Human and Computational Measurement of Lexical Semantic Change

Von der Fakultät Informatik, Elektrotechnik und Informationstechnik der Universität Stuttgart zur Erlangung der Würde eines Doktors der Philosophie (Dr. phil.) genehmigte Abhandlung.

Vorgelegt von

Dominik Schlechtweg
aus Stuttgart, Deutschland

Hauptberichterin:	Apl. Prof. Dr. Sabine Schulte im Walde
Mitberichter:	Dr. Peter Turney
Mitberichter:	Prof. Dr. Jonas Kuhn

Tag der mündlichen Prüfung:	24. März 2022

Institut für Maschinelle Sprachverarbeitung der Universität Stuttgart

2023

Dedicated to those of my family who left us during the creation of this thesis:

Rolf-Dieter Schlechtweg, ✝ *February 2, 2018*
Theresia Edeltraud Hoffmann, ✝ *June 19, 2020*
Volker Axel Detlef Hoffmann, ✝ *July 9, 2021*

I hereby declare that I have created this work completely on my own and used no other sources or tools than the ones listed, and that I have marked any citations accordingly.

Hiermit versichere ich, dass ich die vorliegende Arbeit selbständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt sowie Zitate kenntlich gemacht habe.

Stuttgart, February 25, 2023

Place, Date

Dominik Schlechtweg

Contents

List of Figures

List of Tables

List of Abbreviations

AGL	Agglomerative Clustering
APD	Average Pairwise Distance
ARI	Adjusted Rand Index
BERT	Bidirectional Encoder Representations from Transformers
BZ	Berliner Zeitung Corpus
CCOHA	Clean Corpus of Historical American English
CD	Cosine Distance
CI	Column Intersection
CNT	Raw Count Matrix
COOK	Cooking Corpus
COS	Cosine Distance on mean vectors
DTA	Deutsches Textarchiv Corpus
DURel	Diachronic Usage Relatedness
FD	Frequency Difference
HD	Entropy Difference
JSD	Jensen-Shannon Distance
LND	Local Neighborhood Distance
LSC	Lexical Semantic Change
LSCD	Lexical Semantic Change Detection

ND	Neues Deutschland Corpus
OP	Orthogonal Procrustes
POS	Part-of-Speech
PPMI	Positive Pointwise Mutual Information
RI	Random Indexing
SCAN	Sense ChANge Topic Model
SdeWaC	SdeWaC Corpus
SemEval	Workshop on Semantic Evaluation
SGNS	Skip-Gram with Negative Sampling
SRV	Shared Random Vectors
SURel	Synchronic Usage Relatedness
SVD	Singular Value Decomposition
TD	Type Difference
USG	Use-Sense Graph
VI	Vector Initialization
VSM	Vector Space Model
WI	Word Injection
WiC	Word-in-Context
WUG	Word Usage Graph

Abstract

Human language changes over time. This change occurs on several linguistic levels such as grammar, sound or meaning. The study of meaning changes on the word level is often called **Lexical Semantic Change** (LSC) and is traditionally either approached from an onomasiological perspective asking by which words a meaning can be expressed, or a semasiological perspective asking which meanings a word can express over time. In recent years, the task of automatic detection of semasiological LSC from textual data has been established as a proper field of computational linguistics under the name of **Lexical Semantic Change Detection** (LSCD). Two main factors have contributed to this development: (i) The **digital turn** in the humanities has made large amounts of historical texts available in digital form. (ii) New computational models have been introduced efficiently learning semantic aspects of words solely from text.

One of the main motivations behind the work on LSCD are their applications in historical semantics and historical lexicography, where researchers are concerned with the classification of words into categories of semantic change. Automatic methods have the advantage to produce semantic change predictions for large amounts of data in small amounts of time and could thus considerably decrease human efforts in the mentioned fields while being able to scan more data and thus to uncover more semantic changes, which are at the same time less biased towards ad hoc sampling criteria used by researchers. On the other hand, automatic methods may also be hurtful when their predictions are biased, i.e., they may miss numerous semantic changes or label words as changing which are not. Results produced in this way may then lead researchers to make empirically inadequate generalizations on semantic change. Hence, automatic change detection methods should not be trusted until they have been evaluated thoroughly and their predictions have been shown to reach an acceptable level of correctness.

Despite the rapid growth of LSCD as a field, a solid evaluation of the wealth of proposed models was still missing at the onset of this thesis. The reasons were multiple, but most importantly there was no annotated benchmark test set available. This thesis is thus concerned with the process of providing such an evaluation for LSCD, including

- the definition of the basic concepts and tasks,
- the development and validation of data annotation schemes with humans,
- the annotation of a multilingual benchmark test set,
- the evaluation of computational models on the benchmark, their analysis and improvement, as well as
- an application of the developed methods to showcase their usefulness in the targeted fields (historical semantics and lexicography).

Überblick

Die menschliche Sprache verändert sich im Laufe der Zeit. Dieser Wandel vollzieht sich auf verschiedenen sprachlichen Ebenen wie Grammatik, Laute oder Bedeutung. Die Untersuchung von Bedeutungsänderungen auf der Wortebene wird oft als **lexikalischer Bedeutungswandel** bezeichnet und wird traditionell entweder aus einer onomasiologischen Perspektive mit der Frage angegangen, durch welche Wörter eine Bedeutung ausgedrückt werden kann, oder aus einer semasiologischen Perspektive mit der Frage, welche Bedeutungen ein Wort im Laufe der Zeit ausdrücken kann. In den letzten Jahren hat sich die Aufgabe der automatischen Erkennung von semasiologischem Bedeutungswandel aus Textdaten als ein eigenes Gebiet der Computerlinguistik unter dem Namen **Bedeutungswandelerkennung** etabliert. Zwei Hauptfaktoren haben zu dieser Entwicklung beigetragen: (i) Der **digital turn** in den Geisteswissenschaften hat große Mengen historischer Texte in digitaler Form verfügbar gemacht. (ii) Es wurden neue Computermodelle eingeführt, die semantische Aspekte von Wörtern allein aus Texten effizient erlernen.

Eine der Hauptmotivationen für die Arbeit an der Bedeutungswandelerkennung sind ihre Anwendungen in der historischen Semantik und der historischen Lexikographie, wo sich Forscher unter anderem mit der Klassifizierung von Wörtern in Kategorien des Bedeutungswandels beschäftigen. Automatische Methoden haben den Vorteil, dass sie Bedeutungswandel für große Datenmengen in kurzer Zeit vorhersagen und so den menschlichen Aufwand in den genannten Bereichen erheblich verringern können, während sie in der Lage sind, mehr Daten zu scannen und somit mehr semantische Veränderungen aufzudecken, die gleichzeitig weniger durch die von Forschern verwendeten Ad-hoc-Stichprobenkriterien beeinflusst werden. Andererseits können automatische Methoden auch schädlich sein, wenn ihre Vorhersagen fehlerhaft sind, d. h. sie können zahlreiche semantische Veränderungen übersehen oder Bedeutungswandel bei Wörtern erkennen, die keinen durchlaufen. Die auf diese Weise gewonnenen Ergebnisse könnten dann

Forscher dazu verleiten, empirisch fehlerhafte Verallgemeinerungen über Bedeutungswandel vorzunehmen. Daher sollte man automatischen Methoden zur Erkennung von Bedeutungswandel erst dann vertrauen, wenn sie gründlich evaluiert wurden und ihre Vorhersagen nachweislich einen akzeptablen Grad an Korrektheit erreicht haben.

Trotz des rasanten Wachstums der Bedeutungswandelerkennung als Gebiet der Computerlinguistik fehlte zu Beginn dieser Arbeit noch eine solide Evaluierung der Fülle der vorgeschlagenen Modelle. Die Gründe dafür waren vielfältig, aber am wichtigsten war, dass kein annotierter Benchmark-Testdatensatz verfügbar war. Diese Arbeit befasst sich daher mit dem Prozess der Durchführung einer solchen Evaluierung für die Bedeutungswandelerkennung, einschließlich

- der Definition der grundlegenden Konzepte und Tasks,
- der Entwicklung und Validierung von Datenannotationsprozessen mit Menschen,
- der Annotation eines mehrsprachigen Benchmark-Testdatensatzes,
- der Evaluierung von Computermodellen auf dem Testdatensatz, deren Analyse und Verbesserung, sowie
- der Anwendung der entwickelten Methoden, um ihre Nützlichkeit in den Zielbereichen (historische Semantik und Lexikographie) zu zeigen.

Acknowledgements

First of all, I want to thank my supervisor Sabine Schulte im Walde who has supported me with advice, trust and ressources throughout the five years of creating this thesis.

Next, I want to thank the Konrad Adenauer Foundation for awarding me a three-year Phd scholarship giving me immense independence and freedom to pursue my research ideas.

Special thanks belongs to the group of researchers I have organized the SemEval-2020 shared task with: Barbara McGillivray, Haim Dubossarsky, Nina Tahmasebi and Simon Hengchen. Many ideas presented in this thesis have profited from discussions with them.

The thesis has also profited immensely from collaborations with students implementing ideas and contributing their own ones. Most important to mention are Jens Kaiser, Severin Laicher, Sinan Kurtyigit and Serge Kotchourko.

I also want to thank a number of researchers who have supported me in different ways: Enrico Santus and Vered Shwartz for integrating me into an excellent project at a very early stage of my career; Nikolay Arefyev for extensive discussions and together with Serge Kotchourko and Sean Papay for advice on mathematical matters; Maike Park, Stefanie Eckmann and Anna Hätty for very fruitful collaborations; Cennet Oguz for support and exchange of ideas; Peter Turney and Diana McCarthy for an expert view on my research; My brother Sascha Schlechtweg for helpful discussions and together with my sister Tatjana Schlechtweg for an outsider's perspective on my field; Agnieszka Faleńska for help and support; Jonas Kuhn for ressources to implement innovative research ideas; The IMS staff

for providing an excellent research environment and the members of the former SemRel group including Maximilian Köper, Kim-Anh Nguyen, Sylvia Springorum, Stefan Bott and Jeremy Barnes for giving input to my ideas at the beginning of this thesis.

My gratitude also goes to Michael Schmitz from Konrad Adenauer Foundation for supporting my Phd project.

I want to thank my family for supporting me on the way I chose and bearing the burden that comes with it. My wife Manuela and my son Jannik deserve the greatest thanks for being my rock at all times, but also my sister Tatjana and my brother Sascha.

Chapter 1

Introduction

Human language changes over time. This change occurs on several linguistic levels such as grammar, sound or meaning (Bybee, 2015). The study of meaning changes on the word level is often called **Lexical Semantic Change** (LSC) and is traditionally either approached from an onomasiological perspective asking by which words a meaning can be expressed, or a semasiological perspective asking which meanings a word can express over time (Geeraerts, 2020). The latter perspective is exemplified by considering the German word *Presse* and the senses it expressed around 1800 (Paul, 2002). Before ≈1800 *Presse* was mainly used in the sense of 'press machine'. After 1800 we still observe this sense, and in addition we find a new, clearly distinguished sense as 'news press': The word gained an additional sense and thus changed its meaning.

In recent years, the task of automatic detection of semasiological LSC from textual data has been established as a proper field of computational linguistics under the name of **Lexical Semantic Change Detection** (LSCD) with the number of papers written on this topic exploding since 2016 (Tahmasebi et al., 2021; Schlechtweg et al., 2020). Two main factors have contributed to this development: (i) The **digital turn** in the humanities has made large amounts of historical texts available in digital form (cf. Geeraerts, 2020, pp. 20–21). (ii) New computational models have been introduced efficiently learning semantic aspects of words solely from text (Mikolov et al., 2013a; Peters et al., 2018; Devlin et al., 2019).

One of the main motivations behind the work on LSCD are their applications in historical semantics and historical lexicography, where researchers are concerned with the classification of words into categories of semantic change (cf. Paul, 2002; Blank, 1997). Automatic methods have the advantage to produce semantic change predictions for large amounts of data in small amounts of time and could thus considerably decrease human efforts in the mentioned fields while being able to scan more data and thus to uncover more semantic changes, which are at the same time less biased towards ad hoc sampling criteria used by researchers. On the other hand, automatic methods may also be hurtful when their predictions are biased, i.e., they may miss numerous semantic changes or label words as changing which are not. Results produced in this way may then lead researchers to make empirically inadequate generalizations on semantic change (Dubossarsky et al., 2017). Hence, automatic change detection methods should not be trusted until they have been evaluated thoroughly and their predictions have been shown to reach an acceptable level of correctness.

Despite the rapid growth of LSCD as a field, a solid evaluation of the wealth of proposed models was still missing at the onset of this thesis in 2017, as acknowledged by various authors (Lau et al., 2012; Cook et al., 2014; Frermann and Lapata, 2016). The reasons were multiple, but most importantly there was no annotated benchmark test set available. This thesis is thus concerned with the process of providing such an evaluation for LSCD, including

- the definition of the basic concepts and tasks,
- the development and validation of data annotation schemes with humans,
- the annotation of a multilingual benchmark test set,
- the evaluation of computational models on the benchmark, their analysis and improvement, as well as
- an application of the developed methods to showcase their usefulness in the targeted fields (historical semantics and lexicography).

We put an equal focus on the **human** as well as the **computational** part of the evaluation procedure: In order to annotate LSC data sets, we define a human measurement process for **word senses** based on the concept of **semantic proximity** between word uses. This concept is deeply rooted in Blank (1997)'s theory of LSC and thus provides a direct link between our human measurement process and a widely acknowledged theory from historical semantics. The computational measurement processes that we define then try to model each step of the human process and can thus be related to historical semantic theory. In this way, we aim to standardize the field of LSCD assuring relevance of data sets, tasks and models to historical semantics.

We make two important restrictions: (i) We define tasks of LSCD only as the comparison of a word's meaning between **two time periods**. While this simplifies the LSCD problem, it reduces the number of time periods for which data has to be annotated so that we can annotate larger corpus samples and hence more reliably represent the semantic change of target words. Moreover, it reduces the task complexity allowing different model architectures to be applied to it. (ii) We focus on **unsupervised** LSCD models because they completely remove the bottleneck of human input and only very small amounts of annotated data were available at the onset of this thesis.

1.1 Outline

The remainder of this dissertation is organized as follows:

Chapter 2 provides some general background on the research in LSC. We introduce the view on word meaning, senses and change which we will assume in this thesis. This view will build the basis for human annotation in Chapter 3. Next, we describe the state of research on LSCD as encountered at the beginning of this thesis in 2017.

In **Chapter 3**, we describe the human annotation process for the benchmark data set and analyze its reliability.

In **Chapter 4**, we describe the computational model architectures used to detect LSC. All of these are token- or type-based Distributional Semantic Models exploiting the distributional hypothesis.

In **Chapter 5**, we define the basic tasks which LSCD models should solve based on the annotated data from Chapter 3. Next, we describe the results from three studies evaluating and analyzing token- and type-based model architectures.

In **Chapter 6**, we apply several high-performing models to discover unknown semantic changes of words and evaluate our approach with respect to the usability for historical lexicographers.

Chapter 7 concludes the thesis. We reflect what we learned about LSCD, review the points left open and discuss implications for other research fields.

1.2 Publications

This thesis is based on the following publications:

- Schlechtweg, D., Schulte im Walde, S., and Eckmann, S. (2018). Diachronic Usage Relatedness (DURel): A framework for the annotation of lexical semantic change. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 169–174, New Orleans, Louisiana

  Chapter 3 is partly based on this publication, which is a result of joint work with my supervisor Sabine Schulte im Walde and Stefanie Eckmann. The underlying ideas and the design of the annotation study came from discussions between Sabine and me. Further, my contribution was the implementation of the annotation study, the analysis of the data and the writing of the paper. My co-authors gave feedback on the paper draft. Stefanie helped with the preparation of the study and with the annotation.

- Hätty, A., Schlechtweg, D., and Schulte im Walde, S. (2019). SURel: A gold standard for incorporating meaning shifts into term extraction. In *Proceedings of the 8th Joint Conference on Lexical and Computational Semantics*, pages 1–8, Minneapolis, MN, USA

  Chapter 3 is partly based on this publication, which is a result of joint work with Anna Hätty and Sabine Schulte im Walde. The underlying ideas and the design of the annotation study came from discussions between Anna, Sabine and me. Further, my contribution was the implementation of parts of the annotation study and the analysis of the data.

- Schlechtweg, D., Hätty, A., del Tredici, M., and Schulte im Walde, S. (2019a). A Wind of Change: Detecting and Evaluating Lexical Semantic Change across Times and Domains. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 732–746, Florence, Italy. Association for Computational Linguistics

  Chapter 4 and Chapter 5 are partly based on this publication resulting from joint work with Anna Hätty, Marco del Tredici and Sabine Schulte im Walde. The underlying ideas and the design of the experiments were developed in discussions between all authors. Anna and I took care of the implementation of the models, the experiments, the analysis of the data and the writing of the paper. Marco and Sabine gave feedback on the paper draft. Marco further contributed an implementation of the Vector Initialization approach.

- Schlechtweg, D., McGillivray, B., Hengchen, S., Dubossarsky, H., and Tahmasebi, N. (2020). SemEval-2020 Task 1: Unsupervised Lexical Semantic Change Detection. In *Proceedings of the 14th International Workshop on Semantic Evaluation*, Barcelona, Spain. Association for Computational Linguistics

  Schlechtweg, D., Tahmasebi, N., Hengchen, S., Dubossarsky, H., and McGillivray, B. (2021b). DWUG: A large Resource of Diachronic Word Usage Graphs in Four Languages. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7079–7091, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics

  Chapter 3 and Chapter 5 are partly based on these closely related publications, which are both joint work with Barbara McGillivray, Simon Hengchen, Haim Dubossarsky and Nina Tahmasebi. The idea to organize a shared task resonated from discussions between Barbara, Simon and me. The annotation approach for German, English and Swedish as well as the clustering approach was based on my ideas and refined in discussions with Nina. I implemented the annotation study for these languages with input on the selection of target words by Simon and Haim for English and by Nina for Swedish. The Swedish annotators were recruited by Nina, who also took care of administrative issues regarding the Swedish annotators. Barbara organized the Latin annotation. I took care of the data preparation and publication of the English, German and Swedish data, helped with the Latin data, set up the codalab competition and implemented the baselines for the task. The task design was developed in discussions between all authors. All authors participated in the analysis of the task results and in the writing of both papers.

- Laicher, S., Kurtyigit, S., Schlechtweg, D., Kuhn, J., and Schulte im Walde, S. (2021). Explaining and Improving BERT Performance on Lexical Semantic Change Detection. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 192–202, Online. Association for Computational Linguistics

  Parts of the model descriptions in Chapter 4 and some results in Chapter 5 are based on this publication, which resulted from Severin Laicher's bachelor thesis (Laicher, 2021) supervised by Sabine Schulte im Walde and me. I contributed the basic idea of the study to analyze BERT clusterings according to bias variables. The experimental design was developed in discussions between Severin, Sabine and me. The experiments were implemented by Severin and Sinan Kurtyigit, where all results and next steps were discussed with me. Severin, Sinan and I wrote the paper with feedback from Sabine. Jonas Kuhn had an advisory role.

- Kurtyigit, S., Park, M., Schlechtweg, D., Kuhn, J., and Schulte im Walde, S. (2021). Lexical Semantic Change Discovery. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Online. Association for Computational Linguistics

  Parts of the model descriptions in Chapter 4 and the entire Chapter 6 are based on this publication, which resulted from Sinan Kurtyigit's bachelor thesis (Kurtyigit, 2021) supervised by me. All models and experiments were implemented by Sinan. I contributed the basic idea of the study, the design of the experimental setup, advice on data analysis and participated in the writing of the paper. All results and next steps were closely discussed between Sinan and me. Maike Park contributed the lexicographical analysis. The paper was written by Sinan, Maike and me with feedback from Sabine Schulte im Walde. Jonas Kuhn had an advisory role.

The final form of this dissertation is also a result of other peer-reviewed articles published in the course of my doctoral studies. Although they are not the core parts of it, the thesis refers to them when addressing some less essential concepts:

- Alatrash, R., Schlechtweg, D., Kuhn, J., and Schulte im Walde, S. (2020). CCOHA: Clean Corpus of Historical American English. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6958–6966, Marseille, France. European Language Resources Association

  In this publication, we describe the creation of a clean version of the Corpus of Historical American English (Davies, 2012), which we use in Section 3.2.1.1.

- Baldissin, G., Schlechtweg, D., and Schulte im Walde, S. (2022). DiaWUG: A Dataset for Diatopic Lexical Semantic Variation in Spanish. In *Proceedings of* *the 13th Language Resources and Evaluation Conference*, Marseille, France. European Language Resources Association

This publication describes the application of the annotation framework we develop in Section 3.1 to an onomasiological setting, using data from multiple Spanish varieties.

- Bott, T., Schlechtweg, D., and Schulte im Walde, S. (2021). More than just Frequency? Demasking Unsupervised Hypernymy Prediction Methods. In *Proceedings of the Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Findings)*, Online. Association for Computational Linguistics

  We demonstrate that the predictions of several hypernymy detection models, including a model based on vector entropy (see Section 4.2.3.2), are highly correlated with frequency-based predictions.

- Dubossarsky, H., Hengchen, S., Tahmasebi, N., and Schlechtweg, D. (2019). Time-Out: Temporal Referencing for Robust Modeling of Lexical Semantic Change. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 457–470, Florence, Italy. Association for Computational Linguistics

  We compare OP to WI alignment for type-based embeddings (both described in Section 4.2.2) and find that the latter introduces a lower level of noise to word representations.

- Hätty, A., Schlechtweg, D., Dorna, M., and Schulte im Walde, S. (2020). Predicting Degrees of Technicality in Automatic Terminology Extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Seattle, Washington. Association for Computational Linguistics

  In this publication, we apply LSCD methods to the task of term extraction.

- Hengchen, S., Tahmasebi, N., Schlechtweg, D., and Dubossarsky, H. (2021). Challenges for Computational Lexical Semantic Change. In Tahmasebi, N., Borin, L., Jatowt, A., Xu, Y., and Hengchen, S., editors, *Computational Approaches to Semantic Change*, volume Language Variation, chapter 11. Language Science Press, Berlin

In this paper, we describe the most important challenges for LSCD and outline future directions of research.

- Kaiser, J., Schlechtweg, D., Papay, S., and Schulte im Walde, S. (2020a). IMS at SemEval-2020 Task 1: How low can you go? Dimensionality in Lexical Semantic Change Detection. In *Proceedings of the 14th International Workshop on Semantic Evaluation*, Barcelona, Spain. Association for Computational Linguistics

  We describe the results of our system for the SemEval shared task (see Section 5.2), which applies the SGNS+VI+CD model architecture (see Section 4.2).

- Kaiser, J., Schlechtweg, D., and Schulte im Walde, S. (2020b). OP-IMS @ DIACR-Ita: Back to the Roots: SGNS+OP+CD still rocks Semantic Change Detection. In Basile, V., Croce, D., Di Maro, M., and Passaro, L. C., editors, *Proceedings of the 7th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2020)*, Online. CEUR.org

  We present the results of our participation in the DIACR-Ita shared task on Italian data (Basile et al., 2020), using the SGNS+OP+CD model architecture (see Section 4.2). We win the task with high performance.

- Kaiser, J., Kurtyigit, S., Kotchourko, S., and Schlechtweg, D. (2021). Effects of pre- and post-processing on type-based embeddings in lexical semantic change detection. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 125–137, Online. Association for Computational Linguistics

  This paper describes experiments with preprocessing and postprocessing techniques for type-based LSCD models.

- Laicher, S., Baldissin, G., Castaneda, E., Schlechtweg, D., and Schulte im Walde, S. (2020). CL-IMS @ DIACR-Ita: Volente o Nolente: BERT does not outperform SGNS on Semantic Change Detection. In Basile, V., Croce, D., Di Maro, M., and Passaro, L. C., editors, *Proceedings of the 7th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2020)*, Online. CEUR.org

  In this publication, we present the results of our second participation in the DIACR-Ita shared task, using the BERT+APD model (see Section 4.1).

- Schlechtweg, D., Eckmann, S., Santus, E., Schulte im Walde, S., and Hole, D. (2017). German in flux: Detecting metaphoric change via word entropy. In *Proceedings of the 21st Conference on Computational Natural Language Learning*, pages 354–367, Vancouver, Canada

  This publication describes how we explore vector entropy to detect metaphoric change.

- Schlechtweg, D. and Schulte im Walde, S. (2018). Distribution-based prediction of the degree of grammaticalization for German prepositions. In Cuskley, C., Flaherty, M., Little, H., McCrohon, L., Ravignani, A., and Verhoef, T., editors, *The Evolution of Language: Proceedings of the 12th International Conference (EVOLANGXII)*. Online at http://evolang.org/torun/proceedings/papertemplate.html?p=169

  We test the hypothesis that the degree of grammaticalization of German prepositions correlates with their corpus-based contextual dispersion measured by vector entropy.

- Schlechtweg, D., Oguz, C., and Schulte im Walde, S. (2019b). Second-order co-occurrence sensitivity of skip-gram with negative sampling. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 24–30, Florence, Italy. Association for Computational Linguistics

  In this paper, we show that the type-based word representation model SGNS (see Section 4.2) captures second-order co-occurrence relationships between words.

- Schlechtweg, D. and Schulte im Walde, S. (2020). Simulating Lexical Semantic Change from Sense-Annotated Data. In Ravignani, A., Barbieri, C., Martins, M., Flaherty, M., Jadoul, Y., Lattenkamp, E., Little, H., Mudd, K., and Verhoef, T., editors, *The Evolution of Language: Proceedings of the 13th International Conference (EvoLang13)*

  In this publication, we present a novel procedure to simulate LSC from synchronic sense-annotated data, using the change measures defined in Section 3.1.5, and demonstrate its usefulness for evaluating LSCD models.

- Schlechtweg, D., Castaneda, E., Kuhn, J., and Schulte im Walde, S. (2021a). Modeling sense structure in word usage graphs with the weighted stochastic block model. In *Proceedings of *SEM 2021: The Tenth Joint Conference on Lexical and Computational Semantics*, pages 241–251, Online. Association for Computational Linguistics

  We cluster a subset of the WUGs annotated for SemEval (see Section 3.2.1) with a probabilistic graph clustering technique, enabling us to rigorously compare models of word senses with respect to their fit to the data.

- Shwartz, V., Santus, E., and Schlechtweg, D. (2017). Hypernyms under siege: Linguistically-motivated artillery for hypernymy detection. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, Valencia, Spain*, pages 65–75

  We investigate an extensive number of unsupervised Vector Space Models for hypernymy detection and compare these to state-of-the-art supervised models. Some of the model components are used in Section 4.2.

- Zamora-Reina, F. D., Bravo-Marquez, F., and Schlechtweg, D. (2022). LSCDiscovery: A shared task on semantic change discovery and detection in Spanish. In *Proceedings of the 3rd International Workshop on Computational Approaches to Historical Language Change*, Dublin, Ireland. Association for Computational Linguistics

  In this paper, we present the results of the first LSCD shared task on Spanish data. The data is annotated using the procedure described in Section 3.1 and the task definitions are based on the ones developed in Chapter 5 and Chapter 6.

Chapter 2

Background

In this chapter, we describe the state of the relevant research fields at the onset of this thesis.[1] We start with a historical overview of LSC research pointing out the central role of the concept of **word senses**. We then introduce Blank (1997)'s theory including his definition of LSC based on word senses as it provides the theoretical backbone of this thesis. We continue by describing the lexicographic measurement process of word senses and point out how this process relates to Blank's concept of **semantic proximity**. We show how similar concepts to semantic proximity have been used in practical computational linguistic annotation studies and how such annotations can be used to derive word senses on a human-annotated graph with automatic **clustering** methods. We then give an overview of computational approaches to measure word meaning and meaning changes and finish the chapter with an overview and criticism of evaluation practices.

2.1 Lexical Semantic Change

Research on LSC starts with rhetorics in classical antiquity (Ullmann, 1957, p. 203), where different degrees of habitualization of the literary tropes, including **metaphor**, **metonymy** and **synecdoche**, were recognized (Blank, 1997, p. 9). Cicero realized that rhetorical metaphors are more likely to spread in cases where there is no word for the expressed meaning (p. 9). These observations laid out the basis for the subsequent scientific study of LSC which found its zenith in the 19th and 20th century (p. 8). Reisig (1972, p. 21ff.) observes that LSC is not ran-

[1] Please find an overview of recent developments in LSCD in the major reviews of the field (Kutuzov et al., 2018; Tahmasebi et al., 2021; Hengchen et al., 2021).

dom: he claims that two completely opposite meanings cannot be derived from each other and concludes that there must be a **semantic relation** (association) between old and new meanings (Blank, 1997, p. 10). For Reisig, such relations cover synecdoche (part-whole relation), metonymy (partly) and metaphor. The great rise of historical semantics came with Bréal (1899) providing the first large scientific study on LSC. He aimed to formulate universal **laws of semantic change** similar to the Neogrammarian's laws of sound change (Blank, p. 14). Importantly, he identifies **polysemy** (multiple **senses** of a word) as consequence of LSC and describes the **disambiguation** of polysemy within **uses** of a word (p. 14). Several historical semanticists distinguish between **types** of semantic change based the semantic relations or the levels of meaning involved. Paul (1975, p. 94f.) stresses the central role of metaphor in semantic change while Lehmann (1884) describes changes on the evaluative meaning level, i.e., amelioration and pejoration (Blank, pp. 11, 13).

With de Saussure's ideas giving rise to structural linguistics (de Saussure, 1968) and the influence of Bloomfield (1935)'s "meaning-hostile" linguistic theory (Blank, p. 8), the study of historical semantics was widely neglected in the middle of the 20th century (Geeraerts, 1992, p. 257f.). Structural linguistics had a focus on synchronic language description and its method of analysis was quite complex, and Bloomfield regarded the statement of meaning as the "weak point in language study" (Bloomfield, 1935, p. 140; Blank, p. 24).

A new impulse to historical semantics came from cognitive linguistics in the 1980s (Geeraerts, 1983; cf. Blank, p. 31). In this tradition, researchers try to explain LSC from general cognitive principles such as analogy, association, categorization, etc. (Györi, 2002). Most important is Blank's theory, which has gained wide acceptance (cf. Grzega, 2002; Geeraerts, 2010), captures the canonical examples of semantic change discussed in the historical semantics literature and defines criteria to distinguish word use meanings based on the notion of **semantic proximity** that can be exploited for human word sense annotation (cf. Soares da Silva, 1992; Brown, 2008; Erk et al., 2013).

The last major development in LSC research is the application of computational methods for detection of LSC (Tahmasebi et al., 2021). Around 2010, computational linguists started using computational models of word meaning to automatically detect LSC in text corpora (Sagi et al., 2009; Cook and Stevenson, 2010). This has led to an enormous upsurge in research on LSCD in recent years (Tahmasebi et al., 2021).

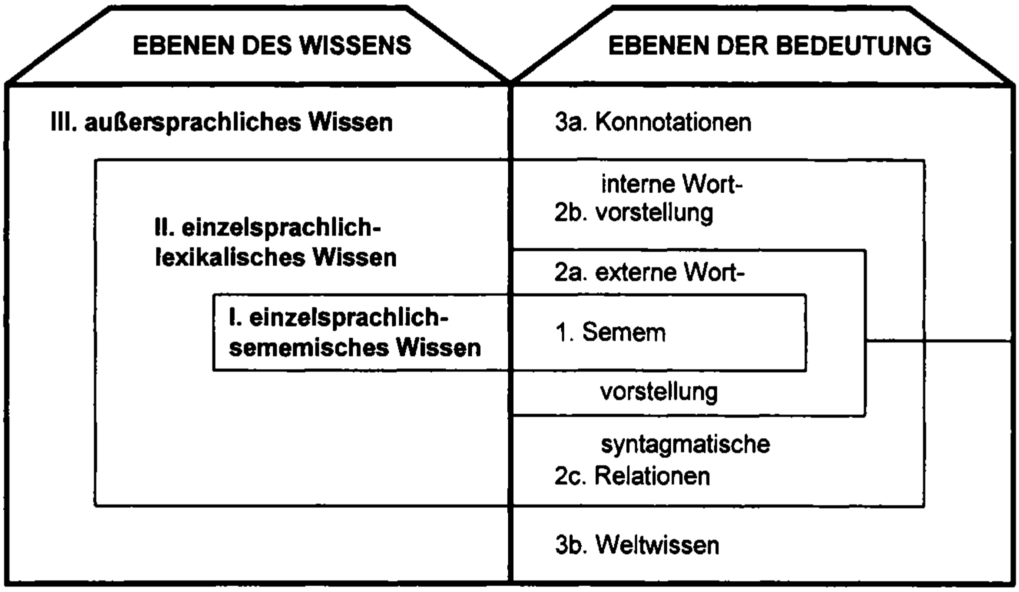

EBENEN DES WISSENS
EBENEN DER BEDEUTUNG
III. außersprachliches Wissen
II. einzelsprachlich-lexikalisches Wissen
I. einzelsprachlich-sememisches Wissen
3a. Konnotationen
2b. interne Wortvorstellung
2a. externe Wortvorstellung
1. Semem
2c. syntagmatische Relationen
3b. Weltwissen

Figure 2.1: Blank's levels of meaning (p. 95).

2.2 Blank's Theory

Blank develops a theory of meaning based on cognitive semantics where polysemy is the synchronic, observable result of LSC. It integrates both traditional and modern approaches to historical semantics and is developed along the lines of a variety of examples of LSC cited in the previous literature (pp. 1, 6).

2.2.1 Levels of Meaning

Blank considers the meaning of a word to be knowledge that humans have about that word (pp. 54ff., 94–96). He identifies three types of "meaning-like" knowledge (see Figure 2.1, left): (i) language-specific semantic, (ii) language-specific lexical and (iii) language-external knowledge. Based on these types, Blank then defines three more fine-grained levels of meaning (see Figure 2.1, right) as "purely linguistic abstractions" over these levels of knowledge for the sake of analysis (pp. 92, 94). Meaning level 1 comprises core semantic knowledge needed to distinguish different **senses** from each other (p. 55). This knowledge corresponds to a set of minimal language-specific semantic attributes, often called **sememe** in structural semantics. From these follow the hierarchical lexical semantic relations between words, e.g. synonymy or hypernymy (p. 56). Meaning level 2 comprises (a) knowledge about the word's rules of use (regional, social, stylistic or diachronic variety), (b) the word's role in the lexicon (part-of-speech, word family or knowledge about polysemy/multiple meanings and semantic relations between meanings) and (c) syntagmatic knowledge such as selectional restrictions, phraseologisms or collocations. Meaning level 3 comprises knowledge about (a) connotation and (b) general knowledge about the world. Here is where Blank locates mental concepts, as we will see in Section 2.2.1.1.

Blank assumes that the knowledge from these three levels is stored in the mind of speakers, which can also change historically in these three levels. An example for change on level 1 is Latin *pipio* 'young bird' > 'young pigeon' which gained the attribute [pigeon-like] (pp. 106–107).[2] Further examples are English *mouse* 'animal' > 'computer mouse' (p. 150) and German *Presse* 'press machine' > 'news press' showing that old and new attributes can be very different (p. 107). Changes on meaning level 1 are often accompanied by changes on the other meaning levels (p. 112), as e.g. in the case of *mouse*, where the new meaning is used in a more specific technical context (level 2a) and the concept of the referent changes (level 3). It is also possible that changes on level 2 and 3 occur independently from the other levels. An example for such an independent change on level 2a is *gota* 'cheek', which changes from being used commonly in Old Italian to being used exclusively in the literary-poetic register in New Italian (p. 107).

2.2.1.1 Signs & Concepts

Blank combines the levels of meaning from Figure 2.1 with Raible (1983)'s model of the semiotic process to describe how the levels of meaning are used in language production (pp. 96–102). This combination is shown in Figure 2.2 (cf. also pp. 99–101): The process starts from a speaker uttering a word (*Konkrete Lautung*), which is the realization of an abstract phonological representation of that word in the speaker's mind (*Zeichenausdruck*). This representation then evokes the corresponding sememe (*Zeicheninhalt*). The **sign** (*Zeichen*) comprises lexical knowledge about the phonological representation and the sememe. It expresses a particular **concept** (*Designat*), which is used to refer to a concrete referent (pp. 99–100, 148).[3] The se-

[2] We will use the notation sense1 > sense2 to mean that sense2 developed from sense1 (see also Section 2.2.2.2).

[3] It is unclear whether Blank assumes a 1-to-1 relationship between signs and concepts (cf .p. 100).

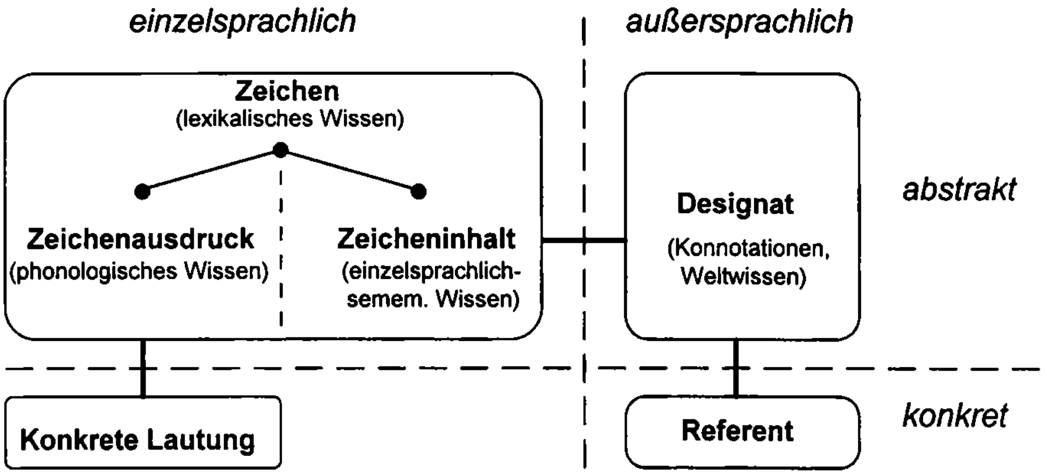


Figure 2.2: Blank's model of the semiotic process (p. 102) derived from Raible (1983).

meme corresponds to meaning level 1 while the sign corresponds to meaning levels 1 and 2 and the concept corresponds to meaning level 3 (cf. Figure 2.1).

Each sign expresses a particular concept, which is used to pick out a referent. The class of referents for which a speaker would use a sign is its **extensional category** (p. 78). The corresponding mental concept is **prototypical** in that it summarizes salient features of the referents in this category (pp. 54, 79f., 415; cf. Rosch and Mervis, 1975). These features are in turn used to determine whether a new referent is assigned to that category, i.e., whether a speaker will use the sign to refer to the new referent (pp. 59–60). In order to belong to an extensional category of a sign, a particular referent does not have to correspond completely to the prototypical concept, but can show different degrees of prototypicality (pp. 81, 415).

A word with multiple senses can be represented in this schema by a sign with different sememes corresponding to different concepts (p. 164, see also Figure 2.3).[4] The sign then expresses multiple concepts through multiple sememes, each corresponding to another extensional category.

Blank follows Aristotle in assuming three principles of psychological association between concepts (or signs) based on human perception of the world (p. 144):

[4] This corresponds to Blank's wording (p. 164). But, he is inconsistent in his use of *Zeichen*. He partially seems to assume that each sense corresponds to a different sign (pp. 110, 98: footnote 131). This would also be more consistent with his view that the whole sign expresses one particular concept (pp. 99, 101).

similarity, **contiguity** and **contrast** (Aristoteles, nd).[5] We will refer to them as **semantic relations**. Blank assumes that these relations structure the storage of information about language in the human mind (p. 137) and that they drive innovative LSC (see Section 2.2.2.2). This means that two concepts sharing one of the described relations are more strongly connected than concepts not sharing such a relation.

A similarity relation holds if the two compared concepts have a "perceptual, functional or intersubjectively experienced commonness" (p. 160). In some cases, this commonness can be described as a concrete overlap of features, as in the case of *mouse* 'animal' and 'computer mouse' from above, explained in more detail in Section 2.2.2.2. This is not easily possible if there are abstract concepts involved, as in German *umwälzen* 'to turn around' > 'to change something radically'. The similarity relation is the basis for the literary trope **metaphor**.

A contiguity relation holds between two concepts belonging to the same field of knowledge, which is a grouping of knowledge perceived to belong together in our cognition (p. 237). Blank equates these fields of knowledge with **semantic frames** (Minsky, 1975; Fillmore, 1975). A contiguity relation holds, for example, between the two concepts 'press machine' and 'news press' of German *Presse*. Both concepts are part of the frame 'news media'. The contiguity relation is the basis for the literary trope **metonymy**.

A contrast relation holds between two opposite concepts (pp. 217f.). It holds, for example, between the two concepts of Italian *famigerato* 'famous' > 'infamous' (p. 220), which have a contrast in connotation. The oppositeness of concepts can lead to lexical antonomy, but not necessarily (p. 219). The contrast relation is the basis for the literary trope **antiphrasis**.

2.2.2 Lexical Semantic Change

2.2.2.1 A Definition

For Blank, only changes on the core meaning level 1 (sememe) are meaning changes in the narrow sense as they can affect all other levels of meaning (p. 112). Only these changes in the meaning of a word lead to the emergence or loss of a full-fledged new sense, i.e., a new sememe. Based on this criterion, meaning changes must be either of the two main types:

[5] Besides associations between concepts, Blank sees associations between signs, sememes and phonological representations to be relevant for semantic change (p. 155).

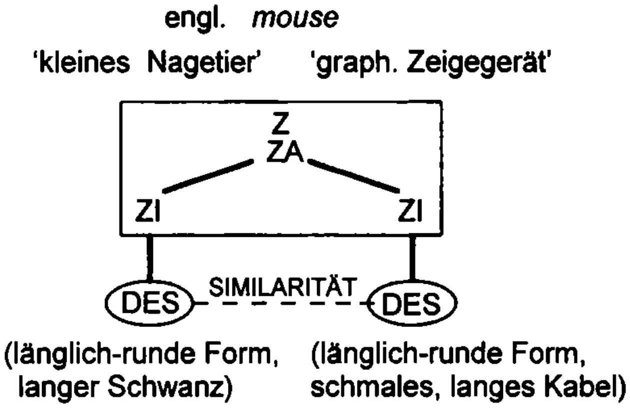

engl. *mouse*
'kleines Nagetier'
'graph. Zeigegerät'
Z
ZA
ZI
ZI
DES
SIMILARITÄT
DES
(länglich-runde Form, langer Schwanz)
(länglich-runde Form, schmales, langes Kabel)

Figure 2.3: Blank's example of concept similarity for *mouse* (p. 151).

- **innovative meaning change**: emergence of a full-fledged additional sense of a word; old and new sense are related by polysemy,
- **reductive meaning change**: loss of a full-fledged sense of a word.

The changes described above for *pipio*, *mouse*, *Presse* and *famigerato* are all innovative meaning changes. An example of reductive meaning change is the German word *Zufall* 'coincidence', 'seizure' losing the sense 'seizure' (Osman, 1971).

2.2.2.2 The Process

Innovative meaning change begins in language discourse by a speaker using a word to refer to a referent for which the word had not been used before (p. 119f.). Such a **semantic innovation** can then either be neglected or adopted by other speakers. If it is adopted, the innovation can become a discourse rule within a particular discourse tradition, which can further be adopted by all speakers of a language or language variety. Blank calls the latter process the **lexicalization** of a semantic innovation, which leads to a new meaning in the mind of speakers and hence to LSC.

As indicated in Section 2.2.1.1, innovations typically occur because of psychological associations in the mind of the speaker between concepts or signs (p. 138). If a new referent (thing or circumstance) should be named, the speaker associates the concept or sign he has gained for this referent with another concept or sign for which he already knows a name. Consider Blank's explanation for the process of change of *mouse* from Section 2.2.1 (p. 150), as depicted in Figure 2.3: the inventor of the computer mouse had a new device in front of him as well as a mental concept for that device including formal properties such as round form, long and thin (and possibly gray) cable. He then associated this new concept with the concept of the animal mouse and started using the word for the animal also for the computer device. We can further speculate that this usage was adopted as discourse rule in a technical context and later, with increased importance of computers in everyday life, lexicalized within the English speaker community. The old and new sense of *mouse* are exemplified by the uses of the word in (2.1) and (2.2). Before the invention of the computer mouse in the 1960s, the word was used as in 2.1. From roughly the 1980s, we then find uses in the computer sense, as displayed in (2.2):

(2.1) [. . .] thought I heard a **mouse** or rat stir in a far corner of the room [. . .] [6]

(2.2) [. . .] does the software require special computer accessories (**mouse**, joystick, sound card, colour monitor, printer)? [7]

The basis for association in the case of *mouse* is the similarity relation (see Section 2.2.1.1). Blank refers to a change based on this relation as **metaphoric change**. Changes based on the contiguity relation are instead referred to as **metonymic change**. According to Blank, these are the two most important types of innovative meaning change (cf. pp. 157ff.).

An example of metonymic change is German *Presse* from Section 2.2.1. Before roughly 1800, *Presse* was mainly used in the sense 'press machine', as in (2.3). After 1800, it gained the additional sense 'news press' as in (2.4):

(2.3) Der zweyte Theil vom Bauernrechte ist schon lange aus der **Presse**; [8]
*'The second part of Farmers' Rights already left the **press**;'*

(2.4) Alle Freiheiten suspendirt! die persönliche Freiheit wie die der **Presse**! [9]
*'All freedoms suspended! the personal freedom as well as the one of the **press**!'*

The press machine used to produce a newspaper and the collectivity of people writing and publishing it are part of the same cognitive frame, which means that a contiguity relation holds between them.

[6] Styron, W.: Set This House on Fire. 1960.

[7] Keep on hackin': kids and computers. Todays Parent. Vol. 10, Iss. 2. 1993.

[8] Rabener, G. W.: Sammlung satyrischer Schriften. Vol. 3. Leipzig, 1752.

[9] Neue Rheinische Zeitung. No. 30. Cologne, June 30, 1848.

Reductive meaning change occurs when a sense of a word becomes unusual (p. 121f.). The sense can either be lost directly or, as with innovation, remain to be used within a discourse tradition or language variety. Usually the older senses of a word are lost (p. 129). As an example, consider the German word *Zufall* from Section 2.2.2.1. It had two senses until around 1850, 'seizure' and 'coincidence', as in (2.5). From 1850, the word occurs less and less often in the former sense until it is exclusively used in the sense 'coincidence', as in (2.6):

(2.5) [. . .] daß sie aus Angst **Zufälle** bekommen und wieder gefährlich krank werden würde. [10]
*'[. . .] that she would have **seizures** out of fear and become dangerously ill again.'*

(2.6) Es muß verbrannt sein, vielleicht durch **Zufall**, vielleicht durch Feindeshand. [11]
*'It must have burned, perhaps by **coincidence**, perhaps by enemy hands.'*

2.2.3 Summary

Blank identifies three levels of meaning: Level 1 comprises the core meaning of a word distinguishing different senses from each other. Level 2 comprises knowledge about the word's rules of use, its role in the lexicon and syntagmatic knowledge. Level 1 and level 2 form the lexical knowledge of a word. Level 3 comprises connotational and world knowledge forming the conceptual knowledge about a word. Concepts are related by psychological associations driving innovative semantic change.

For Blank, only changes on meaning level 1 are actual meaning changes leading to a new sense or loss of an old sense. Correspondingly, he differentiates between innovative and reductive meaning change. Both types imply a change in the polysemy of a word. In order to be regarded as LSC, such changes should be spread widely within a speaker community.

The aim of this thesis is to define measurement processes for LSC. This implies that we need to measure changes of human knowledge on meaning level 1 over time, i.e., we need to measure whether senses were gained or lost. This can be done by measuring a word's senses at two time periods and comparing the results. However, senses (being human knowledge) are not directly observable and thus cannot

[10] Jung-Stilling, J. H. Lebensgeschichte. Stuttgart, 1835.

[11] von Vischer, F. T.: Auch Einer. Eine Reisebekanntschaft. Vol. 1. Stuttgart i.a., 1879.

be measured directly. Instead, **speaker behavior** in the form of produced language (the **usage** of language) is usually analyzed in linguistics as a proxy for speaker knowledge (Chomsky, 1986). In the next section, we will describe approaches to measure word senses based on language use. For this, we will also come back to Blank's own proposal.

2.3 Human Measurement of Word Senses

2.3.1 The Lexicographic Process

The task of measuring word senses is the main goal of the field of lexicography. In modern lexicography word senses are measured as abstractions over the **patterns of usage** of a particular word (Kilgarriff, 2007), where the **usage** of a word is commonly measured as a set of occurrences of that word within a sentence in text form, such as (2.1) or (2.2) from Section 2.2.2.2. Each such occurrence will be called a **word use** and a set of uses will be called a **corpus**. The process of abstracting over the patterns of usage is generally done by a human lexicographer proceeding in the following steps (Kilgarriff, 2007):

1. gather a corpus of uses for a target word,
2. divide the uses into clusters; optimally, all the members of each cluster have more in common with each other than with any member of any other cluster,
3. for each cluster, work out what it is that makes its members belong together and
4. use these conclusions to create a dictionary definition.

We are only interested in steps 1 and 2 of this process (**corpus creation** and **clustering**) as they will provide the information we need to measure LSC, i.e., whether there is a lost or gained sense (see Sections 2.2.3 and 3.1.4).

Step 1 requires us to gather uses for each target word. In order to reflect the target word's usage within a speaker community, these should be gathered from a wide variety of speakers, discourse types, genres, etc. (see Section 2.2.2.2). For our purposes, uses should also come from different time periods. We will describe the creation of our corpora in detail in Chapter 3.

In step 2, the lexicographer gains an interpretation of each word use (**use meaning**) with his speaker (and possibly further) knowledge and determines how much it has **in common with** the other uses based on his interpretations.[12] He then forms **clusters of uses** based on his judgments of commonness between uses. Each cluster formed in this way corresponds to one **lexicographic word sense** and uses from the same cluster are assumed to express the same sense. We will often refer to step 2 as the **lexicographic clustering process**.

The lexicographic clustering process poses multiple problems: (i) The criteria applied for clustering are implicit and may vary between lexicographers (Kilgarriff, 1997, 2007). Hence, we do not know how well word senses derived in this way would reflect speaker knowledge (see Section 2.2). It is thus important to explicitly specify criteria in advance. These criteria should allow us to argue that the derived senses reflect the speakers' (who produced the corpus) knowledge about word senses. (ii) Scanning and comparing all word uses is time-consuming. With more efficient processes we may analyze more data in the same time. (iii) Lexicographers are specialists and thus expensive and rare. Hence, we want the task to be done by less specialized people.

Approaches from lexical semantics try to solve some of these problems with more controlled **word sense annotation** procedures. They usually do not reproduce the full lexicographic process, but either (i) exploit predefined dictionary definitions, or (ii) try to define and annotate what uses "have in common" (see step 2 above) and to obtain clusters from these annotations. We will now shortly discuss each of these two approaches.

2.3.2 Sense Definition Annotation

In this approach, human annotators are presented with predefined dictionary sense definitions and asked to assign either a single best sense per word use (Weaver, 1955; Navigli, 2009) or graded judgments between words uses and sense definitions (Erk et al., 2009, 2013). This approach requires two steps: (i) the lexicographic process to create dictionary definitions, and (ii) an annotation step where annotators assign uses to the definitions. This annotation can be done by non-lexicographers and requires each annotator to compare each word use only to a small number of sense definitions (instead of comparing each word use to each other use). However, it still relies on a full run through the lexicographic process and often leaves the criteria used to create the sense definitions implicit. There has been extensive work on sense definition annotation and several large-scale annotation projects have been

[12] A use meaning best corresponds to sign and concept in Figure 2.2.

↑	
	Identity
	Context Variance
	Polysemy
	Homonymy

Table 2.1: Blank's scale of semantic proximity.

carried out, as e.g. SemCor and OntoNotes (Langone et al., 2004; Hovy et al., 2006), but only a small fraction has a diachronic focus (Bamman and Crane, 2011; Lau et al., 2012; Cook et al., 2014; Tahmasebi and Risse, 2017; Schlechtweg et al., 2017).

2.3.3 Use Pair Proximity Annotation

In this approach, human annotators are asked to judge word use pairs for their **semantic proximity**, sometimes referred to as similarity or relatedness in meaning (Soares da Silva, 1992; Blank, 1997; Brown, 2008; Erk et al., 2009, 2013). All of these can be seen as attempts to measure how much uses **have in common** (see Section 2.3.1) based on some more or less specified criteria. Erk et al. (2013)'s in-depth study reveals a reasonably high inter-annotator agreement as well as a strong correlation with sense definition annotation and annotation of multiple lexical paraphrases. The task is much less complex than the full lexicographic process. Hence, the annotation can be done by non-lexicographers (Erk et al., 2013, Blank, 1997, p. 417). The bottleneck of this approach are the pairwise judgments quadratically increasing the number of annotation instances compared to sense definition judgments.

Rather clearly specified criteria for use pair annotation are given by Blank. His approach also has the advantage that it is developed within a theory of LSC and can thus be expected to cover the canonical examples of LSC.

2.3.3.1 Blank's Concept of Semantic Proximity

Within his theory described in Section 2.2.1, Blank develops a scale of **semantic proximity** of word uses with polysemy located between identity, context variance and homonymy, as depicted in Table 2.1 (pp. 413–418). For each level of the scale, Blank gives criteria derived from Deane (1988) to decide whether a word use pair should be located on that level. The pair (2.7,2.8) is classified as **identical** as the referents of two uses of the word *arm* are both prototypical representatives of the same extensional category (see Section 2.2.1), corresponding to the concept 'a human body part':

(2.7) [. . .] and taking a knife from her pocket, she opened a vein in her little **arm**, and dipping a feather in the blood, wrote something on a piece of white cloth, which was spread before her.

(2.8) [. . .] and though he saw her within reach of his **arm**, yet the light of her eyes seemed as far off [. . .]

The use pair (2.9,2.10) is classified as **context variance** as both referents still belong to the same extensional category, but one is a non-prototypical representative. Hence, there is some variation in meaning, e.g. the arm of a statue loses the function of the physical arm to be lifted:

(2.9) [. . .] and taking a knife from her pocket, she opened a vein in her little **arm**, and dipping a feather in the blood, wrote something on a piece of white cloth, which was spread before her.

(2.10) [. . .] as in "Planet of the Apes," when the disembodied **arm** of the Statue of Liberty jets spectacularly out of the sandy beach.

The use pair (2.11,2.12) would be classified as **polysemy** as the two referents of *arm* belong to different extensional categories, but the corresponding concepts still hold a semantic relation (in this case a similarity relation regarding physical form).

(2.11) [. . .] and taking a knife from her pocket, she opened a vein in her little **arm**, and dipping a feather in the blood, wrote something on a piece of white cloth, which was spread before her.

(2.12) It stood behind a high brick wall, its back windows overlooking an **arm** of the sea which, at low tide, was a black and stinking mud-flat [. . .]

In contrast, the referents of *arm* in the **homonymic** pair (2.13,2.14) belong to different extensional categories and the corresponding concepts do *not* hold a semantic relation:

(2.13) [. . .] and taking a knife from her pocket, she opened a vein in her little **arm**, and dipping a feather in the blood, wrote something on a piece of white cloth, which was spread before her.

(2.14) And those who remained at home had been heavily taxed to pay for the **arms**, ammunition; fortifications, and all the other endless expenses of a war.

We summarize this as follows: Context variance is different from identity in that at least one of the referents is non-prototypical. Polysemy is different from context variance in that the referents do not belong to the same extensional category and there is a semantic relation between the two concepts. Homonymy is different from polysemy in that the two concepts do not hold a semantic relation. The semantic relation can be based on similarity, contiguity or contrast (see Section 2.2.1).[13]

2.3.3.2 Semantic Proximity Clustering

Use pair proximity judgments can be represented in a densely-connected graph and senses can be derived by clustering the nodes (McCarthy et al., 2016).[14] This avoids the need for predefined sense definitions and makes the criteria for the clustering step of the lexicographic process explicit: Use pair proximity judgments measure how much uses **have in common** and the clustering algorithm provides a clearly defined method to aggregate those uses that have more in common with each other than with uses from other clusters.

Blank's concept of semantic proximity (see Section 2.3.3.1) is special in that it provides a concrete cluster criterion, i.e., for whether two word uses have the same sense: the belonging to different extensional categories in combination with the existence of a conceptually motivated relation. If the two word uses are located on level 3 or 4 on the scale in Table 2.1, they should be assigned to the same cluster. If they are located on level 1 or 2, they should be assigned to different clusters. We will exploit this criterion for our clustering approach in Section 3.1.4. Blank's semantic proximity has the further advantage of being conceptually motivated (see Section 2.2.1.1). This allows for a general argument that annotator and speaker knowledge have a correspondence because conceptual structures are similar between speakers of a language (cf. Lakoff and Johnson, 1980).

[13] Note that the criterion used by Blank to distinguish context variance and polysemy (membership in extensional category) is located on meaning level 3 (see Section 2.2.1). However, he defines polysemy as a difference on meaning level 1 (two sememes). Hence, he seems to assume that membership in a different extensional category indicates a different sememe. See also Blank's discussion of the role of sememes and concepts in the process of recognizing referents (pp. 59–60) as well as his discussion of the sememe being part of the concept (p. 100).

[14] There also exist lexical substitution annotation approaches (McCarthy and Navigli, 2009), equally allowing for graph clustering (McCarthy et al., 2016).

However, this approach does not come without its problems: Blank's concept of semantic proximity could be hard to grasp for annotators (cf. Blank, 1997, pp. 417–419). It would then be questionable whether annotator judgments actually reflect the meaning structure assumed by Blank. Moreover, it is generally a problem to infer speaker knowledge from annotator knowledge as we have to rely on a range of assumptions which are very hard to verify. One such assumption is that speaker and annotator have the same use meaning in mind. Another one is that they have the same conceptual organization in their brains. In a historical setting, this is even more of a problem as the conceptual structures in the minds of speakers changes over time (Thagard, 1990).[15] Another, practical, problem is the above-mentioned quadratically increasing number of annotation instances. We will come back to these problems in Chapter 3.

2.4 Lexical Semantic Change Detection

Before we look into computational approaches to the detection of LSC, we need to take a more general look into the computational measurement of word meaning as LSCD models are most often adjustments of these more general models to a diachronic corpus setting.

2.4.1 Computational Measurement of Word Meaning

The most popular family of unsupervised and text-based computational models of word meaning are **Distributional Semantic Models** (Lenci, 2008; Turney and Pantel, 2010). These models assume that a word's meaning can be partly inferred from the contexts it is used in (Harris, 1954). Crucial for our purposes is the distinction into (i) models building one meaning representation for *each* word use (**token-based**) and (ii) models building an aggregated meaning representation *across* a word's uses (**type-based**). Token-based meaning representations provide a model of **use meaning** (see Section 2.3.1) and allow us to model **semantic proximity** (see Section 2.3.3.1) between pairs of word uses by measuring the similarity between their meaning representations (cf. Pilehvar and Camacho-Collados, 2019; Armendariz et al., 2020). This allows to derive **word senses** with a similar procedure to the one described in Section 2.3.3.2. Hence, based on token-based models we can

[15] Note that these problems are not specific to this annotation approach, but generally apply to a number of research fields such as historical linguistics or cognitive science.

model the clustering step of the lexicographic clustering process described in Section 2.3.1. As we will explain further in Section 4.1, this gives us a good argument why such models should be able to measure LSC while this is not easily possible with most of the type-based models.[16]

Most token-based models are **Vector Space Models** (VSMs), i.e., they represent meaning as vectors in a vector space (Turney and Pantel, 2010). Early approaches simply represented each word use as a vector of the frequencies of the co-occurring words (Schütze, 1998) while modern approaches learn vectors as parameters in a language model (Peters et al., 2018; Devlin et al., 2019). Such vectors can then be clustered and clusters can be interpreted as word senses. Such models are commonly applied to solve the task of Word Sense Induction (WSI), which can be seen as a formalization of the lexicographic clustering process (cf. Navigli, 2009).

Type-based models can be distinguished into VSMs, Co-occurrence Graphs (Mihalcea and Tarau, 2004) and Topic Models (Steyvers and Griffiths, 2007; Blei, 2012). Similar to token-based VSMs, early type-based VSM approaches simply represented each word as a vector of the frequencies of the co-occurring words in all contexts (Wilks et al., 1990) or optimizations to these vectors (Landauer and Dumais, 1997) while modern approaches learn vectors as parameters in a language model (Mikolov et al., 2013a; Pennington et al., 2014).

Topic Models infer a probability distribution for each word over different topics (or word senses), which are in turn modeled as a distribution over words. Most Topic Models are probabilistic generative models in the sense that they model the distribution of words over documents in a corpus by assuming a latent topic structure which is inferred from the co-occurrence statistics. In contrast to type-based VSMs, Topic Models do infer a sense-like word representation structure and thus allow a more clear argument why they should be able to measure LSC.

Co-occurrence Graphs represent words as nodes in a graph connected by edges representing their co-occurrence relationships (Mihalcea and Tarau, 2004).[17] Nodes in the graph can be clustered and clusters can be interpreted as word senses. Similar to Topic Models, clustered Co-occurrence Graphs infer a sense-like structure allowing a better argument why they should be able to measure LSC.

[16] Find more detailed explanations of each model type in Chapter 4.

[17] If they are fully connected, their adjacency matrix can be regarded as a VSM.

2.4.2 Computational Measurement of Lexical Semantic Change

Existing approaches for LSCD were nearly exclusively derived from the above-described model types.[18] A standard token-based LSCD approach using clustering to derive word senses similar to Schütze (1998) was completely missing at the onset of this thesis. Only Sagi et al. (2009) use a token-based VSM to infer one representation per word use and then measure changes in dispersion (average distance) within the set of word vectors over time to detect semantic change.

Most other models are type-based: VSMs represent each word with multiple vectors reflecting its co-occurrence statistics at different periods of time (Gulordava and Baroni, 2011; Kim et al., 2014; Basile et al., 2015; Xu and Kemp, 2015; Eger and Mehler, 2016; Hamilton et al., 2016a,b; Hellrich and Hahn, 2016; Rosenfeld and Erk, 2018). LSC is typically measured by the cosine distance (or some alternative similarity metric) between vectors (Salton and McGill, 1983) or by differences in contextual dispersion between the two vectors (Kisselew et al., 2016; Schlechtweg et al., 2017). **Similarity-based** measures can be motivated by the assumption that sense frequency changes (see Section 3.1.5) correlate with changes in the global co-occurrence statistics of a word per time period. **Dispersion-based** measures instead rely on the more specific assumption that sense frequency changes correlate with changes in the predictability of the global co-occurrence statistics (see Section 4.2.3).

Diachronic Topic Models either jointly model corpora from different time periods (Lau et al., 2012; Cook et al., 2014) or explicitly model the change of topics over time (Wang and McCallum, 2006; Wijaya and Yeniterzi, 2011; Frermann and Lapata, 2016). LSC of a word is measured by manually comparing topic densities (Wijaya and Yeniterzi, 2011) or calculating a novelty score for each topic based on their frequency of use and optionally combining this with a relevance score for each sense based on keyword probabilities (Lau et al., 2012; Cook et al., 2014; Frermann and Lapata, 2016). The main aim is to detect words that gained senses.

Mitra et al. (2015) construct a clustered co-occurrence graph for each time period, align clusters across periods and then measure change as gain, loss, splitting or merging of clusters.

Most studies have not been explicit about the underlying concepts which they try to model and how this relates to theories of LSC in historical linguistics. Many studies rely on a vague notion of *degree of LSC* without defining it (Gulordava and Baroni, 2011; Hamilton et al., 2016b; Dubossarsky et al., 2017; Bamler and Mandt,

[18] For a comprehensive overview see Tahmasebi et al. (2021).

2017; Rudolph and Blei, 2018; Rosenfeld and Erk, 2018). This graded notion of LSC seems to diverge from the definition applied in historical linguistics, where LSC is typically not assumed to be graded, but binary (see Section 2.2). That is, either a word gained/lost a sense over time or not while the graded notion seems to imply that slight changes to the frequencies of word senses are also considered as instances of LSC. This deviation is striking as the most straightforward application of LSCD models is their use to aid historical linguists (Hamilton et al., 2016b).

2.4.3 Evaluation

Existing evaluation procedures for LSCD can be distinguished into evaluation on (i) empirically observed data, and (ii) synthetic data or related tasks.

Category (i) includes case studies of individual words (Sagi et al., 2009; Jatowt and Duh, 2014; Hamilton et al., 2016a), stand-alone comparison of a few hand-selected words (Wijaya and Yeniterzi, 2011; Hamilton et al., 2016b; del Tredici and Fernández, 2017), comparison of hand-selected changing vs. semantically stable words (Lau et al., 2012; Cook et al., 2014), and post-hoc evaluation of the predictions of the presented models (Cook and Stevenson, 2010; Kim et al., 2014; Kulkarni et al., 2015; Basile et al., 2015; del Tredici et al., 2016; Eger and Mehler, 2016; Ferrari et al., 2017). Moreover, Gulordava and Baroni (2011) describe a human-annotated data set of 100 English words where annotators were asked to rate the words according to their degree of change in the last 40 years, but without relating these annotations to a corpus. Schlechtweg et al. (2017) evaluate on a small human-annotated data set of metaphoric change in German.

Category (ii) includes studies that simulate LSC (Cook and Stevenson, 2010; Kulkarni et al., 2015; Rosenfeld and Erk, 2018), evaluate sense assignments in WordNet (Mitra et al., 2015; Frermann and Lapata, 2016), identify text creation dates (Mihalcea and Nastase, 2012; Frermann and Lapata, 2016), or predict the log-likelihood of textual data (Frermann and Lapata, 2016).

Overall, the various studies use different evaluation tasks and data, with little overlap. Most evaluation data has not been annotated by humans or does not reflect diachronic corpus data. Models were rarely compared to previously suggested ones, especially if the models differed in meaning representations. Synthetic data sets are problematic because they do not reflect actual diachronic changes. Concepts underlying tasks and experiments are often not clearly defined and related to historical linguistics. Consider the results of two popular LSCD papers in order to better understand why this is problematic: Hamilton et al. (2016b) evaluate various VSM models on 9 English words from previous work and 30 English words yielded by their models' predictions (10 per model) manually assigning LSC labels to words without examining the corpus data directly and based on a vague concept of "semantic shift". Frermann and Lapata (2016) evaluate their Topic Model on a set of tasks of which only two are directly meaning change related: (i) For novel sense detection they evaluate how well their model detects gained target word senses relying on automatic mappings of topics to WordNet. How reliable these automatic mappings are, is not evaluated. (ii) They additionally evaluate on Gulordava and Baroni (2011)'s above-described English data set which does not relate LSC labels to corpus data. Hamilton et al. as well as Frermann and Lapata report a good performance of their models. However, for a user it is unclear which of the models to choose to detect LSC because they have never been directly compared to each other. We do not know whether Hamilton et al.'s performance on their small evaluation data will persist once the model is applied to different (possibly non-English) data. Additionally, some of the reported performances may be irrelevant because the evaluation data on which they were obtained was not validated. It is even unclear whether the evaluation data from both studies reflects the same concept of LSC because they have not defined it or given exact annotation guidelines.

The state of evaluation in the LSCD research field at the onset of this thesis is summarized on point by Tahmasebi et al. (2021):

> When it comes to evaluating methods and systems, there is a general lack of standardized evaluation practices. Different papers use different datasets and testset words, making it difficult or impossible to compare the proposed solutions. Proper evaluation metrics for semantic change detection and temporal analog detection have not been yet established. Furthermore, comparing methods proposed by different groups is difficult due to varying preprocessing details. For example, filtering out infrequent words can impact the results considerably and different papers employ different thresholds for removing rare words (e.g., some filter out words that appear less than 5 times, others less than 200 times).

Similar criticisms are shared by a number of authors (Lau et al., 2012; Cook et al., 2014; Frermann and Lapata, 2016). Starting from this observation, this thesis mainly aims to standardize LSCD evaluation by creating human-annotated, multi-lingual, diachronic data sets, defining standard tasks and evaluation metrics, and identifying high-performing baseline models for evaluation.

Chapter 3

Human Measurement

The question how to detect LSC with computers inevitably presupposes an answer to the question how humans detect it. This is because a standard assumption in computational linguistics is that humans are best in judging the properties of language and that the performance of a computational model should be measured against a human gold standard. In this chapter, we motivate, describe and evaluate our annotation methodology to create such a multilingual human gold standard for LSC. The annotation methodology is based on human semantic proximity judgments of use pairs, which are then represented in a graph and uses are clustered with a principled graph clustering technique, formalizing the lexicographic clustering process described in Section 2.3.1. LSC is (mainly) measured on the inferred clusterings which are equated to senses. The resulting data is then used to evaluate computational models in Chapter 5.

We define two time periods for each data set for which we create two time-specific subcorpora C_1 and C_2 from existing larger corpora (see Sections 3.2.1 and 3.2.2). The tasks we define on this data in Chapter 5 are then consequently also based on the comparison of two time periods. While this simplifies the LSCD problem, it has two main advantages: (i) It reduces the number of time periods for which data has to be annotated so that we can annotate larger corpus samples and hence more reliably represent the sense distributions of target words. (ii) It reduces the task complexity, allowing different model architectures to be applied to it.

3.1 Annotation

For each target word, we aim to derive a label measuring the semantic change between the word's uses from C_1 and C_2. For this, we combine human semantic proximity judgments between use pairs with an automatic clustering procedure and various semantic change measures (recall Section 2.3.3.2).

3.1.1 Word Uses

We can only feasibly annotate a sample U of a word's uses from the corpora C_1 and C_2. (The procedures to sample U will be described in Sections 3.2.1.4 and 3.2.2.4.) This sample will only partly reflect the semantic change which we could have measured if we had annotated all of the word's uses.[1] For instance, we could miss infrequent senses leading to a wrong classification of the word as changing or non-changing. Hence, we consider the change labels measured through the annotation process described in this section as **estimates** of the actual change labels and will try to quantify how much error is introduced through this estimation.

3.1.2 Semantic Proximity

We rely on human semantic proximity annotation to create our data because of the advantages described in Section 2.3, including the direct connection to Blank (1997)'s theory. This annotation approach has been operationalized in various previous studies (Soares da Silva, 1992; Brown, 2008; Erk et al., 2013). These studies do not directly apply Blank's scale, but use less specified proximity, similarity or relatedness scales reflecting very similar ideas. As these scales have been used in previous annotation studies, we can assume that they are implementable and yield sufficient agreement between annotators. For our study, we decide to adopt a relatedness scale similar to Brown's, shown in Table 3.1. This scale reflects the four levels of Blank's scale and the central role of the notion of a semantic relation (see Section 2.2). In this way, we keep the connection to Blank's theory of LSC assuring relevance to historical semantics while adopting a practically tested annotation

[1] Note that also the full corpora only correspond to the limited sample of word uses which was recorded as text and subsequently digitized. Hence, also the full corpora can only approximate a word's usage at a particular point of time. This means that we can never evade the problem of sampling error (cf. Koplenig, 2019).

Identity	4: Identical
Context Variance	3: Closely Related
Polysemy	2: Distantly Related
Homonymy	1: Unrelated

Table 3.1: Blank's scale of semantic proximity (left) and the DURel relatedness scale derived from Blank's scale (right).

scale. We train annotators with adapted guidelines from Erk et al. (2013).[2]

For each target word, the sampled uses U are combined into pairs such as (3.1,3.2) and annotated with their semantic proximity on the DURel scale in Table 3.1:

(3.1) [...] and taking a knife from her pocket, she opened a vein in her little **arm**, and dipping a feather in the blood, wrote something on a piece of white cloth, which was spread before her.

(3.2) It stood behind a high brick wall, its back windows overlooking an **arm** of the sea which, at low tide, was a black and stinking mud-flat [...]

Annotators could also choose the annotation label 'Cannot decide'. We optimize the annotation process in order to reduce the annotation load, as described in Section 3.2.1.5. This has the effect that only a subset of the full set of a word's use pairs are annotated.

3.1.3 Graph Representation

We represent annotated data (semantic proximity judgments of use pairs) in the mathematical structure of a graph which we call **Word Usage Graph** (WUG) in order to relate our data to graph theory and graph clustering. A WUG $\mathbf{G} = (\mathbf{U}, \mathbf{E}, \mathbf{W})$ is a standard weighted, undirected graph where nodes $u \in U$ represent word uses and $W : E \mapsto Z$ where $Z \subset \mathbb{R}$ maps each edge to its weight.[3] Weights $w \in W$ represent the semantic proximity of a pair of uses $(u_1, u_2) \in E$ (cf. McCarthy et al.,

[2] The guidelines are available at https://www.ims.uni-stuttgart.de/data/wugs under DWUG EN/DE/SV and DURel/SURel.

[3] Z is defined to be a subset of the real numbers because we later take the median of annotator judgments as edge weight, which is not always a natural number, and also because we shift weights for clustering having a similar effect.

2016).[4] A WUG G represents the semantic proximity structure underlying a set of word uses U. In principle, the uses in U can be sampled from only one corpus, or from different corpora representing different time periods, genres, authors, language varieties or even languages. In our case, the uses in U were mostly sampled from two different time periods t_1, t_2 corresponding to the subsets $U_1 \subset U$ and $U_2 \subset U$, and comparisons of semantic patterns are performed according to these subsets. Correspondingly, we can define two subgraphs G_1, G_2 containing only uses from U_1 and U_2, respectively, and the edges between them. We also define a subgraph $\mathbf{G_{1,2}} = (\mathbf{U}, \mathbf{E_{1,2}}, \mathbf{W_{1,2}})$ containing all uses U from both time periods, but only those edges (u_1, u_2) and their weights where word uses u_1 and u_2 are from different time periods (COMPARE edges): $E_{1,2} = \{(u_1, u_2)|(u_1, u_2) \in U_1 \times U_2\}$ and $W_{1,2} = [W(u_1, u_2)|(u_1, u_2) \in E_{1,2}]$.

3.1.4 Clustering

As sketched in Section 2.3.3.2, word senses can be derived on a WUG with a graph clustering algorithm. In order to see why clusters derived in such a way should reflect lexicographic sense distinctions, consider once more the lexicographic clustering process from Section 2.3.1: the members of each cluster should **have more in common** with other members of that cluster than with members of other clusters. The semantic proximity between use pairs measures how much meanings of uses have in common: A high semantic proximity indicates that uses have a lot in common, a low proximity indicates that they have little in common. Hence, a clustering procedure assigning uses with high semantic proximity to the same cluster while assigning uses with low semantic proximity into different clusters will generally fulfill the commonness criterion for lexicographic word senses given by Kilgarriff (2007).

The WUGs obtained from the annotation are weighted, undirected, often sparsely observed and noisy (see Section 3.2.1.5). This poses a very specific problem that calls for a robust clustering algorithm. We use a variation of correlation clustering (Bansal et al., 2004) which minimizes the sum of cluster disagreements, i.e., the sum of low edge weights (semantic proximity) within a cluster plus the sum of high edge weights across clusters. For this, we have to choose a threshold h on edge weights deciding which weights will be considered as high and which ones as low. We set $h = 2.5$ as this splits Blank's scale of semantic proximity be-

[4] We use the short form $w \in W$ to refer to all weights in the multiset $W = [W(u_1, u_2)|(u_1, u_2) \in E]$.

tween context variance and polysemy, which is in line with his view that context variance is a variation *within* a sense while polysemy is a relation *between* senses (see Section 2.3.3.1). Consequently, the weight $W(e)$ of each edge $e \in E$ in a WUG $\mathbf{G} = (\mathbf{U}, \mathbf{E}, \mathbf{W})$ is shifted to $W'(e) = W(e) - 2.5$ (e.g. a weight of 4 becomes 1.5). Those edges e with a weight $W'(e) \geq 0$ are referred to as **positive** edges P_E while edges with weights $W'(e) < 0$ are called **negative** edges N_E. Let further $C : U \mapsto L$ be some clustering on U, $\phi_{E,C}$ be the set of positive (high) edges **across** any of the clusters in clustering C and $\psi_{E,C}$ the set of negative (low) edges **within** any of the clusters. We then search for a clustering C that minimizes the sum of weighted cluster disagreements:

$$SWD(C) = \sum_{e \in \phi_{E,C}} W'(e) + \sum_{e \in \psi_{E,C}} |W'(e)| \,.$$

That is, the sum of positive edge weights between clusters and (absolute) negative edge weights within clusters is minimized. Minimizing SWD is a discrete optimization problem which is NP-hard (Bansal et al., 2004). As we have a relatively low number of nodes ($\leq$ 200), we approximate the global optimum with Simulated Annealing (Pincus, 1970), a standard discrete optimization algorithm.[5] In order to reduce the search space, we iterate over different values for the maximum number of clusters ($\leq$ 20). We also iterate over randomly as well as heuristically chosen initial clustering states.[6]

This way of clustering WUGs has several advantages: (i) It finds the optimal number of clusters on its own. (ii) It easily handles missing information (non-observed edges). (iii) It is robust to errors by using the global information on the graph. That is, one wrong judgment can be outweighed by correct ones. (iv) It directly optimizes an intuitive quality criterion on WUGs. Many other clustering algorithms such as Chinese Whispers (Biemann, 2006) make local decisions so that the final solution is not guaranteed to optimize a global criterion such as SWD. (v) By weighing each edge with its (shifted) weight, SWD respects the gradedness of word meaning. That is, edges with $|W'(e)| \approx 0$ have less influence on SWD than edges with $|W'(e)| \approx 1.5$.[7]

[5] This optimization algorithm showed superior performance in the simulation study described in Appendix D.

[6] Find our code at `https://github.com/Garrafao/WUGs`. We use mlrose (Hayes, 2019) to perform the clustering.

[7] Note that, in principle, WUGs can be clustered using any graph clustering algorithm including soft-clustering (e.g. Biemann, 2006; McCarthy et al., 2016; Abbe, 2017).

We finally obtain a clustering $C : U \mapsto L$ mapping each use $u \in U$ to a cluster label $l \in L \subset \mathbb{N}$. From this, we calculate a **cluster (sense) frequency distribution** D encoding the size of each cluster as

$$D = (f(L_1), f(L_2), ..., f(L_i))$$

where $L_i < L_{i+1}$ and $f(L_i)$ is the number of times any use from U was mapped to the cluster label L_i (cf. McCarthy et al., 2004; Lau et al., 2014). Correspondingly, we obtain two distributions D_1, D_2 from C for the two time-specific use sets U_1, U_2. D, D_1 and D_2 are ordered and contain the frequencies for the full set of cluster labels L so that the ith index always corresponds to the same cluster label. (Note that this means that the time-specific sense frequency distributions are obtained from clustering the full graph.) We obtain the corresponding **cluster (sense) probability distributions** P_1, P_2 by dividing D_1 and D_2 by their respective total frequencies.

3.1.5 Change Scores

Assume that $\mathbf{G} = (\mathbf{U}, \mathbf{E}, \mathbf{W})$ is a Word Usage Graph (see Section 3.1.3) of word w containing w's uses U from two time periods. D and E are the time-specific sense frequency distributions (see Section 3.1.4) of length K obtained by clustering the uses in U based on the edge weights in W, and P and Q the corresponding sense probability distributions obtained by dividing D and E by their respective total frequencies. Note that P and Q are probability distributions, i.e., $0 \leq p_i, q_i$ and $\sum_i^K p_i = 1, \sum_i^K q_i = 1$.[8] Further, assume that $\mathbf{G_{1,2}} = (\mathbf{U}, \mathbf{E_{1,2}}, \mathbf{W_{1,2}})$ is the subgraph of G containing all uses U from both time periods, but only those edges (u_1, u_2) where word uses u_1 and u_2 are from different time periods (COMPARE edges, see Section 3.1.3).

3.1.5.1 Binary Change

Analogous to the definition of the annotation scale (see Table 3.1) and the cluster criteria (see Section 3.1.4), our definition of binary semantic change follows from Blank's theory. Recall from Section 2.2.2.1 that he defines two main types of LSC:

- **innovative meaning change**: emergence of a full-fledged additional sense of a word; old and new sense are related by polysemy,

[8] We allow for zero-probability events.

- **reductive meaning change**: loss of a full-fledged sense of a word.

These two types can be inferred directly from the final sense frequency distributions obtained on a target word's WUG (see Section 3.1.4). We define the **binary change** score of the word w as

$$\begin{aligned} B(D,E) = B(w) = 1 &\text{ if for some } i, D_i \leq k \text{ and } E_i \geq n, \\ &\text{or vice versa.} \\ B(w) = 0 &\text{ else.} \end{aligned}$$

where D_i and E_i are the frequencies of sense i in the first and the second time period, and k, n are lower frequency thresholds aimed to avoid that small random fluctuations in sense frequencies caused by sampling variability or annotation error are misclassified as change. According to the definition above, a word is classified as change if it either gains or loses a sense. It gains a sense if it is attested at most k times in the annotation sample from C_1, but attested at least n times in the sample from C_2. (Similarly for words that lose a sense.) Note that $B(w)$ is symmetric, i.e., invariant to switching D and E. We make no distinction between words that gain vs. words that lose senses, both fall into the change class. Equally, we make no distinction between words that gain/lose one sense vs. words that gain/lose several senses.

3.1.5.2 Graded Change

The binary notion of LSC closely corresponds to Blank's definition and is thus well-motivated. However, it has several disadvantages: (i) It has a comparably high expected error (see Section 3.3.3.2). In most cases it is not feasible to annotate or even observe the full use set for a particular word. Consequently, it is possible that words receive the wrong binary change label because we miss (especially infrequent) senses in the annotated sample. Each such wrongly labeled word produces the highest possible per instance error in a binary classification task. The same holds for annotation errors: human annotators make occasional mistakes. Few such mistakes may change the binary change label of a word and thus have a strong influence on the error.[9] Similarly, our clustering operates on sparsely annotated graphs which may produce clustering errors having a similar effect. (ii)

[9] We try to mitigate this with the thresholds mentioned above, but are not able to completely avoid it.

The binary change measure does not capture initial stages of changes where senses decrease or increase in frequency (see Section 2.2.2.2). A measure of such more frequency-related changes can also be interesting for other fields of research interested in sense frequency divergences between domains such as terminology extraction (Hätty et al., 2020). (iii) The binary measure does not capture the graded view of semantic change underlying previous work on LSCD (Hamilton et al., 2016b; Dubossarsky et al., 2017; Bamler and Mandt, 2017; Rudolph and Blei, 2018; Rosenfeld and Erk, 2018).

Hence, we introduce a second, graded measure of LSC. We first normalize the cluster frequency distributions D and E to probability distributions P and Q. The **graded change** score of the word w is then defined as the Jensen-Shannon Distance (JSD) between the two normalized frequency distributions:

$$G(P, Q) = G(w) = JSD(P, Q)$$

where the Jensen-Shannon Distance is the symmetrized square root of the Kullback-Leibler Divergence (Lin, 1991; Donoso and Sanchez, 2017):

$$JSD(P, Q) = \sqrt{\frac{KLD(P \parallel M) + KLD(Q \parallel M)}{2}}$$

where

$$KLD(P \parallel Q) = \sum_{i}^{K} p_i \log_2 \left(\frac{p_i}{q_i}\right),$$

$$M = \frac{(P + Q)}{2}.$$

We prefer the Jensen-Shannon *Distance* over Jensen-Shannon or Kullback-Leibler *Divergence* because the former is a true metric in contrast to the latter. This means that amongst other properties specific to metrics the Jensen-Shannon Distance is symmetric, i.e., $JSD(P, Q) = JSD(Q, P)$.

$G(w)$ ranges between 0 and 1 (we use $\log_2$) and is high if P and Q assign very different probabilities to the same senses. Note that $B(w)$ and $G(w)$ do not necessarily correspond to each other: a word w may show no binary change but high graded change, or vice versa.[10] The graded change notion enables us to compare

[10] Find some formal correspondences between the measures in Appendix B.

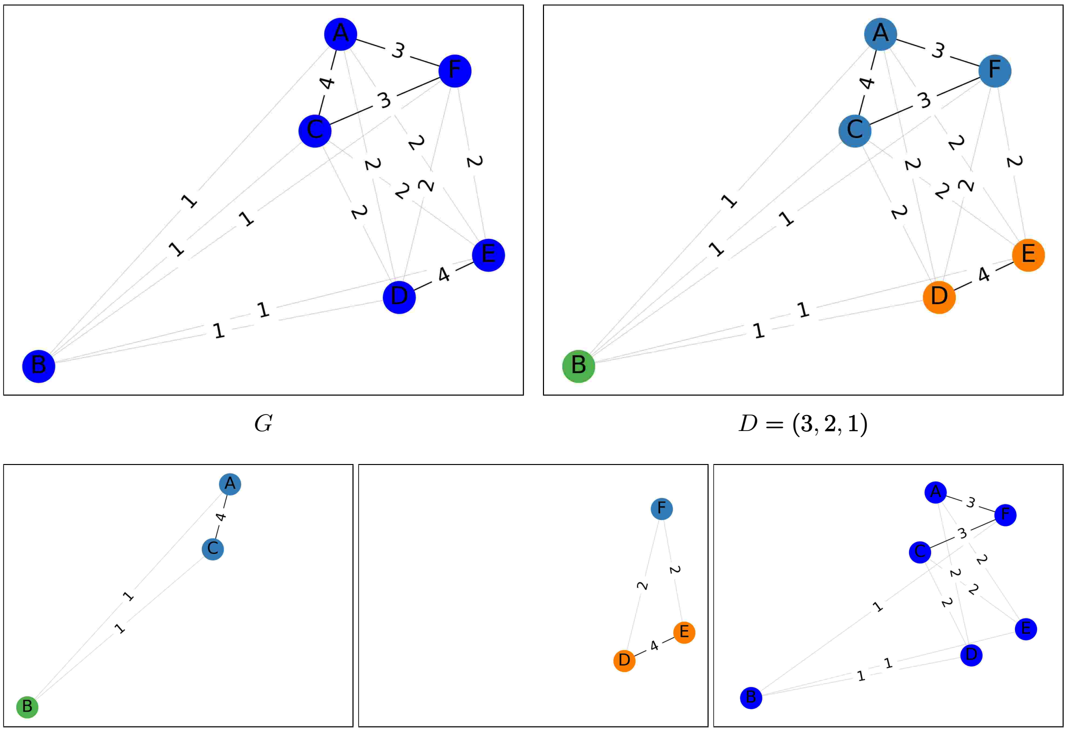
any two words over time and decide which of the words changed more. It provides an answer to questions like: Did a word that lost a very frequent sense change more than a word that lost a very infrequent sense? And, did a word that gained two senses change less than a word that lost three senses? Above that, it solves many of the problems of the binary change notion: missing infrequent senses or occasional annotation and clustering errors produce low error on $G(w)$ as cases in which we wrongly observe $D_n = 0$, $E_n \neq 0$ (or vice versa) do not necessarily produce the highest possible error of 1.0 (see Section 3.3.3.2). Further, there will always be graded change if probabilities for some sense are different (see Lemma 10 in Appendix B). Hence, it also captures changes in sense frequencies beyond the complete loss or gain of senses, which are not captured by the binary notion.

3.1.5.3 Negated COMPARE

Just as the binary notion, the graded notion of change introduced above is defined on sense probability distributions and hence relies on some clustering of word uses. The clustering process is complex and may introduce additional errors due to many reasons such as annotation errors, graph sparsity or non-optimal cluster loss solutions (see Section 3.3.2.2). It also requires a large number of comparisons between word use pairs as each use has to be compared to several other uses to obtain a graph of sufficient density for meaningful clustering. Hence, we introduce a secondary and simple, graded measure **Negated COMPARE**, which does not rely on clustering and captures specific cases of $G(w)$ (see Appendix B). For this, we only consider the edges from the subgraph $\mathbf{G_{1,2}} = (\mathbf{U}, \mathbf{E_{1,2}}, \mathbf{W_{1,2}})$, i.e., we consider only the semantic proximities of uses between (not within) time periods (COMPARE weights, see Section 3.1.3). The Negated COMPARE score of the word w is then defined as the negated mean of the COMPARE weights:

$$C(W_{1,2}) = -\frac{1}{|W_{1,2}|} \sum_{x \in W_{1,2}} x \; .$$

$C(w)$ is easy to compute, symmetric and ranges between -4.0 and -1.0. It exploits the fact that semantic proximity strongly correlates with cluster membership (see Section 3.1.4). We report the *negated* COMPARE score as then the global maximum and minimum will coincide with $G(w)$ for certain cases (see Appendix B). Although the Negated COMPARE measure is not a perfect approximation of the graded change score (cf. Arefyev and Bykov, 2021), we assume that they correlate



G $D = (3, 2, 1)$

G_1, $D_1 = (2, 0, 1)$ G_2, $D_2 = (1, 2, 0)$ $G_{1,2}$

Figure 3.1: Top: WUG G of *arm* (left) and clustered WUG (right). D gives cluster frequency distribution. **black**/gray lines for **high** (≥ 2.5)/low (< 2.5) edge weights. Spatial proximity of nodes loosely corresponds to their semantic proximity annotation. Bottom: Subgraph for 1st time period G_1 and 2nd time period G_2. D_1 and D_2 give corresponding cluster frequency distributions. $G_{1,2}$ contains only edges between time periods (COMPARE edges).

significantly. We experimentally verify this assumption in Section 3.3.3.2.

3.1.6 Example

Consider the example in Figure 3.1: WUG G represents the semantic proximity structure underlying the set of word uses U of the English word *arm* displayed in Table 3.2.[11] The uses $U_1 = \{A, B, C\}$ and $U_2 = \{D, E, F\}$ were sampled from the two time periods 1820–1860 and 1950–1990 respectively (t_1, t_2). We derive lexicographic senses on G by building three clusters of uses with high semantic proxim-

[11] Find the code used for generating WUG visualizations at https://github.com/Garrafao/WUGs.

A	1824	and taking a knife from her pocket, she opened a vein in her little **arm**,
B	1842	And those who remained at home had been heavily taxed to pay for the **arms**, ammunition;
C	1860	and though he saw her within reach of his **arm**, yet the light of her eyes seemed as far off
		...
D	1953	overlooking an **arm** of the sea which, at low tide, was a black and stinking mud-flat
E	1975	twelve miles of coastline lies in the southwest on the Gulf of Aqaba, an **arm** of the Red Sea.
F	1985	when the disembodied **arm** of the Statue of Liberty jets spectacularly out of the

Table 3.2: Sample of diachronic corpus, cf. Deane (1988, p. 347) and Blank (1997, pp. 412–417).

ity and low semantic proximity to other clusters (see Section 3.1.4): $C_1 = \{A, C, F\}$ (blue), $C_2 = \{D, E\}$ (orange), $C_3 = \{B\}$ (green). We then build the time-specific subgraphs G_1 and G_2 and are now able to compare the clusters between time periods, where D, D_1 and D_2 give the cluster frequencies in the full graph and the two subgraphs respectively. For instance, C_3 only exists in the first time period while C_2 only exists in the second time period. Setting thresholds $k = 0$ and $n = 0$, this means that $B(arm) = 1$. The cluster probability distributions for D_1 and D_2 are $P_1 = (0.66, 0.0, 0.33)$ and $P_2 = (0.33, 0.66, 0.0)$, yielding a rather high graded change score $G(arm) = 0.74$ as a sense with high probability is gained and the other senses lose or gain moderately in probability. There are 9 edges between time periods ($G_{1,2}$) with corresponding weights $W_{1,2} = \{3, 2, 2, 3, 2, 2, 1, 1, 1\}$. Hence, we have $C(arm) = -1.89$, also a rather high change score.

Let us look more closely at which senses are represented by the clusters we built: C_1 represents *arm*'s sense 'human upper limb', clearly expressed by uses A and C, having high semantic proximity (4). There is, however, some variation within this cluster, as F expresses a variant of the core sense expressed by A and C, referring to the non-human arm of a statue. Yet, F bears enough similarity to A and C such that they receive a rather high semantic proximity (3). The uses in C_1 are all rather distinct from the uses D and E in C_2, representing the sense 'an inlet of water'. However, they still bear a distant semantic similarity (in form) to each other and hence receive a rather low semantic proximity (2). Note that within C_2 we still have high semantic proximity (4) between D and E as these uses express the same sense. C_3 represents the third sense 'weapon'. B has a low semantic proximity (1) to all other uses as there is no semantic relation e.g. between B and D.

3.2 Data

We apply the above-described framework in several studies to obtain annotated and clustered WUGs and to derive estimated change scores from these. We also carry out a simplified annotation where clustering is omitted. Both types of studies are described in this section. We also describe the annotation of a third data set, using traditional sense definitions, which we create in order to validate the WUG approach.

3.2.1 SemEval 2020

This study was carried out as part of the shared task on Unsupervised Lexical Semantic Change Detection at the Workshop on Semantic Evaluation (SemEval) 2020. We describe the annotation procedure for the English (EN), German (DE) and Swedish (SV) data sets.[12]

3.2.1.1 Corpora

For English, we use the Clean Corpus of Historical American English (CCOHA, Alatrash et al., 2020), a cleaned version of COHA (Davies, 2012), which spans 1810–2010. COHA is balanced by text genre decade by decade. For German, we use the DTA corpus (Deutsches Textarchiv, 2017) and a combination of the BZ and ND corpora (Berliner Zeitung, 2018; Neues Deutschland, 2018).[13] DTA contains texts from different genres spanning the 16th–20th centuries. BZ and ND are newspaper corpora jointly spanning 1945–1993. For Swedish, we use the Kubhist corpus (Språkbanken, 2019), a newspaper corpus containing texts from 18th–20th century. All corpora are lemmatized and POS-tagged. CCOHA and DTA are additionally

[12] The task also includes evaluation on a Latin data set, which was annotated differently and will not be described here, but in Appendix C.

[13] We use the TCF-version of DTA released October 18, 2018: `http://www.deutschestextarchiv.de/download`.

	C_1					C_2				
	corpus	**period**	**tokens**	**types**	**TTR**	**corpus**	**period**	**tokens**	**types**	**TTR**
English	CCOHA	1810–1860	6.5M	87k	13.38	CCOHA	1960–2010	6.7M	150k	22.38
German	DTA	1800–1899	70.2M	1.0M	14.25	BZ+ND	1946–1990	72.3M	2.3M	31.81
Swedish	Kubhist	1790–1830	71.0M	1.9M	47.88	Kubhist	1895–1903	110.0M	3.4M	17.27

Table 3.3: Statistics of SemEval corpora. TTR = Type-Token ratio (number of types / number of tokens * 1000).

spelling-normalized. BZ, ND and Kubhist contain frequent OCR errors (Adesam et al., 2019; Hengchen et al., 2020).

From each corpus, we extract two time-specific subcorpora C_1, C_2, as defined in Table 3.3. The division is driven by considerations of data size and availability of target words (see Section 3.2.1.2). From these two subcorpora, we then sample the released test corpora in the following way: Sentences with < 10 tokens are removed. German C_2 is downsampled to fit the size of C_1 by sampling all sentences containing target lemmas and combining them with a random sample of sentences not containing target lemmas of suited size. An equal procedure is applied to downsample English C_1 and C_2. For Swedish, the full amount of sentences is used. Finally, all tokens are replaced by their lemma, punctuation is removed and sentences are randomly shuffled within each of C_1, C_2. The final corpus files have one sentence per line.

Sentence shuffling and lemmatization are done for copyright reasons. Some corpora require special processing steps: Where Kubhist does not provide lemmatization (through KORP, Borin et al., 2012), we leave tokens unlemmatized. For copyright reasons, CCOHA contains frequent replacement tokens (10 x '@'). We split sentences around replacement tokens and remove them as a first step in the preprocessing pipeline. Further, because English frequently combines various POS in one lemma and many of our target words underwent POS-specific semantic changes (see Section 3.2.1.2), we concatenate targets in the English corpus with their broad POS tag ('target_pos'). Also, the joint size of the CCOHA subcorpora has to be limited to ~10M tokens because of copyright issues.

Find a summary of the released (lemmatized) test corpora in Table 3.3. We also create a tokenized version of the corpora with sentences in the same order as in the lemmatized version.[14]

[14] Find the corpora at `https://www.ims.uni-stuttgart.de/data/sem-eval-ulscd`.

3.2.1.2 Target Words

Target words are either: (i) words that we assume to *change* their meaning (lost or gained a sense) between C_1 and C_2, or (ii) *stable* words that we assume did not change their meaning during that time.[15] A large list of 100–200 changing words is selected by scanning etymological and historical dictionaries (Paul, 2002; Svenska Akademien, 2009; OED, 2009) for changes within the time periods of the respective corpora. This list is then further reduced by one annotator who checks whether there are meaning differences in samples of 50 uses from C_1 and C_2 per target word. Stable words are chosen by sampling a control counterpart for each of the changing words with the same POS and comparable frequency development between C_1 and C_2, and manually verifying their diachronic stability, as described above. Both types of words are annotated, which allows us to verify the a priori choice of changing and stable words. By balancing the target words for POS and frequency we aim to minimize the possibility that model biases towards these factors lead to artificially high performance (Dubossarsky et al., 2017; Schlechtweg and Schulte im Walde, 2020).

3.2.1.3 Annotators

We start out with four annotators per language. Following high annotation loads and dropouts, we hire additional annotators, resulting in 9/8/5 annotators in total for EN/DE/SV, respectively. All annotators are native speakers and current or former university students. The number of annotators with a background in historical linguistics is two for DE and one for EN and SV.[16] [17]

3.2.1.4 Use Sampling

For each target word, 100 sentences are randomly sampled from each of the tokenized versions of C_1 and C_2 (see Section 3.2.1.1). Each sentence contains the target word (possibly in an inflected form) and a minimum of ten tokens, yielding a total of 200 uses per target word.[18] If a target word has less than 100 uses, the full

[15] A target word is represented by its lemma form.

[16] In Section 3.3.1, we find that annotators with and without historical background have reasonably high agreement.

[17] The guidelines for annotator training are available at `https://www.ims.uni-stuttgart.de/data/wugs` under DWUG EN/DE/SV.

[18] Because English frequently combines various POS in one lemma and many of our target words underwent POS-specific semantic changes, we sample only uses of English target words with the

sample is annotated. We then mix the use samples of a target word into a joint set U and annotate U for semantic proximity between word use pairs, as described in Section 3.1.2. These are presented to annotators in random order and annotated on the four-point scale in Table 3.1.

3.2.1.5 Edge Sampling

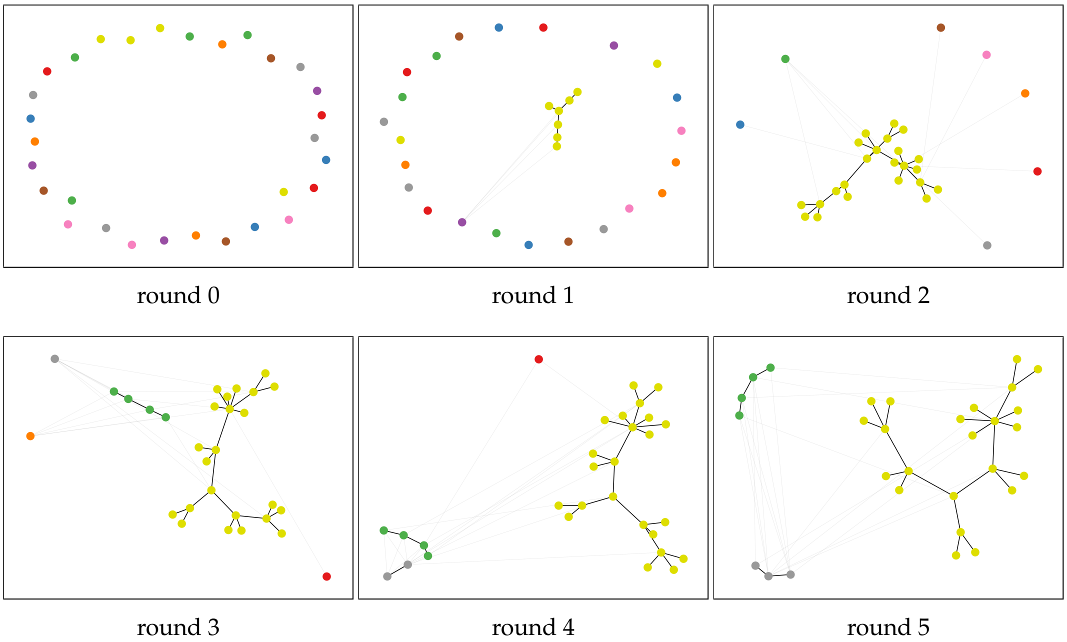
Annotating the full WUG is not feasible even for a small set of n uses as this implies annotating $\frac{n(n-1)}{2}$ edges. Hence, the main challenge with our annotation approach is to annotate as few edges as possible while keeping the information needed to infer the same clustering on the graph as on the fully-annotated graph. For this, we exploit the following observation: Uses from the same cluster (with the same sense) have similar patterns of semantic proximity, i.e., they have similarly low semantic proximity values to uses from another cluster and similarly high values to uses from their own cluster. From this, it follows that it is sufficient to annotate few edges within and between clusters to approximate the patterns of all uses. For instance, consider a situation where use u has high semantic proximity with only one use from some cluster. We can assign use u to that cluster assuming that uses within the cluster have similar patterns of semantic proximity, i.e., assuming that all uses within the cluster have high proximity with use u.

In order to reduce annotation of redundant information, we annotate the data in several rounds. After each round, the WUG of a target word is updated with the new annotations and a new clustering is obtained.[19] Based on this clustering, the edges for the next round are sampled through heuristics similar to Biemann (2013). The annotation load is randomly distributed making sure that roughly half of the use pairs are annotated by more than one annotator.

The first round aims to obtain a small high-quality reference set of clusters. This is achieved through the sampling of 10% of the uses from U and 30% of the edges by a random walk through the sample graph (**exploration**), which guarantees that all nodes are connected by some path. Hence, the first clustering is obtained on a small but richly-connected subgraph ensuring that not too many clusters are inferred as this would lead to a strong increase in annotation instances in the subsequent rounds. In the second round, the reference clusters from the first round serve as a comparison for those uses which were not assigned to a multi-cluster yet

broad POS tag for which a change has been described.

[19] If an edge was annotated by several annotators, the median was retained as an edge weight.



round 0 round 1 round 2

round 3 round 4 round 5

Figure 3.2: Simulated example of annotation pipeline.

(**combination**).[20] In all subsequent rounds, both a combination step and an exploration step are employed. The combination step combines each single use u_1 that is not yet member of a multi-cluster with a random use u_2 from each of the multi-clusters to which u_1 has not yet been compared. The exploration step consists of a random walk on 30% of the edges from the non-assignable uses, i.e., uses which have already been compared to each of the multi-clusters, but were not assigned to any of these by the clustering algorithm. This procedure slowly populates the graph while minimizing the annotation of redundant information. We aim to stop the procedure when each cluster has been compared to each other cluster. The full procedure, including the sample sizes for the random walk, is validated in the simulation study described in Appendix D.

We combine the sampling procedure with further heuristics added after round 1 to increase the quality of the annotation: (i) We sample a low number of randomly chosen edges and edges between already confirmed multi-clusters for further annotation to corroborate the inferred structure. (ii) We detect strong disagreements

[20] We refer to a cluster with ≥ 2 uses as 'multi-cluster'.

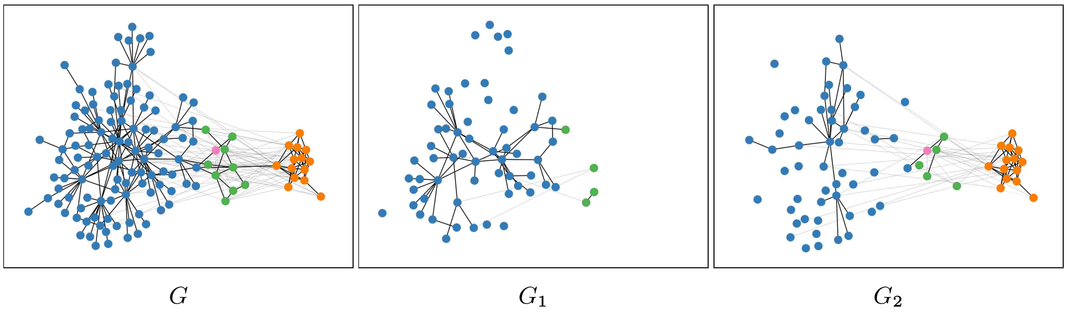


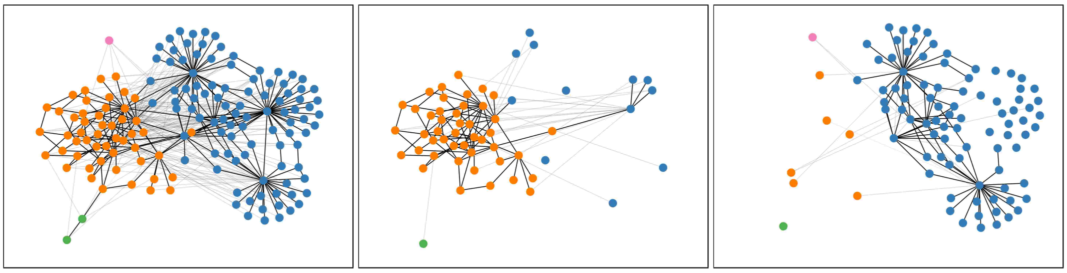
Figure 3.3: WUG of Swedish *ledning* (left), subgraphs for 1st time period G_1 (middle) and 2nd time period G_2 (right). $D_1 = (58, 0, 4, 0)$, $D_2 = (52, 14, 5, 1)$, $B(w) = 1$ and $G(w) = 0.34$.

between annotators, i.e., judgments with a difference of ≥ 2 on the scale or edges with a median ≈ 2.5, and redistribute each such edge to another randomly chosen annotator from the ones who did not annotate the respective edge yet to resolve the disagreement. (iii) We detect clustering conflicts, i.e., positive edges between clusters and negative edges within clusters (see Section 3.1.4) and sample a new edge for each node connected by a conflicting edge. This adds more information in regions of the graph where finding a good clustering is hard. Furthermore, after each round, nodes from the graph whose 'Cannot decide' judgments made up at least half of their total judgments are removed, and in a few cases, whole words are removed if they have a high number of 'Cannot decide' judgments or need a high number of further edges to be annotated. We stop the annotation after four rounds for time constraints.

Figure 3.2 shows an example of our annotation pipeline. As the annotation proceeds through the rounds, the graph becomes more populated and the correct cluster structure is found. In round 1, one multi-cluster is found. Hence, all remaining uses are compared with this cluster in round 2 by the combination step. In rounds 3 and 4, the exploration step discovers more clusters not found in the rounds before.

3.2.1.6 Summary

Figure 3.3 and Figure 3.4 show the annotated and clustered WUG G for Swedish *ledning* and German *Eintagsfliege* (left). Nodes represent uses of the target word. Edges represent the median of semantic proximity judgments between uses (**black**/gray



G G_1 G_2

Figure 3.4: WUG of German *Eintagsfliege* (left), subgraphs for 1st time period G_1 (middle) and 2nd time period G_2 (right). $D_1 = (12, 45, 0, 1)$, $D_2 = (85, 6, 1, 1)$, $B(w) = 0$ and $G(w) = 0.66$.

lines for **positive**/negative edges). Spatial proximity between uses in the plot loosely corresponds to the median of their annotated semantic proximities. Colors make clusters (senses) inferred on G. After splitting G into the two time-specific subgraphs for C_1, C_2, we obtain the two sense frequency distributions D_1, D_2. From these, we infer the binary and the graded change score setting the lower frequency thresholds for the binary score to $k = 2$ and $n = 5$ (see Section 3.1.5). The two words represent semantic changes indicative of the difference between the two scores, respectively: *ledning* gains a sense with rather low frequency in C_2. Hence, it has binary change, but low graded change. For *Eintagsfliege*, however, its two main senses exist in both C_1 and C_2 while their frequencies change dramatically. Hence, it has no binary change, but high graded change. Note that the obtained scores only estimate the binary and graded change scores defined in Section 3.1.5 as the graph clustering is obtained from an incomplete graph.

Find a summary of the annotation outcome in Table 3.4. The final test sets contain between 31 (Swedish) and 48 (German) target words. Throughout the annotation we excluded several targets if they had a high number of 'Cannot decide' judgments or needed a high number of further edges to be annotated. Following Erk et al. (2013), we report the mean of pairwise Spearman correlations (Spearman, 1904) between annotator judgments as agreement measure. Erk et al. report agreement scores of 0.55 and 0.62, which is comparable to ours.[21] Following

[21] Note that because we spread disagreements from previous rounds in each round to further annotators, uses in later rounds become much harder to judge on average, which has a negative effect on agreement. Hence, for comparability reasons, we report the agreement in the first round where

	General					Binary				Graded			
	n	N/V/A	SPR	KRI	$\|U\|$	LSC	FRQ$_\mathrm{d}$	FRQ$_\mathrm{m}$	PLY$_\mathrm{m}$	LSC	FRQ$_\mathrm{d}$	FRQ$_\mathrm{m}$	PLY$_\mathrm{m}$
English	37	33/4/0	.69	.61	193	.43	-.18	-.03	.45	.24	-.29	-.05	.72
German	48	32/14/2	.59	.53	175	.35	-.06	-.11	.68	.31	.00	-.02	.73
Swedish	31	23/5/3	.57	.56	187	.26	-.04	-.29	.45	.16	.00	-.13	.75

Table 3.4: Overview SemEval target words. n = number of target words, N/V/A = number of nouns/verbs/adjectives, SPR = weighted mean of pairwise Spearman in round 1, KRI = Krippendorff's alpha in round 1, $|U|$ = avg. no. uses per word (after cleaning), LSC = mean binary/graded change score, FRQ$_\mathrm{d}$ = Spearman correlation between change scores and target words' absolute difference in log-frequency between C_1, C_2. Similarly for minimum frequency (FRQ$_\mathrm{m}$) and minimum number of senses (PLY$_\mathrm{m}$) across C_1, C_2.

Rodina and Kutuzov (2020), we also report Krippendorff's alpha (Krippendorff, 2018) which is chance-corrected and reach comparable scores to their 0.51 and 0.53. In the calculation of both agreement measures we omit 'Cannot decide' judgments.

The class distribution (column 'LSC') for binary change differs per language as a result of several target words being dropped during the annotation. In Swedish the majority has no binary change. This is also reflected in the mean scores for graded LSC. Despite the excluded target words the frequency statistics are roughly balanced (FRQ$_\mathrm{d}$, FRQ$_\mathrm{m}$). However, we did not control the test sets for polysemy and there are strong correlations for English, German and Swedish between graded change and polysemy (PLY$_\mathrm{m}$). This correlation reduces for binary change, but is still moderate for English and Swedish and remains high for German.

In total, roughly 86k (37k/29k/20k for EN/DE/SV) judgments are made by annotators. $\approx$ 50% of the sampled use pairs are annotated by more than one annotator. Find a selection of WUGs from all data sets in Appendix A.[22]

3.2.2 DURel & SURel

The Diachronic Usage Relatedness (DURel) and the Synchronic Usage Relatedness (SURel) data set compare the semantic proximity of word uses across a time-specific corpus pair (DURel) and a domain-related corpus pair (SURel), respectively. Although SURel does not measure diachronic sense changes (LSC), it gives us an op-

no disagreement detection has taken place. The agreement across all rounds, calculated as weighted mean of agreements, is 0.52/0.60/0.58.

[22] The data is available at `https://www.ims.uni-stuttgart.de/data/wugs` under DWUG EN/DE/SV Version 1.0.0.

	C_1					C_2				
	corpus	**period**	**tokens**	**types**	**TTR**	**corpus**	**period**	**tokens**	**types**	**TTR**
DURel	DTA	1750–1800	26.7M	252k	9.47	DTA	1850–1900	40.3M	796k	19.75
SURel	SdeWaC	general	109.7M	2.4M	22.03	COOK	domain	1M	49k	46.86

Table 3.5: Statistics of DURel/SURel corpora. TTR = Type-Token ratio (number of types / number of tokens * 1000)

portunity to evaluate computational models on synchronic word sense divergences between corpora. We describe the creation process for both data sets in one section as they are carried out completely in parallel.

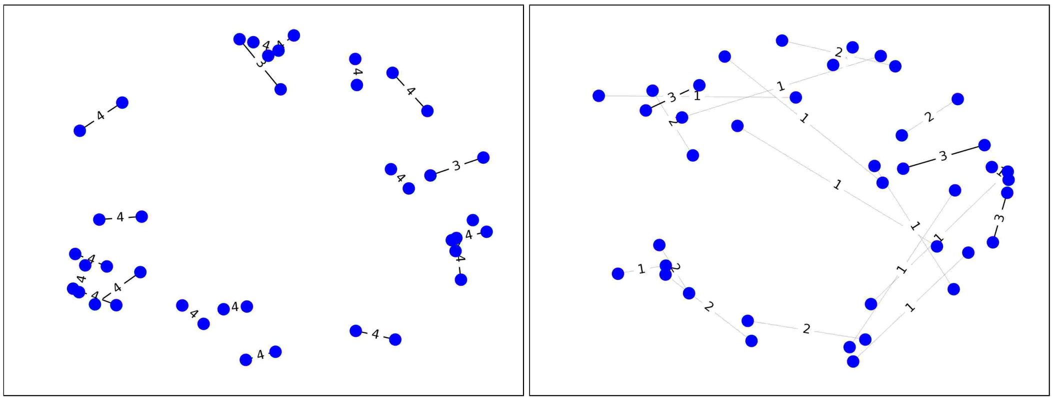
3.2.2.1 Corpora

For DURel, we use the DTA corpus (see Section 3.2.1.1).[23] For SURel, we compare a general language corpus to a domain-specific one. For the general corpus, we subsample SdeWaC (Faaß and Eckart, 2013), a cleaned version of the web-crawled corpus deWaC (Baroni et al., 2009). We reduce SdeWaC to $\frac{1}{8}$th of its original size by selecting every 8th sentence for our general-language corpus. As a domain-specific corpus, we create COOK. For this, we crawl cooking-related texts from several categories (recipes, ingredients, cookware and cooking techniques) from the German cooking recipes websites *kochwiki.de* and *Wikibooks Kochbuch*[24].

From DTA, we extract two time-specific subcorpora C_1 and C_2 (DTA18 and DTA19) for the periods 1750–1800 and 1850–1900. For the domain-related corpora, C_1 and C_2 are given by SdeWaC and COOK respectively. For all corpora, we create preprocessed versions by removing words below a frequency threshold t. For the smallest corpus COOK, we set $t = 2$, and set the other thresholds in the same proportion to the corpus size. This leads to $t = 25, 37, 97$ for DTA18, DTA19 and SdeWaC, respectively. We then create two preprocessed versions of the corpora: (i) a version with minimal preprocessing, i.e., with lemmatization and removed punctuation (L$_{\text{ALL}}$), and (ii) a stronger preprocessed version with only content words: After punctuation removal, lemmatization and POS-tagging, only nouns, verbs and adjectives are retained in the form *lemma:POS* (L/P). Table 3.5 shows statis-

[23] We use the TCF-version of DTA released September 1, 2017: `http://www.deutschestextarchiv.de/download`.

[24] `de.wikibooks.org/wiki/Kochbuch`



German *Abend*, $C(w) = -3.9$ German *Donnerwetter*, $C(w) = -1.8$

Figure 3.5: COMPARE subgraph $G_{1,2}$ for words from DURel data set.

tics for the L_{ALL} corpus versions.[25]

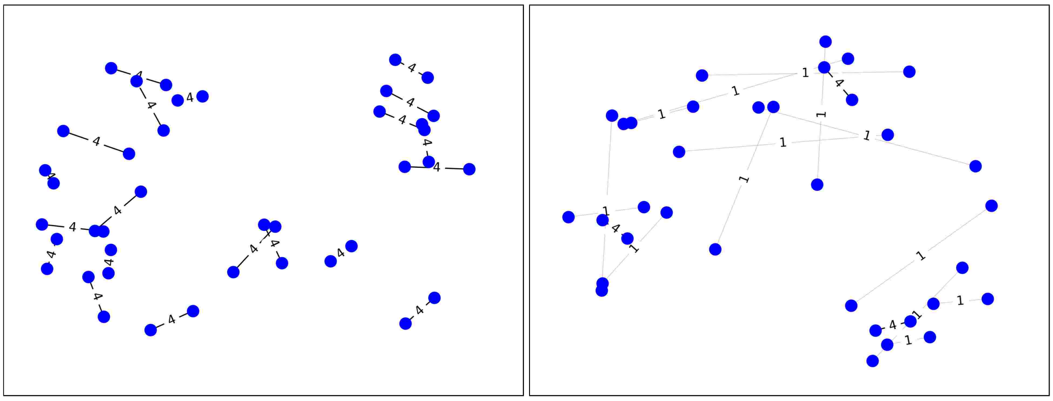
3.2.2.2 Target Words

For DURel, the target words are selected by manually checking DTA for innovative and reductive meaning changes, based on cases of metaphoric and metonymic change and narrowing (innovative), as reported by Paul (2002), and cases of reduction due to homonymy (reductive), as reported by Osman (1971). By focusing on a late time period (19th century), we try to reduce problems coming with historical language data as much as possible. We still normalize special characters to modern orthography.

We include only those words as targets for which we find the change suggested by the literature reflected in the corpus, either weakly or strongly because an annotation relying on a random selection of words suggested to undergo change is likely to produce a set with very similar and rather low values representing small effects. We thus guarantee to include both: (i) words for which we expect weak effects as well as (ii) words for which we expect strong effects. We end up with 22 target words.

For SURel, we select 22 target words which occur in both SdeWaC and COOK, and which we expect to exhibit different degrees of domain-specific meaning change.

[25] Find the L_{ALL} corpora at `https://www.ims.uni-stuttgart.de/data/wocc`.



German *Schnittlauch*, $C(w) = -4.0$ German *Hamburger*, $C(w) = -1.5$

Figure 3.6: COMPARE subgraph $G_{1,2}$ for words from SURel data set.

3.2.2.3 Annotators

For DURel, five native speakers of German are asked to rate use pairs on our 4-point scale of semantic proximity in Table 3.1. All annotators are students of linguistics. We explicitly choose two annotators with a background in historical linguistics in order to see whether knowledge about historical linguistics has an effect on the annotation. Annotators are not told that the study is related to semantic change.

For SURel, four native speakers of German are asked to rate the use pairs. All annotators are university students.[26]

3.2.2.4 Use Sampling

For each target word, we sample all sentences from the source corpus of C_1 and C_2 (see Section 3.2.2.1) searching for the respective lemma and POS and mix them into a joint set U.

3.2.2.5 Edge Sampling

We define three time-specific subsets (groups) of use pairs: EARLIER, containing all pairs with both uses from C_1, LATER, containing all pairs with both uses from

[26] The annotator guidelines for DURel and SURel are available at `https://www.ims.uni-stuttgart.de/data/wugs`.

	General					
	n	N/V/A	SPR	KRI	$\|U\|$	LSC
DURel	22	15/1/6	.59	.54	104	-2.7
SURel	22	19/3/0	.84	.83	104	-2.7

Table 3.6: Overview DURel/SURel target words. n = number of target words, N/V/A = number of nouns/verbs/adjectives, $|U|$ = avg. no. uses per word (after cleaning), SPR = weighted mean of pairwise Spearman in round 1, KRI = Krippendorff's alpha in round 1, LSC = mean Negated COMPARE score.

C_2 and COMPARE, containing only pairs with uses from different time periods. For each target word, we randomly sample 20 use pairs from each of the groups EARLIER, LATER and COMPARE, yielding 60 use pairs per word and 1,320 use pairs for 22 target words in total.

The annotators are provided these use pairs. For DURel, they are provided additionally with the preceding and the following sentence in the corpus. We double-check that each use of a word is only sampled once within each group. If the total number of uses in the group is less than needed, uses are allowed twice across pairs. Before presenting the pairs to the annotators in a spreadsheet, use sequence within pairs is randomized and pairs from all groups are mixed and randomly ordered.

3.2.2.6 Summary

The annotated use pairs are represented in a WUG. However, we skip the clustering step as the graph was not sampled such that nodes are densely connected by edges. Instead, we directly obtain the subsampled COMPARE subgraph $G_{1,2}$ and compute the estimated Negated COMPARE score $C(w)$ for which no clustering step is needed.[27] [28] Find some examples from DURel and SURel with low and high $C(w)$ scores in Figures 3.5 and 3.6. Note that the obtained scores only estimate the Negated COMPARE score defined in Section 3.1.5 as the COMPARE use pair samples are only a random sample from the full set of COMPARE weights $W_{1,2}$.

Find a summary of the annotation outcome in Table 3.6. The final test sets

[27] For our purposes, we ignore the subgraphs G_1 and G_2, i.e., the groups EARLIER, LATER. But, these can in principle be used to calculate further change scores taking the within-period semantic variation into account (Schlechtweg et al., 2018).

[28] Note that in this early version of the data set we computed edge weights as the mean instead of the median of annotator judgments.

contain 22 target words, each. As in Section 3.2.1.6, we report the mean of pairwise Spearman correlations and Krippendorff's alpha as agreement measures. For DURel, we reach comparable scores to previous studies and the SemEval data while for SURel we reach very high agreement above 0.83. Both data sets have a mean Negated COMPARE score of -2.7, where the scores of the target words are widely distributed over the range from -4.0 to -1.0. In total, roughly 12k (6.6k for DURel, 5.28k for SURel) judgments were made by annotators. For DURel, all sampled use pairs were annotated by five annotators while for SURel they were annotated by four annotators. Find a selection of WUGs from all data sets in Appendix A.[29]

3.3 Validation

As the WUG annotation framework described in Section 3.1.2 including (i) annotation of semantic proximity, (ii) clustering and (iii) inference of change scores is rather new, we try to validate each of these levels.

3.3.1 Semantic Proximity

As the clustering operates on semantic proximity annotations, it is important to know whether these annotations reflect anything meaningful about the semantics of word uses. To meet this criterion, the proximity judgments should be intersubjective, i.e., humans should show considerable agreement on their judgments. If this is not the case, we have to assume that the concept of semantic proximity between word uses cannot be grasped by humans or that our particular annotators did not grasp it.

In Tables 3.4 and 3.6, we see the agreement between annotators for the annotation studies from Section 3.2. Overall, the agreement is moderate to high (SPR>= .57) and clearly above chance (KRI>= .53). For SURel, agreement is exceptionally high (SPR= .84, KRI> .83).

In Table 3.7, we can now now take a closer look at Spearman correlations between annotators on the DURel data. The bottom line provides the agreement of each annotator's judgments against the average judgment score across the other annotators. The range of correlation coefficients is between 0.57 and 0.68, with an average correlation of 0.66. All the pairs are highly significantly correlated ($p < 0.01$).

[29] The data is available at `https://www.ims.uni-stuttgart.de/data/wugs` under DURel/SURel V2.0.0.

	1	2	3	4	5
1		0.59	0.63	0.67	0.66
2			0.57	0.64	0.65
3				0.64	0.62
4					0.68
avg	0.71	0.68	0.68	0.75	0.74

Table 3.7: Annotator agreement (Spearman) on DURel; *avg* refers to agreement of the annotator in the respective column against the average of annotations across the other annotators.

The annotators with historical background are annotators 4 and 5, who show the highest pairwise agreement and also the highest agreement with the average of the other annotators. This indicates that historical knowledge makes a positive difference when annotating historical semantic proximity. Yet, the agreements of the non-expert annotators only deviate slightly.

Overall, our correlations are comparable and even moderately higher than the ones found in Erk et al. (2013), who report average correlation scores of 0.55 and 0.62. This difference is remarkable given that annotators had to judge historical data. Note, however, that the studies are not exactly comparable as Erk et al. used a more fine-grained 5-point scale and we presumably excluded a larger number of 'Cannot decide' judgments.

3.3.1.1 Annotator Disagreements

We now manually inspect use pairs where annotators show strong disagreement. We perform this analysis on the WUGs of *abbauen*, *abgebrüht*, *Knotenpunkt*, *Manschette* and *zersetzen* from the SemEval DE data set (see Appendix A), which are chosen to cover different POS and cases with low and high correspondence to the sense definition annotation described in Section 3.3.2 (see Figure 3.7). For this, we extract interactive WUGs displaying only edges where at least one annotator pair diverges by at least two points on the DURel scale in Table 3.1 (e.g. 1/3) and analyze these.[30] We identify 5 main classes of disagreement sources:

[30] Interactive WUGs are interactive HTML files plotting WUGs in 2D, allowing humans to more easily analyze the annotated data. Such interactive plots come with each of our data set uploads.

- ambiguity,
- meaning unfamiliarity,
- misleading context,
- unclear meaning abstraction level and
- different intuitions on semantic proximity.

Most cases of disagreements between annotators can be traced back to ambiguity or meaning unfamiliarity with one of the involved uses. Consider the following uses:

(3.3) [. . .] das war ein finsterer Herr mit dem harten Blick eines **abgebrühten** Schellfisches.
*'[. . .] that was a sinister gentleman with the hard look of a **blanched/hard-nosed** haddock'*

(3.4) Darum hatte Calloway solche **Manschetten**, was?
*'That's why Calloway had **fear/cuffs/collars** like that, huh?'*

(3.5) Vor allem Gregor Strasser war einer der braunen Halbgötter, bis er 1932 kurzerhand von Hitler **abgebaut** wurde.
*'Above all Gregor Strasser was one of the brown demigods until he was **destroyed?/deprived?** by Hitler in 1932.'*

(3.3) is a case of ambiguity: *abgebrüht* modifies an animal which could be 'blanched' in the literal sense, but could also mean 'hard-nosed' as the animal is further attributed with a "hard glance". Often ambiguity is also triggered by missing sentence context: (3.4) is a rather short sentence, which gives little clues on the meaning of the target word. Hence, *Manschetten* is at least ambiguous between a 'fear', a 'cuffs' and a 'collar' reading. (3.5) is a case of meaning unfamiliarity: *abgebaut* occurs in an archaic sense which we only observed once in our data and which is likely unfamiliar to annotators. The context and the word's other senses suggest that *abgebaut* means something around 'to destroy, to deprive', but the exact meaning remains unclear. Further cases include uses with misleading context where a superficial reading or certain key words suggest a specific reading while a deeper reading suggests another one and uses where the meaning of the target word could be described on various abstraction levels. There are also a few cases where the

Find the code for generating the plots at `https://github.com/Garrafao/WUGs`.

above-mentioned categories do not apply. These may be due to (genuinely) different intuitions on semantic proximity.

Our analysis suggests that providing only the sentence containing the target word to annotators (as for the SemEval data sets and SURel) is often not enough to disambiguate the meaning of the target word. More context (e.g. one preceding and one following sentence, as in DURel) and meta-information (such as author information) should be provided to reduce the problem of use ambiguity.

3.3.2 Clustering

Using clustering on WUGs to infer sense clusterings for model evaluation has not been done before. Hence, we created validation data for the obtained sense clusters with an established annotation strategy. Strong divergences from traditional sense definitions may reveal inferred clusterings which do not reflect intuitive sense distinctions. However, some degree of divergence is expected, especially in the case of fine-grained distinctions.

3.3.2.1 Sense Definition Annotation

We choose 24 target words randomly from the SemEval DE data set and extract sense definitions from two historical dictionaries (DWDS, 2021; Paul, 2002).[31] We then randomly sample 50 uses for each target word (25 per time period) and ask three annotators to assign each use to the sense definition best describing the meaning of the target word in this use. The annotators have the option to assign the use to a non-specified sense definition 'other' ('andere') if none of the definitions fits or to make no decision. One annotator is a professional computational linguist, one annotator holds a degree in linguistics and the third annotator is a current university student. The annotators had no access to the data before the annotation. In the first round, only one annotator annotates the data and also provides additional sense definitions for four words (*abdecken*, *Fuß*, *Manschette*, *Schmiere*). These are then added to the previous definitions and presented to the two other annotators.[32]

[31] Figurative meanings listed in DWDS (2021) are treated as separate senses.

[32] The annotated data is available at `https://www.ims.uni-stuttgart.de/data/wugs` under DWUG DE V2.3.0.

	A	B	C	full
A		.84	.89	
B			.89	
full				**.87**

Table 3.8: Agreement (Krippendorff) between annotators on sense definition annotation.

Agreement We present the agreement of the three annotators in Table 3.8, where 134 judgments assigning 'other' are ignored. Krippendorff's alpha is high with .89 on the full data and similar for pairwise agreements. Percentage agreement (ITA) and pairwise Cohen's Kappa (Artstein and Poesio, 2008), which we do not report here, yield similar scores. According to Erk et al. (2013), sense annotation studies relying on the WordNet sense inventory show percentage agreement from .67 to .78. We conclude that our data set is sufficiently reliable to serve as a gold standard. We now use this data set to evaluate the clusterings derived with the clustering algorithm described in Section 3.1.4 on the human annotation of SemEval DE.

Evaluation We exclude 6 uses that have no annotation, 113 uses with at least one 'others' label and 255 uses with at least one disagreement between an annotator pair, leaving 826 labeled uses. We extract clusters from these labels by assigning each use to the same cluster label if they were assigned to the same sense definition by the annotators. Figure 3.7 shows the Adjusted Rand Index (ARI, Hubert and Arabie, 1985) comparing the SemEval DE clusterings to the sense definition clusterings. As we can see, the scores vary considerably between target words. While there is very high correspondence between the two strategies for some words (1.0), it is very low for others ($\approx$ .0). The average correspondence score is moderate to high with ARI=.65.[33]

We further check whether the correspondence can be improved by reducing the sparsity of the SemEval DE WUGs: the multi-round annotation process described in Section 3.2.1 did not converge for a number of words and was stopped for time constraints after round 4. We continue the annotation process with one more round

[33] Using *opt* clusterings from DWUG DE V1.0.0 (new runs of the clustering algorithm with optimized parameters) instead of *semeval* clusterings yields the same results (less than 0.01 improvement).

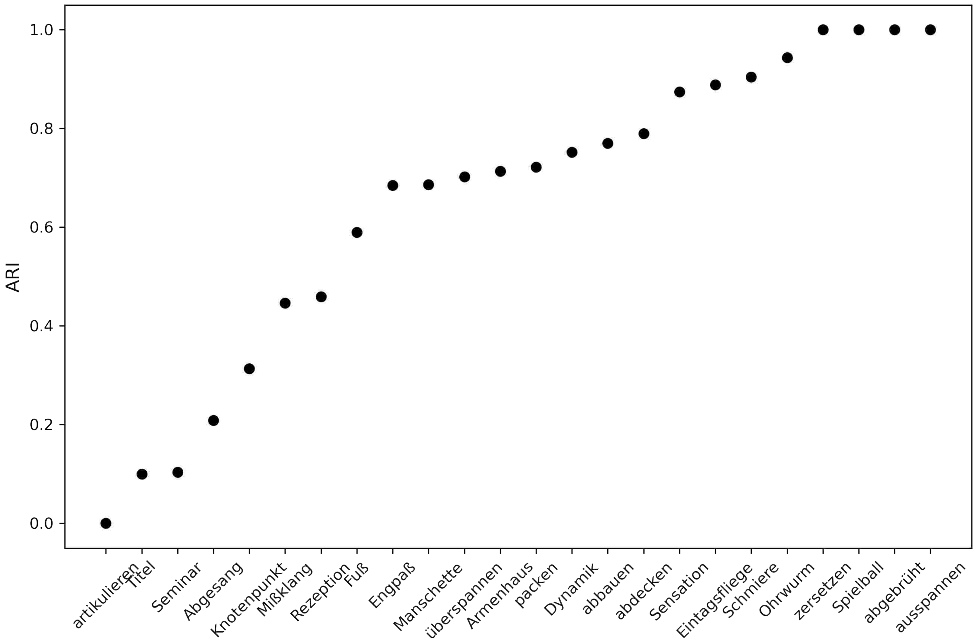


Figure 3.7: Correspondence (ARI) of clusters resulting from sense definitions vs. SemEval DE clusterings.

and obtain new clusterings on the updated graphs.[34] This considerably improves the average correspondence score to 0.74. Hence, the original SemEval DE WUGs likely suffer from sparsity (see also Section 3.3.2.2) resulting in noisy clusterings. They should hence be seen as a silver rather than a gold standard.

3.3.2.2 Manual Analysis

In the following section, we manually analyze what the SemEval DE clusterings reflect. We again choose *abbauen*, *abgebrüht*, *Knotenpunkt*, *Manschette* and *zersetzen* for the analysis to cover different POS and cases with low and high correspondence to the sense definition annotation. (Note that for reasons of the process of this thesis the analysis is done only for sense clusters derived from sense definition annotations of annotatorA, and not on the full annotated data.) For each target word, we find the best mapping between the two cluster assignments (WUG vs.

[34] The additional data is available at `https://www.ims.uni-stuttgart.de/data/wugs` under DWUG DE V2.0.0.

definition) with the Hungarian method (Kuhn, 1955). We then plot both clusterings in interactive WUG form, where corresponding clusters receive the same color (cf. Figures A.15 and A.16 in Appendix A). We first inspect each plot individually and judge whether the clusters represent intuitive sense distinctions. Then, we inspect uses which were assigned to different clusters by the two approaches, judge the assignments' intuitive validity and try to identify edges in the annotated WUGs which may have contributed to the inferred solution by the clustering approach. We identify five sources of disagreement between clustering approaches:

Ambiguity and unfamiliarity of use Many cases of disagreements between clustering strategies involve ambiguous uses or uses with unfamiliar meanings. Such cases of disagreement are not related to the annotation strategy; their cluster assignment rather depends on the interpretation of individual annotators.

Noise from connected use There is a number of clearly interpretable uses receiving surprising cluster assignments by the WUG approach. Many of these cases are connected to ambiguous uses introducing noise into the clustering procedure as the cluster assignment of a use depends on the assignments of connected uses and thus on judgments on these connected uses. This is a crucial challenge as it implies that any type of noise introduced on a particular use (e.g. stemming from ambiguity, vagueness or errors) influences the assignment of other uses connected to it. This can lead to counter-intuitive cluster assignments and in the worst case whole regions of a graph can be affected.

Sub-optimal clustering loss In some cases, based on the observed judgments, we note that the inferred WUG clustering is not optimal in terms of clustering loss (see Section 3.1.4). This is e.g. the case for *Knotenpunkt*. We re-cluster the WUGs and observe that often a better solution can be found. For *Knotenpunkt*, this improves the loss and the correspondence to the sense definition clustering. It does, however, not improve the average correspondence across target words. This is likely due to the fact that for some target words a less optimal solution is found after re-clustering (higher loss). However, this indicates that with various iterations of the clustering algorithm better solutions could be found further minimizing the loss, and possibly increasing correspondence between annotation approaches.

Sparsity Sparsity affects the clustering subtly in many regions of the graphs. It has a strong influence when combined with noisy annotation. But, we also identify cases where sparsity has a concretely observable effect: if not all clusters are guaranteed to have an edge between them, the clustering may split a cluster only because it has no information on their relation. We find this to occur with *Ohrwurm*, where one disagreement between clustering strategies can be traced back to a small cluster which was not compared to one of the two major clusters.

Sense definition bias A number of disagreements can be traced back to a tendency of the sense definition annotator to assign uses to the predefined sense definitions, as e.g. for *zersetzen*. The sense definition approach results in two clusters corresponding roughly to the sense definitions 'to dissolve' and 'to destroy'. The WUG approach infers two corresponding major clusters, one for the concrete chemical or biological meaning which could be described as 'to decompose/to dissolve', and one for the related abstract metaphorical meaning 'to destroy'. Disagreements between the two approaches are exclusively cases of uses from the concrete physical 'to decompose/to dissolve' sense inferred on the WUG annotated with the 'to destroy' sense definition. The annotator reported to perceive 'to dissolve' as a subsense of 'to destroy' and even annotated both definitions in two cases. While the annotator was aware of considerable variation between the uses labeled as 'to destroy' covering concrete physical and abstract metaphorical meanings, the definition still seemed adequate and sufficiently general to her to cover these uses. This example shows how the initial choice, granularity and interpretation of sense definitions can strongly influence the obtained clustering.

3.3.3 Change Scores

In this section, we point out the estimation character of the change scores derived in Sections 3.2.1 and 3.2.2 and their resulting limitations, and show connections and divergences between the change measures. (Find a discussion of formal connections between the measures in Appendix B.)

3.3.3.1 Robustness after Continued Annotation

We test the robustness of change scores after the additional round of annotation described in Section 3.3.2. Note that in this comparison at least annotation sparsity, annotation and clustering errors, as well as frequency thresholds have an influence. The Accuracy (Tharwat, 2020) of binary change labels between original and updated data with slightly different lower frequency thresholds n and k is 0.86/0.81/0.94 for EN/DE/SV while Spearman correlation between original and updated graded change scores is 0.89/0.98/0.82. This shows that both change scores are not completely robust. For binary change scores up to 19% of items change their labels.

3.3.3.2 Expected Error from Sense Sampling Variability

The derived change scores (see Section 3.1.5) depend on the distribution of the senses of the word uses sampled from the source corpus. (Remember that we only annotate a sample of word uses from the source corpus and not the full corpus.) This means that the final change scores derived on the annotated data are only estimates of the change scores that we would obtain from the full source corpus. Hence, there is an expected error stemming from the sampling variability on senses of word uses.[35] In the case where models are evaluated only on the sampled word uses, this is not a problem for us. However, if models are evaluated on the full corpus (see Chapter 5), we want the annotated estimates to be an accurate reflection of the change scores underlying the full corpus. We now estimate the expected error from sampling variability via simulation.

Assume that P and Q are discrete probability distributions giving the true (as obtained from the full corpus) word sense probability distributions for the use sets U_1 and U_2 for a target word w.[36] Now, assume that $\hat{P}$ and $\hat{Q}$ are estimates of these distributions obtained by sampling i.i.d. from P and Q.[37] For each of our change score estimators $\hat{\theta}$ of the true change score θ, we would like to measure the mean squared error:

$$MSE(\hat{\theta}) = E_{\theta}\left[(\hat{\theta} - \theta)^2\right] .$$

For $G(w)$, $\hat{\theta} = JSD(\hat{P}, \hat{Q})$ estimates the parameter $\theta = JSD(P, Q)$. For $B(w)$, we assume thresholds $n = 0$, $k = 0$, in this case $B(D, E) = B(P, Q)$ (see Appendix B). Then, $\hat{\theta} = B(\hat{P}, \hat{Q})$ estimates the parameter $\theta = B(P, Q)$.

[35] Additional error is expected from annotation errors or clustering errors.

[36] For simplicity, we ignore the problem that the derived cluster structure may depend on the sampled uses.

[37] For simplicity, we assume i.i.d. sampling from probability distributions instead of random sampling without replacement from word sense clusters.

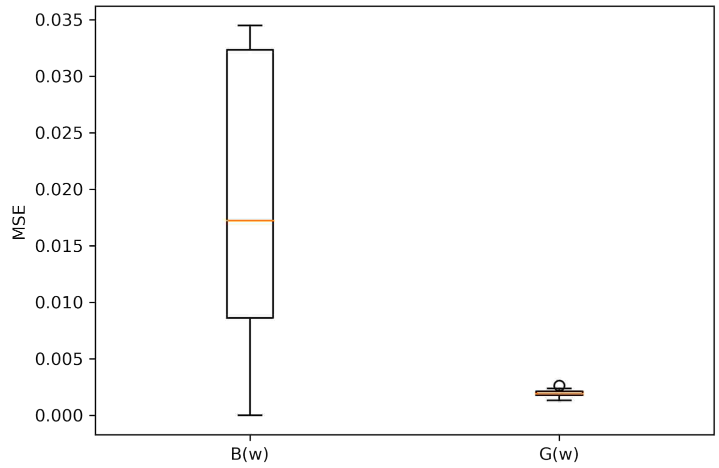


Figure 3.8: MSE of change score estimators on simulated distributions.

We now estimate the MSE by simulating realistic word sense distributions P and Q and sampling $\hat{P}$ and $\hat{Q}$ from these. For this, we take all target words from the SemEval data sets and treat the sense probability distributions which were inferred from the annotated data as P and Q. This yields realistic and often strongly skewed word sense distributions (Kilgarriff, 2004) with realistic changes. From these, we sample $\hat{P}$ and $\hat{Q}$ by drawing 100 samples i.i.d., which corresponds to the maximal sample size per corpus for the SemEval data, and compute the MSE. This process is repeated 50 times. We plot the distribution of the resulting MSE values in Figure 3.8.

$B(w)$ has higher MSE values than $G(w)$ and can thus be expected to be less robust to sampling error. Further skewing the distributions by multiplying the frequencies of the most dominant sense by 2 and 5 respectively increases $B(w)$'s median MSE to 0.03 and 0.06 respectively.

For $C(w)$, we assume that edge weights between uses of different senses in P and Q are constantly 1 while weights between uses of the same sense are constantly 4. This implies that a true probability distribution of COMPARE weights R can be defined, which is given by $R(4) = \sum_i^K p_i q_i$, $R(1) = 1 - \sum_i^K p_i q_i$, and that we can calculate $C(w)$ based on P and Q with the help of this distribution as $C(R) = -\big(3 * \sum_i^K p_i q_i + 1\big)$ (see Appendix B). We estimate $C(R)$ by $C(\hat{R})$. We are interested in how well $\hat{\theta} = C(\hat{R})$ estimates $\theta = JSD(P, Q)$, i.e., how well the es-

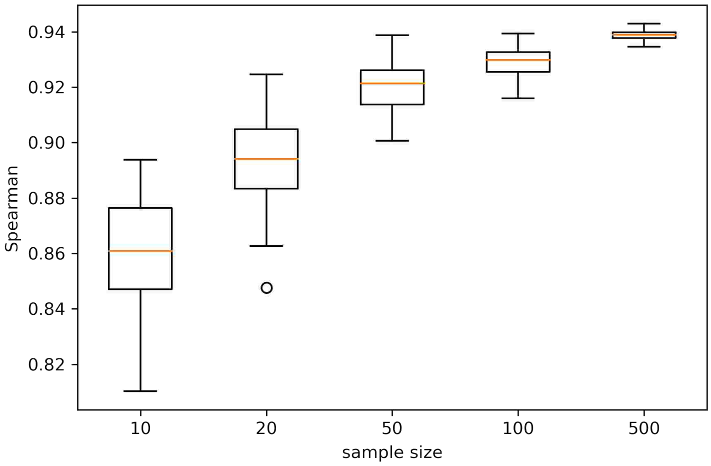


Figure 3.9: Spearman of Negated COMPARE estimator with true graded change score on simulated distributions, over various sample sizes.

timated Negated COMPARE score estimates the true graded change score. This time, we do not calculate the MSE, but measure the Spearman correlation between the two quantities as they cover different value ranges and because for the evaluation with Spearman described in Chapter 5, not absolute values, but the rankings of target words are important. We again simulate word sense distributions P and Q as described above, calculate the corresponding R and sample $\hat{R}$ from this with sample sizes from 10 to 500, yielding correlations between 0.87 and 0.94, as shown in Figure 3.9. Increasing sample sizes yield increasingly higher correlation between the two quantities. With a sample size of 500 per distribution, we reach the true correlation of 0.94 between $C(R)$ and $JSD(P, Q)$. This is a very interesting result because it means that we can estimate $JSD(P, Q)$ well with relatively small sample sizes via $C(\hat{R})$, for which no clustering is needed.[38] This result is further supported by the observed correlation between the graded change scores and the Negated COMPARE scores estimated from annotated data: On some data sets we can derive both scores (provided that they are densely enough annotated to be clustered and edges for annotation were randomly sampled). We do this for DiscoWUG (see Section 6.3) and compute the correlation between the two quantities, yielding 0.9, which confirms the result of the simulation.

[38] Skewing the distributions further, as above, has only a negligible impact on the results.

3.4 Discussion

In this chapter, we introduced a theoretically well-motivated annotation framework for word senses relying on human semantic proximity judgments with an additional automatic clustering step. We chose this novel approach because it built a close connection to Blank's theory of meaning and meaning change. It further formalizes the complex lexicographic clustering process and allows to control parameters on each of the steps in the process, such as granularity of the clustering. In contrast, these steps are hidden and may vary between lexicographers in the traditional sense definition annotation approach. Our approach is also simple because the only human input needed are semantic proximity judgments while for the traditional approach sense definitions have to be extracted beforehand, requiring additional human efforts. The procedure can be easily extended to obtain data for more than two time points.

However, the approach also has major challenges: because judgments are done on pairs of word uses (instead of individual uses), the possible number of annotations increases quadratically. Thus, only a subset of use pairs can be annotated for each word. This leads to sparse graphs which we identified as a reason for clustering errors. There is above-chance, but mostly moderate agreement on the concept of semantic proximity, ambiguity affects the clustering procedure and there are valid alternatives to the clustering procedure we chose. The correspondence with the traditional annotation approach is moderate to high, but could be increased with an additional round of annotation. We hence consider the clusterings and change scores rather a silver than a gold standard. Note also that the change scores resulting from our annotation process by themselves are measurements of **word sense divergences** and do not necessarily imply LSC in the sense of being sufficiently widespread amongst speakers of a language (see Section 2.2). Only in connection with an adequate word use sampling procedure guaranteeing to cover a wide range of speakers, these divergences can be argued to indicate LSC. While some of the corpora we rely on (DTA, CCOHA) do cover a wide range of written language, the representativeness of any text corpus for a particular language is questionable (Koplenig, 2019). Note, however, that we have sampled many target words from historical dictionaries resulting from detailed analyses of historical linguists, which is a (more or less) independent confirmation of the annotated changes for our target words.

The binary change score we defined is prone to sampling error and other sources of noise while the graded change score is more robust. After an additional round of annotation, both scores showed a high robustness (Accuracy and Spearman $> .94$) for some data sets, but considerable variation for others ($\approx .8$). We have shown empirically that the Negated COMPARE score can be used as a simple approximation of the graded change score and shares some formal properties with it (see Appendix B).

We did not train the annotators with explicit reference to Blank's scale and his criteria. Our instructions and guidelines were rather loose adaptions of existing, similar guidelines. The correspondence between our annotation and Blank's theory is thus not strictly given. In the future, it would be interesting to see whether exact implementations of his criteria are possible in practical annotation studies and whether they yield similar results to our annotation.

A general challenge for annotation of historical data is the question how well modern annotators can grasp historical meanings. They could be biased towards their modern interpretations of words and infer the wrong meanings. We tried to reduce this possibility by choosing rather late time periods for annotation and mixed in historical annotators. Although these have higher agreement with each other, non-historical annotators significantly agree with historical annotators.

By inferring hard clusters in Section 3.1.4, we assumed word senses to be discrete. While this is probably a simplification because some uses can be assigned to more than one sense (Erk et al., 2013), it is necessary to obtain a straightforward measurement of semantic change that corresponds to Blank's definition (see Section 3.1.5). It is unclear how to define binary change without relying on discrete sense assignments. Note, however, that alternative clustering algorithms, e.g. using soft-clustering (Jurgens and Klapaftis, 2013), and alternative change score definitions can be applied to our WUGs.

While for the clustering approach described in Section 3.1.4 we did not make the assumption of **true sense clusters**, in the simulations from Section 3.3.3.2 and Appendix D we had to make this assumption. This way to treat word senses is common in probabilistic modeling of word meaning, where sense clusters can be treated as a latent variable (e.g. Blei, 2012; Perrone et al., 2019) that can be inferred from the observed variables, i.e. in our case, from human semantic proximity judgments (Peixoto, 2017). However, aside from simulation purposes, we do not take this view for two reasons: (i) It makes the strong ontological assumption that a hidden distribution of word senses exists. While this may be more in line with Blank's view (see Section 2.3.3.1), it conflicts with the view taken by notable lexicographers such as Kilgarriff (1997) that word senses are **produced** rather than **found** (see Section 2.3.1). (ii) Correlation clustering, as defined in Section 3.1.4, does not make the assumption of latent word senses. It rather optimizes a global heuristic on the graph. However, one could argue that by estimating the semantic proximity between use pairs from the median over annotator judgments we implicitly assumed that there is one **true semantic proximity** value for each use pair. Consequently, we have to treat disagreement between annotators as measurement errors (see also Appendix D). The disagreement on semantic proximity judgments observed in Section 3.3.1 indicates a large measurement error. This raises the question whether annotator disagreements should not rather be seen as different annotator interpretations of the concept of semantic proximity.

The annotation annotation procedure developed in this chapter was applied in several further studies on various languages (Giulianelli et al., 2020; Rodina and Kutuzov, 2020; Kutuzov and Pivovarova, 2021b; Kurtyigit et al., 2021; Zamora-Reina et al., 2022; Baldissin et al., 2022; Kutuzov et al., 2022; Aksenova et al., 2022). It was further validated in a simulation study (Kotchourko, 2021) and has been implemented into an online annotation interface.[39] The annotated data was made part of a LSCD benchmark and has further potential uses, including the fine-tuning of meaning representations (such as contextualized embeddings) and the use for modern tasks, e.g. to improve robustness against timeshifts.[40]

[39] `https://www.ims.uni-stuttgart.de/data/durel-tool`

[40] `https://github.com/ChangeIsKey/LSCDBenchmark`

Chapter 4

Computational Measurement

In this chapter, we describe the computational models used to measure LSC. They can be distinguished into (i) models building one meaning representation for each word use (**token-based**) and (ii) models building an aggregated meaning representation across a word's uses (**type-based**). With the former we can model the annotation process described in Section 3.1 and can thus give a good argument why the model should be able to measure LSC while this is not easily possible with the latter models. All models share that (i) they are based on the distributional hypothesis (Harris, 1954) in the sense that they infer semantic representations from word co-occurrences. (ii) They are trained in an unsupervised way, i.e., do not rely on manually labeled input. (iii) They are all (with few exceptions) Vector Space Models (VSMs) as they represent meaning as vectors in a vector space (Turney and Pantel, 2010).

4.1 Token-based VSMs

Token-based VSMs (Schütze, 1998) are nowadays more commonly known as **contextualized embeddings**. They represent the meaning of each word use as a vector in a vector space, allowing to measure distances between vectors. These distances can be seen as a model of negated semantic proximity as they show significant correlation with human semantic proximity judgments (Giulianelli et al., 2020; Arefyev et al., 2021; Pilehvar and Camacho-Collados, 2019; Armendariz et al., 2020).[1] These vectors can then be clustered and change can be measured on the

[1] Note that the usual distance measures between vectors give continuous values while the annotated semantic proximity from Section 3.1.2 are discrete values between 1 and 4. In the future, this

resulting cluster frequency distributions in the same way as on the human annotation, described in Section 3.1.5. Hence, this approach provides a complete model of the derivation process of lexicographic senses. However, it is also possible to aggregate token-based vectors into a type-level meaning representation or to define semantic change measures directly on the set of word vectors in order to avoid the clustering step, which introduces additional difficulties.

Token-based LSCD models are typically composed of

1. a semantic representation mapping each word use to a vector,
2. a clustering method (optional) and
3. a change measure.

Typically, token-based models do not need an alignment step, serving to make vectors trained on different data comparable (see Section 4.2) as vectors for word uses from different time periods are extracted from the same semantic representation. In the following section, we describe the model components used in our experiments.

4.1.1 Semantic Representations

A pretrained token-based VSM can be seen as a function mapping a word use to a vector reflecting its local co-occurrence statistics (within the particular use). We use token-based VSMs to map a set of word uses U of a target word w to a matrix M, where each row M_{i*} corresponds to the token vector of the ith use in U.[2] The time-specific use subsets U_1 and U_2 then correspond to the time-specific sub-matrices M_1 and M_2.

4.1.1.1 BERT

We map word uses to the token vector matrix M with Bidirectional Encoder Representations from Transformers (BERT, Devlin et al., 2019). BERT is a neural language model based on stacked encoders from transformers (Vaswani et al., 2017), inferring contextualized meaning representations for word uses from co-occurrence statistics. It is trained solving two unsupervised tasks simultaneously by minimizing their combined loss function:

discrepancy could be avoided by training models directly on annotated data (cf. Arefyev et al., 2021).

[2] For simplicity we assume here that U is ordered.

1. **Masked Language Model**: Randomly replace some words in the input use by a [MASK] token. Predict the masked words based on the context provided by the other non-masked words.

2. **Next Sentence Prediction**: Given two sentences A and B, predict whether B is the sentence that comes after A, or a random sentence from the corpus.

As displayed in Figure 4.1, first the input use is split into (subword) tokens which are then mapped onto token vectors (embeddings).[3] These are then passed to the first encoder (layer). Each encoder consists of a self-attention layer passing its output to a feed-forward neural network. The ouput of each encoder are contextualized vector representations for each token in the input. These are then used as input to the next encoder.

The self-attention mechanism is a key component of BERT. It contextualizes the vectors of the words in the input use by allowing them to interact. Each self-attention layer has multiple self-attention heads receiving token vectors as input (see Figure 4.2). For every vector, a key, query and value vector with 64 components is created by matrix multiplication with the respective key, query and value matrices, which are learned during training (Allamar, 2021). Then, the following major steps are executed for every token:

1. calculate the dot (scalar) products between the token's query vector and the key vectors of all tokens,

2. normalize these values using softmax (attention scores),

3. create a new (contextualized) token vector by a linear combination of the value vectors of all tokens where the attention scores are coefficients.

The transformed token vectors now encode information about their context. This self-attention mechanism is run simultaneously in multiple heads in the same layer using different key, query and value matrices. The token vectors from the different attention heads are then concatenated and passed to the feed-forward neural network. We use the *bert-base* model, which has 12 self-attention-heads per layer, resulting in contextualized token vectors with 768 components after concatenation. The model uses 12 layers (see Figure 4.1).

[3] The BERT tokenizer often maps words to subwords.

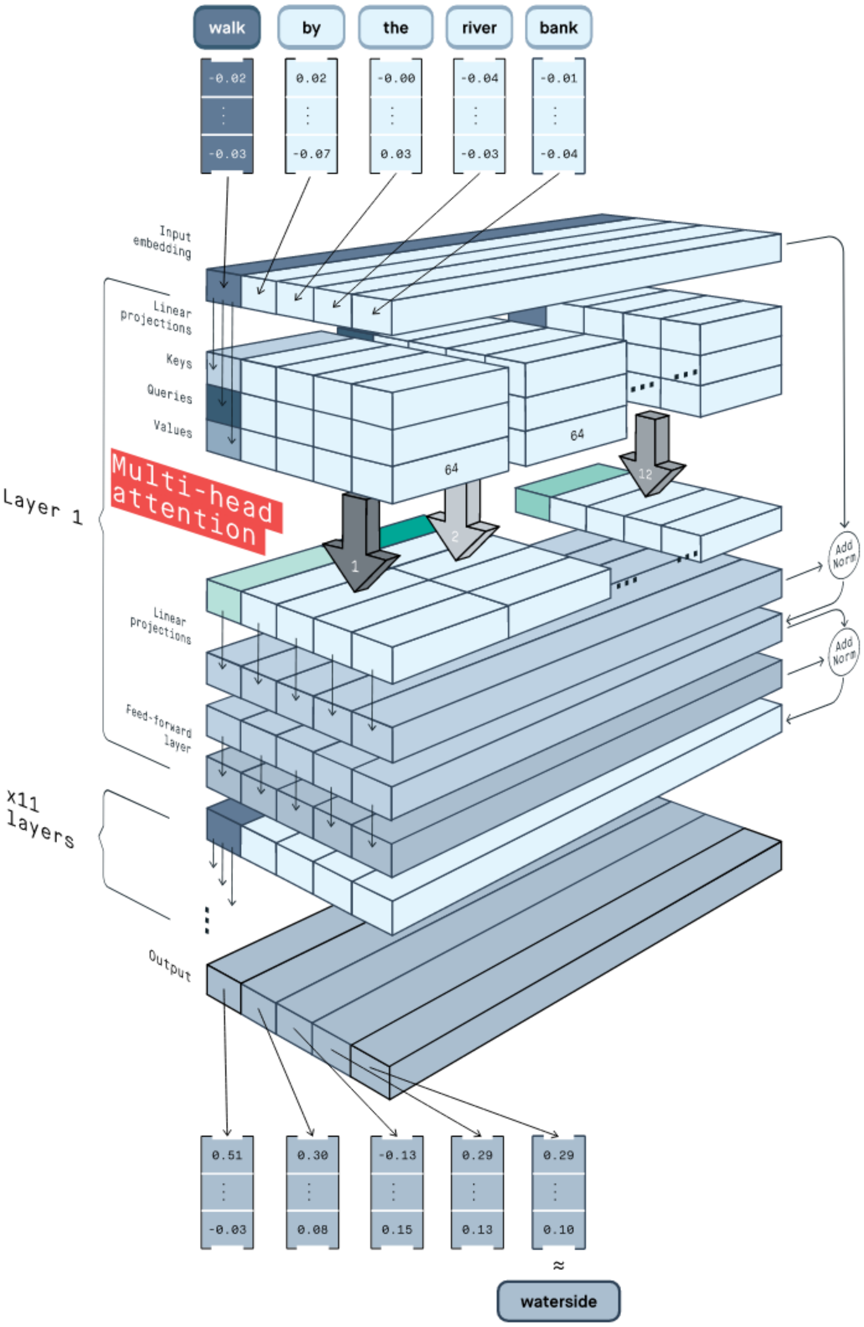


Figure 4.1: The BERT model (Futrzynski, 2021; Kurtyigit, 2021).

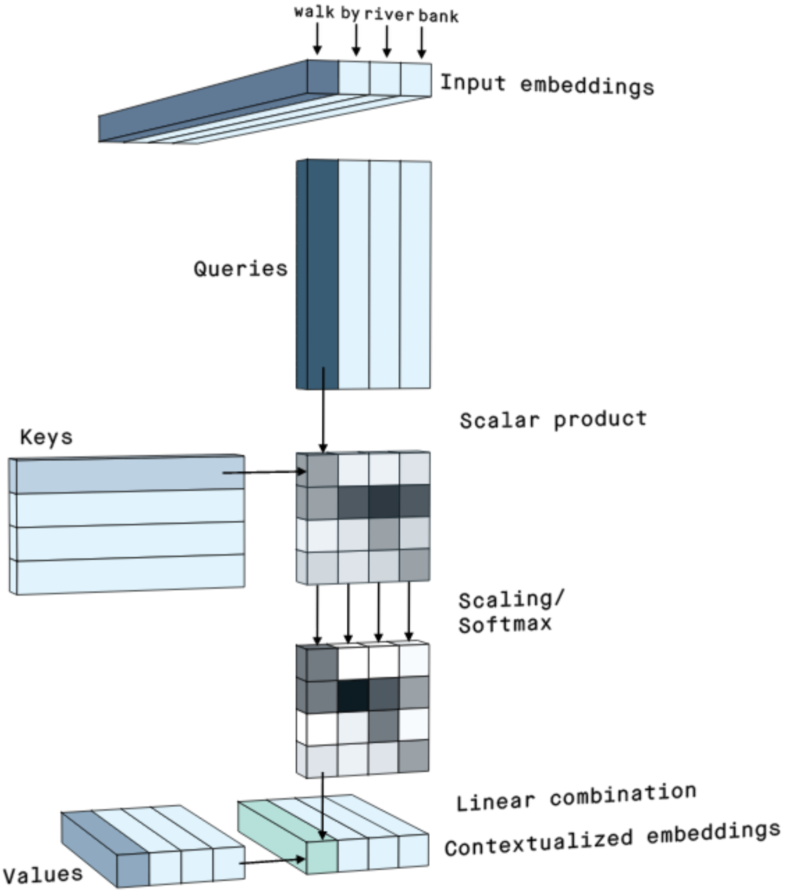


Figure 4.2: BERT's Self-Attention Head (Futrzynski, 2021; Kurtyigit, 2021).

BERT is pretrained on large amounts of data, in contrast to the type-based presentations described in Section 4.2. We create the contextualized token vector matrix M from this pretrained model by feeding it with individual word uses, extracting contextualized token vectors for the target word from one of the 12 different layers or as the average over multiple of those, and storing the resulting vectors in M.[4] In case a target word corresponds to multiple token vectors, we average all subword tokens to obtain the final target word representation.

[4] Uses with more than 512 tokens (max. input sequence length for BERT) are truncated to that length.

4.1.2 Clustering Methods

Token-vectors in a matrix M can be clustered based on their distances using a wealth of clustering algorithms (Aggarwal and Reddy, 2013). We experiment with a very common hierarchical clustering algorithm.

4.1.2.1 Agglomerative Clustering

Agglomerative Clustering (AGL) is a hierarchical clustering algorithm commonly used for Word Sense Induction (Panchenko et al., 2018; Amrami and Goldberg, 2018). We first length-normalize the vectors in the token matrix M and then cluster them with AGL. The algorithm starts with each vector in an individual cluster and then repeatedly merges those two clusters which maximize a predefined criterion, based on distances between vectors. For this, we use Ward's method with Euclidean distances, minimizing the total within-cluster variance (Ward Jr, 1963). At each step it finds the pair of clusters that leads to minimum increase in total within-cluster variance after merging. Within-cluster variance of a cluster c is measured as the error sum of squared distances:

$$ESS = \sum_{i=1}^{n} d(\vec{x}_i, \vec{m}_c)^2$$

where $\vec{x}_i$ is the ith vector in cluster c, $\vec{m}_c$ is the centroid of vectors from cluster c and $d(\vec{x}, \vec{y}) = \sum_{i=1}^{n} \sqrt{(\vec{x}_i - \vec{y}_i)^2}$ is the Euclidean distance. Uses with similar vector representations (small distances) will tend to be assigned to the same cluster because merging them does not strongly increase within-cluster variance.

Cluster merging is performed iteratively until a predefined number of clusters k is obtained. Following Giulianelli et al. (2020) and Martinc et al. (2020a), we estimate the number of clusters k with the **Silhouette Method** (Rousseeuw, 1987): we perform a cluster analysis for each $2 \leq k \leq 10$ and calculate the silhouette index for each k. The number of clusters with the largest index is used for the final clustering.

4.1.3 Change Measures

We define two types of measures on token-vector matrices: (i) Clustering-based measures exploit the cluster structure (derived as described in Section 4.1.2) as in the annotation process (described in Section 3.1.5). (ii) Average measures avoid the clustering step by averaging over vectors or distances between vectors. A common feature with the type-based similarity measures described in Section 4.2.3.1 is the aggregation of similarity information over uses.

4.1.3.1 Clustering-based Measures

From the clustering of matrix M, we obtain an assignment of each word use in U to a cluster and the corresponding cluster frequency distribution (see Section 3.1.4). By splitting the uses into time-specific subsets U_1 and U_2, we obtain the corresponding time-specific cluster frequency distributions D and E as well as the respective cluster probability distributions P and Q. From these, we can measure **binary** and **graded** semantic change in the same way as defined in Section 3.1.5, i.e., as $B(D, E)$ and $JSD(P, Q)$. Hence, these measures can be seen as a model of the binary and graded change scores.

4.1.3.2 Average Measures

Average measures avoid the clustering step by either averaging over distances between vectors from M_1 and M_2 (APD) or by averaging over vectors within M_1 and M_2 (COS). APD models the Negated COMPARE score and can therefore be motivated similarly as approximating graded change (see Section 3.1.5). In contrast, COS can be motivated comparably to similarity measures on type-based embeddings, making the assumption that sense frequency changes correlate with changes in the global co-occurrence statistics for a word between time periods (see Section 4.2.3). We treat both measures as models of graded change.

Average Pairwise Distance (APD) is calculated by averaging over all pairwise distances between the vectors from the two matrices (cf. Schlechtweg et al., 2018; Giulianelli et al., 2020):

$$APD(M_1, M_2) = \frac{1}{|D_{1,2}|} \sum_{x \in D_{1,2}} x$$

where $D_{1,2} = [d(\vec{x}, \vec{y}) | (\vec{x}, \vec{y}) \in M_1 \times M_2]$ (COMPARE distances) and $d = CD$ (cosine distance, see Section 4.2.3.1). Note that this corresponds to the definition of Negated COMPARE in Section 3.1.5. We first randomly sample n vectors from both matrices without replacement, resulting in the sampled matrices $\hat{M}_1$ and $\hat{M}_2$

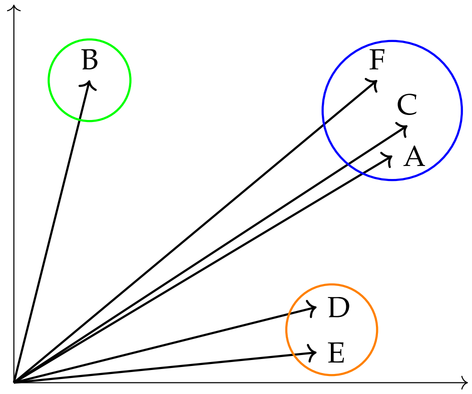


Figure 4.3: Visualization of a 2-dimensional token-based VSM. Vectors represent use meanings, distances between vectors represent negated semantic proximity, and colored clusters represent word senses.

and calculate $APD(\hat{M_1}, \hat{M_2})$. We determine n as the minimum size of M_1 and M_2.

APD-OLD/NEW measure the average of pairwise distances within M_1 and M_2, respectively. They are calculated as the average distance of max. $10,000$ unique combinations of vectors from either M_1 and M_2, i.e., $APD(M_1, M_1)$ and $APD(M_2, M_2)$ when excluding duplicates and ignoring order and reflexive pairs. APD-OLD/NEW measure the within-period semantic variation of a target word.

COS is calculated as the cosine distance of the respective mean (centroid) vectors for M_1 and M_2 (Kutuzov and Giulianelli, 2020):

$$COS(M_1, M_2) = CD(\mu(M_1), \mu(M_2)) .$$

Similar to cosine distance on type-based representations, COS can be treated as a model of the graded change score.

4.1.4 Example

Figure 4.3 shows a simplified example of a typical token-based LSCD model representing the uses of the word *arm* from the corpus shown in Table 3.2 in a two-dimensional vector space. Each use meaning is represented by a vector, where distances between vectors can be interpreted as negated semantic proximity between the corresponding uses. The vectors can be clustered based on their distances and clusters can be interpreted as word senses (compare Figure 4.3 to Figure 3.1). Assuming that time periods are 1820–1860 and 1950–1990, time-specific sense frequency distributions can be extracted from the clustering and LSC can then be measured with binary and graded change as described in Section 4.1.3.1.[5]

4.2 Type-based VSMs

Type-based VSMs do not model the annotation process from Section 3.1 as they only provide one semantic representation per word (type), which can be seen as an average meaning representation over the uses of the word. It is not possible to obtain sense clusters from this representation in a straightforward manner.

Type-based LSCD models are typically composed of

1. a semantic representation mapping each word to a vector,
2. an alignment method and
3. a change measure.

Usually, one vector is learned for each time period, representing the word's meaning aggregated for that period of time. These time-specific representations are then aligned to achieve comparability and finally comparison is performed using a selected change measure.

4.2.1 Semantic Representations

A type-based VSM representation can be seen as a function mapping each word in the vocabulary V of a corpus C (e.g. defined as a set of word uses from different words) to a vector reflecting its global co-occurrence statistics in C (across all its uses in the corpus, cf. Turney and Pantel, 2010). The vectors of all words can be represented in a matrix M where each row vector M_{i*} represents the ith word in the vocabulary V. Target words are usually only a subset of the words in the vocabulary. We construct vector spaces for each time period and compare two state-of-the-art approaches to learn these vectors from co-occurrence data, (i) counting

[5] A possible intermediate step in this process, which we omit here, would be to represent negated pairwise vector distances in a graph in order to apply graph clustering techniques as the one described in Section 3.1.4.

and (ii) predicting. The latter type of vector representations are often called **word embeddings**. All representations are bag-of-words-based, i.e., each word representation reflects a weighted bag of context words. The context words of a target word w are given by the words surrounding it in an n-sized window within its uses in C.[6]

4.2.1.1 Count-based VSMs

In a raw count VSM (CNT), the matrix M is high-dimensional and sparse. The value of each matrix cell $M_{i,j}$ represents the number of co-occurrences of the word w_i and the context word c_j, $\#(w_i, c_j)$. In line with Hamilton et al. (2016b), we apply a number of transformations to these raw co-occurrence matrices as previous work has shown that this improves results on different tasks (Bullinaria and Levy, 2012; Levy et al., 2015).

Positive Pointwise Mutual Information (PPMI) In PPMI representations the co-occurrence counts in each matrix cell $M_{i,j}$ are weighted by the positive mutual information of target w_i and context c_j reflecting their degree of association. The values of the transformed matrix are

$$M^{\text{PPMI}}_{i,j} = \max\left\{\log\left(\frac{\#(w_i, c_j)\sum_c \#(c)^\alpha}{\#(w_i)\#(c_j)^\alpha}\right) - \log(k), 0\right\}$$

where $k > 1$ is a prior on the probability of observing an actual occurrence of (w_i, c_j) and $0 < \alpha < 1$ is a smoothing parameter reducing PPMI's bias towards rare words (Levy and Goldberg, 2014; Levy et al., 2015).

Singular Value Decomposition (SVD) Truncated SVD finds the optimal rank d factorization of matrix M with respect to L2 loss (Eckart and Young, 1936). We use truncated SVD to obtain low-dimensional approximations of the PPMI representations by factorizing M^{PPMI} into the product of the three matrices $U\Sigma V^\top$. We keep only the top d elements of Σ and obtain

$$M^{\text{SVD}} = U_d \Sigma^p_d$$

where p is an eigenvalue weighting parameter (Levy et al., 2015). The ith row of M^{SVD} corresponds to w_i's d-dimensional representation.

[6] Find details on hyperparameter settings in Section 5.4.2.

Random Indexing (RI) is a dimensionality reduction technique based on the Johnson-Lindenstrauss lemma according to which points in a vector space can be mapped into a randomly selected subspace under approximate preservation of the distances between points if the subspace has a sufficiently high dimensionality (Johnson and Lindenstrauss, 1984; Sahlgren, 2004). We reduce the dimensionality of a count-based matrix M by multiplying it with a random matrix R:

$$M^{\mathrm{RI}} = MR^{|\mathcal{V}|\times d}$$

where the ith row of M^{RI} corresponds to w_i's d-dimensional semantic representation. The choice of the random vectors corresponding to the rows in R is important for RI. We follow previous work (Basile et al., 2015) and use sparse ternary random vectors with a small number s of randomly distributed -1s and $+1$s, all other elements set to 0, and we apply subsampling with a threshold t.

4.2.1.2 Predictive VSMs

Skip-Gram with Negative Sampling (SGNS) differs from count-based techniques in that it directly represents each word $w \in V$ and each context $c \in V$ as a dense d-dimensional vector by implicitly factorizing $M = WC^{\top}$ when solving

$$\operatorname*{argmax}_{\theta} \sum_{(w,c)\in D} \log \sigma(v_c \cdot v_w) + \sum_{(w,c)\in D'} \log \sigma(-v_c \cdot v_w)$$

where $\sigma(x) = \frac{1}{1+e^{-x}}$, D is the set of all observed word-context pairs and D' is the set of randomly generated negative samples (Mikolov et al., 2013a,b; Goldberg and Levy, 2014). The optimized parameters θ are $v_c = C_{i*}$ and $v_w = W_{i*}$ for $c, w \in V$, where v_c and v_w are d-dimensional vector representations for w and c. D' is obtained by drawing k contexts from the empirical unigram distribution $P(c) = \frac{\#(c)}{|D|}$ for each observation of (w,c) (cf. Levy et al., 2015). SGNS and PPMI representations are highly related in that the cells of the implicitly factorized matrix M correspond to PPMI values shifted by the constant k (Levy and Goldberg, 2014). Hence, SGNS and PPMI share the hyper-parameter k. The final SGNS matrix is given by

$$M^{\mathrm{SGNS}} = W$$

where the ith row of M^{SGNS} corresponds to w_i's d-dimensional semantic representation. As in RI we apply subsampling with a threshold t. SGNS with particular parameter configurations has shown to outperform transformed count-based techniques on a variety of tasks (Baroni et al., 2014; Levy et al., 2015).

4.2.2 Alignment Methods

In order to make the matrices A and B learned from the corpora C_1 and C_2 with vocabularies V_1 and V_2 comparable, they have to be aligned via a common coordinate axis. Different semantic representations demand different ways of alignment.

Column Intersection (CI) Alignment can be done rather straightforwardly for CNT and PPMI representations because their columns correspond to context words which often occur in both A and B (Hamilton et al., 2016b). In this case, the alignment for A and B is

$$\begin{aligned} A^{\text{CI}}_{*i} &= A_{*w_i} \quad \text{for all } w_i \in V_1 \cap V_2 \\ B^{\text{CI}}_{*i} &= B_{*w_i} \quad \text{for all } w_i \in V_1 \cap V_2 \end{aligned}$$

where A_{*i} denotes the ith column in A and A_{*w_i} denotes that column in A which corresponds to w_i (similarly for B).

Shared Random Vectors (SRV) RI offers an elegant way to align CNT spaces and reduce their dimensionality at the same time (Basile et al., 2015). Instead of multiplying count matrices A and B each by a separate random matrix R_A and R_B (cf. RI from Section 4.2.1), they may be multiplied both by the same random matrix R representing them in the same low-dimensional random space (SRV). Hence, A and B are aligned by

$$\begin{aligned} A^{\text{SRV}} &= AR \\ B^{\text{SRV}} &= BR\,. \end{aligned}$$

We follow Basile et al. and adopt a slight variation of this procedure: Instead of multiplying both matrices by exactly the same random matrix (corresponding to an intersection of their columns), we first construct a unified random matrix and then multiply A and B by the respective sub-matrices.

Orthogonal Procrustes (OP) In the low-dimensional vector spaces produced by SVD, RI and SGNS, the columns may represent different coordinate axes and thus they cannot directly be aligned to each other. Following Hamilton et al. (2016b), we apply OP analysis to solve this problem. Following Artetxe et al. (2017), we define a dictionary D as a binary matrix so that $D_{i,j} = 1$ if $w_i \in V_2$ (the ith word in the vocabulary of C_2) corresponds to $w_j \in V_1$. The goal is then to find the optimal mapping matrix W^* such that the sum of squared Euclidean distances between B's mapping $B_{i*}W$ and A_{j*} for the dictionary entries $D_{i,j}$ is minimized:

$$W^* = \underset{W}{\operatorname{argmin}} \sum_i \sum_j D_{i,j} \|B_{i*}W - A_{j*}\|^2$$

where A_{i*} denotes the ith row of A (similarly for B). Following standard practice, we length-normalize and mean-center A and B in a preprocessing step (Artetxe et al., 2017) and constrain W to be orthogonal, which preserves distances within each time period. Under this constraint, minimizing the squared Euclidean distance becomes equivalent to maximizing the dot product when finding the optimal rotational alignment (Hamilton et al., 2016b; Artetxe et al., 2017). The optimal solution for this problem is then given by $W^* = UV^\top$, where $B^\top DA = U\Sigma V^\top$ is the SVD of $B^\top DA$ (Artetxe et al., 2017). Hence, A and B are aligned by

$$A^{\text{OP}} = A$$
$$B^{\text{OP}} = BW^*$$

where A and B correspond to their preprocessed versions. We also experiment with two variants: (i) OP$_-$ omits mean-centering (Hamilton et al., 2016b), which is potentially harmful if the spaces are not only rotated but also translated in the embedding training process. (ii) OP$_+$ corresponds to OP with additional pre- and postprocessing steps and has been shown to improve performance in research on bilingual lexicon induction (Artetxe et al., 2018a,b). We apply all OP variants only to the low-dimensional matrices.

Vector Initialization (VI) In VI, we first learn A^{VI} using standard SGNS and then initialize the SGNS model for learning B^{VI} on A^{VI} (Kim et al., 2014). The idea is that if a word is used in similar contexts in C_1 and C_2, its vector will be updated only slightly while more different contexts lead to a stronger update.

Word Injection (WI) Finally, we use the WI approach by Ferrari et al. (2017) where target words are substituted by a placeholder in one corpus before learning semantic representations and a single matrix M^{WI} is constructed for both corpora after mixing their sentences. The advantage of this approach is that all vector learning methods described above can be directly applied to the mixed corpus and target vectors are constructed directly in the same space so that no post-hoc alignment is necessary. A very similar approach is Temporal Referencing (Dubossarsky et al., 2019), where the substitution is only done for uses where a word is considered the target and not where it is considered a context word for another target.

4.2.3 Change Measures

From aligned type-based semantic representations, we retain two time-specific vector representations $\vec{x}$ and $\vec{y}$ learned from the time-specific corpora C_1 and C_2 for each target word. Type-based measures predict change scores by comparing $\vec{x}$ and $\vec{y}$. They either capture the similarity of the vectors or changes in their predictability. Similarity-based measures are based on the assumption that sense frequency changes correlate with changes in the global co-occurrence statistics. Dispersion-based measures instead rely on the more specific assumption that sense frequency changes correlate with changes in the predictability of the global co-occurrence statistics. We will treat all measures described in this section as models of the graded change score.

4.2.3.1 Similarity Measures

Cosine Distance (CD) is based on cosine similarity, which measures the cosine of the angle between two non-zero vectors $\vec{x}, \vec{y}$ with equal magnitudes (Salton and McGill, 1983):

$$cos(\vec{x}, \vec{y}) = \frac{\vec{x} \cdot \vec{y}}{\sqrt{\vec{x} \cdot \vec{x}}\sqrt{\vec{y} \cdot \vec{y}}} \,.$$

The cosine distance is then defined as

$$CD(\vec{x}, \vec{y}) = 1 - cos(\vec{x}, \vec{y}) \,.$$

Local Neighborhood Distance (LND) computes a second-order similarity for two non-zero vectors $\vec{x}, \vec{y}$ (Hamilton et al., 2016a). It measures the extent to which $\vec{x}$ and $\vec{y}$'s distances to their shared nearest neighbors differ. First, the cosine similarity of $\vec{x}$ and $\vec{y}$ with each vector in the union of the sets of their k nearest neighbors

$N_k(\vec{x})$ and $N_k(\vec{y})$ is computed and represented as vectors $\vec{s_x}$ and $\vec{s_y}$, whose entries are given by

$$\vec{s}_x(i) = \cos(\vec{x}, \vec{z_i}) \quad \forall \vec{z_i} \in N_k(\vec{x}) \cup N_k(\vec{y})$$
$$\vec{s}_y(i) = \cos(\vec{y}, \vec{z_i}) \quad \forall \vec{z_i} \in N_k(\vec{x}) \cup N_k(\vec{y}) \; .$$

LND is then computed as cosine distance between the two vectors $\vec{s_x}$ and $\vec{s_y}$:

$$LND(\vec{x}, \vec{y}) = CD(\vec{s_x}, \vec{s_y}) \; .$$

LND does not require matrix alignment because it measures the distances to the nearest neighbors in each space separately. It was claimed to capture changes in paradigmatic rather than syntagmatic relations between words (Hamilton et al., 2016a).

4.2.3.2 Dispersion Measures

Frequency Difference (FD) The log-transformed relative frequency of a word w for a corpus C is defined by

$$F(w, C) = \log_2 \frac{\#(w)}{|C|}$$

where $\#(w)$ is the number of occurrences of w in C and $|C|$ is the total corpus size. FD for word w in two corpora C_1 and C_2 is then defined by the absolute difference in F:

$$FD(w, C_1, C_2) = |F(w, C_1) - F(w, C_2)| \; .$$

In Section 5.2.2.1, we also apply this measure without log-transformation.

Type Difference (TD) is similar to FD, but based on vectors. The normalized log-transformed number of context types of a raw count vector $\vec{x}$ learned from corpus C is defined by

$$T(\vec{x}, C) = \log_2 \frac{\sum_{i=1} 1 \quad \text{if } \vec{x}_i \neq 0}{|C_T|}$$

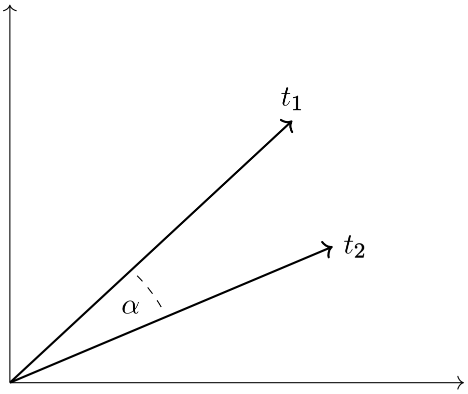


Figure 4.4: Visualization of a 2-dimensional type-based VSM. Vectors represent an aggregated time-specific meaning representation for a word.

where $|C_T|$ is the number of types in corpus C. The TD of two vectors $\vec{x}$ and $\vec{y}$ learned from two corpora C_1 and C_2 is the absolute difference in T:

$$TD(\vec{x}, C_1, \vec{y}, C_2) = |T(\vec{x}, C_1) - T(\vec{y}, C_2)| \, .$$

Entropy Difference (HD) relies on vector entropy as suggested by Santus et al. (2014). The entropy of a non-zero raw count vector $\vec{x}$ is defined by

$$VH(\vec{x}) = -\sum_{i=1} \frac{\vec{x}_i}{\sum_{j=1} \vec{x}_j} \log_2 \frac{\vec{x}_i}{\sum_{j=1} \vec{x}_j} \, .$$

VH is based on Shannon's entropy (Shannon, 1948) and measures the unpredictability of w's co-occurrences (Schlechtweg et al., 2017). HD is defined as

$$HD(\vec{x}, \vec{y}) = |VH(\vec{x}) - VH(\vec{y})| \, .$$

We also experiment with normalizing H dividing it by its maximum value, which is the logarithm of the number of context types in $\vec{x}$.

4.2.4 Example

Figure 4.4 shows a simplified example of a typical type-based LSCD model, representing the meaning of the word *arm* for each time period as a two-dimensional "average" vector learned over the uses from that time period. Uses are given by Table 3.2 and time periods are assumed to be 1820–1860 and 1950–1990. Once the vectors for t_1 and t_2 have been aligned, they can be represented in the same vector space and LSC can be measured with a distance measure such as the cosine distance.

4.3 Topic Models

We also experiment with one Topic Model. Sense ChANGe (SCAN) models LSC via smooth and gradual changes in associated topics (Frermann and Lapata, 2016). If topics are assumed to model word senses, Topic Models can model both, the binary and the graded change score. However, we experiment only with measures for the graded change score.

4.3.1 Semantic Representations

Assuming two time periods, SCAN can be seen as a function mapping a set of word uses U of a target word w to two time-specific K-dimensional distributions $\{\phi_1, \phi_2\}$ over word senses and two time-specific V-dimensional distributions over the vocabulary $\{\psi_1^k, \psi_2^k\}$ for each word sense k, where K is a predefined number of senses for target word w. Each of the word sense distributions reflects w's global co-occurrence statistics across uses from the particular time period. SCAN places parametrized logistic normal priors on ϕ_t and ψ_t^k in order to encourage a smooth change of parameters, where the extent of change is controlled through the precision parameter κ^{ϕ}, which is learned during training.

Although ψ_t^k may change over time for word sense k, senses are intended to remain thematically consistent as controlled by word precision parameter κ^{ψ}. This allows comparison of the topic distribution across time periods. For each target word w, we infer a SCAN model for the two time periods and take the probability distributions ϕ_1 and ϕ_2 as the respective semantic representations.

4.3.2 Change Measures

We compute the Jensen-Shannon Distance, as defined in Section 3.1.5, between topic distributions ϕ_1, ϕ_2 to measure graded change. We also experiment with differences in entropy between topic distributions, which is similar to HD, defined in Section 4.2.3.2.

4.4 Thresholding

In Section 4.2.3, we did not define a method to derive binary change scores from type-based representations. As these representations typically do not yield word-sense-like structures, we need to find another way to infer binary change scores if we want to test whether they have the potential to detect binary change. A straightforward way to do this is to choose a threshold on the graded change predictions above which target words receive label 1 and label 0 below. In an unsupervised setting this can e.g. be done by exploiting the distribution of predicted graded change scores in the corpus from which the test data was sampled (Kaiser et al., 2020b). Kaiser et al. infer a global graded change score distribution by predicting graded change for a large number of words. They then choose the threshold to be at mean + one standard deviation of the global distribution. In a supervised setting, on the other hand, the threshold can be tuned on the development data. We will use this approach in Chapter 6. Note that thresholding can not only be applied to *type*-based measures, but also to graded scores derived from *token*-based representations, such as BERT+APD.

The thresholding approach assumes that graded and binary change are correlated, which can in fact be observed on various data sets (e.g. SemEval). However, this approach also has clear limitations. For instance, it does not capture loss or gain of senses with low frequency.

4.5 Discussion

In this chapter, we described various models used to automatically detect LSC from diachronic text corpora. They fall into the three categories: (i) token-based VSMs, (ii) type-based VSMs and (iii) Topic Models. Token-based VSMs provide a model for (negated) semantic proximity, which is the fundamental concept used in Chapter 3 to define LSC. Hence, they can be argued to model senses and thus allow us to apply the exact same change measures used on the annotated data. This provides a complete model of the binary change score which has a foundation in historical linguistics (see Section 2.2). Topic Models model senses directly and thus allow for a similar argument. This is not easily possible for the type-based VSMs, but also not for the average measures defined on token-based VSMs. While we can specify some vague assumptions motivating why they should model graded change, it is unclear to what extent these assumptions hold. For binary change, we need to introduce the even stronger assumption that graded change correlates with binary change. This is done when using thresholding on graded change predictions to derive binary change predictions. While this correlation is empirically observable for a number of data sets (see Section 4.4), it is not a perfect correlation and certainly has its limits (Zamora-Reina et al., 2022). We can easily construct counterexamples with e.g. strong graded change but no binary change or vice versa. Such counterexamples are given by Figure 3.3 and Figure 3.4.

Another major difference between the token- and type-based models described in this chapter is that the former are pretrained on large amounts of data, which we cannot control, while the latter are trained directly on the diachronic target corpora. Hence, they may contain more specific historical information. However, the contextualization of the token-based models (see Section 4.1.1.1) helps to encode historical context information into the resulting use meaning representations.

Our experiments do not include the use of correlation clustering (see Section 3.1.4). However, in order to precisely model the sense derivation process described in Section 3.1.4, this would be an important and straightforward experiment that should be done in the future. The results of Homskiy and Arefyev (2022), reaching a Spearman correlation of 0.65 on a graded change ranking task by the use of correlation clustering on predicted semantic proximity graphs, indicate that this is a promising direction.

Although there are some interesting LSCD models which we do not cover (e.g. Sagi et al., 2009; Rosenfeld and Erk, 2018), the set of models described in this chapter is extensive and covers all model types described in Section 2.4. In Chapter 5, we will now evaluate the performance of the models introduced in this chapter and analyze their predictions.

Chapter 5

Evaluation

In this chapter, we evaluate the models described in Chapter 4 on the data created in Chapter 3. For this, we define several tasks on the change scores developed in Section 3.1.5: (i) Binary Change Classification, (ii) Graded Change Ranking and (iii) Negated COMPARE Ranking. The main evaluation scenario is that models get the full corpora described in Sections 3.2.1.1 and 3.2.2.1 as input data. This is the most realistic scenario as models should consider all the data available in order to measure the semantic change of a word. Otherwise, they could miss e.g. rare senses. Additionally, we evaluate a token-based model with only the sampled uses as input data in Section 5.3. For these samples, the annotated change scores will more accurately reflect sense changes (see Section 3.1.1). We assume that detecting these sample-specific **word sense divergences** is similar to detecting such divergences in larger samples (full corpora), which is necessary to detect LSC.

5.1 Evaluation Metrics

We use standard metrics for evaluation of model predictions: Binary classification is scored with Accuracy (Tharwat, 2020) ranging between 0 (all items incorrectly classified) and 1 (all items correctly classified). Randomly guessing binary classes will yield an expected performance of 0.5. Ranking is scored with Spearman's rank-order correlation coefficient ρ (Spearman, 1904). Spearman's ρ only considers the order of the predictions, the actual predicted change values are not taken into account. Ties are corrected by assigning the average of the ranks normally assigned to each of the tied values to all tied values, (e.g. two words sharing rank 1 both get assigned rank 1.5). Scores are bounded between -1 (completely opposite to true ranking) and 1 (exact match). A value of 0 means that there is no correlation, i.e., the rankings are independent. Clustering performance is measured with the Adjusted Rand Index (ARI, Hubert and Arabie, 1985), which is chance-corrected. The ARI is equal to 1 only if a clustering completely corresponds to the gold clustering and close to 0 for a random clustering.

5.2 Token- and Type-based Models on SemEval Data

The SemEval data set (see Section 3.2.1) was used in a shared task on LSCD.[1] The participating teams were asked to predict the binary (Subtask 1) and graded change (Subtask 2) scores of the target words. They were provided with the lemma versions of the full corpora described in Section 3.2.1.1. The raw corpora and annotated samples were only published after the shared task.[2] Participants were allowed a total of 10 submissions, the best of which was kept for the final ranking. Performance for each subtask was measured as the average performance across languages. Participants had to submit predictions for both subtasks and all languages. A submission's final score for each subtask was computed as the average performance across all four languages. During the evaluation phase, the leaderboard was hidden.

5.2.1 Task Definition

Given the two time-specific corpora C_1 and C_2 described in Section 3.2.1.1, participants were asked to solve two subtasks:

Subtask 1 Binary classification: For a set of target words, predict the binary change score, i.e., decide which words lost or gained sense(s) between C_1 and C_2, and which ones did not.

Subtask 2 Ranking: Rank a set of target words according to their graded change score between C_1 and C_2, i.e., predict the JSD of their word sense probability distributions.

The submitted predictions were evaluated against the hidden change labels via Accuracy and Spearman.

[1] The task also included evaluation on a Latin data set, which was annotated differently and is described in Appendix C.

[2] Find the post-evaluation data at: `https://www.ims.uni-stuttgart.de/data/sem-eval-ulscd-post`.

5.2.2 Models

5.2.2.1 Baselines

For both subtasks, we have two baselines: (i) FD without log-transformation (Freq. Baseline) first calculates the frequency for each target word in each of the two corpora, normalizes it by the total corpus frequency and then calculates the absolute difference between these values as a measure of graded change in Subtask 2 (cf. Section 4.2.3.2). For Subtask 1, we threshold the graded predictions (see Section 4.4) by choosing some threshold. (ii) CNT+CI+CD (Count Baseline) first learns raw co-occurrence vector representations for each of the two corpora, then aligns them by intersecting their columns and measures graded change for Subtask 2 by cosine distance between the two vectors for a target word (see Section 4.2). For Subtask 1, we again binarize these predictions by setting some threshold. A third baseline for Subtask 1, is the majority class prediction (Maj. Baseline), i.e., always predicting the ‘0’ class (no change).

5.2.2.2 Participants

Thirty-three teams participated in the task, totaling 53 members. The teams submitted a total of 186 submissions. Although the models used by participants have a large overlap with the models introduced in Chapter 4, some of them use components which we did not introduce. Descriptions of these components can be found in the respective participant papers, which are referenced in Schlechtweg et al. (2020).

Participating models mainly fall into the categories of token- and type-based VSMs and show rather slight variations in the components described in Chapter 4. Semantic representations are mainly type-based embeddings and token-based embeddings. Token-based embeddings include BERT, which we described in Section 4.1.1, as well as ELMo (Peters et al., 2018) and variations of BERT such as XLM-R (Conneau et al., 2019). Type-based embeddings include SGNS and transformations of count-based vectors, described in Section 4.2.1, as well as GloVe (Pennington et al., 2014) and Gaussian embeddings (Vilnis and McCallum, 2015). (Table 5.1 shows the type of system for every team’s best submission for both subtasks.) Token embeddings are often combined with a clustering algorithm such as AGL, described in Section 4.1.2, as well as k-means, Affinity Propagation, HDBSCAN, GMM (cf. for an overview Aggarwal and Reddy, 2013). One participating team used a graph-based semantic network, one a Topic Model and several teams also propose ensemble models. Alignment techniques for type-based representations include OP, VI and versions of WI/TR, described in Section 4.2.2, as well as Canonical Correlation Analysis (CCA, Knapp, 1978). A variety of change measures are applied: For token-based models, measures include APD, COS, JSD and binary change on clusterings, which we described in Section 4.1.3, as well as the Kullback-Leibler Divergence (Kullback and Leibler, 1951) and Jensen-Shannon Divergence (Lin, 1991). Type-based measures include CD, LND, FD, TD and HD, which we described in Section 4.2.3, as well as the Euclidean distance.

5.2.3 Results

As illustrated in Table 5.1, **UWB** (Pražák et al., 2020b) have the best performance in Subtask 1 for the average over all languages, closely followed by **Life-Language** (Asgari et al., 2020), **Jiaxin & Jinan** (Zhou and Li, 2020) and **RPI-Trust** (Gruppi et al., 2020).[3] For Subtask 2, **UG_Student_Intern** (Pömsl and Lyapin, 2020) perform best, followed by **Jiaxin & Jinan** and **cs2020** (Arefyev and Zhikov, 2020). Across all systems, good performance in Subtask 1 does not indicate good performance in Subtask 2 (correlation between the system ranks is 0.22). However, and with the exception of **Life-Language** and **cs2020**, most top performing systems in Subtask 1 also excel in Subtask 2, albeit with a slight change of ranking.

Remarkably, all the top performing systems use type-based models based on SGNS. They mainly differ in alignment, change measure and threshold selection (see Chapter 4): **UWB** (SGNS+CCA+CD) align SGNS vectors with CCA and measure graded change with CD. They set the average CD across target words as the threshold for binary change. **Life-Language** (SGNS) use an idea very similar to LND to measure graded change: They measure word meaning with a softmax-normalized vector of similarities to (supposedly) stable pivot words in an SGNS vector space and measure graded change as the Kullback-Leibler divergence between two such vectors obtained from time-specific spaces.[4] They do not provide details on how they select the thresholds for their submissions to Subtask 1. **RPI-Trust** (SGNS+OP) calculate CD, and a variation of LND in OP-aligned SGNS spaces and combine these with a variation of FD into an ensemble score by transforming each individual score to a probability of change by comparing it to the full cor-

[3] **RPI-Trust** submits an ensemble model. As all of the features are derived from the type vectors, we classify it as 'type' in this section.

[4] The team uses a variation of SGNS trained on subword information (Bojanowski et al., 2017).

Team	**Subtask 1**					**System**
	Avg.	EN	DE	LA	SV	
UWB	.687	.622	.750	.700	.677	type
Life-Language	.686	.703	.750	.550	.742	type
Jiaxin & Jinan	.665	.649	.729	.700	.581	type
RPI-Trust	.660	.649	.750	.500	.742	type
UG_Student	.639	.568	.729	.550	.710	type
DCC	.637	.649	.667	.525	.710	type
NLP@IDSIA	.637	.622	.625	.625	.677	token
JCT	.636	.649	.688	.500	.710	type
Skurt	.629	.568	.562	.675	.710	token
Discovery	.621	.568	.688	.550	.677	ens.
Count Bas.	.613	.595	.688	.525	.645	-
TUE	.612	.568	.583	.650	.645	token
Entity	.599	.676	.667	.475	.581	type
IMS	.598	.541	.688	.550	.613	type
cs2020	.587	.595	.500	.575	.677	token
UiO-UvA	.587	.541	.646	.450	.710	token
NLPCR	.584	.730	.542	.450	.613	token
Maj. Bas.	.576	.568	.646	.350	.742	-
cbk	.554	.568	.625	.475	.548	token
Random	.554	.486	.479	.475	.774	type
UoB	.526	.568	.479	.575	.484	topic
UCD	.521	.622	.500	.350	.613	graph
RIJP	.511	.541	.500	.550	.452	type
Freq. Bas.	.439	.432	.417	.650	.258	-

Team	**Subtask 2**					**System**
	Avg.	EN	DE	LA	SV	
UG_Student	.527	.422	.725	.412	.547	type
Jiaxin & Jinan	.518	.325	.717	.440	.588	type
cs2020	.503	.375	.702	.399	.536	type
UWB	.481	.367	.697	.254	.604	type
Discovery	.442	.361	.603	.460	.343	ens.
RPI-Trust	.427	.228	.520	.462	.498	type
Skurt	.374	.209	.656	.399	.234	token
IMS	.372	.301	.659	.098	.432	type
UiO-UvA	.370	.136	.695	.370	.278	token
Entity	.352	.250	.499	.303	.357	type
Random	.296	.211	.337	.253	.385	type
NLPCR	.287	.436	.446	.151	.114	token
JCT	.254	.014	.506	.419	.078	type
cbk	.234	.059	.400	.341	.136	token
UCD	.234	.307	.216	.069	.344	graph
Life-Language	.218	.299	.208	-.024	.391	type
NLP@IDSIA	.194	.028	.176	.253	.321	token
Count Bas.	.144	.022	.216	.359	-.022	-
UoB	.100	.105	.220	-.024	.102	topic
RIJP	.087	.157	.099	.065	.028	type
TUE	.087	-.155	.388	.177	-.062	token
DCC	-.083	-.217	.014	.020	-.150	type
Freq. Bas.	-.083	-.217	.014	.020	-.150	-
Maj. Bas.	-	-	-	-	-	-

Table 5.1: Summary of the performance of systems for which a system description paper was submitted as well as their type of semantic representation for that specific submission in Subtask 1 (left) and Subtask 2 (right). For each team, we report the values of Accuracy (Subtask 1) and Spearman correlation (Subtask 2) corresponding to their best submission (highest Avg. per subtask) in the evaluation phase. Avg. = average across languages, EN = English, DE = German, LA = Latin, and SV = Swedish, type = type-based models, token = token-based models, topic = topic model, ens. = ensemble, graph = graph, UCD = University_College_Dublin, UG_Student = UG_Student_Intern, Discovery = Discovery_Team.

pus distribution of that score. Graded change is then measured as the average of these probabilities. The thresholds for binary change are selected manually choosing values above or equal to 0.5. **Jiaxin & Jinan** (SGNS+WI+CD) use WI to learn aligned SGNS spaces and measure graded change with CD. They choose the threshold for binary change by fitting a Gamma distribution to the observed distribution of target word CDs and set the 75% quantile as the threshold. **UG_Student_Intern** (SGNS+OP+ED) measure graded change using Euclidean distance between two OP-aligned SGNS spaces. Finally, **cs2020** (SGNS+OP+CD) measure graded change using CD between two OP-aligned SGNS spaces.

	Subtask 1		
	ACC	Team	System
EN	.730	NLPCR/UWB	token/type
DE	.812	UWB	type
LA	.700	UWB	type
SV	.774	Random	type

	Subtask 2		
	SPR	Team	System
EN	.440	NLPCR	token
DE	.735	Jiaxin & Jinan	type
LA	.513	Discovery	token
SV	.604	UWB	type

Table 5.2: Top per-language performances in SemEval shared task of systems for which a system description paper was submitted as well as their type of semantic representation for that specific submission in Subtask 1 (left) and Subtask 2 (right). ACC = Accuracy, SPR = Spearman.

Table 5.2 shows the top performances per language. Note that some of these are obtained with a token-based model: **NLPCR** (Rother et al., 2020) obtain their high scores for English in both subtasks with a token-based clustering model that closely models the annotation process (cf. Section 2.3.1): They map word uses to token-vectors with a multilingual variant of BERT, cluster these with HDBSCAN (Campello et al., 2013) and measure change with binary and graded change, as described in Section 4.1.3.1. **Discovery** (Martinc et al., 2020b) obtain their top performance for Latin in Subtask 2 with an ensemble model combining predictions from two token-based models: (i) BERT vectors clustered with Affinity Propagation (Frey and Dueck, 2007) and measuring change with the Jensen-Shannon Divergence. (ii) BERT+COS. Their first approach is very similar to **NLPCR**'s approach and another example of modeling the annotation process. **UWB** obtain their high score for English, German and Latin in Subtask 1 with SGNS+CCA+CD, as described above, and for Swedish in Subtask 2 with SGNS+OP+CD. The threshold for Subtask 1 is either chosen as the average CD across target words (German, Latin) or based on the intersection of nearest neighbors of target word vectors in the two spaces (English). **Random** (Cassotti et al., 2020) obtain their result for Swedish in Subtask 1 with PPMI+SRV+CD and a threshold derived by clustering graded predictions into two clusters. **Jiaxin & Jinan** use SGNS+WI+CD for their high score on German in Subtask 2.

An important finding common to most systems is the difference between their performances across the four languages – systems that excel on one dataset do not necessarily perform well in another. This discrepancy may be due to a range of

System	**Subtask 1**		**Subtask 2**	
	Avg.	Max.	Avg.	Max.
type	**.625**	**.687**	.329	**.527**
ensemble	.621	.621	**.442**	.442
token	.598	.637	.258	.374
topic	.526	.526	.100	.100
graph	.521	.521	.234	.234

Table 5.3: Average and maximum performance (average across languages) of best submissions per subtask for different system types. Submissions that corresponded exactly to the baselines or the sample submission were removed.

factors. (i) The differences in corpus size and quality: As can be seen in Table 3.3, the corpora available for training models have a very different sizes between languages. Type-based models are dependent on large amounts of training data, which may also have influenced the results. Further, English and Latin corpora are clean, Swedish and partly German corpora contain OCR errors. (ii) The target word selection: As can be seen in Table 3.4, the test sets are small and have different average levels of change and polysemy. They are likely to differ also in further properties which we did not measure. (iii) As we saw in Section 3.3, the change scores used for evaluation are noisy. This could affect languages to different degrees. (iv) The availability of tuned hyperparameters might have played a role as well: For German, some teams report following prior work such as Schlechtweg et al. (2019a). (v) Some teams focused on some languages, submitting dummy results for the others.

Type- versus token-based models Tables 5.1 and 5.3 illustrate the gap in performance between type-based models and the token-based ones. Out of the best 10 systems in Subtask 1/Subtask 2, 7/7 systems are type-based compared to only 2/2 systems that are token-based. Contrary to the recent success of token-based models (Peters et al., 2018), they are outperformed by type-based models in our task. This is most surprising for Subtask 1 because type-based models do not distinguish between different senses while token-based models do. Nevertheless, there are some token-based models showing comparably high performance on individual languages, see Table 5.2. This means that they have a potential, but are lacking robustness across data sets. We conjecture that this is related to the fact that

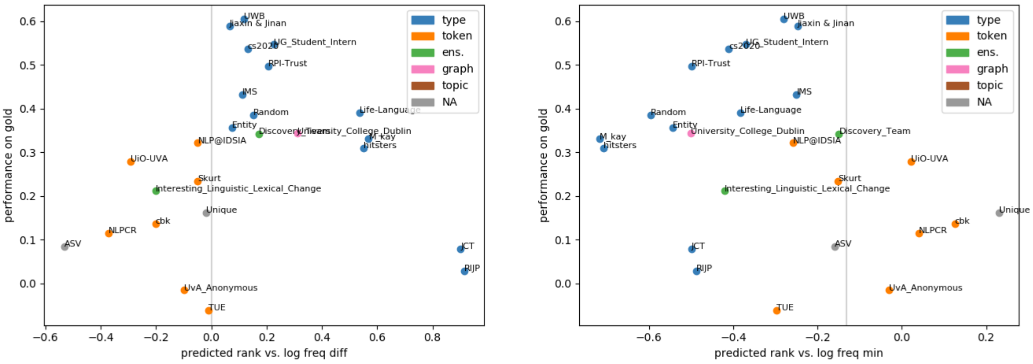


Figure 5.1: Influence of frequency on model predictions in Subtask 2, Swedish. X-axis: correlations of model predictions with FRQ_d (left) and FRQ_m (right), Y-axis: performance on Subtask 2. Gray line gives frequency correlation in gold data (cf. Table 3.4).

contextualized embeddings are a recent technology and as such lack proper usage conventions. For example, it is not clear what impact particular preprocessings (e.g. lemmatization), hyper-parameter choices (e.g. which layers to use) or combinations with clustering methods and change measures can have and whether these are robust across data sets.

We see a range of further points that could influence the performance of token-based models: (i) They are pretrained and cannot exclusively be trained on the relevant historical resources (in contrast to type-based models). As such, they carry additional, and possibly irrelevant, information that may mask true diachronic changes. (ii) They are also sensitive to data preprocessing: Only restricted context is available to the models as a result of the sentence shuffling in the SemEval corpora. Usually, token-based models take more context into account than just the immediate sentence (Martinc et al., 2020b). Also, the corpora were lemmatized while token-based models usually take the raw sentence as input. In order to make the input more suitable for token-based models, we also provided the raw corpora after the evaluation phase and published the annotated uses of the target words with additional context (see Sections 3.2.1.1 and 3.2.1.6). (iii) The corpora contain many historical words for which token-embeddings may have seen little or no training data. The raw corpora additionally contain historical spelling variants which may lead to an incorrect split into subwords in the preprocessing of e.g. BERT. We further investigate some of these questions in Section 5.3.

The influence of frequency In prior work, the predictions of many systems have been shown to be inherently biased towards word frequency, either as a consequence of an increasing sampling error with lower frequency (Dubossarsky et al., 2017) or by directly relying on frequency-related variables (Schlechtweg et al., 2017, 2019a; Bott et al., 2021). We have controlled for frequency when selecting target words (recall Table 3.4) in order to test model performance when frequency is not an indicating factor. Despite the controlled test sets we observe strong frequency biases for the individual models as illustrated for Swedish in Figure 5.1.[5] Model predictions tend to correlate negatively with the minimum frequency of target words between corpora ($\mathrm{FRQ_m}$), and positively with the change in their frequency across corpora ($\mathrm{FRQ_d}$). This means that models predict higher change for low-frequency words and for words with strong changes in frequency. Despite their superior performance, type-based models are more strongly influenced by frequency than token-based models probably because the latter are not trained on the test corpora limiting the influence of frequency. Similar tendencies can be seen for the other languages. For a range of models correlations reach values > 0.8.

The influence of polysemy We did not control the test sets for polysemy. As shown in Table 3.4, the change scores for both subtasks are moderately to highly correlated with polysemy ($\mathrm{PLY_m}$). Hence, it is expected that model predictions would be positively correlated with polysemy. However, correlations are in almost all cases lower than with the change scores and in some cases even negative (Latin and partly English). We conclude that model predictions are only moderately biased towards polysemy on our data.

5.3 Analyzing and Improving a Token-based Model on SemEval Data

In this section, we try to understand better why token-based models showed rather low performance in the SemEval shared task (see Section 5.2). For this, we investigate the influence of a range of variables on clusterings of BERT vectors on the SemEval data and show that it suffers from orthographic information on the target word, which is encoded even in the higher layers of BERT representations. We also

[5] Find the full set of analysis plots at `https://www.ims.uni-stuttgart.de/data/sem-eval-ulscd-post`.

show that lemmatizing the full input use is often not a good solution. By reducing the influence of orthography on the target word while keeping the rest of the input in its natural form, we considerably improve BERT's performance on clustering-based and average methods.[6] In these experiments, we only use the sampled and annotated word uses instead of all uses from the full corpora. These are cleaned by removing uses with many 'Cannot decide' judgments, as described in Section 3.2.1.5. The German uses are further cleaned by replacing a small, manually identified set of historical characters with their modern equivalents.[7] We use the tokenized (Token) and lemmatized (Lemma) versions of the uses.[8]

5.3.1 Models & Measures

For every target word, we feed the uses from the SemEval data set into BERT and use the respective pre-trained cased base model to create token-based models. We then cluster the vectors with AGL and estimate the number of clusters with the Silhouette Method, as explained in Section 4.1.2. We convert the resulting time-specific cluster frequency distributions into probability distribution P and Q and measure their distance $JSD(P, Q)$ to obtain graded change values. We also measure change without clustering using APD and COS, as explained in Section 4.1.3.

5.3.1.1 Cluster Bias

We perform a detailed analysis on what the inferred clusters actually reflect. We test hypotheses on **word form**, **use position**, **number of proper names** and **corpus**. The influence strength of each of these variables on the clusters is measured by the Adjusted Rand Index (ARI) between the inferred cluster labels for each use and a labeling for each use derived from the respective variable. For the variable *word form*, we assign the same label to each use where the target word has the same orthographic form (same string). If ARI = 1, then the inferred clusters contain only uses where the target word has the same form. For *position*, each use receives label 0 if the target word is one of the first three words of the use, label 2 if the target word is one of the last three words, else 1.[9] For *proper names*, a use receives label

[6] Find our code at `https://github.com/Garrafao/TokenChange`.

[7] This is similar to the preprocessing described in Section 6.2 and can be checked at `https://github.com/seinan9/LSCDiscovery`.

[8] Note that we use an early version of the lemmatization where punctuation is replaced by the string '\$'.

[9] We reckon that especially the beginning and ending of a use have a strong influence.

0 if no proper names are in the use, label 1 if one proper name occurs, else 2.[10] The hypothesis that proper names may influence the clustering was suggested by Martinc et al. (2020b). For *corpora*, a use is labeled 0 if it occurs in the first target corpus, else 1.

5.3.2 Results

5.3.2.1 Clustering

Because of the high computational load, we apply the clustering only to the EN and DE parts of the SemEval data set. For this, we use BERT to create token vectors and cluster them with AGL, as described in Section 5.3.1. We then perform a detailed analysis of what the clusters reflect.[11]

We report a subset of the clustering experiment results in Table 5.4, the complete results are provided in Appendix E. Table 5.4 shows JSD performance on Graded Change Ranking (see Section 5.2.1) with Spearman correlation (Graded), clustering performance on the gold clusterings (see Section 3.2.1) measured with ARI (Cluster) as well as the ARI scores for the influence factors introduced above, across BERT layers. For each influence factor, we add two baselines (see Appendix E): (i) The random baseline (Random) measures the ARI score of the influence factor using random cluster labels and (ii) the gold baseline (Gold) measures the ARI score between the gold cluster labels and the influence factor. In other words, (i) and (ii) respectively answer the question of how strong the influence factor is by chance and how strong it is according to the gold annotation. The values of the two baselines are crucial: If an influence factor has an ARI score greater than both baselines, the clustering reflects the influence factor more than expected. Table 5.4 marks influence scores in boldface if they exceed both baseline scores. If additionally the influence score exceeds the actual performance score (Cluster), the clustering reflects the influence factor more than it reflects the gold clustering.

Word form bias As explained above, the word form influence measures how strongly the inferred clusterings represent the orthographic forms of the target word. Table 5.4 shows that for both DE and EN the form bias of the raw (non-

[10] The influence of proper names is only measured for EN since no POS-tagged data was readily available for DE.

[11] We also run most of our experiments with k-means (Forgy, 1965). Both algorithms performed similarly with a slight advantage for AGL. We therefore only report the results achieved using AGL.

	Layer	Token	Lemma	TokLem
Graded	1	-.141	-.033	.100
	12	.205	.154	.168
	9-12	.325	**.345**	.293
Cluster	1	.022	.041	.045
	12	.116	.111	.158
	9-12	.150	.159	**.163**

Form	1	**.907**	.014	.014
	12	**.389**	**.018**	**.077**
	9-12	**.334**	**.018**	**.051**
Position	1	.001	**.026**	**.024**
	12	**.012**	**.012**	**.015**
	9-12	.002	**.007**	**.003**
Corpora	1	**.019**	**.021**	**.033**
	12	**.078**	**.056**	**.082**
	9-12	**.056**	**.044**	**.063**
Names	1	-.007	.010	.010
	12	**.025**	**.027**	**.033**
	9-12	.019	**.022**	**.026**

	Layer	Token	Lemma	TokLem
Graded	1	-.265	-.062	-.170
	12	.123	.427	**.624**
	9-12	.122	.420	.533
Cluster	1	.033	.002	.003
	12	.119	.159	**.161**
	9-12	.155	.142	.154

Form	1	**.706**	.024	.004
	12	**.439**	**.056**	**.150**
	9-12	**.420**	**.047**	**.094**
Position	1	.005	**.023**	**.027**
	12	-.002	.005	-.002
	9-12	**.009**	**.018**	**.012**
Corpora	1	.074	.003	.005
	12	**.110**	**.095**	**.096**
	9-12	**.107**	.068	**.089**
Names	1	-	-	-
	12	-	-	-
	9-12	-	-	-

Table 5.4: Overview of English clustering scores (left) and German clustering scores (right). Bold font indicates best scores for performance on Graded Change Ranking (Graded) and correspondence to gold clusterings (Cluster) (top). For influence variables (bottom), bold font instead indicates scores above all corresponding baselines. See also Appendix E.

preprocessed) token vectors (column 'Token') is extremely high and always yields the highest influence score for each layer combination of BERT. Additionally, the influence of the word form is considerably higher when using lower layers of BERT. This fits well with the observations of Jawahar et al. (2019) that the lower layers of BERT capture surface features, the middle layers capture syntactic features and the higher layers capture semantic features of the input. With the first layer of BERT, the uses are almost exclusively (.9) clustered according to the form of the target word (e.g. plural/singular division). Even in the higher layers word form influence is considerable in both languages (layer 12: $\approx$ .4). This strongly overlays the semantic information encoded in the vectors, as we can see in the low performance scores (Graded, Cluster), which are negatively correlated with word form influence.

The word form bias seems to be lower in DE than in EN (layer 1: .7 vs. .9).

However, this is misleading as our approach to measure word form influence does not capture cases where vectors cluster according to subword forms, as in the case of German *Ackergerät*: Its word forms differ as to whether they are written with an 'h' or not, as in *Ackergerät* vs. *Ackergeräth*. A manual inspection shows that this is strongly reflected in the inferred clustering. However, these forms then further subdivide into inflected forms such as *Ackergeräthe* and *Ackergeräthes*, which is reflected in our influence variable, but not in the inferred clustering. For these cases, our approach tends to underestimate the influence of the variable.[12]

In order to reduce the influence of word form, we experiment with two preprocessing approaches: (i) We feed BERT with lemmatized uses (Lemma) instead of raw ones. (ii) We only replace the target word in every use with its lemma (TokLem). TokLem is motivated by the fact that BERT is pre-trained on raw text. Thus, we assume that BERT is more familiar with non-lemmatized uses and therefore expect it to work better on raw text. In order to continue working with non-lemmatized uses, but at the same time removing word form influence, we only remove the target word form bias by exchanging the target word with its lemma.

As we can see in Table 5.4, lemmatization strongly reduces the influence of word form, as expected.[13] Accordingly, performance (Graded, Cluster) mostly improves. However, it also leads to deterioration in some cases. TokLem also reduces the influence of word form and in most cases yields the overall maximum performance. The Cluster scores for both languages are similar ($\approx .160$) while the Graded performance varies very strongly between languages, achieving a very high score for DE (.624).

Replacing the target word by its lemma form seems to shift the word form influence in the different layers: Especially for DE, layers 1 and 1+12 show the highest influences (.706 and .687) with Token (see also Appendix E). In combination with TokLem, however, both layers are influenced the least (.004 and .046). For EN we see the same effect for layer 1.

Other bias factors We can see in Table 5.4 that most influences are above-baseline. As explained above, the word form bias heavily decreases using higher layers of BERT. For all other influences the bias increases when using higher layers of BERT.

[12] We did not examine whether the bias could be explained exclusively by subword splitting errors because of historical spelling variants. An indication that this is not the case is the high word form bias on the English data where historical spelling variants are infrequent.

[13] In some cases it is, however, above the baselines, indicating that word form is correlated with other use features.

Layer	APD			COS		
	EN	DE	SV	EN	DE	SV
1	.297	.205	.228	.246	.246	.089
12	.566	.359	.529	.339	**.472**	.134
1+12	.455	.316	.280	.365	.373	.077
1-4	.431	.227	.355	**.390**	.297	.079
9-12	**.571**	**.407**	**.554**	.365	.446	**.183**

Table 5.5: Performance of average measures for different layer combinations across languages without preprocessing.

This may be because decreasing the word form influence reveals the existence of further –less strong but still relevant– influences. The same is observable with the Lemma and TokLem results since there the form influence is decreased or even eliminated. While for EN the influence scores mostly increase using Lemma and TokLem, for DE only the position influence increases while corpora influence decreases. This is probably because the corpora influence is to some extent related to word form, which often reflects time-specific orthography, as in *Ackergeräth* vs. *Ackergerät*, where the spelling with the 'h' mostly occurs in the old corpus.

Influence of position and proper names seems to be less important, but the respective scores are still most of the times higher than the baselines. So, overall, the reflection of the two corpora seems to be the most influential factor apart from word form. Often the corpus bias is almost as high as the Cluster performance score.

5.3.2.2 Average Measures

For the average measures, we perform experiments for all three languages (EN/DE/SV).

Layers Because we observe a strong variation of influence scores with layers, as seen in Section (5.3.2.1), we test different layer combinations for the average measures. The following are ones considered: 1, 12, 1+12, 1+2+3+4 (1-4), 9+10+11+12 (9-12). As shown in Table 5.5, the choice of layers strongly affects the performance. We see that for APD the higher layer combinations 12 and 9-12 perform best across all three languages, where the latter is slightly better (.571, .407 and .554). Interestingly, these two are the only layer combinations that do not include layer 1. All

		Layer	Token	Lemma	TokLem
EN	APD	12	**.566**	.483	.494
		1+12	.455	**.483**	.455
		9-12	**.571**	.493	.547
	COS	12	**.339**	.251	.331
		1+12	**.365**	.239	.193
		9-12	**.365**	.286	.353
DE	APD	12	.359	.303	**.456**
		1+12	.316	.643	**.731**
		9-12	.407	.305	**.516**
	COS	12	.472	.693	**.755**
		1+12	.373	.698	**.729**
		9-12	.446	.689	**.726**
SV	APD	12	**.529**	.214	.505
		1+12	.280	.368	**.602**
		9-12	**.554**	.218	.531
	COS	12	.134	-.019	**.285**
		1+12	.077	.012	**.082**
		9-12	.183	-.002	**.284**

Table 5.6: Performance of average measures for three layer combinations with preprocessing variants.

three layer combinations that include layer 1 are considerably worse in comparison. While COS performs best with layer combination 1-4 for EN (.390), for DE and SV we see a similar trend as with APD. Again, the higher layer combinations perform better than the other three, which all include layer 1. For DE, layer combination 12 (.472) performs best while 9-12 yields the highest result for SV (.183). Our results are mostly in line with the findings of Kutuzov and Giulianelli (2020), who observe that APD works best on EN and SV while COS yields the best scores for DE.

Preprocessing As with the clustering, we try to improve the performance of the average measures by using the two above-described preprocessing approaches. We perform experiments only for three layer combinations in order to reduce the complexity: (i) 12 and (ii) 9-12 perform best and are therefore obvious choices. (iii) From the remaining combinations 1+12 shows the most stable performance across measures and languages. Table 5.6 shows the performance of the preprocessings

		Layer	Token	TokLem
EN	APD	1+12	.613	-.026
		9-12	.068	.090
	COS	1+12	.246	-.062
		9-12	.020	.004
DE	APD	1+12	.554	.271
		9-12	.292	.105
	COS	1+12	.387	-.017
		9-12	.205	-.008
SV	APD	1+12	.730	.176
		9-12	.237	.048
	COS	1+12	.429	-.031
		9-12	.277	-.035

Table 5.7: Correlations of word form and predicted change scores.

(Lemma, TokLem) over these three combinations. We can see that both APD and COS perform slightly worse for EN when paired with a preprocessing (exception to this is 1+12 Lemma). In contrast, DE profits heavily: While APD with layer combinations 12 and 9-12 performs slightly worse with Lemma and slightly better with TokLem, we observe an enormous performance boost for layer combination 1+12 (.643 Lemma and .731 TokLem). We achieve a similar boost for all three layer combinations with COS as a measure. We reach a top performance of .755 for layer 12 with TokLem. SV does not benefit from Lemma. We observe large performance decreases, with the exception of combination 1+12 (APD). The APD performance of layers 12 and 9-12 is slightly worse with TokLem. However, layers 1+12, which performed poorly without preprocessing, reaches peak performance of .602 with TokLem. All COS performances increase with TokLem, but are still well below the APD counterparts. The general picture is that DE and SV profit strongly from TokLem. Lemma yields the best performance for only one case. The top performances we reach for EN/DE/SV are .566/.755/.602, obtained with Token/TokLem/TokLem respectively. We see that the performance of preprocessing strongly depends on preprocessing type, language, layer combination and measure. This confirms our observation from Section 5.2 that it is hard to choose robust parameter configurations for BERT.

Word form bias In order to better understand the effects of layer combinations and preprocessing, we compute correlations between word form and model predictions. To lessen the complexity, only layer combination 1+12 (which performed worst overall and includes layer 1), layer combination 9-12 (which performed best overall) in combination with Token and the superior TokLem are considered. The results are presented in Table 5.7. We observe similar findings for all three languages: The correlation between word form and APD predictions is strong (.613, .554 and .730) for layers 1+12 without preprocessing. The correlation is much weaker with layers 9-12 (.068, .292 and .237) or TokLem ($-$.026, .105 and .176). This is in line with the performance development that also increases using layers 9-12 or TokLem, as can be seen in Table 5.6. Both approaches (different layers, preprocessing) result in a considerable performance increase, as described previously. Using layer combination 9-12 with TokLem further decreases the correlation (with the exception of EN). However, the performance is better when only one of these approaches is used. The correlation between word form and COS model predictions is weaker overall (.246, .387 and .429). We see a similar correlation development as for APD, however this time the performance of EN does not profit from the lowered bias (see Table 5.6). Both DE and SV see a performance increase when the word form bias is lowered by either using layers 9-12 or TokLem.

Polysemy bias The SemEval data sets are strongly biased by polysemy, i.e., a perfect model measuring the gold synchronic target word polysemy in either t_1 or t_2 could reach above .7 performance (see Table 3.4). We use APD-OLD and APD-NEW (see Section 4.1.3) to see whether we can exploit this fact to create a purely synchronic polysemy model with high performance. We achieve low to moderate performances for EN and DE (.274/.332 and .321/.450 respectively) and a good performance for SV (.550/.562). While the performance for EN and DE is clearly below the high-scores, the performance is high for a measure that lacks any kind of diachronic information. And in the case of SV, the performance of both APD-OLD and APD-NEW is just barely below the high-scores (cf. Table 5.6). Note that regular APD (in contrast to COS) is, by definition, affected by polysemy (Schlechtweg et al., 2018). It is thus possible that APD's high performance stems at least partly from this polysemy bias. This is supported by comparing the SV results of APD and COS in Table 5.6: COS is weakly influenced by polysemy and performs poorly while APD has higher performance, but only slightly above the purely synchronic measures APD-OLD/NEW.

5.4 Type-based Models on DURel/SURel Data

In this section, we evaluate the type-based models (see Section 4.2) on the DURel/SURel data (see Section 3.2.2). Note that we exclude three targets from the DURel dataset (*flott*, *Kinderstube*, *Steckenpferd*) and one target from the SURel dataset (*Messerspitze*) because they fall below the frequency threshold used for cleaning corpora in Section 3.2.2.1. Models are trained on the full corpora from which the annotated use pairs were sampled.

5.4.1 Task Definition

Given the two time-specific corpora C_1 and C_2 described in Section 3.2.2.1, we aim to solve the following task:

Task Ranking: Rank a set of target words according to their Negated COMPARE score between C_1 and C_2.

Model predictions are scored against the Negated COMPARE scores with Spearman's rank-order correlation coefficient ρ.

5.4.2 Preprocessing & Hyperparameters

We first motivate our settings for corpus preprocessing described in Section 3.2.2.1 and for model hyperparameters described in Section 4.2.

Preprocessing We experiment with the two preprocessed corpus versions $\mathrm{L}_{\mathrm{ALL}}$ and L/P. Both versions include different combinations of common preprocessing steps for VSMs, i.e., lemmatization, removal of low-frequency words/non-content words/punctuation and concatenation with POS tag. Lemmatization is important because LSC is measured on the lemma level (see Chapter 3), but also because German has many inflected word forms which would distribute the vector representation for a lemma across many vectors. Words with very low frequency are often regarded noise in VSMs (Schulte im Walde et al., 2013; Bott and Schulte im Walde, 2014; Levy et al., 2015) while words with very high frequency, non-content words or punctuation can be treated in different ways (Schulte im Walde et al., 2013; Mikolov et al., 2013a; Bott and Schulte im Walde, 2014; Levy et al., 2015). Additional syntactic information has shown to improve performance on some tasks (Padó and Lapata, 2007; Shwartz et al., 2017).

Representation	Alignment					Measure					
	CI	SRV	OP	VI	WI	CD	LND	JSD	FD	TD	HD
CNT	x				x	x	x			x	x
PPMI	x				x	x	x				
SVD			x		x	x	x				
RI		x	x		x	x	x				
SGNS			x	x	x	x	x				
SCAN								x			(x)

Table 5.8: Combinations of semantic representations, alignment types and measures on DURel/SURel. (FD is computed directly from the corpus.)

Context window For all models we experiment with values $n = \{2, 5, 10\}$, as done in Levy et al. (2015). It is important to note that the extraction of context words differs between models because of inherent parameter settings of the implementations. While our implementations of the count-based vectors have a stable window of size n, the SGNS implementation we use (Řehůřek and Sojka, 2010) has a dynamic context window with maximal size n (cf. Levy et al., 2015), and SCAN has a stable window of size n, but ignores all occurrences of a target word where the number of context words on either side is smaller than n. This may affect the comparability of the different models as especially the mechanism of SCAN can lead to very sparse representations on corpora with small sentence sizes, as e.g. the COOK corpus. Hence, this variable should be controlled in future experiments.

VSMs We follow previous work in setting further hyperparameters (Hamilton et al., 2016b; Levy et al., 2015). We set the number of dimensions d for SVD, RI and SGNS to 300. We train all SGNS with 5 epochs. For PPMI, we set $\alpha = .75$ and experiment with $k = \{1, 5\}$ for PPMI and SGNS. For RI and SGNS, we experiment with $t = \{none, .001\}$. For SVD we set $p = 0$. In line with Basile et al. (2015), we set $s = 2$ for RI and SRV. Note though that we have a lower d than Basile et al. who set $d = 500$.

SCAN We experiment with $K = \{4, 8\}$. For further parameters, we follow the settings chosen by Frermann and Lapata (2016): $\kappa^{\psi} = 10$ (a high value forcing senses to remain thematically consistent across time). We set $\kappa^{\phi} = 4$ and the Gamma parameters $a = 7$ and $b = 3$. We use $1,000$ iterations for the Gibbs sampler and set

Dataset	Preproc	Win	Space	Parameters	Align	Measure	Performance m (h, l)
DURel	$\mathrm{L}_{\mathrm{ALL}}$	10	SGNS	k=1,t=None	OP	CD	**.866** (.914, .816)
	$\mathrm{L}_{\mathrm{ALL}}$	10	SGNS	k=5,t=None	OP	CD	.857 (.891, .830)
	$\mathrm{L}_{\mathrm{ALL}}$	5	SGNS	k=5,t=.001	OP	CD	.835 (.872, .814)
	$\mathrm{L}_{\mathrm{ALL}}$	10	SGNS	k=5,t=.001	OP	CD	.826 (.863, .768)
	L/P	2	SGNS	k=5,t=None	OP	CD	.825 (.826, .818)
SURel	L/P	2	SGNS	k=1,t=.001	OP	CD	**.851** (.851, .851)
	L/P	2	SGNS	k=5,t=None	OP	CD	.850 (.850, .850)
	L/P	2	SGNS	k=5,t=.001	OP	CD	.834 (.838, .828)
	L/P	2	SGNS	k=5,t=.001	OP_{-}	CD	.831 (.836, .817)
	L/P	2	SGNS	k=5,t=.001	OP	CD	.829 (.832, .823)

Table 5.9: Top performances (Spearman). Win=Window Size, Preproc=Preprocessing, Align=Alignment, k=negative sampling, t=subsampling, Performance m(h,l): mean, highest and lowest Spearman correlation with gold Negated COMPARE rank.

the minimum amount of contexts for a target word per time period $min = 0$ and the maximum amount to $max = 2000$.

Measures For LND, we set $k = 25$ as recommended by Hamilton et al. (2016a). The normalization constants for FD, HD and TD are calculated on the full corpus with the respective preprocessing (but before deleting words below a frequency threshold).

Find an overview of all tested combinations of semantic representations, alignments and measures in Table 5.8.

5.4.3 Results

First of all, we observe that nearly all model predictions have a strong positive Spearman correlation with the gold Negated COMPARE rank. Table 5.9 presents the overall best results across models and parameters.[14] With .87 for DURel and .85 for SURel, the models reach comparable and unexpectedly high performances on the two distinct datasets. The overall best-performing model is SGNS+OP+CD. The model is robust in that it performs best on both datasets and produces very similar, sometimes the same results, across different iterations.

[14] For models with randomness involved, we compute the average over five iterations.

Dataset	Representation	best	mean
DURel	CNT	.639	.395
	PPMI	.670	.489
	SVD	.728	.498
	RI	.601	.374
	SGNS	**.866**	**.502**
	SCAN	.327	.156
SURel	CNT	.599	.120
	PPMI	.791	.500
	SVD	.639	.300
	RI	.622	.299
	SGNS	**.851**	**.520**
	SCAN	.082	-.244

Table 5.10: Best and mean performances (Spearman) across similarity measures (CD, LND) on semantic representations.

Preprocessing and parameters Regarding preprocessing, the results vary: $\mathrm{L}_{\mathrm{ALL}}$ (all lemmas) dominates in the diachronic task while L/P (*lemma:pos* of content words) dominates in the synchronic task. In addition, L/P preprocessing, which is already limited to content words, prefers shorter windows while $\mathrm{L}_{\mathrm{ALL}}$ (preprocessing where the complete sentence structure is maintained) prefers longer windows. Regarding the preference of L/P for SURel, we blame noise in the COOK corpus, which contains a lot of recipes listing ingredients and quantities with numerals and abbreviations, to presumably contribute little information about context words. For instance, COOK contains 4.6% numerals while DTA only contains 1.2% numerals.

Looking at the influence of subsampling, we find that it does not improve the mean performance for SGNS (with .506, without .517), but clearly for RI (with .413, without .285). Levy et al. (2015) found that SGNS prefers numerous negative samples ($k > 1$), which is confirmed here: mean performance is .487 with $k = 1$ and .535 with $k = 5$.[15] This finding is also indicated in Table 5.9, where $k = 5$ dominates the 5 best results on both datasets. Yet, $k = 1$ provides the overall best result on both datasets.

[15] For PPMI, we observe the opposite preference: mean performance is .549 with $k = 1$ and .439 with $k = 5$.

Dataset	OP	OP_-	OP_+	WI	None
DURel	.618	.557	**.621**	.468	.254
SURel	**.590**	.514	.401	.492	.285

Table 5.11: Mean performances (Spearman) for CD per alignment method. Applies only to RI, SVD and SGNS.

Semantic representations Table 5.10 shows the best and mean results for different semantic representations. SGNS is clearly the best VSM even though its mean performance does not exceed other representations as clearly as its best performance. Regarding count-based models, PPMI and SVD show the best results.

SCAN performs poorly, and its mean results indicate that it is rather unstable. This may be explained by the particular way in which SCAN constructs context windows (see Section 5.4.2): It ignores asymmetric windows, thus reducing the number of training instances considerably, in particular for large window sizes.

Alignments The fact that our modification of Hamilton et al. (2016b) (SGNS+OP, see Section 4.2.2) shows best performance across data sets in Table 5.9 confirms our assumption that column-mean centering is an important preprocessing step in OP analysis and should not be omitted. Additionally, the mean performance in Table 5.11 shows that OP is generally more robust than its variants. OP_+ has the best mean performance on DURel, but performs poorly on SURel. Artetxe et al. (2018a) show that the additional pre- and post-processing steps of OP_+ can be harmful in certain conditions. We tested the influence of the different steps and identified the non-orthogonal whitening transformation as the main reason for a performance drop of ≈20%.

In order to see how important the alignment step is for the low-dimensional embeddings (SVD/RI/SGNS), we also tested the performance without alignment ('None' in Table 5.11). As expected, the mean performance drops considerably. However, it remains positive, which suggests that the spaces learned by the models are not random but rather slightly rotated variants.

Especially interesting is the comparison of WI, where one common vector space is learned, against the OP models, where two separately learned vector spaces are aligned. Although WI avoids (post-hoc) alignment altogether, it is consistently outperformed by OP, which is shown in Table 5.11 for low-dimensional embeddings.[16] We found that OP profits from mean-centering in the preprocessing step: applying mean-centering to WI matrices improves the performance by 3% on WI+SGNS+CD. Further studies have varying results regarding the comparison of OP and WI, showing that their relative performance is also parameter dependent (Dubossarsky et al., 2019; Kaiser et al., 2020a).

The results for Vector Initialization (VI) are unexpectedly low (on DURel mean $-.017$, on SURel mean .082). We later noticed that there was an implementation error in the code and that SGNS+VI can reach high performance, but is difficult to tune and strongly influenced by word frequency (Kaiser et al., 2020a, 2021).

Detection measures CD dominates LND on all vector space and alignment types (e.g., mean on DURel with SGNS+OP is .723 for CD vs. .620 for LND) and hence should be generally preferred if alignment is possible. Otherwise LND or a variant of WI+CD should be used, as they show lower but robust results.[17] In a scenario with development data, LND's k parameter should be tuned first.

Dispersion measures generally exhibit a low performance and previous positive results for them could not be reproduced (Schlechtweg et al., 2017). It is striking that, contrary to our expectation, dispersion measures on SURel show a strong negative correlation (max. $-.79$). We suggest that this is due to frequency particularities of the dataset: SURel's gold LSC rank has a rather strong negative correlation with the targets' frequency rank in the COOK corpus described in Section 3.2.2.1 ($-.51$). Moreover, because COOK is magnitudes smaller than SdeWaC, the normalized values computed in most dispersion measures in COOK are much higher. This also gives them a much higher weight in the final calculation of the absolute differences. Hence, the negative correlation in COOK propagates to the final results. This is supported by the fact that the only measure not normalized by corpus size (HD) has a positive correlation. As these findings show, the dispersion measures are strongly influenced by frequency and very sensitive to different corpus sizes.

[16] We see the same tendency for WI against SRV, but instead variable results for CNT and PPMI alignment (CI).

[17] JSD was not included here, as it was only applied to SCAN and its performance thus strongly depends on the underlying meaning representation.

5.5 Discussion

In this chapter, we evaluated the range of models described in Chapter 4 on three tasks: (i) Binary Change Classification, (ii) Graded Change Ranking and (iii) Negated COMPARE Ranking. The task setup (unsupervised, no genuine development data, different corpora from different languages with very different sizes, varying class distributions) provided an opportunity to test models in heterogeneous learning scenarios, which was very challenging.

The top average performance across languages for Binary Change Classification on the SemEval data was an Accuracy of .69, where the top performance per language was .73/.81/.77 for EN/DE/SV (see Table 5.1). For Graded Change Ranking, the top average was a Spearman correlation of .53 and per language .44/.74/.60. In our experiments with BERT on Graded Change Ranking, we reached top performances for EN/DE/SV of .57/.76/.60 (see Table 5.6), i.e., slightly higher than in the SemEval shared task. However, these were obtained under different conditions, i.e., by tuning on the test data and with the word uses sampled for annotation, which were additionally cleaned. Hence, this only gives a possible upper bound for performance in a realistic scenario. For Negated COMPARE Ranking on DURel and SURel, we reach a mean performance of .50 and .52, and a top performance of .87 and .85 (see Table 5.10). Here again, the top performances can only give an upper bound. Unfortunately, not all models were applied to all tasks for reasons of how this thesis progressed, which puts some limits on the comparisons we can make.

Most type-based models were tested on all three tasks and all data sets. The SemEval results suggest that SGNS in combination with CCA+CD, OP+CD, OP+ED, WI+CD or LND and optional thresholding is a good approach for Binary Change Classification and Graded Change Ranking with high performances on several data sets and robustness across data sets. The results for Negated COMPARE Ranking on the DURel/SURel data showed a similar picture with SGNS+OP+CD dominating on both data sets. We further improved the performance of this approach with the application of mean-centering as a preprocessing step for OP alignment.

The token-based models were clearly outperformed by the type-based approaches on Binary Change Classification and Graded Change Ranking on the SemEval data when considering the mean performance across data sets. This was surprising because type-based approaches do not model the annotation process as described in Sections 2.3.1 and 4.1.3.1. However, some token-based approaches managed to reach the top performance on individual languages, suggesting that they rather lack robustness than ability to measure LSC. Strikingly, these approaches modeled the annotation process, showing that it is possible to closely model the measurement process of LSC with computers.

We then wanted to understand what could have affected the performance of the token-based models on the SemEval data. We tested several hypotheses on potential biases in BERT and found that it is influenced by various factors, but most strongly by target word form if word uses are not preprocessed. Even in higher layers this influence persisted. By removing the form bias, we were able to considerably improve the performance for Graded Change Ranking across languages on the SemEval data. We also found that using the lemmatized word uses (as provided to SemEval shared task participants in the evaluation phase) often had a negative impact on performance compared to other preprocessing variants. The final BERT performances achieved in this way were slightly higher than the top performances per language from the SemEval shared task. Note, however, that they varied strongly (see Table 5.6), confirming the lack of robustness observed before. Thus, it is hard to choose the optimal setting in an unsupervised scenario without development data such as the SemEval shared task.

Although we finally reached a comparably high performance with BERT and clustering for Graded Change Ranking in German, average measures still performed better than the clustering-based approaches. The reasons for this are still unclear and should be addressed in future research. The correspondence of inferred clusterings to the gold standard was very low, indicating that the clustering does not infer word senses but other patterns. Note, however, that the gold clusterings derived on the SemEval data can sometimes be noisy (see Section 3.3.2), which could also explain this rather low correspondence.[18]

We found that polysemy is a strong predictor of the SemEval change scores. It was possible to build a simple, purely synchronic polysemy model with a performance on Graded Change Ranking which was near the best diachronic model for one language. However, model predictions in the SemEval shared task did not show polysemy bias above expectation.

Although we compared a large number of models on several multilingual data sets and tasks, the findings were obtained on rather small data sets affecting their generalizability (cf. Arefyev and Zhikov, 2020). They are small because the annota-

[18] Extended and less noisy data is available at `https://www.ims.uni-stuttgart.de/data/wugs`.

tion of one LSC label requires the comparison of a large number of individual word uses. This is a fundamental problem in LSCD, which we will continue to struggle with in the future (Hengchen et al., 2021). Also, the gold labels we used for evaluation were obtained with a new and imperfect annotation procedure (see Section 3.3). Hence, we need to continue to replicate the results we have obtained here on further data sets and tasks: The good performance of SGNS+OP+CD has been demonstrated in previous studies (Hamilton et al., 2016b) and is supported by several follow-up studies (Kaiser et al., 2020a, 2021) and a shared task on Italian data (Basile et al., 2020; Kaiser et al., 2020b; Pražák et al., 2020a). Shoemark et al. (2019) further show good performance of OP+CD combined with another (but related) predictive VSM approach. Laicher et al. (2020) confirm that (off-the-shelf) BERT generalizes poorly and does not transfer well between data sets. The potential of token-based models, on the other hand, has been shown in two subsequent shared tasks on Russian and Spanish (Kutuzov and Pivovarova, 2021a; Zamora-Reina et al., 2022), where especially Word-in-Context models (Pilehvar and Camacho-Collados, 2019) using BERT (or XLMR) vectors as input features show high performance, dominating type-based models (Arefyev and Bykov, 2021; Arefyev and Rachinskiy, 2021; Arefyev et al., 2021; Homskiy and Arefyev, 2022). Also, clustering approaches are amongst the best systems in the Spanish task (Kashleva et al., 2022; Teodorescu et al., 2022). These developments are crucial because they will allow modeling the annotation process more closely in the future. They also enable working with smaller historic data samples because BERT is pre-trained unlike SGNS, which needs large corpora to learn vectors.

Chapter 6

Application

In the previous chapter, we evaluated various token- and type-based VSMs on different LSCD data sets and tasks. The best-performing models included SGNS+OP+CD (or slight variations) and BERT + average measures (APD, COS). We now use these models to **discover novel instances of semantic change** from the SemEval DE corpora and to evaluate the usefulness of such discovered sense changes for external fields. Such discoveries may be useful in a range of fields (Hengchen et al., 2019; Jatowt et al., 2021), among which historical semantics and lexicography represent obvious choices (Ljubešić, 2020). We validate the model predictions post-hoc with the annotation procedure developed in Section 3.1.2. In this way, we automatically detect previously described semantic changes and discover novel instances of semantic change, which had not been indexed in standard historical dictionaries before. We further evaluate the usability of the approach from a lexicographer's viewpoint and show how intuitive visualizations of human-annotated data can benefit dictionary makers.

6.1 Task Definition

The tasks defined in Chapter 5 require to detect semantic change in a small preselected set of target words. Instead, we are interested in the discovery of changing words from the full vocabulary of the corpus. We define the task of **binary lexical semantic change discovery** as follows.

Task Binary classification: Given a diachronic corpus pair C_1 and C_2, decide for the intersection of their vocabularies which words lost or gained sense(s) between C_1 and C_2, and which ones did not.

This task can also be seen as a special case of Binary Change Classification (see Section 5.2), where the target words equal the intersection of the corpus vocabularies. Note, however, that discovery introduces additional difficulties for models e.g. because a large number of predictions is required and the target words are not preselected, nor balanced or cleaned. Yet, discovery is an important task with applications such as lexicography, where dictionary makers aim to cover the full vocabulary of a language.

6.2 Models

We train SGNS on the lemmatized versions of the SemEval DE corpora. For BERT, we extract uses for every target word by randomly subsampling up to 100 uses from both subcorpora C_1 and C_2.[1] These are then fed into BERT to create contextualized embeddings resulting in two sets of up to 100 vectors M_1 and M_2, for each time period respectively (see Section 4.1.3). We experiment with different preprocessings for BERT's input uses, sampling uses from the relevant preprocessed corpus versions. As in Section 5.3, the tokenized use versions are further cleaned by replacing a small, manually identified set of historical characters with their modern equivalents.

We start the discovery process by generating optimized graded change predictions with SGNS+OP+CD, BERT+APD and BERT+COS on the SemEval DE Subtask 2 data using high-performing parameter configurations following previous work and tuning. Then, we infer binary scores with a principled thresholding technique to obtain binary change predictions. We tune the threshold to find the best-performing type- and token-based approaches for binary classification. These are used to generate two sets of predictions for discovery. We evaluate the graded rankings in Subtask 2 with Spearman's rank-order correlation coefficient ρ, as in Chapter 5. For binary classification, we compute precision, recall and $F_{0.5}$. The latter puts a stronger focus on quality (precision) than quantity (recall) yielding smaller sets of positive predictions, which can be more easily evaluated via human annotation.[2]

[1] We subsample as some words have 10,000 or more uses.

[2] Find the code used for each step of the prediction process at `https://github.com/seinan9/LSCDiscovery`.

6.2.1 Tuning

Graded change SGNS is commonly used (see Section 5.2) and also highly optimized (Kaiser et al., 2020a,b, 2021), so it is difficult to further increase its performance. We thus rely on the work of Kaiser et al. (2020a) and test their parameter configurations on the SemEval DE data set.[3] We obtain three slightly different parameter configurations (see Table 6.2 for more details), yielding competitive $\rho = .690$, $\rho = .710$ and $\rho = .710$, respectively (cf. Table 5.1).

In order to improve the performance of BERT, we test different preprocessings, layer combinations and semantic change measures. With the help of our results from Section 5.3, we are able to drastically increase the performance of BERT on SemEval DE. In a preprocessing step, we replace the target word in every use by its lemma (TokLem). In combination with layer 1+12, both APD and COS perform competitively well on Subtask 2 ($\rho = .690$ and $\rho = .738$).

Binary change Solving Subtask 2 (Graded Change Ranking) is straightforward since both the type-based and token-based approach output distances between representations for C_1 and C_2 for every target word. Like many approaches presented in Section 5.2, we use thresholding to binarize these values. The idea is to define a threshold parameter where all ranked words with a distance greater or equal to this threshold are labeled as changing words (see Section 4.4). For cases where no tuning data is available, Kaiser et al. (2020b) propose to choose the threshold according to the population of CDs of all words in the corpus. Kaiser et al. set the threshold to $\mu + \sigma$, where μ is the mean and σ is the standard deviation of the population. We slightly modify this approach by changing the threshold to $\mu + t * \sigma$. In this way, we introduce an additional parameter t, which we tune on the SemEval DE Subtask 2 test data. We test different values ranging from -2 to 2 in steps of 0.1 obtaining $F_{0.5}$-scores for a large range of thresholds. SGNS achieves peak $F_{0.5}$-scores of .692, .738 and .685, respectively (see Table 6.2). Interestingly, the optimal threshold is at $t = 1.0$ in all three cases. This corresponds to the threshold used in Kaiser et al. (2020b). While the peak $F_{0.5}$ of BERT+APD is marginally worse (.598 at $t = -0.2$), BERT+COS is able to outperform the best SGNS configuration with a peak of .741 at $t = 0.1$.

In order to obtain an estimate on the sampling variability that is caused by sampling only up to 100 uses per word for BERT+APD and BERT+COS (see Section

[3] All configurations use $w = 10$, $d = 300$, $e = 5$ and a minimum frequency count of 39.

4.1.3), we repeat the whole procedure 9 times and estimate mean and standard deviation of performance on the tuning data. In the beginning of every run, the uses are randomly sampled from the corpora. We observe a mean ρ of .657 for BERT+APD and .743 for BERT+COS with a standard deviation of .015 and .012, respectively, as well as a mean $F_{0.5}$ of .576 for BERT+APD and .684 for BERT+COS with a standard deviation of .013 and .038, respectively. This shows that the variability caused by subsampling word uses is negligible.

6.2.2 Discovery

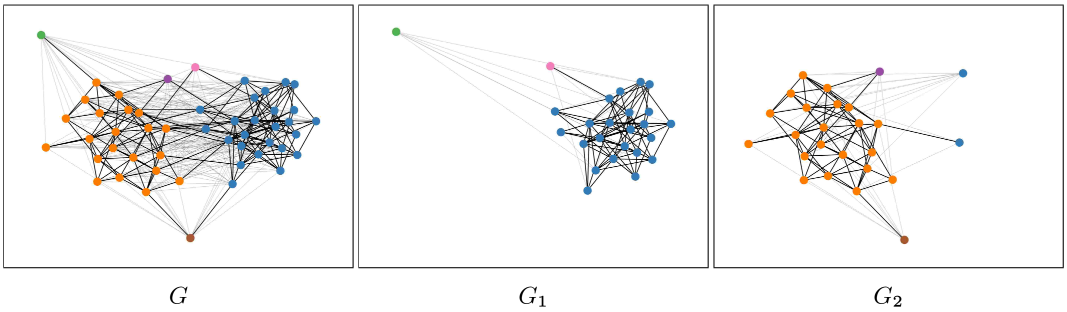
Since SGNS generates type-based vectors for every word in the vocabulary, measuring the distances for the full vocabulary comes with low additional computational effort. Unfortunately, this is much more difficult for BERT: Creating up to 100 vectors for every word in the vocabulary drastically increases the computational burden. Hence, we choose a population of only 500 words for our work, allowing us to test multiple parameter configurations.[4] We sample words from different frequency areas to have predictions not only for low-frequency words. For this, we first compute the frequency range (highest frequency – lowest frequency) of the vocabulary. This range is then split into 5 areas of equal frequency width. Random samples from these areas are taken based on how many words they contain. For example: if the lowest frequency area contains 50% of all words from the vocabulary, then $0.5 * 500 = 250$ random samples are taken from this area. The SemEval DE target words are excluded from this sampling process. For the resulting population of words, we create graded and binary predictions.

We use the top-performing configurations (see Table 6.2) to generate two sets of large-scale predictions for SGNS and BERT by computing graded predictions and thresholding them using the optimal threshold found in Section 6.2.1. While for SGNS we use the matrix obtained for tuning from lemmatized corpora, for BERT we use the raw corpora with lemmatized target words instead. The latter choice is motivated by the previously described performance increases during tuning.

The binary predictions (words labeled as changing) contain proper names, foreign language and lemmatization errors, which we aim to filter out automatically, as such cases are usually not considered to be semantic changes.[5] We only allow

[4] In a practical setting where predictions have to be generated only once, a much larger number may be chosen. Also, possibilities to scale up BERT performance can be applied (Montariol et al., 2021).

[5] We use spaCy for filtering (Honnibal et al., 2020).



G G_1 G_2

Figure 6.1: WUG of German *Aufkommen* (left), subgraphs for 1st time period G_1 (middle) and for 2nd time period G_2 (right).

nouns, verbs and adjectives to pass. Words where over 10% of the uses are either non-German or contain more than 25% punctuation are filtered out as well.

After the filtering, we obtain 27 and 75 words labeled as changing, respectively. We further sample 30 targets from the second set (75) of predictions to obtain a feasible number for annotation. We call the first set SGNS targets and the second one BERT targets, with an overlap of 7 targets. Additionally, we randomly sample 30 words from the population (with an overlap of 5 with the SGNS and BERT targets) in order to estimate the change distribution underlying the corpora. We call these baseline (BL) targets. This baseline will help us to put the discovery result into context and will reveal whether the predictions of the two models improve upon a random word selection procedure from the corpus. Following the annotation process, binary gold data is generated for all three target sets, in order to validate the quality of the predictions.

6.3 Annotation

The model predictions are evaluated by human annotation. For this, we apply the annotation procedure developed in Section 3.1.2 to the uses of the discovered target words from Section 6.2.2. Annotators are asked to judge the semantic relatedness of use pairs, such as the two uses of *Aufkommen* in (6.1) and (6.2), on the scale in Table 3.1.

(6.1) Es ist richtig, dass mit dem **Aufkommen** der Manufaktur im Unterschied zum Handwerk sich Spuren der Kinderexploitation zeigen.

Data set	n	N/V/A	SPR	KRI	|U|	LSC$_B$	LSC$_G$
SemEval DE	48	32/14/2	.59	.53	175	.35	.31
Predictions	75	39/16/20	.64	.58	49	.48	.40

Table 6.1: Overview target words. n = no. of target words, N/V/A = no. of nouns/verbs/adjectives, SPR = weighted mean of pairwise Spearman, KRI = Krippendorff's α, $|U|$ = avg. no. of uses per word, LSC$_{B/G}$ = mean binary/graded change score.

'It is true that with the ***emergence*** *of the manufactory, in contrast to the handicraft, traces of child labor are showing.'*

(6.2) Sie wissen, daß wir für das Vieh mehr Futter aus eigenem **Aufkommen** brauchen.
'They know that we need more feed from our own ***production*** *for the cattle.'*

As described in Section 3.1.3, the annotated data of a word is then represented in a WUG, where vertices represent word uses, and weights on edges represent the (median) semantic relatedness judgment of a use pair. The final WUGs are clustered with Correlation Clustering (see Figure 6.1, left) and split into two subgraphs representing nodes from subcorpora C_1 and C_2 respectively (middle and right). As described in Section 3.1.5, clusters are then interpreted as word senses and changes in clusters over time as LSC.

We use the openly available DURel interface for annotation and visualization.[6] This also implies a change in sampling procedure as the system only implements random sampling of use pairs (without SemEval-style optimization in rounds). For each target word, we sample $|U_1| = |U_2| = 25$ uses per subcorpus (C_1, C_2) and upload these to the DURel system, which presents use pairs to annotators in randomized order. We recruit eight German native speakers with university level education as annotators. Five have a background in linguistics, two in German studies, and one has an additional professional background in lexicography. Similar to the criterion used in Section 3.2.1.5, we ensure the robustness of the obtained clusterings by continuing the annotation of a target word until all multi-clusters (clusters with more than one use) in its WUG are connected by at least one judgment. We finally label a target word as changed (binary) if it gained or lost a cluster over time

[6] https://www.ims.uni-stuttgart.de/data/durel-tool.

	parameters	t	tuning				predictions			
			ρ	$F_{0.5}$	P	R	ρ	$F_{0.5}$	P	R
SGNS	$k = 1, s = .005$	1.0	.690	.692	.750	.529				
	$\mathbf{k = 5, s = .001}$	1.0	.710	**.738**	.818	.529	.295	**.714**	.667	1.0
	$k = 5, s =$ None	1.0	.710	.685	.714	.588				
BERT	APD	-0.2	.673	.598	.560	.824				
	COS	0.1	.738	**.741**	.706	.788	.482	.620	.567	1.0
BL	random sampling							.349	.300	1.0

Table 6.2: Performance (Spearman's ρ, $F_{0.5}$-measure, precission P and recall R) of different approaches on tuning data (SemEval DE targets) and performance of best type- and token-based approach on respective predictions with optimal tuning threshold t, as well as the performance of a randomly sampled baseline.

(see Section 3.1.5). For instance, *Aufkommen* in Figure 6.1 is labeled as change as it gains the orange cluster from C_1 to C_2. As defined in Section 3.1.5, we use k and n as lower frequency thresholds for binary change to avoid that small random fluctuations in sense frequencies caused by sampling variability or annotation error be misclassified as change. For comparability across sample sizes, we propose to generally set $k = 1 \leq 0.01 * |U_i| \leq 3$ and $n = 3 \leq 0.1 * |U_i| \leq 5$, where $|U_i|$ is the number of uses from the respective time period (after removing uses with many 'Cannot decide' judgments from the graphs, see Section 3.2.1.5). This results in $k = 1$ and $n = 3$ for all target words.

For an overview over the final set of WUGs, refer to Table 6.1. We reach a comparable inter-annotator agreement (Spearman's $\rho = .64$) to the SemEval and DURel data sets (cf. Tables 3.4 and 3.6). Find a selection of WUGs from all data sets in Appendix A.[7]

6.4 Results

The evaluation of the discovery predictions from Section 6.2.2 on the annotated data from Section 6.3 is presented in Table 6.2. We achieve a $F_{0.5}$-score of .714 for SGNS and .620 for BERT. Out of the 27 words predicted by the SGNS model, 18 (67

[7] The data is available at `https://www.ims.uni-stuttgart.de/data/wugs` under DiscoWUG V1.0.0.

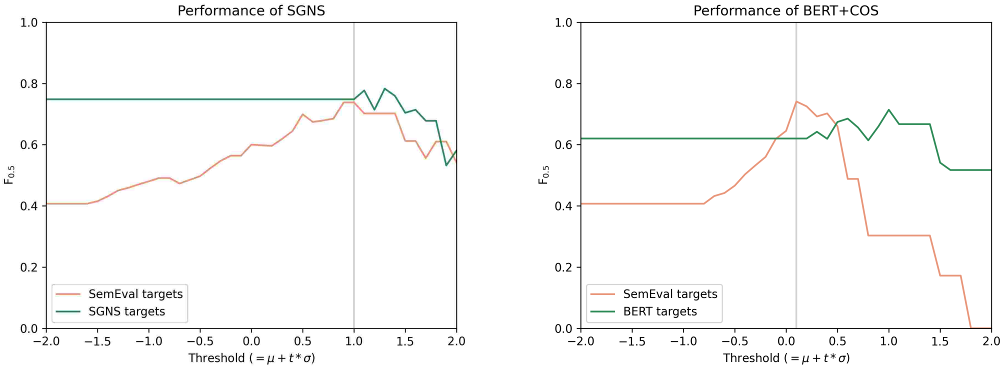
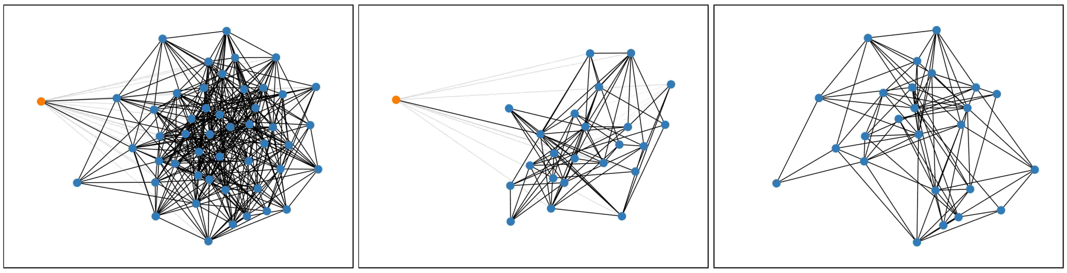


Figure 6.2: $F_{0.5}$ performance on SemEval DE targets (orange) and respective discovery predictions (green) across different thresholds. Left: SGNS. Right: BERT+COS. Gray vertical line indicates optimal performance on SemEval targets.

%) were actually labeled as changing words by the human annotators. In comparison, only 17 out of the 30 (57 %) BERT predictions were annotated as such. The performance of SGNS for discovery (SGNS targets) is even higher than on the tuning data (SemEval targets). In contrast, BERT's performance for discovery drops strongly in comparison to the performance on the tuning data (.741 vs. .620). This replicates our previous results from Chapter 5 that BERT generalizes poorly for LSCD and does not transfer well between data sets. If we compare these results to the baseline, we can see that both models perform much better than the random baseline ($F_{0.5}$ of .349). Only 10 out of the 30 (30 %) randomly sampled words are annotated as changing. This indicates, that the performance of SGNS and BERT is likely not a cause of randomness. Both models considerably increase the chance of finding changing words compared to a random model.

Figure 6.2 shows the detailed $F_{0.5}$ developments across different thresholds on the SemEval targets and the discovered words. Increasing the threshold on the discovered words improves the $F_{0.5}$ for both the type-based and token-based approach. A new high-score of .783 at $t = 1.3$ is achievable for SGNS. While BERT's performance also increases to a peak of .714 at $t = 1.0$, it is still lower than in the tuning phase.



G G_1 G_2

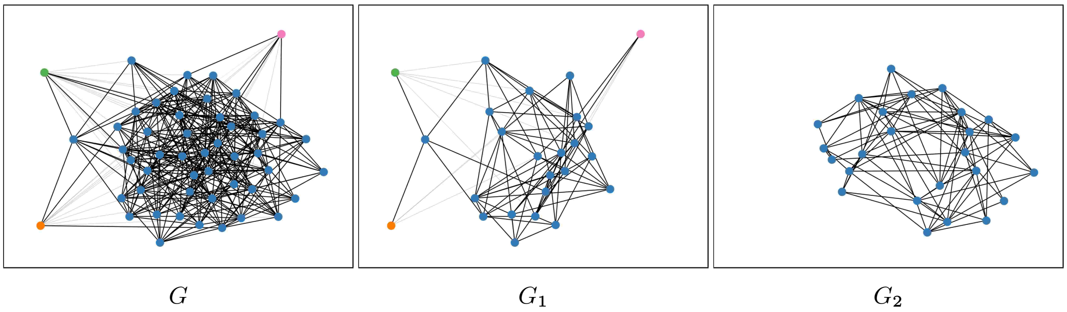
Figure 6.3: WUG of German *Angriffswaffe* (left), subgraphs for 1st time period G_1 (middle) and for 2nd time period G_2 (right).

6.5 Analysis

For further insights into sources of errors, we take a close look at the false positives, their WUGs and the underlying uses. Most of the wrong predictions can be grouped into one out of two error sources (cf. Kutuzov, 2020, pp. 175–182):

Context change The first category includes words where the context in the uses shifts between time periods while the meaning stays the same. The WUG of *Angriffswaffe* ('offensive weapon', see Figure 6.3) shows a single major cluster for both C_1 and C_2. In the first time period, *Angriffswaffe* is used to refer to a hand weapon (such as 'sword', 'spear'). In the second period, however, the context changes to nuclear weaponry. We can see a clear contextual shift while the meaning did not change. In this case both models are tricked by the change of context. Further false positives in this category are the SGNS targets *Ächtung* ('ostracism') and *aussterben* ('to die out') and the BERT targets *Königreich* ('kingdom') and *Waffenruhe* ('ceasefire').

Context variety Words that can be used in a large variety of contexts form the second group of false positives. SGNS falsely predicts *neunjährig* as a changing word. When we take a closer look at its WUG (see Figure 6.4), we observe that there is only one (and the same) cluster in both time periods. The meaning of the target does not change, even though a large variety of contexts exists in both C_1 and C_2. For example: 'which bears oats at **nine years** fertilization', 'courageously, a **nine-year-old** Spaniard did something' and 'after **nine years** of work'. Both models



G G_1 G_2

Figure 6.4: WUG of German *neunjährig* (left), subgraphs for 1st time period G_1 (middle) and for 2nd time period G_2 (right).

are misguided by this large context variety. Further examples include the SGNS target *vorjährig* ('of the previous year') and the BERT targets *bemerken* ('to notice') and *durchdenken* ('to think through').

6.6 Lexicographical Evaluation

We now evaluate the usefulness of the proposed semantic change discovery procedure, including the annotation approach and WUG visualization from a lexicographer's viewpoint. The advantage of our approach lies in providing lexicographers and dictionary makers the choice to take a look into predictions which they consider promising with respect to their research objective (disambiguation of word senses, detection of novel senses, detection of archaisms, describing senses in regard to specific discourses etc.) and the type of dictionary. Visualized annotations for target words may be analyzed in regard to single senses, clusters of senses and the semantic proximity of sense clusters. Random sampling of uses also offers the opportunity to judge underrepresented senses in a sample that might be infrequent in a corpus or during a specific period of time (although currently a high number of overall annotations would be required in order to do so). Most importantly, the use of a variable number of human annotators has the potential to ensure a more intersubjective analysis of large amounts of corpus data. In order to evaluate the potential of the approach for assisting lexicographers with extending dictionaries, we analyze the annotated data for the two sets of model predictions (SGNS, BERT) and compare them to existing dictionary contents.

We consider inter-annotator agreement (Krippendorff's $\alpha >= .5$) and annotated binary change label to select 21 target words for lexicographical analysis. In this way, we exclude unclear cases and non-changing words. The data is analyzed by inspecting interactive cluster visualizations of WUGs (similar to Figure 6.1, see also Section 3.3.1.1) and comparing them to entries in general and specialized dictionaries in order to determine

1. whether a sense derived from the annotation is included in any of the reference dictionaries, indicating the discovery of a previously unknown sense and
2. whether binary change labels derived from the annotation correspond to changes in the entries found in the two reference dictionaries that are consulted for C_1 and C_2.

Three dictionaries are consulted throughout the analysis: (i) the Dictionary of the German language (DWB, 2021) by Jacob und Wilhelm Grimm (digitized version of the 1st print published between 1854–1961), (ii) the Dictionary of Contemporary German (WGD, 2021), published between 1964–1977, now curated and digitized by the DWDS (2021) and (iii) the Duden online dictionary of German language (DUDEN, 2021), reflecting usage of Contemporary German up until today.[8] Additionally, lemma entries in the Wiktionary online dictionary (Wiktionary, 2021) are consulted to verify genuinely novel senses described in Section 6.6.1.

6.6.1 Records of Senses

In the case of 17 target words, all senses identified by the annotation approach are included in at least one of the three dictionaries consulted for the analysis. In the four remaining cases, at least one sense of a word is neither paraphrased nor given as an example of semantically related senses in the dictionaries (see also Figure 6.5):

einbinden Reference to the integration or embedding of details on a topic, event, person in respect to a chronological order within written text or visual presentation (e.g. for an exhibition on an author) is identified by the annotation approach as a gained sense in close semantic proximity to the old sense 'to bind sth. into sth.', e.g. flowers into a bundle of flowers. The word *einbinden* is also used in technical

[8] Only the fully-digitized version of the DWB's first print was consulted for this evaluation since a revised version has not been completed yet and is only available for lemmas starting with letters a–f.

contexts, meaning 'to (physically) implement parts of a construction or machine into their intended slots'.

niederschlagen In cases where the verb *niederschlagen* co-occurs with the verb particle *auf* and the noun *Flügel*, the verb refers to a bird's action of repeatedly moving its wings up and down in order to fly.

regelrecht Used as an adverb, *regelrecht* may refer to something being the usual outcome that ought to be expected due to scientific principles, with an emphasis on the actual result of an action (such as the dyeing of fiber of a piece of clothing following the bleaching process) whereas senses included in dictionaries for general language emphasize either the intended accordance with a rule or something usually happening (the latter being colloquial use).

Zehner The sense 'a winning sequence of numbers in the national lottery', annotated to be gained between C_1 and C_2, is not included in any of the reference dictionaries.

Additionally, we consult a dictionary for Early New High German (FHDW) in order to check whether discovered senses existed at an earlier stage and thus likely may be discovered due to low frequency or sampling error. In two cases, discovered senses that are not included in the DWB (for C_1) are found to be included in the FHDW (2021). Interestingly, one sense paraphrased for *Ausrufung* ('a loud wording, a shout') is neither included in DWB nor in WGD, but in the FHDW (earlier) and DUDEN (as of now). These findings suggest that it might be reasonable to use more than two reference corpora. This would also alleviate the corpus bias, stemming from idiosyncratic data sampling procedures.

Note that some of the senses described in this section might still be included in more specialized dictionaries, which we did not check, e.g. technical usage of *einbinden*.

6.6.2 Records of Changes

For 12 target words, the binary semantic change predicted by the models correlates with the addition or non-inclusion of senses in dictionary entries consulted for the respective period of time (DWB for C_1, WGD for C_2). Notably, lemma lists of the two dictionaries might be lacking lemmas in the headword list. Moreover, lemma entries might be lacking paraphrases or examples of senses of the lemma simply because corpus-based lexicography was not available at the time of their first print and revisions of the dictionaries are a work in progress.

6.7 Discussion

In this section, we tuned the state-of-the-art approaches to LSCD identified in Chapter 5 on the SemEval DE data set for Binary Change Classification to automatically discover semantic changes in the SemEval DE diachronic corpus pair. While the type-based approach showed better performance, both approaches were able to discover various semantic changes with above-random probability, some of them previously undescribed in etymological dictionaries.

We validated model predictions by the annotation process developed in Chapter 3, yielding a comparable inter-annotator agreement and providing convenient ways of visualization. In addition, we evaluated the full discovery process from a lexicographer's point of view. The results endorse that our approach might aid lexicographers in extending and altering dictionary entries. On the other hand, we identified context change and context variability as weak spots of the applied models, suggesting ways to further improve the latter. The lexicographical analysis further showed that it is important to compare more than two reference corpora to detect LSC.

As we have discussed in Section 3.4, the change scores resulting from our annotation process by themselves are merely measurements of word sense divergences and as such can indicate LSC in the entire speaker community only in connection with an adequate word use sampling procedure. Such a procedure should guarantee that it samples widely throughout a speaker population and ideally takes a large random sample from each time period. It should avoid to sample only from specific text genres as this would likely bias the sample towards a particular speaker subpopulation. However, the corpus for the second time period of the SemEval DE data set is composed of texts from only two East German newspapers (see Section 3.2.1.1). It is thus possible that some observed changes in the annotated data would not be observed in a more general sample (and vice versa). While the lexicographical analysis indicates that this is not a frequent problem in our data and newspaper corpora are frequently used in historical lexicography (Gloning, 2017), future studies should aim to alleviate this problem by choosing corpora representing as much language variety as possible.

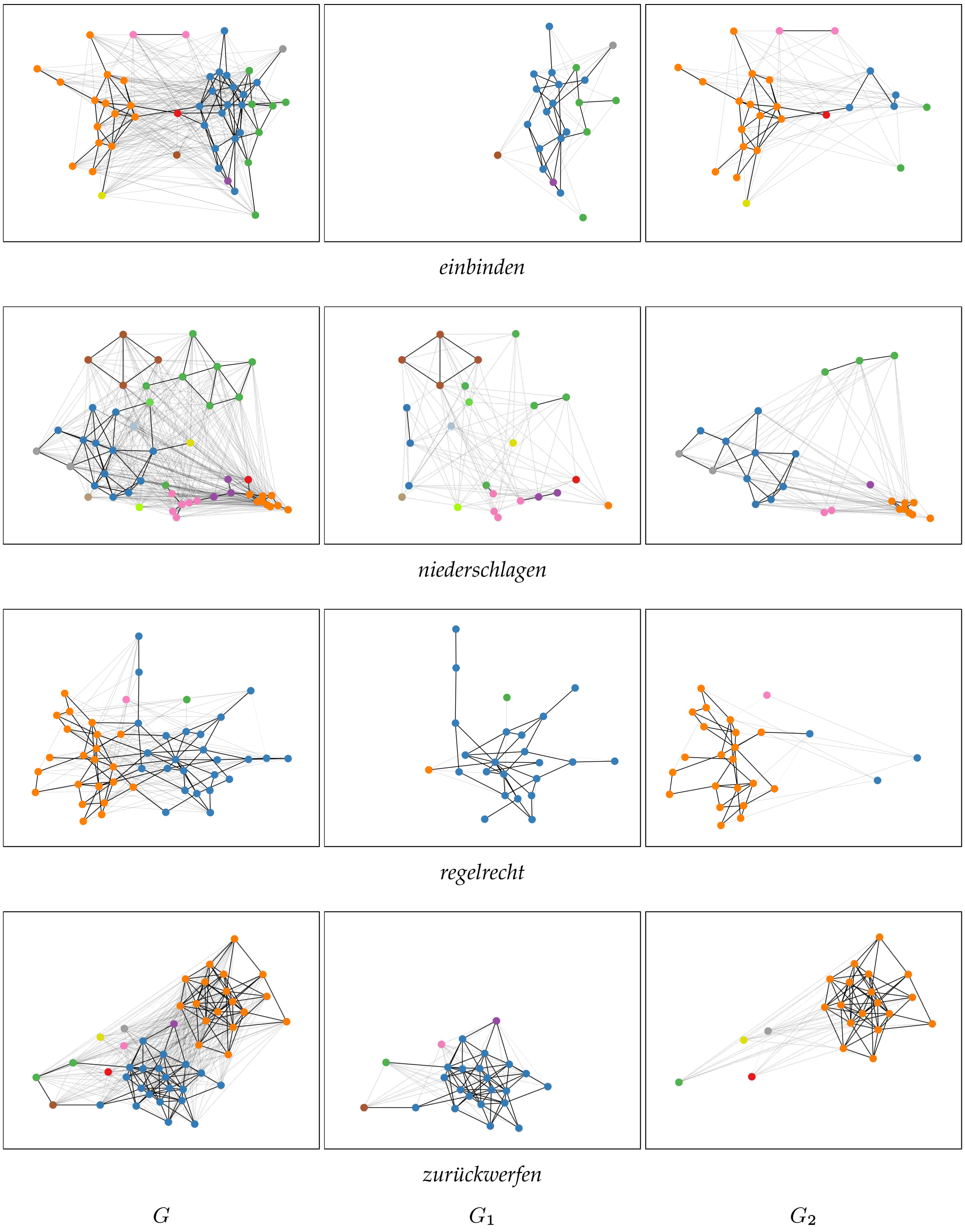


Figure 6.5: WUGs from DiscoWUG with discovered senses. Full graph G (left), subgraphs for 1st time period G_1 (middle) and 2nd time period G_2 (right).

Chapter 7

Conclusion

This dissertation provides a complete evaluation framework for the task of Lexical Semantic Change Detection. We started out by extending and optimizing **word use pair** proximity annotation to gather large amounts of data. This annotation procedure was chosen because it relies on the simple and intuitive concept of **semantic proximity**, which gives an explicit criterion for clustering in the lexicographic process deriving **word senses** and is used in Blank's widely accepted theory of LSC. It further avoids the need for word sense definitions, considerably reducing the efforts needed in the preparation of an annotation study. However, this annotation procedure quadratically increases the annotation load. Hence, we needed to develop sampling procedures iteratively finding necessary comparisons to infer a meaningful clustering on the annotated graphs. We defined and motivated a **binary** and a **graded** measure of semantic change based on these clusterings of word uses. They can be seen as estimates of the change scores underlying the full corpus. While the graded measure can be estimated rather well from feasible sample sizes, the binary measure is less reliable. We also defined and annotated a third change measure avoiding clustering and showed that it empirically approximates the graded measure of semantic change well even with small sample sizes. This is an important result because it suggests simpler annotation strategies. We further validated the levels of the annotation process through annotator agreement on the concept of semantic proximity, correspondence of the clusterings to an independent annotation with sense definitions, a manual analysis of annotator disagreements and the inferred clusterings, as well as robustness of change scores after an additional round of annotation. The overall agreement was reasonably high, clearly above chance and comparable to previous studies. Similarly for the correspondence to sense definitions, but an additional round of annotation showed clear improvements on the correspondence. The manual analysis revealed that **ambiguity** and **sparsity** are the major challenges for our annotation approach. Finally, the change scores showed a high robustness for some data sets, but considerable variation for others. We concluded that the current change scores should be seen as a silver rather than a gold standard.

In the next step, we implemented a range of models from the existing literature including our own extensions. **Token-based** models infer a meaning representation on the word use level while **type-based** models infer a meaning representation on the word level. The former can be seen as a model of the lexicographic clustering process, deriving word senses, while the latter in most cases do not allow such an interpretation straightforwardly. These models were then tested in three studies on one binary classification and two ranking tasks, requiring models to predict the change scores derived on the annotated data. We found good performances for both types of models on individual data sets with an overall advantage for the type-based approaches, being more robust across data sets and outperforming the token-based approaches even on the binary change task. However, the tasks still remain far from being solved. The binary task will be a particular challenge in the future: Type-based approaches to binary change detection use thresholding and are thus exploiting empirically observed correlations between graded and binary change in the existing data sets rather than actually solving the task (Zamora-Reina et al., 2022).

Skip-Gram with Negative Sampling + Orthogonal Procrustes + Cosine Distance (including variations in alignment and change measure) turned out to be a reliable model with high performance on several data sets and high robustness. We further improved the performance of this approach through mean centering in the alignment step. We then tested several hypotheses on potential biases in the commonly applied token-based model BERT and found that it is influenced by various factors, but most strongly by target word form if word uses are not preprocessed. By removing the form bias, we were able to considerably improve the performance across languages, but still found considerable variation of results across parameters and data sets. We also found that using the lemmatized word uses (as provided to SemEval shared task participants in the evaluation phase) often has a negative impact on performance compared to other preprocessing variants. Although we reached a rather high performance with clustering for Graded Change Ranking in German, average measures still performed better than clustering-based approaches. The reasons for this are unclear and should be addressed in future research.

In a final step, we tuned the state-of-the-art approaches to LSCD identified in Chapter 5 for Binary Change Classification to **automatically discover semantic changes** in a diachronic corpus pair. While the type-based approach showed better performance, both approaches were able to discover various semantic changes with above-random probability, some of them previously undescribed in etymological dictionaries. We validated model predictions by the annotation process developed in Chapter 3 and evaluated the full discovery process from a lexicographer's point of view. The results of the analyses endorse that our approach might aid lexicographers with extending and altering existing dictionary entries as well as discovering new changing words. On the other hand, we identified **context change** and **context variability** as weak spots of the applied models, suggesting ways to improve these. The lexicographical analysis further showed that it is important to compare more than two reference corpora to detect LSC.

In summary, we can say that we tried to standardize the field of LSCD, but have to admit that we are only at the beginning of this process. More work has to be done on controlling evaluation data quality and understanding model performance. Especially the binary change score cannot be assumed to generalize well to the full corpora. Hence, Binary Change Classification should preferably be done only on the annotated use samples. The **data quality** should be further improved by (i) adding more annotations to the existing data sets to reduce sparsity or to increase the time ranges covered, (ii) annotating the existing data sets with alternative annotation strategies such as sense definitions or (iii) cleaning the existing data sets by removing words with low agreement, high sparsity or high clustering loss.[1] Because of the late availability of data sets or models in the progression of this thesis, not all models were tested on all data sets and tasks. These missing evaluations should be done in the future to get a full picture of model performances.

By defining LSCD tasks only for two corpora, we have gained **feasibility** of annotation and **simplicity** of the tasks, but also limited the potentials of models. The traditional examples of semantic changes (see Section 2.1) describe developments over long periods of time. Hence, a comparison of change predictions for **multiple time points** may help models to distinguish fluctuations in meaning stemming from sampling variability or bias from true (enduring) semantic changes over

[1] Some of these proposals are already implemented into the LSCD Benchmark: `https://github.com/ChangeIsKey/LSCDBenchmark`.

longer periods of time. Models of such continuous changes have been proposed in recent years (Fremann and Lapata, 2016; Rosenfeld and Erk, 2018; Tsakalidis and Liakata, 2020).

Our results with respect to the dominance of type-based models have to be seen in context of the latest research: In the latest shared tasks on Russian and Spanish data (Kutuzov and Pivovarova, 2021a; Zamora-Reina et al., 2022), type-based approaches such as SGNS+OP+CD were clearly outperformed by **Word-in-Context** (WiC) models (Pilehvar and Camacho-Collados, 2019) using BERT (or XLMR) vectors as input features (Arefyev and Bykov, 2021; Arefyev and Rachinskiy, 2021; Arefyev et al., 2021; Homskiy and Arefyev, 2022). These models learn to distinguish the meanings of word use pairs from (binary) human semantic proximity judgments of such pairs and can thus be seen as an optimized model for semantic proximity. Hence, their good performance can be explained with the theoretical background given in this thesis and the measurement process of LSC developed from it, where semantic proximity is the most fundamental concept. A promising direction of future research will be to exploit the thousands of semantic proximity judgments which we and follow-up studies annotated for optimization of WiC models for LSCD (cf. Arefyev et al., 2021).[2] A major advantage of this development is that while contextualized embeddings are expensive to train from scratch, we can extract contextualized meaning representations reflecting historical word meanings once they are trained without needing large amounts of training data. In contrast, type-based embeddings are usually trained from scratch and need relatively large amounts of training data which is not always available for historical languages.

The evaluation approaches developed in this thesis have inspired a range of follow-up studies. Three shared tasks have been organized on Italian (Basile et al., 2020), Russian (Kutuzov and Pivovarova, 2021a) and Spanish (Zamora-Reina et al., 2022) data, respectively, adopting (parts of) our evaluation setup. Several studies have adopted our annotation framework (Giulianelli et al., 2020; Rodina and Kutuzov, 2020; Kutuzov and Pivovarova, 2021a; Kutuzov et al., 2022; Zamora-Reina et al., 2022; Baldissin et al., 2022; Aksenova et al., 2022). The annotation style using semantic proximity between word uses and the Negated COMPARE measure has inspired the transfer of WiC (Pilehvar and Camacho-Collados, 2019) models to LSCD resulting in the above-described leap of performance for token-based mod-

[2] These data sets are available at `https://www.ims.uni-stuttgart.de/data/wugs`.

els (Arefyev et al., 2021). Furthermore, there has been a range of follow-up research on the annotation and clustering of WUGs (Schlechtweg et al., 2021a; Kotchourko, 2021; Tunc, 2021).

7.1 Implications for other Research Fields

This thesis contributes to research in different fields on multiple levels. We discuss the most important implications, which we see for historical and evolutionary linguistics, cognitive semantics and lexical semantics, lexicography, humanities and social sciences.

We tried to detect **where** word senses are lost or gained, but not **why** this happens. The latter question is a major concern in **historical linguistics** and **evolutionary linguistics**. However, an answer to the why presupposes an answer to the where, i.e., we can only find out why LSC happens if we know where it happens and can analyze the detected instances. Hence, this thesis provides the tools necessary to investigate the **causes** of LSC on a larger scale than previously possible (as done e.g. by Hamilton et al., 2016b). However, it is important to validate models and to investigate their biases before testing scientific hypotheses about LSC (Dubossarsky et al., 2017). Although we have taken a semasiological perspective, the presented methods described in Chapter 3 and Chapter 4 can be used in an **onomasiological** setting, relating the meanings of uses from *different* words (Baldissin et al., 2022). In this way, we can e.g. detect synonyms and their semantic developments over time, which can contribute to understand the causes of LSC (Turney and Mohammad, 2019). With more and more historical text resources being digitized, text data becomes an increasingly important source to study such long-term language developments.

We generated large amounts of data with the annotation process described in Chapter 3. These human semantic proximity judgements can be used to answer **cognitive questions** about word meaning. This can be done, for instance, by formulating precise (e.g. Bayesian) models of word meaning and comparing models encoding different hypotheses regarding their fit to the data (Schlechtweg et al., 2021a). This can help to answer questions about the discreteness of word senses or the universality of semantic proximity (see Section 3.4). Our data is particularly interesting because word uses were *randomly* sampled from corpora, thus representing realistic and rather unbiased language samples in contrast to similar, but strongly cleaned datasets (Pilehvar and Camacho-Collados, 2019; Armendariz et al., 2020). Moreover, the word sense frequency distributions inferred on the annotated data are valuable information to understand statistical properties of word senses in **lexical semantics** (Kilgarriff, 2004) and, importantly, to connect these to their changes.

Chapter 3 describes a theoretically-grounded human annotation approach formalizing the lexicographic clustering process and applying scientific measurement techniques common in the behavioral sciences. This can serve as a starting point to approach **lexicography** more from an empirical scientific perspective (Margalitadze, 2018). The annotation approach has been implemented into an online interface that can be used by lexicographers to measure word meaning and meaning change.[3] The computational models described in Chapter 4, which we used to automate the measurement of word meaning and meaning change, have a huge potential to be used for lexicographical analysis, as we demonstrated in Chapter 6. These models will also be subsequently implemented into the online annotation interface to make them usable for a wider public. Further, the data we created or gathered can be useful to both, **historical linguistics** and lexicography. This includes larger lists of words and (noisy) change scores assembled in the pre-annotation phase for SemEval-2020 Task 1 or the annotated model predictions for changes from Chapter 6.[4]

Both, the human as well as the computational measurement approach we described in Chapter 3 and Chapter 4, are of importance to the **humanities** and **social sciences**, where language use is often seen as an indicator of cultural developments (Hengchen et al., 2019; Hamilton et al., 2016a; Ferrara et al., 2022), short-term social or political developments (Kutuzov et al., 2017), or social or political differences between groups (Ceron et al., 2022; Nanni et al., 2022). The computational methods can be used to predict large-scale semantic divergences while the human annotation may serve to validate a sample of predictions on the particular data.

[3] https://www.ims.uni-stuttgart.de/data/durel-tool

[4] Find the lists at https://www.ims.uni-stuttgart.de/data/wugs under DWUG DE/EN/SV V1.0.0.

Appendix A

Word Usage Graphs

Selected WUGs from each Data Set

Figures A.1–A.14 show WUGs for selected words from each data set described in Chapter 3 and Chapter 6, including for each word the full graph G (left), the subgraph for 1st time period G_1 (middle) and the subgraph for the 2nd time period G_2 (right).

Comparison of Clusterings on SemEval DE

Figures A.15 and A.16 show the comparison of clusterings for German *abbauen*, *abgebrüht*, *Knotenpunkt*, *Manschette*, *zersetzen* from SemEval DE, as described in Section 3.3.2: sense description clustering (left), corresponding semantic proximity clustering for same nodes (middle) and full semantic proximity clustering (right). Node positions for each plot are generated on the full WUG.

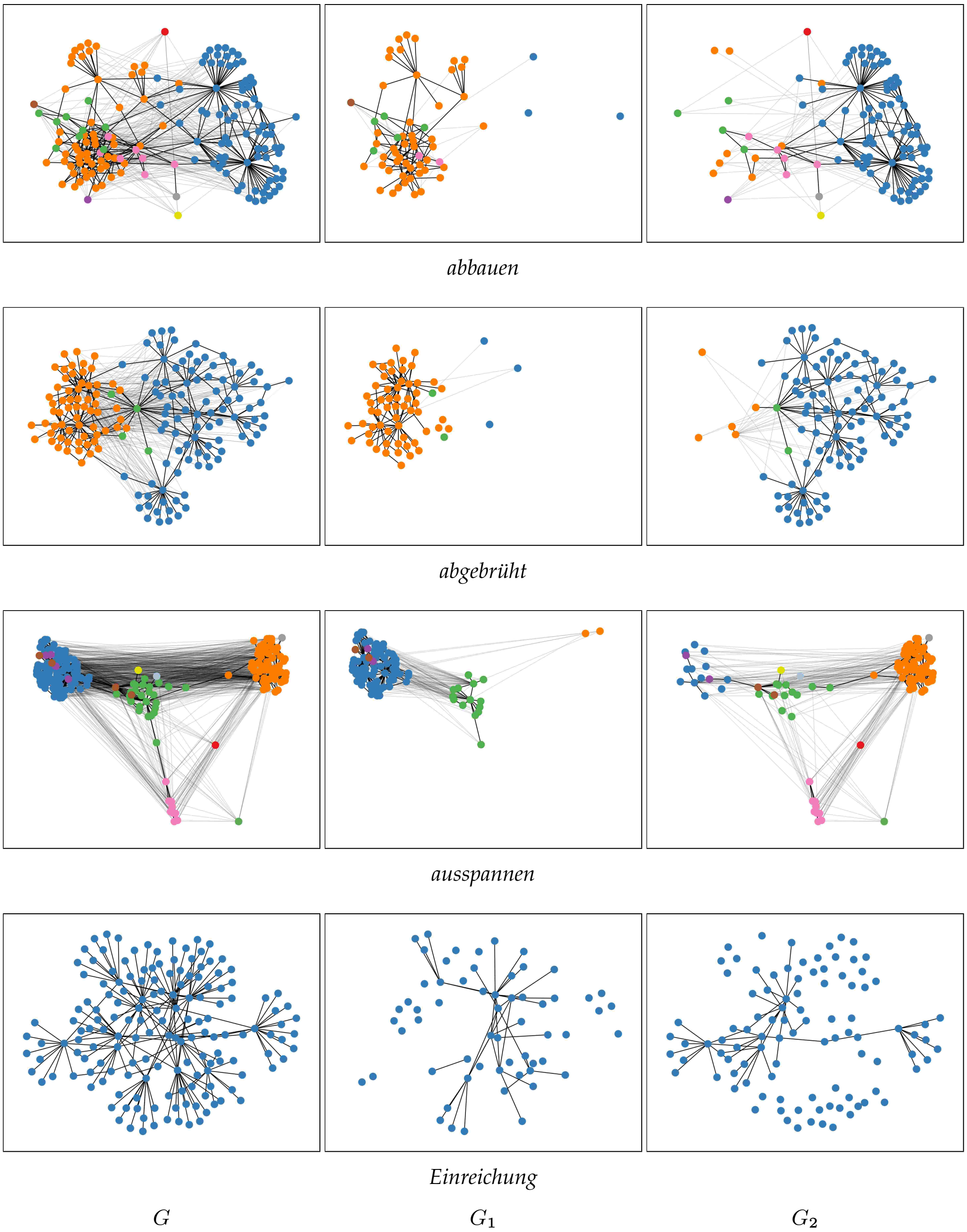
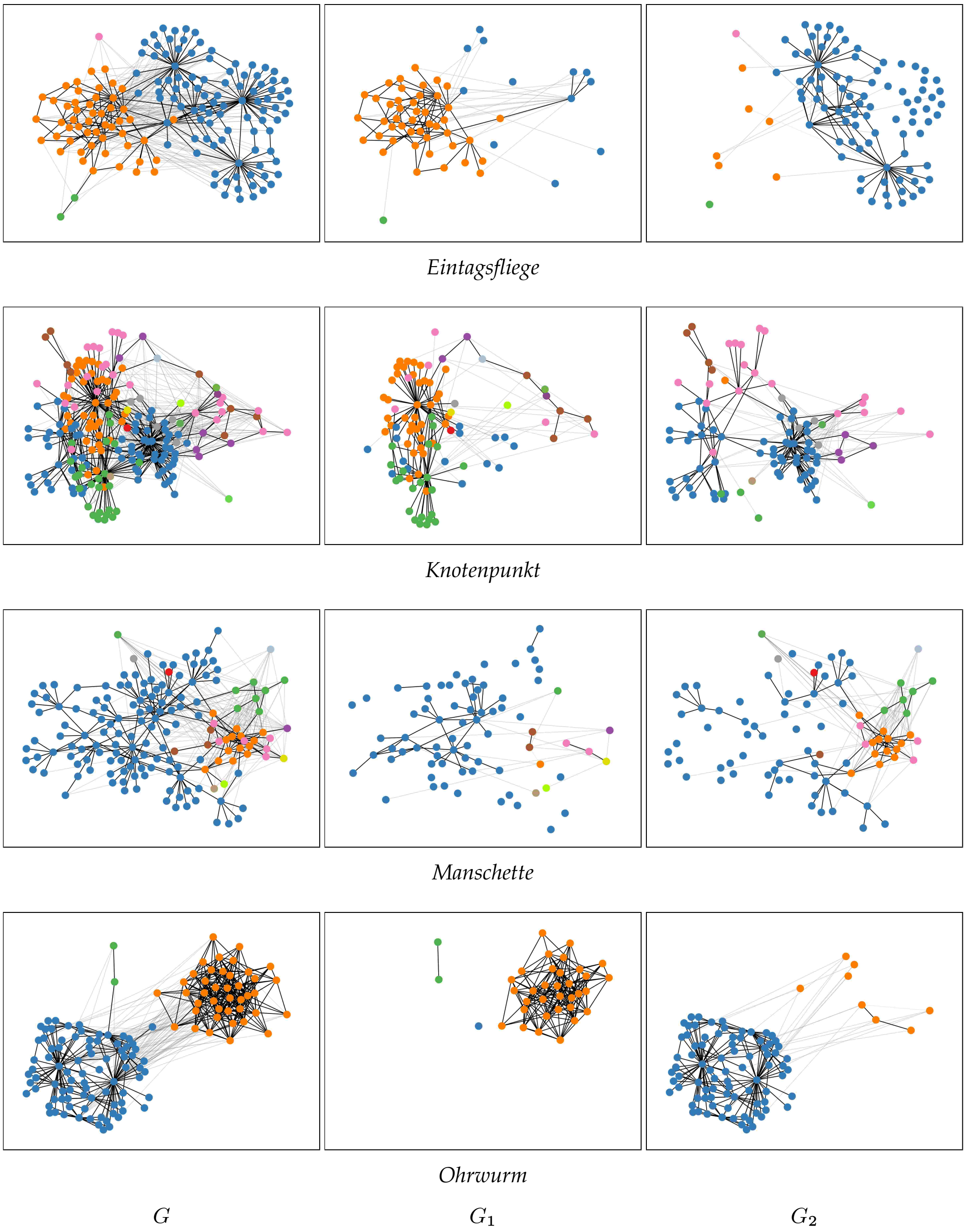


Figure A.1: WUGs from SemEval DE. Full graph G (left), subgraphs for 1st time period G_1 (middle) and 2nd time period G_2 (right).



Eintagsfliege

Knotenpunkt

Manschette

Ohrwurm

G G_1 G_2

Figure A.2: WUGs from SemEval DE. Full graph G (left), subgraphs for 1st time period G_1 (middle) and 2nd time period G_2 (right).

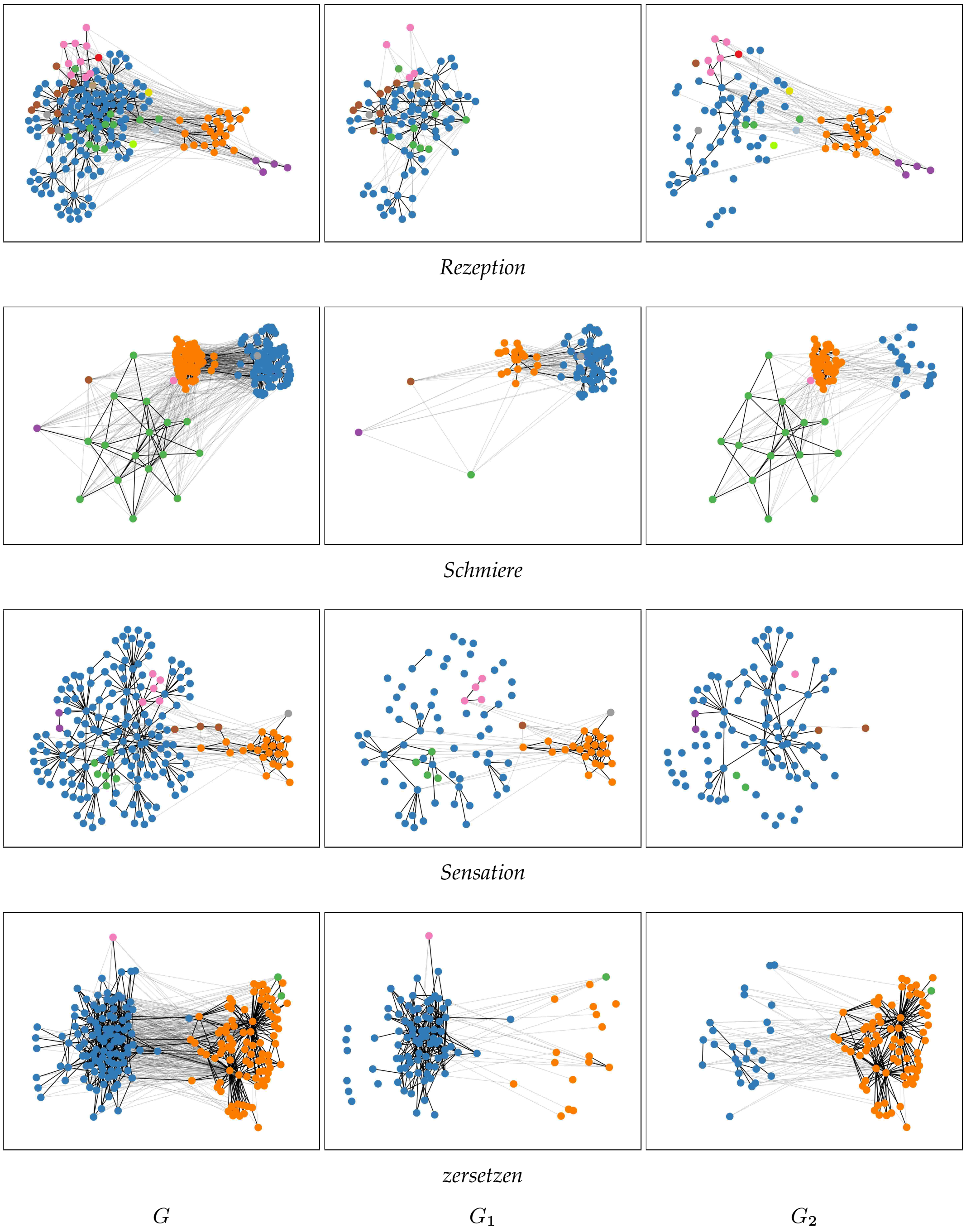


Figure A.3: WUGs from SemEval DE. Full graph G (left), subgraphs for 1st time period G_1 (middle) and 2nd time period G_2 (right).

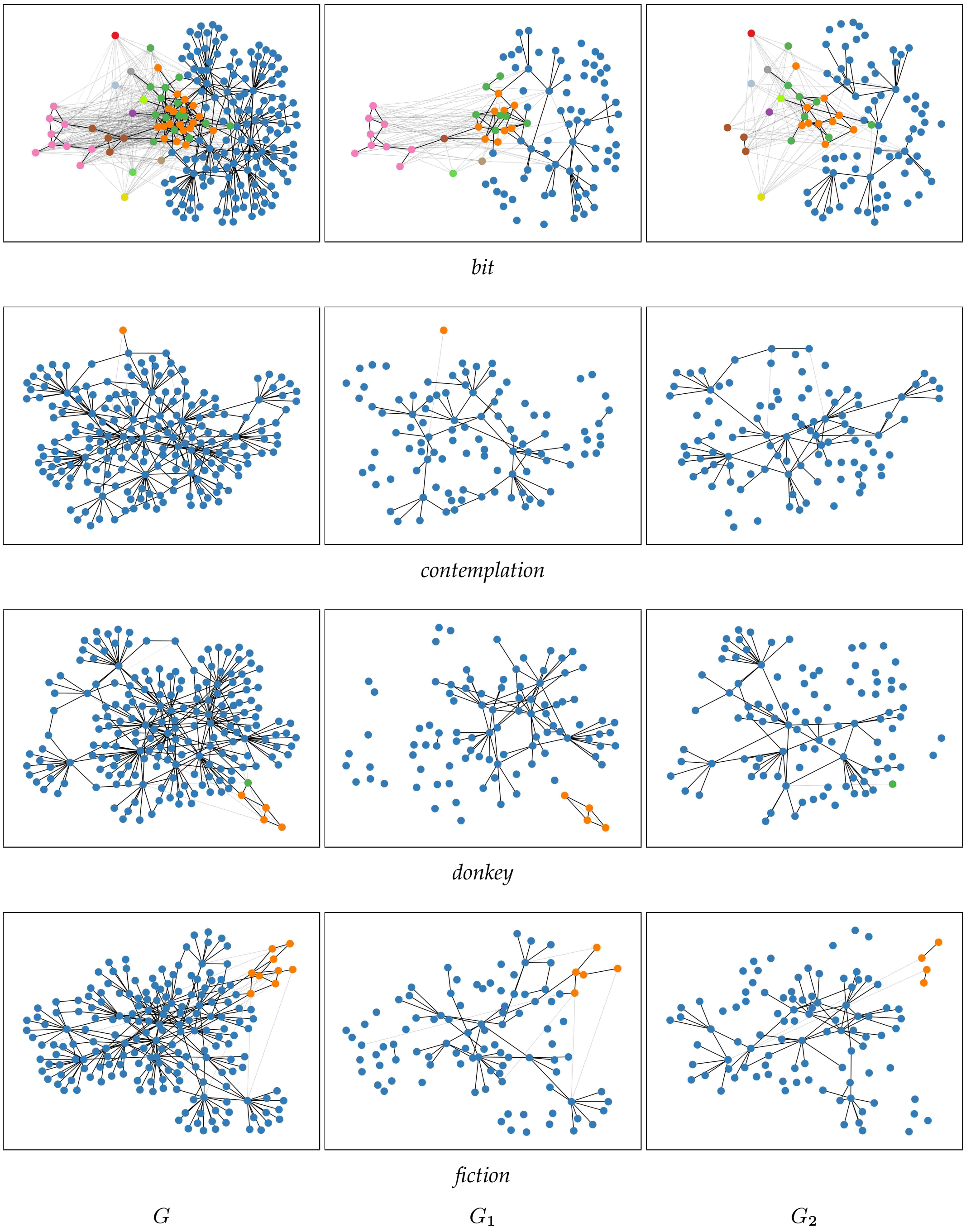
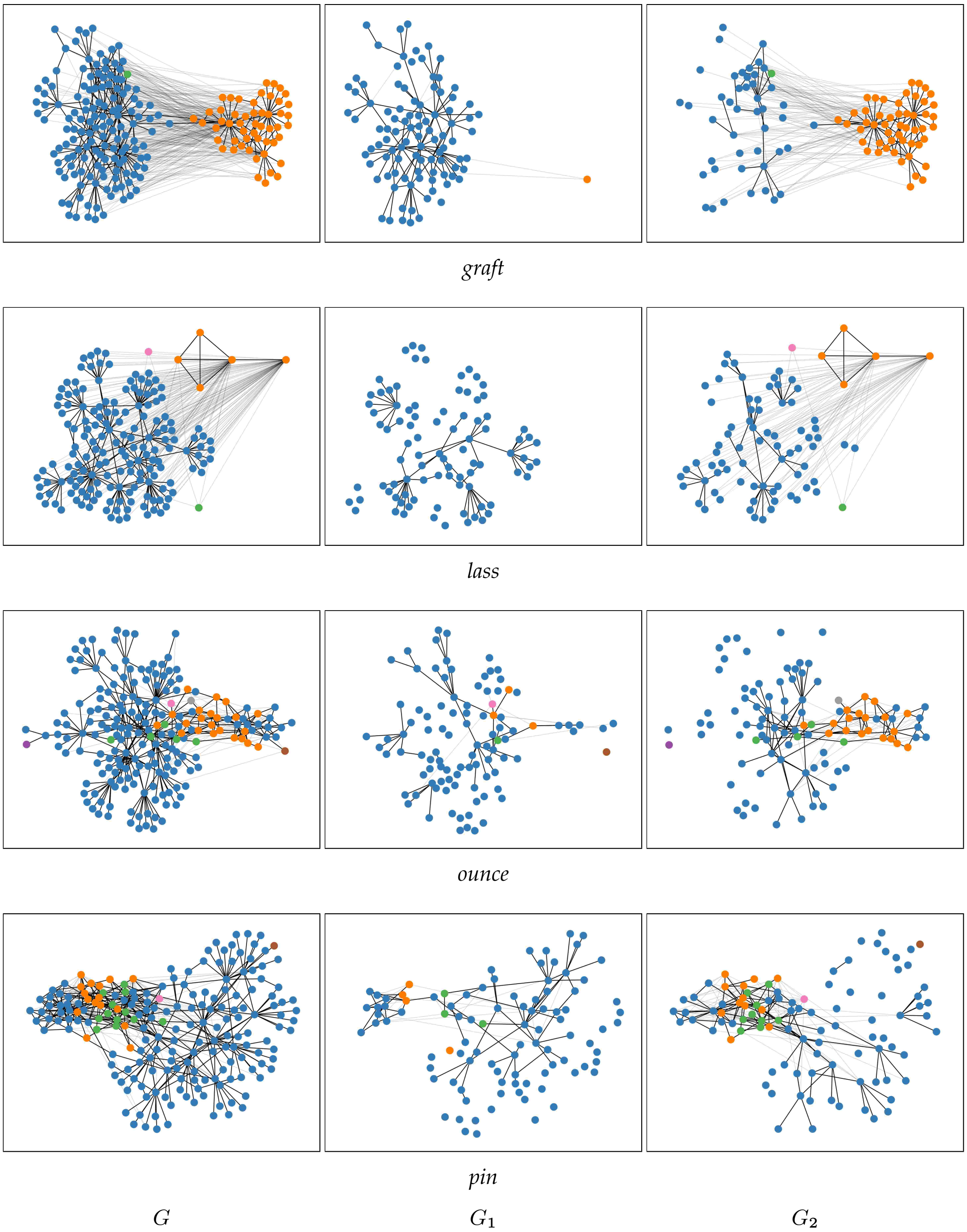


Figure A.4: WUGs from SemEval EN. Full graph G (left), subgraphs for 1st time period G_1 (middle) and 2nd time period G_2 (right).



graft

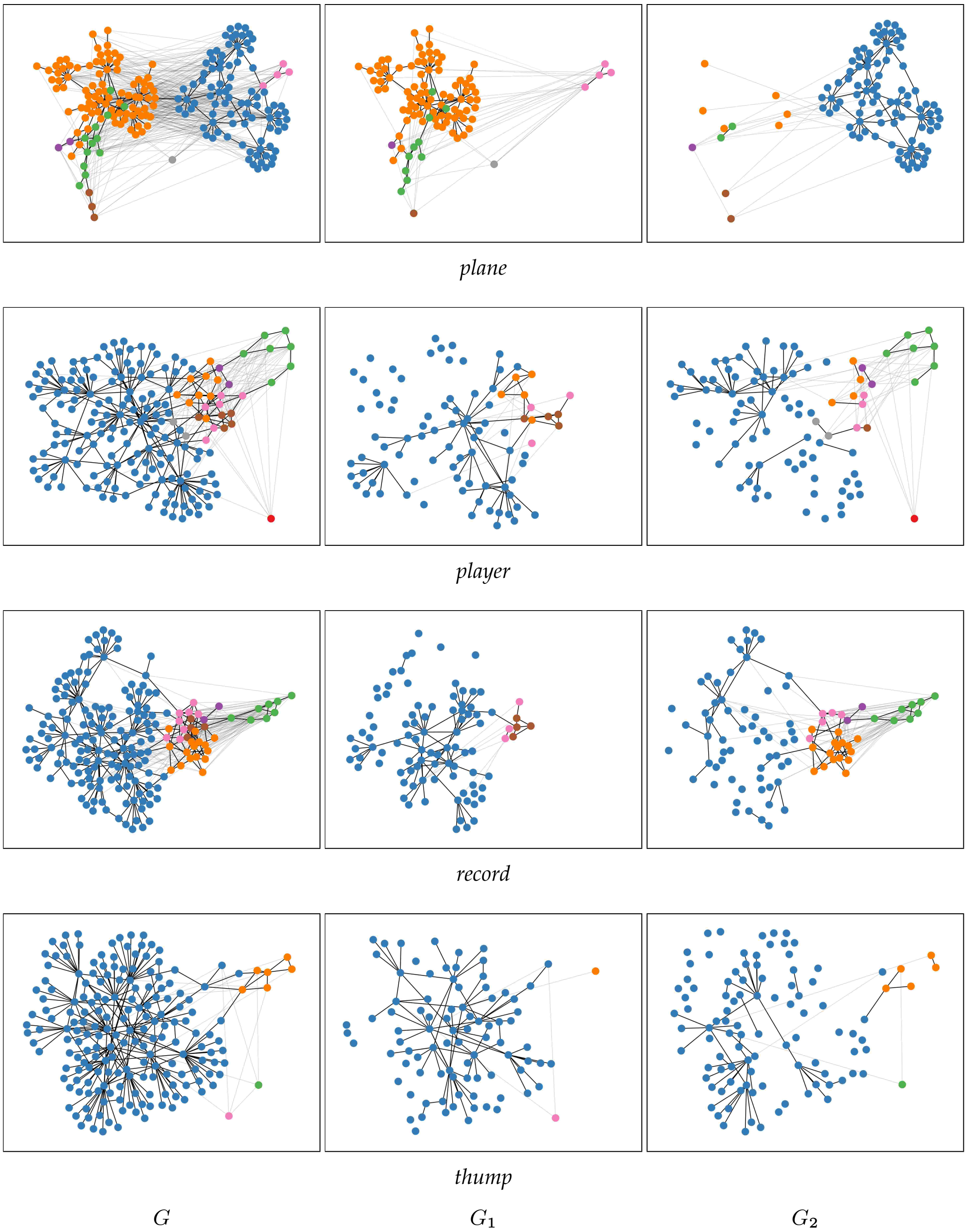
lass

ounce

pin

G G_1 G_2

Figure A.5: WUGs from SemEval EN. Full graph G (left), subgraphs for 1st time period G_1 (middle) and 2nd time period G_2 (right).



plane

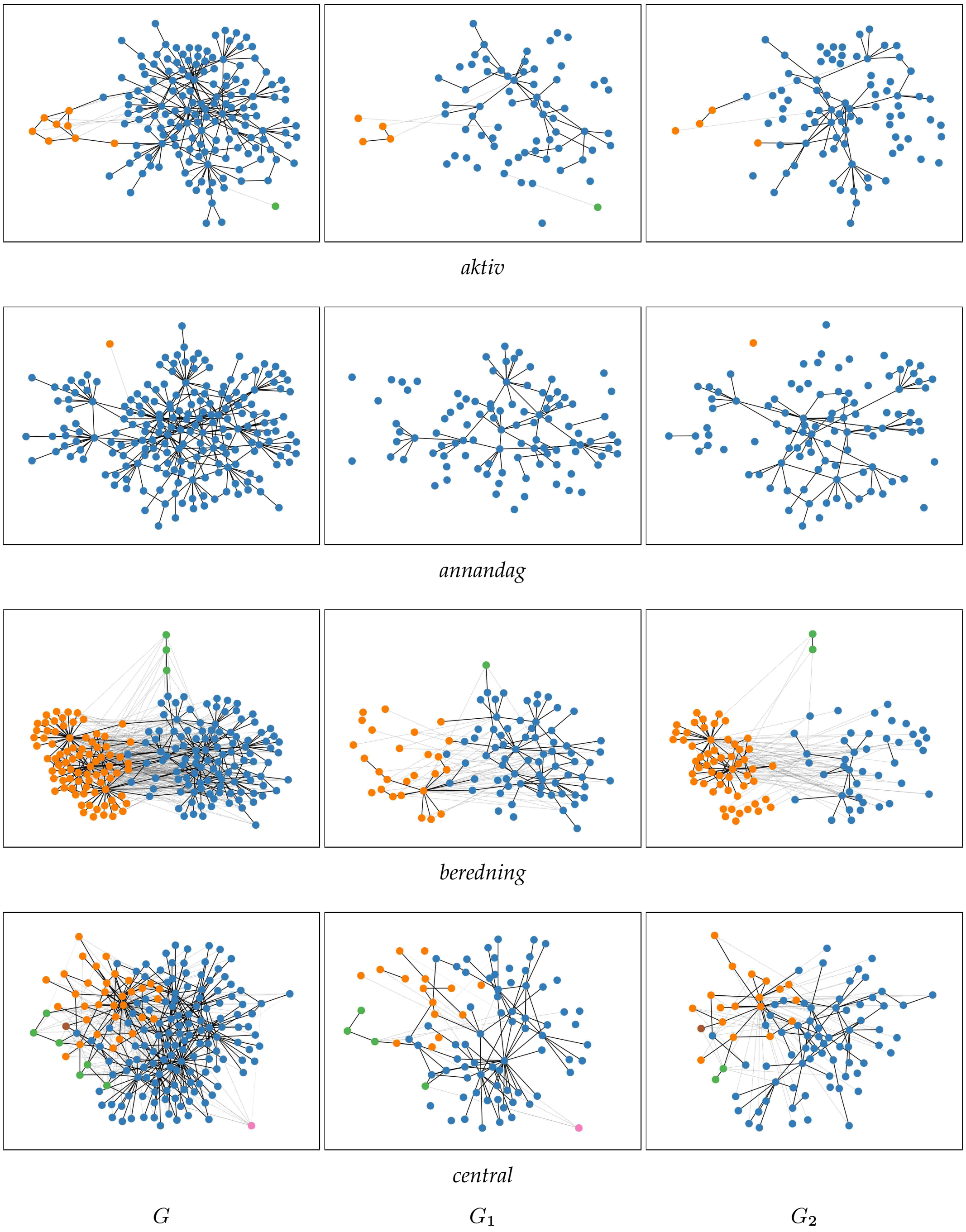
player

record

thump

G G_1 G_2

Figure A.6: WUGs from SemEval EN. Full graph G (left), subgraphs for 1st time period G_1 (middle) and 2nd time period G_2 (right).



aktiv

annandag

beredning

central

Figure A.7: WUGs from SemEval SV. Full graph G (left), subgraphs for 1st time period G_1 (middle) and 2nd time period G_2 (right).

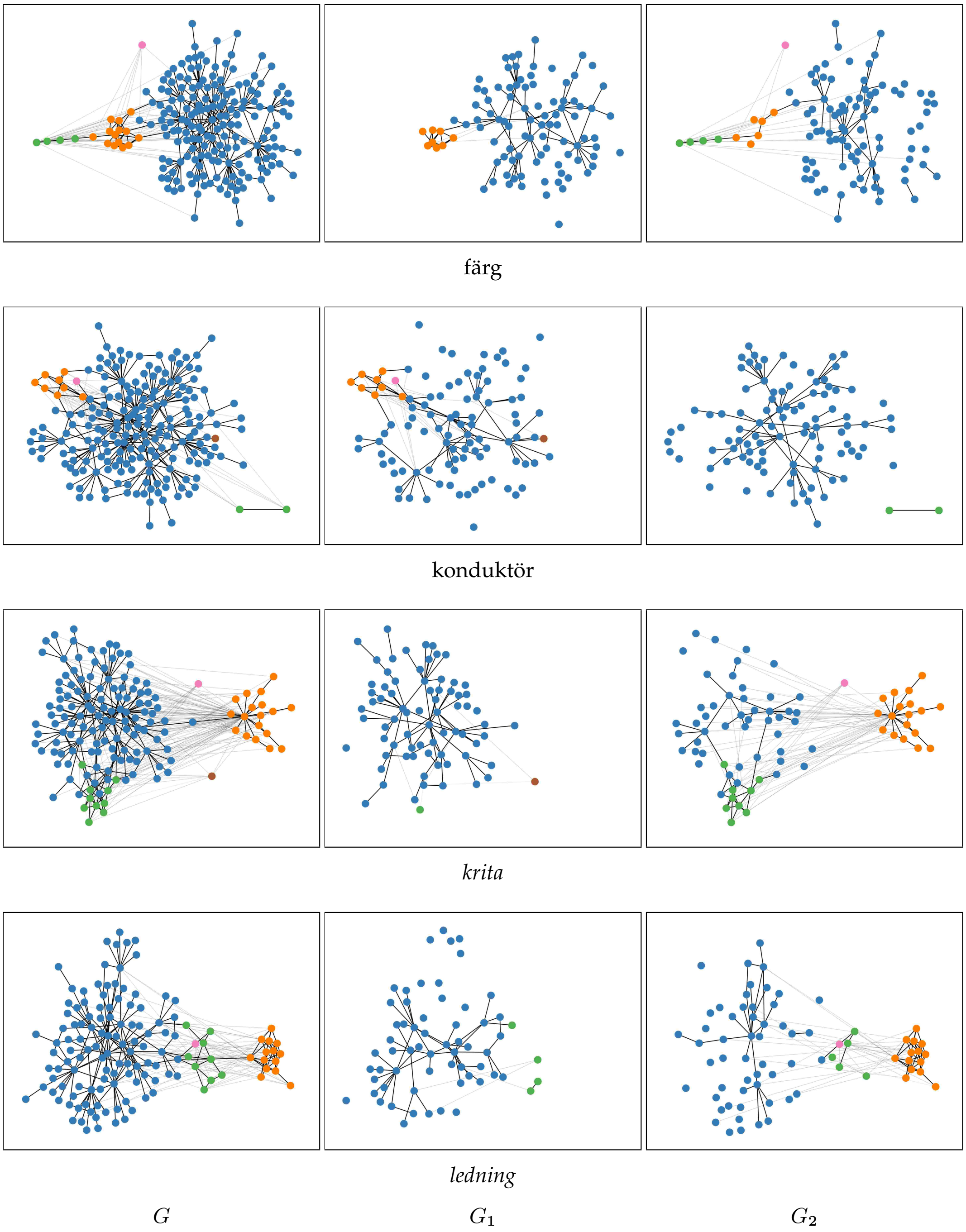


Figure A.8: WUGs from SemEval SV. Full graph G (left), subgraphs for 1st time period G_1 (middle) and 2nd time period G_2 (right).

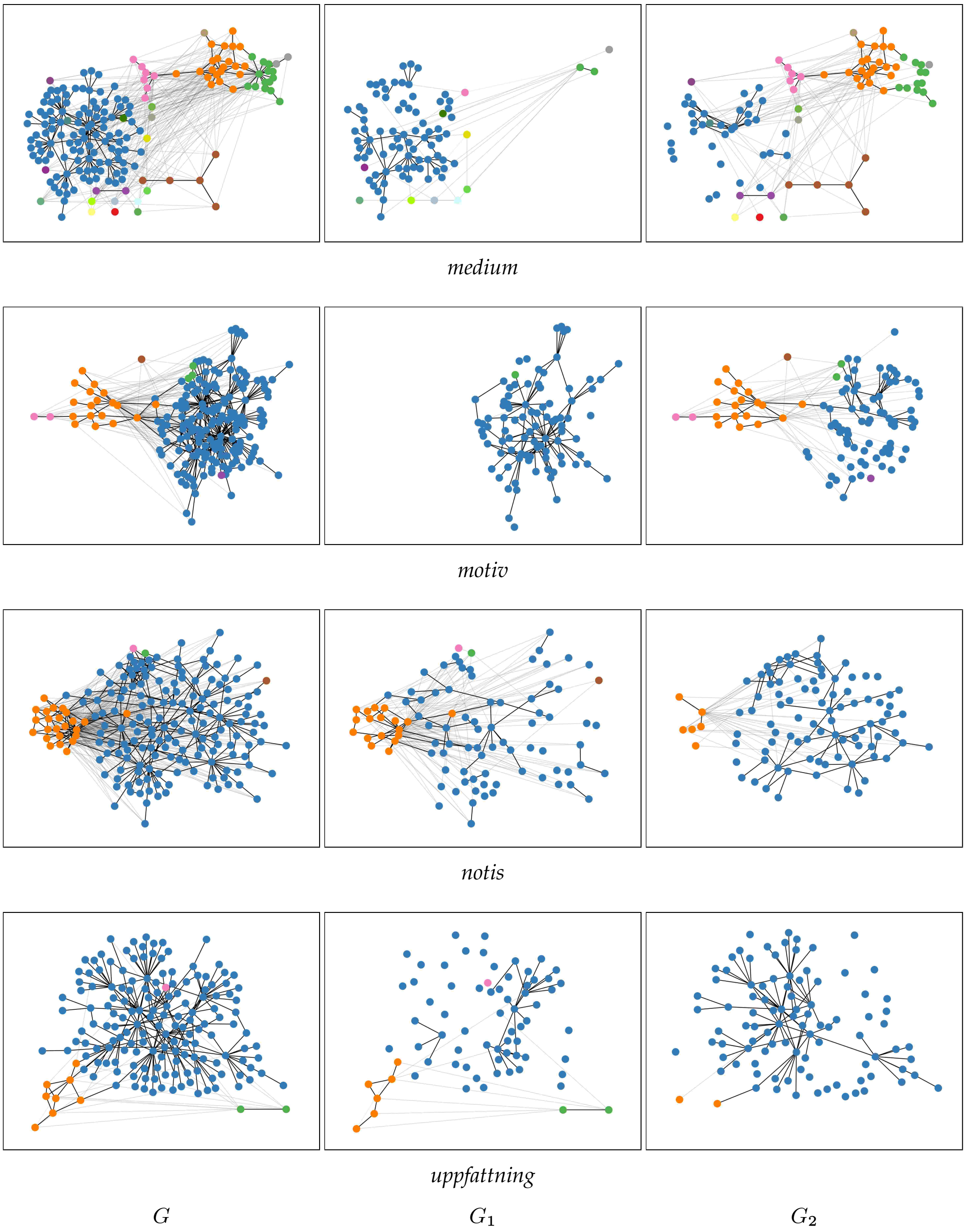


Figure A.9: WUGs from SemEval SV. Full graph G (left), subgraphs for 1st time period G_1 (middle) and 2nd time period G_2 (right).

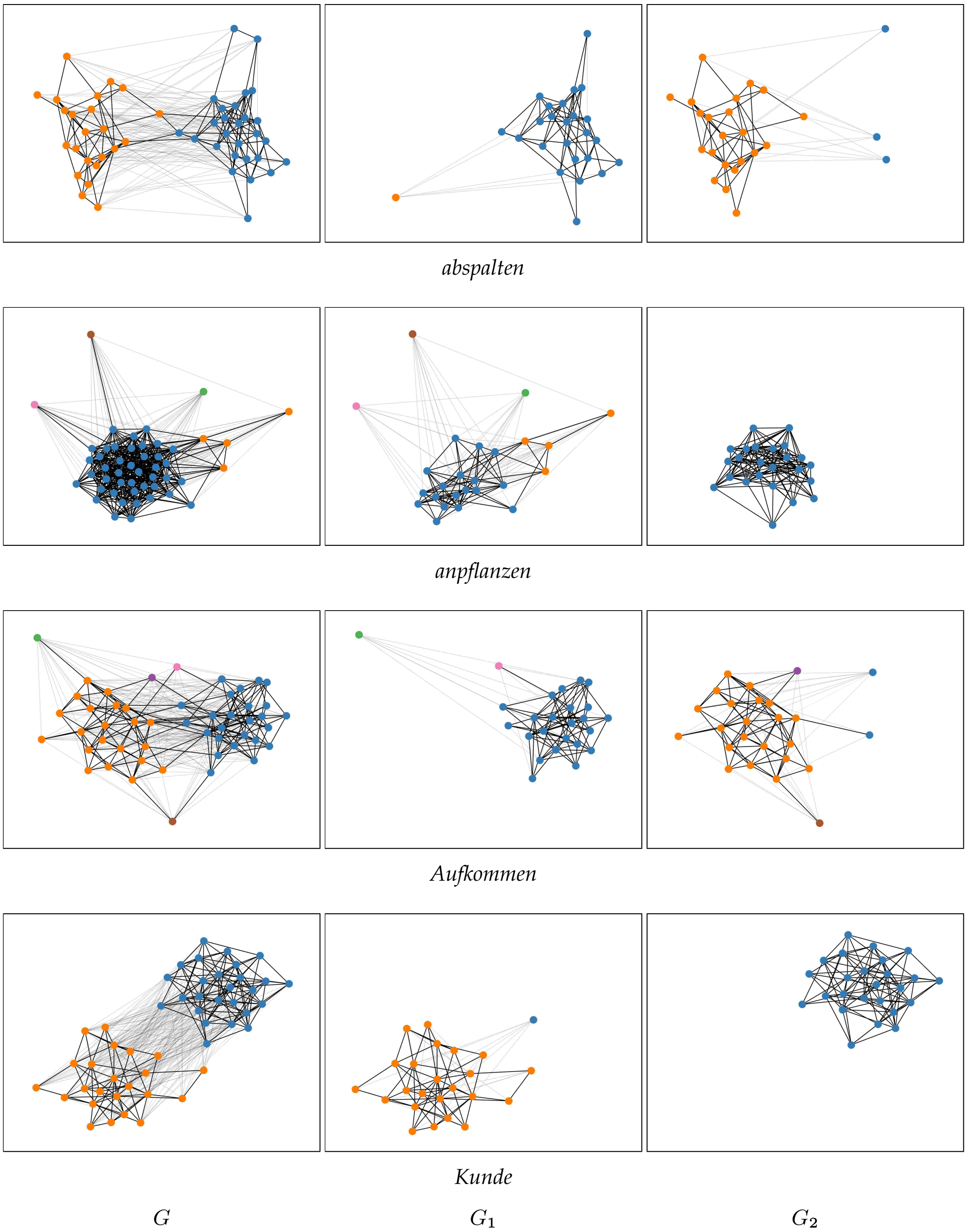
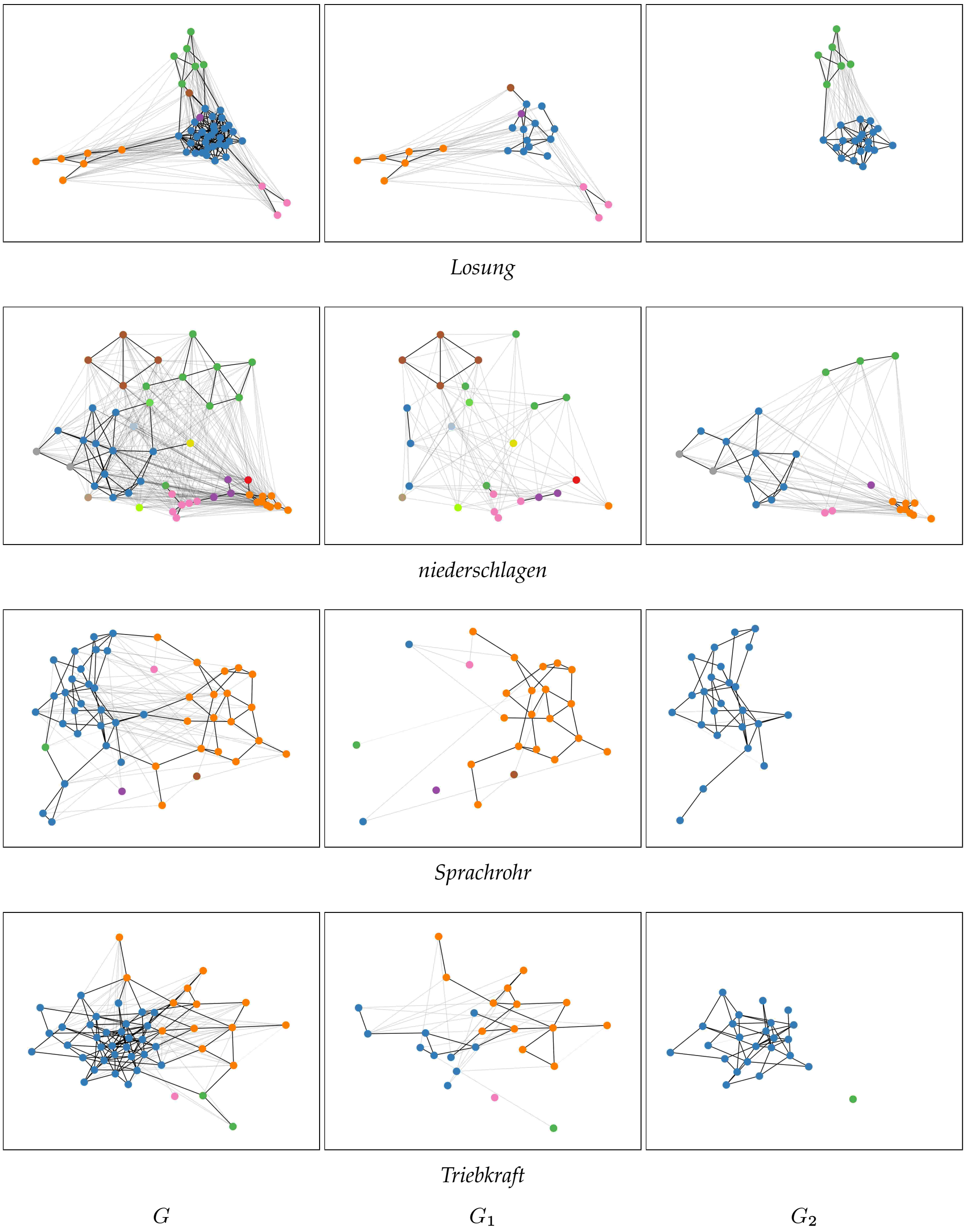


Figure A.10: WUGs from DiscoWUG. Full graph G (left), subgraphs for 1st time period G_1 (middle) and 2nd time period G_2 (right).



Losung

niederschlagen

Sprachrohr

Triebkraft

G G_1 G_2

Figure A.11: WUGs from DiscoWUG. Full graph G (left), subgraphs for 1st time period G_1 (middle) and 2nd time period G_2 (right).

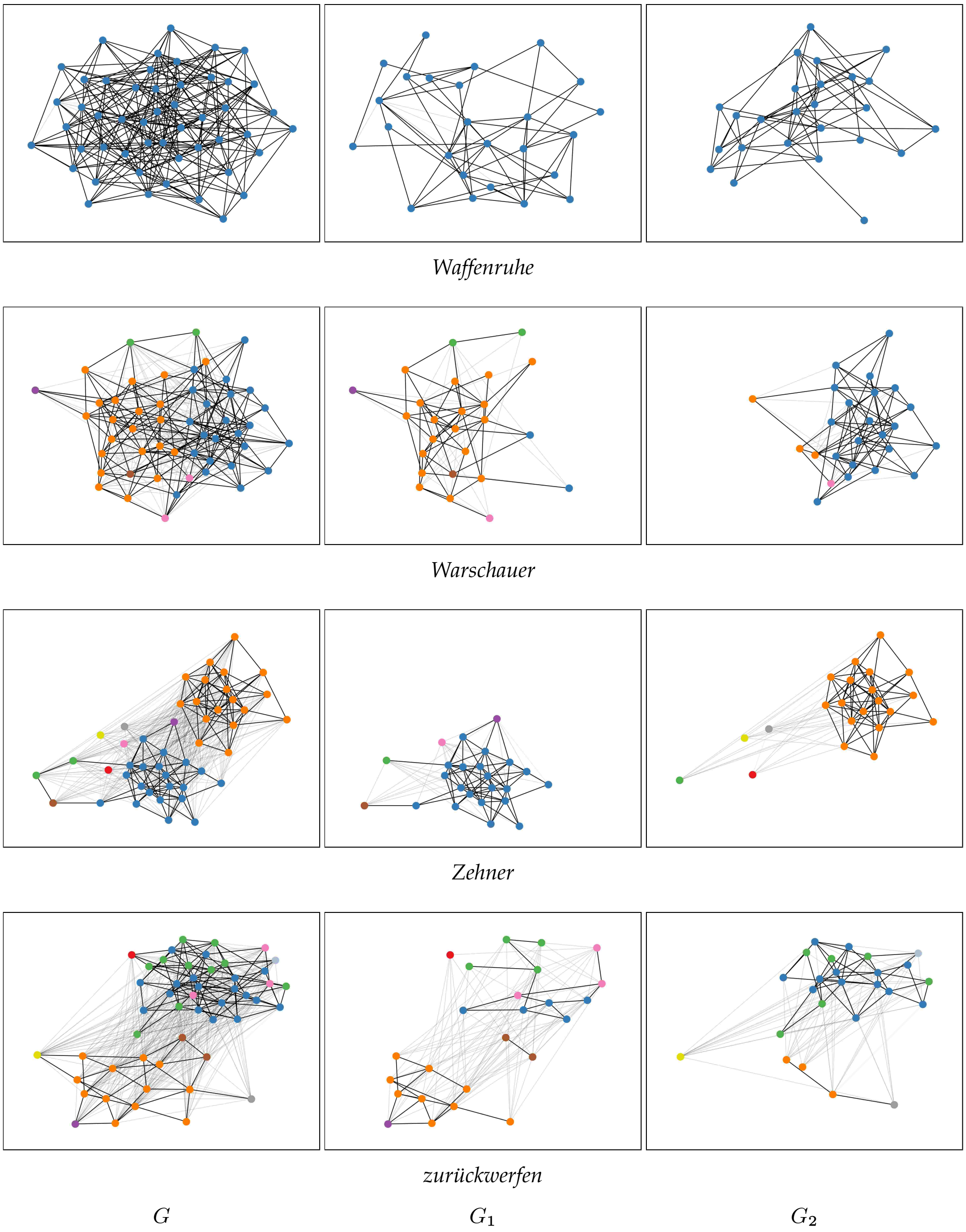
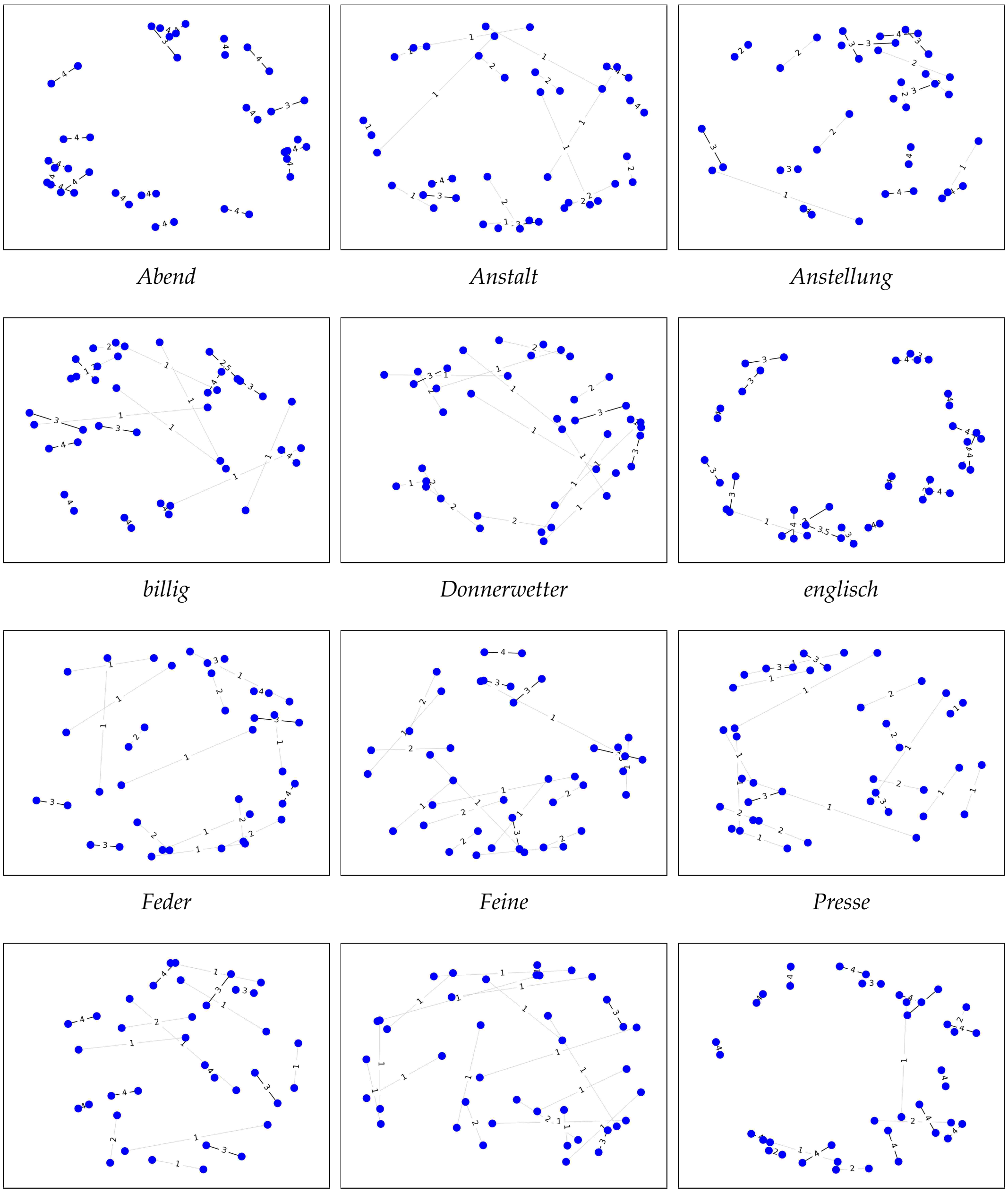


Figure A.12: WUGs from DiscoWUG. Full graph G (left), subgraphs for 1st time period G_1 (middle) and 2nd time period G_2 (right).



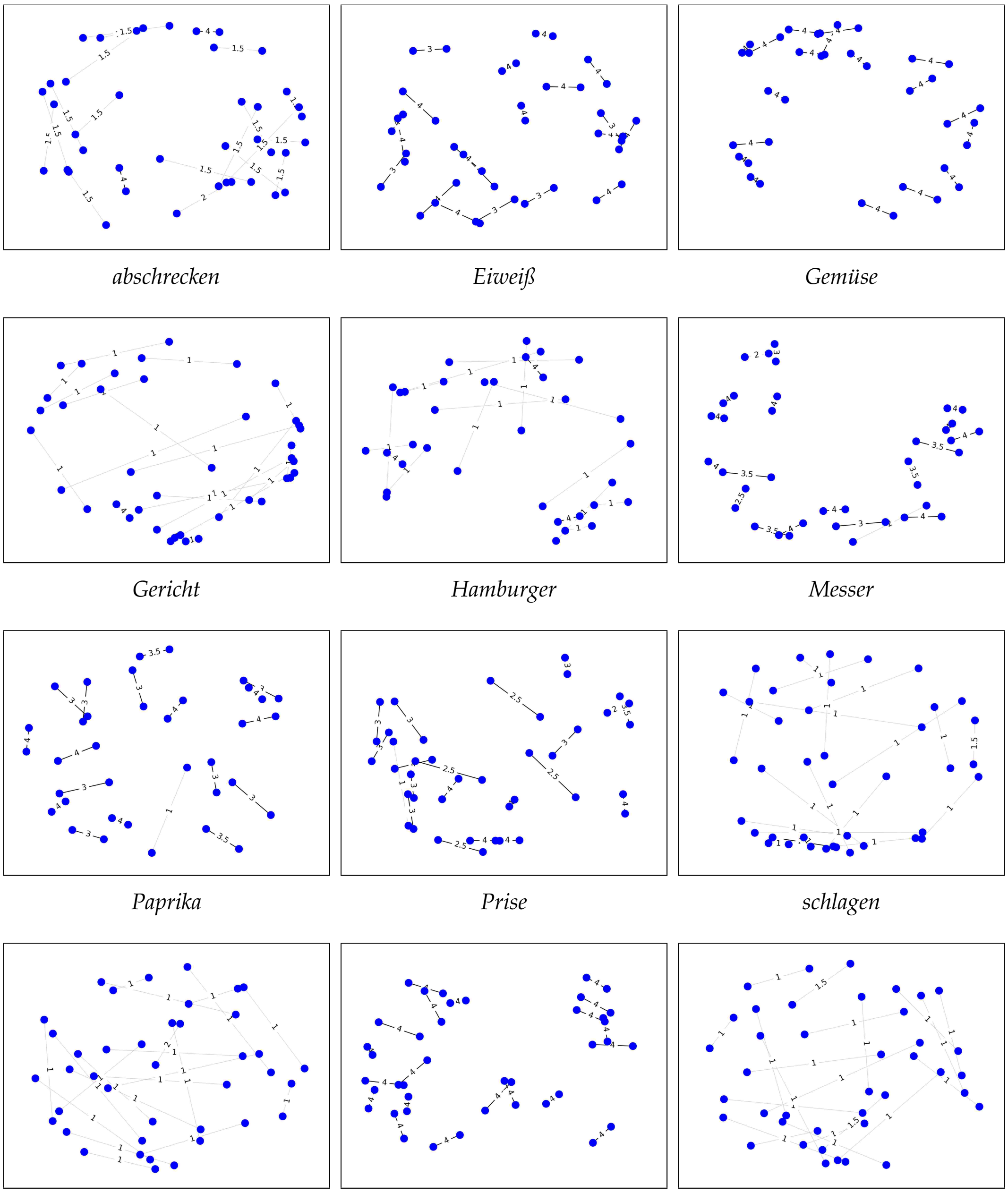
Abend *Anstalt* *Anstellung*

billig *Donnerwetter* *englisch*

Feder *Feine* *Presse*

Steckenpferd *Vorwort* *Zufall*

Figure A.13: WUGs from DURel showing the COMPARE subgraph $G_{1,2}$.



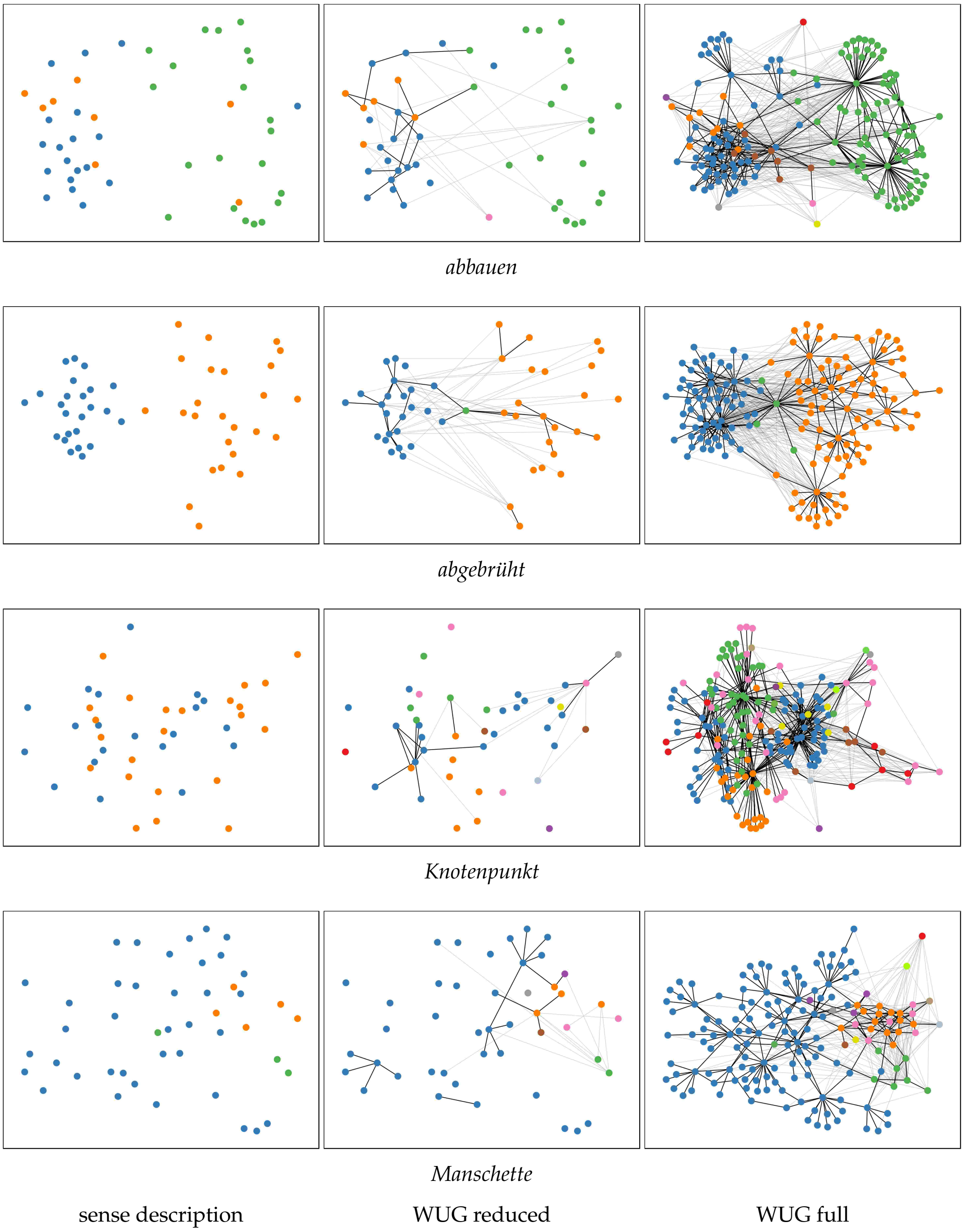
abschrecken *Eiweiß* *Gemüse*

Gericht *Hamburger* *Messer*

Paprika *Prise* *schlagen*

Schnee *Schnittlauch* *Strudel*

Figure A.14: WUGs from SURel showing the COMPARE subgraph $G_{1,2}$.



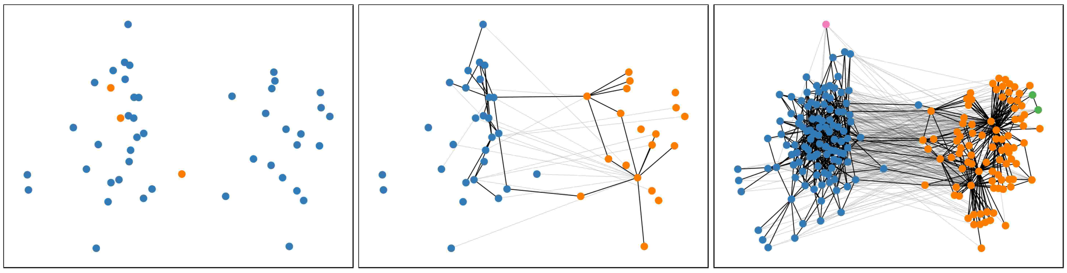
abbauen

abgebrüht

Knotenpunkt

Manschette

Figure A.15: Cluster comparison for words from SemEval DE.



zersetzen

sense description WUG reduced WUG full

Figure A.16: Cluster comparison for words from SemEval DE.

Appendix B

Formal Relations between Change Measures

The change measures described in Section 3.1.5 are mathematically related, e.g., they coincide for certain special cases of sense probability distributions (normalized sense frequency distributions). In this chapter, we show some of these relations under simplifying assumptions. We focus on relations between the measures' maximum and minimum values. Specifically, we show that:

Lemma 1 If $G(P,Q) = 1$, then $B(P,Q) = 1$.

Lemma 2 If $G(P,Q) = 0$, then $B(P,Q) = 0$.

Lemma 3 $G(P,Q) = 1$ iff $C(P,Q) = -1$.

Lemma 4 If $C(P,Q) = -4$, then $G(P,Q) = 0$.

Lemma 5 If $C(P,Q) = -4$, then $B(P,Q) = 0$.

Lemma 6 If $C(P,Q) = -1$, then $B(P,Q) = 1$.

Definitions and Assumptions

Assume that $\mathbf{G} = (\mathbf{U}, \mathbf{E}, \mathbf{W})$ is a Word Usage Graph (see Section 3.1.3) of word w containing w's uses U from two time periods. D and E are the time-specific sense frequency distributions (see Section 3.1.4) of length K, obtained by clustering the uses in U based on the edge weights in W, and P and Q the corresponding sense probability distributions, obtained by dividing D and E by their respective total frequencies. Note that P and Q are probability distributions, i.e., for each p_i, q_i, $0 \leq p_i, q_i$ and $\sum_i^K p_i = 1, \sum_i^K q_i = 1$.[1] Further, assume that $\mathbf{G_{1,2}} = (\mathbf{U}, \mathbf{E_{1,2}}, \mathbf{W_{1,2}})$ is the subgraph of G containing all uses U from both time periods, but only those edges (u_1, u_2) and their weights where word uses u_1 and u_2 are from different time periods (COMPARE edges, see Section 3.1.3).

Binary change $B(w)$ was defined (see Section 3.1.5) as

$$\begin{aligned} &B(D,E) = 1 \text{ if for some } i,\ D_i \leq k \text{ and } E_i \geq n, \\ &\qquad\qquad \text{or vice versa.} \\ &B(D,E) = 0 \text{ else.} \end{aligned}$$

We assume that $k = 0$ and $n = 0$. Hence, $B(D,E) = B(P,Q)$ becomes

$$\begin{aligned} &B(P,Q) = 1 \text{ if for some } i,\ P_i = 0 \text{ and } Q_i > 0, \\ &\qquad\qquad \text{or vice versa.} \\ &B(P,Q) = 0 \text{ else.} \end{aligned}$$

Graded change $G(w)$ was defined as the Jensen-Shannon Distance between P and Q:

$$G(P,Q) = JSD(P,Q)\ .$$

The JSD has two alternative formulations (Lin, 1991):

$$JSD(P,Q) = \sqrt{H(M) - \frac{1}{2}(H(P) + H(Q))}\ ,$$

$$JSD(P,Q) = \sqrt{\frac{KLD(P \parallel M) + KLD(Q \parallel M)}{2}}$$

where

$$H(P) = -\sum_i^K p_i \log_2(p_i),$$

[1] We allow for zero-probability events.

$$M = \frac{P+Q}{2},$$

$$KLD(P \parallel Q) = \sum_{i}^{K} p_i \log_2\left(\frac{p_i}{q_i}\right) .$$

The **Negated COMPARE** score $C(w)$ was defined as the negated mean of the COMPARE weights $W_{1,2}$:

$$C(W_{1,2}) = -\frac{1}{|W_{1,2}|} \sum_{x \in W_{1,2}} x .$$

Now, assume that $R : set(W_{1,2}) \mapsto [0, 1]$ is the probability distribution over the COMPARE weights $W_{1,2}$ mapping each possible unique edge weight to its probability of occurrence given by its normalized occurrence frequency in $W_{1,2}$. Now, note that the negated mean (expectation) of this distribution is equivalent to the mean of COMPARE weights $W_{1,2}$:

$$C(R) = C(W_{1,2}) = -\sum_{x \in set(W_{1,2})} xR(x) .$$

Now, assume that edge weights between uses of different senses (clusters) in P and Q are constantly 1 while weights between uses of the same sense are constantly 4. This implies that the distribution of the COMPARE weights is defined by $R(4) = \sum_i^K p_i q_i$, $R(1) = 1 - \sum_i^K p_i q_i$ (cf. Arefyev and Bykov, 2021). Now, $C(R)$ can be expressed as the negated sum of probabilities of sampling the same sense or different senses of a word from two time-periods multiplied by the corresponding weights (4 or 1):

$$C(P, Q) = -\big(4 * \sum_{i}^{K} p_i q_i + 1 * \big(1 - \sum_{i}^{K} p_i q_i\big)\big) = -\big(3 * \sum_{i}^{K} p_i q_i + 1\big) .$$

Proofs

We now give proofs of the Lemmas under the assumptions mentioned above.

Lemma 1 *If* $JSD(P, Q) = 1$*, then* $B(P, Q) = 1$.

Proof Assume that $JSD(P,Q) = 1$. From Lemma 7, it follows that for every i either $p_i > 0$ and $q_i = 0$, or $q_i > 0$ and $p_i = 0$, or $p_i = q_i = 0$. Also, there must be at least one i, j such that $p_i > 0$ or $q_j > 0$ as otherwise it would not hold that $\sum_i^K p_i = 1, \sum_i^K q_i = 1$.

■

Lemma 2 *If $JSD(P,Q) = 0$, then $B(P,Q) = 0$.*

Proof Assume that $JSD(P,Q) = 0$. Then, by Lemma 10, it holds that $P = Q$. This directly implies that $B(P,Q) = 0$.

■

Lemma 3 *$JSD(P,Q) = 1$ iff $C(P,Q) = -1$.*

Proof Assume that $JSD(P,Q) = 1$. From Lemma 7, it follows that for every i either $p_i > 0$ and $q_i = 0$, or $q_i > 0$ and $p_i = 0$, or $p_i = q_i = 0$. Hence, it holds that for every i, $p_i q_i = 0$. This implies that $C(P,Q) = -\big(3\sum_i^K p_i q_i + 1\big) = -1$.

Now, assume that $C(P,Q) = -\big(3\sum_i p_i q_i + 1\big) = -1$. Then,

$$-\big(3\sum_i^K p_i q_i + 1\big) = -1$$

$$3\sum_i^K p_i q_i + 1 = 1$$

$$3\sum_i^K p_i q_i = 0$$

$$\sum_i^K p_i q_i = 0$$

This means that for every i either $p_i > 0$ and $q_i = 0$, or $q_i > 0$ and $p_i = 0$, or $p_i = q_i = 0$. From Lemma, 7 it follows that $JSD(P,Q) = 1$.

■

Lemma 4 *If $C(P,Q) = -4$, then $JSD(P,Q) = 0$.*

Proof Assume that $C(P, Q) = -4$. Then,

$$-\big(3\sum_{i}^{K} p_i q_i + 1\big) = -4$$

$$3\sum_{i}^{K} p_i q_i + 1 = 4$$

$$3\sum_{i}^{K} p_i q_i = 3$$

$$\sum_{i}^{K} p_i q_i = p_1 q_1 + \ldots p_i q_i \ldots + p_K q_K = 1 \tag{B.1}$$

Note that for each i, it holds that $p_i q_i \leq q_i$ and $p_i q_i \leq p_i$. Now, assume that for some i, $0 < p_i < 1$. From this, it follows that $\sum_i^K p_i q_i < 1$. This contradicts Equation B.1. Similarly for $0 < q_i < 1$. Hence, it follows that for each i, $p_i = 0$ or $p_i = 1$, and $q_i = 0$ or $q_i = 1$. Now, assume that for some i, $p_i \neq q_i$, i.e., either $p_i = 0$ and $q_i = 1$ or vice versa. It again follows that $\sum_i^K p_i q_i < 1$. This again contradicts Equation B.1. It follows that $p_i = q_i$ for every i. This means that $P = Q$. From Lemma 10, it follows that $JSD(P, Q) = 0$.

■

Lemma 5 *If $C(P, Q) = -4$, then $B(P, Q) = 0$.*

Proof Assume that $C(P, Q) = -4$. From Lemma 4, it follows that $JSD(P, Q) = 0$. From Lemma 2, it follows that $B(P, Q) = 0$.

■

Lemma 6 *If $C(P, Q) = -1$, then $B(P, Q) = 1$.*

Proof Assume that $C(P, Q) = -1$. From Lemma 3, it follows that $JSD(P, Q) = 1$. From Lemma 1, it follows that $B(P, Q) = 1$.

■

Lemma 7 *$JSD(P, Q) = 1$ iff for every i either $p_i > 0$ and $q_i = 0$, or $q_i > 0$ and $p_i = 0$, or $p_i = q_i = 0$.*

Proof Assume that $JSD(P,Q) = 1 = \sqrt{H(M) - \frac{1}{2}(H(P) + H(Q))}$. Then,

$$\sqrt{1 - \frac{1}{2}\log_2 \prod_i^K \frac{(p_i + q_i)^{p_i+q_i}}{p_i^{p_i} q_i^{q_i}}} = 1 \qquad \text{(Lemma 8)}$$

$$1 - \frac{1}{2}\log_2 \prod_i^K \frac{(p_i + q_i)^{p_i+q_i}}{p_i^{p_i} q_i^{q_i}} = 1$$

$$-\frac{1}{2}\log_2 \prod_i^K \frac{(p_i + q_i)^{p_i+q_i}}{p_i^{p_i} q_i^{q_i}} = 0$$

$$-\log_2 \prod_i^K \frac{(p_i + q_i)^{p_i+q_i}}{p_i^{p_i} q_i^{q_i}} = 0$$

$$\prod_i^K \frac{(p_i + q_i)^{p_i+q_i}}{p_i^{p_i} q_i^{q_i}} = 1$$

$$\frac{\prod_i^K (p_i + q_i)^{p_i+q_i}}{\prod_i^K p_i^{p_i} q_i^{q_i}} = 1$$

$$\prod_i^K (p_i + q_i)^{p_i+q_i} = \prod_i^K p_i^{p_i} q_i^{q_i}$$

$$\prod_i^K (p_i + q_i)^{p_i+q_i} = \prod_i^K p_i^{p_i} \prod_i^K q_i^{q_i}$$

$$\prod_i^K (p_i + q_i)^{p_i} \prod_i^K (p_i + q_i)^{q_i} = \prod_i^K p_i^{p_i} \prod_i^K q_i^{q_i} \qquad \text{(B.2)}$$

First, note that for each p_i, q_i, it holds that $(p_i + q_i)^{p_i} \geq p_i^{p_i}$ and $(p_i + q_i)^{q_i} \geq q_i^{q_i}$ as $0 \leq p_i, q_i \leq 1$ implies $p_i + q_i \geq p_i$, $p_i + q_i \geq q_i$. Now, assume that $p_i, q_i \neq 0$ for some i. It holds that $(p_i + q_i)^{p_i} > p_i^{p_i}$ and $(p_i + q_i)^{q_i} > q_i^{q_i}$. From this, it follows that $\prod_i^K (p_i + q_i)^{p_i} > \prod_i^K p_i^{p_i}$ and $\prod_i^K (p_i + q_i)^{q_i} > \prod_i^K q_i^{q_i}$. Finally, this implies $\prod_i^K (p_i + q_i)^{p_i} \prod_i^K (p_i + q_i)^{q_i} > \prod_i^K p_i^{p_i} \prod_i^K q_i^{q_i}$. This contradicts Equation B.2. Hence, for every i either $p_i > 0$ and $q_i = 0$, or $q_i > 0$ and $p_i = 0$, or $p_i = q_i = 0$.

Now, assume that for every i, either $p_i > 0$ and $q_i = 0$, or $q_i > 0$ and $p_i = 0$, or

$p_i = q_i = 0$. We know that

$$JSD(P,Q) = \sqrt{1 - \frac{1}{2}\log_2 \prod_i^K \frac{(p_i+q_i)^{p_i+q_i}}{p_i^{p_i} q_i^{q_i}}} \quad \text{(Lemma 8)}$$

We consider all three cases: (i) Assume that $p_i > 0$ and $q_i = 0$. This means that the term $\frac{(p_i+q_i)^{p_i+q_i}}{p_i^{p_i} q_i^{q_i}}$ reduces to $\frac{p_i^{p_i}}{p_i^{p_i}} = 1$. Similarly for (ii), assuming that $q_i > 0$ and $p_i = 0$. And, in case (iii) with $p_i = q_i = 0$, $\frac{0^0}{0^0} = 1$. Hence,

$$JSD(P,Q) = \sqrt{1 - \frac{1}{2}\log_2(1)} = \sqrt{1} = 1$$

■

Lemma 8 $JSD(P,Q) = \sqrt{1 - \frac{1}{2}\log_2 \prod_i^K \frac{(p_i+q_i)^{p_i+q_i}}{p_i^{p_i} q_i^{q_i}}}$.

Proof

$$JSD(P,Q) = \sqrt{H(M) - \frac{1}{2}(H(P) + H(Q))}$$

$$= \sqrt{1 - \frac{1}{2}\sum_i^K (p_i+q_i)\log_2(p_i+q_i) - \frac{1}{2}(H(P)+H(Q))} \quad \text{(Lemma 9)}$$

$$= \sqrt{\frac{1}{2}\Big(2 - \sum_i^K (p_i+q_i)\log_2(p_i+q_i) - (H(P)+H(Q))\Big)}$$

$$= \sqrt{\frac{1}{2}\Big(2 - \sum_i^K (p_i+q_i)\log_2(p_i+q_i) - \Big(-\sum_i^K p_i\log_2(p_i) - \sum_i^K q_i\log_2(q_i)\Big)\Big)}$$

$$= \sqrt{\frac{1}{2}\Big(2 - \sum_i^K (p_i+q_i)\log_2(p_i+q_i) - \Big(-\sum_i^K p_i\log_2(p_i) + q_i\log_2(q_i)\Big)\Big)}$$

$$= \sqrt{\frac{1}{2}\Big(2 - \sum_i^K (p_i+q_i)\log_2(p_i+q_i) + \sum_i^K p_i\log_2(p_i) + q_i\log_2(q_i)\Big)}$$

$$= \sqrt{\frac{1}{2}\Big(2 - \Big(\sum_i^K (p_i+q_i)\log_2(p_i+q_i) - \sum_i^K p_i\log_2(p_i) + q_i\log_2(q_i)\Big)\Big)}$$

$$= \sqrt{\frac{1}{2}\left(2 - \sum_{i}^{K}(p_i + q_i)\log_2(p_i + q_i) - p_i \log_2(p_i) - q_i \log_2(q_i)\right)}$$

$$= \sqrt{\frac{1}{2}\left(2 - \sum_{i}^{K}\log_2((p_i + q_i)^{p_i+q_i}) - \log_2(p_i^{p_i}) - \log_2(q_i^{q_i})\right)}$$

$$= \sqrt{\frac{1}{2}\left(2 - \sum_{i}^{K}\log_2((p_i + q_i)^{p_i+q_i}) - \log_2(p_i^{p_i} q_i^{q_i})\right)}$$

$$= \sqrt{\frac{1}{2}\left(2 - \sum_{i}^{K}\log_2 \frac{(p_i + q_i)^{p_i+q_i}}{p_i^{p_i} q_i^{q_i}}\right)}$$

$$= \sqrt{\frac{1}{2}\left(2 - \log_2 \prod_{i}^{K} \frac{(p_i + q_i)^{p_i+q_i}}{p_i^{p_i} q_i^{q_i}}\right)}$$

$$= \sqrt{1 - \frac{1}{2}\log_2 \prod_{i}^{K} \frac{(p_i + q_i)^{p_i+q_i}}{p_i^{p_i} q_i^{q_i}}}$$

■

Lemma 9 $H(M) = 1 - \frac{1}{2}\sum_{i}^{K}(p_i + q_i)\log_2(p_i + q_i)$.

Proof

$$H(M) = -\sum_{i}^{K} \frac{p_i + q_i}{2} \log_2 \frac{p_i + q_i}{2}$$

$$= -\sum_{i}^{K} \frac{p_i + q_i}{2}\big(\log_2 \frac{1}{2} + \log_2(p_i + q_i)\big)$$

$$= -\frac{1}{2}\sum_{i}^{K}(p_i + q_i)\big(\log_2 \frac{1}{2} + \log_2(p_i + q_i)\big)$$

$$= -\frac{1}{2}\sum_{i}^{K}(p_i + q_i)(-1 + \log_2(p_i + q_i))$$

$$= -\frac{1}{2}\sum_{i}^{K} -(p_i + q_i) + (p_i + q_i)\log_2(p_i + q_i)$$

$$= -\frac{1}{2}\sum_{i}^{K} -(p_i + q_i) - \frac{1}{2}\sum_{i}^{K}(p_i + q_i)\log_2(p_i + q_i)$$

$$= \frac{1}{2}\sum_{i}^{K} p_i + q_i - \frac{1}{2}\sum_{i}^{K}(p_i + q_i)\log_2(p_i + q_i)$$

$$= 1 - \frac{1}{2}\sum_{i}^{K}(p_i + q_i)\log_2(p_i + q_i)$$

■

Lemma 10 $JSD(P, Q) = 0$ *iff* $P = Q$.

Proof Assume that $JSD(P, Q) = 0 = \sqrt{\frac{KLD(P\|M)+KLD(Q\|M)}{2}}$. Then,

$$\frac{1}{2}\big(KLD(P \parallel M) + KLD(Q \parallel M)\big) = 0$$

$$KLD(P \parallel M) + KLD(Q \parallel M) = 0$$

From the first part of Gibbs' inequality (Lemma 11), it follows that $KLD(P \parallel M) = KLD(Q \parallel M) = 0$. The second part of Gibbs' inequality implies that $P = M$ and $Q = M$. From this, it follows that $P = Q$.

Now, assume that $P = Q$. It follows that $P = Q = M$. Hence, $KLD(P \parallel M) = KLD(Q \parallel M) = 0$, which means that $JSD(P, Q) = 0$. ■

Lemma 11 (Gibbs' inequality) $KLD(P \parallel Q) \geq 0$, *where* $KLD(P \parallel Q) = 0$ *iff* $P = Q$.

Proof Find a proof in Cover and Thomas (2006, p. 28).

Appendix C

Annotation of SemEval LA

In this chapter, we describe the procedure devised to annotate the Latin SemEval data. This procedure is different from the other languages as in a trial annotation task the annotators reported difficulties to judge use-use pairs. In consideration of this, annotators were asked to judge use-sense pairs.

Use-Sense Annotation

Following Erk et al. (2013), semantic proximity for use-sense pairs can be measured by human annotator judgments on a similar scale as for use-use pairs. Hence, we ask annotators to judge the semantic relatedness of use-sense pairs using the same scale as for use-use pairs (see Table 3.1 in Chapter 3). (C.1) contains an example of a use-sense pair for *sacramentum*, displaying the older sense 'a civil suit or process'.

(C.1) `Use`: Cum Arretinae mulieris libertatem defenderem et Cotta xviris religionem iniecisset non posse nostrum **sacramentum** iustum iudicari, [...] *'When I was defending the liberty of a woman of Arretium, and when Cotta had suggested a scruple to the decemvirs that our* ***action*** *was not a regular one, [...] '* [1]
`Sense`: 'a cause, a civil suit or process'

Graph Representation

We represent annotated data (semantic proximity judgments of use-sense pairs) in a graph which we call Use-Sense Graph (USG). A USG $\mathbf{G} = (\mathbf{V}, \mathbf{E}, \mathbf{W})$ is a weighted,

[1] M. Tullius Cicero. The Orations of Marcus Tullius Cicero, literally translated by C. D. Yonge, B. A. London. Henry G. Bohn, York Street, Covent Garden. 1856.

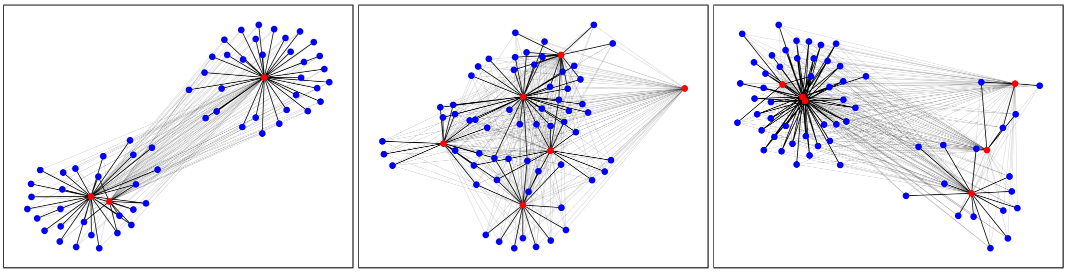


Figure C.1: USG of Latin *pontifex* (left), *potestas* (middle) and *sacramentum* (right). Nodes in blue/red represent uses/senses respectively.

undirected graph whose nodes $v \in V$ represent either word uses or sense descriptions and weights $w \in W$ represent the semantic proximity of a use-sense pair $(u, s) \in E$ (cf. Section 3.1.3).[2] We denote the set of word uses as U and the set of word sense descriptions as S, where $V = U \cup S$.

Figure C.1 shows three USGs resulting from our annotation. The first word, *pontifex*, originally meant 'a member of the college of priests having supreme control in matters of public religion in Rome', and with Christianity it acquired the sense of 'bishop'. The three senses presented to the annotators were 'priest, high priest', 'Roman high-priest, a pontiff, pontifex', and 'bishop'. The first two correspond to the two red nodes in the bottom left corner of the first plot in Figure C.1, and the last one corresponds to the top right red node. The plot of the second word, *potestas* shows a complex and highly related set of senses, which can be summarised as: 'power of doing any thing', 'political power', 'magisterial power', 'meaning of a word' (the isolated sense on the far right of the plot), 'force, efficacy' and 'angelic powers'. The last plot refers to *sacramentum* and shows how the two senses 'military oath of allegiance' and 'oath' are closely together on the top left of the plot while the legal sense 'a civil suit or process' is separated from the others in the top right corner and the Christian sense of 'sacrament' is at the bottom right corner.

Clustering

From the annotation, we obtain USGs where each use is connected to each sense by one edge (see Figure C.1). Therefore, there is a first-order path between each use-

[2] Note that we do not consider the possible cases where E contains additional use-use pairs or sense-sense pairs.

	C_1					C_2				
	corpus	**period**	**tokens**	**types**	**TTR**	**corpus**	**period**	**tokens**	**types**	**TTR**
Latin	LatinISE	-200–0	1.7M	65k	38.24	LatinISE	0–2000	9.4M	253k	26.91

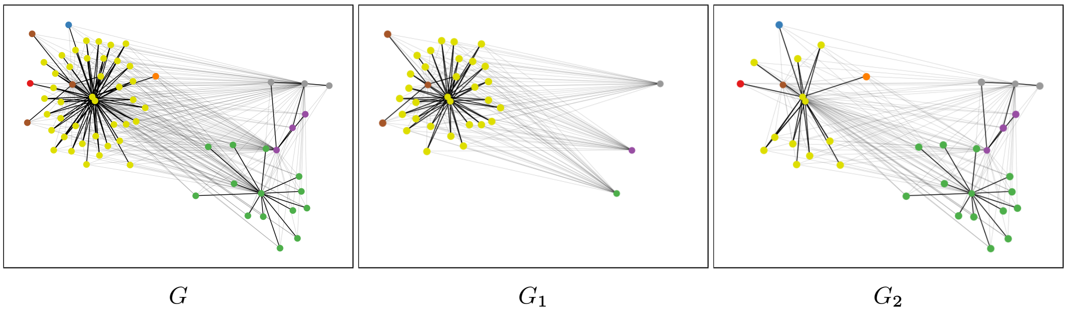
Table C.1: Statistics of SemEval LA corpora. TTR = Type-Token ratio (number of types / number of tokens * 1000).

sense pair and a second-order path between each use-use pair. Similarly to WUGs derived from use-use judgments described in Section 3.1.2, we want to assign uses and senses to the same cluster if they receive high judgments (3, 4) and to different clusters if they receive low judgments (1, 2). For this, we use the same clustering algorithm as for WUGs, defined in Section 3.1.4, to cluster uses and senses in V at the same time. In this way, uses end up in the same cluster if they have high judgments with the same senses. If there are contradictory judgments (e.g. a use has high judgments with several senses), the clustering uses the global information to decide on the cluster assignment by choosing the clustering with the lowest loss. This can also lead to the collapsing of two sense descriptions into one cluster, e.g. for Latin *sacramentum* in Figure C.2.

Data

Corpora

The corpora are created in a similar procedure as the one described in Section 3.2.1.1. We use the lemmatized and POS tagged LatinISE corpus (McGillivray and Kilgarriff, 2013), spanning from the 2nd century B.C. to the 21st century A.D. A study on lemmatization accuracy on a sample of two texts (Cicero's *De Officiis* and Rutilius Taurus Aemilianus Palladius' *Opus agriculturae*) against the PROIEL treebank (Haug and Jøhndal, 2008) as a gold standard showed an accuracy of 92.77% and 80.96%, respectively. We then extract two time-specific subcorpora C_1, C_2, as defined in Table C.1. From these two subcorpora, we then sample the released test corpora in the following way: Sentences with < 2 tokens are removed, tokens are replaced by their lemma, punctuation is removed and sentences are randomly shuffled within each of C_1, C_2. We also create a tokenized version of the corpora with sentences in the same order as in the lemmatized version. Find a summary of



G G_1 G_2

Figure C.2: USG of Latin *sacramentum* (left), subgraphs for 1st time period G_1 (middle) and 2nd time period G_2 (right). $D_1 = (28, 0, 2, 0, 0, 0, 0, 0, 0)$, $D_2 = (10, 13, 0, 2, 2, 1, 1, 1, 1)$, $B(w) = 1$ and $G(w) = 0.69$.

the released (lemmatized) test corpora in Table C.1.[3]

Target Words

We select a range of target words whose meaning has changed between the pre-Christian and the Christian era according to the literature (Clackson, 2011) and in the pre-annotation trial we check that these meanings are present in the corpus data. For each changed word, we select a control word whose meaning did not change from the pre-Christian era and the Christian era, whose POS is the same as the changed word, and whose frequency development between C_1 and C_2 is similar to the changed word.

Annotators

Since we do not have access to native speakers of Latin, eight annotators with a high-level knowledge of Latin are recruited, ranging from undergraduate students to PhD students, post-doctoral researchers, and more senior researchers.

Use and Sense Sampling

For each target word, 30 uses from each of the tokenized versions of C_1 and C_2 are randomly sampled, yielding a total of 60 uses per target word. The sense defini-

[3] Find the extracted corpora at `https://www.ims.uni-stuttgart.de/data/sem-eval-ulscd`.

	General					Binary				Graded			
	n	N/V/A	SPR	KRI	$\lvert U \rvert$	LSC	FRQ$_\text{d}$	FRQ$_\text{m}$	PLY$_\text{m}$	LSC	FRQ$_\text{d}$	FRQ$_\text{m}$	PLY$_\text{m}$
Latin	40	27/5/8	.64	.62	59	.65	.16	.02	.14	.33	.39	-.13	.31

Table C.2: Overview SemEval LA target words. n = number of target words, N/V/A = number of nouns/verbs/adjectives, SPR = weighted mean of pairwise Spearman on *virtus*, KRI = Krippendorff's alpha on *virtus*, $|U|$ = avg. no. uses per word (after cleaning), LSC = mean binary/graded change score, FRQ$_\text{d}$ = Spearman correlation between change scores and target words' absolute difference in log-frequency between C_1, C_2. Similarly for minimum frequency (FRQ$_\text{m}$) and minimum number of senses (PLY$_\text{m}$) across C_1, C_2.

tions are taken from the Latin portion of the Logeion online dictionary.[4] Due to the challenge of finding qualified annotators, each word is assigned to one annotator, apart from *virtus*, which is annotated by four annotators and used for calculation of inter-annotator agreement (see Table C.2). The senses and uses are presented in randomized order to the annotators.

Edge Sampling

The USG annotation procedure has an upper bound on the total number of annotated use-sense pairs of $n \times k$, with k senses for n uses. The number of senses ranges between 2 and 7 with a use sample size of 60 (30 + 30), which yields a feasible number of annotation instances. Hence, no further optimization of the edge sampling procedure is carried out. Note though that a similar optimization as in Section 3.2.1.5 would be possible by annotating the data incrementally or by randomly subsampling edges.

Summary

Find a summary of the annotation outcome in Table C.2. The final test set contains 40 target words. The inter-annotator agreement is comparable to the ones observed in Section 3.2.1.6. Figure C.2 shows the annotated and clustered USG for Latin target *sacramentum* from Figure C.1 along with the two time-specific subgraphs for C_1 and C_2. The process of deriving change scores is identical to the one for WUGs as described in Section 3.1.4: For each target word, we obtain the two time-specific sense frequency distributions D_1 and D_2 from the full clustering. From these, we

[4] `https://logeion.uchicago.edu/`

infer the binary and the graded change score, setting the lower frequency thresholds to $k = 0$, $n = 1$ (see Section 3.1.5).

Appendix D

Annotation Simulation

We validate the annotation procedure and the clustering algorithm described in Chapter 3 in a simulation study by simulating 40 ground truth WUGs with zipfian sense frequency distributions covering roughly the frequency range of the majority of SemEval target words (50–1000). We first simulate the zipfian sense frequency distributions and then introduce change to half of the target words by setting some of its senses' frequencies to 0 in either of D_1, D_2. We then simulate fully-connected (true) graphs sampling weights between clusters uniformly from $\{1, 2\}$ and weights within clusters uniformly from $\{3, 4\}$. Then we sample edge weights from these graphs in several rounds as described in Section 3.2.1.5, simulate human judgments in each round by adding a normally distributed error to sampled edge weights and compare the resulting clustering to the clustering of the true graph. The true clustering can be recovered with high accuracy (average of $> .96$ ARI). We also use the simulation to predict the feasibility of the study and to tune parameters of the annotation such as sample sizes for nodes and edges. With the finally chosen parameters described in Section 3.2.1.5, the algorithm converges on average after 5 rounds and $\approx$ 8000 judgments per annotator. This was within the bounds of our time limits and financial budget.

We also test the clustering algorithm against several standard techniques (Biemann, 2006; Blondel et al., 2008) and vary the loss optimization algorithm. None of these variations had a comparable performance to our approach.

Appendix E

Cluster Bias

In this chapter, we report more detailed results on the clustering bias experiments described in Section 5.3. Please find the full results in Tables E.1–E.3.

		Layer	Token	Lemma	TokLem
Performance	Graded	1	-.141	-.033	.100
		12	.205	.154	.168
		1+12	-.316	.130	.081
		6+7	.075	-.103	.017
		9-12	.325	**.345**	.293
	Cluster	1	.022	.041	.045
		12	.116	.111	.158
		1+12	.022	.141	.149
		6+7	.119	.111	.145
		9-12	.150	.159	**.163**

		Layer	Token	Lemma	TokLem
Performance	Graded	1	-.265	-.062	-.170
		12	.123	.427	**.624**
		1+12	-.252	.235	.401
		6+7	.002	.464	.320
		9-12	.122	.420	.533
	Cluster	1	.033	.002	.003
		12	.119	.159	**.161**
		1+12	.037	.064	.080
		6+7	.101	.158	.152
		9-12	.155	.142	.154

Table E.1: Clustering performances on SemEval EN (left) and DE (right).

		Layer	Token	Lemma	TokLem
Form	Influence	1	.907	.014	.014
		12	.389	.018	.077
		1+12	.881	.020	.057
		6+7	.572	.013	.028
		9-12	.334	.018	.051
	Random	1	.002	.002	.002
		12	-.001	.001	-.001
		1+12	-.002	-.001	-.001
		6+7	.001	.002	.001
		9-12	-.001	-.001	-.002
	Gold	1	.017	.017	.017
		12	.017	.017	.017
		1+12	.017	.017	.017
		6+7	.017	.017	.017
		9-12	.017	.017	.017
Position	Influence	1	.001	.026	.024
		12	.012	.012	.015
		1+12	-.001	.019	.007
		6+7	-.002	.018	-.003
		9-12	.002	.007	.003
	Random	1	.001	.003	.001
		12	.001	-.001	-.001
		1+12	-.001	-.001	-.001
		6+7	.001	-.001	-.001
		9-12	.001	.001	-.001
	Gold	1	-.002	-.002	-.002
		12	-.002	-.002	-.002
		1+12	-.002	-.002	-.002
		6+7	-.002	-.002	-.002
		9-12	-.002	-.002	-.002

		Layer	Token	Lemma	TokLem
Form	Influence	1	.706	.024	.004
		12	.439	.056	.150
		1+12	.687	.039	.046
		6+7	.503	.050	.050
		9-12	.420	.047	.094
	Random	1	-.001	-.002	.020
		12	-.001	.001	.021
		1+12	-.001	-.001	.020
		6+7	.002	.001	.019
		9-12	.001	-.001	.021
	Gold	1	.036	.036	.036
		12	.036	.036	.036
		1+12	.036	.036	.036
		6+7	.036	.036	.036
		9-12	.036	.036	.036
Position	Influence	1	.005	.023	.027
		12	-.002	.005	-.002
		1+12	.002	.021	.013
		6+7	.010	.020	.018
		9-12	.009	.018	.012
	Random	1	.001	.001	.001
		12	.001	-.001	.001
		1+12	-.001	-.001	.002
		6+7	-.001	.001	.001
		9-12	-.001	.001	.001
	Gold	1	.005	.005	.005
		12	.005	.005	.005
		1+12	.005	.005	.005
		6+7	.005	.005	.005
		9-12	.005	.005	.005

Table E.2: Clustering influences of target word form and position on SemEval EN (left) and DE (right).

		Layer	Token	Lemma	TokLem
Corpora	Influence	1	.019	.021	.033
		12	.078	.056	.082
		1+12	.027	.050	.074
		6+7	.034	.035	.050
		9-12	.056	.044	.063
	Random	1	.001	-.001	.003
		12	.001	.001	.001
		1+12	-.001	.001	.001
		6+7	.001	.001	.002
		9-12	.002	.001	.002
	Gold	1	.018	.018	.018
		12	.018	.018	.018
		1+12	.018	.018	.018
		6+7	.018	.018	.018
		9-12	.018	.018	.018
Names	Influence	1	-.007	.010	.010
		12	.025	.027	.033
		1+12	.018	.022	.027
		6+7	.012	.016	.027
		9-12	.019	.022	.026
	Random	1	-.001	-.002	-.002
		12	-.001	.001	.001
		1+12	-.001	.001	-.001
		6+7	-.001	.001	.001
		9-12	-.001	-.001	.001
	Gold	1	.019	.019	.019
		12	.019	.019	.019
		1+12	.019	.019	.019
		6+7	.019	.019	.019
		9-12	.019	.019	.019

		Layer	Token	Lemma	TokLem
Corpora	Influence	1	.074	.003	.005
		12	.110	.095	.096
		1+12	.077	.024	.052
		6+7	.101	.058	.075
		9-12	.107	.068	.089
	Random	1	-.001	-.001	.001
		12	.001	-.001	.001
		1+12	-.001	.001	.002
		6+7	-.001	.001	-.001
		9-12	-.001	.001	-.001
	Gold	1	.083	.083	.083
		12	.083	.083	.083
		1+12	.083	.083	.083
		6+7	.083	.083	.083
		9-12	.083	.083	.083
Names	Influence	1	-	-	-
		12	-	-	-
		1+12	-	-	-
		6+7	-	-	-
		9-12	-	-	-
	Random	1	-	-	-
		12	-	-	-
		1+12	-	-	-
		6+7	-	-	-
		9-12	-	-	-
	Gold	1	-	-	-
		12	-	-	-
		1+12	-	-	-
		6+7	-	-	-
		9-12	-	-	-

Table E.3: Clustering influences of corpora and proper names on SemEval EN (left) and DE (right).

Bibliography

Abbe, E. (2017). Community detection and stochastic block models: recent developments. Cited on page 35.

Adesam, Y., Dannélls, D., and Tahmasebi, N. (2019). Exploring the quality of the digital historical newspaper archive KubHist. In *Proceedings of the 2019 DHN conference*, pages 9–17. Cited on page 43.

Aggarwal, C. C. and Reddy, C. K. (2013). *Data Clustering: Algorithms and Applications*. Chapman and Hall/CRC Data Mining and Knowledge Discovery Series. Taylor and Francis. Cited on pages 74 and 91.

Aksenova, A., Gavrishina, E., Rykov, E., and Kutuzov, A. (2022). Rudsi: graph-based word sense induction dataset for russian. Cited on pages 67 and 134.

Alatrash, R., Schlechtweg, D., Kuhn, J., and Schulte im Walde, S. (2020). CCOHA: Clean Corpus of Historical American English. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6958–6966, Marseille, France. European Language Resources Association. Cited on page 42.

Allamar, J. (2021). The Illustrated Transformer. `https://jalammar.github.io/illustrated-transformer/`. Accessed: 2021-02-26. Cited on page 71.

Amrami, A. and Goldberg, Y. (2018). Word sense induction with neural biLM and symmetric patterns. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4860–4867, Brussels, Belgium. Association for Computational Linguistics. Cited on page 74.

Arefyev, N. and Bykov, D. (2021). An interpretable approach to lexical semantic change detection with lexical substitution. volume 2021-June, pages 31–46. Cited on pages 39, 114, 134, and 157.

Arefyev, N., Fedoseev, M., Protasov, V., Homskiy, D., Davletov, A., and Panchenko, A. (2021). Deepmistake: Which senses are hard to distinguish for a word-in-context model. volume 2021-June, pages 16–30. Cited on pages 69, 70, 114, 134, and 135.

Arefyev, N. and Rachinskiy, M. (2021). Zero-shot cross-lingual transfer of a gloss language model for semantic change detection. volume 2021-June, pages 578–586. Cited on pages 114 and 134.

Arefyev, N. and Zhikov, V. (2020). BOS at SemEval-2020 task 1: Word sense induction via lexical substitution for lexical semantic change detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 171–179, Barcelona (online). International Committee for Computational Linguistics. Cited on pages 92 and 113.

Aristoteles (n.d.). De memoria et reminiscentia. In Sorabij, R., editor, *Aristotle on Memory, 1972*, pages 47–60. Duckworth, London. Cited on page 16.

Armendariz, C. S., Purver, M., Ulčar, M., Pollak, S., Ljubešić, N., and Granroth-Wilding, M. (2020). CoSimLex: A resource for evaluating graded word similarity in context. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5878–5886, Marseille, France. European Language Resources Association. Cited on pages 25, 69, and 136.

Artetxe, M., Labaka, G., and Agirre, E. (2017). Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 451–462. Cited on page 81.

Artetxe, M., Labaka, G., and Agirre, E. (2018a). Generalizing and improving bilingual word embedding mappings with a multi-step framework of linear transformations. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, pages 5012–5019. Cited on pages 81 and 110.

Artetxe, M., Labaka, G., and Agirre, E. (2018b). A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 789–798, Melbourne, Australia. Association for Computational Linguistics. Cited on page 81.

Artstein, R. and Poesio, M. (2008). Inter-coder agreement for computational linguistics. *Comput. Linguist.*, 34(4):555—-596. Cited on page 58.

Asgari, E., Ringlstetter, C., and Schütze, H. (2020). EmbLexChange at SemEval-2020 Task 1: Unsupervised Embedding-based Detection of Lexical Semantic Changes. In *Proceedings of the 14th International Workshop on Semantic Evaluation*, Barcelona, Spain. Association for Computational Linguistics. Cited on page 92.

Baldissin, G., Schlechtweg, D., and Schulte im Walde, S. (2022). DiaWUG: A Dataset for Diatopic Lexical Semantic Variation in Spanish. In *Proceedings of the 13th Language Resources and Evaluation Conference*, Marseille, France. European Language Resources Association. Cited on pages 67, 134, and 135.

Bamler, R. and Mandt, S. (2017). Dynamic word embeddings. In Precup, D. and Teh, Y. W., editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 380–389, International Convention Centre, Sydney, Australia. PMLR. Cited on pages 27 and 38.

Bamman, D. and Crane, G. (2011). Measuring historical word sense variation. In *Proceedings of the 11th Annual International ACM/IEEE Joint Conference on Digital Libraries*, pages 1–10, New York, NY, USA. ACM. Cited on page 22.

Bansal, N., Blum, A., and Chawla, S. (2004). Correlation clustering. *Machine Learning*, 56(1-3):89–113. Cited on pages 34 and 35.

Baroni, M., Bernardini, S., Ferraresi, A., and Zanchetta, E. (2009). The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. *Language resources and evaluation*, 43(3):209–226. Cited on page 50.

Baroni, M., Dinu, G., and Kruszewski, G. (2014). Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. *52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014 - Proceedings of the Conference*, 1:238–247. Cited on page 80.

Basile, P., Caputo, A., Caselli, T., Cassotti, P., and Varvara, R. (2020). Overview of the EVALITA 2020 Diachronic Lexical Semantics (DIACR-Ita) Task. In Basile, V., Croce, D., Di Maro, M., and Passaro, L. C., editors, *Proceedings of the 7th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2020)*, Online. CEUR.org. Cited on pages 114 and 134.

Basile, P., Caputo, A., and Semeraro, G. (2015). Temporal random indexing: A system for analysing word meaning over time. *Italian Journal of Computational Linguistics*, 1:55–68. Cited on pages 27, 28, 79, 80, and 107.

Berliner Zeitung (2018). Diachronic newspaper corpus published by Staatsbibliothek zu Berlin. https://zefys.staatsbibliothek-berlin.de. Cited on page 42.

Biemann, C. (2006). Chinese whispers: An efficient graph clustering algorithm and its application to natural language processing problems. In *Proceedings of the First Workshop on Graph Based Methods for Natural Language Processing*, TextGraphs-1, page 73–80, USA. Association for Computational Linguistics. Cited on pages 35 and 171.

Biemann, C. (2013). Creating a system for lexical substitutions from scratch using crowdsourcing. *Lang. Resour. Eval.*, 47(1):97–122. Cited on page 45.

Blank, A. (1997). *Prinzipien des lexikalischen Bedeutungswandels am Beispiel der romanischen Sprachen*. Niemeyer, Tübingen. Cited on pages xv, 1, 2, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 22, 24, 25, 32, 33, 34, 36, 37, 41, 65, 66, and 131.

Blei, D. M. (2012). Probabilistic topic models. *Commun. ACM*, 55(4):77–84. Cited on pages 26 and 66.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):10008. Cited on page 171.

Bloomfield, L. (1935). *Language (Reprint: London 1967)*. Allen & Unwin, New York. Cited on page 12.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146. Cited on page 92.

Borin, L., Forsberg, M., and Roxendal, J. (2012). Korp — the corpus infrastructure of Språkbanken. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 474–478, Istanbul, Turkey. European Language Resources Association (ELRA). Cited on page 43.

Bott, S. and Schulte im Walde, S. (2014). Optimizing a distributional semantic model for the prediction of German particle verb compositionality. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland. European Language Resources Association (ELRA). Cited on page 106.

Bott, T., Schlechtweg, D., and Schulte im Walde, S. (2021). More than just Frequency? Demasking Unsupervised Hypernymy Prediction Methods. In *Proceedings of the Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Findings)*, Online. Association for Computational Linguistics. Cited on page 97.

Bréal, M. (1899). *Essai de sémantique (first published 1897)*. Hachette. Cited on page 12.

Brown, S. W. (2008). Choosing sense distinctions for WSD: Psycholinguistic evidence. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers*, pages 249–252, Stroudsburg, PA, USA. Cited on pages 12, 22, and 32.

Bullinaria, J. and Levy, J. (2012). Extracting semantic representations from word co-occurrence statistics: Stop-lists, stemming, and svd. *Behavior research methods*, 44:890–907. Cited on page 78.

Bybee, J. L. (2015). *Language change*. Cambridge University Press, Cambridge, United Kingdom. Cited on page 1.

Campello, R. J. G. B., Moulavi, D., and Sander, J. (2013). Density-based clustering based on hierarchical density estimates. In Pei, J., Tseng, V. S., Cao, L., Motoda, H., and Xu, G., editors, *Advances in Knowledge Discovery and Data Mining*, pages 160–172, Berlin, Heidelberg. Springer Berlin Heidelberg. Cited on page 94.

Cassotti, P., Caputo, A., Polignano, M., and Basile, P. (2020). GM-CTSC at SemEval-2020 task 1: Gaussian mixtures cross temporal similarity clustering. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 74–80, Barcelona (online). International Committee for Computational Linguistics. Cited on page 94.

Ceron, T., Blokker, N., and Padó, S. (2022). Optimizing text representations to capture (dis)similarity between political parties. In *Proceedings of the 26th Conference* *on Computational Natural Language Learning*. Association for Computational Linguistics. Cited on page 136.

Chomsky, N. (1986). *Knowledge of Language. Its Nature, Origin, and Use*. Convergence. Praeger, New York/Westport/London. Cited on page 20.

Clackson, J. (2011). *A Companion to the Latin Language*. Wiley-Blackwell. Cited on page 168.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2019). Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*. Cited on page 91.

Cook, P., Lau, J. H., McCarthy, D., and Baldwin, T. (2014). Novel word-sense identification. In *25th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers*, pages 1624–1635, Dublin, Ireland. Cited on pages 2, 22, 27, 28, and 29.

Cook, P. and Stevenson, S. (2010). Automatically Identifying Changes in the Semantic Orientation of Words. In Chair, N. C. C., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Rosner, M., and Tapias, D., editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA). Cited on pages 12 and 28.

Cover, T. and Thomas, J. (2006). *Elements of Information Theory*. Wiley. Cited on page 163.

Davies, M. (2012). Expanding Horizons in Historical Linguistics with the 400-Million Word Corpus of Historical American English. *Corpora*, 7(2):121–157. Cited on page 42.

de Saussure, F. (1968). *Cours de linguistique générale (first published 1916)*. Payot, Paris. Cited on page 12.

Deane, P. D. (1988). Polysemy and cognition. *Lingua*, 75(4):325–361. Cited on pages 22 and 41.

del Tredici, M. and Fernández, R. (2017). Semantic variation in online communities of practice. In *IWCS 2017 - 12th International Conference on Computational Semantics - Long papers*. Cited on page 28.

del Tredici, M., Nissim, M., and Zaninello, A. (2016). Tracing metaphors in time through self-distance in vector spaces. In *Proceedings of the 3rd Italian Conference on Computational Linguistics*. Cited on page 28.

Deutsches Textarchiv (2017). Grundlage für ein Referenzkorpus der neuhochdeutschen Sprache. Herausgegeben von der Berlin-Brandenburgischen Akademie der Wissenschaften. Cited on page 42.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics. Cited on pages 1, 26, and 70.

Donoso, G. and Sanchez, D. (2017). Dialectometric analysis of language variation in twitter. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 16–25, Valencia, Spain. Cited on page 38.

Dubossarsky, H., Hengchen, S., Tahmasebi, N., and Schlechtweg, D. (2019). Time-Out: Temporal Referencing for Robust Modeling of Lexical Semantic Change. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 457–470, Florence, Italy. Association for Computational Linguistics. Cited on pages 82 and 111.

Dubossarsky, H., Weinshall, D., and Grossman, E. (2017). Outta control: Laws of semantic change and inherent biases in word representation models. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1147–1156, Copenhagen, Denmark. Cited on pages 2, 27, 38, 44, 97, and 135.

DUDEN (2021). Duden online. `www.duden.de`. Accessed: 2021-02-01. Cited on pages 125 and 126.

DWB (2021). Deutsches Wörterbuch von Jacob Grimm und Wilhelm Grimm, digitalized edition curated by the Wörterbuchnetz at the Trier Center for Digital Humanities. https://www.woerterbuchnetz.de/DWB. Accessed: 2021-01-07. Cited on pages 125 and 126.

DWDS (2021). Digitales Wörterbuch der deutschen Sprache. Das Wortauskunftssystem zur deutschen Sprache in Geschichte und Gegenwart, hrsg. v. d. Berlin-Brandenburgischen Akademie der Wissenschaften. https://www.dwds.de/. Accessed: 02.02.2021. Cited on pages 57 and 125.

Eckart, C. and Young, G. (1936). The approximation of one matrix by another of lower rank. *Psychometrika*, 1:211–218. Cited on page 78.

Eger, S. and Mehler, A. (2016). On the linearity of semantic change: Investigating meaning variation via dynamic graph models. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 2: Short Papers*. Cited on pages 27 and 28.

Erk, K., McCarthy, D., and Gaylord, N. (2009). Investigations on word senses and word usages. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1*, pages 10–18, Stroudsburg, PA, USA. Cited on pages 21 and 22.

Erk, K., McCarthy, D., and Gaylord, N. (2013). Measuring word meaning in context. *Computational Linguistics*, 39(3):511–554. Cited on pages 12, 21, 22, 32, 33, 48, 55, 58, 66, and 165.

Faaß, G. and Eckart, K. (2013). SdeWaC – A corpus of parsable sentences from the web. In Gurevych, I., Biemann, C., and Zesch, T., editors, *Language Processing and Knowledge in the Web*, volume 8105 of *Lecture Notes in Computer Science*, pages 61–68. Springer Berlin Heidelberg. Cited on page 50.

Ferrara, A., Montanelli, S., and Ruskov, M. (2022). Detecting the semantic shift of values in cultural heritage document collections (short paper). In Damiano, R., Ferilli, S., Striani, M., and Silvello, G., editors, *Proceedings of the 1st Workshop on Artificial Intelligence for Cultural Heritage, AI4CH 2022, co-located with the 21st International Conference of the Italian Association for Artificial Intelligence (AIxIA 2022), Udine, Italy, November 28, 2022*, volume 3286 of *CEUR Workshop Proceedings*, pages 35–43. CEUR-WS.org. Cited on page 136.

Ferrari, A., Donati, B., and Gnesi, S. (2017). Detecting domain-specific ambiguities: An NLP approach based on wikipedia crawling and word embeddings. In

Proceedings of the 2017 IEEE 25th International Requirements Engineering Conference Workshops, pages 393–399. Cited on pages 28 and 81.

FHDW (2021). Frühneuhochdeutsches Wörterbuch. `https://fwb-online.de`. Accessed: 2021-01-07. Cited on page 126.

Fillmore, C. J. (1975). An alternative to checklist theories of meaning. In *Proceedings of the 1st Annual Meeting of the Berkeley Linguistic Society*, pages 123–131, Berkeley. Berkeley Linguistic Society. Cited on page 16.

Forgy, E. W. (1965). Cluster analysis of multivariate data: Efficiency vs. interpretability of classifications. *Biometrics*, 21:768–780. Cited on page 99.

Frermann, L. and Lapata, M. (2016). A Bayesian model of diachronic meaning change. *Transactions of the Association for Computational Linguistics*, 4:31–45. Cited on pages 2, 27, 28, 29, 85, 107, and 134.

Frey, B. J. and Dueck, D. (2007). Clustering by Passing Messages Between Data Points. *Science*, 315(5814):972–976. Cited on page 94.

Futrzynski, R. (2021). Peltarion.
`https://peltarion.com/blog/data-science/self-attention-video`. Accessed: 2021-02-26. Cited on pages 72 and 73.

Geeraerts, D. (1983). Prototype theory and diachronic semantics. a case study. *Indogermanische Forschungen (1983)*, 88(1983):1–32. Cited on page 12.

Geeraerts, D. (1992). The return of hermeneutics to lexical semantics. *Thirty Years of Linguistic Evolution. Studies in Honour of René Dirven on the Occasion of his Sixtieth Birthday*, pages 257–282. Cited on page 12.

Geeraerts, D. (2010). *Theories of Lexical Semantics*. Oxford linguistics. OUP Oxford. Cited on page 12.

Geeraerts, D. (2020). *Semantic Change*, chapter 1, pages 1–24. American Cancer Society. Cited on page 1.

Giulianelli, M., del Tredici, M., and Fernández, R. (2020). Analysing lexical semantic change with contextualised word representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3960–3973, Online. Association for Computational Linguistics. Cited on pages 67, 69, 74, 75, and 134.

Gloning, T. (2017). *Alte Zeitungen und historische Lexikographie. Nutzungsperspektiven, Korpora, Forschungsinfrastrukturen*, pages 121–148. De Gruyter, Berlin, Boston. Cited on page 127.

Goldberg, Y. and Levy, O. (2014). word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method. *arXiv preprint arXiv:1402.3722*. Cited on page 79.

Gruppi, M., Adalı, S., and Chen, P.-Y. (2020). SChME at SemEval-2020 Task 1: A Model Ensemble for Detecting Lexical Semantic Change. In *Proceedings of the 14th International Workshop on Semantic Evaluation*, Barcelona, Spain. Association for Computational Linguistics. Cited on page 92.

Grzega, J. (2002). Some aspects of modern diachronic onomasiology. *Linguistics*, 40(5):1021–1045. Cited on page 12.

Gulordava, K. and Baroni, M. (2011). A distributional similarity approach to the detection of semantic change in the Google Books Ngram corpus. In *Proceedings of the Workshop on Geometrical Models of Natural Language Semantics*, pages 67–71, Stroudsburg, PA, USA. Cited on pages 27, 28, and 29.

Györi, G. (2002). Semantic change and cognition. *Cognitive Linguistics*, 13(2):123–166. Cited on page 12.

Hamilton, W. L., Leskovec, J., and Jurafsky, D. (2016a). Cultural shift or linguistic drift? Comparing two computational measures of semantic change. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2116–2121, Austin, Texas. Cited on pages 27, 28, 82, 83, 108, and 136.

Hamilton, W. L., Leskovec, J., and Jurafsky, D. (2016b). Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1501, Berlin, Germany. Cited on pages 27, 28, 29, 38, 78, 80, 81, 107, 110, 114, and 135.

Harris, Z. S. (1954). Distributional structure. *Word*, 10:146–162. Cited on pages 25 and 69.

Hätty, A., Schlechtweg, D., Dorna, M., and Schulte im Walde, S. (2020). Predicting Degrees of Technicality in Automatic Terminology Extraction. In *Proceedings* *of the 58th Annual Meeting of the Association for Computational Linguistics*, Seattle, Washington. Association for Computational Linguistics. Cited on page 38.

Hätty, A., Schlechtweg, D., and Schulte im Walde, S. (2019). SURel: A gold standard for incorporating meaning shifts into term extraction. In *Proceedings of the 8th Joint Conference on Lexical and Computational Semantics*, pages 1–8, Minneapolis, MN, USA.

Haug, D. T. T. and Jøhndal, M. L. (2008). Creating a parallel treebank of the old indo-european bible translations. In Sporleder, C. and Ribarov, K., editors, *Proceedings of the Second Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2008)*, pages 27–34. Cited on page 167.

Hayes, G. (2019). mlrose: Machine Learning, Randomized Optimization and SEarch package for Python. `https://github.com/gkhayes/mlrose`. Accessed: May 22, 2020. Cited on page 35.

Hellrich, J. and Hahn, U. (2016). Bad Company—Neighborhoods in Neural Embedding Spaces Considered Harmful. In *Proceedings of COLING 2016*, pages 2785–2796, Osaka, Japan. Cited on page 27.

Hengchen, S., Ros, R., and Marjanen, J. (2019). A data-driven approach to the changing vocabulary of the 'nation' in English, Dutch, Swedish and Finnish newspapers, 1750-1950. In *Proceedings of the Digital Humanities (DH) conference 2019*, Utrecht, The Netherlands. Cited on pages 115 and 136.

Hengchen, S., Ros, R., Marjanen, J., and Tolonen, M. (2020). A data-driven approach to studying changing vocabularies in historical newspaper collections. *Digital Scholarship in the Humanities*. Cited on page 43.

Hengchen, S., Tahmasebi, N., Schlechtweg, D., and Dubossarsky, H. (2021). Challenges for Computational Lexical Semantic Change. In Tahmasebi, N., Borin, L., Jatowt, A., Xu, Y., and Hengchen, S., editors, *Computational Approaches to Semantic Change*, volume Language Variation, chapter 11. Language Science Press, Berlin. Cited on pages 11 and 114.

Homskiy, D. and Arefyev, N. (2022). DeepMistake at LSCDiscovery: Can a multilingual word-in-context model replace human annotators? In *Proceedings of the 3rd International Workshop on Computational Approaches to Historical Language Change*, Dublin, Ireland. Association for Computational Linguistics. Cited on pages 87, 114, and 134.

Honnibal, M., Montani, I., Van Landeghem, S., and Boyd, A. (2020). spaCy: Industrial-strength Natural Language Processing in Python. Cited on page 118.

Hovy, E., Marcus, M., Palmer, M., Ramshaw, L., and Weischedel, R. (2006). Ontonotes: The 90% solution. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, NAACL-Short '06, pages 57—-60, USA. Association for Computational Linguistics. Cited on page 22.

Hubert, L. and Arabie, P. (1985). Comparing partitions. *Journal of Classification*, 2:193–218. Cited on pages 58 and 90.

Jatowt, A. and Duh, K. (2014). A framework for analyzing semantic change of words across time. In *Proceedings of the 14th ACM/IEEE-CS Joint Conference on Digital Libraries*, JCDL '14, page 229–238. IEEE Press. Cited on page 28.

Jatowt, A., Tahmasebi, N., and Borin, L. (2021). Computational approaches to lexical semantic change: Visualization systems and novel applications. In Tahmasebi, N., Borin, L., Jatowt, A., Xu, Y., and Hengchen, S., editors, *Computational Approaches to Semantic Change*, Language Variation, chapter 10. Language Science Press, Berlin. Cited on page 115.

Jawahar, G., Sagot, B., and Seddah, D. (2019). What does BERT learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657, Florence, Italy. Association for Computational Linguistics. Cited on page 100.

Johnson, W. B. and Lindenstrauss, J. (1984). Extensions to lipshitz mapping into hilbert space. *Contemporary mathematics*, 26. Cited on page 79.

Jurgens, D. and Klapaftis, I. (2013). SemEval-2013 task 13: Word sense induction for graded and non-graded senses. In *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. Association for Computational Linguistics. Cited on page 66.

Kaiser, J., Kurtyigit, S., Kotchourko, S., and Schlechtweg, D. (2021). Effects of pre- and post-processing on type-based embeddings in lexical semantic change detection. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 125–137, Online. Association for Computational Linguistics. Cited on pages 111, 114, and 117.

Kaiser, J., Schlechtweg, D., Papay, S., and Schulte im Walde, S. (2020a). IMS at SemEval-2020 Task 1: How low can you go? Dimensionality in Lexical Semantic Change Detection. In *Proceedings of the 14th International Workshop on Semantic Evaluation*, Barcelona, Spain. Association for Computational Linguistics. Cited on pages 111, 114, and 117.

Kaiser, J., Schlechtweg, D., and Schulte im Walde, S. (2020b). OP-IMS @ DIACR-Ita: Back to the Roots: SGNS+OP+CD still rocks Semantic Change Detection. In Basile, V., Croce, D., Di Maro, M., and Passaro, L. C., editors, *Proceedings of the 7th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2020)*, Online. CEUR.org. Cited on pages 86, 114, and 117.

Kashleva, K., Shein, A., Tukhtina, E., and Vydrina, S. (2022). HSE at LSCDiscovery in Spanish: Clustering and profiling for lexical semantic change discovery. In *Proceedings of the 3rd International Workshop on Computational Approaches to Historical Language Change*, Dublin, Ireland. Association for Computational Linguistics. Cited on page 114.

Kilgarriff, A. (1997). "I don't believe in word senses". *Computers and the Humanities*, 31(2). Cited on pages 21 and 67.

Kilgarriff, A. (2004). How dominant is the commonest sense of a word? In *International conference on text, speech and dialogue*, pages 103–111. Springer. Cited on pages 63 and 136.

Kilgarriff, A. (2007). *Word Senses*, chapter 2, pages 29–46. Springer. Cited on pages 20, 21, and 34.

Kim, Y., Chiu, Y., Hanaki, K., Hegde, D., and Petrov, S. (2014). Temporal analysis of language through neural language models. In *LTCSS@ACL*, pages 61–65. Association for Computational Linguistics. Cited on pages 27, 28, and 81.

Kisselew, M., Rimell, L., Palmer, A., and Pado, S. (2016). Predicting the direction of derivation in english conversion. In *Proceedings of the 14th SIGMORPHON Work-*

shop on Computational Research in Phonetics, Phonology, and Morphology, Berlin, Germany. Cited on page 27.

Knapp, T. R. (1978). Canonical correlation analysis: A general parametric significance-testing system. *Psychological Bulletin*, 85:410–416. Cited on page 92.

Koplenig, A. (2019). Against statistical significance testing in corpus linguistics. *Corpus Linguistics and Linguistic Theory*, 15(2):321–346. Cited on pages 32 and 65.

Kotchourko, S. (2021). Optimizing human annotation of word usage graphs in a realistic simulation environment. Bachelor thesis, University of Stuttgart. Cited on pages 67 and 135.

Krippendorff, K. (2018). *Content Analysis: An Introduction to Its Methodology*. SAGE Publications. Cited on page 49.

Kuhn, H. W. (1955). The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1-2):83–97. Cited on page 60.

Kulkarni, V., Al-Rfou, R., Perozzi, B., and Skiena, S. (2015). Statistically significant detection of linguistic change. In *Proceedings of the 24th International Conference on World Wide Web, WWW*, pages 625–635, Florence, Italy. Cited on page 28.

Kullback, S. and Leibler, R. A. (1951). On Information and Sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86. Cited on page 92.

Kurtyigit, S. (2021). Lexical semantic change discovery. Bachelor thesis, University of Stuttgart. Cited on pages 72 and 73.

Kurtyigit, S., Park, M., Schlechtweg, D., Kuhn, J., and Schulte im Walde, S. (2021). Lexical Semantic Change Discovery. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Online. Association for Computational Linguistics. Cited on page 67.

Kutuzov, A. (2020). Distributional word embeddings in modeling diachronic semantic change. Cited on page 123.

Kutuzov, A. and Giulianelli, M. (2020). UiO-UvA at SemEval-2020 Task 1: Contextualised Embeddings for Lexical Semantic Change Detection. In *Proceedings of the 14th International Workshop on Semantic Evaluation*, Barcelona, Spain. Association for Computational Linguistics. Cited on pages 76 and 103.

Kutuzov, A., Øvrelid, L., Szymanski, T., and Velldal, E. (2018). Diachronic word embeddings and semantic shifts: a survey. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1384–1397, Santa Fe, New Mexico, USA. Association for Computational Linguistics. Cited on page 11.

Kutuzov, A. and Pivovarova, L. (2021a). Rushifteval: a shared task on semantic shift detection for russian. *Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Dialog conference*. Cited on pages 114 and 134.

Kutuzov, A. and Pivovarova, L. (2021b). Three-part diachronic semantic change dataset for russian. Cited on page 67.

Kutuzov, A., Touileb, S., Mæhlum, P., Enstad, T., and Wittemann, A. (2022). NorDiaChange: Diachronic semantic change dataset for Norwegian. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2563–2572, Marseille, France. European Language Resources Association. Cited on pages 67 and 134.

Kutuzov, A., Velldal, E., and Øvrelid, L. (2017). Tracing armed conflicts with diachronic word embedding models. In *Proceedings of the Events and Stories in the News Workshop*, pages 31–36, Vancouver, Canada. Association for Computational Linguistics. Cited on page 136.

Laicher, S. (2021). Historical word sense clustering with deep contextualised word embeddings. Bachelor thesis, University of Stuttgart.

Laicher, S., Baldissin, G., Castaneda, E., Schlechtweg, D., and Schulte im Walde, S. (2020). CL-IMS @ DIACR-Ita: Volente o Nolente: BERT does not outperform SGNS on Semantic Change Detection. In Basile, V., Croce, D., Di Maro, M., and Passaro, L. C., editors, *Proceedings of the 7th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2020)*, Online. CEUR.org. Cited on page 114.

Laicher, S., Kurtyigit, S., Schlechtweg, D., Kuhn, J., and Schulte im Walde, S. (2021). Explaining and Improving BERT Performance on Lexical Semantic Change Detection. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 192–202, Online. Association for Computational Linguistics.

Lakoff, G. and Johnson, M. (1980). *Metaphors We Live By*. Phoenix books. University of Chicago Press. Cited on page 24.

Landauer, T. K. and Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104(2):211–240. Cited on page 26.

Langone, H., Haskell, B. R., and Miller, G. A. (2004). Annotating wordnet. In *Proceedings of the Workshop Frontiers in Corpus Annotation at HLT-NAACL*, Boston, MA, USA. Cited on page 22.

Lau, J. H., Cook, P., McCarthy, D., Gella, S., and Baldwin, T. (2014). Learning word sense distributions, detecting unattested senses and identifying novel senses using topic models. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 259–270, Baltimore, Maryland. Association for Computational Linguistics. Cited on page 36.

Lau, J. H., Cook, P., McCarthy, D., Newman, D., and Baldwin, T. (2012). Word sense induction for novel sense detection. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 591–601, Stroudsburg, PA, USA. Cited on pages 2, 22, 27, 28, and 29.

Lehmann, H. (1884). *Der Bedeutungswandel im französischen*. A. Deichert. Cited on page 12.

Lenci, A. (2008). Distributional semantics in linguistic and cognitive research. *Rivista di Linguistica*, 20(1):1–31. Cited on page 25.

Levy, O. and Goldberg, Y. (2014). Neural word embedding as implicit matrix factorization. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, pages 2177–2185, Cambridge, MA, USA. MIT Press. Cited on pages 78 and 79.

Levy, O., Goldberg, Y., and Dagan, I. (2015). Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225. Cited on pages 78, 79, 80, 106, 107, and 109.

Lin, J. (1991). Divergence measures based on the shannon entropy. *IEEE Transactions on Information Theory*, 37:145–151. Cited on pages 38, 92, and 156.

Ljubešić, N. (2020). "Deep lexicography" – Fad or Opportunity? "Duboka leksikografija" – pomodnost ili prilika? *Rasprave Instituta za hrvatski jezik i jezikoslovlje*, 46:839–852. Cited on page 115.

Margalitadze, T. (2018). Once again why lexicography is science. *Lexikos*, 28(1). Cited on page 136.

Martinc, M., Montariol, S., Zosa, E., and Pivovarova, L. (2020a). Capturing evolution in word usage: Just add more clusters? In *Companion Proceedings of the Web Conference 2020*, WWW '20, pages 343—-349, New York, NY, USA. Association for Computing Machinery. Cited on page 74.

Martinc, M., Montariol, S., Zosa, E., and Pivovarova, L. (2020b). Discovery Team at SemEval-2020 Task 1: Context-sensitive Embeddings not Always Better Than Static for Semantic Change Detection. In *Proceedings of the 14th International Workshop on Semantic Evaluation*, Barcelona, Spain. Association for Computational Linguistics. Cited on pages 94, 96, and 99.

McCarthy, D., Apidianaki, M., and Erk, K. (2016). Word sense clustering and clusterability. *Computational Linguistics*, 42(2):245–275. Cited on pages 24, 33, and 35.

McCarthy, D., Koeling, R., Weeds, J., and Carroll, J. (2004). Finding predominant word senses in untagged text. In *Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL'04), Main Volume*, pages 279–286, Barcelona, Spain. Cited on page 36.

McCarthy, D. and Navigli, R. (2009). The english lexical substitution task. *Language Resources and Evaluation*, 43(2):139–159. Cited on page 24.

McGillivray, B. and Kilgarriff, A. (2013). Tools for historical corpus research, and a corpus of Latin. In Bennett, P., Durrell, M., Scheible, S., and Whitt, R. J., editors, *New Methods in Historical Corpus Linguistics*, Tübingen. Narr. Cited on page 167.

Mihalcea, R. and Nastase, V. (2012). Word Epoch Disambiguation: Finding How Words Change Over Time. In *Proceedings of the 50th Annual Meeting of ACL*. Cited on page 28.

Mihalcea, R. and Tarau, P. (2004). TextRank: Bringing order into text. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 404–411, Barcelona, Spain. Association for Computational Linguistics. Cited on page 26.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient estimation of word representations in vector space. In Bengio, Y. and LeCun, Y., editors, *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*. Cited on pages 1, 26, 79, and 106.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. In *Proceedings of NIPS*. Cited on page 79.

Minsky, M. (1975). Minsky's frame system theory. In *Theoretical Issues in Natural Language Processing*. Cited on page 16.

Mitra, S., Mitra, R., Maity, S. K., Riedl, M., Biemann, C., Goyal, P., and Mukherjee, A. (2015). An automatic approach to identify word sense changes in text media across timescales. *Natural Language Engineering*, 21(5):773–798. Cited on pages 27 and 28.

Montariol, S., Martinc, M., and Pivovarova, L. (2021). Scalable and interpretable semantic change detection. In *2021 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Cited on page 118.

Nanni, F., Glavaš, G., Rehbein, I., Ponzetto, S. P., and Stuckenschmidt, H. (2022). Political text scaling meets computational semantics. *ACM/IMS Trans. Data Sci.*, 2(4). Cited on page 136.

Navigli, R. (2009). Word sense disambiguation: a survey. *ACM Computing Surveys*, 41(2):1–69. Cited on pages 21 and 26.

Neues Deutschland (2018). Diachronic newspaper corpus published by Staatsbibliothek zu Berlin. https://zefys.staatsbibliothek-berlin.de. Cited on page 42.

OED (2009). *Oxford English Dictionary*. Oxford University Press. Cited on page 44.

Osman, N. (1971). *Kleines Lexikon untergegangener Wörter: Wortuntergang seit dem Ende des 18. Jahrhunderts*. Beck, München. Cited on pages 17 and 51.

Padó, S. and Lapata, M. (2007). Dependency-based construction of semantic space models. *Computational Linguistics*, 33(2):161–199. Cited on page 106.

Panchenko, A., Lopukhina, A., Ustalov, D., Lopukhin, K., Arefyev, N., Leontyev, A., and Loukachevitch, N. (2018). Russe'2018: A shared task on word sense induction for the russian language. *Komp'juternaja Lingvistika i Intellektual'nye Tehnologii*, 2018-May(17):547–564. Cited on page 74.

Paul, H. (1975). *Prinzipien der Sprachgeschichte (first published 1886)*. M. Niemeyer, Halle. Cited on page 12.

Paul, H. (2002). *Deutsches Wörterbuch: Bedeutungsgeschichte und Aufbau unseres Wortschatzes*. Niemeyer, Tübingen, 10. edition. Cited on pages 1, 44, 51, and 57.

Peixoto, T. P. (2017). Nonparametric weighted stochastic block models. *Physical Review E*, 97. Cited on page 66.

Pennington, J., Socher, R., and Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1532–1543, Doha, Qatar. Cited on pages 26 and 91.

Perrone, V., Palma, M., Hengchen, S., Vatri, A., Smith, J. Q., and McGillivray, B. (2019). GASC: Genre-aware semantic change for ancient Greek. In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, pages 56–66, Florence, Italy. Association for Computational Linguistics. Cited on page 66.

Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2227–2237, New Orleans, LA, USA. Cited on pages 1, 26, 91, and 95.

Pilehvar, M. T. and Camacho-Collados, J. (2019). WiC: the word-in-context dataset for evaluating context-sensitive meaning representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1267–1273, Minneapolis, Minnesota. Association for Computational Linguistics. Cited on pages 25, 69, 114, 134, and 136.

Pincus, M. (1970). A monte carlo method for the approximate solution of certain types of constrained optimization problems. *Operations Research*, 18(6):1225–1228. Cited on page 35.

Pömsl, M. and Lyapin, R. (2020). CIRCE at SemEval-2020 Task 1: Ensembling Context-Free and Context-Dependent Word Representations. In *Proceedings of the 14th International Workshop on Semantic Evaluation*, Barcelona, Spain. Association for Computational Linguistics. Cited on page 92.

Pražák, O., Přibáň, P., and Taylor, S. (2020a). UWB @ DIACR-Ita: Lexical Semantic Change Detection with CCA and Orthogonal Transformation. In Basile, V., Croce, D., Di Maro, M., and Passaro, L. C., editors, *Proceedings of the 7th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2020)*, Online. CEUR.org. Cited on page 114.

Pražák, O., Přibáň, P., Taylor, S., and Sido, J. (2020b). UWB at SemEval-2020 Task 1: Lexical Semantic Change Detection. In *Proceedings of the 14th International Workshop on Semantic Evaluation*, Barcelona, Spain. Association for Computational Linguistics. Cited on page 92.

Raible, W. (1983). Zur Einleitung. In Stimm, H. and Raible, W., editors, *Zur Semantik des Französischen*, pages 1–24. Steiner, Wiesbaden. Cited on pages 14 and 15.

Řehůřek, R. and Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA. Cited on page 107.

Reisig, K. (1972). *Semasiologie oder Bedeutungslehre*, pages 21–40. Cited on pages 11 and 12.

Rodina, J. and Kutuzov, A. (2020). RuSemShift: a dataset of historical lexical semantic change in Russian. In *Proceedings of the 28th International Conference on Computational Linguistics (COLING 2020)*. Association for Computational Linguistics. Cited on pages 49, 67, and 134.

Rosch, E. and Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology*, 7:573–605. Cited on page 15.

Rosenfeld, A. and Erk, K. (2018). Deep neural models of semantic shift. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 474–484, New Orleans, Louisiana. Cited on pages 27, 28, 38, 87, and 134.

Rother, D., Haider, T., and Eger, S. (2020). CMCE at SemEval-2020 task 1: Clustering on manifolds of contextualized embeddings to detect historical meaning shifts. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 187–193, Barcelona (online). International Committee for Computational Linguistics. Cited on page 94.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65. Cited on page 74.

Rudolph, M. R. and Blei, D. M. (2018). Dynamic embeddings for language evolution. In *WWW 2018*, pages 1003–1011. ACM. Cited on pages 28 and 38.

Sagi, E., Kaufmann, S., and Clark, B. (2009). Semantic density analysis: Comparing word meaning across time and phonetic space. In *Proceedings of the Workshop on Geometrical Models of Natural Language Semantics*, pages 104–111, Athens, Greece. Association for Computational Linguistics. Cited on pages 12, 27, 28, and 87.

Sahlgren, M. (2004). An introduction to random indexing. *Language*, pages 1–9. Cited on page 79.

Salton, G. and McGill, M. J. (1983). *Introduction to Modern Information Retrieval*. McGraw - Hill Book Company, New York. Cited on pages 27 and 82.

Santus, E., Lenci, A., Lu, Q., and Schulte im Walde, S. (2014). Chasing hypernyms in vector spaces with entropy. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 38–42. Cited on page 84.

Schlechtweg, D., Castaneda, E., Kuhn, J., and Schulte im Walde, S. (2021a). Modeling sense structure in word usage graphs with the weighted stochastic block model. In *Proceedings of *SEM 2021: The Tenth Joint Conference on Lexical and Computational Semantics*, pages 241–251, Online. Association for Computational Linguistics. Cited on page 135.

Schlechtweg, D., Eckmann, S., Santus, E., Schulte im Walde, S., and Hole, D. (2017). German in flux: Detecting metaphoric change via word entropy. In *Proceedings of the 21st Conference on Computational Natural Language Learning*, pages 354–367, Vancouver, Canada. Cited on pages 22, 27, 28, 84, 97, and 111.

Schlechtweg, D., Hätty, A., del Tredici, M., and Schulte im Walde, S. (2019a). A Wind of Change: Detecting and Evaluating Lexical Semantic Change across Times and Domains. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 732–746, Florence, Italy. Association for Computational Linguistics. Cited on pages 95 and 97.

Schlechtweg, D., McGillivray, B., Hengchen, S., Dubossarsky, H., and Tahmasebi, N. (2020). SemEval-2020 Task 1: Unsupervised Lexical Semantic Change Detection. In *Proceedings of the 14th International Workshop on Semantic Evaluation*, Barcelona, Spain. Association for Computational Linguistics. Cited on pages 1 and 91.

Schlechtweg, D., Oguz, C., and Schulte im Walde, S. (2019b). Second-order co-occurrence sensitivity of skip-gram with negative sampling. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 24–30, Florence, Italy. Association for Computational Linguistics.

Schlechtweg, D. and Schulte im Walde, S. (2018). Distribution-based prediction of the degree of grammaticalization for German prepositions. In Cuskley, C., Flaherty, M., Little, H., McCrohon, L., Ravignani, A., and Verhoef, T., editors, *The Evolution of Language: Proceedings of the 12th International Conference (EVOLANGXII)*. Online at http://evolang.org/torun/proceedings/papertemplate.html?p=169.

Schlechtweg, D. and Schulte im Walde, S. (2020). Simulating Lexical Semantic Change from Sense-Annotated Data. In Ravignani, A., Barbieri, C., Martins, M., Flaherty, M., Jadoul, Y., Lattenkamp, E., Little, H., Mudd, K., and Verhoef, T., editors, *The Evolution of Language: Proceedings of the 13th International Conference (EvoLang13)*. Cited on page 44.

Schlechtweg, D., Schulte im Walde, S., and Eckmann, S. (2018). Diachronic Usage Relatedness (DURel): A framework for the annotation of lexical semantic change. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 169–174, New Orleans, Louisiana. Cited on pages 53, 75, and 105.

Schlechtweg, D., Tahmasebi, N., Hengchen, S., Dubossarsky, H., and McGillivray, B. (2021b). DWUG: A large Resource of Diachronic Word Usage Graphs in Four Languages. In *Proceedings of the 2021 Conference on Empirical Methods in Natural*

Language Processing, pages 7079–7091, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Schulte im Walde, S., Müller, S., and Roller, S. (2013). Exploring vector space models to predict the compositionality of German noun-noun compounds. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*, pages 255–265, Atlanta, Georgia, USA. Association for Computational Linguistics. Cited on page 106.

Schütze, H. (1998). Automatic word sense discrimination. *Computational Linguistics*, 24(1):97–123. Cited on pages 26, 27, and 69.

Shannon, C. E. (1948). *A Mathematical Theory of Communication*. CSLI Publications. Cited on page 84.

Shoemark, P., Liza, F. F., Nguyen, D., Hale, S., and McGillivray, B. (2019). Room to Glo: A systematic comparison of semantic change detection approaches with word embeddings. In *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 66–76, Hong Kong, China. Association for Computational Linguistics. Cited on page 114.

Shwartz, V., Santus, E., and Schlechtweg, D. (2017). Hypernyms under siege: Linguistically-motivated artillery for hypernymy detection. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, Valencia, Spain*, pages 65–75. Cited on page 106.

Soares da Silva, A. (1992). Homonímia e polissemia: Análise sémica e teoria do campoléxico. In *Actas do XIX Congreso Internacional de Lingüística e Filoloxía Románicas*, volume 2 of *Lexicoloxía e Metalexicografía*, pages 257–287, La Coruña. Fundación Pedro Barrié de la Maza. Cited on pages 12, 22, and 32.

Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology*, 15:88–103. Cited on pages 48 and 89.

Språkbanken (downloaded in 2019). *The Kubhist Corpus, v2*. Department of Swedish, University of Gothenburg. Cited on page 42.

Steyvers, M. and Griffiths, T. (2007). *Probabilistic Topic Models*. Lawrence Erlbaum Associates. Cited on page 26.

Svenska Akademien (2009). Contemporary dictionary of the Swedish Academy. Cited on page 44.

Tahmasebi, N., Borin, L., and Jatowt, A. (2021). Survey of computational approaches to lexical semantic change detection. Cited on pages 1, 11, 12, 27, and 29.

Tahmasebi, N. and Risse, T. (2017). Finding individual word sense changes and their delay in appearance. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 741–749, Varna, Bulgaria. Cited on page 22.

Teodorescu, D., McIntosh von der Ohe, S., and Kondrak, G. (2022). UAlberta at LSCDiscovery: Lexical semantic change detection via word sense disambiguation. In *Proceedings of the 3rd International Workshop on Computational Approaches to Historical Language Change*, Dublin, Ireland. Association for Computational Linguistics. Cited on page 114.

Thagard, P. (1990). Concepts and conceptual change. *Synthese*, 82(2):255–274. Cited on page 25.

Tharwat, A. (2020). Classification assessment methods. *Applied Computing and Informatics*. Cited on pages 62 and 89.

Tsakalidis, A. and Liakata, M. (2020). Sequential modelling of the evolution of word representations for semantic change detection. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8485–8497, Online. Association for Computational Linguistics. Cited on page 134.

Tunc, B. (2021). Optimierung von Clustering von Wortverwendungsgraphen. Bachelor thesis, University of Stuttgart. Cited on page 135.

Turney, P. D. and Mohammad, S. M. (2019). The natural selection of words: Finding the features of fitness. *PLOS ONE*, 14(1):1–20. Cited on page 135.

Turney, P. D. and Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *J. Artif. Int. Res.*, 37(1):141–188. Cited on pages 25, 26, 69, and 77.

Ullmann, S. (1957). *The Principles of Semantics*. Glasgow University publications. Baresc & Noble. Cited on page 11.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. Cited on page 70.

Vilnis, L. and McCallum, A. (2015). Word representations via Gaussian embedding. In *ICLR*. Cited on page 91.

Wang, X. and McCallum, A. (2006). Topics over time: A non-markov continuous-time model of topical trends. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '06, pages 424–433, New York, NY, USA. ACM. Cited on page 27.

Ward Jr, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301):236–244. Cited on page 74.

Weaver, W. (1949/1955). Translation. In Locke, W. N. and Boothe, A. D., editors, *Machine Translation of Languages*, pages 15–23. MIT Press, Cambridge, MA. Reprinted from a memorandum written by Weaver in 1949. Cited on page 21.

WGD (2021). Wörterbuch der deutschen Gegenwartssprache 1964–1977, curated and provided by the Digital Dictionary of German Language. `https://www.dwds.de/d/wb-wdg`. Accessed: 2021-01-07. Cited on pages 125 and 126.

Wijaya, D. T. and Yeniterzi, R. (2011). Understanding Semantic Change of Words over Centuries. In *Proceedings of the 2011 International Workshop on DETecting and Exploiting Cultural diversiTy on the Social Web*, DETECT '11, pages 35–40, New York, NY, USA. ACM. Cited on pages 27 and 28.

Wiktionary (2021). Wiktionary, das freie Wörterbuch. `https://de.wiktionary.org`. Accessed: 2021-01-07. Cited on page 125.

Wilks, Y., Fass, D., Guo, C.-M., McDonald, J. E., Plate, T., and Slator, B. M. (1990). Providing machine tractable dictionary tools. *Machine Translation*, 5(2):99–154. Cited on page 26.

Xu, Y. and Kemp, C. (2015). A Computational Evaluation of Two Laws of Semantic Change. In *Proceedings of the 37th Annual Meeting of the Cognitive Science Society, CogSci 2015*, Pasadena, California, USA. Cited on page 27.

Zamora-Reina, F. D., Bravo-Marquez, F., and Schlechtweg, D. (2022). LSCDiscovery: A shared task on semantic change discovery and detection in Spanish. In

Proceedings of the 3rd International Workshop on Computational Approaches to Historical Language Change, Dublin, Ireland. Association for Computational Linguistics. Cited on pages 67, 87, 114, 132, and 134.

Zhou, J. and Li, J. (2020). TemporalTeller at SemEval-2020 Task 1: Unsupervised Lexical Semantic Change Detection with Temporal Referencing. In *Proceedings of the 14th International Workshop on Semantic Evaluation*, Barcelona, Spain. Association for Computational Linguistics. Cited on page 92.

Typeset February 25, 2023