘0’ lexeme correspond to concept clusters that do not contain any occurrence in a given time period but that regroup occurrences from other periods.

As we can see, there is a majority of concept clusters containing occurrences of only a single lexeme, and a majority of the lexemes have only one sense. Among clusters appearing in Modern French, the average number of synonyms contained in clusters is 1.78, with 3,102 (62% of the total number of concepts) of them containing a single lexeme. This shows that synonymy is a relatively rare phenomenon, and that for most concepts a single lexeme is used in the considered corpus.

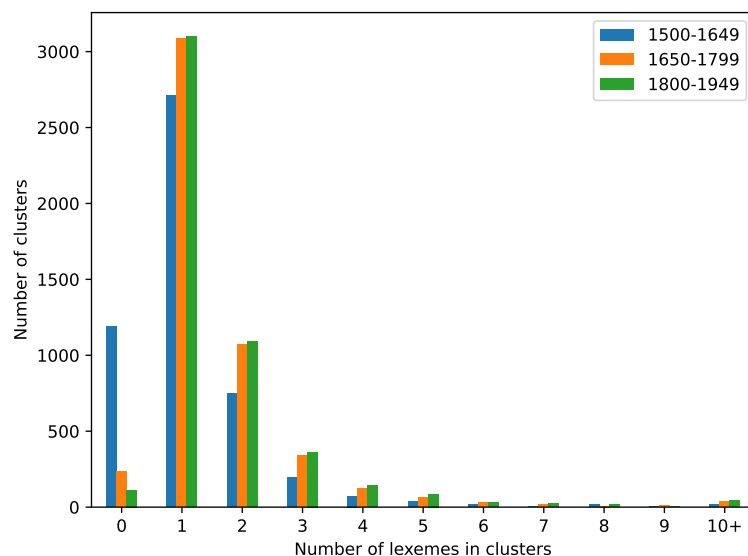


Figure 6.2: Distribution of obtained cluster naming sizes per period.

Evaluation. To assess the quality of obtained clusters, we randomly sample 100 concept clusters among nouns, verbs and adjectives combined. This sample contains 46 noun concepts, 30 verb concepts and 24 adjective concepts. We evaluate these clusters based on the lexemes they contain in Modern French. For a given cluster, we assess whether the grouped words are all synonyms, if they are near-synonyms, if they are more weakly related (antonyms, hypernyms, cohyponyms, or members of the same lexical field) or if they are unrelated. The results are presented in Table 6.3. We observe a majority (65%) of the sampled clusters indeed correspond to (near-)synonyms, showing that the approach, though not perfect, is reliable. Nonetheless, we argue that the appearance, stability or disappearance of weaker relations like antonymy or joint membership of a lexical field may still be indicating of semantic change.

Judgement	# of concepts	Category	# of concepts
Synonyms	37	(Near-)synonyms	65
Near-synonyms	28		
Antonyms	7	Weakly related	32
Co-hyponyms	4		
Lexical field	21		
Unrelated	3	Unrelated	3

Table 6.3: Qualitative evaluation of concept-clusters in the 1800-1949 period.

6.4.5 Onomasiological Changes

Now that we have assessed the synchronic quality of the clustering, let us now turn to the primary object of this study: onomasiological changes. For the sake of simplicity, in the remainder of this chapter, we will focus on the changes between Middle & Preclassical French (1500-1649, PFr) and Modern French (1800-1949, MFr). We leave the investigation of intermediate steps of changes in Classical French for future investigation. It will be shown that observed changes appear to reflect attested changes in most cases.

We first recall the different types of onomasiological changes to observe previously described in Section 6.1.4 and how they should appear in concept cluster's namings, we then discuss the observed distribution of these types of change on our data. Finally we evaluate our approach's ability to identify the different type of onomasiological changes, following the evaluation method outlined in Section 6.3.

Types of onomasiological changes for a concept cluster ψ_k :

- **Unchanged naming:** the concept's naming contains the same lexemes in Middle/Preclassical French and in Modern French, i.e. $N_k^{(\text{PFr})} = N_k^{(\text{MFr})}$.
- **Emerging or new concept:** the concept was not instantiated in Middle & Preclassical French, but is observed in Modern French ; thus $N_k^{(\text{PFr})} = \emptyset$ and $|N_k^{(\text{MFr})}| > 1$.
- **Disappearing concept:** the concept was instantiated in Middle & Preclassical French, but it cannot be observed in Modern French, thus $|N_k^{(\text{PFr})}| > 1$ and $N_k^{(\text{MFr})} = \emptyset$.
- **Expanded naming:** the naming in Modern French contains new lexemes compared to Middle & Preclassical French, i.e. $|N_k^{(\text{PFr})}| < |N_k^{(\text{MFr})}|$.
- **Shrunked naming:** some lexemes that were in the naming in Middle & Preclassical French are missing in the concept's naming in Modern French, i.e. $|N_k^{(\text{PFr})}| > |N_k^{(\text{MFr})}|$.
- **Both:** in Modern French, the concept's naming contains new lexemes and misses some lexemes compared to Middle & Preclassical French, i.e. $N_k^{(\text{PFr})} \not\subseteq N_k^{(\text{MFr})}$ and $N_k^{(\text{MFr})} \not\subseteq N_k^{(\text{PFr})}$.

Quantifications of onomasiological changes. In Table 6.4 we display the evolution paths observed by the 5,018 found concepts. We observe that a vast 49% of the concepts are unchanged between the Middle & Preclassical French and Modern French. Indeed, it is expected that the mapping between words and

Naming Evol.	# of clusters	%
Unchanged	2459	49.0
New concept	1189	23.7
Disappear.	97	1.9
Expanded nam.	782	15.6
Shrunked nam.	269	5.4
Both	222	4.4

Table 6.4: Evolution of the concept-clusters between Middle and Modern French.

meanings remain mostly stable across this period of time, as modern day speakers of French are able to understand the majority of a 16th century text without too much ambiguity.

We observe 1,189 emerging concepts (23.7% of the total). This impressive number of new concepts may be explained by the progresses of science, philosophy, medical sciences and technology after the 16th century, as well as the societal and social changes that occurred in France due to the abolition of monarchy and the many changes of political regimes in the 19th century, the industrial revolution, etc. All of these changes bring many new concepts to the language. Conversely, these changes, combined with the overall loss of importance of the Christianity in the daily life of people, can render some concepts obsolete. Our results indicate that only 97 concepts disappeared (1.9%) Examples from our data will be provided later, when evaluating the observed new concepts.

Another major tendency is that there are more expansions of concept namings than shrinkages (782 expansions for 269 shrinkages). Historical dictionaries indeed reveal that the French vocabulary underwent considerable expansion between the 17th and the 19th centuries, introducing loanwords and neo-classical formations from Greek and Latin. For instance, the word *incendie* (meaning ‘large destructive fire’) was imported from the latin word *incendium* in the early years of the 17th century. The word *tribu* (meaning ‘people’ or ‘social and political group of people sharing ethnic traits’) was also borrowed from Latin *tribus* (denoting an administrative division of the Roman people) during the 18th century. Expansions and shrinkages can either be caused by appearing/disappearing associations between existing words and concepts, or can be caused by the emergence of new words or conversely obsolescence of words (i.e. vocabulary changes). In line with this tendency to expansion rather than shrinkage, we observe that among nouns, verbs and adjectives, 1,980 of the 6,088 lexemes found in Modern French (32%) are absent from Middle & Preclassical French. Only a qualitative analysis of the observed expansions and shrinkages will reveal

whether they reflect vocabulary changes or rather changes in the word-meanings associations.

Evaluating our method for detecting onomasiological change. Let us now assess the accuracy of the observed changes by comparing them to their semasiological counterpart attested in the TLFi. When providing examples, we may use small capital letters to label our interpretation of a cluster’s meaning, based on the sense definition from the TLFi, words in the cluster’s naming and most frequent context words. We may also use double quotation marks to provide a sense definition, as provided in the TLFi. For each kind of onomasiological change, our objective is to assess whether the observed changes correspond to attested changes, are mistakes due to the clustering or are due to artifacts in the data rather than actual linguistic changes.

Evaluating new concepts. We now manually classify cases of new concepts. We sample 50 concept clusters that were not instantiated in Middle & Preclassical French but are observed in Modern French. The classification of these clusters is presented in Table 6.5. This analysis shows that our Diachronic Concept Induction method indeed reflects the emergence of new concepts in the majority of cases, with 36 clusters corresponding to actual emerging concepts whereas only 13 of these clusters correspond to concept that already existed in the data in Middle & Preclassical French.

Classification	Explanation	# of clusters
Societal changes	New concepts emerging from societal changes (in domains of politics, morals, law, etc.)	16
Science/Tech/Health	New concepts emerging from progresses in abstract or natural sciences, technology or health sciences.	20
Not a new concept	Concepts that already existed before and for which we can find occurrences for the clustered lexemes	13
Corpus errors	Pos-tag, word segmentation or lemmatization errors, etc.	1

Table 6.5: Evolution of a sample of 50 *new* concept-clusters between the Middle & Preclassical French and Modern French.

Evaluating Disappearing concepts. Let us now examine the 97 concept clusters that were instantiated in Middle & Preclassical French but not in Modern French. Our manual classification of the observed disappearing concepts is presented in Table 6.6. The 37 corpus-induced errors and sampling artifacts are not relevant for assessing the quality of the method, they are more relevant to the quality and diversity of the data themselves.

Classification	Explanation	# of clusters
Histo./Relig. changes	Concept disappearance caused by historical events/changes or because of the loss of importance of religion in the daily life.	20
Obsolete concepts	Old-fashioned concept for which there is no modern-day equivalent, or a cluster corresponding to a specific context that is no longer relevant	14
Missed concepts	Still-active concept that can even be found in other clusters with the same words	23
Unrelated words	Words that should not have been clustered together	3
Sampling artifacts	Still-active concept absent from the corpus in the Modern French period	23
Corpus errors	Pos-tag, word segmentation or lemmatization errors, etc.	14

Table 6.6: Evolution of the 97 *disappearing* concept-clusters between the Middle & Preclassical French and Modern French.

With this classification, we see that 34 of the observed concept disappearance correspond to actual changes in the language and/or in society (the historical/religious changes and the obsolete concepts), while 26 cases of disappearance actually correspond to cases where the Diachronic Concept Induction system failed to recognize that these concepts were still realized in other clusters, or that the words had no similarity in meaning in Middle & Preclassical French. While the ratio is positive in favor of accurate disappearances, there is still room for improvement, in the clustering method or in the representation technique, to lower the number of cases that do not correspond to actual disappearances.

Noteworthy, the emergence or disappearance of concepts, though reflected in

language change, rather pertains to societal and historical changes than to purely linguistic mechanisms. Diachronic Concept Induction, while focused on lexical semantics, can also be used to identify these extra-linguistic changes as they are reflected in the language; as such, we believe that the proposed method may also be useful to computational social sciences and digital humanities. Such applications are however beyond the scope of this thesis.

Evaluating expanding namings. Onomasiological changes of a more ‘purely linguistic’ nature are the extension or reduction of a concept’s naming. They reflect new conventionalized associations between words and a given concept. We sample 50 cases of expansions of concept namings and we classify them according to whether they correspond to attested semasiological changes in the TLFi. This classification is provided in Table 6.7.

Classification	Explanation	# of clusters
Attested new sense	The observed expansion of the naming corresponds to the attestation of an existing word’s new sense.	11
Attested new word	The observed expansion of the naming corresponds to the emergence of a new word.	13
False expansion	The observed expansion of the naming is contradictory with an already attested word sense.	5
Not synonyms in PFr	The cluster contains words corresponding to different concepts.	11
Sampling artifacts	The appearing word already had this sense in Middle & Preclassical French but did not appear in the corpus with this sense.	10

Table 6.7: Evolution of a sample of 50 *expanding* concept-clusters between the Middle & Preclassical French and Modern French.

We find that 11 of the 50 sampled expansions take place in clusters that conflate, in Middle & Preclassical French, words instantiating different concepts. These rates are in line with the proportions found in our synchronic evaluations in the ‘weakly related’ and ‘unrelated’ categories. Such non-synonymous word clusters decrease the precision of observed results, and future work should improve rep-

resentations and clustering methods to mitigate the formation of such clusters. There are 10 cases of expansions that are caused by the lack of occurrences for an already-existing word sense. Among the remaining clusters that are actual near-synonyms, we can find for 24 of the observed expansions a corresponding attested semasiological development, while we cannot verify such change in only 5 cases.

Interestingly, we find slightly more cases of attested new words (13 cases) than of attested new senses (11 cases), suggesting that the large amount of expanding concepts described previously could indeed be explained by the introduction of many new words to the vocabulary.

Evaluating shrinking namings. Shrinking namings on the contrary represent the loss of conventionalized word-meaning associations. Following the same process of qualitative evaluation using the reference historical dictionary, we sample 50 cases of observed shrinkages and classify them using the TLFi, as shown in Table 6.8.

Classification	Explanation	# of clusters
Attested sense loss	The observed shrink in naming corresponds to a wordsense that became obsolete.	6
Attested word obsolescence	The observed shrink in naming corresponds to a word that became archaic.	11
False shrink	The observed shrink in naming does not correspond to any attestation of an obsolete sense or a disused word.	16
Not synonyms in PFr	The cluster contains words corresponding to different concepts.	12
Corpus errors	Pos-tag, word segmentation or lemmatization errors, etc.	5

Table 6.8: Evolution of a sample of 50 *shrinking* concept-clusters between the Middle & Preclassical French and Modern French.

First, 12 of the 50 sampled shrinkages correspond to ‘weakly related’ and ‘unrelated’ concept clusters, encompassing words that are not synonyms. This proportion is comparable to that of naming expansions. There are also 5 of the sampled shrinkages that are artifacts of corpus-induced errors. Among the remaining

clusters that are actual (near-)synonyms, we find that 17 observed shrinkages correspond to attested phenomena (loss of sense of a word, or obsolescence of a word) while 16 observed shrinkages are found to be mistakes induced by the method.

Proportionally, there are more mistakes in naming shrinkages than in naming expansions. It appears the proposed Diachronic Concept Induction method may be too sensitive to disparities in data, resulting in an over-estimation of cases of ‘losses’, while it is more reliable in detecting positive developments, gains of meanings and lexemes in the naming. We hypothesize that concept-clusters may be over-fragmented (i.e. divided in smaller clusters while they should be reunited in large clusters), as the loss of a lexeme’s occurrences in a cluster is more likely to happen in small clusters than in large ones. Another argument in favor of the over-fragmentation hypothesis is the large amount of small concept-clusters that were discarded by our filtering procedure.

Summary of our findings. We start by examining emerging and disappearing concepts and we find that the errors are mainly due to clusters that should have been merged because they apparently refer to the same meaning but appear as different clusters. We then evaluate expansions and shrinkages of namings, showing that the main source of errors for both is the presence of non-synonymous words within the same cluster. Among clusters grouping genuine synonyms, a part of the observed shrinkages do not correspond to attested changes, while nearly all expansions do, indicating that the proposed approach may not treat these two types of changes equally. Overall, excluding corpus errors and sampling artifacts, the proportion of well-attested changes among observed changes is 56% for disappearing concepts, 61% for new ones, 37% for shrinking namings and 60% for expanding namings.

6.4.6 Semasiological Changes

We have described how Diachronic Concept Induction allows us to investigate onomasiological changes, and shown that the observed changes reflect attested changes in most cases. However, the proposed model of lexical semantic change also allows to explore semasiological changes like most of the existing Natural Language Processing approaches to the field.

For a given lexeme, Diachronic Concept Induction provides a sense distribution for each period, through the distribution of clusters to which its occurrences belong at each time period. Because the concept-clusters are learned in all time periods at once, cluster labels are the same between periods and so comparing

the sense distributions over time is straightforward.

Let us focus on the 4,108 target lexemes that are found in both the Middle & Preclassical French and Modern French (i.e. excluding new lexemes or lexemes that disappeared from the corpus). Among them we find 2,385 nouns, 988 verbs and 735 adjectives.

Measuring semasiological change. Many computational studies of lexical semantic change provide a measure of semasiological semantic change in order to distinguish words that underwent strong changes in meaning and those that remained semantically stable over time. Diachronic Concept Induction too allow to compute such measure: we compute the *Jensen-Shannon Divergence* (JSD) between the sense distribution of Middle & Preclassical French and the distribution of Modern French. The *Average Pairwise Distance* (APD) is the most widely used measures of lexical semantic change in the semasiological setting. Between two time periods T_1 and T_P , it's the average distance in every pairs of an occurrence from T_1 with an occurrence from T_P . Comparing the JSD-based measure to the Average Pairwise Distance between occurrences of Middle & Preclassical French and of Modern French, we find a .46 Spearman correlation between JSD and APD change measures, suggesting that the two measures likely emphasize on different aspects of semasiological change.

Characterizing semasiological change. We have shown that Diachronic Concept Induction allows us to characterize onomasiological changes, distinguishing several types of such changes. Similarly, it can characterize semasiological changes, going beyond the mere quantification of lexical semantic change. Between Middle & Preclassical French and Modern French, each of these lexemes either gains a new meaning, loses one of its early meaning, both gains and loses a meaning, or neither, in which case it is said to be stable in meaning. The distribution of these evolutions is displayed in Table 6.9.

Evolution	# of lexemes	%
Stable meaning	3165	77
Sense gain	567	14
Sense loss	222	5
Both	154	4

Table 6.9: Evolution of lexemes between the Middle & Preclassical French and Modern French.

We observe that the vast majority of lexemes are stable in meaning (reminiscent of the finding that most cluster namings too are stable over time), and that there

Classification	Explanation	# of lexemes
True gain	The presumably gained sense is indeed first attested later than Middle & Preclassical French.	31
False gain	The presumably gained sense is first attested even before Middle & Preclassical French and we have found occurrences with this meaning in Middle & Preclassical French.	14
Sampling artifacts	The presumably gained sense is first attested even before Middle & Preclassical French but we did not find occurrences with this meaning in Middle & Preclassical French.	3
Corpus errors	Pos-tag, word segmentation or lemmatization errors, etc.	2

Table 6.10: Evolution of a sample of 50 lexemes *gaining meanings* between the Middle & Preclassical French and Modern French.

are more lexemes gaining senses than losing them (in line with the observation of a growing number of instantiated concepts between the Middle & Preclassical French and Modern French).

In Tables 6.10 and 6.11, we display the qualitative evaluations conducted on samples of 50 cases of sense gains and 50 cases of sense losses respectively, between the Middle & Preclassical French and Modern French. As for onomasiological changes, we are interested in assessing whether the observed changes correspond to attested changes, are mistakes due to our clustering or are the results of sampling artifacts and errors in the corpus.

Sense gains appear to be accurately identified: 34 of the 50 sampled gains correspond to attested semasiological developments. For instance, the noun *aspect* developed the figurative meaning of SEEMING APPEARANCE, whose first attestation according to the TLFi is in 1684. Only 14 of the 50 reviewed gains could be identified in Middle & Preclassical French part of the corpus and are attested in the TLFi at an earlier date (e.g. *peser* in the figurative meaning of CAUSING A MORAL BURDEN is attested since 1342).

We see that half of the detected sense losses can be matched to senses labeled as *old-fashioned*, *obsolete* or *rare* in our reference dictionary. For instance, the

Classification	Explanation	# of lexemes
True loss	The corresponding word sense is attested as ‘old-fashioned’, ‘obsolete’ or ‘rare’.	24
False loss	The corresponding word sense is attested as still being active.	18
Not a proper sense	We could not infer a corresponding sense from the clusters’ lexemes and context words.	6
Corpus errors	Pos-tag, word segmentation or lemmatization errors, etc.	2

Table 6.11: Evolution of a sample of 50 lexemes *loosing meanings* between the Middle & Preclassical French and Modern French.

adjective *prochain* has, according to our clustering, lost its meaning of NEAR, CLOSE (IN SPACE); this sense is indeed identified as ‘old-fashioned or literary’ in the TLFi. A smaller proportion of the observed losses are not attested in the TLFi (e.g. *troupe* which still has its meaning of CATTLE HERD). We also observe that among these cases of sense loss, some of the identified senses could not be matched to an actual word senses defined in the TLFi.

In the semasiological setting, we can make a similar observation than in the onomasiological setting: gains are more accurately reflected than losses.

Polysemy and semasiological change. Hamilton et al. (2016b) proposed a *law of innovation*, stating that the degree of polysemy correlates positively with semantic change or, in other words, that highly polysemous words are subject to stronger changes than monosemous ones. Let us examine whether we find such an effect in our results. To do so, we compute the Spearman correlation between the number of senses in Middle & Preclassical French and the semantic change measure obtained with Jensen-Shannon Divergence. We find a .72 Spearman correlation between the early number of senses and the Jensen-Shannon Divergence, suggesting that higher degrees of polysemy may indeed be related to stronger semantic shifts and that the Law of Innovation, observed by Hamilton et al. (2016b) and more recently in Romance languages by Kawasaki et al. (2022), may indeed be in effect. The significance of the law of innovation is still under discussion in the NLP community (Dubossarsky et al., 2017; Uban et al., 2019; Kawasaki et al., 2022) and it would require a longer study to verify whether our data can indeed provide evidence in favor of the Law of Innovation.

We leave this perspective to future work, as we simply want to demonstrate that Diachronic Concept Induction opens the way for complete semasiological studies and discussions about laws of semantic change.

Summary of our findings. We have shown that Diachronic Concept Induction, in addition to enabling the characterization of onomasiological changes, also allows comparable investigation than other existing approaches to Lexical Semantic Change in NLP: we can derive a real-valued measure of semantic change from our clustering results, and can also use them to study other statistical laws of semasiological change proposed in the literature. However, our clustering approach allows for a more fine-grained characterization of this semasiological changes. We have evaluated the observed sense losses by comparing them to senses labeled as *old* or *obsolete* in the reference dictionary, and sense gains by comparing them to first attestation dates in the dictionary. Excluding corpus errors and sampling artifacts, our evaluation has shown that the proportion of actually attested changes among observed changes is of 50% for sense losses and 69% for sense gains.

6.5 Summary and Limitations

In this chapter, we have revisited the Concept Induction method proposed in Chapter 4 and adapted it to the diachronic setting, to deal with the multiple time periods and the absence of concept-annotated data. With this renewed framework, **Diachronic Concept Induction**, we aim to automatically uncover onomasiological changes from historical textual data, alongside semasiological changes. To learn concepts across multiple time periods, we have proposed to perform Diachronic Concept Induction on occurrences from all periods at once, thus learning a fixed set of concepts. To select the best set of clustering hyperparameters without concept-annotations, we proposed to maximize the number of retrieved synonym sets.

We have discussed the challenges of evaluating a Diachronic Concept Induction system on its capacity to detect onomasiological and semasiological changes, given that there are no concept-annotated historical data nor reference onomasiological historical resources from which we could extract complete synonym sets over time. Thus, after describing the various scenarios for lexical semantic change in the lexeme-concept mapping under both the semasiological and the onomasiological perspectives, we have proposed to use a semasiological historical dictionary to evaluate both sides of lexical semantic change.

With these adjustments, we have applied this methodology to the PRESTO corpus, an historical collection of French texts spanning from 1500 to 1949, and evaluated changes using the TLFi dictionary of French. Although the TLFi's overall perspective is synchronic, it contains diachronic annotations – such as ‘obsolete’ – as well as the dates of the first attestations of the word and its various meanings. We have shown how this resource, though exclusively semasiological, may be used to evaluate onomasiological changes. We represented occurrences using one of the CALE models presented in Chapter 5 and clustering was performed using the hierarchical bilevel method proposed in Chapter 4.

Our results highlight the significant expansion of the French lexicon and of the set of instantiated concepts between the 17th and the 19th centuries. Our qualitative evaluation reveals that onomasiological and semasiological changes most often correspond to actual changes as attested in the expert resource, and shows that the proposed method allows us to characterize lexical semantic changes and identify concepts whose associated sets of lexemes have expanded, shrunk or remained stable over time, as well as identify lexemes whose sense inventories gained new meanings or lost obsolete ones. Our method is even able to capture newly emergent concepts and disappearing concepts, often reflective of societal changes and historical events, which may be valuable for studies in digital humanities and computational social sciences. We also used our clustering results to compute a measure of lexical semantic change, and we found that this measure of semantic change was correlated with the initial number of senses of a word, echoing the Law of Innovation proposed in previous computational studies of lexical semantic changes.

Overall, the present study accomplishes a first step towards more systematic modeling of Lexical Semantic Change, offering a brand new perspective among computational approaches of semantic change. Diachronic Concept Induction opens the path for studies revisiting laws of lexical semantic change, uncovering new patterns and regularities of semantic change and more generally describing the evolution of the mapping between words and meanings in a more systematic fashion, accounting for both semasiology and onomasiology.

Limitations. One of the main limitations of this study is the lack of a systematic evaluation of the proposed approach, caused by the absence of concept-annotated historical data. Ideally, we would like to compare the choice of learning all concepts at once in the full diachronic data to the alternative of learning clusters in each time period individually and aligning clusters a posteriori, or using an evolutionary clustering algorithm over time. We would also like to

compare our hyperparameter selection criterion to other criteria, such as the widely-used Silhouette score. However, in the absence of semantic annotations on the data, we must rely on a manual qualitative evaluation of the observed changes, and conducting such evaluation on a large number of alternative systems would be prohibitively time-consuming.

The significant number of clusters/occurrences we have excluded when filtering clustering results as well as some of the findings in our qualitative evaluation (false sense loss/gain, missed concepts, ...) underscore that the proposed clustering model tend to overfragment clusters into small clusters even when they correspond to the same meaning and should be merged. Future studies should explore different clustering techniques (algorithms and hyperparameter choices) to mitigate this issue and recover more of the larger yet meaningful clusters. We hypothesize that this overfragmentation is the source of the difference in accuracy between the observed ‘positive changes’ (new concepts, expanded concept namings and word sense gains) and the observed ‘negative’ changes (disappearing concepts, shrinking namings and word sense losses) which contain more mistakes.

The PRESTO corpus covers a large time span of 4.5 centuries. It is uncertain whether the proposed Diachronic Concept Induction method would provide the same results and accuracies for a different time span, whether it may be longer or shorter.

Using the PRESTO corpus provided two significant advantages over many other historical corpora: (i) its *extended* version is very large for a diachronic corpus, allowing us to filter out many occurrences when deemed beneficial for our analyses, and (ii) it provides access to modern lemmatized forms for each word of the corpus, which is essential for comparing occurrences in old periods with occurrences in recent periods when the two periods are separated by several centuries. The proposed method could raise challenges in smaller corpora where filtering many occurrences is not possible or if modern lemmas are not provided.

A use-case for Diachronic Concept Induction. We have seen that Diachronic Concept Induction opens the way for computational studies on the semasiological and onomasiological aspects of lexical semantic change. The present chapter was indeed dedicated to the presentation of the methodology, the discussion of necessary adaptations to the diachronic setting and the general evaluation of the method in an experimental setting. To provide a concrete example of the usefulness of this method in examining complex cases of semantic change, the next chapter investigates two seemingly contradictory laws of semantic change

regarding the evolution of synonymy over time.

7

Case Study: Lexical Semantic Change in Synonyms

In this chapter, we delve into the study of the relation between synonymy and semantic change, and offer this case study to underscore the importance of considering both the semasiological aspect and the onomasiological aspect to better describe lexical semantic change. Our goal is to showcase how computational approaches, and particularly embedding-based models, can be used to tackle complex linguistic questions on the nature and characteristics of lexical semantic change, particularly with respect to its onomasiological aspects.

Synonymy, the fact that different words may be used to describe the same concept, inspired discussion in historical linguistics about its impact on lexical semantic changes. Specifically, some authors like Bréal (1897) posit that synonymous words tend to be reallocated to different semantic functions over time while authors like Stern (1921) argue that associative relations like synonymy are strong between words and tend to be preserved over time, such that if one of the synonyms changes in meaning, the associated words will undergo a parallel change and they will remain synonyms. The first phenomenon is referred to as the ‘Law of Reallocation’ (in French, *loi de répartition* in Bréal 1897, chapter 2) and the second one is called ‘Law of Parallel Change’ (Xu and Kemp, 2015).

These two laws are interesting in several aspects. Firstly, they go beyond the problem of detecting semantic change in individual words, as they concern the question of whether a lexical relationship between words – in this case synonymy – is preserved or not through time. Secondly, these laws make very strong, seemingly opposite, predictions about how synonyms evolve: either their meanings diverge (under the law of reallocation) or they remain close (under the law of parallel change). It is likely that both of these laws might be at work, but they possibly apply to different word classes, correspond to different linguistic or extra-linguistic factors, or operate at different time scales. A large-scale

study, fueled by computational methods over large quantities of texts, would be amenable to statistical analyses addressing these questions.

In this chapter, we discuss in depth the two phenomena and their consequences in the semasiological and the onomasiological perspectives. We review a previous computational study conducted by Xu and Kemp (2015), and then we build upon this first attempt at measuring the two laws by proposing an evaluation framework for computational models of semantic change in synonyms, and assess the ability of several static embeddings-based approaches to detect and explain these laws, showcasing their limitations. Finally, acknowledging the restrictions imposed by the previous study, we demonstrate how the Diachronic Concept Induction method developed in Chapter 6 can be used to shed a new light on the computational investigation of semantic change in synonyms. We leverage the onomasiological observations brought by this model and based on our experimental results, we are able to test some hypotheses about these two laws, suggesting that they may not be contradictory and may operate simultaneously on the same synonym groups.

7.1 Laws of Semantic Change in Synonymy

In Section 2.3.4 of Chapter 2, we have mentioned that various scholars in historical linguistics have identified different paths that synonyms may follow as their meanings evolve over time. Bréal (1897, chapter 2) formulates a ‘law of re-allocation’ (in the original French text, *loi de répartition*), which postulates that different lexemes must instantiate distinct meanings. Consequently, either two synonymous terms must differentiate in meaning, or one of them will fall out of use. Bréal’s hypothesis is in line with the later observations of Clark (1993), that language users naturally create *contrast* in meaning between different forms, and expect a *conventionally* established word for a given meaning.

From an onomasiological point of view, reallocation implies that concept namings, when they contain multiple words, tend to be reduced to singletons over time. From a semasiological point of view, this means that two synonymous w and v words will compete for a given meaning c_k , so that either one of them will eventually change to refer to another meaning $c_{k'}$ such that they no longer compete (in which case we refer to *differentiation*), or one of them will cease to be used by language users (in which case we refer to *disappearance*). Both perspectives are illustrated in Figure 7.1. For instance, *occupy* and *expend* used to be synonyms for the meaning of MAKE USE OF, but the former now only refers to a use of time or space (as opposed to material things); *disperse* and *disgre-*

gate were synonyms for the meaning of SCATTER but the latter word is obsolete, illustrating reallocation by disappearance.

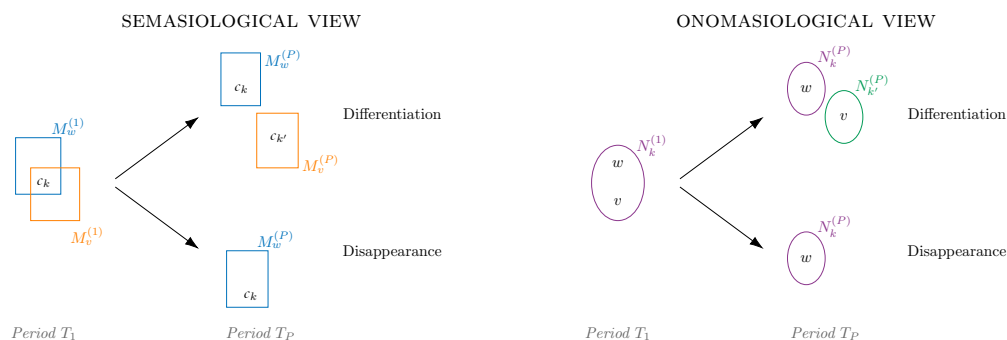


Figure 7.1: Illustration of the Law of Reallocation in the semasiological perspective (left side) and the onomasiological perspective (right side) between two time periods T_1 and T_P , T_P later than T_1 . Blue and orange rectangles correspond to the sense inventories of two lexemes w and v respectively. Purple and green ellipses correspond to the namings of two concepts c_k and $c_{k'}$ respectively. $M_w^{(p)}$ is the sense inventory of lexeme w at time T_p while $N_k^{(p)}$ is the naming of concept c_k at time T_p .

Other authors have observed different semantic change effects among synonyms, and described parallel semantic changes in sets of associated words. Stern (1921) describes how *swiftly* changed in meaning, from RAPIDLY to IMMEDIATELY, and that the same change happened for all of its synonyms (like *quickly*) or their morphologically-related words (like *swift* and *quick* respectively). Lehrer (1985) also observed that words tend to share their meanings to associated lexemes, providing the example of several animal terms metaphorically associated to human qualities (e.g. *snake* developed its secondary sense of TREACHEROUS PERSON after the word *viper* is introduced in English with both the animal and the human meanings). Observations of Stern and Lehrer indicate that while they develop new meanings, (near-)synonyms tend to remain associated and that they do not necessarily compete for meanings. From an onomasiological perspective, it means that if a word enters the naming of a given concept, its synonyms for other concepts are likely to enter this naming as well. From a semasiological perspective, parallel change implies that if two words share a first meaning and one of them has or develops a second meaning, the other word will gain the same meaning over time.

This is illustrated in Figure 7.2. As we can see, there are two possible initial states at time T_1 given two synonyms w and v both referring to a concept c_k : either one of them also refer to $c_{k'}$ (in which case only the other word will develop this new meaning), or none of them refer to $c_{k'}$ (in which case both words will develop

this meaning). Regardless of the initial state, there are also two possible outputs at time T_P : either these synonyms develop the meaning of $c_{k'}$ alongside the meaning of c_k and they end up referring to both concepts, or they completely lose the meaning of c_k in favor of $c_{k'}$.

For instance, when *viper* was introduced to English in a translation of the Bible from French around 1526, it already had both the meaning of *venomous snake* and of TREACHEROUS PERSON, as it is employed metaphorically in the Bible; while *snake* only had the animal meaning. Because of their near-synonymy, *snake* gained the meaning of TREACHEROUS PERSON and they both kept the two meanings. In Figure 7.2, the example of *snake* and *viper* is represented in the upper half of each side, whereas the change of *quickly* and *swiftly* from RAPIDLY to IMMEDIATELY would be illustrated in the bottom half of each side in the figure (both words only had the meaning of RAPIDLY at first, and kept only the meaning of IMMEDIATELY in the end).

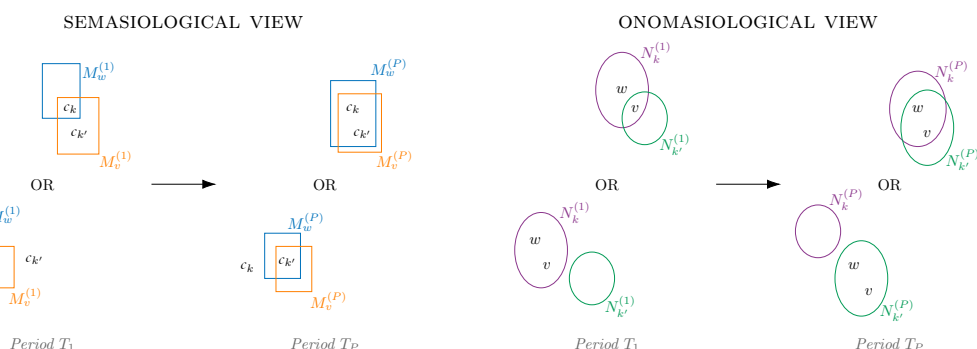


Figure 7.2: Illustration of Parallel Change in the semasiological perspective (left side) and the onomasiological perspective (right side) between two time periods T_1 and T_P , T_P later than T_1 . Blue and orange rectangles correspond to the sense inventories of two lexemes w and v respectively. Purple and green ellipses correspond to the namings of two concepts c_k and $c_{k'}$ respectively. $M_w^{(p)}$ is the sense inventory of lexeme w at time T_p while $N_k^{(p)}$ is the naming of concept c_k at time T_p .

These two laws make seemingly opposite predictions: reallocation states that synonyms are in competition for a given meaning, while parallel change argues that synonyms may develop new shared meanings. Yet we believe that they are not necessarily mutually exclusive. Stern (1921) remarks that the semantic change of *swiftly*, from the meaning of RAPIDLY to that of IMMEDIATELY, occurred after *fast* — originally meaning FIRMLY — developed its new meaning of RAPIDLY. Thus, for a time, *fast* and *swiftly* were synonyms but the latter differentiated in meaning from the former, and this in turn caused the parallel change of other lexemes of similar meaning to *swiftly*.

In this chapter, we are interested in the use of computational methods to identify cases of reallocation and cases of parallel change. The larger (onomasiological) question is whether synonymy endures over time and facilitates parallel evolutions of meaning, or if it is a purely accidental phenomenon that is prompt to disappear. This question is related to the larger on-going debate in linguistics on isomorphism of language systems and the ‘No-synonymy Hypothesis’ (Blythe and Croft, 2012; Uhrig, 2015; Levshina and Lorenz, 2022; Leclercq and Morin, 2023), and on the potential advantages of synonymy and optionality for the language system (Van de Velde, 2014; Fonteyn and Maekelberghe, 2018).

7.2 Computational Studies

Xu and Kemp (2015) proposes a computational study of both phenomena and examines the semantic change occurring in pairs of synonyms, framing the two phenomena into two laws: the *Law of Differentiation* (LD), where synonyms diverge in meaning over time; and the *Law of Parallel Change* (LPC), where synonyms undergo identical changes of meaning, thus remaining synonyms over time. As a pioneering step toward a computational analysis of the phenomena described earlier in Section 7.1, Xu and Kemp make several simplifications in their problem statement. For instance, for Reallocation, they focus solely on Differentiation and have set aside Disappearance.

The main motivation behind the study of Xu and Kemp (2015) was to automatically measure, using Distributional Semantic Models (DSM), which of the two laws was predominant in English between 1890 and 1999. To study which of the two laws actually operates, they focus on word pairs that (i) are synonyms in the 1890s and (ii) where both words changed significantly in meaning between 1890 and the 1990s. It is important to emphasize that the second criterion impacts the coverage of the study regarding the two laws. Excluding pairs in which one of the words did not change in meaning excludes cases of Parallel Change where one of the two words already encoded both the common meaning and the ‘target’ meaning (as with *viper* and *snake*). It also excludes cases of Differentiation, since only one of the two synonymous lexemes had to shift in meaning for the pair to diverge. It is unclear on whether this restriction filters out more cases of Differentiation or more cases of Parallel Changes.

As for their methodology, each word is represented as a probability distribution over context words, computed via a co-occurrence matrix (see Section 3.2). Then, they measure the (synchronic) semantic proximity of words by computing the Jensen-Shannon Divergence (JSD) between the corresponding distributions.

Semantic change for a given word is quantified by comparing the neighborhoods within the semantic spaces in the 1890s and in the 1990s. To obtain pairs of synonyms, Xu and Kemp rely on English dictionaries of synonyms from around 1890, like Fernald (1896). For every extracted synonymous pair where both words have an above-average semantic change measure value, they pick a control pair in order to compare the evolution of semantic proximity between synonyms to semantic proximity between the control pair's elements. Doing so, they avoid comparing JSD in the 1890s to JSD in the 1990s, taking into account that the overall range of values for JSD might not be the same in the 1890's than in the 1990's due to artifacts in the representational space construction.

For each pair of synonyms, they pick a control word pair that has a smaller JSD in the 1890s than the associated synonyms and that has a total degree of semantic change smaller than that of the controlled synonym pair. At a later time in the 1990s, if the JSD between the synonyms is larger than that in the control pair, they decide these synonyms have undergone the Law of Differentiation (LD). Conversely, if the synonym pair ends up closer than its control pair – even though the control pair was initially more semantically similar – they infer that the synonyms underwent the Law of Parallel Change (LPC). The control pairs are to represent an ‘average behavior’ in the representational space in pairs of non-synonym words. If the synonyms end up farther apart than words of the control pairs, it means that they are less semantically close than non-synonyms, so the synonyms should have differentiated. On the contrary, if they end up closer than the non-synonyms, despite the fact that words in the control pairs were closer at the beginning, it should mean that their meanings have converged and that they underwent parallel change.

Conducting their study among Adjectives, Nouns and Verbs, they ultimately found that most pairs (around 60%) have undergone the Law of Parallel Change, which would be the dominant law compared to the Law of Differentiation. Interestingly, they observe a similar effect in antonym pairs.

Xu and Kemp (2015)'s study faces several shortcomings. First, the restriction on pairs that both words in a pair must have changed may affect the actual proportions of Differentiation and Parallel Change. It removes cases where only one lexeme shifts, which are typical of Differentiation, as well as cases where one lexeme already possesses both the original and the target meanings, which are typical of Parallel Change, and we have no indication as to which of the two cases is more impacted by this exclusion. Second, the proposed criterion to select control pairs may select near-synonyms as controls (e.g. *instructive* and *interesting*), because control pairs are constrained to be *closer* in JSD in the 1890s

than the associated synonym pairs. Controlling the change of synonyms with other synonyms should be avoided, as the semantic change path taken by the control synonyms may completely reverse the semantic change path observed in controlled synonyms. Indeed, if both the controlled synonyms and the control synonyms undergo the Law of Differentiation, but the semantic differentiation is stronger in the control pair than in the controlled pairs, then the control JSD will be larger in the 1990s than the synonym JSD, and the synonyms will be wrongly labelled as ‘Parallel Change’.

Crucially, Xu and Kemp (2015) did not compare their predictions to any ground-truth in the 1990s, so there is no evaluation of the reliability of their method. These limitations cast doubt on the obtained proportion of LD and LPC and on the claim that Parallel Change is the dominant tendency..

7.3 Evaluation and Predictions from Static Representations

The goal of this section is to revisit the study of Xu and Kemp, and to provide an evaluation framework to assess a model’s ability to accurately describe semantic changes in synonym pairs. Building upon Xu and Kemp (2015)’s study, our methodology will be limited to static representations, with a single vector per lexeme and per time period.

We revisit the problem statement in a more suitable fashion to provide an evaluation framework, crossing two reference lexical resources to obtain a dataset of pairs of synonyms that diverged in meaning or not over time. Then, in an attempt to explain change with computational models, we propose several approaches to predict whether a pair underwent LD or LPC using static embeddings. Our experiments show that these models are partially predictive of the change observed in synonym pairs, and further analyses suggests that static models struggle to deal with polysemy, and that crossing two lexical resources from different origins and different periods has significant drawbacks, pointing to the need for a different kind of approach.

7.3.1 The task of tracking change in synonymy

Our overarching goal is to develop new computational approaches that are able to automatically predict which pairs of synonymous words underwent LD or LPC. These predictions could be used as a first step towards providing a more refined and statistically robust analysis of the two laws. A crucial step towards develop-

ing such an approach is to compare it to some ground truth. Otherwise, there is no way to assess whether statistics obtained for LD or LPC are indeed reliable, a problem faced by Xu and Kemp (2015).

Unfortunately, there is no existing large-scale resource that records instances of LD/LPC, beyond a handful of examples found in research papers and textbooks in historical linguistics. What does exist, however, are historical lists of synonyms, which we can compare to obtain some form of ground truth. This forces us to consider a slightly different methodological framework, focusing on a more limited prediction task, namely detecting pairs of synonyms at time T_1 that have remained synonymous or that are no longer synonymous at time $T_P (> T_1)$.

Formalization Let us consider two words, w and v in the lexicon W . Let $M_w^{(p)}$ (respectively $M_v^{(p)}$) be the *sense inventory* of w (resp. v) at time T_p , i.e. the set of concepts it can refer to. The words w and v are synonyms at time T_p if they have at least one meaning in common, that is if $M_w^{(p)} \cap M_v^{(p)} \neq \emptyset$. In theory, the Law of Differentiation occurs if $M_w^{(1)} \cap M_v^{(1)} \neq \emptyset$ but $M_w^{(P)} \cap M_v^{(P)} = \emptyset$ (they lost their common meanings). The Law of Parallel change happens if $M_w^{(1)} \cap M_v^{(1)} \neq M_w^{(P)} \cap M_v^{(P)}$ (the set of their common meanings changed between T_1 and T_P) and $|M_w^{(1)} \cap M_v^{(1)}| \leq |M_w^{(P)} \cap M_v^{(P)}|$ (the set of their common meanings is not smaller in T_P than in T_1), assuming that $M_w^{(1)} \cap M_v^{(1)} \neq \emptyset$.

Unfortunately, available resources do not provide access to $M_w^{(p)}$ and $M_v^{(p)}$ over time. From lists of synonyms, we can simply know whether $M_w^{(p)} \cap M_v^{(p)}$ was non-empty for $p = 1$ (i.e. whether they were synonyms at T_1) and whether it was non-empty for $p = 2$ (i.e. whether they are synonyms at T_2). We denote $\mathcal{S}_{(p)}$ the set of synonym pairs at time T_p , i.e. all pairs $\mathcal{S}_{(p)} = \{(w, v) \in W \times W \mid M_w^{(p)} \cap M_v^{(p)} \neq \emptyset\}$.

Task Formulation. Given that we do not have access to the $M_w^{(p)}$ and $M_v^{(p)}$, developing a computational system that attempts to directly predict LD and LPC, and even the construction of an evaluation benchmark for evaluating such a system, are very challenging tasks. The main difficulty is on the discrimination of cases of Parallel Change. With historical lists of synonyms only, considering a pair of synonyms at T_1 that are also found synonyms at T_P , we cannot determine whether it is because the set of their common meanings has not changed or whether it is because of parallel change, or even because they had more than one common meaning at T_1 and lost only some of them but not all.

We suggest to narrow down our target problem to the task of predicting, for a given pair of synonymous words (w, v) at T_1 , whether (w, v) are still synonymous or not at T_P . More formally, we are concerned with the following binary

classification problem:

$$f : \mathcal{S}_{(1)} \rightarrow \{\text{"Syn"}, \text{"Diff"}\}$$

$$(w, v) \mapsto f((w, v)) = \begin{cases} \text{"Syn"} & \text{if } (w, v) \in \mathcal{S}_{(P)} \\ \text{"Diff"} & \text{otherwise} \end{cases}$$

where "Syn" indicates that words (w, v) that were synonymous at T_1 remain synonymous at T_P , while "Diff" signals that they are no longer synonymous at T_P . This simpler problem leads to a more operational evaluation procedure, which does not require access to sets of meanings, but only to lists of synonyms $\mathcal{S}_{(1)}$ and $\mathcal{S}_{(P)}$. It should be clear that predicting which synonym pairs remain ("Syn") or cease to be synonyms ("Diff"), will provide some information about LPC and LD, although the mapping between the two problems is not one-to-one. Even if "Diff" covers fairly well LD, a pair that is still synonymous at T_P could either be a case of LPC (their shared meaning changed the same way for both words) or a pair of words that simply have not changed in meaning at all (or at least whose shared meaning is unchanged).

Now turning to designing a computational system that detects "Syn" vs. "Diff" and building upon the approach of Xu and Kemp (2015), a natural question that emerges is whether Distributional Semantic Models and, more particularly static vector representations, commonly used for detecting LSC in individual words and strong competitors on Lexical Semantic Change Detection benchmarks (Schlechtweg et al., 2020), are able to capture synonym changes. More specifically, the main hypothesis we will be testing in this work is that the evolution of synonymous pairs can be tracked through their word vector representations at T_1 and T_P .

This will be instantiated into different unsupervised and supervised models in Section 7.3.3.

7.3.2 Evaluation Dataset

This section presents a dataset designed to track the evolution of English synonymous word pairs between two time periods T_1 and T_P , with $T_P > T_1$. Specifically, the two time periods considered are the 1890's decade (T_1) and the 1990's decade (T_P). For extracting synonymous pairs in the 1890's (noted $\mathcal{S}_{(1)}$), we use Fernald's *English Synonyms and Antonyms* (Fernald, 1896), as Xu and Kemp (2015) did. Pairs were selected based on a set of specific target words (see appendix A.4.1). As shown in Table 7.1, we obtain 1,507 adjective pairs, 2,689

noun pairs and 1,489 verb pairs. To assess whether these word pairs are still synonyms in the 1990's, we use WordNet (Fellbaum, 1998), which we have already presented in previous chapters. WordNet provides us with $\mathcal{S}_{(P)}$: specifically, we considered that a pair of words/lemmas $(w, v) \in \mathcal{S}_{(1)}$ are still synonymous if they point to at least one common *synset* in WordNet.

Synonyms pairs	ADJ	NOUN	VERB	All
Synonyms at T_1	1507	2689	1489	5685
& synonyms at T_P	202	347	311	860
& synonyms at T_P (%)	13.4	12.9	20.9	15.1
& hypernyms at T_P	0	858	398	1256
& hypernyms at T_P (%)	0.0	31.9	26.7	22.1
& hyp. at T_P (1) (%)	0.0	23.2	22.5	16.9
& hyp. at T_P (2) (%)	0.0	6.9	3.5	4.1
& hyp. at T_P (3) (%)	0.0	1.4	0.5	0.8

Table 7.1: Numbers of synonymous pairs extracted from Fernald (1896) (T_1) displayed by Part-of-Speech, and numbers of those that are also considered as synonyms or hypernyms/hyponyms in WordNet (T_P). For hypernyms, we detail the proportions of hypernym/hyponym pairs that are separated by 1, 2 or 3 nodes in the WordNet graph.

The construction of this dataset relies on two crucial hypotheses. First, both lexical resources rely on the same definition of synonymy. Second, $\mathcal{S}_{(P)}$ meets some exhaustivity criterion, in the sense that $(w, v) \in \mathcal{S}_{(1)}$ not appearing in $\mathcal{S}_{(P)}$ should indicate that w and v are no longer synonymous at T_P , and not be due to a lack of coverage of the resource (i.e., a false negative). WordNet is assumed to be exhaustive enough, as we checked that every word involved in at least one synonymous pair has its own entry in WordNet's database.

Table 7.1 provides some detailed statistics on the evolution of synonymous pairs between 1890s and 1990s decades, overall and for different parts of speech. A first observation on these datasets is that the proportion of pairs that are still synonyms at T_P ("Syn") is globally 15.1%. This implies that most synonymous pairs underwent differentiation. While this does not provide information about how change occurred between T_1 and T_P for the remaining 84.9%, it's a clue that the Law of Differentiation may be a dominant phenomenon among synonyms, which contradicts the findings of Xu and Kemp (2015).

We exploit the structure of the WordNet database to analyze the different cases of "Diff". WordNet includes lexical relations of hyper-/hypo-nymy (e.g., *seat/bench*) as well as holo-/mero-nymy (e.g., *bike/wheel*) and antonymy (e.g., *small/large*)

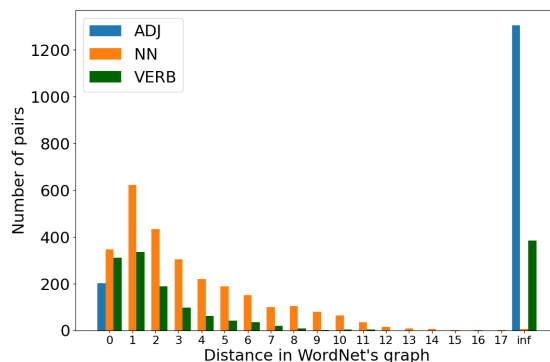


Figure 7.3: Distribution of shortest distances in WordNet’s hypernymy graphs between pairs of words that were synonymous at $T_1 = 1890$ s. This represents the minimum number of nodes (synsets) between the senses of the two words. An infinite distance means that there is no path in WordNet’s hypernymy graph between them.

defined over synsets¹. Note that the hyper-/hypo-nymy relation does not exist in WordNet among adjectives. Among nouns and verbs, we observe that around 30% of pairs that were synonyms at T_1 are in an hyper-/hypo-nymy relation at T_P and two thirds of them are direct hypernyms in WordNet (their synsets are direct parent/child) indicating the preservation of a very close semantic link. Figure 7.3 illustrate the distribution of distances in WordNet’s hypernymy graphs for all pairs of words that were synonyms in T_1 . We considered only shortest distances, i.e. the distance between the closest synsets associated to paired words. An infinite distance means that there is no path in the graph between them (which is, for instance, the case of all adjectives, as hypernymy is not defined in WordNet for this part-of-speech). A distance of 0 means that the paired words are actually synonyms in WordNet, while a distance of 1 implies there is direct hypernymy between them.

One cannot entirely exclude that $\mathcal{S}_{(1)}$ is based on a ‘loose’ notion of synonymy and includes some hyper-/hypo-nyms as synonyms. But even if we extend the notion of synonymy at T_P to include these cases, we would have only around 45% of all pairs still considered synonyms among nouns and verbs. This indicates that “Diff” largely remains the most common phenomenon, with an estimated proportion between 55% and 80%.

In the absence of additional indication that some of these hyper-/hypo-nym cases at T_P are indeed synonyms, or that they may also have been hyper-/hypo-nym at T_1 , we decided to still consider them as instances of “Diff”. Another argument for this decision is precisely that there are well-known reported cases of lexical

¹ As we did for synonyms, we assume that two words w_1 and w_2 are instances of one of these relations R if R holds for one of their corresponding synset pair.

semantic changes in which the meaning of a particular word in effect "widens" to denote a larger subset (i.e., becomes its own hypernym): this is the case of *dog* in English that used to denote a specific breed of dogs and now refer to all domesticated canines (Traugott and Dasher, 2002).

7.3.3 Models

Now that we have curated a dataset for the task of predicting whether synonyms keep their association or not, we can evaluate computational models on their ability to represent such changes. This section presents two classes of computational approaches, unsupervised and supervised, for predicting whether pairs of synonyms at T_1 remain synonyms ("Syn") or cease to be so ("Diff") at a later time T_P . Common to all of these approaches is that they are based on two vector spaces of static representations, one for each time periods.

Temporal Vector Spaces For time periods T_1 and T_P we have two vector spaces $V^{(1)}$ and $V^{(P)}$ encoding the (possibly changing) word meanings at T_1 and T_P . Each word is represented by a single vector in $V^{(1)}$ and $V^{(P)}$. Thus, for each synonym pair (w, v) , we have two pairs of vectors : $(\mathbf{w}^{(1)}, \mathbf{v}^{(1)}) \in V^{(1)} \times V^{(1)}$ and $(\mathbf{w}^{(P)}, \mathbf{v}^{(P)}) \in V^{(P)} \times V^{(P)}$.

Specifically, we use pre-computed SGNS trained by Hamilton et al. (2016b) for each decade between 1800 and 2000 on the *English* part of the GoogleBooks Ngrams dataset² and we keep only that of the 1890s for $V^{(1)}$ and that of the 1990s for $V^{(P)}$. For any word $w \in W$ and any time period T_p , $\mathbf{w}^{(p)} \in V^{(p)}$ is a single 300-dimensional vector. In both periods we observe that the distribution of cosine distances among synonyms from our reference lists is significantly smaller than the distribution of cosine distance of arbitrary word pairs (see Figure 7.4), showing that synonymy is accurately reflected in our representational spaces.

As we have seen in Chapter 3, traditional computational approaches for Lexical Semantic Change Detection using static embeddings rely on self-similarities over time for a given word, e.g. a cosine distance between two representations of the word from two different periods. We refer to this as the *Diachronic Distance*, noted here $DD^{(1,P)}(w)$ for a word w between time periods T_1 and T_P .

There is no canonical distance for comparing *pairs* of word vectors across time, but one can instead rely on comparing the pairwise word vector distance at each time stamp T_p ; we call this *Synchronic Distance*, denoted $SD^{(p)}(w, v)$ at a time T_p between words w and v . By construction of $V^{(p)}$, $SD^{(p)}(w, v)$ is smaller for

²<https://storage.googleapis.com/books/ngrams/books/datasetsv3.html>

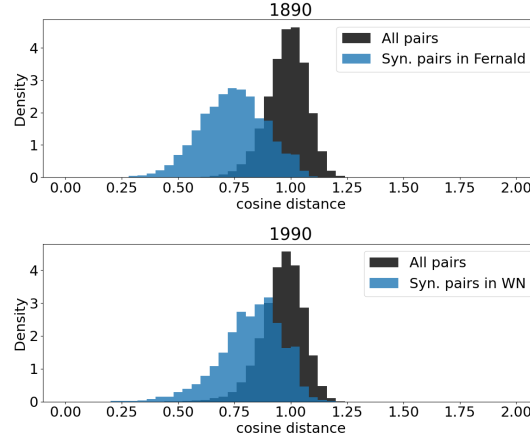


Figure 7.4: Distribution of cosine distances among synonyms (blue) and among other word pairs (black), for the early period (top) and the later period (bottom).

words (w, v) that appear in similar contexts than for unrelated words, which corresponds to words with higher similarity in meaning according to the Distributional Hypothesis (Harris, 1954). The two types of distances for two time stamps T_1 and T_P are described in Figure 7.5.

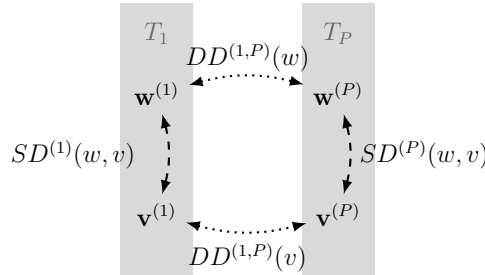


Figure 7.5: Pairs of word embeddings for words w and v at two time periods T_1 and T_P and associated diachronic and synchronic distances.

Unsupervised Methods. Following Xu and Kemp, we first propose an unsupervised model to predict semantic divergence in synonyms. The central idea is that if the set of shared meanings between words w and v changes from T_1 to T_P , this should be reflected in a changes of their usage contexts, and thus in the synchronic distances (SD) between $\mathbf{w}^{(p)}$ and $\mathbf{v}^{(p)}$.

As we have seen earlier, synonyms should be close in the representational space. We assume that there exists a threshold δ_p for each time period T_p such that

$$\forall (w, v) \in \mathcal{S}^{(p)}, SD^{(p)}(w, v) \leq \delta_p.$$

This entails that for a given pair (w, v) :

$$SD^{(p)}(w, v) > \delta_p \Rightarrow (w, v) \text{ are not synonyms.}$$

In this setting, one can compare the synchronic distances in $V^{(1)}$ and in $V^{(P)}$ and decide if the pair has differentiated or remained synonymous. Let (w, v) be a pair of synonyms at T_1 , so that we have that $SD^{(1)}(w, v) \leq \delta_1$. If (w, v) are no longer synonyms at time T_P then $SD^{(P)}(w, v) > \delta_P$. Combining these two inequalities, we would say that a pair of synonyms at T_1 has differentiated at T_P if:

$$\underbrace{SD^{(P)}(w, v) - SD^{(1)}(w, v)}_{= \Delta(w, v)} > \delta_P - \delta_1.$$

Ideally one could imagine that the distance threshold δ_p under which words are considered to be synonyms should be independent of the time period T_p . Empirically however, because word embeddings are not necessarily build with scale constrains, there might be a dilation or shrinking in the overall synchronic distances between T_1 and T_P . Thus, we will assume that $\delta_P = \delta_1 + \tau$ where $\tau \in \mathbb{R}$ is a dilation indicator. We can then estimate the dilation between the representations in the two time periods by the average shift between the synchronic distances in all word pairs.

$$\tau = \frac{1}{|W|^2} \sum_{(w_1, w_2) \in W \times W} \Delta(w_1, w_2)$$

Our decision rule could then be rewritten as:

$$f(w, v) = \begin{cases} \text{“Diff”} & \text{if } \Delta(w, v) \geq \tau \\ \text{“Syns”} & \text{otherwise.} \end{cases} \quad (7.1)$$

This approach is shortly denoted “ Δ ” in our experiments. It diverges from the approach of Xu and Kemp (2015) who rely on control pairs instead of a distance threshold. For the sake of comparison, we implemented their method of selecting control pairs and comparing synonym pairs to them (called “XK controls” in the experimental results). For synonym pair (w, v) , the corresponding control pair $(c_1, c_2) \in W \times W$ is obtained by selecting a pair of words such that $SD^{(1)}(c_1, c_2) < SD^{(1)}(w, v)$ and $DD^{(1,P)}(c_1) + DD^{(1,P)}(c_2) < DD^{(1,P)}(w) + DD^{(1,P)}(v)$ (they start closer than the synonyms and their total amount of semantic change is lower than that of the synonyms).

POS	Control pairs
ADJ	brownish/red, kindly/mild, teeming/agricultural, likeliest/meaningless, various/heterogeneous, barbarous/cruel, abandoned/unsuccessful, trojan/escaping, subjective/relative, reliable/readable.
NN	diphtheria/typhus, muskets/pistol, surgery/appendicitis, beech/apples, accountants/prints, commodity/substances, cups/pots, wife/grandmother, fool/fisherman, obstacles/multiplication.
VERB	moan/groan, divide/span, needed/secured, flowed/flooded, stall/owned, told/asked, mentioned/described, cooperate/accord, copy/filed, increased/diminished.

Table 7.2: Random samples of 10 selected control pairs following the criterion of Xu and Kemp (2015). In italic are control pairs which can be considered (near-)synonyms.

Unlike Xu and Kemp, we did not use co-occurrence counts for representation and Jensen-Shannon Divergence as synchronic distance, we use SGNS word representations and cosine distance to measure synchronic semantic proximity. In Table 7.2 we display randomly chosen samples of 10 control pairs obtained through Xu and Kemp’s method, showing that the presence of near-synonyms is not rare.

Supervised Methods Approaches described so far use the labels in the dataset ("Syn" and "Diff") only for evaluation purposes. Alternatively, we can also use part of the labeled data to learn a *supervised* classifier to predicts semantic divergences. This allows us to assess how well the proposed semantic space models as well as distances computed in them can be predictive of synonymy preservation or loss. Concretely, we trained several Logistic Regression (LR) models³. An advantage of Logistic Regression is its interpretability: if the model indeed manages to accurately predict semantic change in synonyms, it could be interesting to take a look at the features’ weights to see which quantities contributed the most to the prediction. As input features, we use Synchronic Distances in both periods, Diachronic Distances for both words, as well as the corpus frequencies for both words at both time periods⁵. Previous studies have indeed found that frequency changes can be correlated with semantic changes (Kulkarni et al., 2015; Schlechtweg et al., 2019).

Synchronic Distance definitions. In practice, we experiment with two different types of Synchronic Distances SD between words. The first is the cosine

³ Implemented with the *scikit-learn* library for Python⁴

⁵ Word frequencies were estimated from the Corpus of Historical American English (COHA) list,⁶ which has the advantage to be genre-balanced.

distance:

$$SD^{(p)}(w, v) = \text{cos-dist}(\mathbf{w}^{(p)}, \mathbf{v}^{(p)}).$$

Another measure of semantic proximity is based on the shared word neighborhood between the two vectors u and v :

$$SD^{(p)}(w, v) = \text{jacc-dist}(n_{k,w}^{(p)}, n_{k,v}^{(p)}) = 1 - \frac{|n_{k,w}^{(p)} \cap n_{k,v}^{(p)}|}{|n_{k,w}^{(p)} \cup n_{k,v}^{(p)}|}$$

with $n_{k,w}^{(p)}$ being the set of the k -nearest neighbors of the point representing w in the vector space at time T_p , and *jacc-dist* being the Jaccard distance. This measure is ranged between 0 and 1. In experiments,

Diachronic Distance definitions. We also try two different definitions of Diachronic Distance *DD*. First we test **neighborhood-based DD**, comparing sets of neighbors at T_1 and T_P :

$$DD^{(1,P)}(w) = \text{jacc-dist}(n_{k,w}^{(1)}, n_{k,w}^{(P)}).$$

Alternatively, we also compute a **cosine distance-based DD**, defined as the cosine distance between $\mathbf{w}^{(1)}$ and $\mathbf{w}^{(P)}$ after aligning the vector space $V^{(P)}$ to $V^{(1)}$ using Orthogonal Procrustes (Hamilton et al., 2016b; Schlechtweg et al., 2019, 2020). Denoting $\mathbf{w}_{align}^{(P)}$ the vector $\mathbf{w}^{(P)}$ after alignment with Orthogonal Procrustes, we have:

$$DD^{(1,P)}(w) = \text{cos-dist}(\mathbf{w}^{(1)}, \mathbf{w}_{align}^{(P)}).$$

7.3.4 Experiments

Target words selection. We use a unique vocabulary W composed of 6,453 adjectives, 16,135 nouns and 10,073 verbs. The process to select words is described in Appendix A.4.1.

Dataset splits. For every POS tag, we have a set of word pairs that are synonymous at T_1 . We call *ALL* the dataset that comprises all pairs indistinctly of their POS. These datasets (ADJ, NN, VERB or ALL) are individually shuffled and 33% of their samples (pairs) are set aside for testing. For each dataset, a model is trained on the 66% remaining pairs and evaluated on the test part. Presented results are averaged over 20 random train/test splits.

Hyperparameters. For Synchronic Distances, cosine-based SD does not induce any additional hyperparameter. For neighborhood-based SD we use values of k

in $\{5, 10, 15, 20, 40, 100\}$. For neighborhood-based Diachronic Distances, we used a fixed neighborhood size of 100 as in Xu and Kemp (2015); the cosine-based DD using Orthogonal Procrustes does not require any additional parameter. For frequency, we decided to use raw counts or binned frequencies. Frequency bins are computed following the procedure described in Appendix A.4.2. In Appendix A.4.4 we discuss which choices of distances and sets of hyperparameters were optimal for each part-of-speech tags.

Evaluation Metrics. We use two standard evaluation metrics: F_1 score and *Balanced Accuracy (BA)*. F_1 scores are computed for both classes, denoted as “ $F_1(\text{Syn})$ ” for *Syns* and “ $F_1(\text{Diff})$ ” for *Diff*. BA is defined as the average of recalls for both classes, and provides an accuracy measure robust to class imbalance. We also display the percentage of instances predicted as *Diff* (“%D”).

Baselines. We compare against three baselines. The first two baselines are constant output classifiers, only predicting “*Syn*” or “*Diff*”, respectively. They are expected to have a balanced accuracy of 50%, as they would be fully accurate for one class and always wrong for the other. The third baseline (*LR Frequency*) is a Logistic Regression model trained *only* with frequency-related features, without any knowledge on the semantic aspects of the pair (neither *SD* or *DD*).

	ADJ	NN	VERB	ALL			
	Balanced Accuracy			BA	$F_1(\text{Syn})$	$F_1(\text{Diff})$	%(D)
All (<i>Syn</i>)	.50	.50	.50	.50	.48	0	0
All (<i>Diff</i>)	.50	.50	.50	.50	0	.81	100
LR Frequency	.51	.56	.59	.55	.35	.74	75
XK controls	.52	.49	.51	.50	.33	.67	65
Δ	.50	.49	.51	.50	.27	.73	75
LR multi	.62	.64	.65	.62	.51	.71	57
LR multi (without DD)	.61	.64	.63	.62	.51	.71	57
Non-linear multi	.60	.64	.65	.62	.50	.74	63

Table 7.3: Performances of the different approaches. Results are averaged over 20 random splits.

Classification results. Performances on the test parts of the different datasets are displayed in Table 7.3. The results presented here are a subset of those in Liétard et al. (2023), and the more complete table of results investigating the benefit of each feature to the logistic regression models can be found in Appendix A.4.5. ‘LR multi’ refers to the Logistic Regression models with all features (SDs, DDs, Frequencies).

In line with the dataset’s proportions, all models predict a majority of “Diff”, even unsupervised ones (the XK controls and the Δ methods) that are not aware of the proportion of each class in the dataset. While our task does not directly address the opposition between the Law of Differentiation and the Law of Parallel Change but rather distinguishes differentiating synonyms from remaining synonyms, this provides empirical clue in favor of LD, contradicting Xu and Kemp (2015). In our setting, their control-based decision rule also predicts a majority of *Diff*, contrarily to their experiments. However, this discrepancy may simply reflect the differences of representations and synchronic distances. We observe substantial differences in F_1 scores between the two classes, $F_1(\text{Syn})$ being lower than $F_1(\text{Diff})$ across all models. Because the dataset is imbalanced, with more instances of *Diff* than of *Syn*, we observe that improvements to $F_1(\text{Syn})$ have more impact on the overall Balanced Accuracy than improvements of $F_1(\text{Diff})$. Overall, the obtained balanced accuracies range between 0.49 and 0.65, leaving room for improvement.

Unsupervised methods fail to improve over baselines and are outperformed by the frequency-based baseline models which performs surprisingly well given that it has not access to semantic information. In contrast, the Logistic Regression model substantially improves over the baselines, Xu & Kemp’s control pairs and the Δ -based method. For comparison, we included a non-linear classifier (SVM with Gaussian kernel) using the same input features, but there is little to no improvement compared to the linear model. While supervised classifiers perform better than unsupervised models, they do not reach Balanced Accuracy above 0.65. Particularly, they struggle more with capturing the *Syn* class than the *Diff* class, showing that the proposed features, based on static representations, do not fully capture lexical semantic change in synonyms.

The gap between baselines and models is larger for nouns and smaller for verbs. Interestingly, removing Diachronic Distances brings little to no degradation to the model’s performance. Our explanation of this behavior is that individual changes of words are of lesser importance to the problem of *Syn/Diff* than synchronic semantic proximity between words. This observation underscores the importance of an onomasiological perspective: measuring whether words instantiate the same meaning is more informative than measuring whether they individually changed in meaning or not. In future work, diachronic distances could still be used to distinguish pairs of remaining synonyms among the *Syn* class that did not change in meaning and pairs that went under change, which would be cases of LPC.

Using WordNet, we discuss two potential sources of error in detecting a change

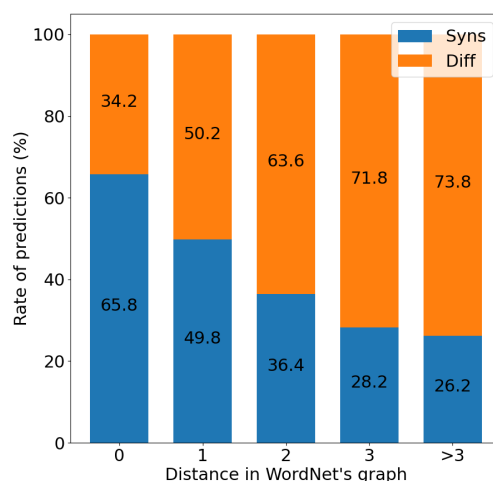


Figure 7.6: Proportions of predictions of the models w.r.t. the actual distance d in WordNet of *noun* pairs. Pairs with $d = 0$ are synonymous pairs in WordNet.

in synonymy: polysemy and hypernymy. We study predictions of our Logistic Regression model on the noun dataset.

Polysemy. WordNet provides different sets of synonyms for each entry, corresponding to different senses or usages, and therefore we can measure the polysemy of a word w at T_P by counting the number of *synsets* associated to w . We found that *Diff* pairs wrongly classified as *Syn* tend to be those whose second term has fewer senses⁷ (6 senses on average, while well-classified *Diff* which have 8 senses on average). As we use static embeddings and no sense disambiguation method, our model is subject to the complexity brought by polysemy. In Zamora-Reina et al. (2022), a shared task of Lexical Semantic Change detection in Spanish, the best performing models are those using sense-disambiguation methods. This finding highlights the importance of addressing polysemy as a potential confounding factor.

Distances in WordNet. Figure 7.6 displays the percentage of prediction with respect to shortest distance between the two words of *noun* pairs in WordNet's graph. The distance d is the minimum number of nodes separating the synsets associated to the two words. As expected, the model predicts *Diff* with increasing frequency as d increases. Interestingly, for $d = 1$ (direct hypernymy), there is still an important proportions of predicted *Syn*. This highlights that our model

⁷ It happens mostly on the second term rather than the first because in Fernald's dictionary, the first term is a head-word that is generally unambiguous (like *kill*) while candidate synonyms, mapped as second terms in pairs, may be more polysemous (like *butcher*).

of semantic change in synonyms has difficulties to distinguish hypernymy from synonymy and may conflate the two relationships.

7.3.5 Interim Conclusion

Summary and discussion. Building on the prior work of Xu and Kemp (2015), we have proposed an evaluation framework to assess a computational model's ability to accurately predict lexical semantic change in pairs of synonyms. Dealing with constraints imposed by the non-availability of rich diachronic lexical resources, we shifted to focus from the opposition between LD and LPC to the simpler opposition between pairs of synonyms that differentiated and those that did not, and remained synonyms. Crossing lists of synonyms from two periods, we have elaborated a dataset of synonym pairs that did or did not differentiated between the 1890s and the 1990s in English, finding that a majority of them did undergo the Law of Differentiation studied by Xu and Kemp. Evaluating a series of unsupervised and supervised computational models of semantic change that rely on static representations, we found that these representations only partially reflect semantic change in synonyms, and that they are crucially sensitive to polysemy and potentially influenced by hypernymy between words in a pair. We have also seen that synchronic similarities between words play a more important role than individual measures of change, advocating for more onomasiological considerations of the problem. Overall, our study provides empirical evidences that differentiating synonyms are more common than persisting synonyms.

Limitations. Yet, our study has its own limitations in its description of the linguistic phenomena. First, the *Syn/Diff* problem does not reflect the initial question of LD/LPC. In particular, the *Syn* class of pairs that remained synonyms contains pairs that underwent LPC and pairs whose shared meaning remained unchanged. The latter does not play a role in the LD/LPC dichotomy. In turn, both LD and LPC are themselves reductions of two larger phenomena. The original *loi de repartition* of Bréal (1897) argues that synonyms may undergo differentiation or disappearance of one of the two terms, a case that has been excluded from both Xu and Kemp's and our computational studies. The Parallel Change studied in Xu and Kemp's study only accounts for cases where the two words changed in meaning, letting aside cases like *snake* and *viper* where one of the two words already had a secondary meaning to share to its synonym. These restrictions are imposed by the semasiological-first approach on this linguistic problem. Both computational studies indeed took only the perspective of words as central entities, without considerations to individual meanings. This is another limitation

of static representations, as they do not readily allow to precisely study of meanings in isolation. We believe that a more onomasiological study of Reallocation and Parallel Change would provide better understanding of the extent of both phenomena.

Second, in both the dataset presentation and experimental results, we discussed how hyper/hypo-nymy could be confused with synonymy. We made the assumption that Fernald (1896) and Wordnet (Fellbaum, 1998) used sufficiently similar notions of synonymy such that our labels *Syn/Diff* are relevant. However, thesauri like Fernald (1896) are created as a tool for writers and authors to avoid repetition and redundancy, thus including broad synonym lists that include hypernyms (e.g. instead of repeating *the bench*, you could say *the seat*). Conversely, WordNet adopts a strict policy in distinguishing synonyms and hypernyms. This creates a potentially unfairness between the two conceptions of synonymy, one that is more loose and one that is more strict. We still miss guidelines or insights about whether to include some cases of hypernymy among synonyms at T_P ; alternatively, automatically detecting pairs of hypernyms to remove them from the source material at T_1 is still an open research question (Camacho-Collados et al., 2018; Bai et al., 2022; Liu et al., 2023). Thus, aligning lists of synonyms from different periods on a common notion of synonymy is challenging.

Towards a new approach. These limitations motivate using the Diachronic Concept Induction method presented in Chapter 6. With this approach, we are able to find groups of synonyms across multiple time periods directly from the data. A first advantage is that we can retrieve groups of synonymous words and not only pairs, allowing us to extend the scope of the study. Secondly, thanks to contextualized representations, words can belong to multiple concept clusters, thus this method natively handles sense differentiation for polysemous lexemes. Moreover, concept clusters are learned in all time periods at once using the same representation process, ensuring that the notion of synonymy captured by these clusters is common to all time periods. Finally, because we have access to both a word’s sense distribution over time and a concept’s naming in multiple time periods, we are able to represent *both* the semasiological and the onomasiological aspects of the problem. This is particularly interesting, as we believe that an onomasiological description of the problem could lead to account for Reallocation and Parallel Change in all the scenarios depicted in Figures 7.1 and 7.2 of Section 7.1.

7.4 Changes in Synonymy with Diachronic Concept Induction

In this section, we elaborate on the use of Diachronic Concept Induction to address the question of semantic change in synonyms, and in particular the question of Reallocation and Parallel Change. This section consists of two parts.

We will first discuss how Reallocation and Parallel Change can be identified with Diachronic Concept Induction systems, and the extent to which the proposed empirical definitions cover both phenomena. Then, we will investigate the cases of Reallocation and Parallel Change found in the PRESTO corpus using the results obtained in the previous chapter. We decide not to use the previous evaluation framework, as we have seen that crossing lists of synonyms from different time periods with potentially two different purposes could create confusion in the analysis. We can't use the same training data neither, because the GoogleBooks Ngrams data does not provide full text data and only n-grams frequency counts. Because our bi-level clustering method is based on a Contextualized Language Model, it is essential to have the complete left and right contexts for each occurrence. Besides, another advantage of the PRESTO corpus is that it allows to explore a different language than English, and French has not yet received a lot of attention in the NLP research community working on Lexical Semantic Change. Investigating laws of semantic change in various languages ensures more robust validations of these laws.

Diachronic Concept Induction allows us to go farther than the mere quantification of both phenomena, thus we will test two hypotheses. First, we believe that Reallocation and Parallel Change can apply simultaneously to the same concept naming, as it appears to have been the case for *RAPIDLY* with *fast*, *swiftly* and *quickly*. Second, if Reallocation is the outcome of a competition between words for a given meaning, as suggested by Bréal's law, Clark's principle of Conventionality and other advocates of the 'No-synonymy' hypothesis, then it might be more frequent in larger synonym sets than in smaller one.

7.4.1 Reallocation and Parallel Change in Concept Clusters

Methodology In this new experiment, instead of starting from a set of identified synonymous words provided by an expert in lexical resource, we will begin with a set of concepts that we have empirically identified from a set of associated lexemes. Let Ψ be the set of concept clusters found by a Diachronic Concept Induction system. We denote as $\Psi_{syn}^{(1)}$ the subset of clusters containing occurrences

of at least 2 distinct words at time period T_1 and that are still instantiated in later time periods. $\Psi_{syn}^{(1)}$ is the set of latent concepts that exhibit synonymy in the earliest time period, and is formally defined as follows (reusing notations from Chapter 6):

$$\Psi_{syn}^{(1)} = \left\{ \psi_k^{(*)} \in \Psi^{(*)} \mid |N_k^{(1)}| \geq 2 \text{ and } N_k^{(P)} \neq \emptyset \right\} \quad (7.2)$$

We exclude concepts that disappeared over time as they are not directly relevant to the study of semantic change in synonymy. Our goal is to quantify the amount of Reallocation and of Parallel Change observed in these clusters.

Identifying Reallocation. Reallocation simply corresponds to concept clusters whose namings (i.e. the set of words whose occurrences can be found in the cluster) have shrunk in the more recent time periods compared to the earlier ones. In comparison to Reallocation as illustrated in Figure 7.1, both Differentiation and Disappearance are captured by this definition: using an onomasiological approach, we simply have to detect whether words that were in the naming at T_1 are now missing at T_P . Distinguishing between Differentiation and Disappearance amounts to checking whether a word missing from the original concept's naming is still found in other clusters or has completely disappeared.

Identifying Parallel Change. Parallel Change is less straightforward to identify. As a rule, we say that we detect a case of Parallel Change for a concept cluster $\psi_k^{(*)}$ with naming $N_k^{(1)} = \{w_1, w_2, \dots\}$ at the first period T_1 if we find any two words $w_i, w_j \in N_k^{(1)}$ such that they both appear in another concept cluster $\psi_l^{(*)}, l \neq k$ at time T_P and that at least one of the two words is new in the naming of this concept cluster. Formally, we have the following decision rule:

$$\left\{ \begin{array}{ll} \text{Parallel Change if } \exists w_i \in N_k^{(1)}, & \overbrace{\begin{array}{l} \exists \psi_l \in \Psi, l \neq k, w_i \notin N_l^{(1)}, w_i \in N_l^{(P)}, \\ \exists w_j \in N_k^{(1)}, i \neq j, w_j \in N_l^{(P)} \end{array}}^{\text{"}w_i \text{ has developped a new meaning"}} \\ \text{No parallel change otherwise.} & \underbrace{\hspace{10em}}_{\text{"that one of its synonyms also has at } T_P"} \end{array} \right. \quad (7.3)$$

Because we are working with concept clusters rather than a simple binary indicator of synonymy, we are able to distinguish between two cases: (i) words have remained synonym simply because the set of their shared meanings hasn't changed and (ii) words that have remained synonymous because they have undergone Parallel Change. This represents an improvement over the study presented in

Section 7.3. Moreover, our definition covers all cases of Parallel Change illustrated in Figure 7.2, including cases where only one term gains a new meaning, as in the pair *viper/snake*.

Thanks to the amount of information provided by Diachronic Concept Induction and thanks to a more onomasiological, cluster-centric approach to the definitions of these two phenomena, we are able to overcome the limitation of scope faced in Section 7.3.

Moreover, we are now able to quantify Differentiation and Parallel Change within larger groups of synonyms and not only pairs. For simplicity, we adopt the less restrictive definitions. We say that Reallocation has occurred in a cluster ψ_k if at least one word in ψ_k has been removed from the concept's naming; and we say that Parallel Change has occurred if at least of one word in ψ_k has developed a new meaning shared with at least one of its initial synonyms.

7.4.2 Experiments

To conduct our experiments, we reuse the same diachronic corpus (PRESTO, a collection of French texts from 1500 to 1949) and Diachronic Concept Induction system as in Chapter 6. We consider the two stages compared in the same chapter, namely Middle & Preclassical French and Modern French, respectively corresponding to 1500-1649 and to 1800-1949. We now explore the obtained results to identify and characterize semantic changes in groups of synonyms following the definitions discussed in the previous section.

Hypotheses about changes in synonyms. Beyond quantifying Reallocation and Parallel Change, our goal is to verify two additional hypotheses. First, we want to test that both phenomena can be simultaneously observed for the same cluster, showing that they are not mutually exclusive. Second, we want to test whether Reallocation applies more systematically when there is a higher number of synonyms in competition.

Applying Diachronic Concept Induction to the PRESTO data, we find 1,099 concept clusters containing occurrences of multiple words in Middle & Preclassical French $T_1 = [1500; 1649]$ (after cluster filtering as described in Chapter 6). These 1,099 concept clusters constitute our initial set $\Psi_{syn}^{(1)}$.

Identified Reallocations. Table 7.4 displays the counts (in number of clusters) per POS tag for each of the three possible evolutions: Reallocation, Parallel Change and Both. Among the 1,099 concept clusters found in $\Psi_{syn}^{(1)}$, 269 (24%)

	All $\Psi_{syn}^{(1)}$	Reallocation (incl. ‘Both’)	Parallel Change (incl. ‘Both’)	Both
Total	1099	269 (24%)	77 (7.1%)	20 (1.8%)
Nouns	548	130 (24%)	19 (3.5%)	3 (0.5%)
Verbs	355	105 (30%)	47 (13%)	14 (3.9%)
Adjectives	196	34 (17%)	11 (0.6%)	3 (1.5%)

Table 7.4: Cluster counts of synonym changes from the PRESTO data. Percentages per row in parentheses.

have shrunk in naming between T_1 and T_P , indicating Reallocation. As an example from our concept clusters, we find that the cluster containing $\{\text{déclarer}, \text{dénoncer}\}$ (‘declare’ and ‘denounce’ respectively) is later reduced $\{\text{déclarer}\}$. The entry for *dénoncer* in the TLFi contains the meaning of DECLARE PUBLICLY with the label ‘archaic’, whereas it is the main sense of *déclarer* and could be the meaning corresponding to this cluster (most frequent context words include *guerre* ‘war’, *roi* ‘king’, *paix* ‘peace’, *loi* ‘law’). Another example is the cluster whose naming contains $\{\text{faveur}, \text{grâce}, \text{merci}\}$ (respectively ‘favor’, ‘grace’ and ‘mercy’). Inspecting the most frequent context words, the cluster appears to correspond to the meaning FAVOR GRANTED TO SOMEONE, and the cluster’s naming has shrunk down to $\{\text{faveur}, \text{grace}\}$. According to the TLFi, *merci* is no longer used in this meaning, labeled as ‘archaic’, and the word now either means CONDITION OF COMPLETE SUBMISSION as in *à la merci de* (‘at the mercy of’), or corresponds to the interjection ‘thank you’. However, as we have seen in the previous chapter, errors in the identified shrinkages may occur. For instance, the cluster whose naming first contains the verbs $\{\text{avertir}, \text{aviser}\}$ and then only $\{\text{avertir}\}$ is labeled as a case of Reallocation, but the inferred common meaning (WARN, MAKE AWARE OF) is still an active sense for both words.

Identified Parallel Changes. 77 (7.1% of the 1,099) concept clusters are sources of Parallel Change. Interestingly, there is a larger proportion of Parallel Change in verbs than in other parts of speech. Let us for instance take the example of the noun concept cluster with naming $\{\text{aqueduc}, \text{canal}, \text{étang}\}$ in Middle & Preclassical French, and with most frequent context words being water-related terms (e.g. *eau* ‘water’, *rivière* ‘river’, *poisson* ‘fish’, ...). We understand that the synonymy between the words *aqueduc* ‘aqueduc’ and *canal* ‘canal’ concerns the concept of ARTIFICIAL STRUCTURE TO CONVEY WATER. In Modern French, both words are found in another cluster related to anatomy, alongside *tube* ‘tube’, *tuyau* ‘tuyau’ and *vaisseau* ‘vessel’. As the most frequent contexts include *lymphatique* ‘lymphatic’, *sang* ‘blood’, *nerf* ‘nerve’, *intestinal* ‘intestinal’, we under-

stand that this cluster correspond to the meaning of PIPE OR CANAL IN WHICH A BODY FLUID IS CONTAINED. The words *aqueduc* and *canal* indeed simultaneously gained this sense with the progresses of health sciences and anatomical descriptions between Middle & Preclassical French and Modern French.

Looking more closely at the apparent mistakes produced by the system, we identify two types of errors: (i) cases where the destination cluster correspond to a meaning that the synonyms already had in common before (e.g. *réveiller* ‘wake up’ and *éveiller* ‘awaken’ for which the identified parallel change would be RE-VIVE A FEELING OR A THOUGHT > CEASE SLEEPING, despite these two meanings being attested in the TLFi for the two words even before 1500); and (ii) cases where we cannot distinguish the semantic difference between the source and destination concept clusters (e.g. a parallel change has been identified between *cause* ‘cause’ and *motif* ‘motive’ but we could not found a semantic difference between the source and destination clusters).

In our previous computational study, we found both in the dataset and in model predictions more cases of ‘Diff’ than of ‘Syn’, indicating that Differentiation was a privileged path of evolution for synonyms. Similarly, we find in this new framework that Reallocation is more frequent than Parallel Change. However, we find that most synonyms groups (773 out of 1,099, i.e. 70.3%) underwent neither of the two phenomena. This is an empirical evidence contradictory to the idea that synonyms *should* differentiate or disappear, as Bréal initially stated. In the on-going debate about whether language systems lean towards a nearly-isomorphic structure (with a one-to-one word-concept mapping) or if they tend to exhibit and preserve advantageous cases of semantic optionality, our results appear to be in line with the latter hypothesis rather than the former. However, this could be an artifact of the considered time periods: examples of Reallocation given by Bréal sometimes involve changes spanning between a language and one of its descendants, and maybe a time frame of 4.5 centuries is not large enough to see Reallocation apply systematically.

Regarding our first research question, on whether Reallocation and Parallel Change can apply simultaneously, we find that in 20 concept clusters (1.8%) we can identify that a Parallel Change has happened among synonyms while the concept’s naming has shrunk simultaneously, indicating that Reallocation also took place. This is obviously a very marginal phenomenon and we would not be able to draw a strong conclusion about it, but given this result we must not reject the idea that the two semantic evolutions could indeed happen in the same synonym group and that the two processes are not necessarily contradictory.

Table 7.5 displays the average size of concept cluster namings (in number of

	All $\Psi_{syn}^{(1)}$	Reallocation	Parallel Change
Total	2.90	2.61	3.47
Nouns	2.83	2.47	3.05
Verbs	2.77	2.74	3.79
Adjectives	3.33	2.74	2.8

Table 7.5: Average number of lexemes per concept-cluster at time T_1 w.r.t. the type of synonym change from the PRESTO data.

lexemes) at time T_1 per POS tag with respect to the evolution path. We find that the average number of lexemes at T_1 in clusters in which Reallocation takes place is of 2.61 on average, which is below the overall average of 2.90 and even below the average for Parallel Change of 3.47. The difference between Reallocation and all synonym groups is weaker in Verbs but the contrast with Parallel Change is stronger; in adjectives, the average size of cluster naming that underwent Parallel Change is below the overall average but still higher than that of cluster that underwent Reallocation. Moreover, looking more closely at the proportion of Reallocation with respect to the initial number of lexemes in the naming, we find 25.4% of Reallocation among the 928 clusters with 2 or 3 lexemes only at T_1 ; this proportion decreases to 22.7% for the 110 clusters of 4 or 5 lexemes, and drops to 13.1% for the 61 clusters of 6 lexemes or more. It appears that we cannot conclude that Reallocation would be more likely in larger synonym groups, thus we must reject this hypothesis based on our numerical results. Our results even suggest the contrary, i.e. that more Reallocation happens in small synonym groups than in large ones.

Summary of the findings. Our results show a larger proportion of Reallocation than of Parallel Change, but that these two phenomena appear marginal compared to the majority of concept clusters whose namings remain unchanged. We have identified cases of concept cluster where *both* Reallocation and Parallel Change happen, but were unable to conclude that Reallocation was more likely to take place in larger synonym groups. Overall, through our experiments using Diachronic Concept Induction on the PRESTO data and comparing concept clusters at time period $T_1 = [1500; 1649]$ and time $T_P = [1800; 1949]$, we have shown that the proposed method was able to (i) quantify the number of cases of Reallocation and Parallel Change in order to assess which of the two was more common in a more comprehensive way than previous studies; (ii) assess whether synonym groups tend to undergo one of these phenomena or if they rather remain stable over time; (iii) test two hypotheses that we made regarding Reallocation and Parallel Change, and gaining insights for future studies.

Limitations. In this section, we have not conducted an evaluation of the proposed method’s accuracy in detecting Reallocation and Parallel Change. We must indeed face the same challenge as in Section 7.3, in that no available resource can directly provide a ground truth for Reallocation and Parallel Change. We did not want to cross synonym lists from different periods either, as we have seen in the previous section that this could induce bias in the analyses. In Chapter 6, however, we have conducted a thorough qualitative evaluation of the clusters obtained by our method in the PRESTO data, showing that observed changes often correspond to attested changes. This evaluation has shed the light on two limitations that are at play in the investigation of Reallocation and Parallel Change.

We have observed in Chapters 4 and 6 that the partition into concept clusters drawn by (Diachronic) Concept Induction might be different than the partitioning in concepts from an expert resource, and that the found concept clusters may include near-synonyms or in some cases semantically related words that are not synonyms.

Furthermore, diachronic evaluation results have shown that on the PRESTO data, the proposed bilevel clustering model struggled to accurately identify actual naming shrinkages, especially compared to other kinds of changes like semasiological gains. As a result, our comparison between Reallocation (which we have defined as naming shrinkages) and Parallel Changes (which is based on naming expansion and/or semasiological gains) may be biased.

Given these limitations, our findings cannot be taken as strong evidences about which kind of change is the dominant one among synonyms. The goal of this section was rather to showcase how a Diachronic Concept Induction model could be employed to address such questions, and make a proof-of-concept of a research methodology that we believe opens more doors than previous ones. Future research should focus on improving the clustering method to obtain better concept clusters and better identification of changes (and in particular cluster naming shrinkages) to allow for a better and more robust investigation of these two laws of lexical semantic change in synonyms.

7.5 Summary

In this chapter, we have explored the case study of synonyms and how the change in meaning over time. Building upon Xu and Kemp (2015)’s study, we have discussed the law of Reallocation, postulating that synonyms cannot stay synonyms for long and that the mapping between words and concepts tend toward a more

isomorphic structure (with only one lexeme for each concept); and the law of Parallel Change, that posits that synonyms tend to share their meanings and develop other similar meanings.

To measure a computational model's ability to represent these changes, we have proposed an evaluation framework based on the conjunction of two lexical resources from two different time periods, to obtain a ground-truth on whether considered pairs of synonyms remain synonymous or not. Inspired by the prior study, we have proposed several computational models using static representations to predict the observed changes. Experimental results underline the importance of evaluating approaches and have shown the limitations of the proposed models. Supervised approaches are only partially predictive of the observed changes, suggesting that static representations may not be powerful enough to be represent semantic changes in the studied setting. Notably, further examination has highlighted that polysemy was a key factor, as words can have multiple meanings and synonyms may share only a portion of their meanings; and the used static representations do not account for such variety of usages. Moreover, it proved challenging to combine the use of two different lexical resources, compiled at different times and with different objectives, where the very concept of synonymy varies according to the resource. The establishment of a rigorous evaluation framework appears to be all the more challenging in the absence of an historical onomasiological reference to obtain lists of synonyms across time periods.

In a second part of the chapter, we have discussed the advantages of using Diachronic Concept Induction to tackle this research question. First, it provides groups of synonyms over time, with a comparable notion of synonymy across all periods. Second, it naturally provides a different perspective, that of concepts, and allows to approach the problem under a new, onomasiological angle. It allows to redefine the targeted phenomena in a more comprehensive way, and we were able to test several hypotheses about the condition in which reallocation and parallel change happen and their alleged oppositeness. Experimental results in a different language than English (namely French) suggest that both phenomena may not be mutually exclusive and can take place simultaneously on a given concept's naming. We could not find evidences in the data that a higher number of competing synonyms in the cluster would facilitate Reallocation, thus invalidating our second hypothesis. Interestingly, both experimental studies have concluded in favor of Reallocation/Differentiation over Parallel Change, in contrast with the results of Xu and Kemp (2015) in two different languages, but our second study on PRESTO has shown that a vast majority of synonym sets are

neither source of Parallel Change nor subject to Reallocation.

Each of the proposed approaches allows to tackle different aspects of the relation between Reallocation and Parallel Change, making different assumptions and having their own limitations. The method proposed by Xu and Kemp (2015) addresses the opposition between Differentiation (a subcase of Reallocation) and Parallel Change, and relies on the assumptions that none of the two words has ceased to be used and that the two words of the studied pairs have undergone change in meaning. The study we conducted in Section 7.3 investigates the opposition between Differentiation and the persistence of synonymy (conflating cases of Parallel Change with cases where there is no change at all in their synonymy relation) and also focuses on synonym pairs where none of the two words has ceased to be used. In contrast, with the approach presented in Section 7.4 using Diachronic Concept Differentiation, we were able to work with more comprehensive definitions of these phenomena, distinguishing between Reallocation (by Differentiation or by Disappearance), Parallel Change (regardless of how many words have changed in meaning) and unchanged synonymy. This method allows to work with synonym groups instead of simple pairs, and by isolating the different meanings of a word, this method is able to treat its different groups of synonyms independently. However, as we have seen in the previous chapter, Diachronic Concept Induction makes the hypothesis that semantic concepts (the common meanings within synonym groups) are stable over time and that change only affects the set of words referring to them. The other two approaches did not require such assumption and could theoretically handle potential concept drifts, as they only focus on relative distances between two synonyms. Nonetheless, this concept-oriented clustering approach using contextualized embeddings widens the scope of computational studies of semantic change in synonyms, and allows for a deeper analysis of the different phenomena.

Overall, through this case study on synonymy, we have demonstrated that computational approaches based on embeddings can indeed be used to address a complex onomasiological question, that of the change among members of a same concept naming. We have also seen that the use of contextualized embeddings and (diachronic) concept induction offers a new, broader perspective on these studies and provides new insights on the data that can inform the scientific discussion on the nature and regularities of semantic change.

Part IV

Conclusions and Appendices

8

Conclusions and Perspectives

8.1 Summary of the Thesis

In this thesis, we have addressed a crucial limitation of current computational models of lexical semantic change: they rely exclusively the semasiological perspective and neglect the onomasiological dimension of semantic change. But the mapping between words and meanings is two-sided, and historical linguistics has relied on both the apparatus of semasiology and that of onomasiology to describe classifications, patterns and regularities of lexical semantic change. Our approach is structured in two steps. First, we explore ways of representing and extracting semantic concepts from data in the synchronic setting with Natural Language Processing (NLP) models. In particular, we use *Contextualized Language Models*, leveraging their ability to compute representations for each occurrence of a word in a corpus to better account for potential polysemy. We then extend the proposed methodology to the diachronic setting as to study, with the tools provided by this method, the semasiological and onomasiological changes over time.

In Chapter 4, we introduced *Concept Induction*, a task that directly models the mapping between words and semantic concepts in data. We proposed a bi-level clustering method, and compared it through a series of experiment to other, more traditional approaches e.g. Word Sense Induction methods. Our results indicated that hierarchical clustering and the proposed bi-level approach outperform alternative techniques, highlighting the multi-level nature of both the problem itself and the spatial organization of semantic information in Contextualized Language Models. Additional results suggested that the global, cross-lexicon aspect of Concept-Induction is also directly beneficial at the local, word-centric level; a concept-inducing system can be a valuable tool for sense-related tasks such as Word Sense Induction or Word-in-Context, especially in settings where there are only few occurrences per target word.

The question of how concepts emerge from data can be approached not only through clustering, but also through the lens of representations. In Chapter 5, we explored the fine-tuning of contextualized language models in order to obtain representations that better encode lexical meaning and similarities/differences in semantic concepts between occurrences. We proposed a method to fine-tune a Contextualized Language Model for a new objective, obtaining models we refer to as *CALE*. Empirical evaluations on several lexical semantic tasks showed that the obtained representations better reflect similarities and differences in lexical meaning, both with occurrences of the same word and among occurrences of different lexemes.

Building upon these two contributions at the synchronic level, Chapter 6 aimed at extending the bi-level methodology to the diachronic setting. We adapted the method to account for temporal variation in historical data, and described how the results produced by *Diachronic Concept Induction* can be used to jointly identify semasiological and onomasiological changes. A manual evaluation on historical data of our proof-of-concept method showed that the results were promising, although we noted that semasiological sense losses and onomasiological naming shrinkages were less accurately identified than other types of changes. We hypothesize that this limitation is related to the high number of very small clusters produced by the method.

Finally, Chapter 7 provided a case study showing the full potential of the Diachronic Concept Induction approach. This study concerned two apparently contradicting phenomena of lexical semantic change regarding the temporal evolution of synonymy, namely the *Law of Reallocation* and *Parallel Change*. We employed two approaches, one based on crossing lists of synonyms from different periods and using static representations, and another based on synonym groups identified by Diachronic Concept Induction using contextualized representations from a CALE model. We discussed the limitations of the two approaches, and showed that the tools provided by Diachronic Concept Induction allow for a more comprehensive description of the two phenomena than other existing frameworks.

Overall, with Diachronic Concept Induction, we have proposed a large-scale method that can assist research in historical linguistics, by extracting semantic concepts from historical data without supervision or external knowledge. The method can simultaneously uncover cases of semasiological changes and of onomasiological ones. Our approach goes beyond the mere detection of semantic change and provide a first characterization, distinguishing between sense gains/sense losses and naming expansions/naming shrinkages. For each uncov-

ered change, it identifies the clusters of occurrences responsible that contributed to the change detection, enabling a manual investigation of said occurrence clusters and qualitative analyses by expert historical linguists. This characterization allows to challenge hypotheses about lexical semantic change in complex scenarios, and opens the path to formulate new hypotheses in the future and uncover new patterns of semantic change.

8.2 Future Work and Perspectives

8.2.1 Clustering Improvements

Hierarchical clustering algorithms and constraints integration. Our results in Chapter 4 suggest that hierarchical algorithms are better suited for the task of Concept Induction than flat clustering approaches. On the one hand, Agglomerative Clustering performed better than KMeans; on the other hand, the bi-level method reached higher precision than direct clusterings. This particular finding underscores the multi-level nature of the problem, as well as its multi-resolution aspect (i.e. clusters may not be evenly distributed across the lexemes, and may vary in number of occurrences and density). Other hierarchical clustering algorithms could thus be explored to improve our approach to Concept Induction. In Chapter 6, we made partial use of HDBSCAN’s process (Campello et al., 2013) to select proto-sense clusters; but we could make a complete use of this density-based hierarchical clustering algorithm, which seems particularly fitting for our problem. Another promising direction for Concept Induction is constrained hierarchical clustering, as in Bair (2013); Chatziafratis et al. (2018); Mauduit and Simonetto (2023) and Zeng et al. (2025), which integrates constraints (often in the form of ‘must-link’ and ‘cannot-link’ associations between instances) in the clustering process. This type of approaches aligns closely with the structured constraints at work between occurrences, lexemes, senses and concepts that we have established in Chapter 4; and future work could attempt to jointly optimize both the local and the global clustering objectives.

Cost functions and priors in hyperparameter selection. In Chapter 6, we selected hyperparameters by choosing the parameter set that maximized the resulting number of concept clusters containing between 2 and n lexemes (with $n = 5$ in the experiments). This pragmatic choice ensured a sufficient recovery of synonymy and therefore of shared concepts, though possibly at the expense of a more accurate recovery of polysemy. Recent research on hierarchical clustering has proposed several objective functions designed to quantify the cost of

a given hierarchical clustering tree, verifying that a tree that minimizes such cost functions fulfills certain desired properties (Dasgupta, 2016; Cohen-Addad et al., 2019; Wang and Wang, 2020; Roch, 2023). Given the hierarchical tree produced by a concept-inducing system (in the bilevel setting, such tree could be obtained by merging the local trees of proto-senses clustering with the global tree learned at the cross-lexicon step) with a given set of hyperparameters, a cost function could be used to determine the most suitable set of parameters for a given dataset. We could even integrate some *prior knowledge* to these cost functions to ensure that the selected tree fulfills some desired properties (e.g. the distribution of sense inventory sizes might be required to approximate Zipf (1949)’s law).

Diachronic Concept Induction without fixed concepts. In Chapter 6, when introducing Diachronic Concept Induction, we have decided to learn concept clusters for all periods at once, thus assuming that there was a set of stable concepts shared across all time periods. Different alternatives could be explored. In the context of semasiological-only approaches to Lexical Semantic Change, Periti and Tahmasebi (2024b) in particular proposed several methods, including learning clusters independently in each period, or using a sliding window of periods. The former implies to link clusters across time periods *a posteriori*, while the latter approaches facilitates the tracking of clusters over time because two consecutive time windows share a part of the occurrences. Periti et al. (2022) adapted the usual sense clustering pipeline to account for multiple time periods using an evolutionary approach, such that recent clusters are derived from older ones. Adapting these approaches to the onomasiological approach of Diachronic Concept Induction would allow us to dispense with the ‘fixed concepts assumption’, thereby accommodating concept drifts. It could improve the detection of lexical semantic change, particularly naming shrinkages and sense losses.

8.2.2 Evaluation and Interpretability

Systematic evaluation for onomasiological change with synthetic data. In Chapter 6, our evaluation of the Diachronic Concept Induction approach was inherently limited by the absence of concept-annotated historical data preventing us to conduct a systematic quantitative evaluation of the method as in Chapter 4. Building a large-scale dataset with annotations from human experts to systematically evaluate the recovery of semasiological and onomasiological changes is appealing, but it would be a long and expensive endeavour. Recently, Cassotti and Tahmasebi (2025) have proposed a generative model that produces time-

specific and sense-specific synthetic occurrences of target words (generating for instance an occurrence of *plane* in the sense of AIRCRAFT as it would be produced in 2010). Combining such a model with a historical dictionary like the TLFi or the Oxford English Dictionary would allow the creation a synthetic benchmark for both semasiological and onomasiological change. For onomasiological changes in particular, if the historical resource does not specify whether a given word has synonyms for a given sense in a given period, we could *simulate* onomasiological changes by generating occurrences of several target lexemes using the same sense definition.

Interpreting Clusters with Feature Embeddings. To interpret the semantic content of concept clusters in Chapter 6 and 7, we have relied on the lexemes contained in each cluster and their most frequent context words. This leaves room for subjective appreciation, especially since these descriptors may be insufficiently informative. In the context of semasiological studies of Lexical Semantic Change, several recent studies employ definition-generation systems to characterize cluster meanings (Giulianelli et al., 2023; Kutuzov et al., 2024; Fedorova et al., 2024a). Generating a definition for each cluster and comparing it to the senses in the reference dictionary would allow for a more systematic and robust evaluation. Adopting a different approach (conceived in the synchronic setting), Chronis et al. (2023) project contextualized embeddings onto psycholinguistically interpretable semantic spaces where each dimension corresponds to a pre-defined feature (e.g. fine-grained features like ‘have legs’ or ‘found in the kitchen’ or coarse-grained features like ‘bright’, ‘taste’ or ‘benefit’), resulting in a spatial organization that is easily interpretable for human users. Projecting our concept clusters onto these interpretable spaces would yield more complete descriptions of lexical semantic changes and would allow to test other hypotheses, for instance investigating whether Reallocation and Parallel Change in synonyms tend to originate in different semantic fields (e.g. maybe one of them is more frequent among words related to abstract things or sensory perception ?).

Diachronic Sense Prototypes. In Chapter 6 we briefly discussed a statistical law of semantic change, the Law of Innovation proposed by Hamilton et al. (2016b). In Chapter 7, we have only tested two laws about semantic change in synonyms, Reallocation and Parallel Change. Other laws and theories about Lexical Semantic Change could be empirically tested with the methods proposed in this thesis. Diachronic Concept Induction could for instance be used to investigate Geeraerts (1997)’s Diachronic Prototype Theory. Assuming that the sense inventory of a word forms a radial set of very prototypical and more pe-

ripheral meanings, this theory states (i) that semasiological changes are more likely to occur in the peripheries than at the core of the set; (ii) that new senses appear as gradual modulations of the existing core senses; and (iii) and that senses may overlap and combine, giving rise to new meanings through interaction between existing meanings. Using Diachronic Concept Induction, we could identify concept clusters and the set of senses for a given word, and then use the distances between them to reveal the structure of the radial set, and its evolution over time. Combined with generated definitions and/or interpretable features, Diachronic Concept Induction could not only test predictions of the Diachronic Prototype Theory, but also examine whether analogous effects can be observed on the onomasiological side, thus potentially broadening the scope of the theory.

Bibliography

Eneko Agirre, Enrique Alfonseca, Keith Hall, Jana Kravalova, Marius Paşca, and Aitor Soroa. A study on similarity and relatedness using distributional and WordNet-based approaches. In Mari Ostendorf, Michael Collins, Shri Narayanan, Douglas W. Oard, and Lucy Vanderwende, editors, *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 19–27, Boulder, Colorado, June 2009. Association for Computational Linguistics. URL <https://aclanthology.org/N09-1003/>.

Reem Alatrash, Dominik Schlechtweg, Jonas Kuhn, and Sabine Schulte im Walde. CCOHA: Clean corpus of historical American English. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6958–6966, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL <https://aclanthology.org/2020.lrec-1.859/>.

Enrique Amigó, Julio Gonzalo, Javier Artiles, and Felisa Verdejo. A comparison of extrinsic clustering evaluation metrics based on formal constraints. *Information retrieval*, 12:461–486, 2009.

Asaf Amrami and Yoav Goldberg. Towards better substitution-based word sense induction, 2019. URL <https://arxiv.org/abs/1905.12598>. arXiv preprint.

Nikolay Arefyev and Vasily Zhikov. BOS at SemEval-2020 task 1: Word sense induction via lexical substitution for lexical semantic change detection. In Aurelie Herbelot, Xiaodan Zhu, Alexis Palmer, Nathan Schneider, Jonathan May, and Ekaterina Shutova, editors, *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 171–179, Barcelona (online), December 2020. International Committee for Computational Linguistics. doi: 10.18653/v1/2020.semeval-1.20. URL <https://aclanthology.org/2020.semeval-1.20/>.

Carlos Santos Armendariz, Matthew Purver, Matej Ulčar, Senja Pollak, Nikola Ljubešić, and Mark Granroth-Wilding. CoSimLex: A resource for evaluating

- graded word similarity in context. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5878–5886, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL <https://aclanthology.org/2020.lrec-1.720/>.
- Naomi Baes, Nick Haslam, and Ekaterina Vylomova. Semantic shifts in mental health-related concepts. In Nina Tahmasebi, Syrielle Montariol, Haim Dubossarsky, Andrey Kutuzov, Simon Hengchen, David Alfter, Francesco Periti, and Pierluigi Cassotti, editors, *Proceedings of the 4th Workshop on Computational Approaches to Historical Language Change*, pages 119–128, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.lchange-1.13. URL <https://aclanthology.org/2023.lchange-1.13/>.
- Amit Bagga and Breck Baldwin. Entity-based cross-document coreferencing using the vector space model. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*, pages 79–85, Montreal, Quebec, Canada, August 1998. Association for Computational Linguistics. doi: 10.3115/980845.980859. URL <https://aclanthology.org/P98-1012/>.
- He Bai, Tong Wang, Alessandro Sordoni, and Peng Shi. Better language model with hypernym class prediction. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1352–1362, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.96. URL <https://aclanthology.org/2022.acl-long.96/>.
- Eric Bair. Semi-supervised clustering methods. *WIREs Computational Statistics*, 5(5):349–361, 2013. doi: <https://doi.org/10.1002/wics.1270>. URL <https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/wics.1270>.
- Gioia Baldissin, Dominik Schlechtweg, and Sabine Schulte im Walde. DiaWUG: A dataset for diatopic lexical semantic variation in Spanish. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the*

- Thirteenth Language Resources and Evaluation Conference*, pages 2601–2609, Marseille, France, June 2022. European Language Resources Association. URL <https://aclanthology.org/2022.lrec-1.278/>.
- Robert Bamler and Stephan Mandt. Dynamic word embeddings. In *International conference on Machine learning*, pages 380–389. PMLR, 2017.
- Pierpaolo Basile, Annalina Caputo, and Giovanni Semeraro. Analysing word meaning over time by exploiting temporal random indexing. In *Proceedings of the First Italian Conference on Computational Linguistics CLiC-it 2014 & and of the Fourth International Workshop EVALITA 2014: 9-11 December 2014, Pisa*, pages 38–42. Pisa University Press, 2014.
- Christin Beck. DiaSense at SemEval-2020 task 1: Modeling sense change via pre-trained BERT embeddings. In Aurelie Herbelot, Xiaodan Zhu, Alexis Palmer, Nathan Schneider, Jonathan May, and Ekaterina Shutova, editors, *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 50–58, Barcelona (online), December 2020. International Committee for Computational Linguistics. doi: 10.18653/v1/2020.semeval-1.4. URL <https://aclanthology.org/2020.semeval-1.4/>.
- James Bergstra, Rémi Bardenet, Yoshua Bengio, and Balázs Kégl. Algorithms for hyper-parameter optimization. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011. URL https://proceedings.neurips.cc/paper_files/paper/2011/file/86e8f7ab32cfd12577bc2619bc635690-Paper.pdf.
- Michele Bevilacqua, Tommaso Pasini, Alessandro Raganato, and Roberto Navigli. Recent trends in word sense disambiguation: A survey. In *International joint conference on artificial intelligence*, pages 4330–4338. International Joint Conference on Artificial Intelligence, Inc, 2021.
- Andreas Blank. *Prinzipien des lexikalischen Bedeutungswandels am Beispiel der romanischen Sprachen*. Max Niemeyer Verlag, Berlin, Boston, 1997. ISBN 9783110931600. doi: 10.1515/9783110931600. URL <https://doi.org/10.1515/9783110931600>.
- Andreas Blank. Why do new meanings occur? a cognitive typology of the motivations for lexical semantic change andreas blank. *Cognitive Linguistics Research*, page 61, 1999.

- Terra Blevins, Mandar Joshi, and Luke Zettlemoyer. FEWS: Large-scale, low-shot word sense disambiguation with the dictionary. In Paola Merlo, Jorg Tiedemann, and Reut Tsarfaty, editors, *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 455–465, Online, April 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.36. URL <https://aclanthology.org/2021.eacl-main.36/>.
- Peter Blumenthal, Sascha Diwersy, Achille Falaise, Marie-Hélène Lay, Gilles Sourvay, and Denis Vigier. Presto, un corpus diachronique pour le français des xvie-xxe siècles. In *Actes de la 24eme conférence sur le Traitement Automatique des Langues Naturelles-TALN*, volume 17, 2017.
- Richard A. Blythe and William Croft. S-curves and the mechanisms of propagation in language change. *Language*, 88(2):269–304, 2012. ISSN 00978507, 15350665. URL <http://www.jstor.org/stable/23251832>.
- Max Boholm, Björn Rönnerstrand, Ellen Breitholtz, Robin Cooper, Elina Lindgren, Gregor Rettenegger, and Asad Sayeed. Can political dogwhistles be predicted by distributional methods for analysis of lexical semantic change? In Nina Tahmasebi, Syrielle Montariol, Andrey Kutuzov, David Alfter, Francesco Periti, Pierluigi Cassotti, and Netta Huebscher, editors, *Proceedings of the 5th Workshop on Computational Approaches to Historical Language Change*, pages 144–157, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.lchange-1.14. URL <https://aclanthology.org/2024.lchange-1.14/>.
- Michel Bréal. *Essai de Sémantique*. Paris: Hachette, 1897.
- Carl Darling Buck. *A dictionary of selected synonyms in the principal Indo-European languages*. University of Chicago Press, 1949.
- John A Bullinaria and Joseph P Levy. Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior research methods*, 39(3):510–526, 2007.
- John A Bullinaria and Joseph P Levy. Extracting semantic representations from word co-occurrence statistics: stop-lists, stemming, and svd. *Behavior research methods*, 44(3):890–907, 2012.
- Joan Bybee. Cognitive processes in grammaticalization. *The New Psychology of Language: Cognitive and Functional Approaches To Language Structure, Volume II*, page 145, 2003.

- Jose Camacho-Collados, Claudio Delli Bovi, Luis Espinosa-Anke, Sergio Oramas, Tommaso Pasini, Enrico Santus, Vered Shwartz, Roberto Navigli, and Horacio Saggion. SemEval-2018 task 9: Hypernym discovery. In Marianna Apidianaki, Saif M. Mohammad, Jonathan May, Ekaterina Shutova, Steven Bethard, and Marine Carpuat, editors, *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 712–724, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/S18-1115. URL <https://aclanthology.org/S18-1115/>.
- Lyle Campbell. *Historical linguistics*. Edinburgh University Press, 2013.
- Ricardo J. G. B. Campello, Davoud Moulavi, and Joerg Sander. Density-based clustering based on hierarchical density estimates. In Jian Pei, Vincent S. Tseng, Longbing Cao, Hiroshi Motoda, and Guandong Xu, editors, *Advances in Knowledge Discovery and Data Mining*, pages 160–172, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg.
- Annalina Caputo, Pierpaolo Basile, and Giovanni Semeraro. Temporal random indexing: A system for analysing word meaning over time. *IJCoL*, 1(1):61–74, December 2015.
- Pierluigi Cassotti and Nina Tahmasebi. Sense-specific historical word usage generation. *Transactions of the Association for Computational Linguistics*, 13:690–708, 2025. doi: 10.1162/tacl_a_00761. URL <https://aclanthology.org/2025.tacl-1.32/>.
- Pierluigi Cassotti, Lucia Siciliani, Marco DeGemmis, Giovanni Semeraro, and Pierpaolo Basile. XL-LEXEME: WiC pretrained model for cross-lingual Lexical sEMantic changE. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1577–1585, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-short.135. URL <https://aclanthology.org/2023.acl-short.135/>.
- Pierluigi Cassotti, Pierpaolo Basile, and Nina Tahmasebi. DWUGs-IT: Extending and standardizing lexical semantic change detection for Italian. In Felice Dell’Orletta, Alessandro Lenci, Simonetta Montemagni, and Rachele Sprugnoli, editors, *Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024)*, pages 190–197, Pisa, Italy, December 2024a. CEUR Workshop Proceedings. ISBN 979-12-210-7060-6. URL <https://aclanthology.org/2024.clicit-1.23/>.

- Pierluigi Cassotti, Stefano De Pascale, and Nina Tahmasebi. Using synchronic definitions and semantic relations to classify semantic change types. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4539–4553, Bangkok, Thailand, August 2024b. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.249. URL <https://aclanthology.org/2024.acl-long.249/>.
- Vaggos Chatziafratis, Rad Niazadeh, and Moses Charikar. Hierarchical clustering with structural constraints. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 774–783. PMLR, 10–15 Jul 2018. URL <https://proceedings.mlr.press/v80/chatziafratis18a.html>.
- Jing Chen, Emmanuele Chersoni, Dominik Schlechtweg, Jelena Prokic, and Churen Huang. ChiWUG: A graph-based evaluation dataset for Chinese lexical semantic change detection. In Nina Tahmasebi, Syrielle Montariol, Haim Dubossarsky, Andrey Kutuzov, Simon Hengchen, David Alfter, Francesco Periti, and Pierluigi Cassotti, editors, *Proceedings of the 4th Workshop on Computational Approaches to Historical Language Change*, pages 93–99, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.lchange-1.10. URL <https://aclanthology.org/2023.lchange-1.10/>.
- Kostadin Cholakov, Chris Biemann, Judith Eckle-Kohler, and Iryna Gurevych. Lexical substitution dataset for German. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC’14)*, pages 1406–1411, Reykjavik, Iceland, May 2014. European Language Resources Association (ELRA). URL <https://aclanthology.org/L14-1450/>.
- Gabriella Chronis and Katrin Erk. When is a bishop not like a rook? when it’s like a rabbi! multi-prototype BERT embeddings for estimating semantic relationships. In Raquel Fernández and Tal Linzen, editors, *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 227–244, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.conll-1.17. URL <https://aclanthology.org/2020.conll-1.17/>.
- Gabriella Chronis, Kyle Mahowald, and Katrin Erk. A method for studying semantic construal in grammatical constructions with interpretable contextual embedding spaces. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki,

- editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 242–261, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.14. URL <https://aclanthology.org/2023.acl-long.14/>.
- Eve V. Clark. *Conventionality and contrast*, page 67–83. Cambridge Studies in Linguistics. Cambridge University Press, 1993. doi: 10.1017/CBO9780511554377.005.
- Eve V Clark. Conventionality and contrast in language and language acquisition. *New directions for child and adolescent development*, 2007(115):11–23, 2007.
- Eve V Clark and Herbert H Clark. Universals, relativity, and language processing. *Universals of human language*, 1:225–277, 1978.
- Vincent Cohen-Addad, Varun Kanade, Frederik Mallmann-Trenn, and Claire Mathieu. Hierarchical clustering: Objective functions and algorithms. *Journal of the ACM (JACM)*, 66(4):1–42, 2019.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.747. URL <https://aclanthology.org/2020.acl-main.747/>.
- Paul Cook, Jey Han Lau, Diana McCarthy, and Timothy Baldwin. Novel word-sense identification. In Junichi Tsujii and Jan Hajic, editors, *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1624–1635, Dublin, Ireland, August 2014. Dublin City University and Association for Computational Linguistics. URL <https://aclanthology.org/C14-1154/>.
- William Croft and D Alan Cruse. *Cognitive linguistics*. Cambridge University Press, 2004.
- D Alan Cruse. *Lexical semantics*. Cambridge university press, 1986.
- D Alan Cruse. Hyponymy and its varieties. In *The semantics of relationships: An interdisciplinary perspective*, pages 3–21. Springer, 2002.

Amaru Cuba Gyllensten, Evangelia Gogoulou, Ariel Ekgren, and Magnus Sahlgren. SenseCluster at SemEval-2020 task 1: Unsupervised lexical semantic change detection. In Aurelie Herbelot, Xiaodan Zhu, Alexis Palmer, Nathan Schneider, Jonathan May, and Ekaterina Shutova, editors, *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 112–118, Barcelona (online), December 2020. International Committee for Computational Linguistics. doi: 10.18653/v1/2020.semeval-1.12. URL <https://aclanthology.org/2020.semeval-1.12/>.

Sanjoy Dasgupta. A cost function for similarity-based hierarchical clustering. In *Proceedings of the Forty-Eighth Annual ACM Symposium on Theory of Computing*, STOC '16, pages 118–127, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450341325. doi: 10.1145/2897518.2897527. URL <https://doi.org/10.1145/2897518.2897527>.

Mark Davies. Expanding horizons in historical linguistics with the 400-million word corpus of historical american english. *Corpora*, 7:121–157, 2012. doi: 10.3366/COR.2012.0024. URL <https://doi.org/10.3366/COR.2012.0024>.

Gerard de Melo. Etymological Wordnet: Tracing the history of words. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1148–1154, Reykjavik, Iceland, May 2014. European Language Resources Association (ELRA). URL <https://aclanthology.org/L14-1063/>.

Ferdinand De Saussure. *Cours de linguistique générale*. Payot, 1916.

Sandra Derrig. Metaphor in the color lexicon. *Papers from the parasession on the lexicon*, pages 85–96, 1978.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Tamar Solorio, editors, *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, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL <https://aclanthology.org/N19-1423/>.

- Haim Dubossarsky, Yulia Tsvetkov, Chris Dyer, and Eitan Grossman. A bottom up approach to category mapping and meaning change. In *NetWordS*, pages 66–70, 2015.
- Haim Dubossarsky, Daphna Weinshall, and Eitan Grossman. Outta control: Laws of semantic change and inherent biases in word representation models. In Martha Palmer, Rebecca Hwa, and Sebastian Riedel, editors, *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1136–1145, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1118. URL <https://aclanthology.org/D17-1118/>.
- Haim Dubossarsky, Simon Hengchen, Nina Tahmasebi, and Dominik Schlechtweg. Time-out: Temporal referencing for robust modeling of lexical semantic change. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 457–470, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1044. URL <https://aclanthology.org/P19-1044/>.
- Philip Edmonds and Scott Cotton. SENSEVAL-2: Overview. In Judita Preiss and David Yarowsky, editors, *Proceedings of SENSEVAL-2 Second International Workshop on Evaluating Word Sense Disambiguation Systems*, pages 1–5, Toulouse, France, July 2001. Association for Computational Linguistics. URL <https://aclanthology.org/S01-1001/>.
- Jeffrey L Elman. An alternative view of the mental lexicon. *Trends in Cognitive Sciences*, 8(7):301–306, 2004. ISSN 1364-6613. doi: <https://doi.org/10.1016/j.tics.2004.05.003>. URL <https://www.sciencedirect.com/science/article/pii/S1364661304001366>.
- Kawin Ethayarajh. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 55–65, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1006. URL <https://aclanthology.org/D19-1006/>.
- Matan Eyal, Shoval Sadde, Hillel Taub-Tabib, and Yoav Goldberg. Large scale substitution-based word sense induction. In Smaranda Muresan, Preslav

- Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4738–4752, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.325. URL <https://aclanthology.org/2022.acl-long.325/>.
- Mariia Fedorova, Andrey Kutuzov, and Yves Scherrer. Definition generation for lexical semantic change detection. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5712–5724, Bangkok, Thailand, August 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.339. URL <https://aclanthology.org/2024.findings-acl.339/>.
- Mariia Fedorova, Timothee Mickus, Niko Partanen, Janine Siewert, Elena Spaziani, and Andrey Kutuzov. AXOLOTL’24 shared task on multilingual explainable semantic change modeling. In Nina Tahmasebi, Syrielle Montariol, Andrey Kutuzov, David Alfter, Francesco Periti, Pierluigi Cassotti, and Netta Huebscher, editors, *Proceedings of the 5th Workshop on Computational Approaches to Historical Language Change*, pages 72–91, Bangkok, Thailand, August 2024b. Association for Computational Linguistics. doi: 10.18653/v1/2024.lchange-1.8. URL <https://aclanthology.org/2024.lchange-1.8/>.
- Christiane Fellbaum. *WordNet: An electronic lexical database*. MIT press, 1998.
- James Champlin Fernald. *English Synonyms and Antonyms*. Funk & Wagnalls Company, 1896.
- John Rupert Firth. A synopsis of linguistic theory 1930-55. *Studies in Linguistic Analysis (special volume of the Philological Society)*, 1952-59:1–32, 1957.
- Antske Fokkens, Serge Ter Braake, Isa Maks, Davide Ceolin, et al. On the semantics of concept drift: Towards formal definitions of semantic change. *Drift-a-LOD@ EKAU*, 2016.
- Lauren Fonteyn and Charlotte Maekelberghe. Competing motivations in the diachronic nominalization of english gerunds. *Diachronica*, 35(4):487–524, 2018.
- Karlién Franco, Mariana Montes, and Kris Heylen. Deconstructing destruction: A cognitive linguistics perspective on a computational analysis of diachronic change. In Nina Tahmasebi, Syrielle Montariol, Andrey Kutuzov, Simon Hengchen, Haim Dubossarsky, and Lars Borin, editors, *Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change*,

- pages 23–32, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.lchange-1.3. URL <https://aclanthology.org/2022.lchange-1.3/>.
- Alexandre François. Lexical tectonics: Mapping structural change in patterns of lexification. *Zeitschrift für Sprachwissenschaft*, 41(1):89–123, 2022. doi: 10.1515/zfs-2021-2041. URL <https://doi.org/10.1515/zfs-2021-2041>.
- Gottlob Frege et al. Über sinn und bedeutung. *Zeitschrift für Philosophie und philosophische Kritik*, 100(1):25–50, 1892.
- Lea Frermann and Mirella Lapata. A Bayesian model of diachronic meaning change. *Transactions of the Association for Computational Linguistics*, 4:31–45, 2016. doi: 10.1162/tacl_a_00081. URL <https://aclanthology.org/Q16-1003/>.
- Olivia Fugikawa, Oliver Hayman, Raymond Liu, Lei Yu, Thomas Brochhagen, and Yang Xu. A computational analysis of crosslinguistic regularity in semantic change. *Frontiers in Communication*, Volume 8 - 2023, 2023. ISSN 2297-900X. doi: 10.3389/fcomm.2023.1136338. URL <https://www.frontiersin.org/journals/communication/articles/10.3389/fcomm.2023.1136338>.
- Tianyu Gao, Xingcheng Yao, and Danqi Chen. SimCSE: Simple contrastive learning of sentence embeddings. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.552. URL <https://aclanthology.org/2021.emnlp-main.552/>.
- Aina Garí Soler and Marianna Apidianaki. MULTISEM at SemEval-2020 task 3: Fine-tuning BERT for lexical meaning. In Aurelie Herbelot, Xiaodan Zhu, Alexis Palmer, Nathan Schneider, Jonathan May, and Ekaterina Shutova, editors, *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 158–165, Barcelona (online), December 2020. International Committee for Computational Linguistics. doi: 10.18653/v1/2020.semeval-1.18. URL <https://aclanthology.org/2020.semeval-1.18/>.
- Dirk Geeraerts. *Diachronic prototype semantics*. Oxford Studies in Lexicography and Lexicology. Oxford University Press, Oxford, England, March 1997.

- Dirk Geeraerts. *Theories of Lexical Semantics*. Oxford University Press, 11 2010. ISBN 9780198700302. doi: 10.1093/acprof:oso/9780198700302.001.0001. URL <https://doi.org/10.1093/acprof:oso/9780198700302.001.0001>.
- Dirk Geeraerts, Dirk Speelman, Kris Heylen, Mariana Montes, Stefano De Pascale, Karlien Franco, and Michael Lang. *Lexical variation and change*. Oxford University Press, London, England, November 2023.
- Thanasis Georgakopoulos and Stéphane Polis. Lexical diachronic semantic maps: Mapping the evolution of time-related lexemes. *Journal of Historical Linguistics*, 11(3):367–420, 2021.
- Thanasis Georgakopoulos and Stéphane Polis. The semantic map model: State of the art and future avenues for linguistic research. *Language and Linguistics Compass*, 12(2):e12270, 2018. doi: <https://doi.org/10.1111/lnc3.12270>. URL <https://compass.onlinelibrary.wiley.com/doi/abs/10.1111/lnc3.12270>. e12270 LNCO-0727.R1.
- Mario Giulianelli, Marco Del Tredici, and Raquel Fernández. Analysing lexical semantic change with contextualised word representations. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3960–3973, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.365. URL <https://aclanthology.org/2020.acl-main.365/>.
- Mario Giulianelli, Andrey Kutuzov, and Lidia Pivovarova. Do not fire the linguist: Grammatical profiles help language models detect semantic change. In Nina Tahmasebi, Syrielle Montariol, Andrey Kutuzov, Simon Hengchen, Haim Dubossarsky, and Lars Borin, editors, *Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change*, pages 54–67, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.lchange-1.6. URL <https://aclanthology.org/2022.lchange-1.6/>.
- Mario Giulianelli, Iris Luden, Raquel Fernandez, and Andrey Kutuzov. Interpretable word sense representations via definition generation: The case of semantic change analysis. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3130–3148, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.176. URL <https://aclanthology.org/2023.acl-long.176/>.

- Keith Green. *Bertrand Russell, language and linguistic theory*. Bloomsbury Publishing, 2007.
- Kristina Gulordava and Marco Baroni. A distributional similarity approach to the detection of semantic change in the Google Books ngram corpus. In Sebastian Pado and Yves Peirsman, editors, *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*, pages 67–71, Edinburgh, UK, July 2011. Association for Computational Linguistics. URL <https://aclanthology.org/W11-2508/>.
- Janosch Haber and Massimo Poesio. Patterns of polysemy and homonymy in contextualised language models. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2663–2676, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-emnlp.226. URL <https://aclanthology.org/2021.findings-emnlp.226/>.
- R. Hadsell, S. Chopra, and Y. LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742, 2006. doi: 10.1109/CVPR.2006.100.
- William L. Hamilton, Jure Leskovec, and Dan Jurafsky. Cultural shift or linguistic drift? comparing two computational measures of semantic change. In Jian Su, Kevin Duh, and Xavier Carreras, editors, *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2116–2121, Austin, Texas, November 2016a. Association for Computational Linguistics. doi: 10.18653/v1/D16-1229. URL <https://aclanthology.org/D16-1229/>.
- William L. Hamilton, Jure Leskovec, and Dan Jurafsky. Diachronic word embeddings reveal statistical laws of semantic change. In Katrin Erk and Noah A. Smith, editors, *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1501, Berlin, Germany, August 2016b. Association for Computational Linguistics. doi: 10.18653/v1/P16-1141. URL <https://aclanthology.org/P16-1141/>.
- Michael Hanna and David Mareček. Analyzing BERT’s knowledge of hypernymy via prompting. In Jasmijn Bastings, Yonatan Belinkov, Emmanuel Dupoux, Mario Giulianelli, Dieuwke Hupkes, Yuval Pinter, and Hassan Sajjad, editors, *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing*

- and Interpreting Neural Networks for NLP*, pages 275–282, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.blackboxnlp-1.20. URL <https://aclanthology.org/2021.blackboxnlp-1.20/>.
- Zellig S. Harris. Distributional structure. *WORD*, 10(2-3):146–162, 1954. doi: 10.1080/00437956.1954.11659520. URL <https://doi.org/10.1080/00437956.1954.11659520>.
- Martin Haspelmath. Coexpression and synexpression patterns across languages: comparative concepts and possible explanations. *Frontiers in Psychology*, 14, 2023. ISSN 1664-1078. doi: 10.3389/fpsyg.2023.1236853. URL <https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2023.1236853>.
- Felix Hill, Roi Reichart, and Anna Korhonen. SimLex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41(4):665–695, December 2015. doi: 10.1162/COLI_a_00237. URL <https://aclanthology.org/J15-4004/>.
- Daniil Homskiy and Nikolay Arefyev. DeepMistake at LSCDiscovery: Can a multilingual word-in-context model replace human annotators? In Nina Tahmasebi, Syrielle Montariol, Andrey Kutuzov, Simon Hengchen, Haim Dubossarsky, and Lars Borin, editors, *Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change*, pages 173–179, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.lchange-1.18. URL <https://aclanthology.org/2022.lchange-1.18/>.
- Paul J Hopper and Elizabeth Closs Traugott. *Grammaticalization*. Cambridge university press, 2003.
- Renfen Hu, Shen Li, and Shichen Liang. Diachronic sense modeling with deep contextualized word embeddings: An ecological view. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3899–3908, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1379. URL <https://aclanthology.org/P19-1379/>.
- Eric Huang, Richard Socher, Christopher Manning, and Andrew Ng. Improving word representations via global context and multiple word prototypes. In Haizhou Li, Chin-Yew Lin, Miles Osborne, Gary Geunbae Lee, and Jong C. Park, editors, *Proceedings of the 50th Annual Meeting of the Association for*

- Computational Linguistics (Volume 1: Long Papers)*, pages 873–882, Jeju Island, Korea, July 2012. Association for Computational Linguistics. URL <https://aclanthology.org/P12-1092/>.
- Lawrence J. Hubert and Phipps Arabie. Comparing partitions. *Journal of Classification*, 2:193–218, 1985.
- Gonia Jarema and Gary Libben. *The Mental Lexicon: Core Perspectives*. Brill, Leiden, The Netherlands, 2007. ISBN 978-00-80-54869-2. doi: 10.1163/9780080548692. URL <https://brill.com/view/title/23295>.
- David Jurgens and Ioannis Klapaftis. SemEval-2013 task 13: Word sense induction for graded and non-graded senses. In Suresh Manandhar and Deniz Yuret, editors, *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 290–299, Atlanta, Georgia, USA, June 2013. Association for Computational Linguistics. URL <https://aclanthology.org/S13-2049/>.
- Joan Casper Kahr. Adpositions and locationals: typology and diachronic development. *Working Papers on Language Universals*, 19:21–54, 1975.
- Vani Kanjirangat, Sandra Mitrovic, Alessandro Antonucci, and Fabio Rinaldi. SST-BERT at SemEval-2020 task 1: Semantic shift tracing by clustering in BERT-based embedding spaces. In Aurelie Herbelot, Xiaodan Zhu, Alexis Palmer, Nathan Schneider, Jonathan May, and Ekaterina Shutova, editors, *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 214–221, Barcelona (online), December 2020. International Committee for Computational Linguistics. doi: 10.18653/v1/2020.semeval-1.26. URL <https://aclanthology.org/2020.semeval-1.26/>.
- Kseniia Kashleva, Alexander Shein, Elizaveta Tukhtina, and Svetlana Vydrina. HSE at LSCDiscovery in Spanish: Clustering and profiling for lexical semantic change discovery. In Nina Tahmasebi, Syrielle Montariol, Andrey Kutuzov, Simon Hengchen, Haim Dubossarsky, and Lars Borin, editors, *Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change*, pages 193–197, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.lchange-1.21. URL <https://aclanthology.org/2022.lchange-1.21/>.
- Yoshifumi Kawasaki, Maëlys Salengre, Marzena Karpinska, Hiroya Takamura, and Ryo Nagata. Revisiting statistical laws of semantic shift in Romance cog-

- nates. In Nicoletta Calzolari, Chu-Ren Huang, Hansaem Kim, James Pustejovsky, Leo Wanner, Key-Sun Choi, Pum-Mo Ryu, Hsin-Hsi Chen, Lucia Donatelli, Heng Ji, Sadao Kurohashi, Patrizia Paggio, Nianwen Xue, Seokhwan Kim, Younggyun Hahm, Zhong He, Tony Kyungil Lee, Enrico Santus, Francis Bond, and Seung-Hoon Na, editors, *Proceedings of the 29th International Conference on Computational Linguistics*, pages 141–151, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics. URL <https://aclanthology.org/2022.coling-1.13/>.
- Christian Kay, Marc Alexander, Fraser Dallachy, Jane Roberts, Michael Samuels, and Irené Wotherspoon. *The Historical Thesaurus of English*. University of Glasgow, 2nd edition, 2020. URL <https://ht.ac.uk/>. version 5.0 accessed in June, 2025.
- Fahad Khan, John P. McCrae, Francisco Javier Minaya Gómez, Rafael Cruz González, and Javier E. Díaz-Vera. Some considerations in the construction of a historical language WordNet. In German Rigau, Francis Bond, and Alexandre Rademaker, editors, *Proceedings of the 12th Global Wordnet Conference*, pages 101–105, University of the Basque Country, Donostia - San Sebastian, Basque Country, January 2023. Global Wordnet Association. URL <https://aclanthology.org/2023.gwc-1.12/>.
- Yoon Kim, Yi-I Chiu, Kentaro Hanaki, Darshan Hegde, and Slav Petrov. Temporal analysis of language through neural language models. In Cristian Danescu-Niculescu-Mizil, Jacob Eisenstein, Kathleen McKeown, and Noah A. Smith, editors, *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pages 61–65, Baltimore, MD, USA, June 2014. Association for Computational Linguistics. doi: 10.3115/v1/W14-2517. URL <https://aclanthology.org/W14-2517/>.
- Ryo Kishino, Hiroaki Yamagiwa, Ryo Nagata, Sho Yokoi, and Hidetoshi Shimodaira. Quantifying lexical semantic shift via unbalanced optimal transport. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15913–15933, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.774. URL <https://aclanthology.org/2025.acl-long.774/>.
- Grzegorz Kleparski. *Semantic change in English: A study of evaluative developments in the domain of HUMANS*. Red. Wydawnictw Kul, 1990.

- Peter Koch. Cognitive onomasiology and lexical change: Around the eye. In *From polysemy to semantic change*, pages 107–137. John Benjamins Publishing Company, 2008.
- Denis Kokosinskii and Nikolay Arefyev. Multilingual substitution-based word sense induction. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11859–11872, Torino, Italia, May 2024. ELRA and ICCL. URL <https://aclanthology.org/2024.lrec-main.1035/>.
- Gerhard Kremer, Katrin Erk, Sebastian Padó, and Stefan Thater. What substitutes tell us - analysis of an “all-words” lexical substitution corpus. In Shuly Wintner, Sharon Goldwater, and Stefan Riezler, editors, *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 540–549, Gothenburg, Sweden, April 2014. Association for Computational Linguistics. doi: 10.3115/v1/E14-1057. URL <https://aclanthology.org/E14-1057/>.
- Saul A Kripke et al. *Naming and necessity*, volume 217. Springer, 1980.
- Anthony Kroch and Beatrice Santorini. Penn-BFM parsed corpus of historical french, version 1.0., 2021. URL <https://github.com/beatrice57/mcvf-plus-ppchf>.
- Heinz Kronasser. *Handbuch der Semasiologie: kurze Einführung in die Geschichte*. Winter, 1952.
- Vivek Kulkarni, Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. Statistically significant detection of linguistic change. In *Proceedings of the 24th International Conference on World Wide Web, WWW '15*, page 625–635, Republic and Canton of Geneva, CHE, 2015. International World Wide Web Conferences Steering Committee. ISBN 9781450334693. doi: 10.1145/2736277.2741627. URL <https://doi.org/10.1145/2736277.2741627>.
- Andrey Kutuzov and Mario Giulianelli. UiO-UvA at SemEval-2020 task 1: Contextualised embeddings for lexical semantic change detection. In Aurelie Herbelot, Xiaodan Zhu, Alexis Palmer, Nathan Schneider, Jonathan May, and Ekaterina Shutova, editors, *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 126–134, Barcelona (online), December 2020. International Committee for Computational Linguistics. doi: 10.18653/v1/2020.emeval-1.14. URL <https://aclanthology.org/2020.semeval-1.14/>.

- Andrey Kutuzov, Erik Velldal, and Lilja Øvrelid. Temporal dynamics of semantic relations in word embeddings: an application to predicting armed conflict participants. In Martha Palmer, Rebecca Hwa, and Sebastian Riedel, editors, *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1824–1829, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1194. URL <https://aclanthology.org/D17-1194/>.
- Andrey Kutuzov, Lidia Pivovarova, and Mario Giulianelli. Grammatical profiling for semantic change detection. In Arianna Bisazza and Omri Abend, editors, *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 423–434, Online, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.conll-1.33. URL <https://aclanthology.org/2021.conll-1.33/>.
- Andrey Kutuzov, Samia Touileb, Petter Mæhlum, Tita Enstad, and Alexandra Wittemann. NorDiaChange: Diachronic semantic change dataset for Norwegian. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2563–2572, Marseille, France, June 2022. European Language Resources Association. URL <https://aclanthology.org/2022.lrec-1.274/>.
- Andrey Kutuzov, Mariia Fedorova, Dominik Schlechtweg, and Nikolay Arefyev. Enriching word usage graphs with cluster definitions. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6189–6198, Torino, Italia, May 2024. ELRA and ICCL. URL <https://aclanthology.org/2024.lrec-main.546/>.
- George Lakoff. *Women, fire, and dangerous things: What categories reveal about the mind*. University of Chicago press, 1987.
- George Lakoff and Mark Johnson. *Metaphors we live by*. University of Chicago press, 1980.
- Ronald W Langacker. *Foundations of cognitive grammar: Volume I: Theoretical prerequisites*, volume 1. Stanford university press, 1987.
- Jey Han Lau, Paul Cook, Diana McCarthy, David Newman, and Timothy Baldwin. Word sense induction for novel sense detection. In Walter Daelemans, editor,

- Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 591–601, Avignon, France, April 2012. Association for Computational Linguistics. URL <https://aclanthology.org/E12-1060/>.
- Marie Leca-Tsiomis. Les dictionnaires en europe. *Dix-Huitième Siecle*, 38(1):4, 2006.
- Benoît Leclercq and Cameron Morin. No equivalence: A new principle of no synonymy. *Constructions*, 15(1), June 2023. doi: 10.24338/cons-535. URL <https://constructions.journals.hhu.de/article/view/535>.
- Christian Lehmann. *Thoughts on grammaticalization*. Language Science Press, 3rd edition, 2015.
- Adrienne Lehrer. *The influence of semantic fields on semantic change*, pages 283–296. De Gruyter Mouton, Berlin, New York, 1985. ISBN 9783110850178. doi: doi:10.1515/9783110850178.283. URL <https://doi.org/10.1515/9783110850178.283>.
- Piroska Lendvai, Uwe Reichel, Anna Jouravel, Achim Rabus, and Elena Renje. Domain-adapting BERT for attributing manuscript, century and region in pre-Modern Slavic texts. In Nina Tahmasebi, Syrielle Montariol, Haim Dubossarsky, Andrey Kutuzov, Simon Hengchen, David Alfter, Francesco Periti, and Pierluigi Cassotti, editors, *Proceedings of the 4th Workshop on Computational Approaches to Historical Language Change*, pages 15–21, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.lchange-1.2. URL <https://aclanthology.org/2023.lchange-1.2/>.
- Yoav Levine, Barak Lenz, Or Dagan, Ori Ram, Dan Padnos, Or Sharir, Shai Shalev-Shwartz, Amnon Shashua, and Yoav Shoham. SenseBERT: Driving some sense into BERT. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4656–4667, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.423. URL <https://aclanthology.org/2020.acl-main.423/>.
- Natalia Levshina and David Lorenz. Communicative efficiency and the principle of no synonymy: predictability effects and the variation of want to and wanna. *Language and Cognition*, 14(2):249–274, 2022. doi: 10.1017/langcog.2022.7.

- Bastien Liétard and Gabriel Loiseau. CALE : Concept-Aligned Embeddings for both within-lemma and inter-lemma sense differentiation, 2025. URL <https://arxiv.org/abs/2508.04494>. arXiv preprint.
- Bastien Liétard, Mikaela Keller, and Pascal Denis. A tale of two laws of semantic change: Predicting synonym changes with distributional semantic models. In Alexis Palmer and Jose Camacho-collados, editors, *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023)*, pages 338–352, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.starsem-1.30. URL <https://aclanthology.org/2023.starsem-1.30/>.
- Bastien Liétard, Pascal Denis, and Mikaela Keller. To word senses and beyond: Inducing concepts with contextualized language models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 2684–2696, Miami, Florida, USA, November 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.156. URL <https://aclanthology.org/2024.emnlp-main.156/>.
- Bastien Liétard, Mikaela Keller, and Pascal Denis. Towards an onomasiological study of lexical semantic change through the induction of concepts. In Nina Tahmasebi, Syrielle Montariol, Andrey Kutuzov, David Alfter, Francesco Periti, Pierluigi Cassotti, and Netta Huebscher, editors, *Proceedings of the 5th Workshop on Computational Approaches to Historical Language Change*, pages 158–167, Bangkok, Thailand, August 2024b. Association for Computational Linguistics. doi: 10.18653/v1/2024.lchange-1.15. URL <https://aclanthology.org/2024.lchange-1.15/>.
- Yuri Lin, Jean-Baptiste Michel, Erez Aiden Lieberman, Jon Orwant, Will Brockman, and Slav Petrov. Syntactic annotations for the Google Books NGram corpus. In Min Zhang, editor, *Proceedings of the ACL 2012 System Demonstrations*, pages 169–174, Jeju Island, Korea, July 2012. Association for Computational Linguistics. URL <https://aclanthology.org/P12-3029/>.
- Chunhua Liu, Trevor Cohn, and Lea Frermann. Seeking clozure: Robust hypernym extraction from BERT with anchored prompts. In Alexis Palmer and Jose Camacho-collados, editors, *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023)*, pages 193–206, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.starsem-1.18. URL <https://aclanthology.org/2023.starsem-1.18/>.

- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019. URL <https://arxiv.org/abs/1907.11692>.
- Suresh Manandhar, Ioannis Klapaftis, Dmitriy Dligach, and Sameer Pradhan. SemEval-2010 task 14: Word sense induction & disambiguation. In Katrin Erk and Carlo Strapparava, editors, *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 63–68, Uppsala, Sweden, July 2010. Association for Computational Linguistics. URL <https://aclanthology.org/S10-1011/>.
- Christiane Marchello-Nizia, Bernard Combettes, Sophie Prévost, and Tobias Scheer. *Grande grammaire historique du français (GGHF)*. de Gruyter, 2020.
- Matej Martinc, Petra Kralj Novak, and Senja Pollak. Leveraging contextual embeddings for detecting diachronic semantic shift. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4811–4819, Marseille, France, May 2020a. European Language Resources Association. ISBN 979-10-95546-34-4. URL <https://aclanthology.org/2020.lrec-1.592/>.
- Matej Martinc, Syrielle Montariol, Elaine Zosa, and Lidia Pivovarova. Capturing evolution in word usage: Just add more clusters? In *Companion Proceedings of the Web Conference 2020*, WWW ’20, page 343–349, New York, NY, USA, 2020b. Association for Computing Machinery. ISBN 9781450370240. doi: 10.1145/3366424.3382186. URL <https://doi.org/10.1145/3366424.3382186>.
- Matej Martinc, Syrielle Montariol, Elaine Zosa, and Lidia Pivovarova. Discovery team at SemEval-2020 task 1: Context-sensitive embeddings not always better than static for semantic change detection. In Aurelie Herbelot, Xiaodan Zhu, Alexis Palmer, Nathan Schneider, Jonathan May, and Ekaterina Shutova, editors, *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 67–73, Barcelona (online), December 2020c. International Committee for Computational Linguistics. doi: 10.18653/v1/2020.emeval-1.6. URL <https://aclanthology.org/2020.semeval-1.6/>.
- France Martineau, Paul Hirschbühler, Anthony Kroch, and Yves Charles Morin.

- MCVF corpus, parsed, version 2.0, 2021. URL <https://github.com/beatrice57/mcvf-plus-ppchf>.
- Eliabelle Mauduit and Andrea Simonetto. Constrained hierarchical clustering via graph coarsening and optimal cuts. In *2023 57th Asilomar Conference on Signals, Systems, and Computers*, pages 844–848, 2023. doi: 10.1109/IEEECONF59524.2023.10476840.
- Diana McCarthy. Lexical substitution as a task for WSD evaluation. In *Proceedings of the ACL-02 Workshop on Word Sense Disambiguation: Recent Successes and Future Directions*, pages 089–115. Association for Computational Linguistics, July 2002. doi: 10.3115/1118675.1118691. URL <https://aclanthology.org/W02-0816/>.
- Diana McCarthy and Roberto Navigli. SemEval-2007 task 10: English lexical substitution task. In Eneko Agirre, Lluís Màrquez, and Richard Wicentowski, editors, *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 48–53, Prague, Czech Republic, June 2007. Association for Computational Linguistics. URL <https://aclanthology.org/S07-1009/>.
- Aaron F McDaid, Derek Greene, and Neil Hurley. Normalized mutual information to evaluate overlapping community finding algorithms. *arXiv preprint arXiv:1110.2515*, 2011. URL <https://arxiv.org/abs/1110.2515>.
- Merriam-Webster dictionary. Entry for “show”, 2025. <https://www.merriam-webster.com/dictionary/show> [Accessed: 2025-15-06].
- Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, The Google Books Team, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak, and Erez Lieberman Aiden. Quantitative analysis of culture using millions of digitized books. *Science*, 331(6014):176–182, 2011. doi: 10.1126/science.1199644. URL <https://www.science.org/doi/abs/10.1126/science.1199644>.
- Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space, 2013a. URL <https://arxiv.org/abs/1301.3781>.
- Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In

- Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, page 3111–3119, Red Hook, NY, USA, 2013b. Curran Associates Inc.
- George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.
- Syrielle Montariol, Matej Martinc, and Lidia Pivovarov. Scalable and interpretable semantic change detection. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou, editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4642–4652, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.369. URL <https://aclanthology.org/2021.naacl-main.369/>.
- Andrea Moro and Roberto Navigli. SemEval-2015 task 13: Multilingual all-words sense disambiguation and entity linking. In Preslav Nakov, Torsten Zesch, Daniel Cer, and David Jurgens, editors, *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 288–297, Denver, Colorado, June 2015. Association for Computational Linguistics. doi: 10.18653/v1/S15-2049. URL <https://aclanthology.org/S15-2049/>.
- Anna Mosolova, Marie Candito, and Carlos Ramisch. Injecting Wiktionary to improve token-level contextual representations using contrastive learning. In Yvette Graham and Matthew Purver, editors, *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 34–41, St. Julian's, Malta, March 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.eacl-short.5. URL <https://aclanthology.org/2024.eacl-short.5/>.
- Gregory L Murphy. *The big book of concepts*. MIT Press, 2002.
- Ryo Nagata, Yoshifumi Kawasaki, Naoki Otani, and Hiroya Takamura. A computational approach to quantifying grammaticization of English deverbal prepositions. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 211–220, Torino, Italia, May 2024. ELRA and ICCL. URL <https://aclanthology.org/2024.lrec-main.19/>.

- Roberto Navigli. Word sense disambiguation: A survey. *ACM Comput. Surv.*, 41(2), February 2009. ISSN 0360-0300. doi: 10.1145/1459352.1459355. URL <https://doi.org/10.1145/1459352.1459355>.
- Roberto Navigli. A quick tour of word sense disambiguation, induction and related approaches. In Mária Bielíková, Gerhard Friedrich, Georg Gottlob, Stefan Katzenbeisser, and György Turán, editors, *SOFSEM 2012: Theory and Practice of Computer Science*, pages 115–129, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg.
- Roberto Navigli, David Jurgens, and Daniele Vannella. SemEval-2013 task 12: Multilingual word sense disambiguation. In Suresh Manandhar and Deniz Yuret, editors, *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 222–231, Atlanta, Georgia, USA, June 2013. Association for Computational Linguistics. URL <https://aclanthology.org/S13-2040/>.
- Teresa Paccosi, Stefano Menini, Elisa Leonardelli, Ilaria Barzon, and Sara Tonelli. Scent and sensibility: Perception shifts in the olfactory domain. In Nina Tahmasebi, Syrielle Montariol, Haim Dubossarsky, Andrey Kutuzov, Simon Hengchen, David Alfter, Francesco Periti, and Pierluigi Cassotti, editors, *Proceedings of the 4th Workshop on Computational Approaches to Historical Language Change*, pages 143–152, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.lchange-1.15. URL <https://aclanthology.org/2023.lchange-1.15/>.
- Allan Paivio. Dual coding theory and the mental lexicon. *The Mental Lexicon*, 5(2):205–230, 2010. ISSN 1871-1340. doi: <https://doi.org/10.1075/ml.5.2.04pai>. URL <https://www.jbe-platform.com/content/journals/10.1075/ml.5.2.04pai>.
- Patrick Pantel and Dekang Lin. Discovering word senses from text. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '02, page 613–619, New York, NY, USA, 2002. Association for Computing Machinery. ISBN 158113567X. doi: 10.1145/775047.775138. URL <https://doi.org/10.1145/775047.775138>.
- Tommaso Pasini, Alessandro Raganato, and Roberto Navigli. Xl-wsd: An extra-large and cross-lingual evaluation framework for word sense disambiguation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(15):

13648–13656, May 2021. doi: 10.1609/aaai.v35i15.17609. URL <https://ojs.aaai.org/index.php/AAAI/article/view/17609>.

Hermann Paul. *Prinzipien der Sprachgeschichte*. Niemeyer, 1880.

Eitan Adam Pechenick, Christopher M. Danforth, and Peter Sheridan Dodds. Characterizing the google books corpus: Strong limits to inferences of socio-cultural and linguistic evolution. *PLOS ONE*, 10(10):1–24, 10 2015. doi: 10.1371/journal.pone.0137041. URL <https://doi.org/10.1371/journal.pone.0137041>.

Yves Peirsman and Dirk Geeraerts. *Metonymy as a prototypical category*. De Gruyter Mouton, 2006.

Francesco Periti and Stefano Montanelli. Lexical semantic change through large language models: a survey. *ACM Comput. Surv.*, 56(11), June 2024. ISSN 0360-0300. doi: 10.1145/3672393. URL <https://doi.org/10.1145/3672393>.

Francesco Periti and Nina Tahmasebi. A systematic comparison of contextualized word embeddings for lexical semantic change. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4262–4282, Mexico City, Mexico, June 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.240. URL <https://aclanthology.org/2024.naacl-long.240/>.

Francesco Periti and Nina Tahmasebi. Towards a complete solution to lexical semantic change: an extension to multiple time periods and diachronic word sense induction. In Nina Tahmasebi, Syrielle Montariol, Andrey Kutuzov, David Alfter, Francesco Periti, Pierluigi Cassotti, and Netta Huebscher, editors, *Proceedings of the 5th Workshop on Computational Approaches to Historical Language Change*, pages 108–119, Bangkok, Thailand, August 2024b. Association for Computational Linguistics. doi: 10.18653/v1/2024.lchange-1.10. URL <https://aclanthology.org/2024.lchange-1.10/>.

Francesco Periti, Alfio Ferrara, Stefano Montanelli, and Martin Ruskov. What is done is done: an incremental approach to semantic shift detection. In Nina Tahmasebi, Syrielle Montariol, Andrey Kutuzov, Simon Hengchen, Haim Dubossarsky, and Lars Borin, editors, *Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change*, pages 33–43, Dublin,

- Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.lchange-1.4. URL <https://aclanthology.org/2022.lchange-1.4/>.
- Francesco Periti, Pierluigi Cassotti, Haim Dubossarsky, and Nina Tahmasebi. Analyzing semantic change through lexical replacements. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4495–4510, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.246. URL <https://aclanthology.org/2024.acl-long.246/>.
- Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In Marilyn Walker, Heng Ji, and Amanda Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1202. URL <https://aclanthology.org/N18-1202/>.
- Mohammad Taher Pilehvar and Jose Camacho-Collados. WiC: the word-in-context dataset for evaluating context-sensitive meaning representations. In Jill Burstein, Christy Doran, and Tamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1267–1273, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1128. URL <https://aclanthology.org/N19-1128/>.
- Mohammad Taher Pilehvar, Dimitri Kartsaklis, Victor Prokhorov, and Nigel Collier. Card-660: Cambridge rare word dataset - a reliable benchmark for infrequent word representation models. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun’ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1391–1401, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1169. URL <https://aclanthology.org/D18-1169/>.
- Martin Pömsl and Roman Lyapin. CIRCE at SemEval-2020 task 1: Ensembling context-free and context-dependent word representations. In Aurelie Herbelot, Xiaodan Zhu, Alexis Palmer, Nathan Schneider, Jonathan May, and Ekaterina Shutova, editors, *Proceedings of the Fourteenth Workshop on Semantic*

- Evaluation*, pages 180–186, Barcelona (online), December 2020. International Committee for Computational Linguistics. doi: 10.18653/v1/2020.emeval-1.21. URL <https://aclanthology.org/2020.emeval-1.21/>.
- B. Pottier. *Vers une sémantique moderne*. Centre de philologie et de littératures romanes de l'Université de Strasbourg, 1964. URL <https://books.google.fr/books?id=7ZcUcgAACAAJ>.
- Sameer Pradhan, Edward Loper, Dmitriy Dligach, and Martha Palmer. SemEval-2007 task-17: English lexical sample, SRL and all words. In Eneko Agirre, Lluís Màrquez, and Richard Wicentowski, editors, *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 87–92, Prague, Czech Republic, June 2007. Association for Computational Linguistics. URL <https://aclanthology.org/S07-1016/>.
- Marko Pranjić, Kaja Dobrovoljc, Senja Pollak, and Matej Martinc. Tracking semantic change in slovene: A novel dataset and optimal transport-based distance, 2024. URL <https://arxiv.org/abs/2402.16596>.
- Ondřej Pražák, Pavel Přibáň, Stephen Taylor, and Jakub Sido. UWB at SemEval-2020 task 1: Lexical semantic change detection. In Aurelie Herbelot, Xiaodan Zhu, Alexis Palmer, Nathan Schneider, Jonathan May, and Ekaterina Shutova, editors, *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 246–254, Barcelona (online), December 2020. International Committee for Computational Linguistics. doi: 10.18653/v1/2020.emeval-1.30. URL <https://aclanthology.org/2020.emeval-1.30/>.
- Maxim Rachinskiy and Nikolay Arefyev. GlossReader at LSCDiscovery: Train to select a proper gloss in English – discover lexical semantic change in Spanish. In Nina Tahmasebi, Syrielle Montariol, Andrey Kutuzov, Simon Hengchen, Haim Dubossarsky, and Lars Borin, editors, *Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change*, pages 198–203, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.lchange-1.22. URL <https://aclanthology.org/2022.lchange-1.22/>.
- Alessandro Raganato, Jose Camacho-Collados, and Roberto Navigli. Word sense disambiguation: A unified evaluation framework and empirical comparison. In Mirella Lapata, Phil Blunsom, and Alexander Koller, editors, *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 99–110, Valencia, Spain, April 2017.

Association for Computational Linguistics. URL <https://aclanthology.org/E17-1010/>.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1410. URL <https://aclanthology.org/D19-1410/>.

Rafael A. Rivera-Soto, Olivia Elizabeth Miano, Juanita Ordonez, Barry Y. Chen, Aleem Khan, Marcus Bishop, and Nicholas Andrews. Learning universal authorship representations. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 913–919, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.70. URL <https://aclanthology.org/2021.emnlp-main.70/>.

Sebastien Roch. Expanding the class of global objective functions for dissimilarity-based hierarchical clustering. *Journal of Classification*, 40(3): 513–526, 2023.

Martina A Rodda, Marco SG Senaldi, and Alessandro Lenci. Panta rei: Tracking semantic change with distributional semantics in ancient greek. *IJCoL. Italian Journal of Computational Linguistics*, 3(3-1):11–24, 2017.

Julia Rodina and Andrey Kutuzov. RuSemShift: a dataset of historical lexical semantic change in Russian. In Donia Scott, Nuria Bel, and Chengqing Zong, editors, *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1037–1047, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.90. URL <https://aclanthology.org/2020.coling-main.90/>.

Julia Rodina, Daria Bakshandaeva, Vadim Fomin, Andrey Kutuzov, Samia Touileb, and Erik Velldal. Measuring diachronic evolution of evaluative adjectives with word embeddings: the case for English, Norwegian, and Russian. In Nina Tahmasebi, Lars Borin, Adam Jatowt, and Yang Xu, editors, *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, pages 202–209, Florence, Italy, August 2019. As-

- sociation for Computational Linguistics. doi: 10.18653/v1/W19-4725. URL <https://aclanthology.org/W19-4725/>.
- Julia Rodina, Yuliya Trofimova, Andrey Kutuzov, and Ekaterina Artemova. Elmo and bert in semantic change detection for russian. In *Analysis of Images, Social Networks and Texts: 9th International Conference, AIST 2020, Skolkovo, Moscow, Russia, October 15–16, 2020, Revised Selected Papers*, page 175–186, Berlin, Heidelberg, 2020. Springer-Verlag. ISBN 978-3-030-72609-6. doi: 10.1007/978-3-030-72610-2_13. URL https://doi.org/10.1007/978-3-030-72610-2_13.
- Eleanor Rosch. Cognitive representations of semantic categories. *Journal of experimental psychology: General*, 104(3):192, 1975.
- Eleanor Rosch and Carolyn B Mervis. Family resemblances: Studies in the internal structure of categories. *Cognitive psychology*, 7(4):573–605, 1975.
- Eleanor H Rosch. Natural categories. *Cognitive psychology*, 4(3):328–350, 1973.
- Alex Rosenfeld and Katrin Erk. Deep neural models of semantic shift. In Marilyn Walker, Heng Ji, and Amanda Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 474–484, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1044. URL <https://aclanthology.org/N18-1044/>.
- Guy D. Rosin and Kira Radinsky. Temporal attention for language models. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz, editors, *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1498–1508, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-naacl.112. URL <https://aclanthology.org/2022.findings-naacl.112/>.
- Guy D. Rosin, Ido Guy, and Kira Radinsky. Time masking for temporal language models. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining, WSDM '22*, page 833–841, New York, NY, USA, 2022a. Association for Computing Machinery. ISBN 9781450391320. doi: 10.1145/3488560.3498529. URL <https://doi.org/10.1145/3488560.3498529>.
- Guy D. Rosin, Ido Guy, and Kira Radinsky. Time masking for temporal language models. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining, WSDM '22*, page 833–841, New York, NY,

- USA, 2022b. Association for Computing Machinery. ISBN 9781450391320. doi: 10.1145/3488560.3498529. URL <https://doi.org/10.1145/3488560.3498529>.
- David Rother, Thomas Haider, and Steffen Eger. CMCE at SemEval-2020 task 1: Clustering on manifolds of contextualized embeddings to detect historical meaning shifts. In Aurelie Herbelot, Xiaodan Zhu, Alexis Palmer, Nathan Schneider, Jonathan May, and Ekaterina Shutova, editors, *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 187–193, Barcelona (online), December 2020. International Committee for Computational Linguistics. doi: 10.18653/v1/2020.semeval-1.22. URL <https://aclanthology.org/2020.semeval-1.22/>.
- Christoph Rzymiski, Tiago Tresoldi, Simon J. Greenhill, Mei-Shin Wu, Nathanael E. Schweikhard, Maria Koptjevskaja-Tamm, Volker Gast, Timotheus A. Bodt, Abbie Hantgan, Gereon A. Kaiping, Sophie Chang, Yunfan Lai, Natalia Morozova, Heini Arjava, Nataliia Hübler, Ezequiel Koile, Steve Pepper, Mariann Proos, Briana Van Epps, Ingrid Blanco, Carolin Hundt, Sergei Monakhov, Kristina Pianykh, Sallona Ramesh, Russell D. Gray, Robert Forkel, and Johann-Mattis List. The database of cross-linguistic colexifications, reproducible analysis of cross-linguistic polysemies. *Scientific Data*, 7(1):13, Jan 2020. ISSN 2052-4463. doi: 10.1038/s41597-019-0341-x. URL <https://doi.org/10.1038/s41597-019-0341-x>.
- Eyal Sagi, Stefan Kaufmann, and Brady Clark. Semantic density analysis: Comparing word meaning across time and phonetic space. In Roberto Basili and Marco Pennacchiotti, editors, *Proceedings of the Workshop on Geometrical Models of Natural Language Semantics*, pages 104–111, Athens, Greece, March 2009. Association for Computational Linguistics. URL <https://aclanthology.org/W09-0214/>.
- Rakia Saidi and Fethi Jarray. Sentence transformers and distilbert for arabic word sense induction. 2023.
- Dominik Schlechtweg, Anna Häty, Marco Del Tredici, and Sabine Schulte im Walde. A wind of change: Detecting and evaluating lexical semantic change across times and domains. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 732–746, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1072. URL <https://aclanthology.org/P19-1072/>.

- Dominik Schlechtweg, Barbara McGillivray, Simon Hengchen, Haim Dubossarsky, and Nina Tahmasebi. SemEval-2020 task 1: Unsupervised lexical semantic change detection. In Aurelie Herbelot, Xiaodan Zhu, Alexis Palmer, Nathan Schneider, Jonathan May, and Ekaterina Shutova, editors, *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1–23, Barcelona (online), December 2020. International Committee for Computational Linguistics. doi: 10.18653/v1/2020.semeval-1.1. URL <https://aclanthology.org/2020.semeval-1.1/>.
- Dominik Schlechtweg, Nina Tahmasebi, Simon Hengchen, Haim Dubossarsky, and Barbara McGillivray. DWUG: A large resource of diachronic word usage graphs in four languages. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7079–7091, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.567. URL <https://aclanthology.org/2021.emnlp-main.567/>.
- Benjamin Snyder and Martha Palmer. The English all-words task. In *Proceedings of SENSEVAL-3, the Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, pages 41–43, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL <https://aclanthology.org/W04-0811/>.
- Pia Sommerauer and Antske Fokkens. Conceptual change and distributional semantic models: an exploratory study on pitfalls and possibilities. In Nina Tahmasebi, Lars Borin, Adam Jatowt, and Yang Xu, editors, *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, pages 223–233, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-4728. URL <https://aclanthology.org/W19-4728/>.
- Maja Stanojevic. Cognitive synonymy: A general overview. *Facta Universitatis, Series: Linguistics and Literature*, 7(2):193–200, 2009.
- James H Steiger. Tests for comparing elements of a correlation matrix. *Psychological bulletin*, 87(2):245, 1980.
- G. Stern. *Meaning and change of meaning; with special reference to the English language*. Meaning and change of meaning; with special reference to the English language. Wettergren & Kerbers, Oxford, England, 1931.

- Gustaf Stern. *Swift, swiftly, and their synonyms: A contribution to semantic analysis and theory*. Wettergren & Kerber, 1921.
- Eve Sweetser. *From etymology to pragmatics: Metaphorical and cultural aspects of semantic structure*, volume 54. Cambridge University Press, 1990.
- Eve Eliot Sweetser. *SEMANTIC STRUCTURE AND SEMANTIC CHANGE: A COGNITIVE LINGUISTIC STUDY OF MODALITY, PERCEPTION, SPEECH ACTS, AND LOGICAL RELATIONS (LEXICAL, CONDITIONALS, HISTORICAL, INDO-EUROPEAN, ETYMOLOGY)*. University of California, Berkeley, 1984.
- Nina Tahmasebi, Lars Borin, and Adam Jatowt. Survey of computational approaches to lexical semantic change detection. *Computational approaches to semantic change*, 6(1), 2021.
- Wayne A Taylor. Change-point analysis: a powerful new tool for detecting changes, 2000.
- Elizabeth C. Traugott and Richard B. Dasher. *Regularity in Semantic Change*. De Gruyter Mouton, Berlin, Boston, 2002. ISBN 0521583780. URL <https://www.degruyter.com/database/COGBIB/entry/cogbib.12064/html>. 2010.
- Elizabeth Closs Traugott. On the expression of spatio-temporal relations in language. *Universals of human language*, 3(369-400), 1978.
- Elizabeth Closs Traugott. Semantic change. In *Oxford research encyclopedia of linguistics*. Oxford University Press, 2017.
- Jost Trier. *Das sprachliche feld: Eine auseinandersetzung*. Verlag nicht ermittelbar, 1934.
- Peter D. Turney and Patrick Pantel. From frequency to meaning: vector space models of semantics. *J. Artif. Int. Res.*, 37(1):141–188, January 2010. ISSN 1076-9757.
- Ana Uban, Alina Maria Ciobanu, and Liviu P. Dinu. Studying laws of semantic divergence across languages using cognate sets. In Nina Tahmasebi, Lars Borin, Adam Jatowt, and Yang Xu, editors, *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, pages 161–166, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-4720. URL <https://aclanthology.org/W19-4720/>.
- Peter Uhrig. Why the principle of no synonymy is overrated. *Zeitschrift für Anglistik und Amerikanistik*, 63(3):323–337, 2015. doi: doi:10.1515/zaa-2015-0030. URL <https://doi.org/10.1515/zaa-2015-0030>.

- Stephen Ullmann. *Semantics: An introduction to the science of meaning*. Oxford: Blackwell, 1962.
- Tim Van de Cruys and Marianna Apidianaki. Latent semantic word sense induction and disambiguation. In Dekang Lin, Yuji Matsumoto, and Rada Mihalcea, editors, *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1476–1485, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. URL <https://aclanthology.org/P11-1148/>.
- Freek Van de Velde. Degeneracy: The maintenance of constructional networks. *Extending the scope of construction grammar*, 1:141–179, 2014.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
- Dan John Velasco, Axel Alba, Trisha Gail Pelagio, Bryce Anthony Ramirez, Jan Christian Blaise Cruz, Unisse Chua, Briane Paul Samson, and Charibeth Cheng. Towards automatic construction of Filipino WordNet: Word sense induction and synset induction using sentence embeddings. In Derry Wijaya, Alham Fikri Aji, Clara Vania, Genta Indra Winata, and Ayu Purwarianti, editors, *Proceedings of the First Workshop in South East Asian Language Processing*, pages 1–12, Nusa Dua, Bali, Indonesia, November 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.sealp-1.1. URL <https://aclanthology.org/2023.sealp-1.1/>.
- Ivan Vulić, Simon Baker, Edoardo Maria Ponti, Ulla Petti, Ira Leviant, Kelly Wing, Olga Majewska, Eden Bar, Matt Malone, Thierry Poibeau, Roi Reichart, and Anna Korhonen. Multi-SimLex: A large-scale evaluation of multilingual and crosslingual lexical semantic similarity. *Computational Linguistics*, 46(4):847–897, December 2020. doi: 10.1162/coli_a_00391. URL <https://aclanthology.org/2020.cl-4.5/>.
- Ekaterina Vylomova, Sean Murphy, and Nicholas Haslam. Evaluation of semantic change of harm-related concepts in psychology. In Nina Tahmasebi, Lars Borin, Adam Jatowt, and Yang Xu, editors, *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, pages 29–34, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-4704. URL <https://aclanthology.org/W19-4704/>.
- Dingkang Wang and Yusu Wang. An improved cost function for hierarchical cluster trees. *Journal of Computational Geometry*, 11(1):283–331, 2020.

- Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference, 2024. URL <https://arxiv.org/abs/2412.13663>.
- Zhibiao Wu and Martha Palmer. Verb semantics and lexical selection. In *32nd Annual Meeting of the Association for Computational Linguistics*, pages 133–138, Las Cruces, New Mexico, USA, June 1994. Association for Computational Linguistics. doi: 10.3115/981732.981751. URL <https://aclanthology.org/P94-1019/>.
- Aotao Xu, Jennifer E. Stellar, and Yang Xu. Evolution of emotion semantics. *Cognition*, 217:104875, 2021. ISSN 0010-0277. doi: <https://doi.org/10.1016/j.cognition.2021.104875>. URL <https://www.sciencedirect.com/science/article/pii/S0010027721002961>.
- Yang Xu and Charles Kemp. A computational evaluation of two laws of semantic change. In David C. Noelle, Rick Dale, Anne S. Warlaumont, Jeff Yoshimi, Teenie Matlock, Carolyn D. Jennings, and Paul P. Maglio, editors, *Proceedings of the 37th Annual Meeting of the Cognitive Science Society, CogSci 2015, Pasadena, California, USA, July 22-25, 2015*. Cognitive Science Society, 2015. URL <https://cogsci.mindmodeling.org/2015/papers/0463/index.html>.
- Zijun Yao, Yifan Sun, Weicong Ding, Nikhil Rao, and Hui Xiong. Dynamic word embeddings for evolving semantic discovery. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining, WSDM '18*, page 673–681, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450355810. doi: 10.1145/3159652.3159703. URL <https://doi.org/10.1145/3159652.3159703>.
- Lei Yu and Yang Xu. Word sense extension. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3281–3294, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.184. URL <https://aclanthology.org/2023.acl-long.184/>.
- Anna A Zalizniak, Maria Bulakh, Dmitrij Ganenkov, Ilya Gruntov, Timur Maisak, and Maxim Russo. The catalogue of semantic shifts as a database for lexical semantic typology. *Linguistics*, 50(3):633–669, 2012a.

- Anna A Zalizniak, Maria Bulakh, Dmitrij Ganenkov, Ilya Gruntov, Timur Maisak, and Maxim Russo. The catalogue of semantic shifts as a database for lexical semantic typology. *Linguistics*, 50(3), January 2012b.
- Frank D. Zamora-Reina, Felipe Bravo-Marquez, and Dominik Schlechtweg. LSCDiscovery: A shared task on semantic change discovery and detection in Spanish. In Nina Tahmasebi, Syrielle Montariol, Andrey Kutuzov, Simon Hengchen, Haim Dubossarsky, and Lars Borin, editors, *Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change*, pages 149–164, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.lchange-1.16. URL <https://aclanthology.org/2022.lchange-1.16/>.
- Guangjie Zeng, Hao Peng, Angsheng Li, Jia Wu, Chunyang Liu, and Philip S. Yu. Scalable semi-supervised clustering via structural entropy with different constraints. *IEEE Transactions on Knowledge and Data Engineering*, 37(1):478–492, 2025. doi: 10.1109/TKDE.2024.3486530.
- Zhuosheng Zhang, Shuohang Wang, Yichong Xu, Yuwei Fang, Wenhao Yu, Yang Liu, Hai Zhao, Chenguang Zhu, and Michael Zeng. Task compass: Scaling multi-task pre-training with task prefix. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5671–5685, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-emnlp.416. URL <https://aclanthology.org/2022.findings-emnlp.416/>.
- Jinan Zhou and Jiaxin Li. TemporalTeller at SemEval-2020 task 1: Unsupervised lexical semantic change detection with temporal referencing. In Aurelie Herbelot, Xiaodan Zhu, Alexis Palmer, Nathan Schneider, Jonathan May, and Ekaterina Shutova, editors, *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 222–231, Barcelona (online), December 2020. International Committee for Computational Linguistics. doi: 10.18653/v1/2020.emeval-1.27. URL <https://aclanthology.org/2020.emeval-1.27/>.
- George Kingsley Zipf. Human behavior and the principle of least effort. 1949.

A

Appendix

A.1 Appendices for Chapter 4

A.1.1 Splits and dataset statistics

In Table A.1 we display statistics over the different splits we used. Dev is a subset containing a sample of 10% of concepts and their occurrences. Synon. is a subset containing only concepts instantiated with 2 lemmas or more, and their occurrences.

	#Occs	#Lemmas	#Concepts	#Occs/Concept	#Occs/Lemmas	d_{Lex}	d_{Polysemy}
Full data	52'997	1'560	3'855	13.75	33.97	1.14	2.83
Dev	4795	389	386	12.42	12.33	1.14	1.13
Synon	13'158	630	447	29.44	20.89	2.24	1.59

Table A.1: Statistics on the different data splits in annotated SemCor. The split “Synon” only contains occurrences of concepts instantiated with multiple lemmas (cases of synonymy). d_{Lex} is the average number of unique lemmas per concept, d_{Polysemy} is the average number of distinct concepts per lemma.

A.1.2 Used hyperparameters and layers

Systems	Best hyperparameters
Local-only Kmeans	$k = 3$
Local-only Agglo	linkage = average, $\nu = 1.0$
Global-only Kmeans	$\pi = 120\%$
Global-only Agglo	linkage = average, $\nu = 3.5$
Bi-level Kmeans	$k = 8, \pi = 120\%$
Bi-level Agglo	linkage = average (both), $\nu_{\text{local}} = 0.0, \nu_{\text{global}} = 4.5$
Bi-level Kmeans (local Agglo)	linkage = average $\nu_{\text{local}} = 0.0, \pi = 120\%$
Bi-level Agglo (local Kmeans)	$k = 10$, linkage = average, $\nu_{\text{global}} = 4.5$

Table A.2: Best hyperparameters on the Dev split.

CLM layers Prior work like Ethayarajh (2019) showed that later layers usually correlates with deeper levels of contextualization and more semantic information, Chronis and Erk (2020) showed that moderately-late were preferred for lexical similarity while very last layers were preferred for semantic relatedness. To get embeddings, we try 4 sets of layers corresponding to different depths: first layers (1 to 4), moderately early layers (8 to 11), moderately late (14 to 17), and last layers (21 to 24). To get the representation of a word’s occurrence, we simply average its embeddings from the four chosen layers into one single 1024-dimensional embedding. For Concept Induction, we find that best results were obtained using layers 14 to 17, that are the reported results.

Hyperparameters For Eyal et al. (2022), we tried different resolution, varying it from $1e-3$ to 10, for the Louvain clustering but found very little to no effect.

For Kmeans at the local level, we varied the number of clusters k between 2 and 10. For Agglomerative clustering at both levels, we tried single, average and complete linkage.

The distance threshold in Agglo τ was indexed on the distribution of distances. We fixed an hyperparameter ν and derived $\tau = \text{avg}(d) - \nu \cdot \text{std}(d)$ with d the distribution of distances between clustered instances. We made ν vary between -4 and +8. For global Kmeans, the number of clusters was indexed using a proportion π on the number of lemmas (e.g. $120\% \times W$), π varying from 40% to 400%. This may help transferring hyperaparameters to other dataset in future research.

Best hyperaparameters choices are in Table A.2

A.1.3 Concept Clusters Size Distribution

The distribution of the concept cluster size (in number of lemmas) obtained with Bi-level Agglo system can be found in Figure A.1

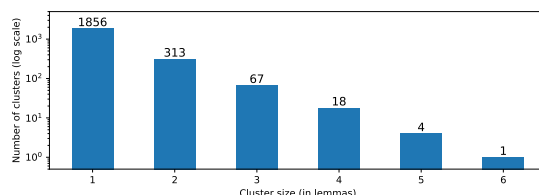


Figure A.1: Distribution of cluster size (in number of lemmas) obtained by the Bi-level Agglo system.

A.2 Appendices for Chapter 5

A.2.1 SPCD Dataset

To extract occurrences from the SemCor corpus Fellbaum (1998), we proceeded as follows:

- We used SemCor as distributed through the NLTK¹ library.
- We only used SemCor sentences that had semantic annotations (of WordNet’s synsets) and contained minimum 10 words and no more than 100.
- For each sentence, we extract an occurrence for each tagged word in the sentence (retrieved occurrences are delimited by the sentence).
- We focus on Adjective, Nouns and Verbs, using the Part-of-Speech tag indicated by the WordNet’s synsets. We merge PoS-tags “a” and “s” as they both cover Adjectives.
- For each PoS independently (to filter-out potential noisy PoS annotations), we keep only lemmas that meet the following conditions:
 - has minimum 3 letters.
 - occurs at least 10 times with the PoS.
 - is a single word (no compound) and contain only letters (no symbol/number).
 - is not a proper noun.
- We discard all occurrences of non-selected lemmas for each PoS tag, and then join the 3 remaining sets of lemmas and their occurrences. Now that extraction and filtering are done, we no longer differentiate on PoS tags (we allow pairs of occurrences from different PoS tags).

Table A.5 provides the number of pairs in each split and category.

A.2.2 Fine-Tuning Hyperparameters

We conducted hyperparameter optimization using Optuna’s² Tree-structured Parzen Estimator (TPE) Bergstra et al. (2011) algorithm to identify the best-performing configuration for our model. The optimal hyperparameters found through this search are presented in Table A.3.

¹<https://www.nltk.org/install.html>

²<https://optuna.org/>

Hyper-parameter	Value
Margin	0.7
Learning Rate	6.02e-06
Warmup Ratio	0.24
Weight Decay	0.05
Adam β_1	0.9
Adam β_2	0.999
Adam ϵ	1e-08
Epochs	1
Seed	42
Embedding Size	1024

Table A.3: CALE models hyper-parameters.

A.2.3 Test Metrics in SPCD

Table A.6 presents more detailed evaluation results on the SPCD dataset, distinguishing between the different pair categories (Same concept/Different concepts and Same lemma/Different lemma).

Table A.4 shows the F1 score in the *test* split of SPCD.

	All	SL	DL
1L1C (baseline)	62.8	69.5	0.0
XLM-R (first)	50.2	57.8	0.0
XLM-R (mid)	67.1	69.7	53.3
XLM-R (last)	64.2	68.2	42.8
XL-LEXEME	73.5	72.5	78.2
CALE (XLM-R)	76.2	74.5	83.6
CALE (MBERT)	75.4	74.4	79.9
CALE (XL-LEX)	76.0	74.8	81.2

Table A.4: F1 scores on test pairs from SemCor. “SL” (resp. “DL”) refers to pairs of occurrences of the same lemma (resp. of different lemmas) and “SC” (resp. “DC”) to pairs of occurrences referring to the same concept (resp. to different concepts).

A.2.4 APD and PRT results on DWUGs

In addition to APD, we also tried another measure from Kutuzov and Giulianelli (2020), the Prototype Distance (PRT), defined as follows:

$$\text{PRT}(w) = d\left(\frac{1}{n} \sum_{i=1}^n e_{w,i}^1, \frac{1}{m} \sum_{j=1}^m e_{w,j}^2\right)$$

with d the cosine distance operator, $e_{w,i}^k$ the embedding of the i -th occurrence of word w at time T_k , for a target word w with n occurrences at T_1 and m occurrences at T_2 .

Side-by-side results of APD and PRT can be found in Table A.7.

Split	Total	Concept		Lemma		Cross-categories			
		Same	Different	Same	Different	SC&SL	SC&DL	DC&SL	DC&DL
Train	156,387	63,752 (41%)	92,635 (59%)	91,551	64,836	48,517	15,235	43,034	49,601
Val	20,346	8,114 (40%)	12,232 (60%)	12,115	8,231	6,387	1,727	5,728	6,504
Test	44,891	18,317 (41%)	26,574 (59%)	26,318	18,573	14,018	4,299	12,300	14,274

Table A.5: SPCD dataset description in number of pairs. The train split used 49,601 unique occurrences, 6,504 for validation and 14,274 for test. “SL” (resp. “DL”) refers to pairs of occurrences of the same lemma (resp. of different lemmas) and “SC” (resp. “DC”) to pairs of occurrences referring to the same concept (resp. to different concepts).

	All [†]	SL [†]	DL [†]	SC&SL	DC&SL	DC&DL	SC&DL
1L1C (baseline)	65.1	50.0	50.0	100	0.0	100	0.0
XLM-R (first)	60.8	54.2	50.0	54.1	51.5	100	0.0
XLM-R (mid)	70.9	59.1	68.4	84.6	30.4	95.5	39.5
XLM-R (last)	69.1	59.6	63.5	77.9	38.4	95.8	28.8
XL-LEXEME	76.7	62.4	82.7	89.7	35.0	98.2	67.4
CALE (XLM-R)	79.3	65.9	86.4	89.1	38.0	98.5	73.1
CALE (MBERT)	78.7	66.6	83.5	87.2	41.9	98.9	67.0
CALE (XL-LEX)	79.2	67.2	84.4	87.2	43.8	98.9	69.4

Table A.6: Accuracies on test pairs from SemCor (Balanced Accuracies are indicated with [†]). The higher the better. “All” refers to all pairs, “SL” (resp. “DL”) to pairs of occurrences of the same lemma (resp. of different lemmas) and “SC” (resp. “DC”) to pairs of occurrences referring to the same concept (resp. to different concepts).

A.3 Appendices for Chapter 6

A.3.1 HDBSCAN cluster selection

HDBSCAN (Campello et al., 2013) is a density-based hierarchical clustering algorithm. By computing a *mutual reachability* distance between data points, it is

	EN		DE		ES		IT		LA		SV	
	APD	PRT	APD	PRT	APD	PRT	APD	PRT	APD	PRT	APD	PRT
XLM-R (first)	.221	.313	.143	.307	.078	.308	.180	.327	.004	.196	.098	.031
XLM-R (mid)	.568	.500	.543	.756	.495	.510	.316	.538	.237	.481	.575	.466
XLM-R (last)	.564	.527	.342	.404	.526	.562	.270	.605	.001	.521	.518	.307
XL-LEXEME	.786	.671	.798	.803	.633	.610	.512	.637	.142	.549	.854	.739
CALE (XLM-R)	.705	.533	.777	.823	.655	.543	.755	.753	.108	.461	.833	.720
CALE (MBERT)	.655	.527	-	-	-	-	-	-	-	-	-	-
CALE (XL-LEX)	.730	.570	.826	.799	.685	.639	.543	.676	.195	.485	.859	.766

Table A.7: Spearman correlations between gold change scores and predicted change scores for LSCD on the DWUG datasets, for APD and PRT measures.

able to construct a hierarchy in a provided dataset. While the definition of distance in HDBSCAN is original, the obtained hierarchy is comparable to that obtained with Agglomerative Clustering with *minimum* (or *single*) linkage. For the sake of our research, we are interested in HDBSCAN’s cluster selection method. This method does not impose conditions on the hierarchy, and can be applied even if the hierarchy has been produced by a different algorithm.

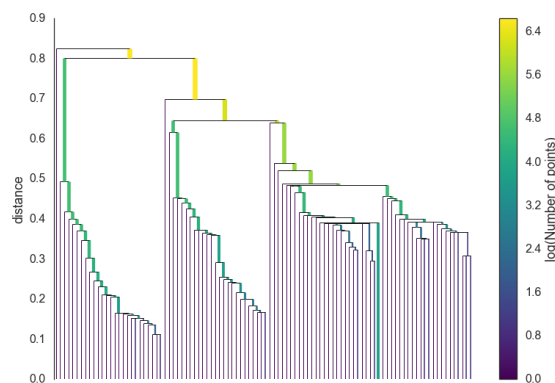


Figure A.2: Example of an HDBSCAN hierarchy (from https://hdbscan.readthedocs.io/en/latest/how_hdbscan_works.html)

Let us consider the hierarchy in Figure A.2. The hierarchy is a binary tree, where each leaf is a data point from the dataset. At the lowest level of the hierarchy, each datapoint is its own cluster. Going through the hierarchy from bottom to top, nodes correspond to *merge operations* between two clusters. Each merge operation is associated to a distance on the y -axis (from which the hierarchy is constructed; in Agglomerative Clustering, it is defined by the choice of linkage). Conversely, if we explore the hierarchy from top to bottom, nodes are *division operations*: at a certain distance level, a large cluster is divided into two smaller clusters, until each datapoint is its own cluster.

Condensing the hierarchy. The method relies on a single parameter m , the *minimum cluster size*. Going through the hierarchy, m enables a condensing of the hierarchy: if a cluster division results in two clusters of more than m data points each, it is a valid split. If one of the resulting cluster has less than m instances, then the larger cluster keeps the identity of the parent cluster, and the smaller cluster is ignored (these are datapoints that have simply “fallen out” of the cluster). If both children of the division operation have less than m instances, then the split is ignored. This process results in a condensed tree, as show in Figure A.3 The y -axis of the condensed tree is relabeled, such that it no longer indicates the distance d at which split happens. Instead, the y -axis shows the value of $\lambda = \frac{1}{d}$. Deeper splits are associated to larger value of λ .

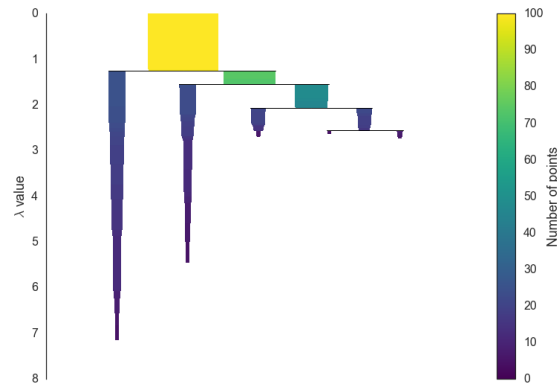


Figure A.3: Example of an HDBSCAN condensed tree (from https://hdbscan.readthedocs.io/en/latest/how_hdbscan_works.html)

Stability of a cluster. Two values λ_+ and λ_- are defined for each cluster kept in the condensed tree. λ_+ is the value on the y -axis at which the cluster splits from its parent and becomes its one cluster. λ_- is the value on the y -axis at which the cluster splits into two clusters, if any. To each datapoint x_i in the cluster is associated a value λ_i , the value on the y -axis at which ‘fell out’ of the cluster ($\lambda_i = \lambda_-$ if the point x_i remained in the cluster until it split into two clusters). For each cluster K , a **stability** value is defined as $S_K = \sum_{x_i \in K} (\lambda_i - \lambda_+)$.

Selecting clusters. All clusters receive the label ‘1’. Going from bottom to top, the stability of a cluster is replaced by sum of the stabilities of its children if the latter quantity is former than the latter. Otherwise, if a cluster has a higher stability than the sum of its children, then this cluster is still labeled ‘1’ but all of its descendants receive the label ‘0’. Once the top of the tree is reached, clusters that are still labeled ‘1’ are the *selected clusters*. This stability-based and iterative bottom-up method allow to select clusters at different levels of the hierarchy (see Figure A.4), unlike a simple horizontal cut with a distance threshold.

After that, any point that is not in a selected cluster is identified as a *noise* point.

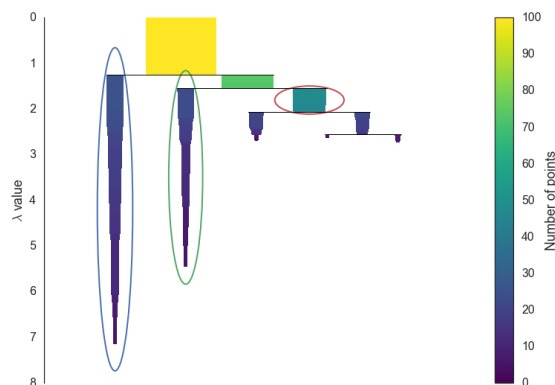


Figure A.4: Example of an HDBSCAN condensed tree with selected clusters in circles (from https://hdbscan.readthedocs.io/en/latest/how_hdbscan_works.html)

A.4 Appendices for Chapter 7

A.4.1 Target words selection

Among words represented in the embeddings provided by Hamilton et al. (2016b), we keep only words following these three requirements. The first is to be POS-tagged as an *adjective*, a *noun* and/or as a *verb* in the COHA. For a given POS-tag among these three, the second requirement is to appear at least 3 times in every decade between 1890 and 1999. Lastly, we require words to be composed of 3 letters or more. If a word appears with multiple POS-tags in the COHA and fulfills the minimum frequency requirement with each of these tags, the same embedding is used as its representation, as Hamilton et al. (2016b)’s training data aggregated POS-tags.

A.4.2 Frequency groups

The procedure to create a fixed number M of frequency group is the following. At a time T_p , the list of target words is sorted by frequency in ascending order, we label as group ‘0’ the first 50% of the list. In the remaining 50%, The first half is labeled as group ‘1’, and so on until group $M - 2$ is created. The still unlabeled words are labeled group $M - 1$, for a total of M groups. Group labels are therefore positively correlated with occurrences counts. This method creates frequency groups that reflect that the way word frequencies are distributed according to Zipf’s law: the largest group is that of unfrequent words (group 0), and groups of higher frequencies contain fewer and fewer words.

A.4.3 Unsupervised models

In Figure A.5, we observe that the quantity Δ does not reflect a clear separation between *Syn* pairs and *Diff* pairs. This explains why the unsupervised methods proposed in Sec. ?? fail to significantly outperform baselines.

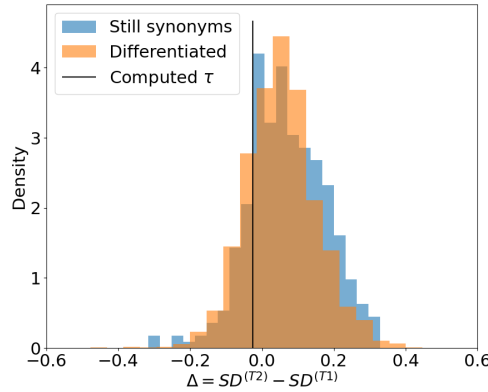


Figure A.5: Histograms of the value of divergence Δ of synonymous pairs, depending whether they differentiated (orange) or stayed synonyms (blue).

A.4.4 Components of selected models

Depending on the POS tag, the implementation strategies of *SD*, *DD* and frequency variables were different. Recall that these strategies were chosen given the average performances over 20 random train/test splits. In the following section, ‘SD(cd)’ refers to synchronic distance with cosine distance, while ‘SD(nk)’ refers to synchronic distance with k -nearest neighbors.

On adjectives, neighborhood based *DD* as well as raw frequency counts was found to be better than alternatives. For *SD*, cosine distance provides slightly higher performances than neighborhood-based measures. Among every method using a neighborhood-based *SD*, a small/mid ranged k was preferable (between 10 and 20).

For nouns, SD(cosine distance) was also the best choice. Overall, the best range for the value of k for neighbors-based *SD* was smaller than for adjectives (5 to 15). Frequency groups worked better than raw frequency counts, while there was no difference in performance between the two definitions of *DD*.

For verbs, SD(nk) with $k = 40$ actually outperforms cosine distance, except for the unsupervised Δ method for which SD(cd) was preferred across all POS tags. Diachronic Distances using Orthogonal Procrustes alignment and cosine distance (Hamilton et al., 2016b) lead to better results than the definition relying on

comparisons local neighborhoods. Both types of frequency variables (raw counts and groups) worked equally well.

Finally, on the ALL dataset reuniting pairs accross POS tags, raw frequencies provide better results than groups. Cosine distance is better than neighborhoods for synchronic distances, and both techniques of diachronic distances performed similarly. For Logistic Regression models using both implementations of SD (cd and nk), the choice of k did not really change the results as the weights of $SD(nk)$ was often negligible compared to the weight on $SD(cd)$.

A.4.5 Full classification results

		ADJ	NN	VERB	ALL			
		Balanced Accuracy			BA	$F_1(Syn)$	$F_1(Diff)$	%(D)
Base.	All (<i>Syn</i>)	.50	.50	.50	.50	.48	0	0
	All (<i>Diff</i>)	.50	.50	.50	.50	0	.81	100
	LR F	.51	.56	.59	.55	.35	.74	75
Unsup.	XK controls	.52	.49	.51	.50	.33	.67	65
	Δ (cd)	.50	.49	.51	.50	.27	.73	75
	Δ (nk)	.48	.49	.49	.50	.32	.67	66
Supervised	LR SD	.60	.62	.59	.60	.48	.69	56
	LR SD + DD	.61	.62	.60	.60	.48	.69	56
	LR SD + F	.61	.64	.63	.62	.51	.71	57
	LR SD + DD + F	.62	.64	.63	.62	.50	.70	57
	LR multi	.62	.64	.65	.62	.51	.71	57
	LR multi. (polynomial)	.56	.63	.62	.62	.50	.70	60
	SVM (gaussian)	.60	.64	.65	.62	.50	.74	63

Table A.8: Performances of the different approaches. Results are averaged over 20 random splits.

A.4.6 Predictive variables in our model

In this supplementary section, we conduct a study about the role of some predictive variables in our best-performing Logistic Regression model. The studied model uses SD with cosine-distance, both implementations of DD and frequency counts.

For a selected number of variables, we look for significant differences between well-classified pairs and pairs with wrong prediction, in both classes separately. For a given variable, we estimate if a difference is significant between the well-classified and the misclassified samples of this class using a t -test for Gaussian distributed variables, or a Mann-Whitney U test for other variables. A difference

	Pred.		$y = \text{Syn}$		$y = \text{Diff}$	
	Syn	Diff	TS	FD	TD	FS
$SD^{(TT_1)}$	.64	.83	.62	.84	.83	.64
$DD(w)$	.46	.46	.45	.47	.46	.47
$DD(v)$	.50	.54	.48	.54	.54	.50
$FG_w^{(T_2)}$	2.3	2.2	2.4	2.1	2.2	2.2
$FG_v^{(T_2)}$	1.9	1.4	2.1	1.5	1.4	1.8

Table A.9: Average values of some variables for data subset based on the prediction of our best-performing LR model. TS,FS,TD,FD stand for True/False Syn/Diff. $FG_w^{(T)}$ stands for Frequency Groups of word w at time T . Significant difference within a pair of columns are in bold.

is significant if the p -value of the test is below 5%. Results are reported in table A.9.

We observe significant differences of SD in pairs that are predicted as *Syn* and those predicted as *Diff* by our model, the first having a smaller SD at T_1 than the latter. Because our model relies mostly on these SD to separate both classes, we wrongly classify *Syn* pairs whose $SD^{(TT_1)}$ is close to that of *Diff*, and conversely *Diff* pairs whose $SD^{(TT_1)}$ is close to that of *Syn* are misclassified. This indicates that our model still misses some subtleties that are now reflected by SD.

A similar non-separability of the distribution of "Syns" and "Diff" appears on DD and Frequency variable for the second word pair of the pair. While it seems logical for our model to behave so regarding to the definition of LD, it is a clue that our input variables reflect noisy information that is confusing to the model. In the same idea, ? remarked that recent LSC detection models tend to raise False Positive, drawing attention to the limit of current models for LSC.