UNIVERSITÀ
DEGLI STUDI DI BARI
ALDO MORO

Department of Computer Science

PhD Programme in
Computer Science and Mathematics

XXXVI Cycle

Academic Discipline INF/01 - Informatics

A Thesis Submitted in Fulfillment
of the Requirements for the Degree of

Doctor of Philosophy

# COMPUTATIONAL APPROACHES TO LANGUAGE CHANGE

*PhD Candidate:*
Pierluigi Cassotti

*Coordinator:*
Prof. Francesca Mazzia

*Supervisor:*
Prof. Marco de Gemmis

*Cosupervisor:*
Prof. Pierpaolo Basile

Final Exam 2024

©2023, *Pierluigi* CASSOTTI

SWAP Research Laboratory
Department of Computer Science
University of Bari Aldo Moro

# Contents

# List of Figures

# List of Tables

UNIVERSITY OF BARI ALDO MORO

# *Abstract*

PhD Programme in Computer Science and Mathematics, XXXIII Cycle
Department of Computer Science

**COMPUTATIONAL APPROACHES TO LANGUAGE CHANGE**

by Pierluigi CASSOTTI

Language is the vital medium through which people communicate and express their needs. To make our communication meaningful, it is essential that we use language in the most effective and efficient way. As humans evolve, so too do languages, adapting to better encapsulate and convey information content. The underlying dynamics and mechanics of language change are often intricate and difficult to disentangle. Linguistic studies tend to focus on small samples, requiring high levels of skill and effort to consult and analyze thousands of historical documents.

The computational interpretation and processing of natural language is a complex endeavor. This complexity escalates when transitioning from a synchronic to a diachronic, or across-time dimension. In the past few years, there has been a significant upswing in interest towards computational strategies for understanding language change. This surge can be attributed to two converging phenomena: the remarkable growth in computational power and a large surge in the availability of textual data.

The core objective of this thesis is to delve deep into computational methods for understanding Language Change, with an emphasis on Lexical Semantic Change. This will be achieved by (i) systematically reviewing and comparing current state-of-the-approaches relying on Temporal Word Embeddings on different languages and benchmarks covering different historical periods (ii) designing and implementing novel models nurtured on synchronic data and evaluating their applicability in a diachronic context, (iii) devising datasets and tools to set performance standards for these models and further linguistic-aided analysis, and (iv) offering methodological perspectives on expansive, quantitative, longitudinal studies of Language Change, highlighting the critical points and the advantages to be gained from this type of analysis.

The initial section of this thesis elucidates foundational concepts from the field of Historical Linguistics and Natural Language Processing. Additionally, it offers a comprehensive overview of cutting-edge computational strategies geared towards understanding Language Change and their consequential applications in Culturomics. Then, the main contributions are presented and discussed in the last chapter of the thesis addressing the Research Questions introduced in the first chapter.

# Chapter 1

# Introduction

Languages constantly change over time, shaped by social, technological, cultural and political factors that influence how people express themselves. Language change is a phenomenon well-known in Linguistics and widely analysed in Historical Linguistics. Over time, tools and methods have been developed to analyse this phenomenon, including systematic categorisations of the types of change, the causes and the mechanisms underlying the different types of change. With the advent of the digital era and the resulting exponential increase in the amount of textual data, we have access to unprecedented opportunities to analyse and understand the complex patterns of language change.

Indeed, this vast amount of data can be used to feed computational models that, powered by the latest cutting-edge technological developments, offer a new paradigm in language change, characterised by large-scale quantitative studies. Traditional linguistic methods, while informative, are often based on small, carefully curated samples. Linguistic analysis using computational models not only speeds up our understanding of how languages change but also provides broader and more detailed insights, opening up the study of vast corpora - from historical archives to social media feeds.

Not all languages change at the same pace, and the same language can change at different rates over time. The increased ease and speed with which we can now communicate digitally, removing all barriers between linguistic and cultural communities, opens up new and novel scenarios of language change. Tracking change in a timely manner requires high-rate snapshots of the language and an explosion of time-sensitive combinatorial analysis that is no longer within the reach of manual inspection. The comparison order is magnified when, for example, social media is considered. The language of social media is characterised by rapid and changing trends, which can lead to an intensified rate of change in the lexicon, the number of changes in meaning, and the introduction of new syntactic structures.

Understanding language change carries profound implications that span different domains. Linguistic changes reflect evolving cultural landscapes and societies. When two cultures interact, they often exchange more than goods and services; they exchange words, reflecting historical interactions. For instance, the Maltese language has borrowed the words *Allah* and *Għid* from Arabic, nowadays used by the Christian communities to refer to God and Easter, respectively. Popular culture, including movies, music, and literature, can introduce and popularise new slang or terms. Words like *selfie* were popularised by their frequent use in pop culture and have since become a recognised part of the English lexicon. As societies evolve, so do their values. Once acceptable, words can become derogatory or outdated, leading to a change in their use or meaning. For example, the term *gay* historically meant *happy* or *joyful*, but its meaning has evolved over the past century, primarily denoting one's sexual orientation. Major social movements can lead to linguistic shifts. The feminist movement, for instance, has influenced language by promoting gender-neutral terms like *firefighter* instead of *fireman* and *chairperson* instead of *chairman*.

Languages worldwide, including Scottish Gaelic, Quechua, Ainu, and Komi, have experienced significant influences from dominant neighbouring languages such as English, Spanish, Japanese, and Russian respectively. As speakers of these languages become bilingual, often due to historical, political, or societal pressures, they incorporate vocabulary, sentence structures, and idioms from the dominant tongues. This borrowing is particularly evident in modern terminology related to technology, governance, and daily life, leading to an evolution of these languages and, in some cases, concerns about their preservation. Computational tools can map how languages interact and identify changes led by words or structures borrowed from dominant languages, highlighting areas where an endangered language may lose its distinctiveness. By studying language change, psychologists and sociologists can better understand group dynamics, identity formation, and even cognitive processes related to language comprehension and production.

The implications also affect the technological side. Natural Language Processing (NLP) is the branch of artificial intelligence concerned with developing tools that enable computers to understand, interpret and generate human language. NLP is concerned with transforming human language into a form that machines can understand, involving tasks such as tokenisation (breaking down text into words or phrases), parsing (identifying the underlying syntactic structures of sentences), and word sense discrimination (recognising differences in the meanings of words).

The last decades has seen a shift from rule-based approaches to more data-driven methods, particularly with the advent of deep learning. Modern NLP models, often based on neural networks, can process vast amounts of textual data, learning patterns and structures without explicit programming. It has advanced exponentially over the past few decades, enabling machines to perform sentiment analysis, machine translation and chatbot interactions. While these achievements are laudable, an essential aspect of human language that remains a challenge for NLP is its inherent and constant change. As language changes, so must the computational models that attempt to understand it. Thus, the intersection of computational approaches and the study of language change becomes paramount.

NLP models trained on older data may become outdated, leading to reduced performance and inaccuracies, or conversely, models trained on modern data may be ineffective in handling historical data. Model adaptation is not straightforward; as models become larger, involving billions of parameters, updating requires strategies capable of learning new patterns without *forgetting* the stored ones [242]. To ensure that these models remain robust and relevant, it is crucial to understand how languages change over time and to develop techniques to adapt models accordingly.

## 1.1 Research Questions

The objective of this thesis is to advance the state of the art regarding the creation of Natural Language Processing tools and resources for Language Change, with particular regard to Lexical Semantic Change. Following, we report the core Research Questions will be answered in this thesis.

**RQ1. How do different models perform across diverse languages and datasets when subjected to benchmarks specifically designed for Lexical Semantic Change Detection?** We aim to evaluate the efficacy of different computational models for Lexical Semantic Change Detection (LSCD). Exposing these models to benchmarks tailored specifically for LSCD, we can gain insights into their performance metrics across different languages and datasets. The underlying objective is to discern which models are universally effective, which ones excel in particular linguistic environments, and the potential reasons for any observed disparities in their performance.

RQ2. **To what extent are synchronic models equipped to understand and track diachronic language changes?** While synchronic models are not primarily designed to work on historical corpora, we can explore their adaptability and efficiency in tracking language change that occur over time and consequently their efficacy on historical corpora. We seek to understand the inherent capacities and potential limitations of synchronic models when applied to diachronic analysis. The overarching goal is to determine if, and how, these models can be harnessed or modified to study language change. In this thesis a synchronic model is presented that obtain state-of-the-art performance for the Lexical Semantic Change Detection task.

RQ3. **How can benchmarks and resources for Language Change be effectively designed to ensure comprehensive and accurate results?** The development of reliable benchmarks and resources is crucial for advancing research in language change. This question delves into the methodologies, criteria, and best practices for designing these tools. It emphasizes the need for comprehensive coverage, ensuring that benchmarks encompass diverse linguistic phenomena and datasets, and that the results generated are both accurate and replicable. This exploration will also consider potential pitfalls and challenges in benchmark design and how they can be circumvented.

RQ4. **How effectively can large-scale longitudinal quantitative studies capture and quantify the influence of socio-cultural events on language change over time?** Computational approaches to Language Change unveil a new paradigm of Language Change analysis, allowing for large-scale longitudinal quantitative studies. From a technical perspective this kind of studies provide challenging aspects requiring to analyse complex interactions between language and social-cultural events underlying the history. We aim to understand the methodologies best suited for analyzing such vast amounts of data and the potential challenges, such as nuances or biases. We asses this by testing the feasibility and value of using newspaper archives as a primary resource for studying language change in the context of societal and cultural shifts.

## 1.2 Contributions

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

**Temporal Aligned Language Models** Various approaches based on the temporal alignment of language models for Lexical Semantic Change Detection have been proposed over the years. In the first contribution, we propose a systematic classification of the approaches proposed in the literature and analyse the performance of these models on two benchmarks involving five languages: English, Swedish, German, Latin and Italian. Furthermore, we evaluated graded Lexical Semantic Change Models using thresholds based on the Gaussian distribution of the cosine similarity. We considered several models: Dynamic Word Embeddings, Temporal Random Indexing, Temporal Referencing, OP-SGNS and Temporal Word Embeddings with a Compass. The review of the approaches and the results obtained are reported in [41], published in the Italian Journal of Computational Linguistics (IJCoL).

A second contribution describe the system we proposed to the shared task SemEval-2020 Task 1: Unsupervised Lexical Semantic Change Detection [43]. The system reached the first place in the binary SubTask for Swedish. We reported a comparison of some of the most recent approaches to model Lexical Semantic Change with Temporal Word Embeddings, and we experimented with an automatic unsupervised procedure to classify changing and stable words. Results show that implicit alignment works generally better in modelling the lexical semantic change.

Finally, the last contribution addresses a longitudinal study for the analysis of the Lexical Semantic Change, involving several models, including Dynamic Word Embeddings, Dynamic Bernoulli Embeddings, Procrustes and Temporal Random Indexing. The results indicate that detecting lexical semantic changes is a intricate endeavour. Systems identify a significant number of change points, which in turn impacts performance. A qualitative examination of word time-series reveals that certain change points are identified slightly before or after the precise time frame. This observation necessitates deeper linguistic analysis to comprehend the underlying causes. The work [39] was published in the proceedings of the 4th Workshop on Natural Language for Artificial Intelligence (NL4AI 2020).

**Linguistic Knowledge Graph Databases** Graph databases are a straightforward schema-less technology for storing knowledge graphs. We exploit the GraphBRAIN Schema (GBS) format to describe a new time-sensitive Linguistic Knowledge in a graph database.

We first introduce a new time-sensitive model of linguistic knowledge based on graph databases, the work was presented at the 1st Workshop on

Artificial Intelligence for Cultural Heritage (AI4CH 2022) [15]. This model can be used to investigate word histories and conduct etymological research, as well as the analysis of quantitative patterns in the distribution of word senses not only across time (semantic change), but also according to the authors of the texts and other textual features (semantic variation). Moreover, the model's ability to connect word meaning instances in texts with lexical concepts also enables applications in the growing area of word sense disambiguation from historical texts, which aims to associate the most likely meaning of a word to an instance of usage of that word in a historical text [21, 141].

Then, we introduce an application of the proposed LKG for Latin data. Focusing on the case of Latin, a high-resource language among historical languages, we present initial results from integrating Latin corpus data, Latin WordNet, and Wikidata into a graph database via a GraphBRAIN Schema and show the potential offered by this model for diachronic semantic research. Differently from previous approaches, it gives rise to explainable results since we take advantage of explicit relationships modelled as graphs. The outcomes and the methodology used is desribed in McGillivray et al. [154] and McGillivray et al. [153] published respectively in the proceedings of the 19th Italian Research Conference on Digital Libraries (IRCDL 2023) and in the proceedings of the 4th Conference on Language, Data and Knowledge (LDK 2023).

**Benchmarking Unsupervised Lexical Semantic Change Detection** The first contribution concerns the presentation of an Italian diachronic corpus based on the newspaper "L'Unità". The corpus spans 67 years (1948-2014) and is provided both in plain text and in an annotated format that includes PoS-tags, lemmas, named entities, and syntactic dependencies. The corpus and the pre-computed data represent a valuable source of information both for linguists and researchers interested in diachronic analysis of the Italian language, and for historians, political scientists, and journalists as a digital resource enriched with automatic text analysis technologies. The paper [14] is published in the proceedings of the Seventh Italian Conference on Computational Linguistics (CLiC-it 2020) and introduces the methodology used for the collection and processing of textual documents. In particular, important steps in the processing of historical corpora are introduced, including the handling of possible OCR errors.

Another contribution concerns the construction of a benchmark for the detection of lexical semantic change for the Italian language. The goal of the benchmark is to develop systems able to automatically detect if a given word has changed its meaning over time, given contextual information from the corpora. We created two corpora for two different time periods $T_1$ and $T_2$, and we manually annotated a set of target words that change/do not change meaning across these two periods. Which work presents details concerning the selection of candidate target words, the finding of the uses of these words and the annotation of the meaning of the uses. An annotation method is then introduced that differs from those found in the literature and is shown to be effective. This is evident from the fact that the models obtain better results when applied to this benchmark compared to other benchmarks. The benchmark was used to evaluate systems at the shared task DIACR-Ita, hosted in the Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2020) [18].

We introduce Diachronic Engine (DE), a tool for the analysis of Lexical Semantic Change. The paper was presented in the Seventh Italian Conference on Computational Linguistics (CLiC-it 2020) [40]. DE integrates and extends current tools for corpus analysis enabling the study of corpus diachronic features. DE includes tools not included in other systems, such as time-series and change points detection based on state-of-the-art models for the analysis of semantic change. The tool is foundational in analyses based on traditional historical linguistics approaches, offering the possibility of investigating changes in detail through the inspection of concordances.

**Lexical Semantic Change Detection** The first contribution introduce XL-LEXEME, a model for LSC Detection. XL-LEXEME is pre-trained on a large WiC dataset to mirror sentence-level encoders focusing on specific words in contexts. We evaluated our model on two Lexical Semantic Change Detection datasets: SemEval-2020 Task 1 and RuShiftEval. XL-LEXEME outperforms state-of-the-art models for LSC Detection in English, German, Swedish, and Russian datasets, with significant differences from the baselines. The XL-LEXEME effectiveness and efficiency make it reliable for LSC Detection on large diachronic corpora and has become the defacto standard to beat in LSCD. The model marks a new direction in the field of the study of lexical meaning change by achieving significantly higher levels of correlation with the test set than approaches so far presented in the literature. The model demonstrates its effectiveness although it was trained on modern data, supporting the thesis that synchronic models are effective in recognizing language change. The paper describing XL-LEXEME [45] has been published in the proceedings of 61st Annual Meeting of the Association for Computational Linguistics (ACL 2023).

A second contribution concerns an analysis of contextual models with particular emphasis in studying how much information learned during training affects the relationship between the target word and its context. We conduct analysis via a replacement schema, which generates replacement sets with graded lexical relatedness, allowing examination of the models' degree of contextualisation. Using this schema, we also propose a novel approach to lexical semantic change detection and are able to surpass the results achieved by existing state-of-the-art models in the task of LSC. The replacement schema gives us an automatic way of providing labels for the change that has occurred, offering us a way to do explainable semantic change detection. The paper was submitted and currently under review to the 18th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2024).

**Computational Social Science and Cultural Analytics** The use of automatic methods for the study of lexical semantic change (LSC) has led to the creation of evaluation benchmarks. Benchmark datasets, however, are intimately tied to the corpus used for their creation questioning their reliability as well as the robustness of automatic methods. This contribution investigates these aspects showing the impact of unforeseen social and cultural dimensions. We also identify a set of additional issues (OCR quality, named entities) that impact the performance of the automatic methods, especially when used to discover LSC. The paper describing the experiments and the results were published in [19].

The last contribution investigates the usage of gender-specific forms of occupational titles in the Italian language in a diachronic corpus of 3 billion tokens extracted from two popular Italian newspapers. The work [42] was published in the proceedings of the Eighth Italian Conference on Computational Linguistics (CLiC-it 2021). The hypothesis is that the usage of gender-specific forms might be influenced by socio-cultural aspects. We automatically collect a set of occupational titles and perform a diachronic analysis exploiting the frequency of gender-specific forms. Results show a correlation between changes in the usage of gender-specific forms and socio-cultural events. Through this analysis, we show that there are significant changes in the way newspaper articles refer to the masculine and feminine form of an occupational title and that they are consistent with the occurring historical events, such as changes in the employment policy. Moreover, we performed a more fine-grained analysis by extracting the most influential figures that have guided this shift.

## 1.3 Outline of the thesis

The thesis outline is structured as follows:

**Chapter 2** introduces the reader to the basic concepts discussed in the rest of the thesis: those related to Historical Linguistics and Natural Language Processing.

**Chapter 3** contains a review of the state of the art related to the main design choices for developing Lexical Semantic Change Models. It discusses the evolution, the challenges, and the solutions that have been developed for Language Change, including state-of-the-art approaches and datasets for Lexical Semantic Change and applications of computational approaches to Language Change

**Chapter 4** through **8** discuss the contributions presented in Section 1.2 in detail.

**Chapter 9** concludes the thesis by summarizing the outcome of each contribution, provides an answer to each of the research questions presented in 1.1, and outlines lessons learned and avenues for future research.

# Chapter 2

# Background

## 2.1 Historical Linguistics

### 2.1.1 Introduction

Historical linguistics is a branch of linguistics that studies the history and development of languages. It looks at how languages change over time, analysing and describing different language communities and revealing patterns of change within languages and between language groups. It also includes the study of etymology, with the aim of reconstructing the origins of words. Historical linguistics describes and analyses changes of all kinds, e.g. phonological, syntactic, morphological and orthographic.

At the phonological level, language change can involve changes in the sounds that make up words. The Great Vowel Shift was a historical sound change that occurred in the English language between 1400 and 1700 and is considered one of the most important events in the history of English. It involved a shift in the pronunciation of long vowels and the silencing of some consonants [127]. Syntactically, language change can involve changes in the way words are put together to form sentences. For example, in English the word *do* has undergone a process of grammaticalisation in which verbs or nouns become grammatical markers. The use of *do* as an auxiliary verb began in the fourteenth century and was mainly used to form negations and questions [185].

At the semantic level, language change can involve changes in the meaning of words. For example, the word *graft*, initially used in the horticultural field, has over time been extended into medical terminology to refer to a surgical procedure. Finally, the most significant semantic shift is the use of *graft* to describe illegal or unethical practices, especially in politics and business. Morphological changes may involve changes in the way words are inflected to show grammatical relationships. For example, in Old English nouns were case inflected, but in Modern English nouns are no longer case inflected.

Historical linguistics has developed a variety of tools and methods to unravel the evolution of languages. At the heart of these investigations is the *comparative method*, a cornerstone of historical linguistics, which systematically compares languages to uncover common ancestry and trace semantic and phonological trajectories over time. *Internal reconstruction* complements this by examining inconsistencies within a single language, revealing historical layers and meanings that have shifted over time.

A powerful lens for understanding semantic change is the *semantic field theory*, which involves groups of words that are semantically related. This theory posits that the semantic fields of a word within a language are not static but dynamic, subject to cultural and social currents that can reshape entire semantic domains.

Glottochronology [215] is a method used by linguists to estimate the time when two languages diverged. The central idea behind glottochronology is that the core vocabulary of languages changes at a constant rate over time. By comparing the core vocabulary of two related languages and determining how much of this basic vocabulary they still share, linguists can calculate an approximate time at which the two languages began to diverge. The method is based on the Swadesh List, a list of words considered resistant to borrowing and change, such as personal pronouns, body parts and natural elements.

To map the relationships between languages and capture the complexity of linguistic evolution, linguists have developed models such as tree diagrams, which depict languages as branches growing from a common trunk representing their common proto-language. But the tree model, with its neat branches, sometimes oversimplifies the intricacy of language contact and borrowing. In response, the wave model captures the ripple effects of linguistic change emanating from multiple centres, recognising that languages can share features through contact, not just common ancestry. This model is particularly useful for illustrating the diffusion of grammatical and semantic features across geographical and linguistic boundaries.

The linkage model [72] goes a step further, illustrating the web of connections between languages and dialects within a language family. It recognises that languages do not simply diverge, but can converge and diverge again over time, forming a complex network of linguistic relationships that challenges the notion of linear, tree-like evolution.

### 2.1.2 Semantic Change

#### Onomasiological and semasiological change

Two main approaches to understanding semantic change are semasiological and onomasiological perspectives. Semasiological change focuses on how meaning changes while form remains relatively constant, but allows for phonological or morphosyntactic change. In contrast, onomasiological change focuses on changes in the linguistic forms used to express a particular concept. Lexical replacement is a common onomasiological process in which a new word or phrase is created to replace an existing one. For example, the Old English term "eorþe" gave way to the Middle English *world*, demonstrating lexical replacement reflecting changes in language and worldview.

Although semasiological and onomasiological changes are often discussed separately, they are inherently related. Semasiological changes can lead to onomasiological changes and vice versa, highlighting the dynamic relationship between meaning and form in language evolution. Consider the change of the term *gay* from its original meaning of *happy* to its contemporary sense of homosexuality. This semasiological change has led to onomasiological changes, with the emergence of new terms such as *homosexual* to refer specifically to same-sex attraction.

#### Taxonomies of semantic change

Several different taxonomies of semantic change have been proposed in the linguistic literature, each of which sheds light on different facets of how meanings evolve over time. The earliest structured attempt to classify semantic change can be traced back to the work of Reisig [184], who provided a distinction of semantic change based on *synecdoche*, which deals with shifts between part and whole; *metonymy*, which deals with shifts between cause and effect; and *metaphor*, where meanings shift based on perceived similarities.

Building on the early foundations, Paul [167] presented a more nuanced classification, recognising generalisation, where the meaning of a word is broadened, and specialisation, where it is narrowed. Darmesteter [53]'s contributions further refined the understanding of metaphor and metonymy, introducing the concept of narrowing and widening of meaning. This expansion and contraction of the word is expressed in terms of the change between whole and part. Bréal [33] extended these ideas by discussing the narrowing of meaning, where the meaning of a word becomes more specific, and the widening of meaning, where the opposite occurs. He also discussed metaphor and introduced the concept of *thickening* of sense, where meanings shift from the abstract to the more concrete. Stern [212]'s taxonomy is more comprehensive, including terms such as substitution, analogy, truncation and nomination, among others. This approach showed a keen understanding of the multifaceted nature of semantic change, recognising that factors such as changes in object use, knowledge and social attitudes all play a role. Bloomfield [30]'s taxonomy is one of the most widely recognised in the English-speaking academic world. It includes narrowing and widening, metaphor, metonymy and synecdoche, as well as hyperbole and meiosis, which deal with changes in the intensity of meaning. Bloomfield's inclusion of degeneration and elevation shows an awareness of the value judgements often attached to semantic shifts. Ullmann [227]'s distinction between the nature and consequences of semantic change represented a more analytical approach, recognising the effects of metaphor and metonymy as well as processes such as folk etymology and ellipsis. His taxonomy also takes into account the consequences in terms of broadening or narrowing of meaning and the qualitative shifts of improvement and degradation.

Finally, Blank [29]'s categorisation, increasingly accepted in recent years, offers a sophisticated framework that includes metaphor, metonymy and synecdoche, as well as specialisation and generalisation. Blank introduces cohyponymic transfer, antiphrasis, auto-antonymy and auto-converse, offering a detailed view of the horizontal, vertical and even opposite shifts that words can undergo.

## 2.2 Natural Language Processing

### 2.2.1 Introduction

Natural Language Processing (NLP) refers to the analysis of natural language text to infer the lexical and semantic characteristics contained within it. NLP attempts to imbue the computer with language skills in order to design computer programs and systems to assist humans in linguistic tasks, such as automatic translators, spell checkers, document and knowledge management; to develop computer systems that use natural language to interact with humans in a natural way, to automatically extract information from texts or other media, and to dynamically extend its own linguistic competence.

Text processing is necessary for the realization of functionalities that bring the machine closer to man and that enable new ways of interaction with computers. The text processing takes place on different levels which are dependent on each other. The aim is to elaborate the language starting from a purely syntactic point of view up to a more semantic level. The semantic level depends on previous elaborations and generally cannot ignore the syntactic analysis of the text; therefore, the processing starts from a purely symbolic level (phonemes, letters) up to a more semantic level (meanings). Five basic levels can be identified:

- sounds and letters: everything related to the articulation and decoding of the sounds of a language
- lexicon and morphology: knowing the words of a language, their structure and organization
- syntax: composing words into complex constituents semantics: assigning meanings to simple and complex linguistic expressions
- pragmatics: using sentences in contexts, situations and ways appropriate to communicative purposes

The study of the first level is necessary for all those applications that must understand sounds and interpret them in letters (e.g. recognition of spoken language) or that must produce sounds starting from the written text. By lexicon and morphology we mean the identification of the words that make up the language, or the search for lemma and lexemes. By the lemma, we mean each of the entries to which the single definitions of a dictionary are dedicated and by lexeme the minimum unit with meaning. The morphological analysis deals with identifying the gender of a word (singular/plural) or the ways and times of a verb.

Syntax analysis aims to identify the parts of speech, i.e. part of speech: verbs, nouns, adjectives, adverbs, prepositions, pronouns, etc. identify groups of words that represent a single meaning: hot dog, look for identify the elementary parts of speech: the nominal parts and the verbal parts (shallow parsing) derive the complete parse tree (full parsing) Semantic analysis aims to add semantics to the identified parts-of-speech.

One of the common strategies is to associate a vector representation to the words, exploiting the distributional hypothesis: words that appear in the same contexts are more similar. A geometric space is created, which tends to bring the vectors of the words that appear in the same contexts closer together, e.g. in the resulting vector space vectors of similar words are close to each other. In the following sections, the basic operations of text processing will be illustrated in detail, e.g. NLP basic pipeline:

- Tokenization
- Stop-words removal
- Part-Of-Speech tagging
- Semantics modelling

**Tokenization** Tokenization is the process of dividing text into smaller units called tokens, and its complexity depends on the level of abstraction considered. A *graphic word* is defined as a continuous alphanumeric string separated by spaces, which may include hyphens and apostrophes but not other punctuation. When tokenization is based on graphic words, the task involves identifying all substrings separated by non-alphanumeric characters. For instance, "28/05/2005" would be split into three tokens ("28", "05", "2005"), losing the understanding that they collectively represent a date. The same issue arises for proper names and complex numbers. Implementing a more abstract tokenizer involves recognizing compound words with apostrophes or hyphens, as well as identifying dates, entities, and numbers using regular expressions. It also requires understanding the roles of non-alphanumeric characters, which can vary across languages. For example, superscripts differ in function between Italian and English. Additionally, language-specific strategies must be considered; German and some Asian languages lack spaces between words, making tokenization more challenging.

Modern tokenization techniques have evolved to address the challenges presented by various languages and complex text structures. Methods like SentencePiece [112] and Byte-Pair Encoding (BPE) [207] have gained popularity for their ability to handle a wide range of languages and tokenization tasks. SentencePiece is a data-driven approach that divides text into smaller units, treating entire sentences or subword units as tokens. This technique is particularly effective in languages with no clear word boundaries or where tokenization rules may be ambiguous. BPE, on the other hand, operates by iteratively merging frequent character sequences to create subword units.

**Stop-words removal** Not all tokens contribute to the informational content of the text. To address this, a technique called stopword removal is employed, involving the identification and removal of common, uninformative words like articles, prepositions, and conjunctions. These words can introduce noise into the representation of textual content.

By removing stop-words, the size of the dictionary is reduced by approximately 40%, leading to improved performance. This is because during the matching process between a query and a document, non-informative tokens are disregarded. Nevertheless, complete removal of all stop-words can be challenging, as some of them might contain information. In such cases, a common approach is to use a limited stop-word list and leverage language statistics to manage these words more effectively.

Creating a stop-word list typically involves a combination of linguistic knowledge and statistical analysis. The goal is to strike a balance between reducing noise and retaining essential information in text representations.

**Lemmatization and stemming** Another crucial text processing stages in Natural Language Processing are lemmatization and stemming. Lemmatization involves reducing words to their base or dictionary form, known as the lemma. This process helps to capture the core meaning of a word by removing inflections and variations. For example, the words *running*, *ran*, and *runs* would all be lemmatized to *run*. Stemming, on the other hand, aims to reduce words to their root or stem by removing common suffixes. While both lemmatization and stemming contribute to reducing the dimensionality of the text data and aiding in information retrieval, lemmatization generally produces more linguistically accurate results compared to the more aggressive stemming, which may sometimes result in non-dictionary words. The choice between these techniques depends on the specific requirements of a natural language processing task.

**Part-Of-Speech tagging** Part-of-Speech Tagging (PoS) is the process of assigning grammatical roles to words in a text. This involves associating each token in the text with its appropriate part of speech, based on a language-specific lexicon. When a word can take on multiple grammatical roles, it becomes ambiguous, such as the word "watches" in English, which can be a verb or a noun. Disambiguation is crucial, and it is achieved by analyzing the lexical context of the word within the text.

**Named Entity Recognition** Named Entity Recognition (NER) is another essential aspect of natural language processing. NER is the process of identifying and classifying named entities, such as names of people, organizations, locations, dates, and more, within a given text. NER plays a vital role in information extraction, text summarization, and question-answering systems, as it helps in pinpointing specific pieces of information within a text.

Modern NER and PoS Tagging approaches often leverage deep learning models, such as Recurrent Neural Networks (RNNs) [193] and transformer-based models like BERT [57], which are pretrained on large text corpora to recognize and categorize named entities with high accuracy.

### 2.2.2 Computational Lexical Semantics

**Distributional semantics**

Most of the Natural Language Processing algorithms that deal with semantics rely on the distributed hypothesis, as Firth puts it, "you shall know a word by the company it keeps " [71]. In Distributed Semantic Models (DSM), words are mapped to high dimensional vectors in a geometric space. The first DSMs were count-based, they compute word vectors by counting how many times a word appears in a context, sentence, paragraph or document, according to the chosen granularity. Early DSMs merely compute word co-occurences using weighting schemas such as tf-idf (term frequency and inverse document frequency) in order to improve the representation and overcome issues related to Zipf distribution of words [100]. Since this type of representation are affected by sparsity issues, dimensionality reduction techniques are introduced such as Latent Semantic Analysis (LSA) [126]. On the other hand, prediction-based models use a continuous representation of word embeddings to predict the probability distribution $P = (w_t|context) \quad \forall t \in V$ of a target word $w_t$ given the context words $context$, for all the words in the vocabulary $V$.

Neural networks started to become dominant in this field with Word2Vec [157], which was introduced mainly to increase the efficiency and scale the dimensionality reduction stage of traditional approaches keeping theoretically the same efficacy in the representation [130]. Following Word2Vec, research interest focuses on overcoming two main drawbacks of these kind of approaches: the collapsing word semantics in a single point and the absence of the modeling of word position in the sentence. Modeling sequences in Neural Networks always posed a challenge in dealing with arbitrary long sequences. For instance, Recurrent Neural Networks (RNNs) [193], which are usually employed to deal with sequences, suffer of a problem called vanishing gradient, i.e. the Long-short term Memory (LSTMs) [92] overcomes this issue exploiting a forget-gate mechanism. ELMo (Embeddings from Language Model) is an example of language model based on Bi-LSTM (Bidirectional LSTM). Althought, the LSTM improved the RNNs they still struggling with handling long-range dependency.

### Contextualized Representations

The attention mechanism represents a new paradigm to sequence modeling, allowing to overcome previous issues underlying RNN models. Transformer [228] is an encoder-decoder Neural Network architecture which exploits the attention mechanism to elaborate sequence-to-sequence tasks. While several different fields of AI benefit from the introduction of Transformers, NLP is the one which has seen a major escalation. BERT (Bidirectional Encoder Representations from Transformers) [58] is the first example of use of Trasformers for NLP tasks. In particular, BERT only uses the encoder component of the Transformer architecture and exploits two specific training objectives for language modeling: Masked Language Modeling (MLM) and Next Sentence Prediction (NSP). After BERT, several different variants were proposed, such as RoBERTa [138], DistilBERT [198], ALBERT [125], and XLM-RoBERTa [51].

While BERT-based models can solve multiple traditional NLP tasks, they are not suitable for generation tasks, i.e. sequence to sequence tasks, since they miss the decoder component of the original Transformer architecture. For instance, T5 [179] and BART [133] are extension of BERT-based models involving the autoregressive decoder as in the original Transformer architecture. The GPT (Generative Pre-trained Transformers) [34] instead only employ the autoregressive decoder. The capabilities of models like T5, BART and GPT are heavily influenced by the the size of the model itself, i.e. number of layers and consequently the number of parameters of the model. While the pretraining of these models can be expensive in terms of time and resources, results show that the larger the model, the higher the performance. Generative models can solve traditional NLP task in zero-shot or few-shot learning settings with results above state-of-the-art results. With the introduction of these models the prompting paradigm [136] become prevalent, which better suits for generative models.

### Sense modelling

Word Sense Disambiguation [163] is a Natural Language Processing task with a long history and is extremely interesting for the Computational Linguistics community. In Word Sense Disambiguation (WSD), the goal is to disambiguate each word occurrence assigning to it the correct sense from a predefined sense inventory, such as WordNet [158]. The strategies for tackling WSD can be categorized mainly into supervised methods, which treat it as a classification task where the algorithm learns from a labeled dataset to predict the correct sense of new instances, and knowledge-based approaches that rely on databases like WordNet to infer the meaning based on semantic relationships.

The introduction of contextualized models, such as BERT, allowing the representation of a word in different contexts, steers the research focus to new tasks, such as the Word in Context (WiC) task [172].

WSD and the WiC task are highly related: while the former models in an explicit way the relationship between the target word and its sense (taken from a predefined sense inventory), the latter reduces it to a binary task. The WiC task requires determining if a word occurring in two different sentences has the same meaning or not. In recent years, there has been a growing interest in the WiC task, demonstrated by the creation of several different resources and shared tasks covering more than 20 languages.

Several datasets for the WiC task have been proposed throughout the years: the first one [172] being the proposal of the WiC task, which also came along with a dataset but was limited to English. For this reason, it was followed by the XL-WiC [181] dataset which tried to tackle this issue by taking into account a total of 15 languages. Next, the MCL-WiC [144] was the first WiC dataset to introduce the Cross-lingual task. The main motivation behind this particular choice was to cover scenarios where systems have to deal with different languages simultaneously, further highlighting the importance of this task in real-world applications. With AM$^2$iCo [137], the main aim was to focus on low-resource languages and to ensure participating models must consider both the target word and the context to achieve good performances. Finally, in CoSimLex [9], the task is extended to *pairs* of words that appear in a shared context, and the goal is to determine to which degree they refer to the same concept. This is done to capture the word polysemy as well as the context-dependency of words.

The Lexical Substitution task model the word's senses using another perspective. Lexical substitution aims to generate words that can replace a given word in a textual context. For instance, the word *car* in a sentence can be replaced with *automobile*, *bike*, or *motor vehicle*, depending on the desired nuance.

The challenge for the Lexical Substitution task has been the scarcity of annotated data, which has limited the use of large pre-trained models in a supervised context. ALaSca (Automated approach for Large-Scale Lexical Substitution) [121] is a dataset proposed to address this limitation by generating large-scale datasets for English lexical substitution, enabling the full potential of neural architectures like transformers to be utilized for the task.

# Chapter 3

# State of the art in Computational Approaches to Semantic Change

## 3.1 Benchmarks for Lexical Semantic Change

Several data sets and tasks are used to evaluate LSC models. Common tasks in which LSC models are evaluated are: (i) solving temporal analogies, which consists of detecting word analogies across time slices; (ii) lexical semantic change detection in a fixed target set, which requires assigning a label (stable or changed) to each word in a predefined set; (iii) lexical semantic change ranking, which ranks a target set of words according to their degree of semantic change; (iv) lexical semantic change discovery.

In the following, we describe LSC datasets for binary classification, ranking and discovery. The statistics of the datasets are given in Table 3.1.

| **Language** | **Time periods** | **Diachronic Corpus** | **# targets** | **Reference** |
|---|---|---|---|---|
| EN | $C_1$: 1810 – 1860<br>$C_2$: 1960 – 2010 | $C_1$: CCOHA, $C_2$: CCOHA | 46 | [204] |
| SV | $C_1$: 1790 – 1830<br>$C_2$: 1895 – 1903 | $C_1$: Kubhist, $C_2$: Kubhist | 44 | [218] |
| DE | $C_1$: 1800 – 1899<br>$C_2$: 1946 – 1990 | $C_1$: DTA, $C_2$: BZ+ND | 50 | [203] |
| LA | $C_1$: 200 – 0<br>$C_2$: 0 – 2000 | $C_1$: LatinISE, $C_2$: LatinISE | 40 | [152] |
| ES | $C_1$: 1810 – 1906<br>$C_2$: 1994 – 2020 | $C_1$: PG, $C_2$: TED2013, NC, MultiUN, Europarl | 100 | [238] |
| RU | $C_1$: 1700 – 1916<br>$C_2$: 1918 – 1990<br>$C_3$: 1992 –2016 | $C_1$, $C_2$, $C_3$: RNC | 111 | [114] |
| NO | $C_1$: 1929 –1965<br>$C_2$: 1970 – 2013 | $C_1$: NBdigital, $C_2$: NBdigital | 40 | [119] |
| NO | $C_1$: 1980 – 1990<br>$C_2$: 2012 – 2019 | $C_1$: NBdigital, $C_2$: NAK | 40 | [119] |
| ZH | $C_1$: 1954 – 1978<br>$C_2$: 1979 – 2003 | $C_1$, $C_2$: People's Daily | 40 | [46] |

TABLE 3.1: LSC benchmarks for Graded Change Detection

**SemEval 2020 Task 1** is the first task on unsupervised lexical-semantic change detection in English, German, Swedish and Latin. SemEval 2020 Task 1 [205] addresses the lack of a systematic approach for the evaluation of automatic methods for diachronic language analysis by proposing a common evaluation framework consisting of two tasks and covering historical corpora written in four different languages, namely German [240, 221], English [4], Latin [150] and Swedish [31]. Given two corpora $C_1$ and $C_2$ for two periods $t_1$ and $t_2$, Subtask 1 requires participants to classify a set of target words into two categories: words that have lost or gained senses from $t_1$ to $t_2$ and words that have not, while Subtask 2 requires participants to rank the target words according to their degree of lexical-semantic change between the two periods. For the annotation process, target words are selected based on whether they changed in meaning over time, using historical and etymological dictionaries as references. The annotators, who are native speakers and university students, are then asked to judge the semantic relatedness of the sampled uses of the target words. Relatedness is determined on a scale from *identical* to *unrelated* based on the context of use [200]. Special attention is given to Latin, due to the lack of native speakers, using a different approach where annotators compared word usage with dictionary sense definitions. The labels for evaluation, both binary and graded, are derived from the sense frequency distributions of the target words. These distributions represented how often different senses of a word are used. An annotated graph $G(V, E)$ is obtained for each word, where the vertices $V$ represent the uses and the edges $E$ indicate the relatedness of pairs of uses. Each graph is clustered using correlation clustering [11] to create usage graphs representing the semantic relatedness between word uses over time. Change scores are then calculated from these graphs to determine the degree of semantic change for each word. Finally, teams are scored on the accuracy of their predictions against hidden labels for Subtask 1 and Spearman correlation for Subtask 2.

**RuShiftEval** [115] is a shared task for detecting semantic shifts in Russian. The goal of the task is to detect changes in the meaning of Russian words over time, using three subcorpora from different periods (pre-Soviet, Soviet and post-Soviet). The RuShiftEval dataset consists of 111 Russian nouns, and participants have to rank them according to the degree of meaning change observed in the three different periods. This shared task introduced two novel aspects with respect to SemEval 2020 Task 1 [205], namely the splitting of the annotated semantic change dataset into more than two time periods and a training set.

In particular, out of the 111 Russian nouns, 99 are used in the test set and 12 are provided in the development set. These nouns are manually annotated to assess the degree of change in their meaning in three time pairs: pre-Soviet to Soviet, Soviet to post-Soviet, and pre-Soviet to post-Soviet.

The annotation is crowdsourced, following the DuReL framework [200]. Annotators score two sentences containing a target word from different time periods. The scores range from 1 to 4, with 1 indicating that the senses are *unrelated* and 4 indicating that they are *identical*. The individual scores are then averaged to produce a mean semantic relatedness score, known as `COMPARE`, which reflects the degree of semantic change - closer to 1, the stronger the change.

The task is structured as a ranking problem; no binary decisions are made about whether a word changed its meaning. Each word had to be assigned three semantic change scores, one for each pair of time periods. Lower scores indicated stronger changes, while higher scores indicated weaker changes. The systems are evaluated using Spearman rank correlations between the system-generated word rankings and the manually annotated gold standard rankings.

**LSC Discovery** [239] is a shared task on the discovery and detection of semantic changes in Spanish. The aim of the task is to detect and discover semantic changes in the Spanish language. The task is divided into two phases: Graded Change Discovery and Binary Change Detection. It introduces a new approach by requiring predictions and evaluations for all the vocabulary words in the corpus, rather than just a pre-selected set of target words. The evaluation is carried out on two reference subcorpora, the Old and the Modern Corpus. The time reference is 1810-1906 for the Old Corpus and 1994-2020 for the Modern Corpus. The text is extracted from various sources, including Project Gutenberg and OPUS, and contains both raw and lemmatised versions of the texts.

Annotators, who are native Spanish speakers with diverse backgrounds, used the DURel framework to assess the semantic relatedness of word usage pairs. The results are represented in Word Usage Graphs (WUGs), which are then clustered to interpret changes in word senses over time. A total of 4385 words are considered for the graded change detection task, and a subset of 100 words are annotated for semantic change. Of these, 20 are discarded due to low inter-annotator agreement. The remaining 80 words are split into two subsets, i.e. 20 words are used for the development set and 60 for the evaluation set.

For Graded Change Discovery, participants ranked words from the diachronic corpus pair based on the degree of change between the two time periods. The actual degree of semantic change is measured by the Jensen-Shannon divergence between word sense frequency distributions derived from human-annotated word usage samples. For binary change detection, participants classified words into two categories: no change or change. The classification is based on whether the words have gained or lost sense between two time periods. The binary labels are determined from the frequency distributions of the word senses.

**NorDiaChange** [120] is the first dataset dedicated to the study of diachronic semantic change in Norwegian. The dataset contains two subsets of about 80 Norwegian nouns, manually annotated for graded semantic change. These subsets reflect changes over time periods that are important in Norwegian history, such as pre- and post-war events, the discovery of oil and gas, and technological advances. The dataset is developed using the DURel framework and is based on two large Norwegian historical corpora. The corpora used to support this dataset are the NBdigital corpus from the National Library of Norway and the Norwegian Newspaper corpus, which cover a range of texts including books, reports and news articles in both Bokmål and Nynorsk. The two subsets within NorDiaChange compare the periods 1929-1965 with 1970-2013 (Subset 1) and 1980-1990 with 2012-2019 (Subset 2).

The target words for annotation are chosen on the basis of the authors' linguistic intuition as native Norwegian speakers and existing linguistic research. These words are expected to have undergone semantic changes during the specified time periods. The annotation is carried out by three native Norwegian speakers with expertise in linguistics or language technology.

In a qualitative analysis of the annotated data, the annotators and authors review the semantic graphs and word usage clusters to determine which senses are clustered and how. This process highlighted several cases where word senses had shifted significantly, such as *stryk* shifting from *rapids* to *fail*, and *kanal* evolving from *channel* to include *TV and radio channels*. The dataset contains both binary and graded change scores for each word.

**ZhShiftEval** is the first dataset for assessing semantic change in Chinese, particularly in the context of Reform and Opening up, covering a period of

50 years in modern Chinese. The dataset is called ZhShiftEval and is based on the DURel framework, which involves the collection of human judgments to assess semantic change.

The corpus used for the dataset is derived from the People's Daily, one of the most popular newspapers in China, from the 1950s to the early 2000s. The dataset contains texts from two time-specific subcorpora, representing the pre- and post-reform and opening-up periods. The annotation word list contains 20 words: 10 that have changed meaning over time and 10 stable words. The words are chosen on the basis of frequency and changes reflected in the corpus, as suggested by linguistic references. For each target word, 20 usage pairs are randomly selected from the subcorpora. Annotation is carried out by five native speakers of Mandarin Chinese with linguistic backgrounds.

**Diachronic Word Usage Graphs** [202] is the largest resource for graded contextualised diachronic word meaning annotations in four languages: English, German, Swedish and Latin. This resource, based on 100,000 human semantic proximity judgments, allows the identification of changes in word usage graphs (WUGs) over time. The final dataset contains 168 diachronic WUGs (DWUGs) for these languages. Examples of DWuGs for English, German and Latin are shown in Figure 3.1 and Figure 3.2.

Data for annotation came from historical subcorpora specific to each language. English data come from the Corpus of Historical American English [4], German from the Deutsches Textarchiv and newspaper corpora [221, 240], Latin from the LatinISE corpus [150], and Swedish from the Kubhist corpus. Approximately half of the target words for each language are selected on the basis of changes described in etymological or historical dictionaries, and the other half are control counterparts. The annotators are native speakers and university students, some with a background in historical linguistics.

The annotation process involved constructing usage-usage graphs by sampling 100 usages of each target word in two different time periods. Annotators judged the semantic proximity between pairs of usages without knowing their time period, allowing the construction of usage-usage graphs through multiple rounds of annotation. The resource contains a significant number of annotated usage pairs, with English, German and Swedish having approximately 50% of usage pairs annotated by more than one annotator. For Latin, each word is mostly annotated by a single annotator.

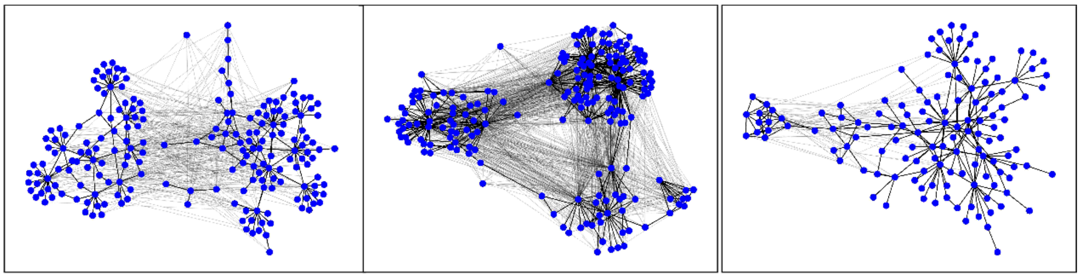


FIGURE 3.1: Usage-usage graphs of English plane (left), German ausspannen (middle) and Swedish ledning (right).

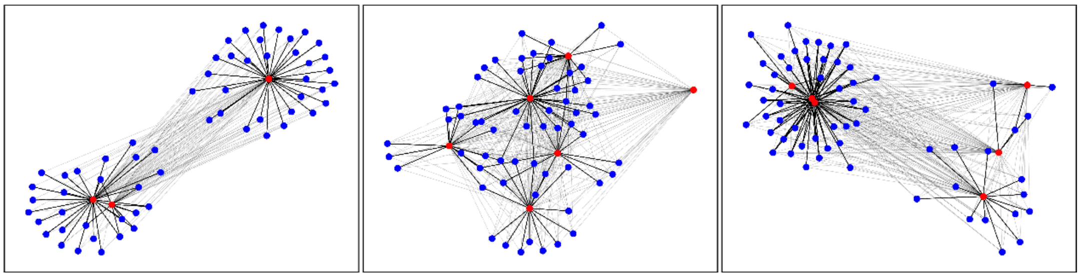


FIGURE 3.2: Usage-sense graphs of Latin pontifex (left), potestas (middle) and sacramentum (right). Nodes in blue/red represent usages/senses respectively

## 3.2 Models of meaning

### 3.2.1 Temporal Aligned Language Models

In general, Distributional Semantic Models (DSMs) approaches produce word vectors that are not comparable across time due to the stochastic nature of low-dimensional reduction techniques or sampling techniques. To overcome this issue a widely adopted approach is to align the spaces produced for each time period, based on the assumption that only few words change their meaning across time. Words that turn out to be not aligned after the alignment, changed their semantics.

Alignment models can be classified in post-alignment and jointly alignment models. *Post-alignment* models first train static word embeddings for each time slice and then align them. *Jointly Alignment* models train word embeddings and jointly align vectors across all time slices. Further, *Jointly Alignment* models can be distinguished in *Explicit alignment* models and *Implicit alignment* models. The objective function of *explicit* alignment models involves constraints on word vectors. Typically those constraints require that the distance of two-word vectors in two consecutive periods is the smallest possible. In the *implicit* alignment, there is no need for *explicit* constraint since the alignment is automatically performed by sharing the same word context vectors across all the time spans.

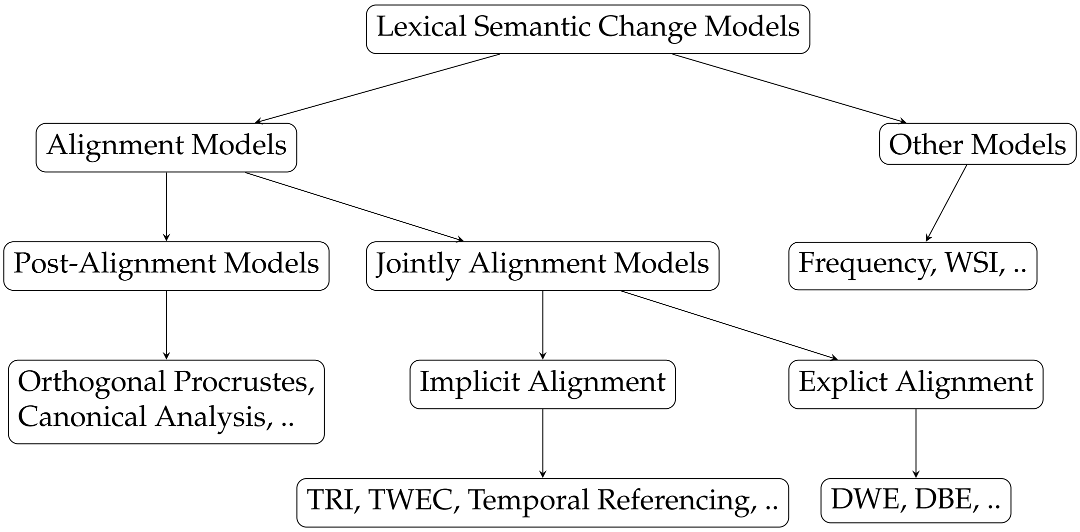


FIGURE 3.3: A classification of Lexical Semantic Change models.

**Post-alignment Models**

Orthogonal Procrustes (OP) [86] is a Post-alignment model, which aligns word embeddings with a rotation matrix. Word embeddings are computed using traditional approaches such as Singular Value Decomposition (SVD) of Positive Point-wise Mutual Information (PPMI) matrices, FastText [99] or Word2vec. The assumption of the OP method is that each word space has axes similar to the axes of the other word spaces, and two-word spaces are different due to a rotation of the axes. In this work, we use Skip-grams with Negative Sampling (SGNS) [157] to compute word embeddings and align them using Orthogonal Procrustes (OP-SGNS). In order to align SGNS word emebddings we compute the orthogonal matrix

$$R = \arg\min_{Q^T Q = I} \left\| QW^t - W^{t+1} \right\|_F$$

where $W^t$ and $W^{t+1}$ are two word spaces for time slices $t$ and $t+1$, respectively. We normalize the length of the matrices $W^t$ and $W^{t+1}$ and mean centre them. $Q$ is an orthogonal matrix that minimizes the Frobenius norm of the difference between $W^t$ and $W^{t+1}$. The aligned matrix is computed as

$$W^{align} = RW^t$$

### Jointly Alignment Models

Dynamic word embeddings (DWE) [236] is a Jointly Alignment Model. DWE is based on the PPMI matrix factorization. In a unique optimization function, DWE produces embeddings and tries to align explicitly them according to the following equation:

$$\min_{U(t)} \frac{1}{2} \left\| Y(t) - U(t)U(t)^T \right\|_F^2 + \frac{\lambda}{2} \left\| U(t) \right\|_F^2 +$$
$$\frac{\tau}{2} \left( \left\| U(t-1) - U(t) \right\|_F^2 + \left\| U(t) - U(t+1) \right\|_F^2 \right)$$

where the terms are, respectively, the factorization of the PPMI matrix $Y(t)$, a regularization term and the alignment constraint that keeps the word embeddings similar to the previous and the next word embeddings.

Temporal Word Embedding with a Compass (TWEC), Temporal Referencing (TR) and Temporal Random Indexing (TRI) are instances of *Jointly Implicit Alignment Models*.

TWEC [38] relies on the two Word2Vec models SGNS and CBOW. TWEC freezes the target and the context embeddings, respectively in CBOW and SGNS model across time, initializing them with the atemporal compass, i.e. word embeddings trained on the whole corpus. TWEC learn temporal specific word embeddings, training only the context or the target embeddings, respectively in CBOW and SGNS models across time.

TR [61] replace a subset of words in the dictionary (target words) with time-specific tokens. Temporal referencing is not performed when the word is considered a context word. Since TR is a generic framework, authors in [61] applied TR to both low-dimensional embeddings learned with SGNS and high-dimensional sparse PPMI vectors. In this work, we focus on the implementation based on SGNS (TR-SGNS). TR requires to fix a set of target words, this makes it impossible to compare words that are not in the target words set.

Finally, we investigate Temporal Random Indexing (TRI) [12] that is able to produce aligned word embeddings in a single step. Unlike previous approaches, TRI is a count-based method. TRI is based on Random Indexing [196], where a word vector (word embedding) $sv_j^{T_k}$ for the word $w_j$ at time $T_k$ is the sum of random vectors $r_i$ assigned to the co-occurring words taking into account only documents $d_l \in T_k$. Co-occurring words are defined as the set of $m$ words that precede and follow the word $w_j$. Random vectors are vectors initialized randomly and shared across all time slices so that word spaces are comparable.

With the increasing use of contextualized word embeddings, numerous approaches employing BERT-base models have been developed for LSC Detection [161, 123]. In TempoBERT [190], the authors exploit the concept of Masked Language Modeling (MLM), where the goal is to train a language model to predict a masked portion of text given the remaining part. In particular, they employ this technique to encode the concept of time into a BERT model. This is done by concatenating a specific token representing time to the text sequence. At inference time, TempoBERT can be used to predict the year of a sentence, masking the time reference, or to predict a masked token of the sentence conditioned by the time reference. In the same line of research, in Temporal Attention [191], the authors investigate the effect of modifying the model instead of the input sentence like in TempoBERT. This is done by extending the model's attention mechanism to consider the time when computing the weight of each word. The time dimension is encoded using a different query embedding matrix for each timestamp.

### 3.2.2 Synchronic Supervised Models

With the growing availability of models capable of producing effective contextualised representations, a number of innovative approaches to Lexical Semantic Change (LSC) Detection have been developed that make use of information derived from other tasks. This transfer learning strategy has yielded interesting results, contributing to the growing body of knowledge in the field.

GlossReader, as described in [178], is a striking example of this strategy. The core of GlossReader's methodology lies in the use of the XLM-R model [51], initially trained for Word Sense Disambiguation (WSD) on the English SemCor dataset [159] with glosses from WordNet 3.0 [158]. The authors take advantage of the model's zero-shot cross-lingual capabilities, allowing it to seamlessly transition to LSC Detection in the Russian language.

DeepMistake [7], another notable work into the field of LSC Detection, takes a slightly different route. In this approach, the authors turn to the Word-in-Context (WiC) task as an alternative to word sense disambiguation (WSD). First, a cross-encoder is trained on the MCL-WiC training and development dataset [144], using the XLM-R model as its foundational language model. This is followed by fine-tuning on the RuSemShift dataset [187], which allows the model to focus on the specific nuances of LSC detection in Russian.

Both systems participated in RuShiftEval and LSC Discovery. In both cases they achieved top rankings. In particular, both models show an effectiveness in the LSC Graded Detection task that is far superior to all other models, and in the LSC Discovery task they also achieve high results in the gain and loss of meaning detection tasks. Both GlossReader and DeepMistake exemplify the essence of task and language transfer learning in the context of LSC Detection. They represent innovative solutions that leverage pretrained models to tackle challenges beyond their original scope. The success of these approaches not only expands our understanding of LSC Detection but also underscores the versatility of modern natural language processing techniques in addressing complex linguistic phenomena.

Giulianelli et al. [78] propose a novel approach to automatically generate natural language definitions of contextualized word usages. This work introduce a specialized Flan-T5 language model for generating these definitions. By selecting the most prototypical definition in a usage cluster as the sense label, the aim is to make existing approaches to semantic change analysis more interpretable and allow users like historical linguists, lexicographers, or social scientists to explore and intuitively explain the diachronic trajectories of word meaning.

## 3.3 Applications

The interdisciplinary nature of computational modelling of semantic change has the potential to reshape our understanding of language evolution across diverse domains. From politics and law to science, literature, and societal issues, the application of computational tools enriches our insights into the dynamic nature of language and its role in shaping and reflecting societal changes. As researchers continue to explore new avenues for semantic analysis, the field is poised to make significant contributions to various academic disciplines and beyond.

One noteworthy exploration into the connections between political ideologies and language comes from Marjanen et al. [143]. Their work delves into the semantic shifts associated with "isms" such as liberalism, socialism, and conservatism, shedding light on the progression of political language throughout history. This research not only contributes to political discourse analysis but also showcases the applicability of computational models in uncovering subtle nuances in language use.

Turning to scientific writing, Bizzoni et al. [27] investigate changes in scientific discourse over time. By applying computational models, they uncover shifts in terminology and semantic structures within scientific literature. This research aids in understanding the dynamic nature of scientific language, an essential aspect for scholars and researchers in various scientific disciplines.

Haider and Eger [85] direct their focus towards poetry studies. Computational models enable them to analyze semantic changes in poetic language, providing a unique perspective on how meanings and connotations in literature evolve over time. This interdisciplinary approach showcases the versatility of computational tools in uncovering patterns within creative and expressive forms of communication.

Moving beyond the arts and humanities, computational modelling of semantic change has found application in addressing societal issues. Wevers [234] and Garg et al. [74] investigate the presence and evolution of gender biases and ethnic stereotypes in textual data. These studies contribute to a growing body of literature on social biases, demonstrating how computational tools can assist in identifying and understanding societal shifts reflected in language use.

The study conducted by Vylomova, Murphy, and Haslam [230] focuses on harm-related concepts within psychology. By examining the semantic transformations of terms like *addiction*, *bullying*, *harassment*, *prejudice*, and *trauma*, the researchers aim to determine if these concepts have broadened in scope over the past four decades. This research has implications for psychologists and mental health professionals, offering insights into the changing landscape of psychological discourse.

In a broader societal context, Tripodi et al. [225] trace the evolution and prevalence of antisemitic biases across various domains, including religion, economics, and socio-politics. Their data reveals an alarming rise in antisemitism, particularly in France, from the mid-80s onward. This study underscores the potential of computational modelling to unveil societal trends and prejudices embedded in language, providing a valuable tool for researchers and policymakers alike.

The Google Ngrams Dataset [79] is a dataset of n-grams extracted by 3.5 million books published between 1520 and 2008. Aiden and Michel [3] exploit the huge quantity of information contained in the Google Ngrams

Dataset to analyze the evolution of the language lexicon over time. In particular, the work offers interesting culturomics results, such as highlighting the spread of the term *influenza* during historical pandemic periods. Kutuzov, Velldal, and Øvrelid [117] exploit diachronic word embeddings to track wars and conflicts that took place from 1994 to 2010 all around the world. Diachronic word embeddings are trained on the English Gigaword news corpus [165] and used to predict conflict states: *peace*, *war* and *stable*. Laine and Watson [124] analyze the linguistic sexism occurring in *The Times* newspaper over five decades (1965-2005), relying on the classification of linguistic sexism proposed in [109]. The authors hypothesize that occupational titles and agents would be more resistant to change than other forms of sexism over the decades. They confirm their hypothesis by exploring the frequencies of masculine and feminine affixes, showing that they keep stable.

# Chapter 4

# Temporal Aligned Language Models

## 4.1 Analyzing Gaussian distribution of semantic shifts in Lexical Semantic Change Models

In this work, we focus on the Lexical Semantic Change Detection, using the data provided by both SemEval-2020 Task 1 Subtask 1 and DIACR-Ita. We compare several approaches: Temporal Random Indexing (TRI) [12], Temporal Word Embeddings with a Compass (TWEC) [38], Orthogonal Procrustes Alignment (OP) [86], Temporal Referencing (TR) [61] and Dynamic Word Embeddings (DWE) [236]. We evaluate all the models against both DIACR-Ita and SemEval-2020 Task 1 since some of these models, currently, have been evaluated in only one of the two tasks.

All the models evaluated in this work are graded, which means that they output a *degree* of semantic change. The degree of semantic change is typically expressed as the cosine between word vectors (embeddings) computed at different time, assuming that the lowest value of cosine similarity corresponds to the highest degree of change. A common strategy to map the degree of change to discrete stable/change label is:

- Compute the degree of change $\delta$ (cosine similarities) for each target word in the target set $T$, $\Sigma = \{\delta | w \in T\}$
- Compute the Gaussian $\mathcal{N}(\mu, \sigma)$ parameters of $\Sigma$
- Use $\mu, \sigma$ to assign a label to the target words (e.g. target words with degree of change less than $\mu - \sigma$ are labeled as change)

This work aims to get an overview of how thresholds based on the Gaussian parameters (e.g. $\mu - \sigma, \mu, \mu + \sigma$) work over different Lexical Semantic Change models and languages.

DIACR-Ita and SemEval-2020 Task 1 Subtask 1 require to assign a label (stable or changed) to each word in a predefined set. Most Lexical Semantic Change models produce graded scores that need to be labeled in one of the two classes. Choose a threshold is a crucial phase in binary classification since we need a strategy that should be independent by different Lexical Semantic Change models and languages. Systems that participated in SemEval-2020 Task 1 and DIACR-Ita employed several strategies to label graded scores (e.g. cosine similarities) obtained by Lexical Semantic Change Models.

The simplest approach is based on the idea that stable and changed words are equally distributed. In this case, it is possible to sort the words by the cosine similarity (in ascending order) and the first portion of the set is labelled as change. However, this is a weak approach since the equal distribution assumption does not fit the real-world.

Another common solution is to use an empirically chosen threshold, that, however, could be model-dependent. For instance, models such as DWE or TR produce smoothness changes than OP applied to vectors computed with Skip-grams with Negative Sampling [157]. In [23], authors use TWEC to compute word vectors and the *move* measure that is a linear combination of the cosine similarity and the similarity of local neighbourhoods. The authors empirically set the *move* threshold to 0.7. The system ranked 3rd in the DIACR-Ita task.

More advanced solutions involve unsupervised approaches to compute the threshold. In [43], target words are clustered using Gaussian Mixture Clustering [96] to form two clusters: the cluster of change targets and the cluster of stable targets. TRI with Gaussian Mixture Clustering ranked 1st in SemEval-2020 Task 1 Subtask 1 for the Swedish language. In [243] authors hypothesize that the target words cosine distances follow a Gamma distribution. Target words at the peak are classified as stable, while those at the tail are classified as change.

In [175] and [174] SGNS vectors are aligned by exploiting Canonical Analysis [88] and Orthogonal Procrustes [86] as Post-alignment models. The authors exploit two different thresholds over the cosine distances: the binary-threshold and the global threshold. The former is computed averaging the target cosine distances, while the latter is computed averaging over the binary-threshold computed for each language. The system based on the binary-threshold ranked 1st in both SemEval-2020 Task 1 Subtask 1 and DIACR-Ita. The experiments in [101], following the same approach in

DIACR-Ita, confirm the results obtained by [174].

In general, we can distinguish three different approaches used by systems proposed in SemEval-2020 Task 1 Subtask 1 and DIACR-Ita to compute thresholds by exploiting the change degree Gaussian distribution:

- Approach 1: Compute the Gaussian parameters over the target set.
- Approach 2: Compute the Gaussian parameters over all the dictionary.
- Approach 3: Compute the Gaussian parameters over the targets and get final thresholds averaging across different languages.

### 4.1.1 Data

In this work, we consider data coming from both SemEval and EVALITA.

SemEval-2020 Task 1 [205] comprises two tasks and covers corpora written in four different languages, namely German [240, 221], English [4], Latin [150], and Swedish [31]. Corpus statistics are reported in Table 4.1. Given two corpora $C_1$ and $C_2$ for two periods $t_1$ and $t_2$, Subtask 1 requires participants to classify a set of target words in two categories: words that have lost or gained senses from $t_1$ to $t_2$ and words that did not, while Subtask 2 requires participants to rank the target words according to their degree of lexical semantic change between the two periods.

DIACR-Ita focuses on the Unsupervised Lexical Semantic Change Detection for the Italian language. DIACR-Ita exploits the "L'Unità" corpus [18] that consist of two corpora $C_1$ and $C_2$. $C_1$ covers the period 1945-1970, while $C_2$ covers the period 1990-2014. An important aspect that distinguishes DIACR-Ita from SemEval is the annotation method. While, SemEval uses the DUREL framework for the annotation, DIACR-Ita relies on a sense-aware method guided by annotation retrieved by the Sabatini Coletti Dictionary [13]. The method consists of a selection and filtering of candidate words followed by manual annotation. The gold standard is obtained by checking that attested semantic change in the Sabatini Coletti dictionary is present in the training corpus.

### 4.1.2 Experimental setting

In order to estimates results, avoiding errors due to stochastic parameters initialization, we bootstrap ten runs for each model and language, respectively, averaging the results across the runs. We set the hyper-parameters according

| **Language** | **Corpus** | **Period** | **#Tokens** |
|---|---|---|---|
| English | CCOHA | 1810-1860 | 6.5M |
| English | CCOHA | 1960-2010 | 6.7M |
| German | DTA | 1800-1899 | 70.2M |
| German | BZ+ND | 1946-1990 | 72.3M |
| Swedish | Kubhist | 1790-1830 | 71.0M |
| Swedish | Kubhist | 1990-2014 | 110.0M |
| Latin | LatinISE | -200-0 | 1.7M |
| Latin | LatinISE | 0-2000 | 9.4M |

TABLE 4.1: SemEval-2020 Task 1 statistics.

| **Corpus** | **Period** | **#Tokens** |
|---|---|---|
| L'Unità | 1948-1970 | 52.2M |
| L'Unità | 1990-2014 | 196.5M |

TABLE 4.2: DIACR-Ita statistics.

to the findings of works proposed for DIACR-Ita and SemEval. For all the models, we set the number of iterations over the data to 5. In particular, for TWEC we set the number of static iterations to 3 and the number of dynamic iterations to 2.

We use a *context-window* of 5 for all the models. We set the number of *negatives* to 5 in all the models that use negative sampling. We set the vector dimension (*dim*) to 300 in all the models, except that for DWE. In DWE, we set the vector dimension *dim* to 100. We use a down-sampling (*sampling*) of 0.001 for all the models: TRI, TWEC, OP-SGNS and TR-SGNS. Table 4.3, reports models and hyper-parameters values. Where not specified, we adopt default values used by the authors of the models reported in SemEval or DIACR-Ita reports.

In particular, in DWE we specify the number of the alignment weight $\tau$, the regularization weights $\lambda$ and $\gamma$ as reported in Table 4.3. In TRI, we set the number of *seeds* to the default value 10.

| **DWE** | | **TRI** | | **TWEC** | | **OP-SGNS** | | **TR-SGNS** | |
|---|---|---|---|---|---|---|---|---|---|
| Param. | Value | Param. | Value | Param. | Value | Param. | Value | Param. | Value |
| dim | 100 | dim | 300 | dim | 300 | dim | 300 | dim | 300 |
| window | 5 | window | 5 | window | 5 | window | 5 | window | 5 |
| iter | 5 | iter | 5 | iter | 5 | iter | 5 | iter | 5 |
| $\lambda$ | 10 | sampling | 0.001 | sampling | 0.001 | sampling | 0.001 | sampling | 0.001 |
| $\gamma$ | 100 | seeds | 10 | negatives | 5 | negatives | 5 | negatives | 5 |
| $\tau$ | 50 | | | | | | | | |

TABLE 4.3: Models hyper-parameters.

| Model | English | German | Swedish | Latin | Italian | Model Avg. |
|---|---|---|---|---|---|---|
| TRI | .51±.18 | .48±.16 | .49±.17 | .63±.18 | .55±.24 | **.53±.18** |
| DWE | .86±.07 | .56±.17 | .66±.13 | .80±.08 | .56±.17 | **.69±.13** |
| TWEC | .65±.10 | .54±.12 | .56±.12 | .61±.10 | .59±.15 | **.59±.12** |
| OP-SGNS | .55±.14 | .41±.16 | .45±.14 | .50±.13 | .43±.21 | **.47±.16** |
| TR-SGNS | .48±.10 | .42±.11 | .42±.11 | .41±.08 | .50±.15 | **.45±.11** |
| **Language Avg.** | **.61±.12** | **.48±.15** | **.52±.13** | **.59±.12** | **.53±.18** | |

TABLE 4.4: Target words cosine similarities mean and standard deviation across different models and languages, computed on the target set.

| Model | English | German | Swedish | Latin | Italian | Model Avg. |
|---|---|---|---|---|---|---|
| TRI | .24±.22 | .34±.23 | .30±.20 | .25±.22 | .46±.26 | **.32±.22** |
| DWE | .72±.12 | .51±.16 | .47±.15 | .69±.11 | .50±.17 | **.58±.14** |
| TWEC | .69±.09 | .56±.10 | .54±.11 | .64±.10 | .63±.11 | **.61±.10** |
| OP-SGNS | .51±.16 | .40±.15 | .36±.17 | .44±.15 | .43±.17 | **.43±.16** |
| **Language Avg.** | **.54±.15** | **.45±.16** | **.42±.16** | **.51±.14** | **.50±.18** | |

TABLE 4.5: Target words cosine similarities mean and standard deviation across different models and languages, computed on the overall dictionary.

### 4.1.3 Results

In SemEval-2020 Task 1, systems are evaluated against three baselines. The Frequency Distance Baseline is based on the absolute difference of the normalized frequency in the two corpora as a measure of change. The Count Baseline implements the model described in [201], while the Majority Baseline always predicts the majority class. DIACR-Ita, as SemEval, provides the frequency distance baseline. Moreover, DIACR-Ita proposes the Collocations baseline. Collocations baseline, introduced in [13], computes the time-dependent representation of targets words using Bag-of-Collocations related to the two different periods. In this work, we use only the frequency baseline. In both SemEval and DIACR-Ita systems are evaluated using the Accuracy.

Tables 4.4 and 4.5 report, respectively, the statistics about cosine similarity over the target set and the overall dictionary[1]. The language average cosine computed on the target set is greater than the language average cosine computed on the overall dictionary, even when the target set consists of a greater

[1] TR-SGNS temporal-aware representations are available only for target words, for this reason it is not possible to compute the cosine similarities for the overall dictionary.

number of change words, as in the Latin language. It appears that the language average cosine computed on the target set is not correlated with the class balance reported in Table 4.8.

We test three Gaussian thresholds: $\mu - \sigma$, $\mu$, $\mu + \sigma$ computed over the target set for each language and for each model, as reported in Table 4.4. We plot the Accuracy obtained by each model, averaging over all the languages in Figure 4.1. The $\mu - \sigma$ threshold outperforms in every case the $\mu$ and $\mu + \sigma$ thresholds. We report results obtained in SemEval in Table 4.6, while results obtained in DIACR-Ita in Table 4.7 using the $\mu - \sigma$ threshold. Moreover, to test the efficacy of the Gaussian threshold, we compute the optimal threshold, maximising the accuracy, of $\lambda$ for each model and language. In particular, we test different values of $\lambda$ in order to find the optimal value that maximize the accuracy.

In DIACR-Ita task, all models outperform the baseline when $\mu - \sigma$ threshold is used. In SemEval, for the German and Swedish languages, the baseline obtains an accuracy very close to the considered models. This fact is more evident if we consider the optimal threshold. The accuracy obtained by any of the considered models with the Gaussian threshold remains above the accuracy obtained by the Baseline with the optimal threshold. In SemEval, the baseline with the optimal threshold outperforms all the models in combination with the Gaussian threshold in both Swedish and Latin languages.

An important consideration is that the target set of the DIACR-Ita task is smaller than about 50% of the English, German, Swedish and Latin target sets. On the other hand, the class balance of DIACR-Ita is very close to the class balance of German and Swedish languages in SemEval.

The class balance, reported in Table 4.8, may have affected the effectiveness of the used threshold. The $\mu - \sigma$ threshold never fits the optimal threshold. In particular, the accuracy of all the models using the $\mu - \sigma$ threshold decreases dramatically for the Latin language. We can hypothesize that the $\mu - \sigma$ threshold is affected by the unbalancing of the target set for the Latin language. The Latin language target set consists of only 35% of stable words. Some considerations for the Latin language:

- The target set for the Latin language consists of a greater number of change words rather than stable words, but most of the models rely on the assumption that only few words change their meaning, while the majority remain stable.

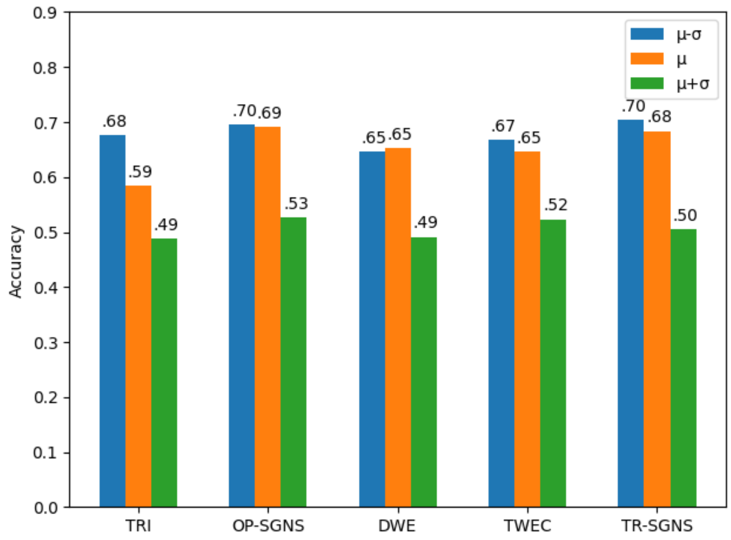


FIGURE 4.1: Models accuracy with different Gaussian thresholds: $\mu - \sigma$, $\mu$, $\mu + \sigma$ computed over the target set for each language and for each model. Accuracy is averaged across English, German, Swedish, Latin and Italian language.

- The Latin dataset is challenging, since the first corpus refers to the ancient Latin, while the second one refers to the Latin of the Catholic Church.

These peculiarities make it challenging to compare the results obtained in the Latin language against the other languages.

| | English | | German | | Swedish | | Latin | |
|---|---|---|---|---|---|---|---|---|
| **Model** | $\mu - \sigma$ | $\lambda$ | $\mu - \sigma$ | $\lambda$ | $\mu - \sigma$ | $\lambda$ | $\mu - \sigma$ | $\lambda$ |
| TRI | .65±.03 | .67±.02 | .65±.02 | .70±.04 | **.80±.02** | .83±.02 | .48±.01 | .66±.01 |
| DWE | .66±.03 | .69±.01 | .69±.02 | .73±.03 | .74±.02 | .81±.02 | .40±.02 | .67±.01 |
| TWEC | .65±.02 | .67±.01 | .74±.02 | .78±.02 | .74±.01 | .77±.00 | **.49±.03** | .70±.01 |
| OP-SGNS | .64±.02 | .66±.02 | .75±.02 | .80±.01 | .75±.03 | .79±.02 | .44±.02 | .69±.01 |
| TR-SGNS | **.71±.01** | .73±.02 | **.80±.01** | .87±.02 | .73±.02 | .79±.02 | .45±.02 | .70±.02 |
| Baseline | .62±.00 | .68±.00 | .65±.00 | .65±.00 | .74±.00 | .81±.00 | .35±.00 | .62±.00 |

TABLE 4.6: Accuracy obtained in SemEval-2020 Task 1 Subtask 1.

## 4.2 Gaussian Mixtures Cross-temporal similarity clustering

SemEval 2020 Task 1 [205] proposes a common evaluation framework that comprises two tasks and covers corpora written in four different languages,

| | Italian | |
|---|---|---|
| **Model** | $\mu - \sigma$ | $\lambda$ |
| TRI | .81±.04 | .83±.02 |
| DWE | .76±.04 | .84±.02 |
| TWEC | .73±.02 | .88±.02 |
| OP-SGNS | **.91±.04** | .96±.02 |
| TR-SGNS | .83±.00 | .95±.02 |
| Baseline | .67±.00 | .67±.00 |

TABLE 4.7: Accuracy obtained in DIACR-Ita.

| Language | Stable | Changed |
|---|---|---|
| English | 43% | 57% |
| German | 67% | 33% |
| Swedish | 74% | 26% |
| Latin | 35% | 65% |
| Italian | 67% | 33% |

TABLE 4.8: Classes balance for each language.

namely German [240, 221], English [4], Latin [150], and Swedish [31]. Given two corpora $C_1$ and $C_2$ for two periods $t_1$ and $t_2$, Subtask 1 requires participants to classify a set of target words in two categories: words that have lost or gained senses from $t_1$ to $t_2$ and words that did not, while Subtask 2 requires participants to rank the target words according to their degree of lexical semantic change between the two periods. We tackle the problem of automatically detecting lexical semantic changes with approaches that rely on temporal word embeddings. In this work, we focus on dynamic word embeddings by exploring methods based on both explicit, such as Dynamic Word2Vec [236], and implicit alignment, namely Temporal Random Indexing [12] and Temporal Referencing [61]. We analyse the use of different similarity measures to determine the extent of a word semantic change and compare the cosine similarity with Pearson Correlation and the neighborhood similarity [209]. While these similarity measures can be directly employed to generate a ranked list of words for Subtask 2, their adoption in Subtask 1 requires further manipulation. We introduce a new method to classify changing vs. stable words by clustering the target similarity distributions via Gaussian Mixture Models. We describe the embedding models and the clustering algorithm in Section 2, while Section 3 provides details about the hyper-parameter selection. Section 4 reports the results of the task evaluation followed by some concluding remarks in Section 5.

### 4.2.1 System description

We model the problem of automatic detection of semantic change by exploiting temporal word embeddings $E_i : w \rightarrow \mathbb{R}^d$ that project each word $w$ in the vocabulary $V$ into a $d$-dimensional semantic space. Given two different time periods $t_1$ and $t_2$, we create two embeddings $E_1$ and $E_2$. We investigate several models to compute temporal word embeddings:

**Dynamic Word2Vec (DW2V)** [236] simultaneously learns time-aware embeddings by aligning and reducing the dimensionality of time-binned Positive Point-wise Mutual Information matrices.

**Temporal Random Indexing (TRI)** [12] implicitly aligns co-occurrence matrices by using the same random projection for all the temporal bins.

**Collocations** extracts for each word and each time period the set of relevant collocations through the Dice score. As similarity function, we measure the cosine similarity between the sets of collocations belonging to the two different time periods. More details are reported in Basile, Semeraro, and Caputo [13].

**Temporal Referencing (TR)** [61] used only in the post-evaluation, it consists in a modified version of Word2Vec Skipgram that adds a temporal referencing to target vectors, keeping context vectors unchanged.

A similarity measure between vectors in the two temporal spaces is adopted to compute the extent of the semantic drift of the target words. We explored several similarity measures:

**Cosine similarity (CS)** is the cosine of the angle between two vectors.

**Pearson correlation (PC)** measures the linear correlation between two variables, in case of centred vectors (with zero means) is equivalent to the cosine similarity.

**Neighborhood similarity (NS)** computes two $k$-neighbour sets $nbrs_k(E_1(w))$ and $nbrs_k(E_2(w))$ and the union set $\mathcal{U} = nbrs_k(E_1(w)) \cup nbrs_k(E_2(w))$. Two second-order vectors, one for each word representation $u_j$, are created. The components of $u_i$ are the cosine similarity between the vector $v_j$[2] and the i-th element of $\mathcal{U}$: $u_{j_i} = cos(v_j,\mathcal{U}(i))$. The Neighborhood similarity is the cosine similarity between the second-order vectors. In all the experiments we set $k = 25$.

[2] Where $v_j$ is the vector representation for the word generated by $E_j$ and $j$ is the time period.

### Subtask 2

In Subtask 2, we use one of the three similarity measures ($CS$, $PC$, $NS$) to compute the set of target similarities $\mathcal{S} = \{sim(E_1(w), E_2(w)) \mid w \in T\}$. Then, we rank the target words according to the distance, computed as: $1- \mid sim(E_1(w), E_2(w)) \mid$.

### Subtask 1: Gaussian Mixture Clustering

Subtask 1 requires a further step: given $\mathcal{S}$, the set of target similarities, we need to predict the target labels. The aim is to assign either of the two classes, 0 (stable) or 1 (change), to each target word of a given language. Once we compute the set of target similarities $\mathcal{S}$, we want to find a way to assign the corresponding label. We assume that low similarities suggest changing words and high similarities indicate stable words.

Gaussian Mixture Models (GMMs) allow us to build probabilistic models for representing the Gaussian distribution of stable and changing targets. We use GMMs[3] to model the density of the distributions of the similarities of targets as a weighted sum of two Gaussian densities [96]:

$$f(\mathcal{S}) = \sum_{m=0}^{M} \pi_m \phi(\mathcal{S}|\mu_m, \Sigma_m) \tag{4.1}$$

where $M$ is the number of mixture components, $\phi(\mathcal{S}|\mu_m, \Sigma_m)$ is the Gaussian density with mean vector $\mu_m$ and covariance matrix $\Sigma_m$, and $\pi_m$ is the prior probability for the $m$-th component. Additional constraints can be applied to the covariance matrix in Eq. 4.1. In our experiments, we allow each component to have its own covariance matrix.

For our purpose, we speculate that the distribution of target similarities is a mixture of two densities, i.e. representing the stable and changing words. Consequently, we fix the number of the mixture components in the GMMs to two. We initially randomly assign a label (stable/changing) to each density distribution. Let $\mu_0$ and $\mu_1$ be the means of the two Gaussians associated with the “stable” and “changing” labels respectively. If $\mu_0 < \mu_1$ (i.e. the similarity mean of the distribution labelled as “stable” is lower than the mean of distribution labelled as “changing”), we invert the labels. Alg. 1 can be used to properly label each word of the target vocabulary.

[3] https://scikit-learn.org/stable/modules/generated/sklearn.mixture.GaussianMixture.html

```
input  : $\mathcal{S}$
output: labels
$\mathcal{N}(\mu_0, \sigma_0), \mathcal{N}(\mu_1, \sigma_1), labels \longleftarrow GaussianMixtures(\mathcal{S})$;
if $\mu_0 < \mu_1$ then
    $labels \longleftarrow 1 - labels$;
end
```

**Algorithm 1:** Assign labels

In order to set the best parameters for each language and model, we rely on the GMMs log likelihood, which is generally used for estimating the clusters quality:

$$\ell(\theta \mid \mathcal{S}) = log \sum_{m=0}^{M} \pi_m \phi(\mathcal{S} \mid \mu_m, \Sigma_m) \tag{4.2}$$

where $\theta$ are the parameters of the GMM. For each language, we select the best model configuration to submit at the challenge using the GMMs log likelihood $\ell(\theta \mid \mathcal{S})$. This means that hyper-parameters across different languages are tuned using GMMs log likelihood. We improperly use this approach for choosing parameters across different models (different sets of similarities $\mathcal{S}$), as we do not have validation set for tuning the parameters. We will investigate this limitation as future work. The selected models and hyper-parameters are reported in Tab. 4.9. In particular, we use cosine similarity, Pearson correlation and Neighborhood similarity for computing the targets similarities in $Overall_{CS}$, $Overall_{PC}$ and $Overall_{NS}$ runs, respectively. In $DW2V$ and $TRI$ runs we use always cosine similarity.

### 4.2.2 Experimental Setup

In all the runs, we do not pre-process data and we use a context window size of 5 while analyzing sentences. The $TR$ model[4] has been adopted into its original implementation[5], as the $TRI$[6] approach and $DW2V$[7] one. For runs involving $TRI$, we experimented with a varying vector size from $200$ to $1,000$. Moreover, we investigated (1) the initialization of the count matrix at time $j$ with the matrix at time $j-1$, (2) the contribution of positive-only projections, and (3) the application of PPMI weights, as explained in QasemiZadeh and Kallmeyer [177]. For $DW2V$, we use the parameter setting proposed in Yao et al. [236]. We set $\lambda = 10$, $\tau = 50$, $\gamma = 100$, $\rho = 50$ and

[4] We add this model during the post-evaluation.

[5] https://github.com/Garrafao/TemporalReferencing

[6] https://github.com/pippokill/tri

[7] https://github.com/yifan0sun/DynamicWord2Vec

experimented with a number of iterations from one to five. As vocabulary, we kept the top 50,000 most frequent tokens for both $TRI$ and $DW2V$. In the $TR$ runs, we set the vector size to $100$, and we experimented eight iterations for English and Latin, and four for German and Swedish. ytick pos=left, All the other parameters used for configuring the models are reported in Tab. 4.9.

| Run | Configuration | English | German | Latin | Swedish |
|---|---|---|---|---|---|
| $Overall_{CS}$ | Model | DW2V | Collocation | DW2V | DW2V |
| | Parameters | it=3 | - | it=3 | it=4 |
| $Overall_{PC}$ | Model | DW2V | DW2V | DW2V | DW2V |
| | Parameters | it=3 | it=4 | it=3 | it=4 |
| $Overall_{NS}$ | Model | DW2V | DW2V | DW2V | DW2V |
| | Parameters | it=3 | it=1 | it=3 | it=4 |
| $TRI$ | Parameters | k= 400<br>pw=False | k=1000<br>pw=True | k=1000<br>pw=True | k=1000<br>pw=True |
| $DW2V$ | Parameters | it=3 | it=4 | it=3 | it=4 |

TABLE 4.9: Hyper-parameters and models selected for each run. *it* is the number of iterations, *k* is the embedding size, *pw* the use of PPMI weights

### 4.2.3 Results

SemEval 2020 Task 1 provide three baselines, namely Freq. Baseline, which uses the absolute difference of the normalized frequency in the two corpora as a measure of change; Count Baseline, which implements the model described in [201]; and Maj. Baseline that always predicts the majority class. Tab. 4.10 reports the main results obtained by the different models. It shows the results obtained from the official submissions at the challenge and the results obtained by the $TR$ approach performed during the post-evaluation phase. The results obtained for Subtask 1 are reported using the accuracy metric, while for Subtask 2, the Spearman's rank-order correlation coefficients are used.

Considering the results of the evaluation phase, the models show inconsistent behaviors. $TRI$ showed the best performance when considering "all the languages" for both Subtasks, although in Subtask 1 it is not able to overcome *Count Baseline* and *Maj. Baseline*. Focusing on Subtask 1, if we consider each language in isolation, we see that $DW2V$ gives the best results for English[8] while $Overall_{PC}$ is our best system for German language, although it is not able to overcome *Count Baseline*. $Collocation$ is the best system for Latin (although outperformed by *Freq. Baseline*) while $TRI$ is our best system for Sweden language. In Subtask 2, the best English score was reported by $Overall_{NS}$. $Overall_{CS}$ ($Collocation$) performed the best in German language. For Latin and Sweden, $TRI$ provided the best results, and interestingly, it is one of the few systems that did not generate a negative correlation, although outperformed by $CountBaseline$ in Latin language.

At the end of the challenge, when the labelled test set was released, we performed more experiments reported in the *post-evaluation* row. In this phase, we run an additional system, $TR$, which outperformed all the previous reported approaches, including all baselines. The only exception is for Latin, in which for Subtask 1 *Freq. Baseline* achieves $0.650$ accuracy in comparison to $0.525$ of $TR$. Comparing $TR$ and $TRI$, which are both based on implicit alignment, the former is a prediction-based model while the latter is a count-based one. Moreover, $TR$ creates a temporal word embedding only for the target words rather than for the whole vocabulary. Consequently, this results in better word embeddings for all the words in the vocabulary that do not have a temporal reference, because they are represented by using all occurrences in $C_1$ and $C_2$. We suppose that these differences allow $TR$ to achieve better results than the other models.

Tab 4.11 reports the best results for each language among all participants to Task 1. UWB obtains the best result for German language, tied with Life-Language and RPI-Trust, and the best average result over all languages. Our official submission $TRI$ gives the best result in the Swedish language, whereas Jiaxin & Jian results first for Latin and NLPCR for English language. In Subtask 2, NLPCR and UWB obtain the best results for English and German languages respectively, confirming results obtained in Subtask 1. Concerning the Latin language, also Jiaxin & Jian confirm results obtained in Subtask 1, outperformed only by RPI-Trust, while in Swedish UWB obtain the best result. In general, each system achieved the best performance in one language while performing differently on the remaining others.

During the post-evaluation, we decided to investigate also the role of GMMs for class labeling (Sec. 4.2.1). We compared GMMs with semi-manual thresholds $\mu_S$, $\mu_S - \sigma_S$, $\mu_S + \sigma_S$ and Winsorizing [110] computing $\mu_{\mathbf{S}}$ and $\sigma_{\mathbf{S}}$

[8] Please, note that for EN, LA and SW $Overall_{CS}$ and $DW2V$ coincide

| | **Subtask 1** | | | | | **Subtask 2** | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **System** | **All Lang.** | **EN** | **DE** | **LA** | **SV** | **All Lang.** | **EN** | **DE** | **LA** | **SV** |
| $Freq.Baseline$ | 0.439 | 0.432 | 0.417 | ***0.650*** | 0.258 | -0.083 | -0.217 | 0.014 | 0.020 | -0.150 |
| $CountBaseline$ | *0.613* | 0.595 | *0.688* | 0.525 | 0.645 | 0.144 | 0.022 | 0.216 | *0.359* | -0.022 |
| $Maj.Baseline$ | 0.576 | 0.568 | 0.646 | 0.350 | 0.742 | NaN | NaN | NaN | NaN | NaN |
| $Overall_{CS}$ | 0.509 | *0.622* | 0.500 | 0.400 | 0.516 | 0.111 | 0.252 | *0.415* | -0.183 | 0.041 |
| $Overall_{PC}$ | 0.533 | 0.595 | 0.646 | 0.375 | 0.516 | 0.056 | 0.272 | 0.168 | -0.135 | -0.080 |
| $Overall_{NS}$ | 0.508 | 0.568 | 0.542 | 0.375 | 0.548 | 0.035 | *0.298* | -0.059 | -0.179 | 0.078 |
| $Collocation$ | 0.513 | 0.486 | 0.500 | 0.550 | 0.516 | 0.273 | 0.144 | *0.415* | 0.194 | 0.340 |
| $DW2V$ | 0.541 | *0.622* | 0.625 | 0.400 | 0.516 | 0.098 | 0.252 | 0.366 | -0.183 | -0.041 |
| $TRI^{*}$ | 0.554 | 0.486 | 0.479 | 0.475 | *0.774* | 0.296 | 0.211 | 0.337 | 0.253 | *0.385* |
| $TR$ *(post-eval.)* | **0.704** | **0.703** | **0.812** | 0.525 | **0.774** | **0.496** | **0.304** | **0.722** | **0.395** | **0.562** |

TABLE 4.10: Results obtained by our models during the official competition and during the post-evaluation phase. For the Subtask 1 the results represent the accuracy score. Spearman's rank-order correlation coefficients are used for the Subtask 2. $TRI^{*}$ is the official submission in the evaluation phase since it obtained the best score in the Subtask1.

on data provided for Subtask 1, where $\mu_{\mathcal{S}}$ and $\sigma_{\mathcal{S}}$ are the mean and the standard deviation computed on the similarity set $\mathcal{S}$. Figure 4.2 reports the different accuracy scores obtained by the five methods for the $TRI$, $Collocation$, $DW2V$, $TR$ approaches. The scores for the GMMs strategy are close to those obtained by $\mu_{\mathcal{S}}$ for TRI and Collocation. While GMMs outperforms $\mu_{\mathcal{S}} + \sigma_{\mathcal{S}}$ in every run, $\mu_{\mathcal{S}} - \sigma_{\mathcal{S}}$ seems to work better than GMMs except that in $TR$. Winsorizing works better than GMMs in $TRI$ and $Collocation$. GMMs outperforms Winsorizing in $DW2V$ and $TR$. These results are not clear enough to advocate for a specific threshold. Consequently, further analysis will be part of future work in order to understand what is the better threshold that

| | **Subtask 1** | | | | | **Subtask 2** | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **System** | **All Lang.** | **EN** | **DE** | **LA** | **SV** | **All Lang.** | **EN** | **DE** | **LA** | **SV** |
| $TRI$ | .55 | .49 | .48 | .47 | **.77** | .30 | .21 | .34 | .25 | .38 |
| NLPCR | .58 | **.73** | .54 | .45 | .61 | .29 | **.44** | .45 | .15 | .11 |
| UWB | **.69** | .62 | **.75** | .70 | .68 | .48 | .37 | **.70** | .25 | **.60** |
| Jiaxin & Jinan | .66 | .65 | .73 | **.70** | .58 | **.52** | .32 | .72 | .44 | .59 |
| Life-Language | .68 | .70 | **.75** | .55 | .74 | .22 | .30 | .21 | -.02 | .39 |
| RPI-Trust | .66 | .65 | **.75** | .50 | .74 | .43 | .23 | .52 | **.46** | .50 |

TABLE 4.11: Best results obtained in Subtask 1 for each language: TRI is compared with results submitted by all participants to the SemEval-2020 Task 1.

could be included in the GMMs process.

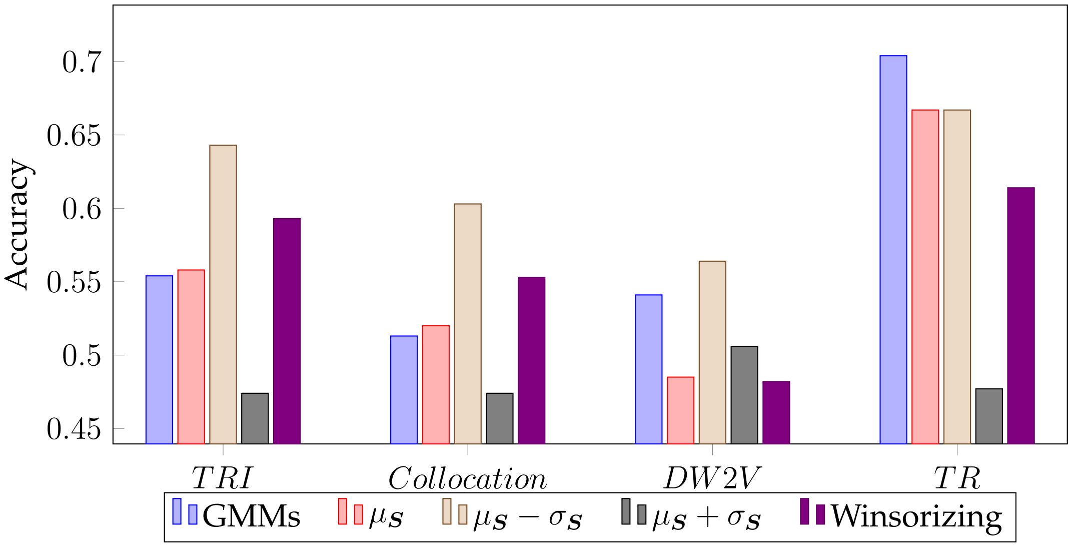


FIGURE 4.2: Accuracy scores in Subtask 1 using different class labeling strategies: GMMs, $\mu_S$, $\mu_S - \sigma_S$, $\mu_S + \sigma_S$ and Winsorizing using mean and standard deviation.

## 4.3 A Comparative Study of Approaches for the Diachronic Analysis of the Italian Language

In this work, we compare state-of-the-art approaches in computational historical linguistics, and we present the results of an in-depth analysis conducted using an Italian diachronic corpus. Specifically, several approaches based on both static embeddings and dynamic ones are implemented and evaluated by using the Kronos-It dataset. We train all word embeddings on the Italian Google n-gram corpus. The main result of the evaluation is that all approaches fail to significantly reduce the number of false-positive change points, which confirms that lexical semantic change is still a challenging task.

Previous works about the Italian Google Ngram corpus and Kronos-it are described in [17, 13], but they are limited to the Temporal Random Indexing model [12] and simple baselines based on word frequencies and collocations ignoring recent approaches based on word embeddings.

### 4.3.1 Methodology

Figure 4.3 shows the pipeline used for the evaluation, it consists of five modules: corpus pre-processing, computation of bins, bins alignment, construction of time-series and change point detection. The framework is written in Python, we adopt Procrustes[9], DBE[10], DWE[11] and TRI[12] using their original implementation.

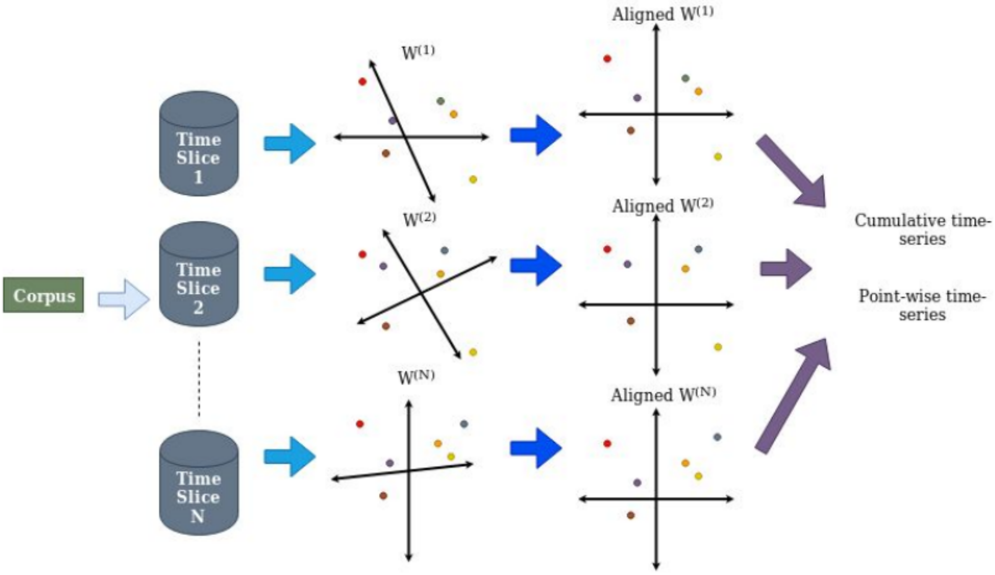


FIGURE 4.3: The evaluation pipeline.

#### Corpus pre-processing

The corpus pre-processing module receives as input a corpus annotated with the time label of each document. The first operation is the corpus splitting into temporal slices. During the splitting, the dictionary is computing by keeping track of each new token encountered and its occurrence. The final dictionary is built with all tokens present in each time slice and selecting the first $n$ tokens sorted by the number of occurrences. In our evaluation, we consider $n = 50,000$.

#### Bins building

The second module takes as input tokenized documents for each time slice and generates for each approach preliminary information useful for the next

[9] https://github.com/williamleif/histwords
[10] https://github.com/mariru/dynamic_bernoulli_embeddings
[11] https://github.com/yifan0sun/DynamicWord2Vec
[12] https://github.com/pippokill/tri

steps. It has an execution mode for each approach namely Word2Vec, PPMI, Static Bernoulli and Temporal Random Indexing. Word2Vec mode trains a Word2Vec model on each sub-corpus using Gensim[13], an open-source library for unsupervised topic modelling and natural language processing. The PPMI mode constructs a PPMI matrix for each time slice, which will then be used to create Dynamic Word Embedding. The Bernoulli mode builds static Bernoulli embedding for each time slice that will later be used to construct Dynamic Bernoulli embeddings. The Temporal Random Indexing mode saves the occurrences of words and contexts that we will later be used to create word embeddings.

### Alignment

The aim of the alignment module is the alignment of the bins produced as output in the previous module, and it is composed of several sub-modules: Procrustes Aligner, Bernoulli Aligner, Dynamic word embeddings construction and the TRI sub-module. The Bernoulli Aligner constructs Dynamic Bernoulli Embeddings starting from the static Bernoulli output. Procrustes Aligner is the sub-module that takes each Word2Vec model and applies Procrustes to each time slice. The Dynamic Word Embeddings sub-module takes the PPMI matrices previously created for building the Dynamic Word embeddings model. The TRI sub-module produces word vectors for each time slice by relying on the co-occurrences information built in the previous step.

### Time-series and change point detection

We compute time-series by exploiting the word embeddings created for each time slice. A time-series for each word is built, this result in a matrix $W^{VxT}$ where $V$ is the dictionary size and $T$ is the number of time slices.

We explore two approaches for the computation of the time-series, namely point-wise and cumulative. In the point-wise approach, the element $i, j$ of $W^{VxT}$ represent the cosine similarity

$$W_{i,j} = cos(v_{w_i}^{j-1}, v_{w_i}^{j})$$

[13] https://radimrehurek.com/gensim/

where $w_i$ is the i-th word in the dictionary and $j$ is the j-th time slice. While, in the cumulative approach, the element $i, j$ of $W$ is

$$W_{i,j} = cos(\frac{\sum_{k=1}^{j} v_{w_i}^{k-1}}{j}, v_{w_i}^{j})$$

In order to detect change points, we use the algorithm proposed in [220]. According to this model, we define a mean shift of a general time-series $W_i$ pivoted at time period $j$ as:

$$K(W_i) = \frac{1}{l-j} \sum_{k=j+1}^{l} W_{i,k} - \frac{1}{j} \sum_{k=1}^{j} W_{i,k} \tag{4.3}$$

To understand if a mean shift is statistically significant at time $j$ we use a bootstrapping [62] approach under the null hypothesis. The null hypothesis states there is no change in the mean. We sample $B$ bootstrap examples by permuting $W_{i,j}$. For each bootstrap sample P, $K(P)$ is calculated to provide its corresponding bootstrap statistic and statistical significance (p-value) of observing the mean shift at time $j$ compared to the null distribution. Finally, we estimate the change point by considering the time point $j$ with the minimum p-value score.

Change points together with the year, the p-value and the word are stored in a file used for the evaluation.

### 4.3.2 Evaluation

#### Data

For the training, we use the Google Ngram, a dataset of ngrams extracted by 305,763 Google Books. Google Ngram covers the period from 1500 to 2012. OCR errors can occur more in older historical documents, then we extract a sub-corpus concerning the period 1900-2010. We split Google Ngram corpus into ten slices with a range of ten years, starting from 1900 to 2010. We chose a time span of ten years for reducing the computational complexity since semantic changes are not frequent and generally require a large time span to be observed. Since the full text is not available in the Google Ngram, we use the method described in [77] for extracting co-occurrences between words. As gold standard, we use Kronos-it [13], a dataset for the Italian lexical change detection task. Kronos-it provides for each lemma a set of years indicating the semantic change for that lemma. Kronos-it is extracted by the Sabatini

| DWE | | TRI | | DBE | | Procrustes | |
|---|---|---|---|---|---|---|---|
| Parameter | Value | Parameter | Value | Parameter | Value | Parameter | Value |
| dimension | 300 | dimension | 300 | dimension | 300 | dimension | 300 |
| window | 4 | window | 4 | window | 4 | window | 4 |
| iters | 5 | down-sampling | 0.001 | negatives | 2 | min-count | 1 |
| $\lambda$ | 10 | seeds | 10 | minibatch | 1000 | negatives | 20 |
| $\gamma$ | 100 | | | n epochs | 4 | sample | 1e-5 |
| $\tau$ | 50 | | | | | iter | 4 |

TABLE 4.12: Models hyper-parameters.

Coletti, an Italian dictionary that contains for some word meanings the year of the first appearance. The Kronos-it dataset contains 13,818 lemmas and 13,932 change points. Lemmas reported in Kronos-it have, on average, one change point.

### Hyper-parameters

We use the same hyper-parameters values shared by two or more models. We use the same values for the *context-window* and the *dimension* of the embeddings. Table 4.12 reports training strategies and hyper-parameters values. We adopt default values used by the authors of the models.

In particular, in DWE we specify the number of *iterations* over the data, the alignment weight $\tau$, the regularization weights $\lambda$ and $\gamma$. In TRI, we set the *down sampling factor*, and the number of *seeds*. In DBE, we set the number of *negative samples*, the *minibatch* size and the *number of epochs*. In Procrustes, we set the minimum number of occurrences a token must have to appear in the dictionary *min-count*, the number of *negative samples*, the downsampling parameter *sample* and the number of *iterations* over the data.

### Metrics

We compute the performance of each approach by using Precision, Recall and F-measure. In the evaluation, a true positive is a change point for a word reported in the gold standard that belongs to the range of the ten years predicted by the system for that word. Change points provided by the systems are compared to the change points reported in the gold standard. The false negatives (FN) are the number of change points in the gold standard minus the true positives. The false positives (FP) are the number of change points provided by the system minus the true positives.

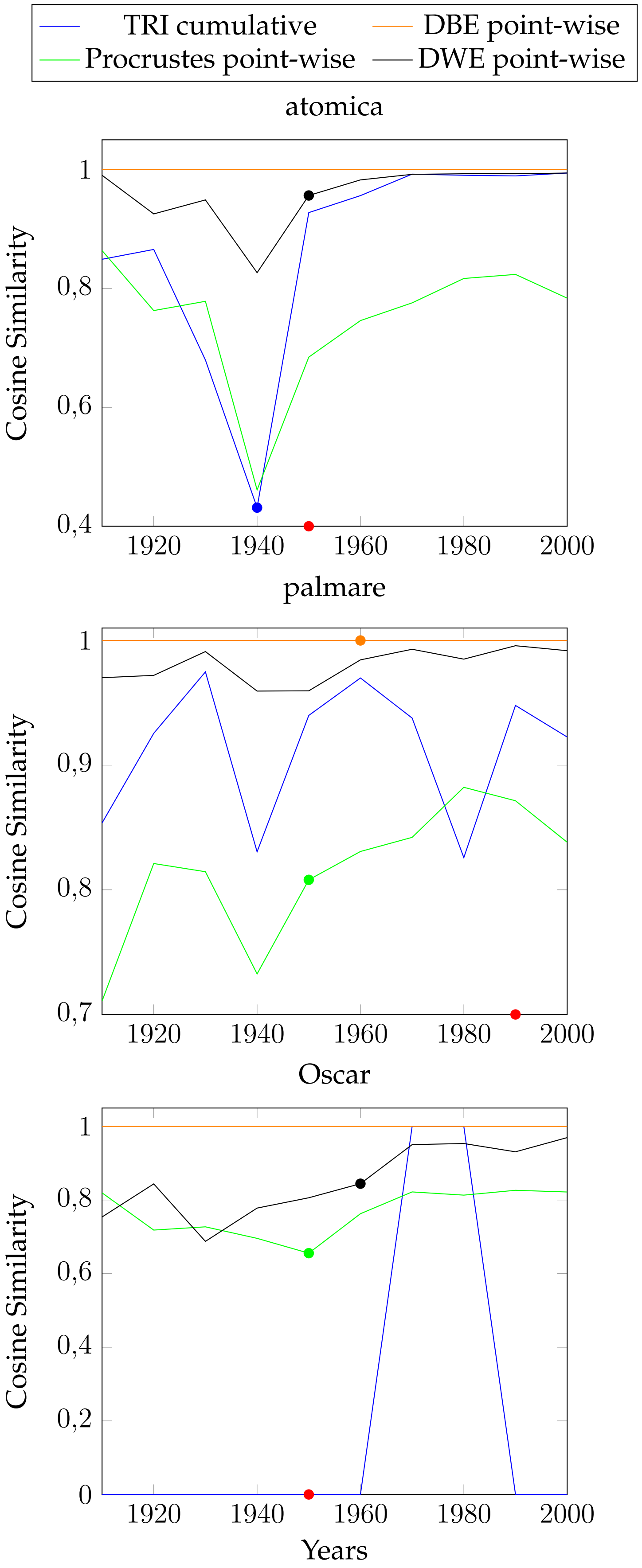


FIGURE 4.4: Example of semantic shifts detected. Red points marks change points in the gold standard. Change points detected in the time-series are shown.

### Results

Table 4.13 reports Precision (P), Recall (R) and F-measure (F) for each system. We can observe that generally, we obtain a low F-measure. This is due to the large numbers of change points detected by each system (false positive). We can observe that the best approach is DWE point-wise. However, the results of DWE point-wise are close to those obtained by Procrustes point-wise and TRI cumulative. A remarkable aspect is the worse performance of DBE respect those of TRI and DWE, the entries of DBE time-series are very close to 1, this highlights a heavy alignment. This is maybe due to the choice of hyper-parameters used to train the DBE. We use, as mentioned above, the default hyper-parameters and the type of datasets used by the authors is different from Google Ngrams, mainly due to the large amount of data in the Google Ngrams. This could have affected results obtained by DBE. The results of the evaluation prove that the task of semantic change detection is very challenging, in particular, the large number of detected change points (false positive) drastically affects the performance. Sometimes change points are detected before or after the change point reported in the gold standard, this supports the hypothesis that the change of semantics of a word is a continuous process, which involves long periods before reaching a stabilization. More studies are necessary to understand which component affects the performance, such an in-depth and explicit analysis of time-series. Moreover, it is important to underline that the year reported in the dictionary may be incorrect.

In Figure 2, we show some examples of time-series. For the word 'atomica', DWE cumulative is the only approach that fits the change point in the gold standard, indicating the change point as the decade 1950-1959, after 1945, year of Hiroshima and Nagasaki. We do not detect change points in the time-series produced by Procrustes point-wise and DBE point-wise, while we find a change point in the TRI-cumulative time-series in the 1950-1959 decade. For the word 'palmare', in the DBE point-wise and Procrustes cumulative time-series, two change points are detected that are too early compared to the change point in the gold standard 1998. Procrustes provided the right range 1950-1959 for the word 'Oscar', years in which for the first time an Italian film director, Vittorio De Sica, won the Oscar. TRI cumulative and DBE point-wise do not detect change points, while in the DWE point-wise time-series a change point is founded in the decade 1960-1969.

| Model | Precision | Recall | F-Measure | Change points detected |
|---|---|---|---|---|
| DWE cumulative | .0016 | .0840 | .0031 | 13207 |
| DWE point-wise | **.0020** | **.0880** | **.0039** | 11115 |
| TRI cumulative | .0017 | .0680 | .0033 | 10233 |
| TRI point-wise | .0016 | .0680 | .0032 | 10315 |
| DBE cumulative | .0000 | .0000 | .0000 | 255 |
| DBE point-wise | .0019 | .0200 | .0035 | 2815 |
| Procrustes cumulative | .0016 | .0640 | .0033 | 9652 |
| Procrustes point-wise | .0019 | .0200 | .0036 | 2757 |

TABLE 4.13: Results of the evaluation.

# Chapter 5

# Linguistic Knowledge Graph Databases

## 5.1 A New Time-sensitive Model of Linguistic Knowledge for Graph Databases

ICT provides an unprecedented opportunity to foster and support the preservation and research on immaterial Cultural Heritage. A large part of research in the Humanities and Cultural Heritage (H&CH) sector involves the collection and analysis of the material of cultural and/or historical interest. Semantic Web technologies have been used successfully in a number of humanities projects such as the Pelagios project [97] and the Mapping the Manuscripts project [35]. Given the relevance of textual materials in this research, it is not surprising that significant progress has been made in the design of linked data models for language data (see, for example, the excellent survey in Khan et al. [106]). A notable example of a multilingual synchronic language resource that has had a profound impact on the research community is BabelNet [162], a semantic network which connects the English computational lexicon WordNet [66] with a range of Open Linked Data resources such as Wikipedia and Wikidata, and many others. Alongside such resources, the research community has developed Semantic Web ontologies such as LeMON [147] particularly designed for the encoding of linguistic information.

The ability to model (language) data *diachronically*, is particularly important as a large part of H&CH work deals with historical data with a view to model change over time. In this line of research, some work has started on the modelling of cognate words and loan relations between words [2]. Related to this is the treatment of semantic change, the phenomenon concerned with the change in the meaning of words over time. The automatic detection of such changes has seen a very rapid development in Natural Language

Processing (NLP) research in recent years [216, 226, 205], with the majority of the approaches relying on distributional semantics, i.e. on representations of the semantics of words trained from corpus data covering different time intervals via embedding technologies. Some studies, e.g. [8], have advocated for the integration of such distributional approaches with Linked Open Data technologies, stressing how this best caters for the heterogeneous nature of the data relevant to this phenomenon, which includes not only language data, but also data on historical events and entities, as well as of bibliographic and geographic nature. However, Linked Open Data technologies have some limitations which we propose to address, as explained below.

Data Bases (DBs) aim at efficient storage, management and retrieval of data. Knowledge Bases (KBs), investigated in AI, are aimed at supporting formal reasoning on the available information. A *Knowledge Graph* (KG) is a kind of KB [63] where an ontology acts as the data model, and the data are organized in a graph structure [206]:

$$\text{ontology} + \text{data} = \text{knowledge graph}.$$

Combining the advantages of Database Management Systems (DBMSs) for handling individuals (scalability, storage optimization, efficient handling, mining and browsing of the data, etc.) with the high-level functionalities available in KBs would endow applications with much more power than allowed by the DB's query language alone.

An opportunity for such combination comes from the recent development of *Graph Databases*, a kind of NoSQL DBs of which Neo4j [186] is the most popular representative. Neo4j has been adopted by many big companies and governmental organizations for several different and relevant use cases, including Recommendation, Biology, Artificial Intelligence and Data Analytics, Social Networks, Data Science and Knowledge Graphs[1]. Neo4j comes with a powerful query language (Cypher) and extensive libraries for advanced data manipulation (APOC).

Unfortunately, formal ontologies and graph DBs refer to different graph models, which cannot straightforwardly be combined together. The standard formalism for expressing ontologies and KGs is based on the Resource Definition Framework (RDF)[2]. RDF graphs are built upon RDF Triples of the form:

$$\text{(Subject, Predicate, Object)}$$

[1] `https://neo4j.com/use-cases/`, consulted September 8, 2021.

[2] `https://www.w3.org/RDF/`

representing arcs between the Subject and Object nodes. A more general structure is provided by the Labeled Property Graphs (LPGs) model [188] (adopted by Neo4j), ensuring great flexibility and expressive power. In LPGs, both nodes and arcs are associated with unique identifiers, may be labeled, and can store *properties* represented as key/value maps. Relevant advantages brought by LPGs over RDF graphs are[3]:

- In RDF graphs nodes are atomic, while in LPGs they carry information; this ensures a much more compact structure in the latter. Consequently, RDF graphs are much less readable and they also cause a significant decay in efficiency, especially in browsing-intensive tasks such as Social Network Analysis or Graph Mining algorithms;
- RDF cannot distinguish different occurrences of the same relationship between the same pair of entities; this is possible in LPGs thanks to the unique identifiers of relationships instances;
- RDF cannot attach properties to instances of relationships; the reification solution (transforming a relationship instance into an object which has relationships to the original Subject and Object and to the additional properties) worsens readability; another partial solution is via annotations.

One limitation of Neo4j is that it is schema-less: the user may apply any label/type or property to each single node or arc. While ensuring great flexibility, this means that there is no clear semantics for the graph contents. Developing LPG-based KGs requires the definition of an LPG-based ontological formalism for expressing graph DB schemas, so as to allow data interpretability and applications interoperability, and of a mapping between this model and the standard ontological model adopted in the literature. Research on this topic resulted in the *GraphBRAIN* technology, whose peculiarities and advantages are discussed in [68]. In GraphBRAIN the KB designers must provide pre-specified data schemas, expressed in the form of LPG-based ontologies, that will drive all subsequent accesses to a Neo4j graph DB. By referring to a schema, the applications will commit to be compliant with it, as in traditional databases. In this work, we will adopt GraphBRAIN technology to model time-sensitive linguistic knowledge in a graph database.

[3] `https://neo4j.com/blog/rdf-triple-store-vs-labeled-property-graph-difference/`, consulted September 8, 2021.

### 5.1.1 GraphBRAIN Graph Database Scheme Format

The *GraphBRAIN Schema* (GBS) format to define graph DB schemas consists of an XML file whose tags allow us to exploit the representational features provided for by the LPG model. Here we will recall its main components (more details can be found in [68]).

The main structure of the XML tags is reported in Figure 5.1, where the universal entity *Entity* and the universal relationship *Relationship*, acting respectively as the roots of the entity and relationship hierarchies, are implicitly assumed (recall that in ontological terminology entities correspond to classes and relationships correspond to object properties). Therefore, entities and relationships are to be specified only starting from the first level of specialization, which we will call *top level*. Since each node (resp., arc) in the graph must be associated with one top-level entity (resp., relationship), the top-level entities (resp., relationships) are to be considered as disjoint. They may be the roots of specialization hierarchies of sub-entities (resp., sub-relationships). The set of direct specializations of a (sub-)entity or (sub-)relationship are in turn disjoint and are not to be intended as a partition: instances that do not fit any of the specializations of a parent (sub-)entity or (sub-)relationship may be directly associated with the parent. This design choice prevents multiple inheritances, i.e. associating an instance to many classes belonging to different branches in the hierarchy. We partially recover this at the level of instances: when two instances of different (sub-)entities represent the same object, we link them using an *aliasOf* relationship. The single reference object represented by all these instances takes the union of their attributes.

```
domain // tag enclosing the overall ontology
    [imports]
    entities // tag enclosing the classes
        {entity} // see (*)
    relationships // tag enclosing the relationships
        {relationship} // see (*)
```

FIGURE 5.1: Main structure of GBS files.

Entities and relationships are specified using the structure shown in Figure 5.2. **Reference** is used only in relationships to specify their possible domain-range pairs, **taxonomy** allows us to conveniently represent the specialization-type assertions; all other object properties are to be specified in the **relationships** section. **Attributes** are mandatory for entities (an entity instance must be described by some attribute) and optional for relationships (a relationship may carry information in its very linking two instances). **Specialization** is a recursive tag, allowing us to define hierarchies of sub-entities or sub-relationships. In addition to its own attributes, each specialization inherits all the attributes of the (sub-)entities (resp., (sub-)relationships) on the hierarchy path from its specific **specialization** section up to the corresponding top-level entity (resp., relationship).

```
(∗) ( entity | relationship | specialization ) tag
      [references]
          {reference}
      [taxonomy]
          {specialization} // see (∗) (recursive)
      [attributes] // specifying the data properties
          {attribute}
```

FIGURE 5.2: Structure for describing entity and relationship hierarchies in GBS files.

Regarding datatypes, attributes of type *integer*, *real*, *boolean*, *string*, *text* take an atomic value of the corresponding type, where *text* is intended for free text of any length. This is different from *string*, which has a limited maximum length that can be specified in the 'length' attribute. Attributes of type *date* take values in one of the following forms: year; year/month; year/month/day. Attributes of type *select* denote a choice in an enumeration of values; attributes of type *tree* denote a choice in a tree of values; attributes of type *entity* denote 1:1 relationships between an instance of the current entity and an instance of another entity (specified in the 'target' attribute of the tag), e.g., the birthplace of an entity Person would be modeled as an attribute of type *entity* with target='Place'.

Each GBS schema is intended to describe one domain. However, sometimes wider domains involve ontological elements that are already described in more 'basic' schemas: for example, the schemas for Cultural Heritage, Food and Transportations might be exploited in the ontology aimed at supporting a touristic application. In such cases, it might be useful to reuse such schemas, both to standardize the definitions and to build on existing knowledge. The combination of multiple schemas is more powerful a representation than the simple juxtaposition of their elements. Indeed, their shared entities act as bridges that allow, through the relationships available in those domains, to connect proprietary entities of each domain that would not otherwise have a chance to be related with each other. In the GBS framework, classes and relationships in different ontologies are considered the same (and thus are shared) if they have the same name. They may have, however, different attributes, reflecting the different perspectives associated with the different domains. If an attribute is present in different domains it must have the same type in all of them. Moreover, additional cross-schema relationships (and entities) may be defined in the overall ontology, building on the existing ones. GBS schemas support such scenarios by providing for an optional section in which existing schemas can be imported.

### 5.1.2 Mapping onto DB and Ontology

Since graph DBs are naturally suited to express knowledge graphs, i.e., knowledge bases based on given ontologies, a fundamental requirement of our approach is that our schemas can be mapped onto both the DB and to an OWL representation which can then be processed by a reasoner. In this section, we report how these two mappings work in practice.

As said, part of the main motivation for defining GBS schemas is to endow LPG-based graph DBs with a schema that ensures a clear semantics to the information pieces they contain and provides directions for their management and interpretation. In this perspective, the DB users will be required to work according to pre-specified data schemas expressed in the form of ontologies. In our approach we allow a single graph DB to underlie several domains (schemas), provided that their elements (entities and relationships) are compatible. Each such schema would provide a partial view of the DB contents, perhaps representing a different perspective.

Let us now show how the GBS elements are implemented using LPG features. Leveraging the possibility of using many labels for nodes, each node is labeled with the specific entity it belongs to and with all the domains for which it is relevant (e.g., 'Herbert Simon' would be labeled with 'Person' for the entity and with 'economy' and 'computing' for the domains). On the other hand, since each arc may take at most one type, we use it for specifying the relationship it expresses.

Concerning attributes, a reserved attribute *notes* is implicitly assumed for both nodes and arcs, which allows us to add information not accounted for by the other, domain-specific attributes. Attribute values of types *integer*, *real*, *boolean*, *string* and *text* are stored as literal values for the corresponding DB types, e.g., Neo4j provides the following types matching GBS types: Integer and Float, Boolean, and String. For types *select* and *tree* the string corresponding to the selected value in the list or tree is stored. An attribute of type *entity* actually corresponds to a relationship between the current instance and an instance of the target entity and thus it is stored in the DB as an arc, connecting the nodes corresponding to these two instances and having the attribute name as type. Finally, albeit Neo4j provides for temporal types, including 'Date', following [186] we propose to model attributes of type *date* as relationships to one of the following four entities: **Day** (representing a specific day of a specific year, with integer attributes *day*, *month*, *year*); **Month** (representing a specific month of a specific year, with integer attributes *month*, *year*); **Year** (representing a year, with a single integer attribute *year*).

### 5.1.3 The Linguistic Knowledge Graph

The Linguistic Knowledge Graph (LKG) aims to capture different aspects of lexical resources, such as relations between words and concepts, morphological, and syntactical information. Moreover, LKG covers diachronic aspects of language, such as the date of publication of a document, and the birth and death of an author. The schema we designed takes inspiration from the ontological lexicon model LeMON [54]. For space constraints, we report in Table 5.1 node types and in Table 5.2 the relationships adopted for diachronic analysis. The lexical unit is represented as node of type *InflectedWord* or *Lemma*, which are subclass of *Word*, i.e. *Lemma IS_A Word* and *InflectedWord IS_A Word*. The *Lemma* can be a multi-word expression (mwe), in this case, the flag mwe is set to True. The respective lemma of an *InflectedWord* can be retrieved exploiting the relationship *HAS_LEMMA* between *InflectedWord* and *Lemma*. The *LexiconConcept* is used to represent the word's meanings, and each instance of *LexiconConcept* represents a different meaning. For example, the *LexiconConcept* can represent the senses reported on a sense inventory, e.g. synsets in WordNet [158]. The relationship between a word and its meaning is expressed using the relationship *HAS_CONCEPT* among instances of *Word* and instance of *LexiconConcept*. Multiple relationships can be defined over couples of *LexiconConcept* using the reflexive relationship *SEM_RELATION*. At the same time, reflexive relationships over the Word instances can be described by the *LEX_RELATION* relationship.

The document structure from which words are extracted can be represented

at different levels of granularity: *Sentence*,*Text*, *Document*, and *Corpus*. In particular, each excerpt can be represented as *Text* or *Sentence*, which is a subclass of *Text*. A *Text* may belong to (*BELONG_TO*) a *Document* and a *Document* can be part of (*BELONG_TO*) a *Corpus*. The occurrences of a word in a particular *Text* are traced by the relationship *HAS_OCCURRENCE* among *Word* and *Text*. In the case of sense-annotated corpora, such as SemCor, is possible to specify the occurrences of senses using the relationship *HAS_EXAMPLE* among *LexiconConcept* and *Text*. Currently, the LKG takes into account two types of metadata: author and language. The relationship *HAS_AUTHOR* among nodes of type *Text* and nodes of type *Person* determines the author of a *Text*. The relationship *HAS_LANGUAGE* among nodes of type *Text*, *Document*, *Corpus*, and *Word* to nodes of type *Language* specifies the respective language.
The time is modelled using two classes of nodes: *TimeInterval*, and *TimePoint*, both subclasses of *TemporalSpecification*. The *TimeInterval* type is used when the date is not precisely stated, while the *TimePoint* is used in cases where the date is fixed. The start and end extremes of the *TimeInterval* nodes can be specified using the respective relationships *startTime* and *endTime*. In the current version of the LKG, time specification is supported for *Person* and *Text*. More specifically, the date of birth and death of authors is specified using the relationship *BORN* and *DIED* between *Person* and *TemporalSpecification*. The publishing date of a text is specified by the relationship *PUBLISHED_IN* among *Text* nodes and *TemporalSpecification* nodes.

### 5.1.4 Use case

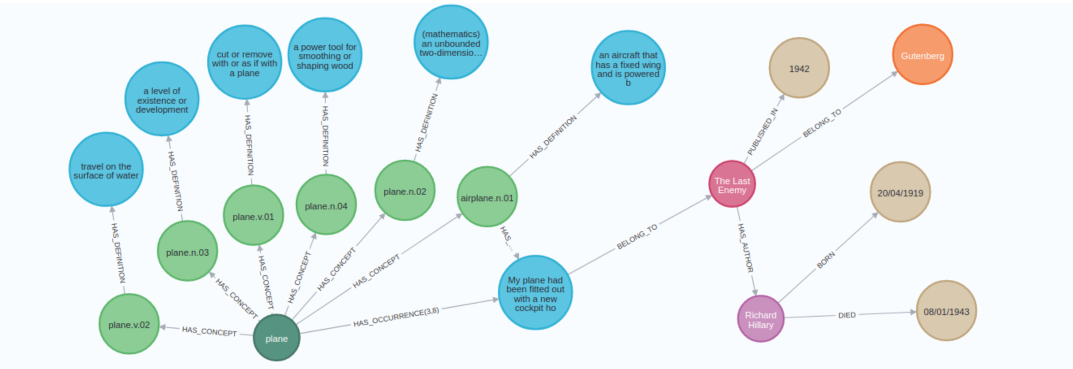


FIGURE 5.3: Example of a sub-graph for the Lexicon Entry *plane*.

| Class | Superclass | Attributes |
|---|---|---|
| Word | | value:String |
| Lemma | Word | value:String |
| | | postag:String |
| | | mwe:Boolean |
| InflectedWord | Word | value:String |
| Stem | | value:String |
| LexiconConcept | Concept | id:String |
| | | resource:String |
| Text | | value:String |
| Sentence | Text | |
| Document | | title:String |
| Corpus | | name:String |
| TemporalSpecification | | name:String |
| | | description:String |
| TimePoint | TemporalSpecification | Year:Integer |
| | | Month:Integer |
| | | day:Integer |
| TimeInterval | TemporalSpecification | |
| Person | | name:String |
| | | lastname:String |
| Language | | iso639-1:String |
| | | iso639-2:String |
| | | enName:String |
| Category | | id:String |

TABLE 5.1: LKG classes with their respective superclasses and attributes.

Figure 5.3 shows the sub-graph related to the Lexicon Entry *plane*. The extracted sub-graph shows the concepts associated with the Lexicon Entry by the referring lexical resource, in this case WordNet, using the HAS_CONCEPT relationship. The concepts sketched are the synsets *airplane.n.01*, *plane.n.02*, *plane.n.03*, *plane.n.04*, *plane.v.01*, *plane.v.02*. For each Lexicon Concept, the WordNet glosses are provided by the relation *HAS_DEFINITION*.

Moreover, the example sub-graph includes a sentence extracted by the book *The Last Enemy* and containing the word *plane*, i.e.

*"My plane had been fitted out with a new cockpit hood".*

The book is represented as a Document instance and belongs to the corpus Gutenberg (rel. BELONG_TO). The rel. *HAS_AUTHOR* connects the book with the author *Richard Hillary*, who was born on the 20th of April 1919 (rel. *BORN*) and died on the 08th of January 1943 (rel. *DIED*). The book publishing date, i.e. 1942, can be obtained via the rel. *PUBLISHED_IN*.

The extracted sentence is connected to both the Lexicon Entry *plane* and the Lexicon Concept *airplane.n.01* respectively by the rels. *HAS_OCCURRENCE*

| **Relationship** | **Subject** | **Object** | **Attributes** |
|---|---|---|---|
| IS_A | Sentence | Text | id:Integer |
| | Lemma ∪ InflectedWord | Word | id:Integer |
| BELONG_TO | Text | Document | id:Integer |
| | Document | Corpus | id:Integer |
| | Text | Category | |
| HAS_OCCURRENCE | Word | Text | begin:Integer<br>end:Integer |
| {LEX_RELATION} | Word | Word | |
| HAS_LEMMA | Word | Lemma | |
| HAS_CONCEPT | Word | LexiconConcept | grade:Float |
| HAS_EXAMPLE | LexiconConcept | Text | |
| HAS_DEFINITION | LexiconConcept | Text | |
| REFER_TO | LexiconConcept | Concept | |
| {SEM_RELATION} | LexiconConcept | LexiconConcept | |
| PUBLISHED_IN | Text ∪ Document ∪ Corpus | TemporalSpecification | |
| HAS_AUTHOR | Text ∪ Document ∪ Corpus | Person | |
| BORN | Person | TemporalSpecification | |
| DIED | Person | TemporalSpecification | |
| startTime | TimeInterval | TimePoint | |
| endTime | TimeInterval | TimePoint | |
| HAS_LANGUAGE | Text ∪ Document ∪ Corpus ∪ Word | Language | |

TABLE 5.2: LKG relationships with their respective subject, object and attributes.

and *HAS_EXAMPLE*. The former rel. addresses the occurrence of the word *plane* in the sentence, the latter that *plane* occurs with the meaning specified by the Lexicon Concept *airplane.n.01*, i.e.

> *"an aircraft that has a fixed wing and is powered by propellers or jets".*

In both relations, the offsets of the word *plane* are specified by the relationship attributes, i.e. 3 and 8.
In the proposed example, the time dimension is elicited by three components: the book publishing date, the date of birth and the date of death of the author *Richard Hillary*. The time specifications acquire a relevant role in the context of Diachronic Linguistics. From the publishing date of *The Last Enemy*, we can infer that the occurrence of *plane* in the extracted sentence is one of the earlier appearances of the word *plane* with the *airplane.n.01* sense. Furthermore, information about the Author, such as his influences, and the historical period in which he lived, can enable deeper analyses, guiding the study and the definition of the cause-effects relationship in Lexical Semantic Change phenomena.

## 5.2 Using Graph Databases for Historical Language Data: Challenges and Opportunities

In 5.1, we adopted GraphBRAIN technology to model time-sensitive linguistic knowledge in a graph database, describing a time-sensitive model of linguistic knowledge that can be used for graph databases. Here, we show an application of this model to the lexical semantic analysis of Latin data, i.e. the analysis of the meanings of Latin words. Differently from previous approaches, such as Basile, Caputo, and Semeraro [12], Hamilton, Leskovec, and Jurafsky [86], and Carlo, Bianchi, and Palmonari [38], we exploit graph database potentialities to detect semantic changes in specific concepts.

Latin is in a particularly favourable position among historical languages for the large-scale analysis of semantic change processes, thanks to a number of factors. First, Latin researchers now enjoy unprecedented access to digital data covering over two thousand years of history. Thanks to the ERC-funded LiLa project [4], seven Latin language resources and six corpora have been linked at the level of word lemmas so far, making Latin a unique case among historical languages. Second, we have access to extensive computational language resources for Latin, Latin WordNet [160], and digitised dictionaries of Latin, which provide rich information about words' semantics and examples of usage. Finally, focussing on Latin allows us to investigate semantic change processes over long time spans. Latin has one of the longest recorded histories of any human language, making it naturally suitable for quantitative studies [173]. The first inscriptional records date from the sixth century BCE, and Latin continues to be used to the current day by the Catholic Church and some academic and legal institutions around the world. Written Latin diverged from the spoken vernaculars in the second half of the first millennium of the Christian era, but it remained in use as one of the principal channels of communication across most of Europe for the next thousand years. The humanists' conscious effort to reproduce Classical Latin led to a range of interesting developments, particularly affecting the neo-Latin lexicon to enable the expression of new concepts. This extensive chronological span has raised the question of the extent to which Latin is seen as a dead or fossilised language (e.g. Herman [89]). However, it remains an open question to what extent this fossilisation affected the semantics of words, as we know that the Latin lexicon, in this respect, has remained dynamic (over 4,500 words have acquired new meanings since the Renaissance; Demo 2022). The extent

[4] `https://lila-erc.eu/`

to which post-classical Latin can really be considered as a “fixed” language (Roelli [189]) from the point of view of its ability to generate new meanings of words is still largely unknown beyond anecdotal evidence.

### 5.2.1 Latin data

The data we loaded into the graph consists of a portion of the LatinISE corpus [150] annotated at the level of dictionary senses. LatinISE is a Latin corpus covering the period from the fifth century BCE to the twenty-first century and contains 10 million word tokens, semi-automatically lemmatised and part-of-speech tagged. The metadata fields in LatinISE indicate text identifier, author, title, dates, century, genre, url of source, and optionally book title/number and character names (for plays). The annotated dataset was produced as part of the SemEval shared task on Unsupervised Lexical Semantic Change Detection [205]. 40 Latin lemmas (“target words”) are selected, of which 20 are known to have changed their meaning with the advent of Christianity (for example, *beatus*, which shifted its meaning from ‘fortunate’ to ‘blessed’) and 20 are known to not have changed their meaning between the BCE era and the CE era. For each of the 40 lemmas, 60 sentences are randomly extracted from LatinISE, 30 of them are from texts dated in the BCE era, and 30 from texts dated in the CE era. Each sentence was annotated by at least one expert annotator, according to the DuReL framework [200]. The annotators were asked to judge the semantic relatedness of an instance of usage of a target word with respect to the list of its dictionary definitions using a four-point scale (Unrelated, Distantly Related, Closely Related, and Identical). The definitions were taken from the Latin portion of the Logeion online dictionary (`https://logeion.uchicago.edu/`) containing Lewis and Short’s *Latin-English Lexicon* (1879) [132], Lewis’ *Elementary Latin Dictionary* (1890) [131], and Du Fresne Du Cange et al. [60]. See McGillivray et al. [151] for further details about the dataset and its annotation framework.

```
MATCH
(centuryNode:TimeInterval)-[:startTime]->(startCentury:TimePoint),
(centuryNode:TimeInterval)-[:endTime]->(endCentury:TimePoint),
(pubNode:TimeInterval)-[:startTime]->(startPub:TimePoint),
(pubNode:TimeInterval)-[:endTime]->(endPub:TimePoint),
(text:Text)-[:PUBLISHED_IN]->(pubNode)
WHERE
centuryNode.description="century"
WITH text,
centuryNode,
CASE WHEN endPub.Year > endCentury.Year THEN endCentury.Year ELSE
↪  endPub.Year END as minEnd,
CASE WHEN startPub.Year > startCentury.Year THEN startPub.Year ELSE
↪  startCentury.Year END as maxStart
```

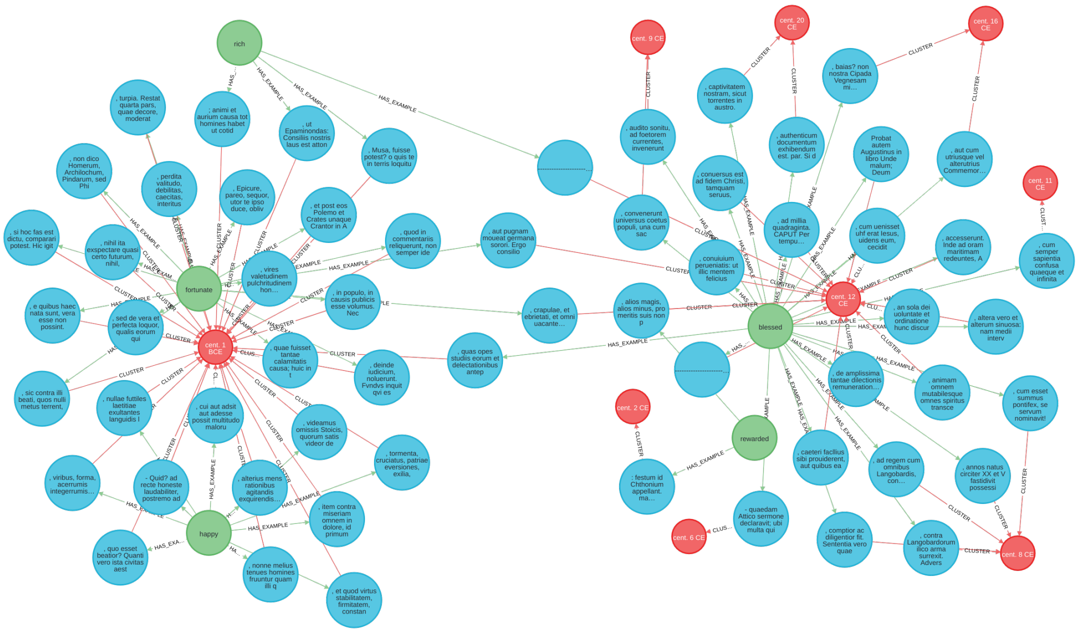


FIGURE 5.5: Graph for the Latin word *beatus*.

```
WITH *,
CASE WHEN minEnd-maxStart+1 > 0 THEN minEnd-maxStart+1 ELSE 0 END as
↪  time_overlap
ORDER BY time_overlap DESC
WITH text,
collect({century:centuryNode})[0] AS max
WITH *,
max.century as century
CREATE (text)-[r:CLUSTER]->(century)
RETURN text,century
UNION ALL
MATCH
(centuryNode:TimeInterval)-[:startTime]->(startCentury:TimePoint),
(centuryNode:TimeInterval)-[:endTime]->(endCentury:TimePoint),
(text:Text)-[:PUBLISHED_IN]->(point:TimePoint)
WHERE
centuryNode.description="century" and
point.Year>=startCentury.Year and
point.Year<=endCentury.Year
WITH text, centuryNode as century
CREATE (text)-[r:CLUSTER]->(century)
RETURN text, century;
```

FIGURE 5.4: Clustering publishing date by centuries

### 5.2.2 Loading the Latin data in the Linguistic Knowledge Graph

For each instance of the target words in the Latin corpus we encode:

- the author as *Person*,

- the manuscript as *Document*,
- the year as *TimePoint* if the date is certain, *TimeInterval* otherwise,
- the sentence (left context, target word and right context) as *Text*,
- the definitions of the Lewis and Short Dictionary as *LexiconConcept*,
- the word lemma as *Lemma*,
- the inflected forms of the target words as *InflectedWord*,
- the scores associated with each *LexiconConcept* as properties of the *HAS_EXAMPLE* and *HAS_OCCURRENCE* relationships.

In order to simplify and make the visualisation more effective, we created the *HAS_EXAMPLE* relationship only in cases where the annotation reported a score of 4. In addition, to make more evident the distribution of senses with respect to centuries, we associate each date of publication of the texts with the reference century. We do this via the query given in Figure 5.4. In case a *Text* is not associated with a specific *TimePoint*, it will be linked with the century having the greatest overlap with the *TimeInterval* of the text itself. On the other hand, for texts for which a precise date is specified, the query associates the *Text* with the respective century of its year. The centuries are represented as *TimeInterval*, and the description attribute is validated with “century”. A new relationship, called *CLUSTER*, is so created among nodes of type *Text* and nodes of type *TimeInterval* to indicate the century.

A subgraph for the word *beatus* is shown in Figure 5.5. The graph shows the nodes representing the texts from which the word *beatus* is extracted, the centuries and the senses given in the Lewis and Short Dictionary. The relationships among these nodes are *CLUSTER* and *HAS_EXAMPLE*. The former connects nodes of type *TimeInterval* and nodes of type *Text*, see 5.4. The latter links *LexiconConcept*s and *Text*s. Most occurrences of the word *beatus* in the reference corpus are dated 1st century BCE and 11th century CE. One can immediately notice a difference in the distribution of the senses: “happy” and “fortunate” on the one hand are associated with the time period BCE (see the cluster of nodes on the left of Figure 5.5), and “blessed”, on the other hand, is associated with the time period CE (see the cluster of nodes on the right of Figure 5.5). In fact, only one sentence in the dataset displays the sense “blessed” in the first century BCE. Similarly, only two sentences dated CE contain the word *beatus* with the meaning of “fortunate”, the latter, on the other hand, is dated 1079-1142 CE and is an excerpt from the Sermones of Petrus Abaelardus.

## 5.3 Graph Databases for Diachronic Language Data Modelling

Research in empirical historical semantics requires access to various sources, from dictionaries and lexicons to encyclopedic information and diachronic texts. While several scholars have recognized the corpus-based nature of diachronic semantics, particularly for corpus languages like Latin [173, 76], quantitative corpus-based studies are yet to pervade historical semantics research. A critical barrier to this is that corpus and lexical resources for historical languages tend to exist in data siloes. While significant progress on linking lexical resources, tools, and corpora at the level of lemmas has been made (cf. Passarotti et al. [166] for Latin), linking at the level of word senses is still missing.

Given the remarkable work done in the design of linked data models for language data [106], some studies such as Armaselu et al. [8] have already advocated for integrating corpus approaches with Linked Open Data technologies to study lexical semantic change, i.e., the phenomenon concerned with the change in the meaning of words over time. One crucial strategy for representing the results of research into language change as linked data is by modeling and publishing them as knowledge bases using a lexicon-based model, usually OntoLex-Lemon and its various extensions. This includes the soon-to-be-published Frequency Attestations and Corpus (FrAC) module, which proposes a new series of classes and properties for linking elements of a lexicon with corpora [47]. Previous work in this area includes a proposal to modify the core organizing principles of wordnets in order to represent semantic shift phenomena [105], as well as work on the representation of etymologies as Resource Description Framework (RDF) graphs using OntoLex-Lemon [104] and the integration of temporal information into linguistically linked datasets via a so-called *four-dimensionalist* approach [107].

Integrating lexical resources and semantically-annotated corpus data at scale would allow us to gather corpus data on sense distribution information, essential for fully implementing the quantitative turn in historical semantics [149]. This integration, however, requires efficient handling of large datasets. An opportunity to combine the efficient storage, management, and retrieval of data offered by Data Base Management Systems (DBMSs) with the support for formal reasoning offered by Knowledge Bases (KBs) comes from the recent development of *Graph Databases*. Graph DBMS are intrinsically designed to store schemaless data, making them suitable to dynamic systems in which merging information is relevant. Unlike traditional DBMSs such as relational [111] or object-oriented [25] ones, Graph DBMS lack predefined structures. Neo4j [5] is among the most common graph DBMSs. GraphBRAIN[6] technology [69] provides intelligent information retrieval functionalities on a graph database. Its interface provides end users with access to data employing schema definitions. Schemes (available in terms of classes, relationships, and attributes) coordinate how data is presented in the interface. In Basile et al. [16], we proposed the *Linguistic Knowledge Graph*, a model based on graph DBMSs. The Linguistic Knowledge Graph models relations between concepts and words, information about word occurrences in corpora, and diachronic information on both concepts and words. In McGillivray et al. [154], we show an application of this model to the lexical-semantic analysis of Latin data.

Our choice to focus on Latin is motivated by several factors. First, Latin has one of the longest recorded histories of any human language, making it naturally suitable for quantitative studies [173]; this, in turn, allows for corpus-driven analyses of semantic change processes over long periods. Second, this language has a particularly favourable position among historical languages: there is a high availability of extensive Latin corpora in digital form (some of which have been linked to language resources at the level of word lemmas in the context of the LiLa project [7]) and of computational language resources such as Latin WordNet [160] and digitized dictionaries such as the Lewis Short Latin dictionary[8].

Focusing on the development of the Latin language, in this work we expand the range of Latin language resources included in the Linguistic Knowledge Graph for the study of lexical semantic change in Latin.[9] Our contributions include: (i) the ingestion of Latin WordNet into the Linguistic Knowledge Graph; (ii) a new curated linking between existing resources for Latin, namely Latin WordNet [160, 26] and the SemEval 2020 Task 1 Latin dataset [148], a sense-annotated portion of the LatinISE diachronic corpus of

[5] `https://neo4j.com/`

[6] http://193.204.187.73:8088/GraphBRAIN/

[7] `https://lila-erc.eu/`

[8] `https://lila-erc.eu/data/lexicalResources/LewisShort/Lexicon`

[9] Our code and data are available at `https://github.com/linguisticGraph/latin-graph`

Latin [151];[10] (iii) the integration of external contextual information (Wikidata) about the occupations of Latin authors. The term 'occupation' is here used in a broad sense, to refer to various types of political, cultural and societal profiles that identify authors in Wikidata. These could be e.g., priests, philosophers, historians, hagiographers, among others.

### 5.3.1 Resources

#### Dataset

LatinISE contains approximately 10 million word tokens from texts dating from the fifth century BCE to the contemporary era; it has been semi-automatically lemmatized and part-of-speech tagged. The corpus includes metadata fields indicating text identifier, author, title, dates, century, genre, URL of the source, and book title/number and character names (for plays). The semantically annotated dataset we use here was created as part of the SemEval shared task on Unsupervised Lexical Semantic Change Detection [205] and will be henceforth referred to as the SemEval Latin dataset. It contains in-context annotations for 40 Latin lemmas, 20 of which are known to have changed their meaning concerning Christianity (for example, *beatus*, which shifted its meaning from 'fortunate' to 'blessed'), and 20 are known not to have changed their meaning between the BCE era and the CE era. For each of these lemmas, 60 sentences were annotated, of which 30 were randomly extracted from BCE texts and 30 from CE texts. The annotation was conducted following a variation of the DuReL framework [200] described in Schlechtweg et al. [205]: the degree by which a usage instance of a target word is related to each of its possible dictionary definitions was annotated using a four-point scale (Unrelated, Distantly Related, Closely Related, and Identical). The definitions were drawn from the Logeion online dictionary (`https://logeion.uchicago.edu/`), which contains Lewis and Short's *Latin-English Lexicon* (1879) [132], Lewis' *Elementary Latin Dictionary* (1890) [131], and the dictionary by Du Fresne Du Cange et al. [60]. The details of the annotation are described in McGillivray et al. [151].

#### Curated Linking

We manually linked each word sense of the SemEval Latin dataset to one or more WordNet synsets. We started with the dataset provided by the LiLa

[10] Openly available at `https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-2506`.

project [73], which contains a sample of 10,314 lemmas from Latin WordNet (LWN) [160, 26]. The LiLa team verified and corrected, where necessary, the synsets associated with each lemma of the sample and linked them to version 3.0 of Princeton WordNet (PWN) [67, 158]. However, as the LiLa dataset only covers 22 of the 40 lemmas in our dataset, we used LWN as a reference for the remaining 18 lemmas. We converted the synset codes 1.6 used by LWN to version 3.0 of PWN for consistency.

The senses assigned to the target words in the SemEval Latin dataset often condensed multiple meanings into a single definition, requiring multiple synsets to be linked to the same meaning to capture all nuances. For example, the sense "understanding, judgment, wisdom, sense, penetration, prudence" of the lemma *consilium* was linked to four synsets.

In some cases, a particular sense could not be described by any of the assigned synsets in the LiLa dataset. In such cases, we searched for the lemma in LWN and selected a more appropriate synset. This was the case e.g. for the adjective *acerbus* and one of its meanings in the SemEval Latin dataset "(of things) heavy, sad, bitter". For this meaning we selected the synset 01650376-a "psychologically painful" from LWN. When we could not find the synset in either LWN or the LiLa dataset, we looked for the most suitable synset in PWN. However, for some meanings specific to Roman culture and institutions, we could not find a suitable synset, such as with the meaning 'Virtue, personified as a deity' of *virtus*. In these cases, we did not link the sense to WordNet.

### Contextual Information

In some instances, the metadata field of the SemEval Latin dataset (which indicates the author and title of the text, dating, and genre) was noisy, incorrectly structured, or incomplete. Wikidata is an extensive, collaboratively maintained knowledge base [229], hosting more than one hundred million items. We exploited Wikidata for de-noising and linking the authors of the documents containing the sentences in our dataset.

First, we extracted the Wikidata entities for which the author's occupation is specified (wdt:P106, *occupation*), and Latin (wd:Q397, *Latin*) is one of the writing languages for the author (wdt:P6886, *writing language*). We retrieve information about each author in the form of key/value properties. Author names in the SemEval Latin dataset can occur in different languages and different forms, for example *praenomen* and *nomen* followed by *cognomen* e.g., Marcus Tullius Cicero; *cognomen* followed by *praenomen* and *nomen* e.g., Cicero, Marcus Tullius; only *cognomen* e.g., Cicero; only *praenomen* and *nomen* e.g., Marcus Tullius. We processed the author's mentions in the SemEval Latin dataset and the writer labels and aliases extracted from Wikidata, performing lowercase and punctuation removal. Matching is realized by computing the Levenshtein distance [199] between the author reported in the SemEval dataset and all the collected surface forms (i.e., labels/aliases) from Wikidata. The surface forms are then ranked by decreasing Levenshtein distance. If the Levenshtein distance between the author's mention and the top-ranked surface form is less than a fixed threshold, i.e., $\delta = 0.1$, the entity referenced by the surface form is linked to the author's mention. For each author, Wikidata provides rich information, such as biographical data, the author's works, and events that influenced their life and production. In this study, we focus on occupation information: we encode the information provided by Wikidata about the occupations of the author exploiting the property wdt:P106 (*occupation*). In particular, we create nodes of type Occupation for each occupation retrieved in Wikidata, generating a relationship between the author and their respective occupation.

**Latin WordNet Ingestion**

The Latin WordNet (LWN) project is an initiative to create and share a common lexico-semantic database of the Latin language. The project originated as a branch of the MultiWordNet [171] project. For diachronic analyses, linking linguistic resources with temporal information allows us to uncover instances of semantic changes in the usage of words. Hence, we provide a mechanism to enrich the Linguistic Knowledge Graph with Latin WordNet and exploit the hierarchical structure of the relationships between synsets.
In Section 5.1.1, we described GraphBRAIN technology and its reliance on schemes/ontologies to deliver information extraction and reasoning functionalities. We mapped the Latin WordNet data with the portion of our ontology specifically devoted to linguistic analysis and understanding. Further details about scheme specifications for document representation are available in [70]. Here we describe the mapping between the lexical database and our schema. In LWN, we identified the following resources, grouped into separate Comma Separated Value (CSV) files: lemma, lexical_relation, literal_sense, metaphoric_sense, metonymic_sense, phrase, semantic_relation, synset. Each resource has features that may be seen as classical columns in a relational database. From now on, we refer to specific fields as *resource.field*

to uniquely identify them and motivate how we map them. The alignment process is as follows:

- lemma: a specific lemma is embedded in our class **Lemma**. A **Lemma** is characterized by a unique id, a lemma (its value), and a PoS tag (modelled as a relationship). For our purposes, the class **PartOfSpeech** collects all the pos tags used, following the Universal PoS Tags standard[11]. We can represent other fields expressed in LWN, such as *lemma.uri*.
- lexical_relation: this represents a relationship between two **Lemma**s. The field *lexical_relation.type* specifies the type of relationship. We modelled the present ones with some explicit names which express their meanings: **ANTONYMOUS_OF**, **PERTAINS_TO** (to refer to the type of relation indicated by the attribute of the relations), with their corresponding inverses, e.g. **IS_PAST_PARTICIPLE_OF**.
- literal_sense: this represents a relationship between a lemma, identified by the field *literal_sense.lemma*, and a synset, identified by *literal_sense.synset*. We call this relationship **expresses**. We highlight that the relationship has a "literal" sense by adding a specific attribute **sense**. Additional information about the period and genre is available.
- metaphoric_sense: similarly to the previous one, this represents a relationship between a lemma and a synset, where the **sense** is "metaphoric".
- metonymic_sense: as before, but the **sense** is "metonymic" in this case.
- phrase: a phrase is a word or a multi-word expression. In both cases, the concept is expressed by the class **Lemma** since for our purposes both concepts play an equally important role when analysing semantic changes. Again, we have the PoS tag information, which is modelled in the same way described above.
- semantic_relation: a relationship between two synsets. Based on the *semantic_relation.type* several relationships may be expressed. They are mapped into the following ones and their corresponding inverses: **PART_OF**, **HAS_SUBCLASS**, **ATTRIBUTE_OF**, **SIMILAR_TO**, **ANTONYMOUS_OF**, **PERTAINS_TO**, **PART_PARTICIPLE_OF**, **CAUSES**, and **ENTAILS**.

[11] `https://universaldependencies.org/u/pos/`

- synset: a synset is embedded in **LexiconConcept** while its property synset.gloss, which is the description of the synset, is represented as the attribute **description** of the class **LexiconConcept**. *synset.gloss* is the description of the synset and is mapped onto the attribute **description**.

Thanks to this mapping, we can acquire the LWN resource and represent it in our formalism, which allows us to leverage the connections between the different datasets, as explained via examples in the next section.

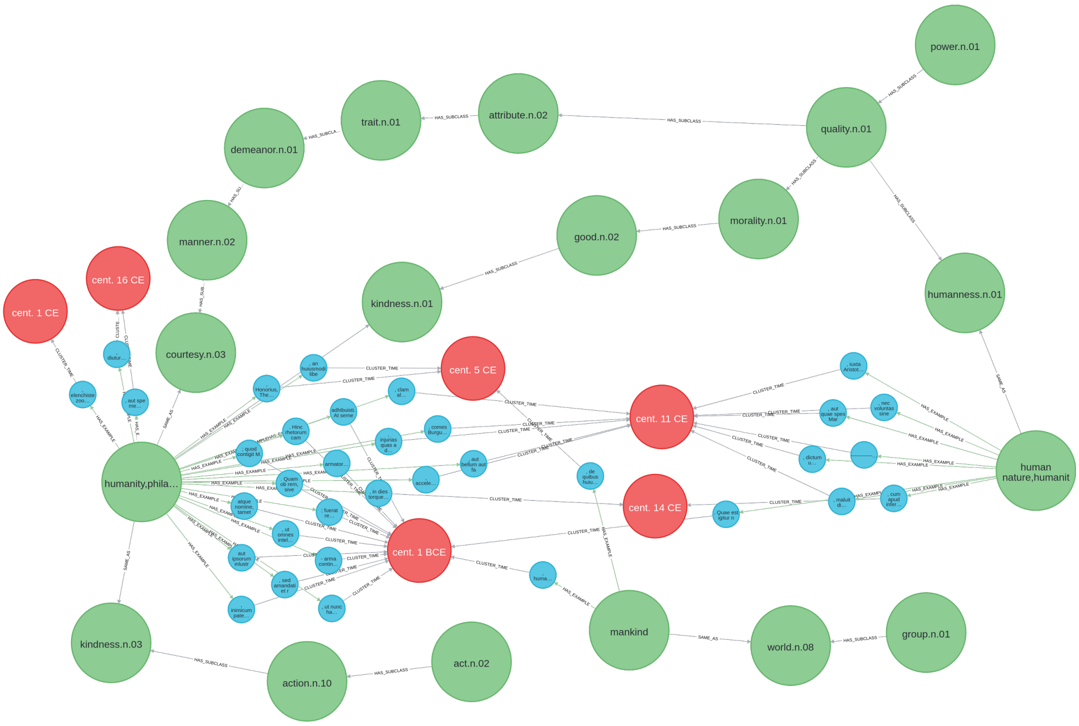
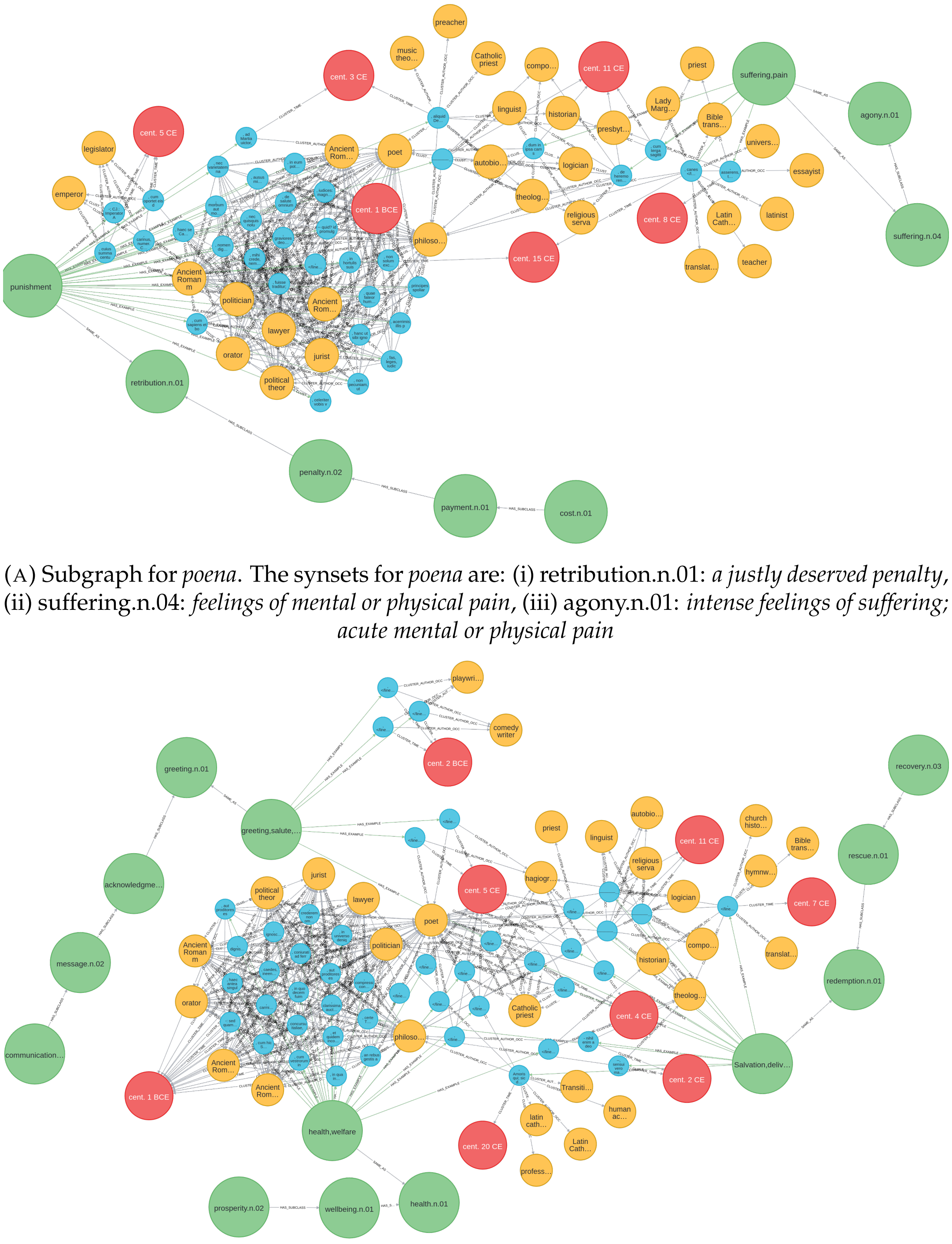


FIGURE 5.6: Subgraph for the word *humanitas*, including the sentences in which the lemma *humanitas* occurs in the SemEval Latin dataset, the century of the works from which the sentences were extracted, the annotated senses in the SemEval Latin dataset, and the curated links between the senses and the synsets in Latin WordNet. The sentences are represented as Text nodes (in blue), the senses and the synsets as LexiconConcept nodes (in green), and the centuries as TimePoint nodes (in red).

### 5.3.2 Analysis and Discussion

Figure 5.6 shows the subgraph for the word *humanitas*. The occurrences of *humanitas* are annotated in the SemEval dataset with three senses: (i) 'human nature, humanity', (ii) 'humanity, philanthropy', and (iii) 'mankind'.[12]

[12] A fourth sense 'liberal education, good breeding, the elegance of manners or language, refinement' was annotated in the Latin dataset, but not encoded in the graph, since the author matching described in Section 2.3 failed.



(A) Subgraph for *poena*. The synsets for *poena* are: (i) retribution.n.01: *a justly deserved penalty*, (ii) suffering.n.04: *feelings of mental or physical pain*, (iii) agony.n.01: *intense feelings of suffering; acute mental or physical pain*

(B) Subgraph for *salus*. The synsets for *salus* are: (i) health.n.01: *a healthy state of well-being*, (ii) redemption.n.01: *(Christianity) the act of delivering from sin or saving from evil*, (iii) greeting.n.01: *an acknowledgment or expression of goodwill*

FIGURE 5.7: Sub-graphs: (a) beatus. (b) poena (c) salus

In the curated link, we associate the sense (i) to the humanness.n.01 synset, the sense (ii) to the synsets kindness.n.01, kindness.n.03, and courtesy.n.03 and sense (iii) to the synset world.n.08. According to the *Thesaurus Linguae Latinae* [222], which confirms the first attestation of all senses in the 1st century BCE, the sense (ii) 'humanity, philanthropy' developed from the more general sense (i) 'human nature, humanity' which refers to human nature in general. The subgraph shows that the three senses are attested at least once in passages dated 1st century BCE. However, the graph shows that the sense of 'philanthropy' dominates all other senses in the 1st century BCE. In the transition to the CE period, the sense of 'humanity' prevails regarding the number of annotations, and the two meanings coexist in the CE period.

By ascending the WordNet hierarchy, we can gain deeper insight into the relationship between the two senses. The sense (ii) 'humanity, philanthropy' and the sense (i) 'human nature' are connected via two paths: sense (ii) originates from the quality.n.01 synset (i.e. 'an essential and distinguishing attribute of something or someone'); sense (i) from the attribute.n.02 synset (i.e., 'an abstraction belonging to or characteristic of an entity'). The two senses have in common the quality.n.01 synset, but the sense (ii) 'humanity, philanthropy' is directly linked to kindness.n.01 synset, and to a higher degree of the WordNet hierarchy to the morality.n.01 synset (i.e., 'concerned with the distinction between good and evil or right and wrong'). The additional information provided by including the WordNet hierarchy in the graph allows us to show the type of semantic relationship between the two predominant senses of *humanitas*. The more general sense (i) 'human nature' specializes in its meaning in the sphere of morality, originating the sense (ii) 'philanthropy'. In the example of *humanitas* shown in Figure 5.6, the injected information from WordNet was exploited to analyze the semantic relationship between the meanings of the lemma *humanitas*. While the synset taxonomy in this example helps us track and classify phenomena of semantic change, including other types of information retrievable from the metadata can help gain further insights into the context of the semantic change. We add information about the authors' occupations in the examples shown in Figure 5.7.

In Figure 5.7, three examples of subgraphs are shown. The three graphs refer, respectively, to the encoded information for the Latin lemmas *beatus*, *poena*, and *salus*. In particular, we filtered for nodes of type Text (blue nodes), Century (red nodes), Synset (green nodes), and Occupation (yellow nodes). We grouped the Text nodes by occupation and century, i.e., we created an explicit link between nodes of type Text and nodes of type TimePoint and between nodes of type Text and nodes of type Occupation.

Combining queries at the level of the annotated senses, WordNet synsets, text metadata and textual data at once, users can have access to rich nuanced information, which is very valuable for quantitative diachronic semantic analyses, both on specific words and whole lexical fields. The graphs in Figure 5.7 seem to show some trends in semantic change, all related to Christianity. The lemma *beatus* was annotated in the SemEval dataset with five senses: (i) 'happy,' (ii) 'fortunate', (iii) 'rewarded', (iv) 'rich', and (v) 'blessed'. The graph shows that the senses (i) 'happy', (ii) 'fortunate', (iii) 'rewarded', and (iv) 'rich' all emerge starting from the 1st century BCE in the annotated dataset. On the other hand, sense (v) 'blessed' emerges later with the advent of Christianity, as we can see in correspondence with the CE nodes. In this case, there seems to be a replacement of the previous senses in favour of the Christian sense.

Additionally, if we consider the nodes of type Occupation, a noticeable difference emerges between the two (groups of) meanings: in the cluster of occupation nodes connected to the Christian sense, we can observe profiles related to theological and religious activity, e.g., priests, hagiographers, which do not appear to be connected to the other senses. The same type of observations can be made for *salus*, which initially has the meanings (i) 'health' and (ii) 'greeting', and, subsequently, develop the Christian sense of (iii) 'salvation, deliverance from sins'. However, in this case, we can notice the difference with *beatus* in the type of semantic change, as the new meaning (iii) 'salvation' replaces or dominates the previously attested meanings but continues to coexist with them.

The lemma *poena* also presents an example of semantic change in which the new meaning does not entirely replace the previous ones. The new sense of 'suffering, pain', which emerges in the CE nodes, continues to coexist with the sense of 'punishment', which was attested from the 1st century BCE in the annotated dataset. In the case of *poena*, the contrast between the two clusters of occupation nodes is even more evident. The sense of punishment is often associated with authors classified as related to the legal world, e.g., legislator, lawyer, and jurist. In contrast, nodes related to the Christian and theological world appear in the case of salvation, e.g., theologian, priest, and presbyter.

The graphs in Figure 5.7 are in line with that we know about semantic changes prompted by the advent of Christianity, which invested many words already in use in pre-Christian Latin with new meanings closely related to the Christian world [36]. Moreover, the lemmas shown in Figure 5.7 illustrate the different types of interaction between older and new senses described in literature [224, pp. 10–12]: in some cases, the two senses can continue to coexist, as for the lemmas *salus* and *poena* (a phenomenon called 'layering' [94, p. 22]); in others, as for the lemma *beatus*, the relationship between the new sense and the older ones is unbalanced as the new sense becomes more prominent in a society invested in Christian values.

# Chapter 6

# Benchmarking Unsupervised Lexical Semantic Change Detection

## 6.1 A diachronic Italian corpus based on "L'Unita"

### 6.1.1 Motivation and Background

Diachronic linguistics, as proposed by de Saussure in his *Cours de linguistique générale*, is one of the two major temporal dimensions of language study and has a long tradition in Linguistics. Recently, the increasing availability of diachronic corpora as well as the development of new NLP techniques for representing word meanings has boosted the application of computational models to investigate historical language data [86, 216, 219]. This culminated in SemEval-2020 Unsupervised Lexical Semantic Change Detection [205], the first attempt to systematically evaluate automatic methods for language change detection.

Italian is a Romance language which has undergone a large amount of changes in its history. Its official adoption as a national language occurred only after the Unification of Italy (1861), having previously been a literary language. Diachronic corpora of Italian are currently available and accessible to the public (e.g., DiaCORIS and MIDIA). Unfortunately, restricted access/distribution of these resources limits their utilisation and actually prevents the investigation of more recent NLP methods to the diachronic dimensions.

To obviate this limit, we collect and make freely available[1] a new corpus based on the newspaper "L'Unità". Founded by Antonio Gramsci on February, 12th 1924, "L'Unità" was the official newspaper of the Italian Communist Party (PCI [2], henceforth). The newspaper had a troubled history: with the

[1] `https://github.com/swapUniba/unita/`

[2] It is the acronym of *Partito Comunista Italiano*.

dissolution of PCI in 1991, the newspaper continued to live as the official newspaper of the new Democratic Party of the Left (PDS/DS) until July, 31th 2014. After that date, it ceased its publication until June, 30th 2015, and it was definitely closed on June, 3rd 2017.

Since 2017, the historical archive of "L'Unità" has been made again visible and available on the Web.[3] One of the main issues of this resource is the lack of information about who owns the rights of the original archive. To our knowledge, the online version of the archive was legally obtained by downloading the original archive before the closure of the newspaper. The current archive, available online, does not contain the local editions of the newspaper and the photographic archive.

The main contribution of this work lies in the resource itself and its accessibility to the research community at large. The corpus is distributed in two formats: raw text and pre-processed. The validity of the corpus for the automatic study of language change is tested as part of the DIACR-Ita task [4] at EVALITA 2020. However, we illustrate some further potential applications of the use of the corpus.

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Ehud | Ehud | PROPN | SP | nsubj | 3 | B-PER | False | False | False | Xxxx |
| 2 | Barak | Barak | PROPN | SP | flat:name | 1 | I-PER | False | False | False | Xxxxx |
| 3 | scende | scendere | VERB | V | ROOT | 0 | O | False | False | False | xxxx |
| 4 | direttamente | direttamente | ADV | B | advmod | 3 | O | False | False | False | xxxx |
| 5 | in | in | ADP | E | case | 6 | O | False | False | True | xx |
| 6 | campo | campire | NOUN | S | obl | 3 | O | False | False | False | xxxx |
| 7 | per | per | ADP | E | mark | 8 | O | False | False | True | xxx |
| 8 | ufficializzare | ufficializzare | VERB | V | advcl | 3 | O | False | False | False | xxxx |
| 9 | la | la | DET | RD | det | 10 | O | False | False | True | xx |
| 10 | candidatura | candidatura | NOUN | S | obj | 8 | O | False | False | False | xxxx |
| 11 | dell' | dell' | DET | DD | det | 13 | O | False | False | False | xxxx' |
| 12 | ex | ex | ADJ | A | amod | 13 | O | False | False | True | xx |
| 13 | premier | premier | NOUN | S | obj | 8 | O | False | False | False | xxxx |
| 14 | laburista | laburista | PROPN | SP | amod | 13 | O | False | False | False | xxxx |

TABLE 6.1: An example of generated token features for the sentence: "*Ehud Barak scende direttamente in campo per ufficializzare la candidatura dell'ex premier laburista.*" [Ehud Barak takes the field to announce the candidacy of the former labour leader.]

### 6.1.2 Corpus Creation

The corpus creation consists of several steps:

**Downloading** All PDF files are downloaded from the source site and stored into a folder structure that mimics the publication year of each article.

[3] https://archivio.unita.news/

[4] https://diacr-ita.github.io/DIACR-Ita/

**Text extraction** The text is extracted from the PDF files by using the Apache Tika library.[5] First, the library tries to extract the embedded text if present in the PDF; otherwise the internal OCR is exploited. It is important to notice that during this step several OCR errors occur. In particular, during the processing of the early years, the newspaper has an unconventional format where a few large pages contain many articles split into several columns. Due to this format, the OCR is not able to correctly identify the column boundaries.

**Cleaning** In this step, we try to fix some text extraction issues. The previous step leaves an empty line when the end of a paragraph is reached. However, a paragraph can be composed of multiple lines which sometimes contain a word break at the end of the line. We manage word breaks in order to obtain a paragraph on a single text line; we still retain the empty line for delimiting paragraphs. Moreover, we remove noisy text by adopting two heuristics: (1) paragraphs must contain at least five tokens composed by only letter characters; (2) 60% of the paragraph must contain words that belong to a dictionary. The dictionary is built by extracting words that occur into the Paisà corpus [140] taking into account only words composed by letters. The output of this process is a plain text file for each year where each paragraph is separated by an empty line.

**Processing** All plain text files produced by the cleaning step are processed by a Python script that splits each paragraph into sentences and analyses each sentence by performing several natural language processing tasks. We rely on the spaCy[6] Python library for performing: tokenization, PoS-tagging, lemmatization, named entity recognition and dependency parsing. The spaCy library provides performance comparable to the state-of-the-art approaches with a good processing speed when compared to other NLP tools.[7] We also provide the plain text in order to allow the processing with other tools. Each plain text file is analysed and transformed in vertical format adding two tags: `<p>...</p>` for the begin and the end of a paragraph, and `<s>...</s>` for delimiting sentences. The vertical format is compliant to the CONLL representation standard and the tag-set for the Italian[8] is

[5] `https://tika.apache.org/`

[6] `https://spacy.io/`

[7] `https://spacy.io/usage/facts-figures`

[8] `https://spacy.io/api/annotation`

automatically mapped to the Universal Dependencies scheme[9].

| Feature | Description |
|---|---|
| Position | The token position in the sentence starting from 1 |
| Token | The token |
| Lemma | The lemma |
| PoS-tag | The PoS tag |
| Tag | Additional tags, such as morphological tags |
| Dependency | Dependency type |
| Head position | Head position of the dependency |
| IOB2 NE | IOB2 tag of the named entity |
| Punctuation | Boolean indicating if punctuation |
| Space | Boolean indicating if space character |
| Stop word | Boolean indicating if stop word |
| Shape | The word shape – capitalisation, punctuation, digits |

TABLE 6.2: Description of token features.

The corpus spans 67 years from 1948 to 2014. For each year, we provide two files: (1) the plain text file containing the cleaned text extracted from PDF where each paragraph is delimited by an empty line; (2) a vertical file. In the vertical file format, exemplified in Table 6.1, each paragraph is split in sentences and tokens occurring in each sentence are annotated with 12 features, whose symbols and descriptions are reported in Table 6.2.

### 6.1.3 Corpus Statistics

In this section, we report some corpus statistics. Table 6.3 illustrates the total number of occurrences and the dictionary size for each feature (token, lemma, and named entity, respectively).

| | dict. size | occurrences |
|---|---|---|
| token | 4,177,128 | 425,833,098 |
| lemma | 4,053,561 | 425,833,098 |
| named entity | 5,429,470 | 26,330,273 |

TABLE 6.3: Dictionary size and total number of occurrences.

The corpus contains more than 400 million occurrences and more than 25 million named entities occurrences. The most frequent entities are *Italia*, *Roma* and *PCI*. This result is expected since "L'Unità" was the newspaper of the Italian Communist Party.

[9] http://universaldependencies.org/u/pos/

Figure 6.1 shows the PoS-tags[10] frequency over time for open-class tags: NOUN, VERB, ADJective, ADVerb and PROPer Noun. The most frequent tag is NOUN followed by VERB, PROPN, ADJ and ADV. We observe that the frequency of PoS-tags is almost constant over time (excluding PROPN) underlying a stable language style that is typical for the news domain. We observe a variable usage of proper nouns, that may be related to the different types of events narrated over time that do not depend on a particular language style. Moreover, after the 1976, we observe a complementary trend between the adjectives and adverbs frequencies: the former slightly increase over time, while the latter decrease. This may denote a change in the language style that has varied to prefer the usage of adjectives over adverbs in more contemporary writing.

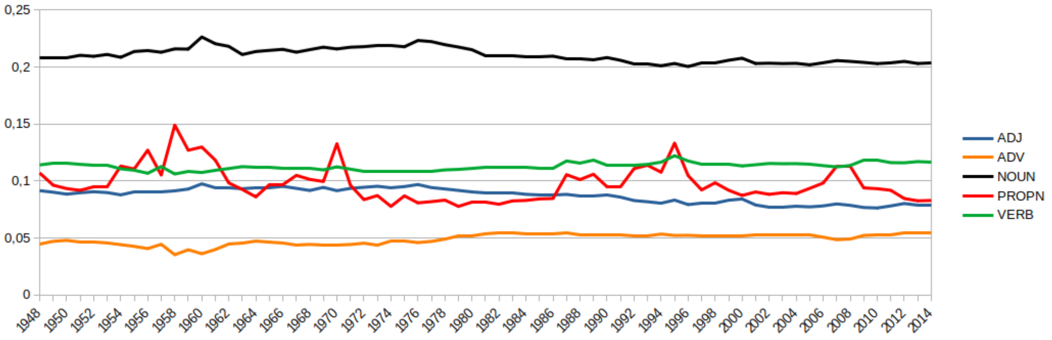


FIGURE 6.1: PoS tags frequency over time for: NOUN, VERB, ADJective, ADVerb

An interesting analysis concerns the tokens occurrences per year, whose result is plotted in Figure 6.2. We observe a low number of occurrences in the period (1948-1970), probably due to two factors: (1) the first period contains many OCR errors and noise removed during the cleaning step; (2) the number of pages of the newspaper increases over time. The latter may also explain the lower number of tokens for some of the years, such as 1981, 1995, 2000, 2007-2008, 2014. In particular, the latest years are characterised by management issues (e.g. the newspaper liquidation in July 2000) that were reflected in the newspaper format.

We also compute the time series of normalised occurrences (frequency) for each token, lemma, and named entity. All the aforementioned statistics are distributed in separate files together with the corpus.

As an illustrative example of the potential use of the corpus, in Figure 6.3 we plot the time series for two keywords. The first, *comunismo* [comunism],

[10] The used tag-set is described here `https://universaldependencies.org/u/pos/`

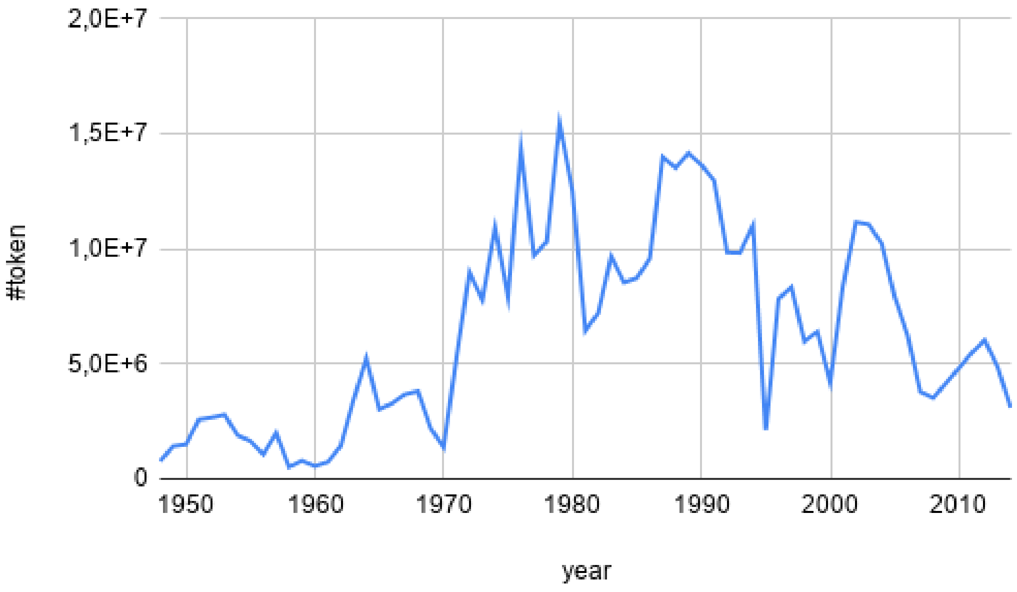


FIGURE 6.2: The plot of token occurrences per year.

is assumed to be pivotal to this corpus due to the specific role played by the newspaper in relation to the PCI. The second keyword, *antipolitica* [anti-politics], is particularly interesting as it is a term used to describe the current state of the political life in Italy, characterised by a high level of distrusts in parties and, more generally, in politics.

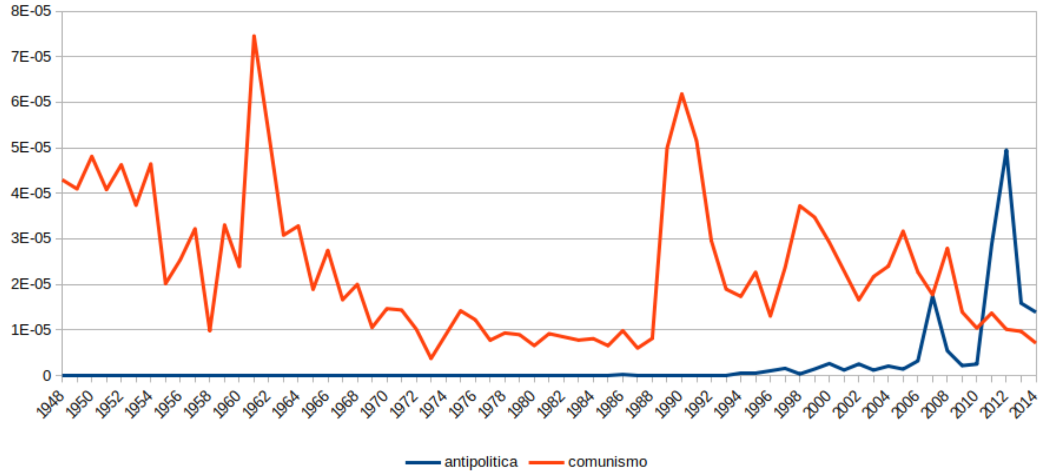


FIGURE 6.3: Plot of the time series for the words *comunismo* [comunism] and *antipolitica* [anti-politics].

The lifespan of *comunismo* [comunism] appears to be extremely influenced and characterised by history. We observe two big spikes in the time series. The first is around 1962, one of the harshest year of the Cold War, witnessing the Cuban missile crisis. The second spike is between 1989 and 1991,

corresponding to the beginning of the worldwide crisis of the communist movement and the dissolution of PCI. After 1991, the frequency of the term constantly decreases. Interestingly, the frequency for *comunismo* [comunism] is low between 1968 and 1988, a period of time that witnessed a cultural hegemony of leftist movements and strong criticism against the U.S.S.R. On the other hand, we observe that *antipolitica* [anti-politics] is a recent term whose first appearance dates back to 1977. The word frequency starts to increase slowly from 1999 and it reaches its peak in 2012 with the unexpected electoral success of the populist 5 Star Movement at the local elections in May.
Using the same approach, we plot the time series for two named entities: *PCI* and *Berlusconi*. We notice that the frequency of *PCI* start dropping in 1986, few years before its dissolution in 1991, while the name *Berlusconi* has a substantial increase in 1994 when he became the Italian Prime Minister.

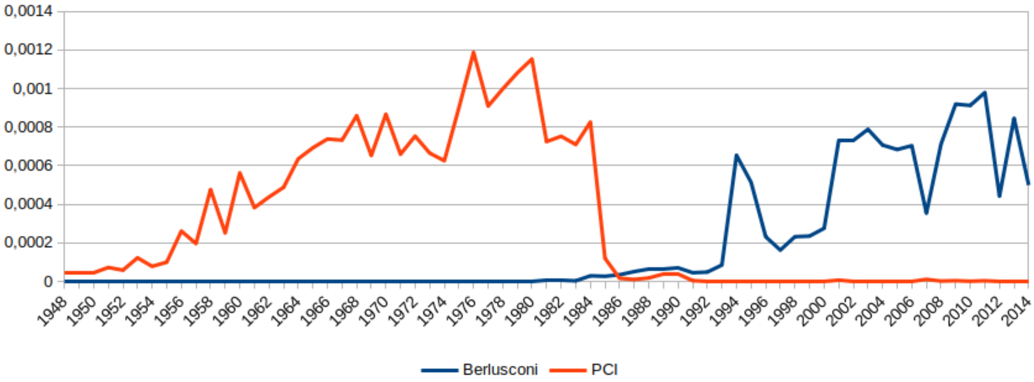


FIGURE 6.4: Plot of the time series for the entities *PCI* and *Berlusconi*.

Finally, we investigate how the vocabulary changes between two periods: $T_1 = [1948 - 1958]$ and $T_2 = [2004 - 2014]$. For each period we build the vocabulary $V_i$ taking into account only words that occur at least 10 times. Then, we compute the differences between the two dictionaries, $V_1 \setminus V_2$ and $V_2 \setminus V_1$, and sort the words in descending order by occurrences. We observe that the words *agrari, imperialisti, mezzadri, monarchici*[11] appear frequently in $T_1$ and never appear in $T_2$, conversely the words *euro, centrosinistra, centrodestra, immigrati*[12] appear only in $T_2$. A similar analysis was executed on named entities[13] and shows that *Scelba, D.C., PSI, U.R.S.S.* are specific to $T_1$, while *Berlusconi, PD, Bush, Obama* to $T_2$, revealing differences in topics and people covered by the newspaper.

[11] In English: *agrarians, imperialists, sharecroppers, monarchists*.
[12] In English: *euro, centre-left politics, centre-right politics, immigrants*.
[13] In this case we consider only entities that appear at least 5 times.

## 6.2 DIACR-Ita: A benchmark for Lexical Semantic Change Detection for the Italian language

### 6.2.1 Background and Motivation

The Diachronic Lexical Semantics (DIACR-Ita) task focuses on the automatic recognition of lexical semantic change over time, combining together computational and historical linguistics. The aim of the task can be shortly described as follows: given contextual information from corpora, systems are challenged to detect if a given word has changed its meaning over time.

Word meanings can evolve in different ways. They can undergo *pejoration* or *amelioration* (when meanings become respectively more negative or more positive) or they can be subject of *broadening* (also referred to as *generalization* or *extension*) or *narrowing* (also known as *restriction* or *specialization*). For instance, the English word *dog* is a clear case of broadening, since its more general meaning came from the late Old English "dog of a powerful breed" [223]. On the contrary, the Old English word *deor* with the general meaning of "animal" became *deer* in present-day English. Semantic changes can be further classified on the basis of the cognitive process that originated them, i.e. either from *metonymy* or *metaphor*. Lastly, it is possible to distinguish among changes due to language-internal or language-external factors [93]. The latter usually reflects a change in society, as in the case of technological advancements (e.g. *cell*, from the meaning of "prisoner cell" to "cell phone").

The problem of the automatic analysis of lexical semantic change is gaining momentum in the Natural Language Processinng (NLP) and Computational Linguistics (CL) communities, as shown by the growing number of publications on the diachronic analysis of language and the organisation of related events such as the 1st International Workshop on Computational Approaches to Historical Language Change[14] and the project "Towards Computational Lexical Semantic Change Detection"[15]. Following this trend, SemEval 2020 has hosted for the first time a task on automatic recognition of lexical semantic change: the SemEval 2020 Task 1 - Unsupervised Lexical Semantic Change Detection[16] [205]. While this task targets a number of different languages, namely Swedish, Latin, and German, Italian is not present.

Many are the existing approaches, data sets, and evaluation strategies used to detect semantic change, or drift. Most of the approaches rely on diachronic

[14] https://languagechange.org/events/2019-acl-lcworkshop/

[15] https://languagechange.org/

[16] https://competitions.codalab.org/competitions/20948

word embeddings, some of these are created as post-processing of static word embeddings, such as Hamilton, Leskovec, and Jurafsky [86]; while others create dynamic word embeddings where vectors share the same space for all time periods [55, 236, 192, 61]. Recent work exploits word sense induction algorithms to discover semantic shifts [217, 95] by analyzing how induced senses change over time. Finally, Gonen et al. [80] propose a simple approach based on the neighbors' intersection between two corpora. The neighborhood of a word is separately computed in each corpus, then the intersection is exploited to compute a measure of the semantic shift. The neighborhood in each corpus can be computed using the cosine similarity between word embeddings built on the same corpus without using vectors alignment. A more complete state of the art is described in a critical and concise way in the latest surveys [216, 118, 219].

Almost all of the previously mentioned methods use English as the target language for the diachronic analysis, leaving the other languages still underexplored. To date, only one evaluation has been carried out on Italian using the Kronos-it dataset [13].

The DIACR-Ita task at the EVALITA 2020 campaign [20] fosters the implementation of new systems purposely designed for the Italian language. To achieve this goal, a new dataset for the evaluation of lexical semantic change on Italian has been developed based on the "L'Unità" corpus [14]. This is the first Italian dataset manually annotated with semantic shifts between two different time periods.

### 6.2.2 Task Description

The goal of DIACR-Ita is to establish if a set of *target* words change their meaning across two time periods, $T_1$ and $T_2$, where $T_1$ precedes $T_2$.

Following the SemEval 2020 Task 1 settings, we focus on the comparison of two time periods. In this way, we tackle two issues:

1. We reduce the number of time periods for which data has to be annotated;

2. We reduce the task complexity, allowing for the use of different models' architectures, and thus widening the range of potential participants.

During the test phase, participants have been provided with two corpora $C_1$ and $C_2$ (for the time periods $T_1$ and $T_2$, respectively), and a list of target words. For each target word, systems have to decide whether the word changed or not its meaning between $T_1$ and $T_2$, according to its occurrences in sentences in $C_1$ and $C_2$. For instance, the meaning of the word "imbarcata" (i.e. *embarked*) is known to have expanded, i.e. it has acquired a new sense, from $T_1$ to $T_2$. The word originally referred to an acrobatic manoeuvre of aeroplanes. Nowadays, it is also used to refer to the state of being deeply in love with someone. This will be reflected in different occurrences of the word usage in sentences between $C_1$ and $C_2$.

The task is formulated as a closed task, i.e. participants must train their model only on the data provided in the task. However, participants may rely on pre-trained word embeddings, but they cannot train embeddings on additional diachronic Italian corpora, they can use only synchronic corpora.

### 6.2.3 Data

This section provides an overview of the datasets that were made available to the participants in the two different stages of the evaluation challenge, namely **trial** and **test**.

#### Trial data

The trial phase corresponds to the evaluation window in which the participants have to build their systems before the official test data are release. The following data were provided:

- An example of 5 trial target words for which predictions are needed;
- An example of gold standard for the trial target words;
- A sample submission file for the trial target words;
- Two trial corpora that participants could use to develop their models and check the compliance of the generated output to the required format;
- An evaluation and some additional utility scripts for managing corpora.

Trial data do not reflect the actual data from $C_1$ and $C_2$. The sample training corpora and target words were artificially built just to provide an example of the data format for developing their systems. Since the training corpus is publicly available on the Internet, we decided not to release these data during the trial phase to prevent participants from identifying the source data and consequently potential set of target words.

### Test data

For the test phase, the following data were provided:

- A diachronic split of the "L'Unità" corpus into the two sub-corpora, $C_1$ and $C_2$, each belonging to a specific time period;
- 18 target words, among which 6 were identified as target of semantic meaning change between the two time periods.

**Corpus Creation** The "L'Unità" diachronic corpus [14] is a collection of documents extracted from the digital archive of the newspaper "L'Unità".[17]
For the task, the corpus has been initially split into two sub-corpora, $C_1$, corresponding to the time period $T_1 = [1945 - 1970]$, and $C_2$, corresponding to the time period $T_2 = [1990 - 2014]$.
To facilitate participants in the closed-task formulation, the corpora were provided in a pre-processed format. In particular, we adopted a tab separated format, with one token per line. For each token, we provided its corresponding part-of-speech and lemma. Sentences are separated by empty lines. Data were pre-processed with UDPipe[18] using the ISDT-UD v2.5 model.
Participants are free to combine the available information as they want. Furthermore, to facilitate the generation of word embeddings, we made available a script for generating a format containing one sentence per line.
The whole "L'Unità" diachronic corpus has been built, cleaned and annotated automatically. This process consisted of several steps, namely:

**Step 1: Downloading** All PDF files are downloaded from the source site and stored into a folder structure that mimics the publication year of each article.

**Step 2: Text extraction** The text is extracted from the PDF files by using the Apache Tika library.[19] First, the library tries to extract the embedded text if present in the PDF. If this process fails, the internal OCR system is used. It is important to notice that during this step several OCR errors may occur due to different reasons. The processing of the early years of publications, i.e., between 1945–1948, represented a non trivial challenge for the extraction of the

[17] https://archivio.unita.news/
[18] http://lindat.mff.cuni.cz/services/udpipe/run.php
[19] https://tika.apache.org/

textual data. In particular, we noticed that the page format had a major impact on the quality of the OCR. In these period, the newspaper has quite an unconventional format where a few large pages contain many articles scattered into several columns. This affected the performance of the OCR due to its failure in properly identifying the column boundaries.

**Step 3: Cleaning** In this step, we try to fix some text extraction issues. We identified two lines of actions, the first dealing with paragraph splits and the second with noisy text. In the text extraction process, paragraphs are separated by means of an empty line. However, word hyphenation can trigger errors in the paragraph segmentation phase by wrongly adding empty lines. We addressed this issue by reconstructing the paragraph on a single text line, thus ensuring that empty lines are only used to delimit the actual paragraphs. In our case, noisy text corresponds to tokens whose composing characters are wrongly interpreted by the OCR mixing together alphabetical characters with numbers or symbols. Two heuristics were implemented to limit the amount of noisy text. The first heuristic requires that paragraphs must contain at least five tokens composed by only alphabetical characters. The second heuristic requires that at least 60% of each paragraph must contain words that are attested in a dictionary. For this, we did not use a reference dictionary, but we automatically created it by extracting tokens from the Paisà corpus [140]. Numbers were excluded and only alphabetical strings were retained. The output of the cleaning process is a plain text file for each year where each paragraph is separated by an empty line.

**Step 4: Processing** All plain text files produced by the cleaning step are processed by a Python script that splits each paragraph into sentences and analyses each sentence with UDPipe [20] ISDT-UD v2.5 model. In this way, we obtain tokens, part-of-speech tags, and lemmas. The processed data are then stored in a vertical format as illustrated is Section 6.2.3.

After these preparation steps, the valid and retained data for the task span over a temporal period between 1948 and 2014. We revised the initial split of the two sub-corpora as follows: $C_1$ ranges between $T_1 = [1948 - 1970]$, and $C_2$ between $T_2 = [1990 - 2014]$. Table 6.4 illustrates the distributions of the tokens across the two time periods for the sub-corpora. The difference in the number of tokens between $C_1$ and $C_2$ reflects differences in the trends in

[20] `http://lindat.mff.cuni.cz/services/udpipe/run.php`

the number of daily published articles, due to cheaper printing costs and the availability of new technologies such as the World Wide Web.

| Corpus | Period | #Tokens |
|---|---|---|
| L'Unità | 1948-1970 | 52,287,734 |
| L'Unità | 1990-2014 | 196,539,403 |

TABLE 6.4: Official Training Corpora: Occurrence of Tokens.

**Creation of the Gold Standard** The selection of the target words that compose the Gold Standard data required a manual annotation. Identifying words that have undergone a semantic change is not an easy task. To boost the identification of candidate target words, we adopted a semi-automatic method. In the following paragraphs we illustrate in detail our approach.

**Step 1: Selection of candidate words.** The initial selection of potential candidate words was based on Kronos-IT [13]. Kronos-IT is a dataset for the evaluation of semantic change point detection algorithms for the Italian language automatically built by using a web scraping strategy. In particular, it exploits the information presents on the online dictionary "Sabatini Colletti"[21] to create a pool of words that have undergone a semantic change. In the dictionary, some lemmas are tagged with the year of the first attestation of its sense. In some cases, associated with the lemma there are multiple years attesting the introduction of new senses for that word. Kronos-IT uses this information to identify the set of semantic changing words. We retained those words that were predicted to have changed their meaning after 1970, so as to match the temporal periods of the sub-corpora. In this way, we obtained 106 candidate lemmas.

**Step 2: Filtering candidate targets.** A challenging issue is the attestation of the potential candidate words in both sub-corpora with a relatively high number of occurrences to account for different contexts of use. Frequency, indeed, plays a quite relevant role for the task: infrequent tokens must be discarded because they affect the quality of word representations.The initial list of candidate targets has been further cleaned by removing all tokens that occur less than 20 times in each corpora. . Moreover, we conducted a further analysis by manually inspecting some randomly sampled lemma contexts. The aim of this analysis was

[21] https://dizionari.corriere.it/dizionario_italiano/

to remove targets for which the lemmas occurrences are affected by OCR errors. This analysis was performed by the means of the Sketch Engine[22], in particular we analyze concordances of the target word in order to discover OCR errors. One of such words was "toro" derived from the mistaken OCR of "loro". At the end of this process, we obtained a list of 27 candidate targets for the annotation.

**Step 3: Manual Annotation.** For each target, we randomly extracted up to 100 sentences from each of the sub-corpus[23]. Each sentence was then annotated by two annotators: they were asked to assign each occurrence to one of the meaning of the lemma according to those reported in the Sabatini-Coletti dictionary. In case the meaning of the word in a sentence was not present in the list of senses reported in the reference dictionary, the annotators were allowed to add the sense to the word. In total, we annotated 2,336 occurrences of the candidate target words.

**Step 4: Annotation check.** All cases of disagreement were collectively discussed among all of the annotators to reach a final decision. We observed that some disagreements were also due to a biased interpretation of the context of occurrence by one of the annotators. These cases mainly concerned short ambiguous sentences that prevented a clear identification of the word meaning. As a result of this step, a few candidates were removed from the pool of candidates because occurring in too ambiguous context.

**Step 5: Creation of the gold standard.** We retained as valid instances of lexical semantic change all those targets that had occurrences of one specific sense only in $T_2$, and never in $T_1$. In other words, in the context of this task, a valid lexical semantic change corresponds to the acquisition of a new meaning by a target word. Out of the 23 candidate target words, only 6 of them show a semantic change in $T_2$. All the other targets did not show a diachronic meaning change. In the final Gold Standard, we kept 12 candidate target words that did not change meaning obtaining a final set of 18 target words.

The Gold Standard contains 18 targets listed as lemmas, one lemma per line, with an accompanying label to mark whether the lemmas has undergone

[22] `https://www.sketchengine.eu/`

[23] This means that in case a target words occurs less than 100 times, all occurrences were annotated.

semantic change (label 1) or not (label 0). Participants were given a file containing the 18 target lemmas, one per each line, without annotation. The expected system output is a modification of this file where the participant had to annotate each target lemma with the system prediction (0 or 1).

### 6.2.4 Evaluation

The task is formulated as a binary classification problem. Systems predictions are evaluated against the change labels annotated in the Gold Standard by using accuracy.
The test set ($G$) contains both positive ($P$) and negative ($N$) examples, i.e. $G = P \cup N$. For example:

$$P = \{pilotato, lucciola, ape, rampante\}$$

$$N = \{brama, processare\}$$

Negative words are those that did not undergo a change in their meaning. Systems' predictions involve both positive and negative classified targets $Pr = Pr_{pos} \cup Pr_{neg}$. Then, true positives (positive targets classified as positive) are $TP = P \cap Pr_{pos}$, true negatives (negative targets classified as negative) are $TN = N \cap Pr_{neg}$, false negatives (positive targets classified as negative) are $FN = P \cap Pr_{neg}$ and false positives (negative targets classified as positive) are $FP = N \cap Pr_{pos}$.
We can then compute the accuracy as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

#### Baselines

We provided two baseline models:

- Frequencies: The absolute value of the difference between the word frequencies in the two sub-corpora;
- Collocations: For each word, we build two vector representations consisting of the Bag-of-Collocations related to the two different time periods ($T_0$ and $T_1$). Then, we compute the cosine similarity between the two BoCs. It is the same approach evaluated in [13].

In both baselines, we use a threshold to predict if the word has changed its meaning. While for the frequencies, a change is detected when the difference is higher than the average. For the collocations a semantic change occurs when the similarity between the two time periods drops under the average plus the variance. Both the average and the variance are computed on the set of target words.

| System | Type |
|---|---|
| OP-IMS | Post-alignement |
| UWB Team | Post-alignement |
| CIC-NLP | PoS tag features |
| UNIMIB | Jointly alignment |
| QMUL-SDS | Jointly alignment |
| VI-IMS | Jointly alignment |
| CL-IMS | Contextual Embeddings |
| unipd | Contextual Embeddings |
| SBM-IMS | Graph |

TABLE 6.5: Systems types.

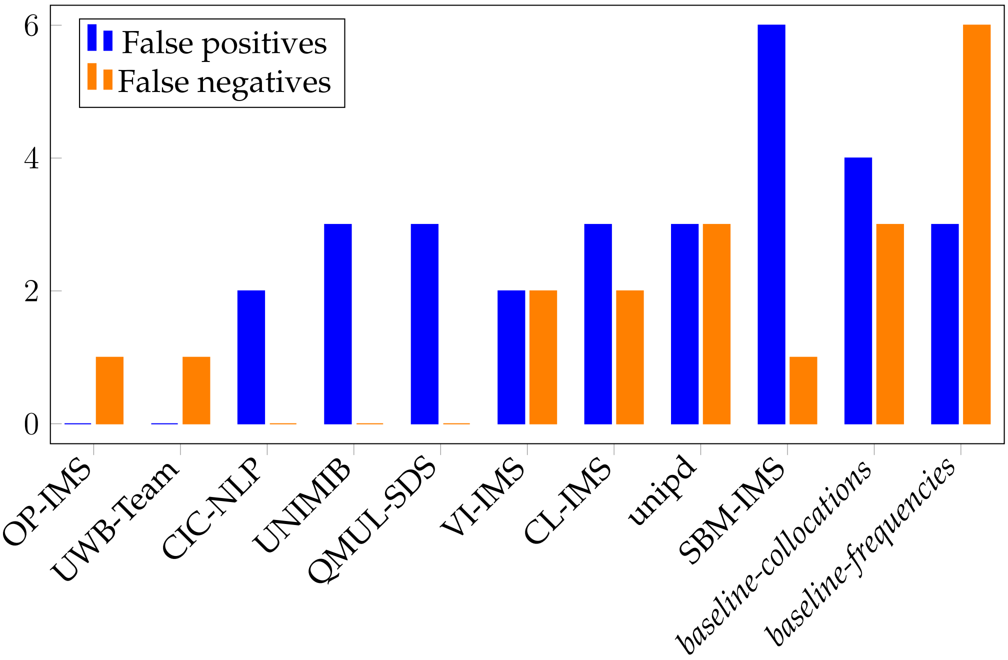


FIGURE 6.5: Number of false positives and false negatives for each system.

### 6.2.5 Systems

21 teams registered to the DIACR-Ita task. However, 9 teams participated in the final task for a total of 36 submitted runs. Based on the algorithms

employed, we can group systems into four categories: Post-alignment, Joint Alignment, Contextual Embeddings, Graph-based and PoS tag features (see Table 6.5). The first two classes are characterised by the type of alignment used. Post-alignment systems first train static word embeddings for each time periods, and then align them. Joint Alignment systems train word embeddings and jointly align vectors across all time slices. Contextual Embeddings systems use contextualized embeddings, such as BERT [59]; while Graph-based systems rely on graph algorithms. PoS tag features system rely on the distribution of targets PoS tags across the two time periods. The majority of participating systems use cosine distance as a measure of semantic change, i.e. compute the cosine distance between the vectors of the target lemmas among time periods. Other systems use the Average Pairwise Cosine Distance or the Average Canberra Distance, since the cosine distance does not fit contextual embeddings representations. The last group of systems uses graph-based measures.
We report a short description of each team (best submission) as follows:

**OP-IMS** [102] This team uses Skipgram model with Negative sampling (SGNS) to compute word embeddings, the resulting matrices are mean-centred. Word embeddings are aligned using Orthogonal Procrustes. They choose cosine similarity to compare vectors of different word spaces and a threshold based on mean and standard deviation to classify target words.

**UWB Team** [174] The team maps semantic spaces using linear transformations, such as Canonical Correlation Analysis and Orthogonal Transformation and cosine similarity as a measure to decide if a target word is stable or not. They use a threshold based on mean.

**CIC-NLP** [6] This team analyses the Part-Of-Speech distribution over the two corpora and create vectors with information about the most common word POS-tags. Then, they obtain a score using pairs of vectors of the two time periods and the sum of Euclidean, Manhattan and cosine distance. They rank targets in discerning order. Finally, they label first upper-third targets as changed words.

**UNIMIB** [22] The team creates temporal word embeddings using Temporal Word Embeddings with a Compass (TWEC) [38]. They use the move measure, i.e. a weighted linear combination of the cosine and Local Neighbors, introduced by [86]. They label targets as stable if the move measure is greater than 0.7.

**QMUL-SDS** [5] The team uses TWEC [38] to compute temporal word embeddings with TWEC C-BoW model (Continuous Bag of Words) default settings. They use a cosine similarity as measure of change and a threshold based on mean.

**VI-IMS** The team uses SGNS to create word embeddings exploiting Vector Initialization [108]. They use cosine distance as a measure of semantic change and a threshold based on the mean and the standard deviation to classify targets words.

**CL-IMS** [122] The team creates word vectors using different combinations of the first and last four layers of BERT. They rank targets according to Average Pairwise Cosine Distance, and label the first 7 targets as changed words.

**unipd** [24] This team uses contextualised word embeddings and an linear combination of distances metrics to measure semantic change, namely Euclidean Distance, Average Canberra distance, Hausdorff distance, as well as Jensen–Shannon divergence between cluster distributions. They rank targets according to the score obtained, and label the first half as changed words.

**SBM-IMS** The team compute token vectors using BERT. They create a graph where the vertices are the vectors extracted from BERT, while the edges are the cosine distance between word vectors. They cluster the graph with Weighted Stochastic Block Model. Then, they consider the number of incoming edges from the first and second period as a measure of semantic change.

### 6.2.6 Results

Table 6.6 reports the final results. The best result has been achieved by two systems: *OP-IMS* and *UWB-Team*. Both systems exploit post-alignment strategy. The second system *CIC-NLP* uses an approach based on PoS tag features. QMUL-SDS and VI-IMS are based on joint alignment, while *unipd* and *SBM-IMS* use contextual embeddings. The last system *SBM-IMS* is the only graph-based approach. Moreover, we report both false negative and false positives in Figure 6.5. Both post-alignment systems share the same unique false negative: the target "tac", while *CIC-NLP* detects two false positives. Joint-alignment systems have a number of false positives higher or at least

| **Team** | **Accuracy** |
|---|---|
| **OP-IMS** | 0.944 |
| **UWB Team** | 0.944 |
| CIC-NLP | 0.889 |
| UNIMIB | 0.833 |
| QMUL-SDS | 0.833 |
| VI-IMS | 0.778 |
| CL-IMS | 0.722 |
| unipd | 0.667 |
| SBM-IMS | 0.611 |
| *baseline-collocations* | 0.611 |
| *baseline-frequencies* | 0.500 |

TABLE 6.6: Results.

equal to the number of false negatives. *CL-IMS* and *unipd* produce respectively 2 and 3 false negatives and both misclassify three stable words. The only graph-based approach, *SBM-IMS*, reports the highest number of false positives. In conclusion, the results show that systems based on post/joint alignment and PoS tag features achieve the best performance, while contextual embeddings do not perform as good in this type of task. However all the systems outperform both the baselines.

# Chapter 7

# Lexical Semantic Change Detection

## 7.1 XL-LEXEME: WiC Pretrained Model for Cross-Lingual LEXical sEMantic changE

The LSC Detection task implicitly aims to disambiguate synchronic word sense occurrences and then find differences in the word sense frequencies in different periods. Word Sense Disambiguation (WSD) is a long-studied task in Natural Language Processing [163], which consists of associating the correct sense to a word occurring in a specific context. WSD involves some crucial issues, such as relying on a fixed sense inventory. Fixed sense inventories ignore the diachronic aspect of language because they can miss older unused senses or be outdated and missing new senses.

The Word in Context task (WiC) [172] aims to overcome these issues. In this work, we train a model on the WiC task and then use it to perform LSC Detection. In the WiC task, given the word *w* and two different contexts *C1*, *C2*, the systems have to determine whether the meaning of *w* is the same in the two contexts or not. Our approach is grounded on the assumption that models trained on the WiC tasks are robust enough to transfer the knowledge learned in a synchronic setting to a diachronic one. We summarise the main contribution of this work as follows: (i) We propose a pre-trained bi-encoder model, called XL-LEXEME, on a large-scale dataset for the WiC task, which allows us to obtain comparable lexical-based representations; (ii) We assert the effectiveness of XL-LEXEME despite the computational limitation compared to the cross-encoder architecture for the LSC Detection task; (iii) Experiments on the LSC Detection task show that XL-LEXEME outperforms state-of-the-art LSC Detection models for English, German, Swedish, and Russian.

### 7.1.1 XL-LEXEME

Generally, for pairwise sentence similarity tasks, BERT models use a cross-encoder, in which the pairwise sequences are jointly encoded, and the overall vectors are used for the classification. However, in several tasks, the cross-encoder is not suitable since it cannot provide a distinct meaningful representation for each sentence. An approach to overcome this issue involves pooling the BERT output encoded vectors, which often results in worse performance. Sentence-BERT (SBERT) [183] overcomes the limitation of cross-encoders using a Siamese Network, i.e., the weights of the underlying networks are shared. SBERT encodes the two sequences separately in the BERT model exploiting the Siamese architecture. The sequence-level representation is obtained by averaging the output encoded vectors, which are directly compared using similarity measures such as cosine similarity.
Meanwhile, cross-encoders perform better since they are trained to profit from the attention over the whole input. In this work, we introduce XL-LEXEME[1] which mirrors models for pairwise sequence similarity tasks and adapts them to the WiC task, giving prominence to the target word, i.e. the word for which we want to detect the LSC. The model takes as input two sequences $s_1$ and $s_2$. The sequences are tokenized using subwords tokenizer, such as Sentence Piece [112], and the special tokens <t> and </t> are used as target word delimiters [235]:

$$\begin{aligned} s_1 &= w_1, ..., \text{<t>}, w_i^t, ..., w_{i+k}^t, \text{</t>}, ..., w_N \\ s_2 &= w_1, ..., \text{<t>}, w_j^t, ..., w_{j+p}^t, \text{</t>}, ..., w_M \end{aligned} \tag{7.1}$$

where $N$ and $M$ represent the number of subwords of the sequence $s_1$ and $s_2$ respectively, while $w_i^t, ..., w_{i+k}^t$ and $w_j^t, ..., w_{j+p}^t$ are the subwords of the target words. In the following, we describe the baseline cross-encoder and XL-LEXEME based on a bi-encoder. For the cross-encoder, the two input sequences are concatenated by the special token $[SEP]$ in an overall sequence $s = [CLS]\ s_1\ [SEP]\ s_2\ [SEP]$. If the length of $s$, i.e. $N + M + 3$, is greater than the maximum sequence length $\lambda$, then the sequence $s$ is cut such that the length of $s_1$ and $s_2$ is less than $\lambda^* = \frac{\lambda-3}{2}$. To comply with the maximum length, the left and right contexts of the sequence are truncated. For instance,

[1] The XL-LEXEME code is available on GitHub `https://github.com/pierluigic/xl-lexeme`. The XL-LEXEME model is available in the Hugging Face Model Hub `https://huggingface.co/pierluigic/xl-lexeme`.

$s_1$ is truncated as follows:

$$s_1 = w_{n_0}, ..., \text{<t>}, w_i^t, ..., w_{i+k}^t, \text{</t>}, ..., w_{n_1} \tag{7.2}$$

where $n_0 = \max(0, i - 1 - \frac{\lambda^* - k - 2}{2})$ and $n_1 = \min(N, i + k + 1 + \frac{\lambda^* - k - 2}{2})$. The truncated sequence has a length $\gamma < \lambda$. The encoded representations of each subword $(v_1, v_2, ..., v_\gamma)$ are summed to get the encoded representation of the overall sequence, i.e. $s^{enc} = \sum_i^\gamma v_i$. Finally, the vector $s^{enc}$ is used to compute the logits:

$$logit = \log \sigma(W s^{enc}) \tag{7.3}$$

where $W \in \mathbb{R}^{1 \times d}$. The model is trained to minimize the Binary Cross-entropy loss function.

XL-LEXEME is a bi-encoder that encodes the input sequences using a Siamese Network into two different vector representations. Each sequence is tokenized and truncated according to the maximum length $\lambda^*$, using Equation (7.2). We thus obtain the new lengths $\gamma_1, \gamma_2$. The vector representation is computed as the sum of the encoded subwords $(v_1, v_2, ..., v_\gamma)$, i.e. $s_1^{enc} = \sum_i^{\gamma_1} v_i$ and $s_2^{enc} = \sum_j^{\gamma_2} v_j$.

XL-LEXEME is trained to minimize the Contrastive loss [84]:

$$\ell = \frac{1}{2} \left[ y \cdot \delta^2 + (1 - y) \cdot \max(0, m - \delta)^2 \right] \tag{7.4}$$

where we adopt a margin $m = 0.5$. We use as default distance $\delta$ the cosine distance between the encoded representations of $s_1$ and $s_2$, i.e. $\delta = \cos(s_1^{enc}, s_2^{enc})$. The main advantage of XL-LEXEME concerning models based on the cross-encoder architecture is efficiency. The time cost can be directly derived from the different architectures that exploit XL-LEXEME and the cross-encoder baseline. The self-attention time complexity $O(N^2 * d)$ depends on the vector dimension $d$ and the sequence length, which is $N$ for the cross-encoder and $\frac{N}{2}$ for XL-LEXEME. For XL-LEXEME, the time complexity is reduced to $O((\frac{N}{2})^2 * 2d)$.

## 7.1.2 Experimental setting

### Training details

XL-LEXEME and the cross-encoder are trained using XLM-RoBERTa (XLM-R) [51] large as the underlying Language Model[2] and using an NVIDIA

[2] The XLM-R model is fine-tuned during the training.

GeForce RTX 3090. As for training data, the model uses the training data of MCL-WiC [144], AM$^2$ICO [137], and XL-WiC datasets [181] merged with the randomly sampled 75% of the respective development data of each dataset. The remaining 25% of the development data is used to fine-tune hyper-parameters. Moreover, we augment training data for the cross-encoder by swapping the order of sentences in the training set [144].

We use AdamW optimizer and linear learning warm-up over the 10% of training data. We perform a grid search for the hyper-parameters optimization, tuning the learning rate in {1e-6, 2e-6, 5e-6, 1e-5, 2e-5} and the weight decay {0.0, 0.01}. Table A.1 (Appendix A.1) shows the selected hyper-parameters. We sample 200 sentences containing the target word for each language and each period. The sampling is repeated ten times, and the results are averaged over the ten iterations. We use the same methodology of Rachinskiy and Arefyev [178] for sampling sentences from the RuShiftEval corpora. We sample sentences in which we find the exact match with the target words with no pre-processing of the SemEval dataset. The LSC score is computed as the average distance between the vectors over the two different periods:

$$\text{LSC}(s^{t_0}, s^{t_1}) = \frac{1}{N \cdot M} \sum_{i=0}^{N} \sum_{j=0}^{M} \delta(s_i^{t_0}, s_j^{t_1}) \tag{7.5}$$

where $\delta$ is the distance measure, i.e. $\delta = 1 - \log \sigma(W s^{enc})$ for the cross-encoder baseline and $\delta = \cos(s_1^{enc}, s_2^{enc})$ for XL-LEXEME.

| | *SemEval-2020 Task 1 Subtask 2 Leaderboard* | | | | | | Temporal BERT | | cross-encoder | XL-LEXEME |
|---|---|---|---|---|---|---|---|---|---|---|
| Lang. | UG_Student _Intern | Jiaxin & Jinan | cs2020 | UWB | Count baseline | Freq. baseline | TempoBERT | Temporal Attention | | |
| EN | 0.422 | 0.325 | 0.375 | 0.367 | 0.022 | -0.217 | 0.467 | †0.520 | †0.752 | **0.757** |
| DE | 0.725 | 0.717 | 0.702 | 0.697 | 0.216 | 0.014 | - | †0.763 | †0.837 | **0.877** |
| SV | †0.547 | †0.588 | †0.536 | †0.604 | -0.022 | -0.150 | - | - | †0.680 | **0.754** |
| LA | 0.412 | **0.440** | 0.399 | 0.254 | 0.359 | †0.020 | 0.512 | 0.565 | †0.016 | -0.056 |
| Avg. | 0.527 | 0.518 | 0.503 | 0.481 | 0.144 | -0.083 | - | - | 0.571 | **0.583** |

TABLE 7.1: Results (Spearman correlation) on the SemEval-2020 Task 1 Subtask 2 test set. The symbol † indicates there is no statistical difference with the correlation obtained by XL-LEXEME.

### 7.1.3 Results

Table 7.1 and Table 7.2 report the results on the SemEval-2020 Task 1 Subtask 2 and the results on the RuShiftEval test set. The results of the best systems are in bold. XL-LEXEME achieves the best score for English, German, Swedish, RuShiftEval1, RuShiftEval2, and RuShiftEval3. XL-LEXEME

| | *RuShiftEval Leaderboard* | | | | cross-encoder | XL-LEXEME | XL-LEXEME (Fine-tuned) |
|---|---|---|---|---|---|---|---|
| Dataset | GlossReader | DeepMistake | UWB | Baseline | | | |
| RuShiftEval1 | †0.781 | †0.798 | 0.362 | 0.314 | †0.727 | 0.775 | **0.799** |
| RuShiftEval2 | †0.803 | †0.773 | 0.354 | 0.302 | †0.753 | 0.822 | **0.833** |
| RuShiftEval3 | †0.822 | †0.803 | 0.533 | 0.381 | †0.748 | 0.809 | **0.842** |
| Avg. | 0.802 | 0.791 | 0.417 | 0.332 | 0.743 | 0.802 | **0.825** |

TABLE 7.2: Results (Spearman correlation) on the RuShiftEval test set. The symbol † indicates there is no statistical difference with the correlation obtained by XL-LEXEME.

achieves a strong Spearman correlation for English and Swedish languages and a solid correlation on the German dataset, obtaining a significative correlation ($p$ < 0.001). XL-LEXEME obtains no significant results in the Latin language since the predicted scores for the target words are not correlated with the test set. Latin is underrepresented in the training data of XLM-R, and there are no similar languages in the WiC dataset that we use for training XL-LEXEME. Moreover, the Latin dataset is more challenging as it involves the first corpus written in ancient Latin, which differs in many aspects from modern Latin. For this reason, XL-LEXEME could be ineffective in ancient languages and, in general, in languages that are not widely covered by the WiC dataset.

We report the statistical significance of the difference between the performance of XL-LEXEME concerning the other models. The statistical significance of the difference is computed using Fisher's $z$-transformation [176]. XL-LEXEME obtains stronger correlations than the cross-encoder, but the differences are not significant. The correlations obtained on the English and the German datasets are significantly different ($p$ < 0.05) for all the systems that participated in the SemEval-2020 Task 1 but not for TempoBERT and Temporal Attention. On the other side, TempoBERT and Temporal Attention obtain a Spearman correlation on English and German that is not statistically different from the systems on the SemEval-2020 Task 1 leaderboard. In the Swedish language, XL-LEXEME is the only one obtaining a significantly different correlation from the Count baseline results. XL-LEXEME showed its effectiveness also in Swedish, although the WiC dataset does not cover this language. Presumably, Swedish benefits from the presence of other languages descending from the Old Norse language, namely Danish and Norwegian.

XL-LEXEME obtains competitive results for the Russian language in the RuShiftEval leaderboard. Contrary to XL-LEXEME, Deep Mistake and Gloss Reader are fine-tuned on the RuSemShift dataset. The differences between XL-LEXEME and the best two systems in the leaderboard are not statically significant. Moreover, in Table 7.2, the results of XL-LEXEME fine-tuned on the RuSemShift are shown. Although the fine-tuned model achieves the best correlation scores in the three datasets, the difference between DeepMistake and GlossReader is not significant.

### 7.1.4 Conclusion

XL-LEXEME is pre-trained on a large WiC dataset to mirror sentence-level encoders focusing on specific words in contexts. We evaluated XL-LEXEME on two Lexical Semantic Change Detection datasets: SemEval-2020 Task 1 and RuShiftEval. XL-LEXEME outperforms state-of-the-art models for LSC Detection in English, German, Swedish, and Russian datasets, with significant differences from the baselines. The XL-LEXEME effectiveness and efficiency make it reliable for LSC Detection on large diachronic corpora.

### 7.1.5 Limitations

While the vector representations obtained using XL-LEXEME for different languages are potentially comparable, lying on the same geometric space, the evaluation of cross-lingual semantic changes cannot be performed for lacking cross-lingual LSC Detection resources. SemEval 2020 Task 1 datasets consist of small sets of target words, i.e., the number of target words for English, German, Latin, and Swedish is 37, 48, 40, and 31, respectively. The example of the Latin language highlights that XL-LEXEME can perform poorly on languages that are underrepresented in the training set of XLM-R and not covered by the WiC dataset. Generally, at the moment is not possible to state precisely how and how much XL-LEXEME performance is affected by the language distribution in the XLM-R training set and the WiC dataset.

## 7.2 Tug of War: Studying the Contextualisation of Pre-trained BERT models

The major novelty and advancement that contextualised BERT models have brought to language modelling is the ability to dynamically generate vector representations based on specific usage context: we can thus easily differentiate between *sitting on a rock* and *listening to rock*.

While BERT models are *pre-trained* on huge all-purpose corpora, typically with a large emphasis on web corpora, researchers and practitioners employ them for diverse text applications. Regardless of how well the information and language in the studied text align with the pre-training text, the models are used to generate word vector representations (i.e. embeddings) for any input sequence. However, this implies a gap between where the models have learned all of their parameters, and the data on which they are applied. Indeed, BERT models serve as the lens through which we view the studied texts: if our texts are contemporary with the pre-training, the gap is likely to be minimal; if, however, we intend to study historical or other out-of-domain (OOD) corpora, this gap can be arbitrarily large and have major effects on follow-up studies.

The effects can be exemplified in the area of Lexical Semantic Change (LSC) where the state of the art postulates a standard recipe: a word is modelled by its contextualised embeddings for different time periods [161]. These representations are then compared over time to detect change. However, if BERT models impose too much pre-trained contemporary knowledge in a historical context, they are unable to generate accurate representations of the historical meaning of a word. For instance, the historical expression *sit at meat* denotes taking a seat at the dining table or joining others for a meal[3], as *meat* originally referred to "any kind of food". Over time, the term underwent semantic narrowing and now typically refers to animal flesh used as food. Models relying heavily on contemporary pre-trained knowledge may not faithfully capture the authentic historical meaning of "meat". This could result in an underestimation of semantic change. Similar effects may arise when studying language variation across speaker communities or in out-of-domain contexts, such as detecting hate speech [164], radicalization [197], and dog whistles [56]. By studying the degree of influence that the model exerts versus the context, we can offset this in subsequent modelling, and build more contextualised models.

Some previous exploration of the contextualisation has been done, typically through probing tasks [98], or by analysing the geometry of the vector representations [64]. We investigate to what degree the representation of a target word is determined by BERT's pre-trained knowledge and contra the context in which it appears by analysing the effect of substituting a *target* word with

[3] An example is the passage from the King James Bible (Luke 14:10): "But when thou art bidden, go and sit down in the lowest room; that when he that bade thee cometh, he may say unto thee, Friend, go up higher: then shalt thou have worship in the presence of them that *sit at meat* with thee."

a *replacement* word and studying the impact on the corresponding contextualised representations.[4]

**Our contributions:**

- Using a replacement schema, we analyse how the "tug of war" between the context and BERT's pre-trained knowledge affects the contextualisation. We find that the degree of BERT's contextualisation is not fixed but depends on the linguistic relations between the words, thereby reflecting the model's sensitivity to these linguistic features. We establish that contextualisation changes as a function of semantic relatedness which we systematically test across different linguistic relations, namely synonymy, antonymy, and hypernymy.
- We investigate BERT's contextualisation across different PoS classes and find that they are affected differently. We validate these findings through the Word-in-Context (WiC) task, revealing discrepancies between verbs and nouns. This implies that assessments for tasks like WiC should be tailored to different PoS to provide a more comprehensive evaluation, thereby enhancing the robustness of BERT models.
- We challenge the use of BERT embeddings to capture semantic changes involving meanings beyond the pre-trained knowledge of BERT. We demonstrate that approaches relying on the clustering of BERT embeddings fall short in capturing semantic changes, as they struggle to correctly contextualise untuned word meanings.
- We propose a new interpretable approach to Lexical Semantic Change (LSC) by leveraging contextualised BERT embeddings. Our approach leverage lexical replacements and outperforms existing state-of-the-art (SOTA) achieving the top score for English.

### 7.2.1 Related Work

BERT-like models leverage the Transformer encoder to capture the semantics of words [59, 228]. Their success in solving NLP tasks has prompted numerous studies to explore the nature and characteristics of their contextualised architectures. Ethayarajh [64], Coenen et al. [50], Cai et al. [37], and Jawahar, Sagot, and Seddah [98] shed light on the geometry of the embedding space. Serrano and Smith [208], Bai et al. [10], and Guan et al. [81] investigate the

[4] To reduce the amount of free variables, like spelling variations and OCR errors, in this study, we focus on modern texts and will verify our findings on historical texts in future work.

interpretability of the attention mechanism. Yenicelik, Schmidt, and Kilcher [237], Garí Soler and Apidianaki [75], Kalinowski and An [103], and Haber and Poesio [83] examine the clusterability of word representations. Abdou et al. [1], Hessel and Schofield [90], Mickus et al. [156], and Wang et al. [232] analyse the impact of word position in the embeddings generation. Coenen et al. [50], Levine et al. [129], and Pedinotti and Lenci [168] study how word meaning are represented in the embedding space.

Most of the current work involves probing tasks, as proposed by Hewitt and Liang [91]. These tasks consist of training an auxiliary classifier on top of a model, where the contextualised embeddings serve as features to predict syntactic (e.g. PoS) and semantic (e.g. word relations) properties of words [49, 135, 231, 135, 182]. If the auxiliary classifier accurately predicts a linguistic property, the property is assumed to be encoded in the model.

Our work extends previous work that focus on the contextualisation of BERT-like models. Instead of using probing tasks, we leverage a replacement schema according to PoS and semantic categories with graded lexical relatedness. Like Ethayarajh [64] and Jawahar, Sagot, and Seddah [98], we analyse the contextualisation across all layers of BERT. Building on the work of Zhao et al. [241], our research focuses on the degree of contextualisation. However, while they assess the inference of semantic word classes from contextualised embeddings, we analyze the contextualisation of different PoS that models exert under LSC and for OOD words.

### 7.2.2 Methodology

In our experiments, we leverage a replacement schema to investigate the pre-trained contextualisation of BERT. This involves analyzing the variations in embedding representations when a target replacement is introduced.[5] For instance, by replacing a target like *cat* with a replacement like *chair* in a specific context like *The <target/replacement> was purring loudly*.

#### The replacement schema

We use WordNet to generate different classes of replacements for a specific word [66], which correspond to a varying degree of plausibility (i.e. suitability of a specific replacement) between the target word and its replacement.

[5] Given our primary focus on words and their replacements when these words are split into multiple sub-words by the model, we calculate the average embeddings of the corresponding sub-words. This approach ensures the preservation of the same number of tokens in the original and artificial sentences and enables accurate distance calculations.

Thus, we hypothesise that each class is associated with a different impacts on contextualisation. Each class of replacements also has diachronic relevance, as the synchronic, semantic relation can be considered to have a parallel in semantic change [233]. To ensure accurate linguistic replacements, we maintain PoS agreement with the target words; that is, *nouns* are replaced with *nouns*, and so forth.

- `synonyms` (e.g. *sadness* ← *unhappiness*) are used to evaluate the stability in contextualisation; that is, we hypothesise similar embeddings between target and replacement words. Indeed, synonyms are considered equally likely alternatives in BERT's pre-trained knowledge. On the diachronic level, they emulate the absence of any semantic change of the replacement word;
- `antonyms` (e.g. *hot* ← *cold*) are used to evaluate a light change in contextualisation; that is, we hypothesise slightly less similar embeddings between target and replacement words. Indeed, antonyms are sometimes equally plausible alternatives[6] while other times are likely to surprise the model[7]. On the diachronic level, they emulate the a contronym change[8] of the replacement word;
- `hypernyms` (e.g. *bird* ← *animal*) are used similarly to `antonyms`. However, on the diachronic level, they emulate a broadening semantic change of the replacement word;
- `random` words (e.g. *sadness* ← *eld*) are used to evaluate a change in contextualisation. If BERT places high importance on the context, then the replacement should receive a similar representation to the target word. Otherwise, if BERT heavily relies on its pre-trained knowledge, the replacement will exhibit dissimilarity to the target word despite the identical context, as well as dissimilarity to the typical replacement representations. On the diachronic level, `random` emulates the presence of strong semantic change of the replacement word, that is the emergence of a homonymic sense.
- `synthetic` words (e.g. *love* ← *new-token*) are used as a baseline to evaluate the contextualisation of word meanings regardless of pre-trained knowledge. Indeed, we add a new token in the BERT's pre-trained vocabulary, and as such, it does not have any associated pre-trained

[6] For example: I *love/hate* you

[7] For example: I burned my tongue because the coffee was too *hot/cold*

[8] A contronym change occurs when a word's new meaning is the opposite of its original meaning (e.g. *sanction* in English)

knowledge (i.e., untuned weights). We hypothesise very dissimilar embeddings between target and synthetic words. On the diachronic level, this class mimics the adaptation to newly emerging concepts beyond their specific training time frame or to new domains.

### Data

To avoid introducing noise into our experiments resulting from the conflation of senses, we replace words with contextually appropriate replacements based on the intended sense of the word within a specific sentence (e.g, *stone* and *music* for *sitting on a rock* and *listening to rock*, respectively). We therefore leverage the SemCor dataset [159], still the largest and most commonly used sense-annotated corpus for English. To select candidate replacements, we consider different PoS tags, namely *verbs*, *nouns*, *adjectives* and *adverbs*, and semantic classes, namely *synonyms*, *hypernyms* and *antonyms*. We randomly sample a set of synsets for each PoS tag occurring in SemCor, and for a specific synset, we extract a subset of sentences where a word is annotated with that synset. We sample a maximum of 10 sentences per synset to prevent oversampling of high-frequency synsets. For each sentence, we generate the *synonym* and *antonym* replacements for all PoS, and *hypernym* replacements only for nouns and verbs[9] (see Table 7.3).

| **PoS** | **N. target words** | **Avg. sampled senteces per target word** | **N. examples** |
|---|---:|---:|---:|
| noun | 360 | 3.55 | 1277 |
| verb | 433 | 3.45 | 1494 |
| adjective | 393 | 3.39 | 1334 |
| adverb | 158 | 3.46 | 546 |

TABLE 7.3: Data statistics over PoS, sampled from SemCor.

### Tug of War in BERT

We delve into the intricate dynamics of the tug of war that occurs within the contextualisation of words by focusing on the word contextualisation, and the use of replacements as a proxy for semantic change. In our experiments, we utilise word embeddings generated by the monolingual BERT

[9] WordNet lacks hypernym information for other PoS

base model[10] due to its widespread usage. Additional analyses for XLM-R[11] and mBERT[12] are presented in appendix.

### 7.2.3 Word contextualisation

We analyse the word contextualisation by comparing the embeddings of a word $w$ in the original sentence to those in the same sentence when $w$ is replaced by $r$. To perform this comparison, we rely on the cosine distance between the embeddings of $w$ and $r$. We refer to this distance as the *self-embedding distance* (SED), i.e.,

$$\begin{array}{ccccccc} w^{-n} & ... & w^{-1} & w & w^{+1} & ... & w^{+m} \\ & & & \updownarrow & & & \\ w^{-n} & ... & w^{-1} & r & w^{+1} & ... & w^{+m} \end{array}$$

where $w^{-n}, ..., w^{-1}$ and $w^{+1}, ..., w^{+m}$ denotes the embeddings of $n$ and $m$ neighbouring words to the left and right of the target word $w$, respectively.[13] The higher the SED, the less BERT leverages context to contextualise words but instead relies on pre-trained knowledge.

#### Self-embedding distance

For each pair of original and artificial sentences, we computed SED across each layer. We then analysed the average SED for each class of replacement and PoS at each BERT layer. To address the anisotropic nature of BERT's space[14] and ensure comparability across layers, we normalise the average SED score for each class of replacements with the average SED score obtained with `synthetic` replacements. We report average SED scores in Figure 7.1. Like Ethayarajh [64], we observe that the word contextualisation increases across layers as the SED decreases. However, we find a similar degree of contextualisation in the last layers of BERT, indeed the SED becomes somewhat stable for each class of replacements and PoS. Adverbs represents an exceptional case, as the word contextualisation is less stable than other class of replacements.

[10] *bert-base-uncased*

[11] *xlm-roberta-base*

[12] *bert-base-multilingual-cased*

[13] Given our primary focus on words and their replacements when these words are split into multiple sub-words by the model, we calculate the average embeddings of the corresponding sub-words. This approach ensures the preservation of the same number of tokens in the original and synthetic sentences and enables accurate distance calculations.

[14] Embeddings occupy a narrow cone within the vector space [64].

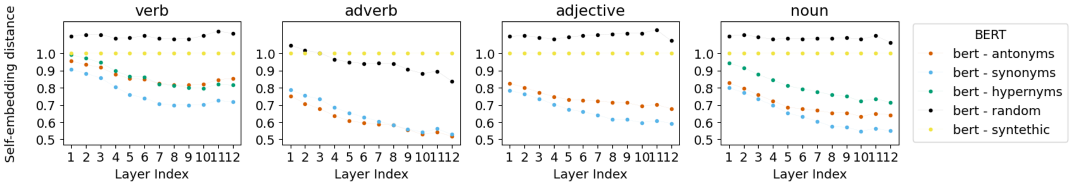


FIGURE 7.1: Average SED over layers.

For `adverb`, the `synonyms` and `antonyms` are closer than for other PoS, while being, along with `random`, associated with lower SED; that is, adverbs are more contextualised than other PoS and less pre-trained knowledge is used for their representation. This is most likely because adverbs in English are more context-dependent than the other PoS. Indeed, they can modify verbs, adjectives, adverbs, and entire sentences, in contrast to adjectives, which may only modify nouns and pronouns, and to verbs and nouns, which constitute essential components of sentences. This finding is in line with the work of Lorge and Pierrehumbert [139], which shows *weak differentiation amongst the semantic classes of adverbs*.

For other PoS, the SED score for `random` is above 1.0 in all layers, leading us to argue that BERT falls short in contextualising out-of-context words when pre-trained knowledge is available. That is, the representation of a random word does not mimic the representation of the target word that it replaces. The context thus has minimal effect in determining the representation of the replacement word.

For `verb`, we note a higher SED for `antonyms`, `synonyms`, and `hypernyms` in comparison to other PoS. Additionally, there is a narrower gap between the SED for `random` and the SED for `antonyms, synonyms, hypernyms`. These observations suggest that the contextualisation of verbs is less pronounced than that of other PoS and that the model relies more on pre-trained knowledge. As a result, the embeddings of verbs exhibit greater similarity to the embeddings of random words in context.

As for `adjective` and `noun`, we note that they exhibit similar contextualisation across layers. Additionally, for `noun`, `hypernyms` are less similar to the target than `antonyms` and `synonyms`. This aligns with the recent findings of Hanna and Mareček [87], suggesting that BERT's understanding of noun hypernyms is limited.

All in all, our results suggest that models exhibit varying degrees of contextualisation for different PoS, with lower contextualisation observed for verbs.

### PoS contextualisation through WiC

To verify that the variations in contextualisation across PoS are not an artifact of the replacement schema, we utilise Word-in-Context (WiC), a popular task in NLP for which various benchmarks are available in different languages. WiC is a recent task designed to assess the effectiveness of contextualised embeddings in capturing word usages by determining whether a target word $w$, occurring in a pair of sentences $\langle s1, s2 \rangle$, has the same meaning or not [172]. PoS information is available for most WiC benchmarks but, thus far, existing approaches to WiC have treated all PoS as a bundle. For our experiments, we consider benchmarks for different languages: English (WiC-en [172], MCL-WiC- [144]), French (MCL-WiC-fr [144]), Italian (XL-WiC-it [181], WiC-ITA-it [44]), and German (DWUG-de [205])[15]. We split the WiC benchmarks into distinct sub-corpora for nouns, verbs, and adjectives[16]. Details on the benchmarks are available in Appendix B.3.

Following Pilehvar and Camacho-Collados [172], we tune a threshold-based classifier on the development set (Dev) of each WiC benchmark, PoS, and BERT layer, totaling 216 classifiers[17]. For each sentence pair available in a Dev set, we computed the cosine similarity between the contextualised embeddings of the target word $w$ extracted individually from a layer of a BERT model[18]. The threshold for a classifier is determined by optimising the F1-score over Dev and then applied to both the Train and Test sets. Instead of using accuracy as an evaluation metric, we employ the F1-score to account for the inherent class imbalance in these split benchmarks.

For the sake of brevity, we report in Table 7.4 results for classifiers based on last layer embeddings. Additional and consistent results for each layers are reported in Appendix B.3. According with the finding of Figure 7.1, we find that BERT consistently performs better at distinguishing usages of nouns than verbs, with adjectives falling in between.

As *verbs are approximately twice as polysemous as nouns* [134], we want to exclude that WiC performance over PoS is not influenced varying levels of polysemy instead of the models' contextualisation capability. We analysed the polysemy distributions of target words for each PoS and benchmark, finding

[15] Although the benchmark for French and German provided by Raganato et al. [181] is considerably larger, we have made the decision not to use them due to the presence of noisy instances that could potentially impact our analysis negatively.

[16] We excluded adverbs due to the limited number of examples.

[17] 6 benchmarks, 3 PoS, 12 layers

[18] We use *bert-base-uncased* for English, *dbmdz/bert-base-french-europeana-cased* for French, *dbmdz/bert-base-french-italian-uncased* for Italian, *bert-german-cased* for German. The models are base versions of BERT with 12 attention layers and a hidden layer of size 768.

| | **WiC-en** | | **MCL-WiC-en** | | **XL-WiC-fr** | | **XL-WiC-it** | | **WiC-ITA-it** | | **DWUG-de** | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *Train* | *Test* | *Train* | *Test* | *Train* | *Test* | *Train* | *Test* | *Train* | *Test* | *Train* | *Test* |
| *nouns* | **0.691** | **0.683** | **0.796** | **0.847** | **-** | **0.677** | **0.678** | **0.689** | 0.682 | **0.759** | **0.808** | **0.799** |
| *verbs* | 0.615 | 0.611 | 0.752 | 0.805 | **-** | 0.622 | 0.639 | 0.593 | 0.459 | 0.615 | 0.630 | 0.640 |
| *adjectives* | **-** | **-** | 0.738 | 0.819 | - | 0.653 | **-** | **-** | **0.733** | 0.695 | **-** | **-** |
| *Acc.* | 0.638 | | 0.840 | | 0.787 | | 0.726 | | Macro-F1 0.670 | | - | |
| *Reference* | Pilehvar and Camacho-Collados | | Martelli et al. | | Martelli et al. | | Raganato et al. | | Periti and Dubossarsky | | | |

TABLE 7.4: Scores obtained by threshold-based classifiers for WiC trained on Dev sets. Results are provided for PoS and WiC benchmarks, with the best result in each considered benchmark and dataset highlighted in **bold**. The bottom rows show the SOTA accuracy and corresponding reference reported for the benchmark.

that a comparable degree of polysemy for each PoS in the considered WiC benchmarks. This assessment involved checking the polysemy of each target word in the various WiC benchmarks using Open Multilingual Wordnet, for English, French, Italian [66], and Odenet for German [210]. See Appendix B.4 for details. These results lead us to conclude that BERT exhibits varying degree of contextualisation for different PoS.

### Form-based and sense-based approaches

Approaches to LSC relying on contextualised word embeddings are typically distinguished into two main categories: *form-based* approaches and *sense-based* approaches [161]. The former captures semantic change by solely relying on similarities among raw embeddings without depending on sense disambiguation and representation. A common strategy involves aggregating all the embeddings of a word in $C_1$ and $C_2$ by averaging, and modeling the change as the cosine similarity of the average representations (PRT) [145] . The latter generally use clustering algorithm like Affinity Propagation to identify senses and subsequently model the change as divergence of cluster distributions (JSD) [146].

We argue that our findings in Section 7.2.3 has important implications on these approaches when pre-trained models are used. Specifically, consider semantic changes involving the acquisition or loss of new/old meanings. When these semantic changes occurs, if BERT models lack pre-trained knowledge of the involved meanings, they tend to position the corresponding word embeddings instances far apart in the space from the other embeddings of the same word, regardless of the context in which the word occurs (see `random` in Fig 7.1). While *form-based* approaches may still detect semantic changes by identifying low-contextualised word occurrences, *sense-based* approaches fall short in accurately detecting the same semantic changes. This is because they require modelling meanings outside the model's pre-trained knowledge before detecting those changes. Since these meanings cannot be adequately modeled, given the model's low degree of contextualisation, the performance of *sense-based* approaches is reduced compared to that of *form-based* approaches.

We further tested these implications in the LSC task by comparing PRT and JSD on an artificial diachronic corpus spanning two time periods (see details in Appendix B.5). Essentially, we introduced random replacements in $C_2$ with varying probabilities to emulate different degrees of change for a set of 46 target words. Subsequently, we compared the Spearman Correlation between the scores obtained with PRT and JSD with the artificially graded score of emulated semantic change. Results using BERT are presented in Figure 7.2 (see Appendix B.5 for additional results). Our hypothesis is that while PRT can accurately predict changes, JSD falls short because it can only correctly model the meanings that BERT is already aware of.

As shown in the figure, we can accurately model artificial semantic changes, even from layer 2, using PRT. This is not the case for JSD, where we observe statistically significant correlations for only a few layers. However, the significance of performance for JSD is an artifact of BERT embeddings and does not authentically represent the simulated change. We verify this by examining the modelled clusters. While, in general, the number of clusters of AP is large [146, 170], representing *sense nodules*[19] rather than word meanings [116], we find that the injected confusion in the model due to the `random` replacements results in a very low number of clusters (typically 2, maximum of 4). We report similar results in Appendix for other languages (i.e. German, Swedish, Spanish)

### Addressing LSC through replacements

We propose a novel supervised[20] approach to Graded Change Detection building upon the replacement schema. Our approach leverages a curated set of word replacements from WordNet and Wiktionary.

We denote $T = \{w_1, w_2, ..., w_N\}$ as the set of target words. For each target word, we extract a set of possible replacements $\rho(w_i) = \{r_1, r_2, ..., r_M\}$, resulting in $N * M$ replacement pairs. The set of replacements is obtained by considering the lemmas of synonyms and hypernyms associated with the target word $w_i$ in WordNet and words extracted from the Wiktionary page

[19] Lumps of meaning with greater stability under contextual changes [52]

[20] According to the classification framework presented by Montanelli and Periti [161]

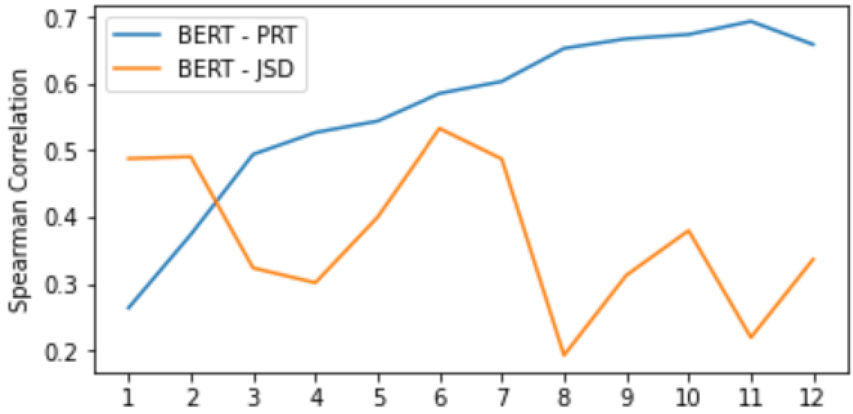


FIGURE 7.2: Spearman Correlation over layers for artificial semantic change.

corresponding to the target word. For each target word $w_i$, we sample up to 200 sentences from each period that remain stable regardless of the replacement word $r_j$. For each replacement pair $(w_i, r_j)$, we denote the set of sentences for a time period $t \in \{1, 2\}$ as $S^t(w_i, r_j)$.
For each sentence $s \in S^t(w_i, r_j)$ we measure the self-embedding distance of the target and replacement word and denote it $sed(s)$. The average self-embedding distance of a target-replacement pair is defined as

$$awd^t(w_i, r_j) = \frac{1}{|S^t(w_i, r_j)|} \sum_{s \in S^t(w_i, r_j)} sed(s)$$

The absolute difference in $awd$ over time is denoted TD$(w_i, r_j)$. Finally, we rank the replacements $\rho(w_i)$ according to their degree of time difference:

$$R(\rho(w_i)) = \{r_1, r_2, ..., r_M |\, \text{TD}(w_i, r_{i+1},) \leq \text{TD}(w_i, r_i)\}$$

and we compute a semantic change score $lsc_w$ as the average TD considering the top $k$ replacements:

$$lsc_w = \frac{1}{k} \sum_{r \in R(\rho(w_i))_k} \text{TD}(w_i, r)$$

We evaluate our approach on the SemEval-2020 Task 1, Subtask 2 dataset for English. We compute the Spearman Correlation between the graded score reported in the gold truth and the $lsc$ scores. Figure 7.3 reports the correlation computed for different values of $k$. The highest correlation of 0.741 is

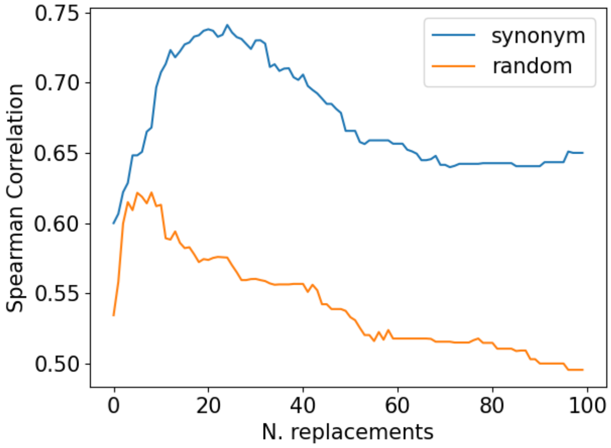


FIGURE 7.3: Top-k replacement vs Spearman Correlation.

| | Model | Spearman Correlation |
|---|---|---|
| | Rosin and Radinsky | 0.629 |
| | Kutuzov and Giulianelli | 0.605 |
| | Laicher et al. | 0.571 |
| | Periti et al. | 0.512 |
| **Synonym Replacement** | Replacement Min. Corr. | 0.600 |
| | Replacement Max. Corr. | 0.741 |
| | Replacement Avg. Corr. | 0.674 |
| **Random Replacement** | Replacement Min. Corr. | 0.495 |
| | Replacement Max. Corr. | 0.622 |
| | Replacement Avg. Corr. | 0.542 |

TABLE 7.5: Spearman Correlation on SemEval-2020 Task 1 (Eng), where our results outperform current SoTA.

achieved when considering the first 22 replacements, while the lowest correlation of 0.600 is obtained using only the first replacement (see Table 7.5). Interestingly, the minimum correlation obtained using the replacements is competitive with SOTA results. Moreover, on average, the correlation is higher than the SOTA model's performance. The replacements are reported in Table B.5. We used the linguistically-aware replacement strategy in the LSC task to assess and quantify the semantic changes undergone by words over time. By replacing the target words with different semantically related words, we generate contextual variations that enable the detection of semantic shifts. In the case of words like *record* and *land* that have undergone semantic change through narrowing and generalisation, respectively, linguistically aware replacement scan provide valuable insights. The replacement process generates a list of replacements that can be used as labels for the types of semantic change observed. By associating each replacement with a specific semantic category or change type, it becomes possible to analyse and quantify the semantic shifts experienced by words over time.

### Random replacements

In this section, we present results using randomly selected words with the same Part of Speech (PoS) as the target word, i.e. `random` replacement as introduced in Section 7.2.2. This approach generates a list of substitute words contextually unrelated to the target word. Some interesting patterns emerge when these results are compared with those obtained using synonym replacement. In the case of semantic change detection, the use of synonyms can provide more contextually relevant replacements, as they share semantic relationships with the target word. However, using random word replacements can still yield reasonable results, as evidenced by an average correlation of 0.542. These results is in line with the finding of Section 7.2.3.

In these approach, although random replacements tend to perform worse than synonym replacements, they have one distinct advantage: they do not rely on external lexical resources. This way, the approach is also suitable for unsupervised scenarios. While synonym replacements can improve contextualisation and semantic relevance, they are not always readily available or reliable for languages with limited linguistic resources. In such cases, random word replacements can still provide reasonable results and serve as a practical and resource-efficient approach for tasks where synonym information is scarce or unavailable.

In Section 7.2.3, by using SemCor, we effectively account for the nuances of different word senses, thereby improving the contextualisation and semantic relevance of synonym replacements. This approach is more targeted as synonyms are selected based on their association with a particular sense, leading to higher quality contextualisation in the context of that sense. As a result, synonym replacements are more finely tuned to the specific meaning of the target word, reducing noise and improving correlation with semantic change labels.

The lower correlations observed with random replacements indicate that contextualisation effects vary significantly when unrelated words are introduced into the context. This emphasizes the crucial role of context in the performance of language models when they have prior knowledge of input word meanings. However, it should not be assumed that contextualisation is equally effective when modeling new meanings outside the scope of the model's pre-trained knowledge.

### 7.2.4 Conclusions

Ethayarajh's work (2019), which was the first to study contextualisation in a methodical way, has clearly set the stage for our fine-grained study of contextualisation. We extend this line of work by studying the "tug of war" between BERT's contextualisation and its pre-trained knowledge respect to the area of LSC and OOD.

We analyse the degree of contextualisation using a replacement schema that can be used for any pre-trained LLMs. We focus on BERT and find that contextualisation effects differ across PoS. We support this finding by splitting up WiC datasets and find that the results are consistently better for nouns than verbs, across several languages, an effect that cannot be attributed to polysemy effects. We also conclude that the pre-training models significantly influence the representation of a word, and BERT models are not capable of capturing contextualised word meanings beyond their pre-training. Thus, our results indicate that the low degree of contextualisation can have severe limiting effects when pre-trained BERT models are applied to out-of-domain text. We will further verify these findings using, for example, historical, sense-tagged texts.

Using the same replacement schema, we demonstrate that *form-based* approaches are more suitable than *sense-based* approaches for detecting semantic changes when pre-trained models are employed. Additionally, we propose a novel approach to LSC and are able to surpass the results achieved by existing state-of-the-art models in the task of LSC. Our replacement schema gives us an automatic way of providing labels for the change that has occurred, offering us a way to do explainable semantic change detection. In future work, we will use models that utilise the complete encoder-decoder architecture, such as T5 [180], or exclusively the decoder architecture such as GPT-like models [34], to generate replacement categories without relying on existing resources like WordNet. We will also evaluate the degree of contextualisation on these models.

#### Limitations

One potential limitation of our study lies in the use of the replacement schema in conjunction with lexical replacements generated from WordNet. As a matter of fact, inherent limitations of WordNet, such as potential gaps, inaccuracies, or ambiguities in the semantic relationships may influence our analysis. WordNet also limits the data sources from which we can draw sentences, since we need a corpus with sense annotations corresponding to a lexicon.

Furthermore, in our experiment, the lexical replacement process involves substituting a *word* occurrence in the original sentence with a related *lemma* extracted from WordNet. As a result, providing the model with synthetic sentences containing the lemma instead of the inflected word may influence the generation of word embeddings and the contextualisation of every word in the sentences. However, we assume that this limitation equally affects every class we consider. For example, while the lemma of a verb may reduce the third singular verb form, the plural forms of adjectives and nouns can also be simplified to singular lemma forms. Additionally, to mitigate these issues and ensure that all PoS are equally affected by the replacement procedure, we replaced both the target and replacement words with lemmas in the original and synthetic sentences, respectively.

Finding the correct form of a replacement requires advanced morphological analysis and carries the risk of leading to errors. For now, we therefore opted to circumvent this by replacing targets and lemmas alike. Furthermore, we would like to highlight a relevant study by [123] that delves into the influence of various linguistic variables on the use of BERT embeddings for the LSC task. This research demonstrates that by reducing the influence of orthography through lemma usage, significant enhancements in BERT's performance were observed for German and Swedish, while maintaining comparable results for English. This underscores the potential benefits of lemma-based contextualisation and that linguistic features like orthography can sometimes be minimised without substantial loss of performance.

# Chapter 8

# Computational Social Science and Cultural Analytics

## 8.1 The corpora they are a-changing: a case study in Italian newspapers

The use of automatic methods for the study of lexical semantic change (LSC) has led to the creation of evaluation benchmarks. Benchmark datasets, however, are intimately tied to the corpus used for their creation questioning their reliability as well as the robustness of automatic methods. This contribution investigates these aspects showing the impact of unforeseen social and cultural dimensions. We also identify a set of additional issues (OCR quality, named entities) that impact the performance of the automatic methods, especially when used to discover LSC.

### 8.1.1 Methodology

To test benchmark independence and models' robustness for LSC, we design a set of experiments using two source corpora, a common benchmark, and a common architecture for LSC detection.

The first corpus is the "L'Unità" corpus [14]. It covers a time span between 1945–2014 and it has been collected, pre-processed, and released for the DIACR-Ita (Diachronic Lexical Semantics in Italian) task [18], a LSC change shared task for Italian. Texts were extracted from PDF files by using the Apache Tika library[1] and pre-processed with spaCy[2] for tokenization, PoS-tagging, lemmatization, named entity recognition and dependency parsing. The second corpus was obtained by crawling a publicly available digital archive of the Italian newspaper "La Stampa". The corpus covers a shorter

[1] `https://tika.apache.org/`

[2] `https://spacy.io/`

time period (1945–2005) and it was pre-processed using the same tools and pipeline of "L'Unità". Each corpus is split into two sub-corpora, $C_1$ and $C_2$, covering different time periods. Table 8.1 summarises the basic statistics of corpora and the time periods of each sub-corpus.

| Corpus | Subcorpus | Tokens |
|---|---|---|
| L'Unità | $C_1$ [1945 – 1970] | 52,287,734 |
| L'Unità | $C_2$ [1990 – 2014] | 196,539,403 |
| La Stampa | $C_1$ [1945 – 1970] | 670,281,513 |
| La Stampa | $C_2$ [1990 – 2005] | 1,193,959,080 |

TABLE 8.1: Corpora statistics.

The corpora present two major differences. First, as shown in Table 8.1, the number of tokens in "La Stampa" is consistently larger than "L'Unità". Second, the political and social orientations of the two newspapers are different. Historically, "L'Unità" has been the official newspaper of the Italian Communist Party and of its successors PDS/DS. "La Stampa" is the oldest newspaper in Italy, traditionally it has voiced centrist and liberal positions.

The only benchmark for Italian has been proposed in the context of DIACR-Ita. The dataset contains 18 target lemmas, 6 of which are instances of a LSC. The dataset was manually created using the "L'Unità" corpus, where a valid LSC corresponds to the acquisition of a new meaning by a target word in $C_2$. As architecture for automatic LSC detection, we obtain comparable diachronic representations of word meanings by re-implementing the Word2Vec Skipgram model [157] with Orthogonal Procrustes (OP-SGNS) [86]. In particular, we adopted the implementation proposed by Kaiser, Schlechtweg, and Walde [101], a state-of-the-art system that ranked 1$^{st}$ both at DIACR-Ita and at SemEval 2020 Task 1: Unsupervised Lexical Semantic Change Detection [205]. Model parameters are reported in Appendix C.1. Word embeddings were generated using lemmas to reduce sparseness and facilitate the evaluation against the benchmark.

### 8.1.2 Testing for Robustness and Independence

Testing for robustness and consistency for LSC is not trivial since it requires to distinguish between two strictly connected dimensions: (i) reliability of the benchmark (dataset dimension), and (ii) variations in data distributions (corpora dimension). The first dimension (dataset) is analysed by comparing on the DIACR-Ita benchmark the performances of the same model trained on the two corpora. The corpora dimension is investigated by manually inspecting the disagreements between the model predictions. All 18 target words in the benchmark satisfy a minimal frequency threshold of 10 both in $C_1$ and $C_2$ in "La Stampa", allowing us to reliably compare the results.

To study the reliability of the DIACR-Ita benchmark with respect to the underlying corpus, for each target word in the benchmark, we computed the cosine similarity of its embedding representation from each sub-corpus ($C_1$ and $C_2$). To account for the random initialisation of the OP-SGNS parameters, we ran 10 experiments with different initialisations and averaged the results. The system accuracy is computed as the fraction of correctly predicted words over the total number of words in the benchmark. A target word is deemed as an instance of LSC when its cosine similarity across the two time periods is below a given threshold $\lambda^*$.

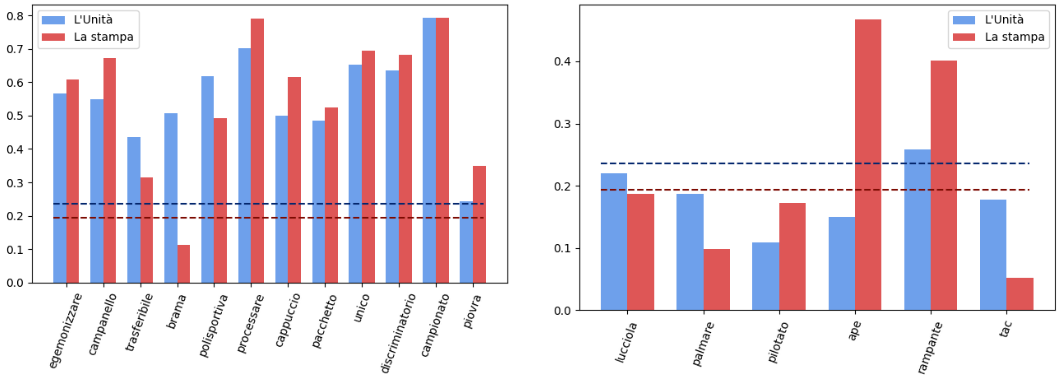
Since the focus is on the reliability of the benchmark across corpora, and not the system performances, the threshold $\lambda^*$ for each corpus is set up to the value that maximises the system performance on the corpus.

Using the optimal threshold, our implementation of OP-SGNS obtained an accuracy of $0.96 \pm .02$ when trained on "L'Unità" and $0.83 \pm .00$ when trained on "La Stampa", a difference spawned by the incorrect classification of the words *ape* (LSC), *rampante* (LSC), and *brama* (stable).

To understand the role of the two corpora, we compare the target word similarities between $C_1$ and $C_2$ on the two corpora. Figure 8.1a and Figure 8.1b illustrate the similarities of the stable and LSC target words, respectively. Overall, the identification of LSC target words seems consistent among the two corpora, and lets us assume that the benchmark is reliable and the algorithm is robust.

We further analyse the system's disagreements by manually exploring their occurrences in each corpus for every time period.[3] For the target *brama* ('yearning'), "La Stampa" indicates a potential LSC. The manual inspection, however, has confirmed the annotation in the benchmark (i.e., a stable meaning) showing that the change is triggered by the presence of this word in band names in the $C_2$ portion of the corpus. *Ape* ('bee') is listed in the benchmark as an LCS, since in $C_2$ it refers not only to the insect, but also to a three-wheeled vehicle. Despite this new sense is present in the $C_2$ sub-corpus of "La Stampa", the difference in similarity is above the threshold. Interestingly, in this corpus we observe the three-wheeled vehicle sense also in $C_1$, especially as part of paid advertisements. This points to a bias in the corpus (i.e,

[3] We use NoSketch Engine `https://nlp.fi.muni.cz/trac/noske`.



(A) Stable targets.

(B) LSC targets.

FIGURE 8.1: LSC change scores computed using cosine similarity on both "L'Unità"and "La Stampa"corpus. The dashed lines indicate the $\lambda^*$ thresholds, computed respectively on "L'Unità"and "La Stampa"corpus. Similarities below the thresholds trigger an LSC.

"L'Unità") used to create the benchmark, namely the lack of (or extremely limited) presence of advertisements, which has obfuscated the occurrence of the three-wheeled vehicle sense and suggested *ape* as a good candidate for an LSC. *Ape* is interesting also for another reason: the discrepancy between when it was first on the market (1948) and its first attestation in the Sabatini Coletti dictionary (1983). Further related to the more commercial nature of the "La Stampa"newspaper is the higher difference in similarity with respect to the "L'Unità" for the word *rampante* ('rampant'/'high-flying'). In "La Stampa", the word occurs also in $C_1$ as part of the book title "Il barone rampante"; this has mitigated the variation in context of usage with the occurrences of *rampante* in $C_2$.

### 8.1.3 Models into the Wild

We further extended the analysis to the whole common vocabulary of the two corpora to test the robustness of the computational model. In particular, we consider the vocabulary intersection $V$ of the two sub-corpora, that consists of 48,681 lemmas. Then, we compute the two sets $X$ and $Y$ of cosine similarities for all the words in $V$. A first analysis was conducted to understand to which extent the rank order of the two sets $X$ and $Y$ are correlated. The Spearman Correlation between the two sets is $0.67$ (p-value $< 0.01$), which indicates a positive correlation between the two rank orders, suggesting that the output of OP-SGNS is similar across the two corpora. The plots of the correlations are reported in Figure C.1 in Appendix C.2.

In this analysis, the optimal thresholds cannot be computed due to the lack of a gold-standard for the whole vocabulary intersection $V$. Potential LSC instances are identified by using as threshold the difference between the average of the cosine similarities ($\mu$) and the standard deviation ($\sigma$) over the set $V$:

$$LCS(X) = \{t_i \in V \mid x_i < \mu(X) - \sigma(X)\}$$

Where $t_i$ is the term associated with the i$^{th}$ similarity $x_i \in X$. Similarly, we compute the set $LCS(Y)$. The intersection of the two LCS sets consists of 2,283 lemmas. A quick inspection of the proposed LSCs immediately triggers observations concerning two aspects: (i) the well formedness of a lemma; and (ii) the presence of named entities (NEs). By well formedness, we refer to the lemma being an actual word attested in a reference dictionary of Italian. Indeed, some of the lemmas with the lowest similarity scores, e.g., *gaucha*, *bwa*, *bill*, *-anche*, do not appear to be well formed Italian words. Reasons for this are to be found in the quality of the digitized versions of the documents of the two corpora, the presence of foreign words (e.g., *frere*, French for 'brother'), as well as tokenization errors of the pre-processing tool. We use the list of lemmas in the Sabatini Coletti dictionary to filter out all of these entries.

NEs appear to be an additional source of noise. Lemmas like *albertarelli*, *beraudo*, *napoleoni*, *armellini*, are all instances of NEs referring to people's surnames. We automatically filter NEs in two steps: (i) for each word in a sequence tagged as NE by spaCy, we retrieve and store separately the corresponding lemma; (ii) every candidate LSC lemma is matched against the list generated in (i), greedily filtering all lemmas found to be part of a NE.

After the filtering, only 232 lemmas remain. We sample 50 lemmas (approx. 20%), for a manual inspection. For each lemma, we collected its definitions and the associated year(s) of first attestation from the Sabatini Coletti. Then we manually explored the context of occurrence of each lemma in each time period for each corpus. The manual validation followed a similar approach to the creation of DIACR-Ita gold standard: a lemma is considered to be undergone an LSC only if the definition(s) of the sense are attested in $C_2$ and not in $C_1$.

By simply using the date of first attestation in the dictionary, 37 lemmas do not qualify as having undergone LSC between the two time periods. Of the remaining 13 lemmas: three have no date of first attestation; five lemmas have a date of first attestation after $1970$ (i.e. these lemmas were not used before); and five lemmas present new senses. However, when considering only those lemmas with a new sense attested after $1970$, this list reduces to two lemmas.

The manual exploration of the contexts of occurrence in both corpora of the 50 lemmas showed that only four of them (8% of the total sample) can be considered correct examples of an LSC. Two of them, *palmare* ('obvious'/'palmar'/'hand-held computer') and *patteggiare* ('negotiate'/'plea'), are also attested in the Sabatini Coletti. The remaining two, *handicappare* ('to handicap') and *orgasmo* ('orgasm'), indicate a change of use rather than an actual change of meaning. In particular, *handicappare*, and namely its participial form, was used during the 80s/90s to refer to people with disabilities, extending the initial meaning in $C_1$ of "to assign an handicap to a team". The use of the word with this meaning is now derogatory and it is not attested in the dictionary. On the other hand, *orgasmo* was used in $C_1$ in its figurative meaning of great or extreme anxiety, e.g. "nell'orgasmo del momento" ('in the excitement of the moment'). On the other hand, in $C_2$ is used with reference to sex and sexuality. Three additional lemmas are signalled as lexical changes: *pula*, *doc*, and *tac*. However, they are officially attested as different lemmas in the Sabatini Coletti, thus implying homonymy. All remaining entries are false positives being either NEs or OCR errors. For the NEs, these are cases where the NE also corresponds to a lemma in the reference dictionary. A good example of this is *borsellino*. In $C_1$, both corpora present context of use with the dictionary meaning of "a small purse". However, in $C_2$, the contexts of use refer to the judge Paolo Borsellino[4], killed in a terrorist attack by the Mafia.

NEs introduce additional challenges while constructing a benchmark for LSC, especially when they are homonyms with common nouns. A viable solution to this problem would be to detect and disregard from the corpus those entities that are homonyms of common nouns. This also calls for the development of more robust systems for NE detection: besides our efforts at filtering NEs, lots of them have remained as potential targets of LSC.

## 8.2 Analysis of lexical semantic changes in corpora with the Diachronic Engine

With the growing availability of digitized diachronic corpora, the need for tools capable of taking into account the diachronic component of corpora becomes ever more pressing. Recent works on diachronic embeddings show

[4] https://en.wikipedia.org/wiki/Paolo_Borsellino

that computational approaches to the diachronic analysis of language seem to be promising, but they are not user friendly for people without a technical background. *Diachronic Engine* is a system for the diachronic analysis of corpora lexical features. *Diachronic Engine* computes word frequency, concordances and collocations taking into account the temporal dimension. It is also able to compute temporal word embeddings and time-series that can be exploited for lexical semantic change detection.

### 8.2.1 Motivation and Background

In recent works about computational diachronic linguistics, techniques based on word embeddings produce promising results. In Semeval Task 1 [205], for instance, type embeddings rich high performances on both subtasks. However, these techniques are not included in any aforementioned linguistic tool. In order to bridge this gap, we try to build a tool that includes approaches for the analysis of diachronic embeddings. The result of our work is Diachronic Engine (DE), an engine for the management of diachronic corpora that provides tools for change detection of lexical semantics from a frequentist perspective. DE includes tools for extracting diachronic collocations, concordances in different time periods as well as for computing semantic change time-series by exploiting both word frequencies and word embeddings similarity over time.

The rest of this section is organized as follows: Section 8.2.2 describes the technical details of DE, while Section 8.2.3 shows some use cases of our engine that encompass that address time-series. We also present the results of a preliminary evaluation about the system's usability in Section 8.2.4.

### 8.2.2 Diachronic Engine

Diachronic Engine (DE) is a web application for lexical semantic change analysis in diachronic corpora. The DE pipeline needs diachronic corpora to compute statistics about the corpus. A diachronic corpus must include a temporal feature (e.g., year or timestamp of the publication date); DE exploits that feature to sort the documents.

We adopt the vertical format to represent word information, as specified for the IMS Corpus Workbench (CWB). In a vertical corpus, each word is in a new line. In each line, fields, called p-attributes, are separated by tabs. In

DE the default p-attributes are word, lemma, PoS tag and syntactic dependency. Non-recursive XML tags (s-attributes) on a separate line can be used for representing sentences, paragraphs and documents.

Corpora can be served in vertical format[5] or in plain-text mode; in the latter case, the plain-text is transformed in vertical format using the Spacy UDPipe[6] [213] tool, which splits plain-text into sentences and then predicts the PoS-tag, the lemma and the syntactic dependency for each token. UDPipe is a dependency parser that provides models for several languages. Models are built by using the Universal Dependencies [7] datasets as training data. Input files' names must contain the temporal tag of the period to which they refer. DE automatically detects temporal patterns in the name of the files. In particular, the last sequence of numbers in the file name is used to sort the documents.

Corpora are stored and managed by the CWB, a tool for the manipulation of large, linguistically annotated corpora. In particular, DE relies on the Corpus Query Processor (CQP) [48], a specialized search engine for linguistic research.

For building temporal word embeddings, DE exploits Temporal Random Indexing (TRI) [12, 17] that computes a word vector for each time period by summing shared random vectors over all the periods. TRI is able to produce aligned word embeddings in a single step and it is based on Random Indexing [196], where a word vector (word embedding) $sv_j^{T_k}$ for the word $w_j$ at time $T_k$ is the sum of random vectors $r_i$ assigned to the co-occurring words taking into account only documents $d_l \in T_k$. Co-occurring words are defined as the set of $m$ words that precede and follow the word $w_j$. Random vectors are vectors initialized randomly and shared across all time slices so that word spaces are comparable.

Future versions will include other approaches, such as Procustes [86], Dynamic Word Embeddings [236], Dynamic Bernoulli Embeddings [192] and Temporal Referencing [61].

The DE architecture is based on the client-server paradigm. The back-end of DE has been developed with Flask, a web framework written in Python. Concordances are retrieved by CQP, that indexes the corpus as soon as it is uploaded to the server, while collocations and frequencies are computed in Python. The back-end provides a set of services by a REST API where the input/output is based on JSON messages.

[5] https://www.sketchengine.eu/my_keywords/vertical/

[6] https://pypi.org/project/spacy-udpipe/

[7] http://universaldependencies.org

The back-end consists of three macro components: User Handler, Corpus Handler and Diachronic Operations. The User Handler manages registered users information such as username and passwords. Admitted operations on users are creation, read, update and delete. The Corpus Handler Component manages corpora information such as name, language, the list of fields in the vertical files, corpus visibility. Moreover, it deals with corpora types: each corpus has a label indicating if it is synchronic or diachronic. For diachronic corpora also the temporal range is stored. Operations admitted on corpora are: creation, update, delete, search and read. The Diachronic Operations component shows frequency lists, collocations of words, time-series, change-points and concordances. This component relies on CWB that indexes vertical files.

The Diachronic Operations component architecture is sketched in Figure 8.2. The front-end of DE has been developed with JHipster[8], using Spring[9] for server-side applications and Angular for client-side applications. The front-end communicates with the back-end by the means of the REST API.

The front-end design is inspired by the Google's Material Design and the Sketch Engine interface. The user interface provides multilingual support in Italian and English, but we plan to extend it to other languages.

This architecture allows the independence between the back-end and the front-end, in this way is possible to develop a different front-end or connect the front-end to a different implementation of the back-end. The only constraint is the REST API interface.

The homepage provides an easy access to all corpora owned by the logged user with links to available tools. The front-end provides also tools for creating and managing users and corpora. In particular, it is possible to define different grant permissions for each corpus.

The tool is distributed as open-source software under the GNU v3 license[10].

### DE tools

DE provides a set of tools for managing and querying diachronic corpora. The core of the back-end is based on the IMS Open Corpus Workbench (CWB) [11], which allows querying the indexed corpora by using the powerful CQP. Other tools have been integrated to facilitate the analysis of a diachronic corpus:

[8] `https://www.jhipster.tech/`

[9] `https://spring.io/`

[10] `https://github.com/swapUniba/Diachronic-Engine`

[11] `http://cwb.sourceforge.net/`

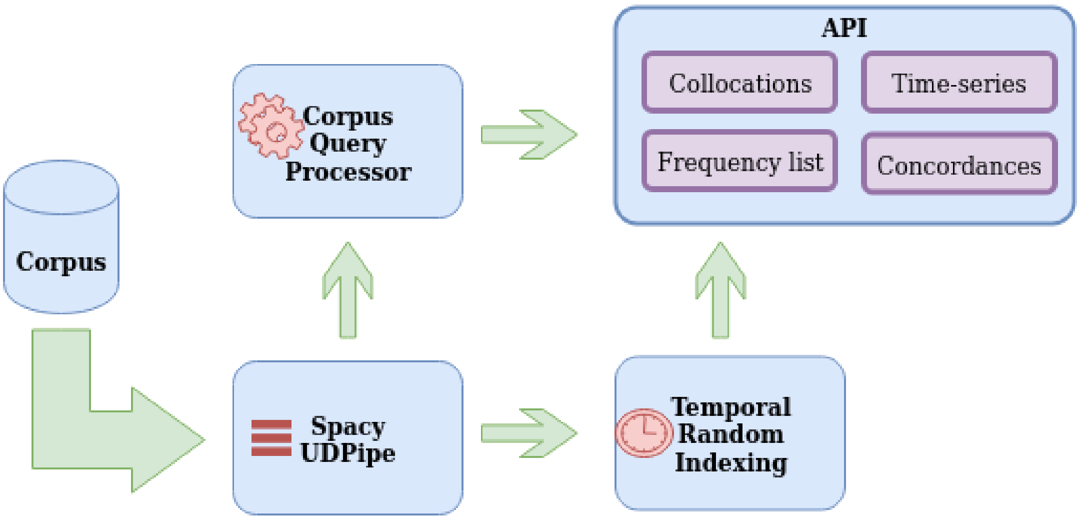


FIGURE 8.2: Diachronic Engine corpora manager.

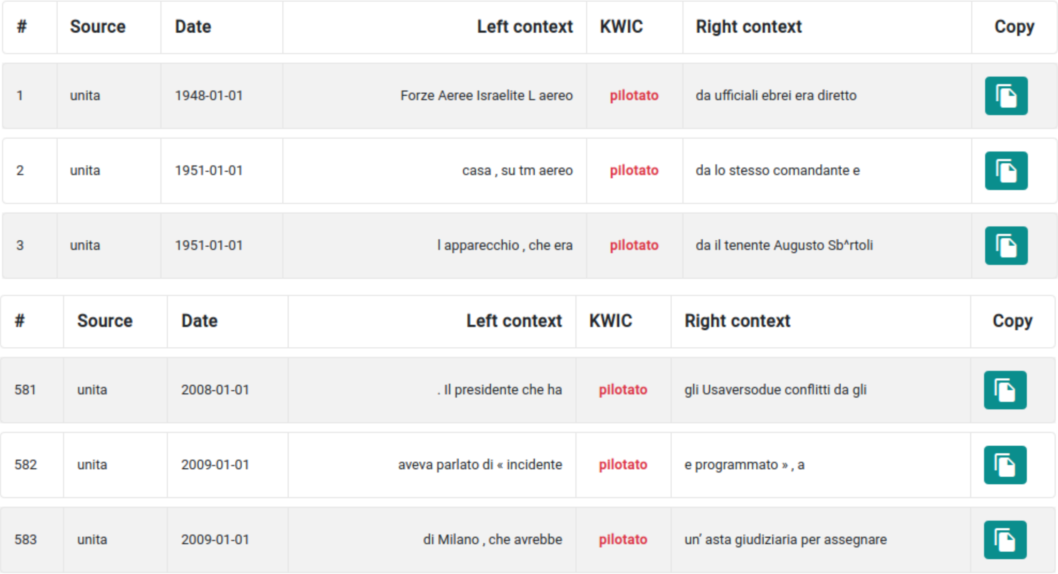
**Word frequency** Many works show a correlation between lexical semantic change and frequency differences between time periods. Google Ngram Viewer [155] uses n-grams frequencies over time to show the change in the semantics of n-grams. SketchEngine exposes the Trends tool, which uses a linear regression of frequencies to predict words that appear to be changed. In DE, queries can be filtered by part-of-speech, as well as by time periods. We use normalized frequencies, that can be filtered by time period.

**Collocations** Collocations have shown to be an effective tool in diachronic analysis [13]. A collocation is a sequence of words that occurs more often than would be expected. In order to compute the collocation strength we use the logDice [194]:

$$log\frac{2f_{xy}}{f_x + f_y}$$

logDice takes into account the frequency of the word $f_x$, of the collocate $f_y$ and the frequency of the whole collocation $f_{xy}$. Collocation results can be grouped by the PoS tag.

**Concordances** Concordances offer a way to find "the evidence" directly in the text by exploiting the context. The Concordances tool lists instances of a word with its immediate left and right context and the period the collocation belongs to. An example of concordances from "L'Unità" [14], is shown in Figure 8.3.



| # | Source | Date | Left context | KWIC | Right context | Copy |
|---|---|---|---|---|---|---|
| 1 | unita | 1948-01-01 | Forze Aeree Israelite L aereo | pilotato | da ufficiali ebrei era diretto | 📋 |
| 2 | unita | 1951-01-01 | casa , su tm aereo | pilotato | da lo stesso comandante e | 📋 |
| 3 | unita | 1951-01-01 | l apparecchio , che era | pilotato | da il tenente Augusto Sb^rtoli | 📋 |

| # | Source | Date | Left context | KWIC | Right context | Copy |
|---|---|---|---|---|---|---|
| 581 | unita | 2008-01-01 | . Il presidente che ha | pilotato | gli Usaversodue conflitti da gli | 📋 |
| 582 | unita | 2009-01-01 | aveva parlato di « incidente | pilotato | e programmato » , a | 📋 |
| 583 | unita | 2009-01-01 | di Milano , che avrebbe | pilotato | un' asta giudiziaria per assegnare | 📋 |

FIGURE 8.3: DE shows the KWIC (Keyword in the context) "pilotato", shifted from meaning *driven* to meaning *manipulated*.

**Time-series** A time-series $\Gamma(w)$ of a word $w$ is an ordered sequence of cosine similarities between the word vector at time $k$ ($v_w^k$) and the previous one at time $k-1$ ($v_w^{k-1}$):

$$\Gamma(w)_k = \frac{v_w^k \cdot v_w^{k-1}}{|v_w^k||v_w^{k-1}|}$$

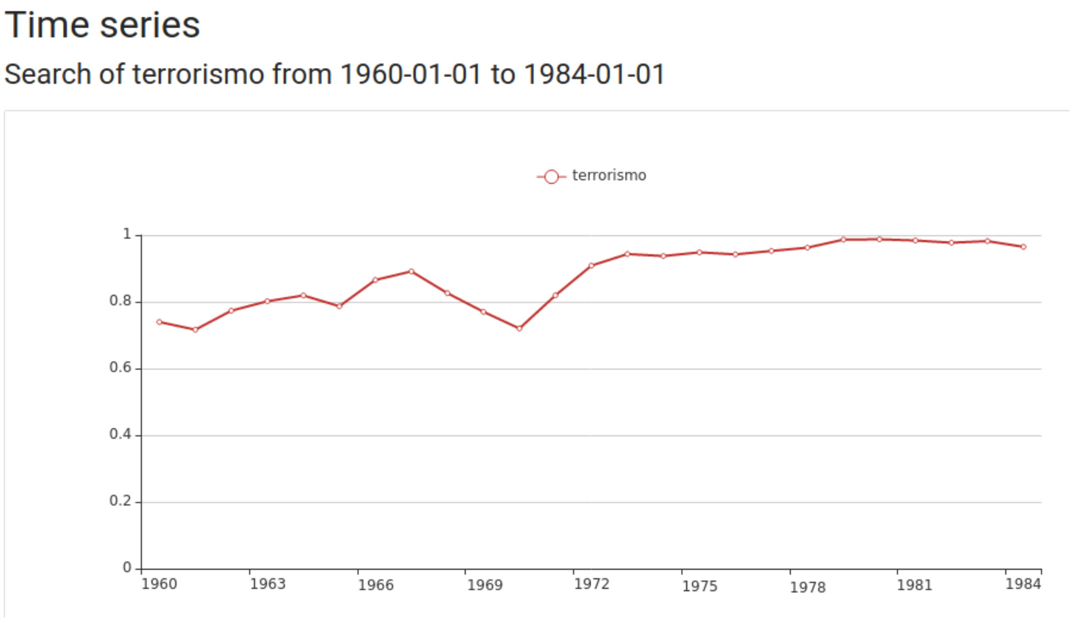
Diachronic Engine relies on word vectors computed by Temporal Random Indexing, but it is possible to integrate other approaches. In order to detect change points, we use the Mean Shift algorithm [220]. According to this model, we define a mean shift of a general time series $\Gamma$ pivoted at time period $j$ as:

$$K(\Gamma) = \frac{1}{l-j}\sum_{k=j+1}^{l}\Gamma_k - \frac{1}{j}\sum_{k=1}^{j}\Gamma_k \tag{8.1}$$

In order to understand if a mean shift is statistically significant at time $j$, a bootstrapping [62] approach under the null hypothesis that there is no change in the mean is adopted. In particular, statistical significance is computed by first constructing $B$ bootstrap samples by permuting $\Gamma(t_i)$. Second, for each bootstrap sample P, $K(P)$ is calculated to provide its corresponding bootstrap statistic and statistical significance (p-value) of observing the mean shift at time $j$ compared to the null distribution. Finally, we estimate the change point by considering the time point $j$ with the minimum p-value score. The output of this process is a ranking of words that potentially have changed meaning. Time-series is able to compare multiple words at the same time and allows to filter words by time period.

### 8.2.3 Use cases

In this section, we describe two use cases concerning both historical and computational linguistics. DE is an extension of existing tools for synchronic corpora. It shares many of the use cases already available on those tools, such as applications in lexicography, terminology and linguistics.



Time series

Search of terrorismo from 1960-01-01 to 1984-01-01

FIGURE 8.4: DE shows time-series of the word “terrorismo”.

#### Event detection through time-series

Lexical semantic changes can reveal aspects of real-world events, such us global armed conflicts [117]. DE provides several tools to help events detection through time-series:

- the comparison of two time-series for highlighting potential correlations between lexical-semantic changes
- the plot of the time-series of cosine similarity between two word vectors over time, showing how the relatedness between two words changes over time
- the detected change points can bring out hidden information

In Figure 8.4, the time-series of "terrorismo" (*terrorism*) is shown. The time-series appears to be influenced by real-world events happening in Italy. In particular, we can observe a decrease in similarity starting in 1968 and culminating in 1970 during a crucial moment in Italy: "Anni di piombo" (*Years of Lead*), years marked by terrorism and violent clashes carried out by political activists.

**Annotation of semantic shifts**

The manual annotation of lexical-semantic shifts can be very expensive. Although robust frameworks [200] for the annotations already exist and are successfully used in evaluation tasks [205], no tools for facilitating the annotation are available yet.
DE can provide useful tools for the annotation of semantic shifts:

1. Frequencies over time can be preliminary exploited to filter words that have good coverage in the years under analysis;

2. Change points in time-series offer an overall and intuitive idea of the potential semantic shifts;

3. Diachronic concordances and collocations can support the identification of the type of change [28], such as when a word gains or loses a meaning.

### 8.2.4 Evaluation

We place a particular focus on the usability of our tool by giving a satisfactory experience. To understand the strength and weakness of the user interface, we conduct a preliminary usability test, according to the eGLU protocol [211]. We use 21 participants. As a first step of the evaluation, we want to test the system's usability by measuring the task success rate: the ratio of users able to accomplish a set of predefined tasks. We ask participants to perform four tasks and we compute the average task success over all the 21 participants. During the evaluation, all participants complete their tasks without difficulties except for the showing frequency list task, where they had some problems with the corpus selection. We have already fixed this issue: the user is warned to choose a corpus from those available if no corpus is selected.
Results of the evaluation are reported in Table 8.2.

| Task | Avg. task success |
|---|---|
| User registration | 1 |
| Login and show user information | 1 |
| Add a corpus | 1 |
| Show frequency list | .8095 |
| Overall | **.9523** |

TABLE 8.2: Results of the usability evaluation.

Moreover, we designed and dispensed a questionnaire for measuring user satisfaction. The questionnaire is composed of ten questions about the usability and the design of DE with a Likert scale of five values. The questionnaire results return an average score of 84.05/100. The system appear likeable to use.

## 8.3 Emerging Trends in Gender-Specific Occupational Titles in Italian Newspapers

Throughout history, the prerogative use of specific gender forms over particular professions can fade away by introducing changes in the language lexicon (e.g., neologisms) or in the language usage (e.g., word frequencies). The way the lexicon is affected by those changes depends on the grammatical gender system, i.e. the set of rules that define the agreement between noun classes forms and the other parts-of-speech. Grammatical gender systems can vary dramatically from one language to another. Gygax et al. [82] propose a classification of languages based on their grammatical gender system. In this work, we focus on the Italian language, a grammatical gender language in which all nouns must be classified for gender. The Italian gender system admits three categories for nouns: gender-specific ending nouns, mobile gender nouns, and nouns where the gender is specified through determiners and adjectives [142]. In gender-specific ending nouns, the gender forms are expressed through completely different lexical roots (e.g., *genero/nuora*). In mobile gender nouns, the specific gender forms share the same lexical root, and the semantic gender is instead represented by different suffixes (e.g., *scrittore/scrittrice*). In other cases, the semantic gender of a noun is inferred only by the determiner and/or adjective (e.g., *il* giudice, *la* giudice). The peculiar characteristic found in the Italian language has strong repercussions in the way people refer to occupational titles, because a specific gender form might be preferred over the other due to historical reasons, regardless of the gender of the actual person being talked about [195]. This has become a hot-button issue in the last years, especially as a result of the United Nations Resolution "Transforming our world: the 2030 Agenda for Sustainable Development" with its global indicator framework for Sustainable Development Goals (SDGs), and specifically of SDG 5 *Achieve gender equality and empower all women and girls* (sub-goal 5.1 *End all forms of discrimination against all women and girls everywhere*) [128].

The objective of this work is to monitor how the use of gender-specific occupational titles has changed in the Italian language over the years through the use of diachronic analysis tools. We would like to emphasize that the goal is not to map the composition of men and women for each profession over time, as this cannot be reliably inferred from text. Instead, we are interested in gauging the cultural relevance of the gender-specific titles over time, as reflected in the news domain. Accordingly, the contributions in this work can be summarized as follows:

(i) We analyze emerging trends in the use of gender-specific occupational titles in the Italian language in a corpus of newspaper articles.[12]

(ii) We perform a deep-dive analysis of the figures that have guided a significant shift for two professions in particular.

### 8.3.1 Corpus

Occupational titles occurrences are extracted from a diachronic corpus that comprises two sub-corpora. The former corpus is the "L'Unità" corpus [14] that covers the time period 1945-2014. The latter is crawled by the publicly available digital archive of the Italian newspaper "La Stampa" covering the period 1945-2005 and processed using the same methodology mentioned in [14]. In order to align the two sub-corpora time ranges, we consider a subportion of the "L'Unità" corpus that spans the period 1948-2005. The overall corpus contains 3,529,820,155 tokens and spans the period 1948-2005. Corpus statistics are reported in Table 8.3. The corpus presents two main critical issues. First, despite having performed pre-processing and filtering, the documents from the earlier periods suffer from several OCR errors and noise.

[12] All data collected in this experiment is available here: `https://github.com/pierluigic/igsot`

Second, data is not equally distributed, the number of tokens drops dramatically in the first years. Text is processed using the UDPipe model [214] included in spaCy[13]. The UDPipe model is trained on the Italian Stanford Dependency Treebank [32]. Each sentence is tokenized, lemmatized and annotated with PoS-tags, named entity tags and dependency relations. Moreover, the UDPipe model provides information about inflectional features of nouns exploited in the occupational titles extraction pipeline.

| Corpus | Tokens | Period |
|---|---:|---|
| L'Unità | 425,833,098 | 1948-2014 |
| La Stampa | 3,145,959,127 | 1948-2005 |
| Overall | 3,529,820,155 | 1948-2005 |

TABLE 8.3: Corpus statistics.

### 8.3.2 Extracting Occupational Titles

The first step of our investigation consists of extracting a list of occupational titles from a common Knowledge Base. Specifically, we have exploited Wikidata [229], since it has collected a wide range of entities related to professional activities. We first extracted a list of all entities that are an instance of *profession* (wd:Q28640), or of an entity that is a subclass of it, for which a label in the Italian language is present. This label commonly contains the masculine gender form of the occupational title. Then, we filtered the list of professions by only including those that possess the *feminine form of label* (wdt:P2521) property for the Italian language. This property denotes the feminine variant of the occupational title, where applicable. The next step consists of filtering out occupational titles for which the gender is not easily distinguishable from text, such as those in which both gender variants share the same lexical root (e.g. the aforementioned *il giudice/la giudice*), or those that do not feature gender variants at all (e.g. *la guardia*, i.e. the guard). We also removed all occupational titles that consist of two or more tokens. Then, we reduced the list by filtering out polysemous words. A common example of polysemy in the Italian language occurs when an occupational title shares the same lexical form as the discipline to which it belongs, such as *matematica* (feminine form of *mathematician*), or *fisica* (feminine form of *physicist*). For each occupational title, we used WordNet to find all synsets in which it appears and then removed it if the synset is a hyponym of the *discipline.n.01*

[13] `https://spacy.io/`

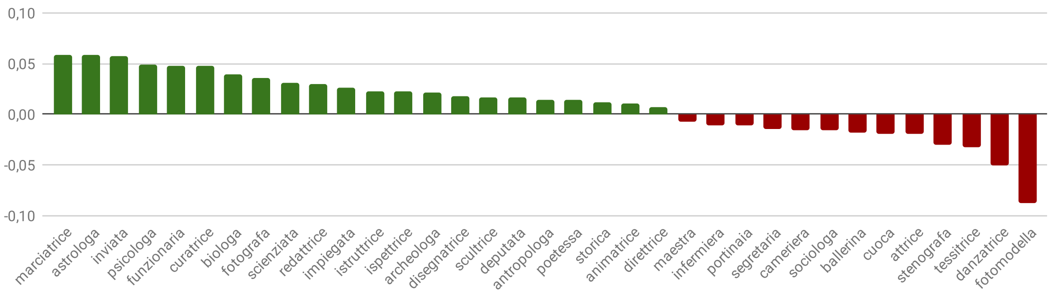


FIGURE 8.5: Final set of occupational titles (the feminine form is reported) and the slope of $odds(w)^t$.

synset. Moreover, we manually analyzed the list of remaining occupational titles and removed other instances of polysemy, which would otherwise hinder the quality of the results. For instance, we filtered the word *editrice* (feminine form of *editor*) as it can also appear in the phrase *casa editrice* (i.e. publishing house), and the word *tecnica* (feminine form of *technician*), which can also refer to the word *technique* depending on context. We also decided to remove words that have additional figurative meanings, such as *cacciatrice* (feminine form of *hunter*) and guerriera (feminine form of *warrior*). This process was undertaken by two independent annotators and then checked for agreement. The final result of this process is $T$, a set of tokens that unequivocally refer to occupational titles, and that feature distinct masculine and feminine gender variants which can reliably be extracted from text.

### 8.3.3 Experimental setup

Once we have acquired the set of occupational titles $T$, the next step of the analysis consisted of measuring the frequency with which each term $w \in T$ occurs for each year in the corpus described in Section 8.3.1. We also make use of the lexical information contained in said corpus in order to eliminate any remaining ambiguity in the words. In fact, for each occupational title, we counted a hit in the corpus if it appears with the NOUN tag. This allows us to avoid counting occupational titles that can be confused with verbs or adjectives, such as *impiegato/impiegata*, which can refer to the noun *employee* in Italian, but also to the past participle conjugation of the verb *to employ*.
Moreover, we only counted a hit if the word has been registered with the singular form. This is done for two reasons: first, occurrences of the plural form are outside the scope of this investigation, because in Italian the masculine plural form is traditionally used as the default, while the feminine variant of the plural is only used in exceptional cases, such as when referring to a

group that is composed entirely of women. Second, this strategy filters out cases where the plural form shares the same lexical root as one of the gender variants. An example of this is the word *infermiere* (i.e. *nurse*), which can refer to both the singular masculine form (as in *l'infermiere*), or the plural feminine form (as in *le infermiere*).

Since the objective of this study is to observe the trends in the use of masculine and feminine forms for occupational titles, we are interested in analyzing how their frequency changes from one year to the other. However, measuring the absolute frequency in each year for both forms would be misleading, as it heavily depends on the amount of data that is available for each year in the corpus. Instead, we compute the smoothed relative frequency $p_w^t$ for each word $w$ and each year $t$ using the following formula:

$$p_w^t = \frac{f_w^t + 1}{C^t + \mid V^t \mid} \tag{8.2}$$

where $f_w^t$ is the frequency of word $w$ in the year $t$, $C^t$ is the count of tokens occurring in the corpus the year $t$ and $|V^t|$ is the vocabulary length computed on the year $t$. We compute $p_w^t$ for both gender forms of each occupational title. Then we compute $odds(w)^t$ which represents the log ratio of the smoothed relative frequency of the feminine and masculine forms respectively:

$$odds(w)^t = log\frac{p_{w_f}^t}{p_{w_m}^t} \tag{8.3}$$

Operationally, $odds(w)^t$ specifies the probability that the feminine variant will appear in a text relative to the masculine form in the specified year $t$. We then obtain the time-series by concatenating the $odds(w)$ values computed for each year: $(odds(w)^{1948}, odds(w)^{1949}, .., odds(w)^{2004})$. Assuming a linear course of the time-series, three different scenarios can occur: *(i)* the occurrences of the feminine form are growing; *(ii)* the occurrences of the masculine form are growing; *(iii)* the ratio of the masculine and feminine form of an occupational title are stable over time. We computed the regression line of the time-series, using the linear least-squares regression method provided by the SciPy library[14]. We use the slope of the regression line to determine whether the values of $odds(w)^t$ are changing over time. If the slope is positive/negative, $odds(w)^t$ is increasing/decreasing over time, which means that the frequency of $w_f$ is increasing/decreasing faster than that of $w_m$, or that the frequency of $w_m$ is decreasing/increasing faster than that of $w_f$. For each regression line,

[14] https://www.scipy.org/

we also compute the statistical significance of the slope parameter relying on the Wald Test [65]. Specifically, the null hypothesis states that the slope parameter of the regression line is zero. In this stage, occupational titles for which we get a $p - value > 0.1$ are filtered out.

### 8.3.4 Results

Figure 8.5 describes the value of the slope for each occupational title. Depending on the sign of the slope, we can identify two distinct groups of occupational titles. Green bars indicate that the slope of $odds(w)^t$ is positive, i.e. the frequency of the feminine form is increasing relative to that of the masculine form. On the other hand, red bars indicate that the slope is negative, thus the frequency of the feminine form is decreasing relative to that of the masculine form. Out of 35 occupational titles, 22 have a positive slope, while 11 result in a negative slope. In particular, the most positive slope is the one associated to *marciat-ore/-rice* (i.e. *racewalker*), while the most negative slope is *fotomodell-o/-a* (i.e. *fashion model*).

For many of these titles, the resulting slope can be mapped to specific social changes. An interesting example in this regard is *infermiere* (i.e. *nurse*), to which a negative slope is recorded: indeed, in Italy the position of nurse has been opened to men starting from 1971[15]. The odds(w) time series of infermiera/infermiere is reported in Figure 8.6.

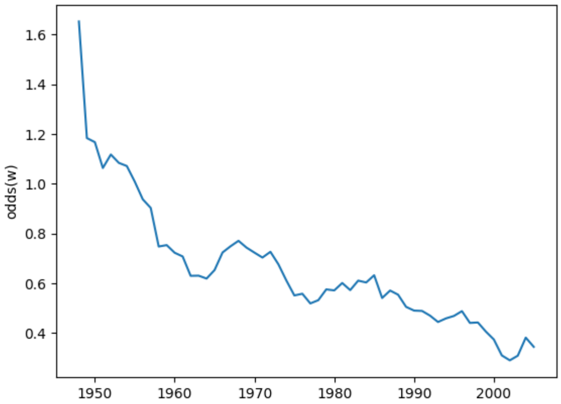


FIGURE 8.6: 10-year moving average of odds(w) for infermiera/infermiere.

Moreover, results show that managerial roles such as *funzionaria* (i.e. *civil servant*), *ispettrice* (i.e. *inspector*), *direttrice* (i.e. *director*) are associated to a

[15] https://www.gazzettaufficiale.it/eli/id/1971/04/03/071U0124/sg

positive slope, which is indicative of a stronger perception of women in such roles.

A similar push can be observed also in the scientific domain, with a positive trend for the words *biologa* (i.e. *biologist*), *scienziata* (i.e. *scientist*), as well as the artistic one. On the other hand, we observe an increase in the usage of the masculine form for *segretario* (i.e. *secretary*), *ballerino* (i.e. *dancer*), and *stenografo* (i.e. *stenographer*).

In the second part of the experiment, we attempt to identify the people that have driven the change in the usage of the feminine and masculine forms of an occupational title. To do this, we retrieve the Named Entities (NEs) to which the occupational titles refer for each year, and monitor their frequency. In particular, we exploit the UDPipe annotations to extract valid NEs, i.e. entities that are directly connected to an occupational title via a dependency relation.

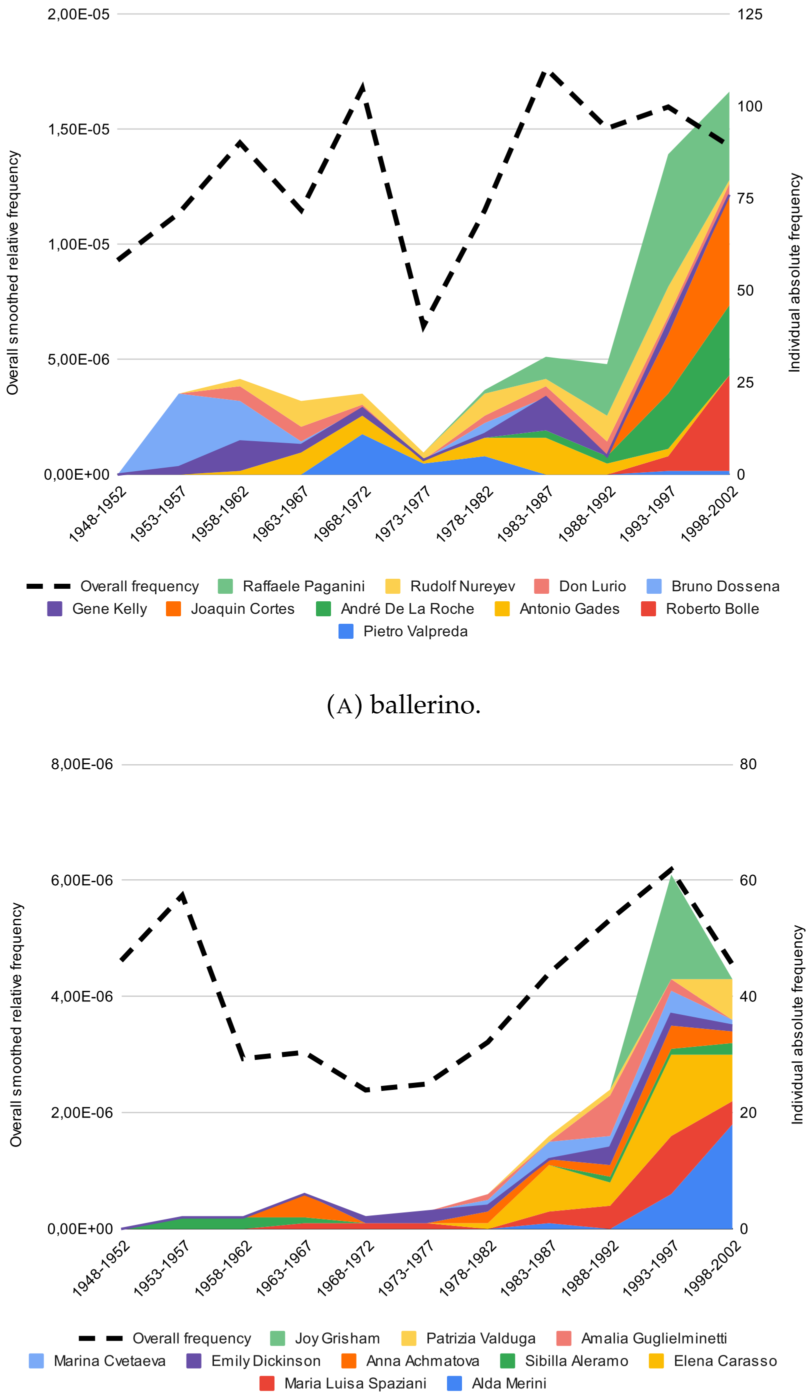
In Figure 8.7, we report the NEs extracted for two particular occupational titles: *ballerino* (i.e. male dancer) and *poetessa* (i.e. feminine poet). We have chosen these titles because they feature the largest number of occurrences of NEs in the corpus. The data is presented in the form of stacked line charts, which report the absolute frequency of each NE so that the height of a coloured line represents how many times a NE has been mentioned within a specified period. The dotted black line reports the overall smoothed relative frequency for the occupational title. Both the absolute frequency of NEs mentions and the overall smoothed relative frequency are aggregated in bins of 5 years.

Three male dancers are referenced over a wide period due to their historical role in the field: *Rudolf Nureyev*, *Antonio Gades* and *Gene Kelly*. However, the last years have seen a rise in popularity of new figures such as *Raffaele Paganini*, *Joaquin Cortes*, *André de La Roche* and *Roberto Bolle*.

Occurrences of specific female poets in the corpus keep low until the late '70s. Ignoring a spike in 1953-1957, probably due to the quality issues in the data collected, the individual absolute frequency of NE mentions seems to agree with the overall smoothed relative frequency of the noun *poetessa*. In the 1988-2002 period, four figures overwhelm the scene: *Joy Grisham*, *Elena Carasso*, *Maria Luisa Spaziani* and *Alda Merini*. Even though the first work of *Maria Luisa Spaziani* dates back to 1954, we observe a significant rise in the occurrences in the early '90s, when she is nominated three times for the Nobel Prize for Literature [16]. The increase in NE mentions over time is even more apparent in this case, however, it follows a different trend compared

[16] https://en.wikipedia.org/wiki/Maria_Luisa_Spaziani

to that of the overall frequency of the noun *poetessa*, which suggests that the word may have been used differently in the earliest period.



(A) ballerino.

(B) poetessa.

FIGURE 8.7: Occurrences of Named Entities associated to two occupational titles. The X-axis reports the time periods. The left Y-axis reports the overall smoothed relative frequency of the occupational title. The right Y-axis reports the absolute frequency of each Named Entity.

# Chapter 9

# Conclusions

## 9.1 Summary of the Thesis

In this thesis we have approached the problem of modelling language change using computational approaches from different perspectives. Particular attention has been paid to lexical-semantic change, which poses significant challenges from both a computational and a linguistic point of view and allows the full potential of computational models to be expressed. Based on well-established linguistic theories of language change, we aim to derive computational models that can automatically track change, and to develop methodologies for evaluating these models.

Specifically, we aim to use computational techniques to quantify, predict and potentially explain patterns of semantic shift in meaning and usage. The core premise of our study underscores the belief that understanding the trajectory of word meanings can provide essential insights into the historical, cultural and social contexts from which they emerge and evolve.

As illustrated in the previous chapters, addressing the multifaceted nature of language change computationally involves numerous challenges that cut across different research disciplines. These complexities have guided the trajectory of this work. As such, the contributions in this thesis address several key concerns:

- We have proposed a systematic review and classification of Temporal Aligned Language Models
- With XL-LEXEME we have advanced the state-of-the-art for Lexical Semantic Change Detection for English, German, Swedish and Russian languages

- We have provided methodological insights on the creation of Diachronic resources such as historical corpora preprocessing and the creation of a benchmark for Lexical Semantic Change Detection
- We have introduced longitudinal studies of Language Change, for both morphological and semantic changes

## 9.2 Answers to the Research Questions

This section provides a comprehensive response to each of the RQs defined in chapter 1.1, based on the results of the work presented in chapters 4 to 8.

### 9.2.1 RQ1. How do different models perform across diverse languages and datasets when subjected to benchmarks specifically designed for Lexical Semantic Change Detection?

In the leading lexical-semantic change detection methods, models that use temporal alignment have gained prominence. This is evident from the results of SemEval-2020 Task 1 and DIACR-Ita, where many of the best performing models adopted the temporal alignment approach. Despite the growing popularity of contextualised models in other areas of NLP, they consistently underperform in lexical-semantic change detection. However, the correlations obtained are relatively weak.
The systems presented in RuShiftEval represented a paradigm shift. These systems heralded a new era in the field, consistently demonstrating strong correlations with the test set, especially in the graded task of lexical-semantic change. In particular, the best performers among them are synchronous systems, which are mainly trained on Word Sense Disambiguation or Word-in-Context datasets with synchronous data.
Following these developments, XL-LEXEME was introduced. XL-LEXEME (Section 7.1) stands out as the first model to achieve significant correlation in the graded lexical-semantic change task, with results not seen in related work. Interestingly, XL-LEXEME excels in state-of-the-art performance across all languages, save for Latin. This exception can be attributed to the absence of ancient Latin both in the foundational language model (XLM-R) and the WiC training dataset utilized for training XL-LEXEME. This challenge with Latin underscores the continued relevance of Temporal Language

Models, which, despite requiring less data, often outshine in performance when dealing with low-resource languages.

The road to progress in lexical-semantic change is not without its challenges. A major area of concern remains the binary task. There is a noticeable lack of research aimed at the binary task. In this Thesis to bridge this gap, we have introduced two pioneering studies: (i) the Gaussian Mixture model, as discussed in Section 4.2, and (ii) a comprehensive exploration of the distribution of change scores across various models and languages, detailed in Section 4.1.

### 9.2.2 RQ2. To what extent are synchronic models equipped to understand and track diachronic language changes?

In Section 7.1, we introduced a model designed to investigate lexical semantic changes using synchronic data. This model utilized a dataset sourced from the Word-in-Context task, predominantly featuring contemporary corpora. Our findings underscored the capability of such models to detect semantic shifts effectively, even when relying on data spanning diverse historical periods.

A striking attribute of the proposed model is its adaptability. The model demonstrated versatility across varied contexts: being repurposed for unrelated tasks, applied to unfamiliar languages, and most notably, when analyzing data from epochs distinct from their training period. A distinctive edge of the approach introduced in Section 7.1 is its independence from a sense inventory, commonly associated with WSD-based methods. This unique feature possibly bolsters its prowess in generalizing unseen word usage scenarios during training. We conclude from the results that models nurtured on synchronic data provide excellent performance in modelling historical data and are effective in detecting changes in language.

A demonstration of the ability of synchronic models to detect linguistic change is given in 8.3. In this work, we used an NLP pipeline of synchronic models to detect changes in the morphology of the Italian language. Specifically, the pipeline includes several tools from the spaCy[1] library, including a tokenizer, a lemmatizer, a morphological analyser, a named entity recogniser and a PoS tagger.

[1] https://spacy.io

While the outcomes of this study were largely qualitative, they resonated with our preliminary assumptions, further affirming the proficiency of synchronic model strategies. Through this integrated approach, the study successfully identified instances of occupational titles within an expansive historical newspaper archive, spanning 1948 to 2004. This enabled a comprehensive examination of the morphological variants of the occupational titles, leading to the derivation of a morphological change score.

### 9.2.3 RQ3. How can benchmarks and resources for Language Change be effectively designed to ensure comprehensive and accurate results?

In Section 6.1, we elucidated the process of creating a historical corpus, a significant feature of which is the pronounced occurrence of OCR errors. Such errors, while a common fixture of historical corpora, can be particularly disruptive to analyses, especially when examining language change over time.
In 6.2, we introduced a benchmark dedicated to exploring shifts in the meanings of words in the Italian language. Notably, our methodology diverges from conventional ones cited in literature. For instance, we opted not to employ the concept of semantic proximity, a staple in datasets like SemEval 2020 Task 1 for English, Swedish, and German, as well as the RuShiftEval and RuSemShift datasets for Russian.

Our benchmark, DIACR-Ita, draws inspiration from standard WSD annotation techniques. The unique aspect of our annotation demands the elucidation of a word's meaning in each of its instances. The fruition of this method was largely owed to the Sabatini-Coletti dictionary for Italian. This lexicon pinpoints the first instance in which word senses appear. The Sabatini-Coletti proved instrumental during the preliminary stages, assisting in earmarking potential words with change meanings, and later during the annotation phase by offering a comprehensive inventory of potential word senses.

However, while lexicographic resources such as the Sabatini Coletti are invaluable for detecting semantic shifts, they sometimes fall short when used in combination with corpus linguistics. This shortcoming is mainly due to the fact that such resources may not cover the full range of word senses, their sense annotations may be outdated, or the data available to the author at the time of editing may be insufficient. Furthermore, certain senses recorded in these resources may not be arranged within the specific corpus under study.

This discrepancy could be due to the fact that the sense belongs to a different domain than the corpus, or simply due to its rarity, making it elusive to sampling methods.

To bridge this gap, the integration of qualitative linguistic analysis with computational techniques is recommended. Echoing this sentiment, we introduced in 8.2 the Diachronic Engine. This tool augments traditional corpus linguistic instruments, incorporating temporal data masks and state-of-the-art methods to explore into the time series of language change.

### 9.2.4 RQ4. How effectively can large-scale quantitative studies capture and quantify the influence of socio-cultural events on language change over time?

Large-scale longitudinal studies offer invaluable insights into the intricate interplay between historical socio-cultural events and language transformation. Typically, linguistic shifts occur subtly and continuously. By examining vast repositories of documents that stretch over extensive periods, we obtain a vantage point that closely mirrors the natural progression of these changes. While such an analysis holds immense potential for various disciplines, numerous intricacies await exploration and resolution. In Section 4.3, we assessed different models designed to generate time series of semantic shifts in the Italian lexicon. The outcome revealed a notable variance between the produced results and the annotations found in lexicographic resources. This discrepancy could stem from inaccuracies within the lexicographic sources, noise present in the analyzed corpus, or, most significantly, the precision of computational methodologies employed. A pressing challenge in this domain involves pinpointing tools adept at detecting change points in time series. Our exploration in 4.3 was somewhat hampered by our wide search approach, probing diverse semantic changes without specific boundaries.

Conversely, the outcomes of the work presented in Section 8.3, harmonized with our preliminary hypotheses. We deliberately narrowed our focus, targeting specific linguistic variations. This approach enabled us to discern correlations between shifts in the morphology of the Italian language and the socio-cultural dynamics of the given era. Specifically, we delved into the morphology of occupational titles, seeking to discern if languages with gender-based grammatical systems, like Italian, exhibit morphological evolutions influenced by societal occurrences.

For instance, post the 1970s, there was a marked rise in the masculine form of the word *infermiere* (i.e. *nurse*), correlating with the period when the nursing profession opened its doors to men. Simultaneously, there was a surge in feminine grammatical occurrences associated with managerial positions, signaling an uptick in the female presence in these roles. A subsequent qualitative assessment revealed that shifts in occupational titles often aligned with renowned or influential figures.

Large-scale longitudinal studies are those that allow us to analyse how historical cultural or social events intervene in language change. Language change usually occurs gradually over time, and the ability to observe it over large collections of documents spanning decades provides a perspective that is very close to the real situation. Although this type of analysis is very interesting for various applications, several aspects remain to be understood and resolved.

## 9.3 Limitations and Future Work

In this thesis we have proposed various computational approaches to language change, with a particular focus on semantic change over time. It quickly becomes apparent that there is a need for approaches that also cover other types of change, primarily at other levels of language, such as syntax, which has been extensively studied in linguistics, as in the case of grammaticalisation, or at the phonetic level, with various approaches present in the linguistic literature and significant phonetic changes that have occurred over time, such as the Great Vowel Shift. In addition, changes in the language that are not related to time, such as geographical changes, in particular those due to linguistic variation and the spread and development of dialects, remain under-researched.

In particular, there has been little or no effort to develop computational approaches for studying the interaction between languages over time and space. Regarding the specific problem of semantic change over time, although research in this area has progressed rapidly in recent years, important open questions remain that require greater effort.

In particular, while there are now high performance models for ranking words in order of semantic change, the question of accurately detecting semantic change remains open and unsolved. The problem of recognition becomes much greater when moving from a simplified setting of two time periods to more extensive spans involving hundreds of time periods. In this particular setting, there is a significant gap in the literature, both in terms of models and in terms of data and evaluation. Moving to multiple time periods opens up several avenues, with two main dimensions of the problem that require aggregation: the temporal dimension and the dimension of meanings. The temporal dimension can be complex to model because it involves different sub-factors, such as the granularity of the temporal data and the level of aggregation, or the different possibilities for comparison, such as comparison with the first or last period, or comparison with successive periods. At the same time, however, the temporal dimension can provide interesting insights, and looking at time series as a whole can reveal patterns that are not apparent when only two time periods are considered. The dimension of meanings, on the other hand, is a variable that is difficult to process because the senses of a word may be less or more represented in a given historical period, and their absence may only be due to an under representation of the rarer senses of a word, making the frequency of senses in general a major obstacle. The granularity of meaning is also a difficult hurdle to overcome, if not by imposing a priori constraints on the analysis.

On the other hand, the possibility of building models that can explain the predictions remains unexplored. The most powerful models today only provide a score that says little about what kind of semantic change has occurred and why. In particular, there is a need to develop models that can sufficiently differentiate the senses of a word over time to highlight precisely which meaning has changed. It is also necessary to develop models capable of classifying types of semantic change, which, as described in the Background chapter, can vary according to the linguistic literature. It is necessary to identify the cause of the semantic change, which can be of various kinds, social, cultural or other. And once the nature of the cause has been identified, it is necessary to identify possible triggering factors, such as particular historical events, e.g. the introduction of a new technology or the outbreak of a war.

Finally, there is a need to develop new technologies for studying language change, including for under-represented languages. Language change is a complex phenomenon to analyse because it requires many resources, and existing models for making robust predictions contribute to increasing the amount of data needed. It is therefore necessary to introduce new methods and models that can overcome these obstacles. One solution could be to use the latest Large Language Models for the synthetic generation of large-scale linguistic phenomena. This solution would allow current studies to be extended to different time periods and languages not currently covered, or pilot studies to be carried out before embarking on large and costly data collections.

# Appendices

# Appendix A

## A.1 Hyper-parameters

| **Hyper-parameter** | **Value** |
|---|---|
| hidden act | gelu |
| hidden dropout prob | 0.1 |
| hidden size | 1024 |
| initializer range | 0.02 |
| intermediate size | 4096 |
| layer norm eps | 1e-05 |
| max position embeddings | 514 |
| num attention heads | 16 |
| num hidden layers | 24 |
| position embedding type | absolute |
| vocab size | 250004 |
| learning rate | |
| cross-encoder | 1e-05 |
| XL-LEXEME | 1e-05 |
| weight decay | |
| cross-encoder | 0.01 |
| XL-LEXEME | 0.00 |
| max sequence length | |
| cross-encoder | $\lambda = 256$ |
| XL-LEXEME | $\lambda* = 128$ |

TABLE A.1: XL-LEXEME and cross-encoder hyper-parameters.

# Appendix B

## B.1 Tug of War

## B.2 Self-embedding distance (SED)

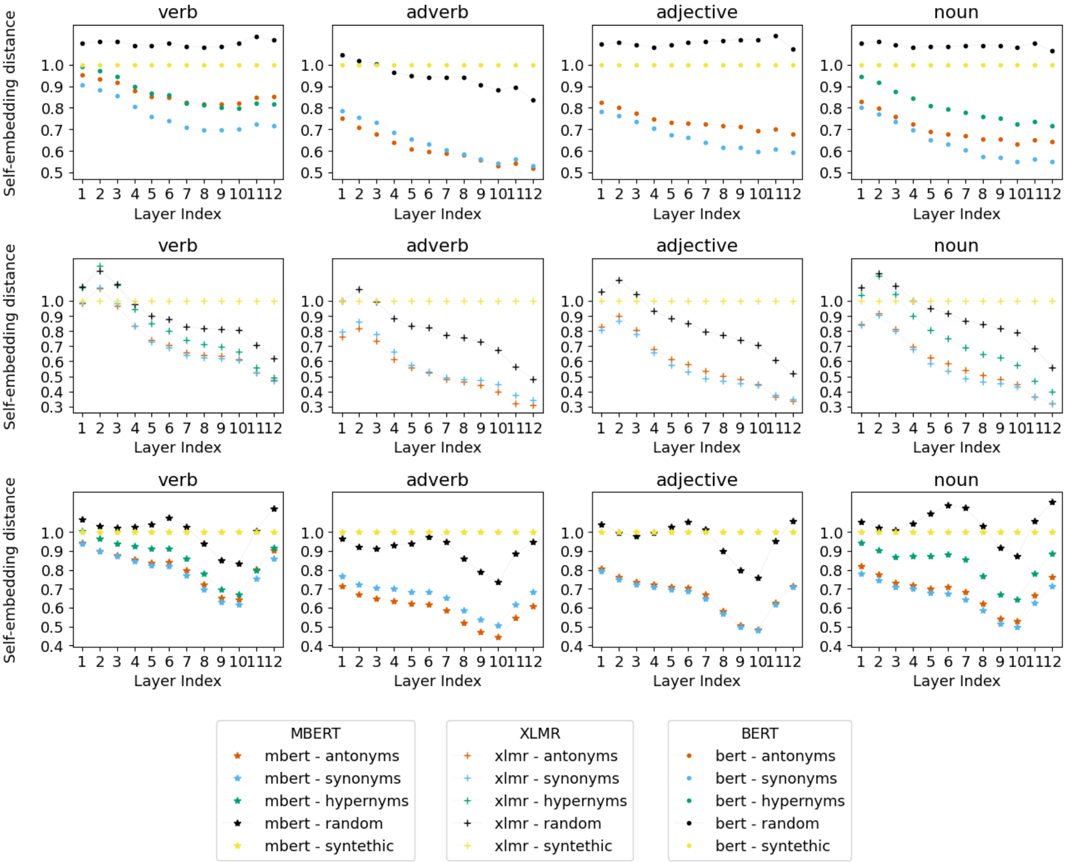


FIGURE B.1: Average SED over layers

| | | WiC-en | | | MCL-WiC-en | | | XL-WiC-fr | | | XL-WiC-it | | | WiC-ITA-it | | | DWUG-de | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Instances | *Dev* | *Train* | *Test* | *Dev* | *Train* | *Test* | *Dev* | *Train* | *Test* | *Dev* | *Train* | *Test* | *Dev* | *Train* | *Test* | *Dev* | *Train* | *Test* |
| **nouns** | *TP* | 196 | 1474 | 422 | 290 | 2093 | 263 | 275 | - | 261 | 65 | 340 | 180 | 166 | 1153 | 174 | 778 | 2402 | 783 |
| | *TN* | 199 | 1320 | 409 | 292 | 2031 | 265 | 273 | - | 253 | 62 | 312 | 180 | 156 | 463 | 175 | 168 | 465 | 163 |
| **verbs** | *TP* | 123 | 1240 | 278 | 126 | 1095 | 149 | 120 | - | 117 | 34 | 228 | 115 | 36 | 414 | 31 | 349 | 1081 | 326 |
| | *TN* | 120 | 1394 | 291 | 120 | 1175 | 149 | 142 | - | 155 | 37 | 259 | 116 | 50 | 200 | 49 | 249 | 680 | 270 |
| **adjectives** | *TP* | - | - | - | 78 | 724 | 76 | 84 | - | 101 | - | - | - | 46 | 347 | 27 | - | - | - |
| | *TN* | - | - | - | 80 | 706 | 68 | 72 | - | 83 | - | - | - | 38 | 119 | 25 | - | - | - |
| **ALL** | *TP* | 319 | 2714 | 700 | 500 | 4000 | 500 | 500 | - | 500 | 99 | 568 | 295 | 250 | 1999 | 250 | 1127 | 3483 | 1109 |
| | *TN* | 319 | 2714 | 700 | 500 | 4000 | 500 | 500 | - | 500 | 99 | 571 | 296 | 250 | 806 | 250 | 417 | 1145 | 433 |

TABLE B.1: True Positives (TP) and True Negatives (TN) for different benchmarks and part-of-speech

| | | WiC-en | | MCL-WiC-en | | XL-WiC-fr | | XL-WiC-it | | WiC-ITA-it | | DWUG-de | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Layer | *Train* | *Test* | *Train* | *Test* | *Train* | *Test* | *Train* | *Test* | *Train* | *Test* | *Train* | *Test* |
| **nouns** | *1* | **0.593** | **0.606** | **0.656** | **0.727** | - | **0.591** | **0.642** | **0.679** | **0.644** | **0.555** | **0.808** | **0.799** |
| | *6* | **0.697** | **0.656** | **0.720** | **0.812** | - | **0.661** | **0.678** | **0.713** | 0.694 | **0.699** | **0.808** | **0.800** |
| | *12* | **0.691** | **0.683** | **0.796** | **0.847** | - | **0.677** | **0.678** | **0.689** | 0.682 | **0.759** | **0.806** | **0.809** |
| **verbs** | *1* | 0.558 | 0.513 | 0.572 | 0.606 | - | 0.530 | 0.569 | 0.562 | 0.478 | 0.539 | 0.597 | 0.597 |
| | *6* | 0.657 | 0.613 | 0.683 | 0.782 | - | 0.561 | 0.581 | 0.582 | 0.575 | 0.586 | 0.643 | 0.633 |
| | *12* | 0.615 | 0.611 | 0.752 | 0.805 | - | 0.622 | 0.639 | 0.593 | 0.459 | 0.615 | 0.630 | 0.640 |
| **adjectives** | *1* | - | - | 0.650 | 0.699 | - | 0.539 | - | - | 0.666 | 0.653 | - | - |
| | *6* | - | - | 0.703 | 0.789 | - | 0.574 | - | - | **0.706** | 0.687 | - | - |
| | *12* | - | - | 0.738 | 0.819 | - | 0.653 | - | - | **0.733** | 0.695 | - | - |

TABLE B.2: **BERT**: Comparison of F1 score scores obtained by threshold-based classifiers for Word in Context (WiC) trained on development sets. The threshold is determined on the development set and applied to both the train and test sets. Results are provided for different parts of speech and WiC benchmarks. We report in bold the best result for each considered layer, benchmark, and data sets.

## B.3 Word-in-Context (WiC)

In the WiC experiments, we leveraged pre-trained models without performing any kind of fine-tuning (therefore not using the Train set). The Dev set, as in traditional experimental settings, was used to tune the only parameter involved in our experiments: the threshold used to determine if the words in the two sentences have the same meaning or not. In other words, pairs of examples for which the cosine similarity is lower than the threshold are classified as negative; otherwise, they are classified as positive. In our experiments, we follow Pilehvar and Camacho-Collados [172] in tuning the threshold-based classifier on the development set (Dev), thus being able to compare with their results. Additionally, for French, a Train set is not available. PoS-specific thresholds have been computed to guarantee a fair evaluation across different PoS tags. Regarding the size of the PoS-specific sets, we argue that the number of examples, even if small, is sufficient to establish a binary threshold. This is supported by our results across different datasets, which achieve comparable outcomes to the SOTA result on the full set (not PoS split, as shown in Table 7.4). We would like to highlight that similar results are obtained when using the Train set instead of the Dev set for the

| | | WiC-en | | MCL-WiC-en | | XL-WiC-fr | | XL-WiC-it | | WiC-ITA-it | | DWUG-de | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Layer | *Train* | *Test* | *Train* | *Test* | *Train* | *Test* | *Train* | *Test* | *Train* | *Test* | *Train* | *Test* |
| **nouns** | *1* | 0.530 | 0.521 | **0.592** | **0.632** | - | **0.588** | **0.612** | **0.645** | 0.571 | 0.582 | **0.819** | **0.811** |
| | *6* | **0.624** | **0.628** | 0.675 | **0.738** | - | **0.625** | **0.690** | **0.675** | 0.691 | 0.691 | **0.807** | **0.794** |
| | *12* | **0.644** | **0.611** | 0.705 | **0.780** | - | **0.707** | **0.644** | **0.680** | **0.691** | **0.713** | **0.813** | **0.812** |
| **verbs** | *1* | **0.563** | **0.547** | 0.564 | 0.576 | - | 0.520 | 0.562 | 0.524 | 0.477 | 0.601 | 0.618 | 0.617 |
| | *6* | 0.604 | 0.555 | 0.622 | 0.682 | - | 0.571 | 0.596 | 0.554 | 0.555 | 0.566 | 0.669 | 0.627 |
| | *12* | 0.623 | 0.586 | 0.698 | 0.720 | - | 0.662 | 0.552 | 0.543 | 0.578 | 0.506 | 0.667 | 0.675 |
| **adjectives** | *1* | - | - | 0.548 | 0.568 | - | 0.531 | - | - | **0.633** | **0.614** | - | - |
| | *6* | - | - | **0.676** | 0.736 | - | 0.598 | - | - | **0.715** | **0.750** | - | - |
| | *12* | - | - | **0.707** | 0.735 | - | 0.701 | - | - | 0.600 | 0.683 | - | - |

TABLE B.3: **mBERT**: Comparison of F1 score scores obtained by threshold-based classifiers for Word in Context (WiC) trained on development sets. The threshold is determined on the development set and applied to both the train and test sets. Results are provided for different parts of speech and WiC benchmarks. We report in bold the best result for each considered layer, benchmark, and data sets.

threshold selection.

### B.3.1 The DWUG benchmark

In the original SemEval DWUG benchmarks [205], human annotators were provided with a target word and a pair of sentences and asked to estimate the degree of change on a scale from 1 (change) to 4 (identical). The task is analogue to a multi-class WiC task. We converted the original annotations to a binary WiC benchmark. Specifically, we transformed the ground truth for a specific pair of sentences as follows: if the average agreement of the Lexical Semantic Change annotations was greater than 3.5, we considered the meaning to be the same (assigned a label of 1); for sentence pairs with an average agreement lower than 1.5, we considered the meaning to be different (assigned a label of 0).

### B.3.2 Size of WiC benchmarks

We split the considered WiC benchmarks in distinct sub-corpora for nouns, verbs, and adjectives. In TableB.1, we provide the number of true (i.e. 1, same meaning) and negative (i.e. 0, different meaning) instances for each sub-corpus.

### B.3.3 Macro-F1 across BERT layers

We examine the PoS sensitivity using WiC by comparing embedding similarities across all BERT and mBERT layers. For sake of simplicity, we provide additional results using Macro F1-score for layers 1, 6, and 12 across various

PoS categories and corpora. Results are consistent across the layers, confirming the best performance for the noun syntactic class. We have omitted the WiC results for XLM-R due to its tokenisation characteristics. For numerous pairs of sentences, the XLM-R tokeniser generates more tokens than BERT and mBERT, surpassing the input limit and causing the exclusion of target words.

## B.4 WiC polysemy per language

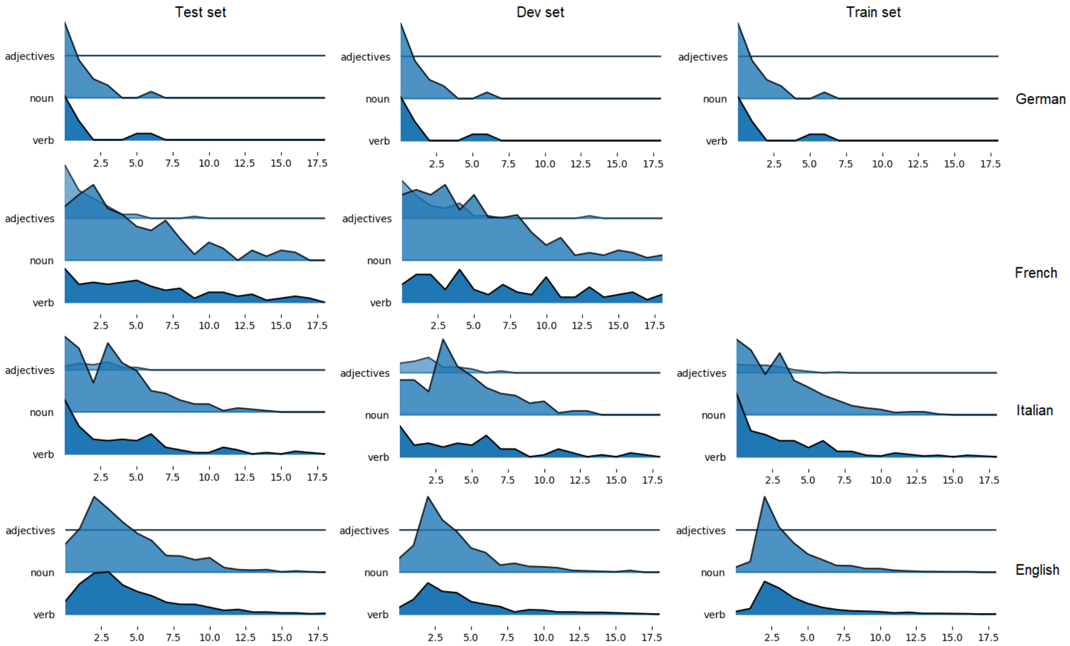


FIGURE B.2: Distribution of polysemy (i.e., number of senses on x-axis) of the target words in each Dev, Test, and Train WiC set for each considered language

| References | Benchmark | # targets |
|---|---|---|
| [205] | DWUG-English | 46 |
| [205] | DWUG-German | 50 |
| [205] | DWUG-Swedish | 44 |
| [239] | DWUG-Spanish | 100 |

TABLE B.4: References and number of targets for each consider artificial corpus

## B.5 `Random` Lexical Semantic Change

### B.5.1 Artificial diachronic corpus

We generated an artificial diachronic corpus for LSC by utilising the SemEval and LSCDiscovery benchmakrs for LSC in DWUG format[1] (see Table B.4). Instead of incorporating data from both time periods, $T_1$ and $T_2$, we discarded information from the first time period as it is more likely to contain word meanings outside the pre-trained knowledge of the models under examination. We created two distinct artificial sub-corpora, $C_1$ and $C_2$, by randomly sampling occurrences from the data of the second time period $T_2$. The DWUG English dataset contains data for 46 target words.

For each target $t$, we considered all sentences where another target $t1$, with $t1 \neq t$, appeared as potential candidates to emulate instances of semantic change. We simulated a change instance through a *random* replacement, that is by replacing $t$ in the sentence where $t1$ occurred – i.e., $t1 \leftarrow t$. We sample a varying number of sentences and perform replacements for each target, thereby emulating a varying degree of semantic change.

[1] English: `https://zenodo.org/records/5796878`, German: `https://zenodo.org/records/5796871`, Swedish: `https://zenodo.org/records/5090648`, Spanish: `https://zenodo.org/records/6433667`

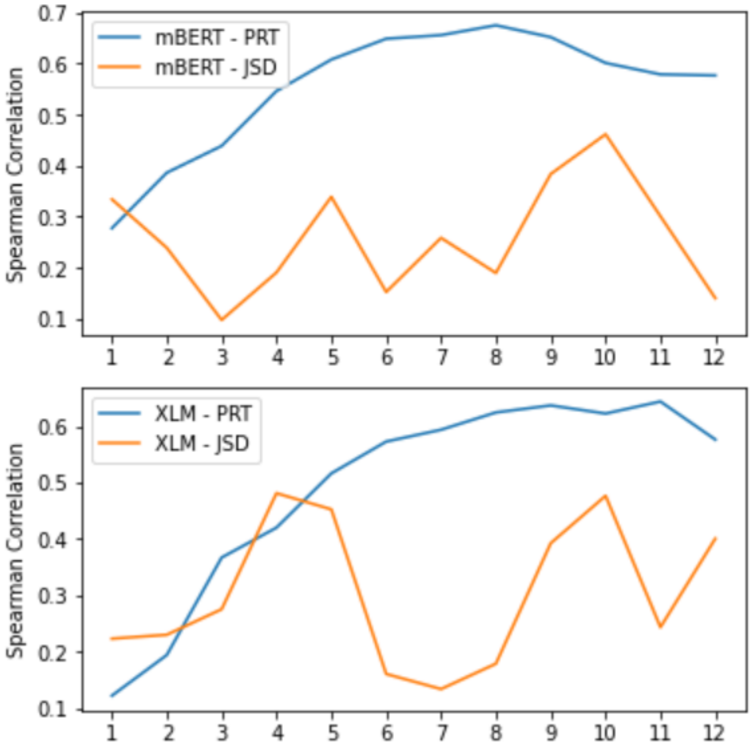


FIGURE B.3: PRT and JSD performance on the artificial LSC dataset

## B.6 Lexical Semantic Change

| Word | Time span | (Ranked) Farthest replacements | $lsc_w$ (k=1) |
|---|---|---|---|
| attack | T1 | **physical**, degeneration, blast, crime, disease, death, condition, plane, affliction, birthday attack | -0.036 |
| | T2 | **approach**, force, onslaught, assault, exploit, challenge, commencement, aim, worth, signal | 0.059 |
| bit | T1 | **nominative case**, accusative case, cryptography, information theory, bdsm, time,point, binary digit, sociologic, sublative | -0.018 |
| | T2 | **saddlery**, chard, illative case, iron, bevelled, tack, small, gun, cut, elative case | 0.067 |
| circle | T1 | **wicca**, circumlocution, encircle, astronomy, tavern, semicircle, around, logic, go,wand | 0.002 |
| | T2 | **pitch**, place, graduated, figure, disk, territorial, enforce, worship, line, bagginess | 0.064 |
| edge | T1 | **brink**, cricket, instrument, margin, polytope, side, edge computing, verge, demarcation line, demarcation | -0.015 |
| | T2 | **data**, production, climax, division, superiority, organization, sharpness, graph, win, geometry | 0.047 |
| graft | T1 | **lesion**, bribery, felony, politics, bribe, corruption, autoplasty, surgery, nautical, illicit | -0.047 |
| | T2 | **branch**, stock, tree, fruit, shoot, join, cut, graft the forked tree, stem, portion | 0.103 |
| head | T1 | **headland**, head word, capitulum, syntactic, pedagogue, fluid dynamics, hip hop, headway, pedagog, word | 0.004 |
| | T2 | **leader**, organs, implement, top, tail, foreland, chief, bolt, axe, forefront | 0.084 |
| land | T1 | **real estate**, real property, surface, property,build, physical object, Edwin Herbert Land, electronics, landing, first person | -0.032 |
| | T2 | **realm**, country, kingdom, province, domain, people, homeland, territory, nation, region | 0.076 |
| lass | T1 | **sweetheart**, girl, missy, woman, yorkshire, lassem, lasst, lassie, loss, miss | 0.014 |
| | T2 | **fille**, dative case, jeune fille, loose, lasses, unattached, young lady, young woman, north east england, past participle | 0.099 |
| plane | T1 | **airplane**, aeroplane, pt boat, heavier-than-air craft, glide , boat, lycaenidae, lift, bow, hand tool | -0.197 |
| | T2 | **geometry**, point, shape, surface, flat, degree, form, range, anatomy, smooth | 0.205 |
| player | T1 | **media player**, idler, soul, thespian, person, individual, trifler, performer, somebody, histrion | -0.065 |
| | T2 | **contestant**, performing artist, actor, musician, musical instrument, music, gamer, theater, player piano, play the field | 0.042 |
| prop | T1 | **props**, airscrew, astronautics, actor, airplane propeller, seashell, stagecraft, stage, property, art | -0.042 |
| | T2 | **around**, rugby, imperative mood, about, singular, scrum, ignition, roughly, ballot, manually | 0.088 |
| rag | T1 | **ragtime**, nominative case, accusative case, rag week, terminative case, inflectional, sublative, piece of material, tag, sanitary napkin | -0.049 |
| | T2 | **clothes**, exhaustion, university, society, silk, ragged, journalism, haze, ranking, torment | 0.071 |
| record | T1 | **attainment**, track record, achievement, accomplishment, struct, number, intransitive, record book, criminal record, disc | -0.036 |
| | T2 | **evidence**, document, information, audio, recollection, storage medium, memory, electronic, sound recording, data | 0.089 |
| stab | T1 | **thread**, staccato, feeling, nominative case, sheet, chord, bacterial, culture, twinge, sensation | -0.046 |
| | T2 | **wound**, tool, knife thrust, weapon, plaster, criticism, wire, pierce, thrust, try | 0.029 |
| thump | T1 | **clunk**, throb, clump, thud, pound, thumping, rhythmic, sound, blow, hit | -0.036 |
| | T2 | **muffled**, hit, blow, sound, rhythmic, thumping, pound, thud, clump, throb | 0.033 |
| tip | T1 | **gratuity**, first person, forty, bloke, singular, overturn, stringed instrument, unbalanced, taxi driver, sated | -0.031 |
| | T2 | **brush**, tap, strike, gift, tram, flex, tumble, heap, full, hint | 0.070 |

TABLE B.5: Words annotated as changed in SemEval 2020 Task 1: Binary Subtask and retrieved farthest replacements for each time span.

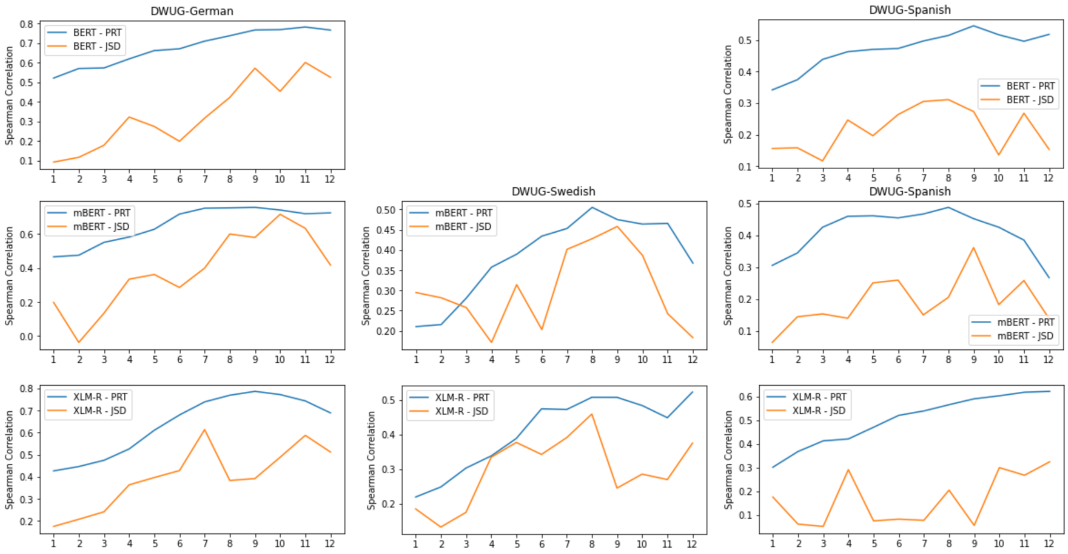


FIGURE B.4: PRT and JSD performance on the artificial LSC dataset

# Appendix C

## C.1 OP-SGNS Parameters

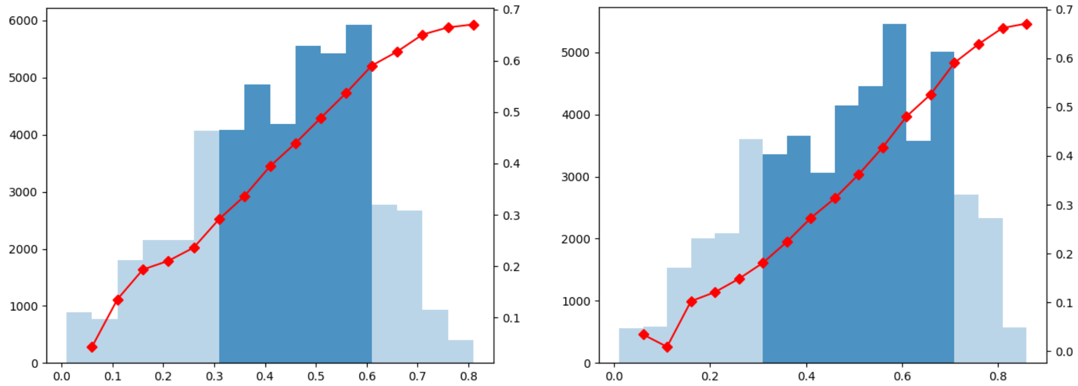
| Parameter | Value |
|---|---|
| learning rate | 0.025 |
| min. frequency | 10 |
| downsampling rate | 0.001 |
| training epochs | 5 |
| negative sampling | 5 |
| context window | 5 |
| vector dimension | 300 |

TABLE C.1: OP-SGNS Parameters for the creation of the word embeddings.

The initial learning rate is set to $0.025$, with a negative sampling of $5$ and a context window size fixed to $5$.

## C.2 Cosine similarities: Spearman Correlations

Figure C.1 shows the plots of the Spearman correlations between the two sets of ranked similarities computed over the two sub-corpora, $C_1$ and $C_2$, of "L'Unità" and "La Stampa", respectively. The cosine similarities are binned in bin of size $0.05$ in the interval $[0.0, 0.9]$. The background histogram reports the binned cosine similarity distribution for "L'Unità" (Figure (a)) and "La Stampa" (Figure (b)). The foreground red plot shows the corresponding Spearman correlation values when computed against the "La Stampa" (Figure (a)) and "L'Unità" (Figure (b)), respectively.



(A) L'Unità - La Stampa

(B) La Stampa - L'Unità

FIGURE C.1: Correlation plots.

# Bibliography

[1] Mostafa Abdou et al. "Word Order Does Matter and Shuffled Language Models Know It". In: *Proc. of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 6907–6919.
DOI: 10.18653/v1/2022.acl-long.476. URL: https://aclanthology.org/2022.acl-long.476.

[2] F. Abromeit et al. "Linking the Tower of Babel: Modelling a Massive Set of Etymological Dictionaries as RDF". In: *Proceedings of the 5th Workshop on Linked Data in Linguistics (LDL-2016): Managing, Building and Using Linked Language Resources*. 2016, pp. 11–19.

[3] Erez Lieberman Aiden and Jean-Baptiste Michel. "Culturomics: Quantitative Analysis of Culture Using Millions of Digitized Books". In: *6th Annual International Conference of the Alliance of Digital Humanities Organizations, DH*. Stanford, CA, USA: Stanford University Library, June 2011, p. 8. URL: http://xtf-prod.stanford.edu/xtf/view?docId=tei/ab-003.xml.

[4] Reem Alatrash et al. "CCOHA: Clean Corpus of Historical American English". In: *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020*. Ed. by Nicoletta Calzolari et al. Marseille, France: European Language Resources Association, May 2020, pp. 6958–6966. URL: https://www.aclweb.org/anthology/2020.lrec-1.859/.

[5] Rabab Alkhalifa et al. "QMUL-SDS @ DIACR-Ita: Evaluating Unsupervised Diachronic Lexical Semantics Classification in Italian". In: *Proceedings of the 7th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2020)*. Ed. by Valerio Basile et al. Online: CEUR.org, 2020.

[6] Jason Angel et al. "CIC-NLP @ DIACR-Ita: POS and Neighbor Based Models for Lexical Semantic Change in Diachronic Italian Corpora".

In: *Proceedings of the 7th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2020)*. Ed. by Valerio Basile et al. Online: CEUR.org, 2020.

[7] Nikolay Arefyev et al. "DeepMistake: Which Senses are Hard to Distinguish for a WordinContext Model". In: *Computational Linguistics and Intellectual Technologies - Papers from the Annual International Conference "Dialogue" 2021*. Vol. 2021-June. Section: 20. 2021.

[8] Florentina Armaselu et al. "LL(O)D and NLP perspectives on semantic change for humanities research". In: *Semantic Web* 13.6 (2022), pp. 1051–1080.
DOI: `10.3233/SW-222848`. URL: `https://doi.org/10.3233/SW-222848`.

[9] Carlos Santos Armendariz et al. "CoSimLex: A Resource for Evaluating Graded Word Similarity in Context". In: *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*. Ed. by Nicoletta Calzolari et al. European Language Resources Association, 2020, pp. 5878–5886. URL: `https://aclanthology.org/2020.lrec-1.720/`.

[10] Bing Bai et al. "Why Attentions May Not Be Interpretable?" In: *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 2021, pp. 25–34.

[11] Nikhil Bansal, Avrim Blum, and Shuchi Chawla. "Correlation Clustering". In: *Mach. Learn.* 56.1-3 (2004), pp. 89–113.
DOI: `10.1023/B:MACH.0000033116.57574.95`. URL: `https://doi.org/10.1023/B:MACH.0000033116.57574.95`.

[12] Pierpaolo Basile, Annalina Caputo, and Giovanni Semeraro. "Temporal Random Indexing: a Tool for Analysing Word Meaning Variations in News". In: *Proceedings of the First International Workshop on Recent Trends in News Information Retrieval co-located with 38th European Conference on Information Retrieval (ECIR 2016)*. Ed. by Miguel Martinez-Alvarez et al. Vol. 1568. CEUR Workshop Proceedings. Padua, Italy: CEUR-WS.org, pp. 39–41. URL: `http://ceur-ws.org/Vol-1568/paper7.pdf`.

[13] Pierpaolo Basile, Giovanni Semeraro, and Annalina Caputo. "Kronos-it: a Dataset for the Italian Semantic Change Detection Task". In: *Proceedings of the Sixth Italian Conference on Computational Linguistics*.

Ed. by Raffaella Bernardi, Roberto Navigli, and Giovanni Semeraro. Vol. 2481. CEUR Workshop Proceedings. Bari, Italy: CEUR-WS.org, Nov. 2019. URL: `http://ceur-ws.org/Vol-2481/paper3.pdf`.

[14] Pierpaolo Basile et al. "A Diachronic Italian Corpus based on "L'Unità"". In: *Proceedings of the Seventh Italian Conference on Computational Linguistics, CLiC-it 2020, Bologna, Italy, March 1-3, 2021*. Ed. by Johanna Monti, Felice Dell'Orletta, and Fabio Tamburini. Vol. 2769. CEUR Workshop Proceedings. CEUR-WS.org, 2020. URL: `http://ceur-ws.org/Vol-2769/paper\_44.pdf`.

[15] Pierpaolo Basile et al. "A New Time-sensitive Model of Linguistic Knowledge for Graph Databases". In: *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*. Ed. by Rossana Damiano et al. Vol. 3286. CEUR Workshop Proceedings. CEUR-WS.org, 2022, pp. 69–80. URL: `https://ceur-ws.org/Vol-3286/08\_paper.pdf`.

[16] Pierpaolo Basile et al. "A New Time-sensitive Model of Linguistic Knowledge for Graph Databases". In: *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*. Ed. by Rossana Damiano et al. Vol. 3286. CEUR Workshop Proceedings. CEUR-WS.org, 2022, pp. 69–80. URL: `https://ceur-ws.org/Vol-3286/08\_paper.pdf`.

[17] Pierpaolo Basile et al. "Diachronic Analysis of the Italian Language exploiting Google Ngram". In: *Proceedings of Third Italian Conference on Computational Linguistics (CLiC-it 2016) & Fifth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2016)*. Ed. by Pierpaolo Basile et al. Vol. 1749. CEUR Workshop Proceedings. Napoli, Italy: CEUR-WS.org, Dec. 2016. URL: `http://ceur-ws.org/Vol-1749/paper9.pdf`.

[18] Pierpaolo Basile et al. "DIACR-Ita @ EVALITA2020: Overview of the EVALITA2020 Diachronic Lexical Semantics (DIACR-Ita) Task". In: *Proceedings of the Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020), Online event, December 17th, 2020*. Ed. by Valerio Basile et al. Vol. 2765.

CEUR Workshop Proceedings. CEUR-WS.org, 2020. URL: http://ceur-ws.org/Vol-2765/paper158.pdf.

[19] Pierpaolo Basile et al. "The Corpora They Are a-Changing: a Case Study in Italian Newspapers". In: *Proceedings of The 2nd International Workshop on Computational Approaches to Historical Language Change 2021, LChange@ACL-IJCNLP 2021, Online, August 6, 2021*. Ed. by Nina Tahmasebi et al. Association for Computational Linguistics, 2021, pp. 14–20.
DOI: 10.18653/v1/2021.lchange-1.3. URL: https://doi.org/10.18653/v1/2021.lchange-1.3.

[20] Valerio Basile et al. "EVALITA 2020: Overview of the 7th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian". In: *Proceedings of Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020)*. Ed. by Valerio Basile et al. Online: CEUR.org, 2020.

[21] Kaspar Beelen et al. "When Time Makes Sense: A Historically-Aware Approach to Targeted Sense Disambiguation". In: *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. Online: Association for Computational Linguistics, Aug. 2021, pp. 2751–2761.
DOI: 10.18653/v1/2021.findings-acl.243. URL: https://aclanthology.org/2021.findings-acl.243.

[22] Federico Belotti, Federico Bianchi, and Matteo Palmonari. "UNIMIB @ DIACR-Ita: Aligning Distributional Embeddings with a Compass for Semantic Change Detection in the Italian Language (short paper)". In: *Proceedings of the Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020)*. Ed. by Valerio Basile et al. Vol. 2765. CEUR Workshop Proceedings. Online event: CEUR-WS.org. URL: http://ceur-ws.org/Vol-2765/paper147.pdf.

[23] Federico Belotti, Federico Bianchi, and Matteo Palmonari. "UNIMIB @ DIACR-Ita: Aligning Distributional Embeddings with a Compass for Semantic Change Detection in the Italian Language (short paper)". In: *Proceedings of the Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020)*. Ed. by Valerio Basile et al. Vol. 2765. CEUR Workshop Proceedings. Online event: CEUR-WS.org, Dec. 2020. URL: http://ceur-ws.org/Vol-2765/paper147.pdf.

[24] Wang Benyou, Emanuele Di Buccio, and Massimo Melucci. "University of Padova at DIACR-Ita". In: *Proceedings of the 7th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2020)*. Ed. by Valerio Basile et al. Online: CEUR.org, 2020.

[25] Elisa Bertino and Lorenzo Martino. "Object-oriented database management systems: concepts and issues". In: *Computer* 24.4 (1991), pp. 33–47.

[26] Erica Biagetti, Chiara Zanchi, and William Michael Short. "Toward the creation of WordNets for ancient Indo-European languages". In: *Proceedings of the 11th Global Wordnet Conference*. University of South Africa (UNISA): Global Wordnet Association, Jan. 2021, pp. 258–266. URL: `https://aclanthology.org/2021.gwc-1.30`.

[27] Yuri Bizzoni et al. "Linguistic Variation and Change in 250 Years of English Scientific Writing: A Data-Driven Approach". In: *Frontiers in Artificial Intelligence* 3 (2020).
ISSN: 2624-8212.
DOI: `10.3389/frai.2020.00073`. URL: `https://www.frontiersin.org/articles/10.3389/frai.2020.00073`.

[28] Andreas Blank. *Prinzipien des lexikalischen Bedeutungswandels am Beispiel der romanischen Sprachen*. Vol. 285. Walter de Gruyter, 2012.

[29] Andreas Blank. "Why do new meanings occur? A cognitive typology of the motivations for lexical semantic change". In: *Historical semantics and cognition* ().

[30] Leonard Bloomfield. *Language*. Motilal Banarsidass Publ., 1994.

[31] Lars Borin, Markus Forsberg, and Johan Roxendal. "Korp - the corpus infrastructure of Spräkbanken". In: *Proceedings of the Eighth International Conference on Language Resources and Evaluation*. Ed. by Nicoletta Calzolari et al. Istanbul, Turkey: European Language Resources Association (ELRA), May 2012, pp. 474–478. URL: `http://www.lrec-conf.org/proceedings/lrec2012/summaries/248.html`.

[32] Cristina Bosco et al. "The EVALITA 2014 dependency parsing task". In: *The Evalita 2014 Dependency Parsing task* (2014), pp. 1–8.

[33] Michel Bréal. *Essai de sémantique (science des significations)*. Hachette, 1904.

[34] Tom B. Brown et al. "Language Models are Few-Shot Learners". In: *Proc. of NeurIPS*. Ed. by Hugo Larochelle et al. 2020.

[35] Toby Burrows et al. "Mapping Manuscript Migrations: Digging into Data for the History and Provenance of Medieval and Renaissance Manuscripts". In: *Manuscript Studies: A Journal of the Schoenberg Institute for Manuscript Studies* 3.1 (2018), pp. 249–252.

[36] Philip Burton. "Christian Latin". In: *A companion to the Latin language*. Ed. by James Clackson. Oxford: Wiley-Blackwell, 2011, pp. 485–501.

[37] Xingyu Cai et al. "Isotropy in the contextual embedding space: Clusters and manifolds". In: *Proc. of the International Conference on Learning Representations*. 2021.

[38] Valerio Di Carlo, Federico Bianchi, and Matteo Palmonari. "Training Temporal Word Embeddings with a Compass". In: *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI*. Honolulu, Hawaii,USA: AAAI Press, Jan. 2019, pp. 6326–6334.
DOI: `10.1609/aaai.v33i01.33016326`. URL: `https://doi.org/10.1609/aaai.v33i01.33016326`.

[39] Pierluigi Cassotti et al. "A Comparative Study of Approaches for the Diachronic Analysis of the Italian Language". In: *Proceedings of the 4th Workshop on Natural Language for Artificial Intelligence (NL4AI 2020) co-located with the 19th International Conference of the Italian Association for Artificial Intelligence (AI*IA 2020), Anywhere, November 25th-27th, 2020*. Ed. by Pierpaolo Basile et al. Vol. 2735. CEUR Workshop Proceedings. CEUR-WS.org, 2020, pp. 130–140. URL: `https://ceur-ws.org/Vol-2735/paper40.pdf`.

[40] Pierluigi Cassotti et al. "Analysis of Lexical Semantic Changes in Corpora with the Diachronic Engine". In: *Proceedings of the Seventh Italian Conference on Computational Linguistics, CLiC-it 2020, Bologna, Italy, March 1-3, 2021*. Ed. by Johanna Monti, Felice Dell'Orletta, and Fabio Tamburini. Vol. 2769. CEUR Workshop Proceedings. CEUR-WS.org, 2020. URL: `https://ceur-ws.org/Vol-2769/paper\_71.pdf`.

[41] Pierluigi Cassotti et al. "Analyzing Gaussian distribution of semantic shifts in Lexical Semantic Change Models". In: *IJCoL. Italian Journal of Computational Linguistics* 6.6-2 (2020), pp. 23–36.

[42] Pierluigi Cassotti et al. "Emerging Trends in Gender-Specific Occupational Titles in Italian Newspapers". In: *Proceedings of the Eighth Italian Conference on Computational Linguistics, CLiC-it 2021, Milan, Italy, January 26-28, 2022*. Ed. by Elisabetta Fersini, Marco Passarotti, and Viviana Patti. Vol. 3033. CEUR Workshop Proceedings. CEUR-WS.org, 2021. URL: https://ceur-ws.org/Vol-3033/paper52.pdf.

[43] Pierluigi Cassotti et al. "GM-CTSC at SemEval-2020 Task 1: Gaussian Mixtures Cross Temporal Similarity Clustering". In: *Proceedings of the Fourteenth Workshop on Semantic Evaluation, SemEval@COLING2020*. Ed. by Aurélie Herbelot et al. Barcelona (online): International Committee for Computational Linguistics, Dec. 2020, pp. 74–80. URL: https://www.aclweb.org/anthology/2020.semeval-1.7/.

[44] Pierluigi Cassotti et al. "WiC-ITA at EVALITA2023: Overview of the EVALITA2023 Word-in-Context for ITAlian Task". In: *Proc. of EVALITA*. Parma, Italy: CEUR.org, Sept. 2023.

[45] Pierluigi Cassotti et al. "XL-LEXEME: WiC Pretrained Model for Cross-Lingual LEXical sEMantic changE". In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Toronto, Canada: Association for Computational Linguistics, July 2023, pp. 1577–1585.
DOI: 10.18653/v1/2023.acl-short.135. URL: https://aclanthology.org/2023.acl-short.135.

[46] Jing Chen et al. "ChiWUG: Diachronic Word Usage Graphs for Chinese". Version 1.0.0. In: (Oct. 2023).
DOI: 10.5281/zenodo.10023263. URL: https://doi.org/10.5281/zenodo.10023263.

[47] Christian Chiarcos et al. "Modelling Collocations in OntoLex-FrAC". In: *Proceedings of Globalex Workshop on Linked Lexicography within the 13th Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, June 2022, pp. 10–18. URL: https://aclanthology.org/2022.gwll-1.3.

[48] Oliver Christ et al. "The IMS Corpus Workbench: Corpus Query Processor (CQP): User's Manual". In: *University of Stuttgart* 8 (1999).

[49] Kevin Clark et al. "What Does BERT Look at? An Analysis of BERT's Attention". In: *Proc. of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Florence, Italy: ACL, Aug. 2019, pp. 276–286.

[50] Andy Coenen et al. "Visualizing and Measuring the Geometry of BERT". In: *Proc. of NeurIPS*. Red Hook, NY, USA: Curran Associates Inc., 2019.

[51] Alexis Conneau et al. "Unsupervised Cross-lingual Representation Learning at Scale". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*. Ed. by Dan Jurafsky et al. Association for Computational Linguistics, 2020, pp. 8440–8451.
DOI: 10.18653/v1/2020.acl-main.747. URL: https://doi.org/10.18653/v1/2020.acl-main.747.

[52] D. Alan Cruse. "Aspects of the Micro-Structure of Word Meanings". In: *Polysemy: Theoretical and Computational Approaches*. Ed. by Yael Ravin and Claudia Leacock. Oxford University Press, 2000, pp. 30–51.

[53] Arsène Darmesteter. *La vie des mots étudiée dans leurs significations*. C. Delagrave, 1893.

[54] Thierry Declerck et al. "Lemon: An Ontology-Lexicon model for the Multilingual Semantic Web." In: (2010).

[55] Marco Del Tredici, Malvina Nissim, and Andrea Zaninello. "Tracing metaphors in time through self-distance in vector spaces". English. In: *CEUR Workshop Proceedings*. 3rd Italian Conference on Computational Linguistics, CLiC-it 2016 and 5th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, EVALITA 2016 ; Conference date: 05-12-2016 Through 07-12-2016. 2016.

[56] Quentin Dénigot and Heather Burnett. "Dogwhistles as Identity-based interpretative variation". In: *Proceedings of the Probability and Meaning Conference (PaM 2020)*. Gothenburg: Association for Computational Linguistics, June 2020, pp. 17–25. URL: https://aclanthology.org/2020.pam-1.3.

[57] Jacob Devlin et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *CoRR* abs/1810.04805 (2018). arXiv: 1810.04805. URL: http://arxiv.org/abs/1810.04805.

[58] Jacob Devlin et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *Proc. of NAACL-HLT*. ACL, June 2019, pp. 4171–4186.

[59] Jacob Devlin et al. “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)*. Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 4171–4186.
DOI: `10.18653/v1/N19-1423`. URL: `https://www.aclweb.org/anthology/N19-1423`.

[60] Charles Du Fresne Du Cange et al. *Glossarium mediæet infimælatinitatis*. Niort: L. Favre, 1883-1887.

[61] Haim Dubossarsky et al. “Time-Out: Temporal Referencing for Robust Modeling of Lexical Semantic Change”. In: *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Volume 1: Long Papers*. Ed. by Anna Korhonen, David R. Traum, and Lluís Màrquez. Florence, Italy: Association for Computational Linguistics, pp. 457–470.
DOI: `10.18653/v1/p19-1044`. URL: `https://doi.org/10.18653/v1/p19-1044`.

[62] Bradley Efron and Robert Tibshirani. *An Introduction to the Bootstrap*. Springer, 1993.
ISBN: 978-1-4899-4541-9.
DOI: `10.1007/978-1-4899-4541-9`. URL: `https://doi.org/10.1007/978-1-4899-4541-9`.

[63] Lisa Ehrlinger and Wolfram Wöß. “Towards a Definition of Knowledge Graphs”. In: *Joint Proceedings of the Posters and Demos Track of the 12th International Conference on Semantic Systems - SEMANTiCS2016 and the 1st International Workshop on Semantic Change & Evolving Semantics (SuCCESS'16) co-located with the 12th International Conference on Semantic Systems (SEMANTiCS 2016), Leipzig, Germany, September 12-15, 2016*. Ed. by Michael Martin, Martí Cuquet, and Erwin Folmer. Vol. 1695. CEUR Workshop Proceedings. CEUR-WS.org, 2016. URL: `http://ceur-ws.org/Vol-1695/paper4.pdf`.

[64] Kawin Ethayarajh. “How Contextual are Contextualized Word Representations? Comparing the Geometry of BERT, ELMo, and GPT-2 Embeddings”. In: *Proc. of EMNLP-IJCNLP*. Hong Kong, China: ACL, Nov. 2019, pp. 55–65.

[65] Ludwig Fahrmeir et al. *Regression*. Springer, 2007.

[66] Christiane Fellbaum, ed. *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press, 1998.

[67] Christiane Fellbaum. *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press, 1998. URL: `https://mitpress.mit.edu/9780262561167/`.

[68] Stefano Ferilli. "Integration Strategy and Tool between Formal Ontology and Graph Database Technology". In: *Electronics* 10.21 (2021). ISSN: 2079-9292.
DOI: `10.3390/electronics10212616`. URL: `https://www.mdpi.com/2079-9292/10/21/2616`.

[69] Stefano Ferilli and Domenico Redavid. "The GraphBRAIN system for knowledge graph management and advanced fruition". In: *Foundations of Intelligent Systems: 25th International Symposium, ISMIS 2020, Graz, Austria, September 23–25, 2020, Proceedings*. Springer. 2020, pp. 308–317.

[70] Stefano Ferilli, Domenico Redavid, and Davide Di Pierro. "Holistic graph-based document representation and management for open science". In: *International Journal on Digital Libraries* (2022), pp. 1–23.

[71] J. R. Firth. "A synopsis of linguistic theory 1930-55." In: *Studies in linguistic analysis* 1952-59 (1957), pp. 1–32.

[72] Alexandre François. "Trees, waves and linkages". In: *The Routledge Handbook of Historical Linguistics* (2015), p. 161.

[73] Greta Franzini et al. "*Nunc Est Aestimandum*: Towards an Evaluation of the Latin WordNet". In: *Proceedings of the Sixth Italian Conference on Computational Linguistics*. Bari: Accademia University Press, Nov. 2019.
DOI: `10.5281/zenodo.3518774`. URL: `https://doi.org/10.5281/zenodo.3518774`.

[74] Nikhil Garg et al. "Word embeddings quantify 100 years of gender and ethnic stereotypes". In: *Proceedings of the National Academy of Sciences* 115.16 (2018), E3635–E3644.

[75] Aina Garí Soler and Marianna Apidianaki. "Let's Play Mono-Poly: BERT Can Reveal Words' Polysemy Level and Partitionability into Senses". In: *Transactions of the Association for Computational Linguistics* 9 (2021), pp. 825–844.

DOI: 10.1162/tacl_a_00400. URL: https://aclanthology.org/2021.tacl-1.50.

[76] Dirk Geeraerts, Caroline Gevaert, and Dirk Speelman. "Current methods in historical semantics". In: *Current methods in historical semantics* (2012). Ed. by Kathryn Allan and Justyna Robinson, pp. 73–109.

[77] F. Ginter and J. Kanerva. "Fast Training of word 2 vec Representations Using N-gram Corpora". In: (2014). URL: https://www2.lingfil.uu.se/SLTC2014/abstracts/sltc2014_submission_27.pdf.

[78] Mario Giulianelli et al. "Interpretable Word Sense Representations via Definition Generation: The Case of Semantic Change Analysis". In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki. Toronto, Canada: Association for Computational Linguistics, July 2023, pp. 3130–3148. DOI: 10.18653/v1/2023.acl-long.176. URL: https://aclanthology.org/2023.acl-long.176.

[79] Yoav Goldberg and Jon Orwant. "A Dataset of Syntactic-Ngrams over Time from a Very Large Corpus of English Books". In: *Atlanta, Georgia, USA* (2013), p. 241.

[80] Hila Gonen et al. "Simple, Interpretable and Stable Method for Detecting Words with Usage Change across Corpora". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, July 2020, pp. 538–555. DOI: 10.18653/v1/2020.acl-main.51. URL: https://www.aclweb.org/anthology/2020.acl-main.51.

[81] Yue Guan et al. "How Far Does BERT Look At: Distance-based Clustering and Analysis of BERT's Attention". In: *Proceedings of the 28th International Conference on Computational Linguistics*. Barcelona, Spain (Online): International Committee on Computational Linguistics, Dec. 2020, pp. 3853–3860. DOI: 10.18653/v1/2020.coling-main.342. URL: https://aclanthology.org/2020.coling-main.342.

[82] Pascal Mark Gygax et al. “A Language Index of Grammatical Gender Dimensions to Study the Impact of Grammatical Gender on the Way We Perceive Women and Men”. In: *Frontiers in Psychology* 10 (2019), p. 1604.
ISSN: 1664-1078.
DOI: 10.3389/fpsyg.2019.01604. URL: https://www.frontiersin.org/article/10.3389/fpsyg.2019.01604.

[83] Janosch Haber and Massimo Poesio. “Patterns of Polysemy and Homonymy in Contextualised Language Models”. In: *Findings of the Association for Computational Linguistics: EMNLP 2021*. Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 2663–2676.
DOI: 10.18653/v1/2021.findings-emnlp.226. URL: https://aclanthology.org/2021.findings-emnlp.226.

[84] Raia Hadsell, Sumit Chopra, and Yann LeCun. “Dimensionality Reduction by Learning an Invariant Mapping”. In: *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2006), 17-22 June 2006, New York, NY, USA*. IEEE Computer Society, 2006, pp. 1735–1742.
DOI: 10.1109/CVPR.2006.100. URL: https://doi.org/10.1109/CVPR.2006.100.

[85] Thomas Haider and Steffen Eger. “Semantic Change and Emerging Tropes In a Large Corpus of New High German Poetry”. In: *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*. Florence, Italy: Association for Computational Linguistics, Aug. 2019, pp. 216–222.
DOI: 10.18653/v1/W19-4727. URL: https://aclanthology.org/W19-4727.

[86] William L. Hamilton, Jure Leskovec, and Dan Jurafsky. “Diachronic word embeddings reveal statistical laws of semantic change”. In: *54th Annual Meeting of the Association for Computational Linguistics, ACL 2016 - Long Papers*. Vol. 3. May 2016, pp. 1489–1501.
ISBN: 9781510827585.
DOI: 10.18653/v1/p16-1141. arXiv: 1605.09096. URL: http://arxiv.org/abs/1605.09096.

[87] Michael Hanna and David Mareček. "Analyzing BERT's Knowledge of Hypernymy via Prompting". In: *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*. Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 275–282.
DOI: `10.18653/v1/2021.blackboxnlp-1.20`. URL: `https://aclanthology.org/2021.blackboxnlp-1.20`.

[88] David R. Hardoon, Sandor Szedmak, and John Shawe-Taylor. "Canonical Correlation Analysis: An Overview with Application to Learning Methods". In: *Neural Computation* 16.12 (2004), pp. 2639–2664.
DOI: `10.1162/0899766042321814`.

[89] J. Herman. *Vulgar Latin. Translated by Roger Wright*. The Pennsylvania State University, 2000.

[90] Jack Hessel and Alexandra Schofield. "How effective is BERT without word ordering? Implications for language understanding and data privacy". In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. Online: Association for Computational Linguistics, Aug. 2021, pp. 204–211.
DOI: `10.18653/v1/2021.acl-short.27`. URL: `https://aclanthology.org/2021.acl-short.27`.

[91] John Hewitt and Percy Liang. "Designing and Interpreting Probes with Control Tasks". In: *Proc. of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*. Ed. by Kentaro Inui et al. Association for Computational Linguistics, 2019, pp. 2733–2743.
DOI: `10.18653/v1/D19-1275`. URL: `https://doi.org/10.18653/v1/D19-1275`.

[92] Sepp Hochreiter and Jürgen Schmidhuber. "Long Short-Term Memory". In: *Neural Computation* 9.8 (1997), pp. 1735–1780.

[93] Willem Hollmann. "Semantic change". In: *English Language: Description, Variation and Context*. Basingstoke: Palgrave, 2009, pp. 301–313.

[94] Paul J. Hopper. “On some principles of grammaticalization”. In: *Approaches to grammaticalization*. Ed. by Elizabeth Closs Traugott and Bernd Heine. Amsterdam, Philadelphia: John Benjamins Publishing, 1991, pp. 17–35.

[95] Renfen Hu, Shen Li, and Shichen Liang. “Diachronic Sense Modeling with Deep Contextualized Word Embeddings: An Ecological View”. In: *57th Annual Meeting of the Association for Computational Linguistics*. 2019, pp. 3899–3908.
DOI: `10.18653/v1/p19-1379`. URL: `https://en.oxforddictionaries.com/`.

[96] Tao Huang, Heng Peng, and Kun Zhang. “MODEL SELECTION FOR GAUSSIAN MIXTURE MODELS”. In: *Statistica Sinica* 27.1 (2017), pp. 147–169.
ISSN: 10170405, 19968507. URL: `http://www.jstor.org/stable/44114365`.

[97] Leif Isaksen et al. “Pelagios and the emerging graph of ancient world data”. In: *Proceedings of the 2014 ACM conference on Web science*. 2014, pp. 197–201.

[98] Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. “What Does BERT Learn about the Structure of Language?” In: *Proc. of ACL*. Florence, Italy: ACL, July 2019, pp. 3651–3657.
DOI: `10.18653/v1/P19-1356`.

[99] Armand Joulin et al. “Bag of Tricks for Efficient Text Classification”. In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017, Volume 2: Short Papers*. Ed. by Mirella Lapata, Phil Blunsom, and Alexander Koller. Valencia, Spain: Association for Computational Linguistics, Apr. 2017, pp. 427–431.
DOI: `10.18653/v1/e17-2068`. URL: `https://doi.org/10.18653/v1/e17-2068`.

[100] Daniel Jurafsky and James H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. 1st. USA: Prentice Hall PTR, 2000.
ISBN: 0130950696.

[101] Jens Kaiser, Dominik Schlechtweg, and Sabine Schulte im Walde. "OP-IMS @ DIACR-Ita: Back to the Roots: SGNS+OP+CD still Rocks Semantic Change Detection". In: *Proceedings of the Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020)*. Ed. by Valerio Basile et al. Vol. 2765. CEUR Workshop Proceedings. Online event: CEUR-WS.org, Dec. 2020. URL: `http://ceur-ws.org/Vol-2765/paper133.pdf`.

[102] Jens Kaiser, Dominik Schlechtweg, and Sabine Schulte im Walde. "OP-IMS @ DIACR-Ita: Back to the Roots: SGNS+OP+CD still Rocks Semantic Change Detection". In: *Proceedings of the Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020)*. Ed. by Valerio Basile et al. Vol. 2765. CEUR Workshop Proceedings. Online event: CEUR-WS.org. URL: `http://ceur-ws.org/Vol-2765/paper133.pdf`.

[103] Alexander Kalinowski and Yuan An. "Exploring Sentence Embedding Structures for Semantic Relation Extraction". In: *2021 International Joint Conference on Neural Networks (IJCNN)*. IEEE. 2021, pp. 1–7.

[104] Anas Fahad Khan. "Towards the Representation of Etymological Data on the Semantic Web". In: *Information* 9.12 (2018). Publisher: MDPI AG, p. 304.
ISSN: 2078-2489.
DOI: `10.3390/info9120304`. URL: `http://dx.doi.org/10.3390/info9120304`.

[105] Anas Fahad Khan et al. "Some Considerations in the Construction of a Historical Language WordNet". In: (2023).

[106] Anas Fahad Khan et al. "When linguistics meets web technologies. Recent advances in modelling linguistic linked data". In: *Semantic Web* (2022), pp. 1–64.
DOI: `10.3233/SW-222859`.

[107] Fahad Khan. "Representing Temporal Information in Lexical Linked Data Resources". English. In: *Proceedings of the 7th Workshop on Linked Data in Linguistics (LDL-2020)*. Marseille, France: European Language Resources Association, May 2020, pp. 15–22.
ISBN: 979-10-95546-36-8. URL: `https://aclanthology.org/2020.ldl-1.3`.

[108] Yoon Kim et al. “Temporal Analysis of Language through Neural Language Models”. In: *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*. Baltimore, MD, USA: Association for Computational Linguistics, June 2014, pp. 61–65. DOI: 10.3115/v1/W14-2517. URL: https://www.aclweb.org/anthology/W14-2517.

[109] Ruth Elizabeth King. *Talking gender: A guide to nonsexist communication*. Copp Clark Professional, 1991.

[110] PN Kokic and PA Bell. “Optimal winsorizing cutoffs for a stratified finite population estimator”. In: *Journal of Official Statistics* 10.4 (1994), p. 419.

[111] Hans-Peter Kriegel et al. “The paradigm of relational indexing: A survey”. In: *BTW 2003–Datenbanksysteme für Business, Technologie und Web, Tagungsband der 10. BTW Konferenz*. Gesellschaft für Informatik eV. 2003.

[112] Taku Kudo and John Richardson. “SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing”. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Ed. by Eduardo Blanco and Wei Lu. Brussels, Belgium: Association for Computational Linguistics, Nov. 2018, pp. 66–71. DOI: 10.18653/v1/D18-2012. URL: https://aclanthology.org/D18-2012.

[113] Andrey Kutuzov and Mario Giulianelli. “UiO-UvA at SemEval-2020 Task 1: Contextualised Embeddings for Lexical Semantic Change Detection”. In: *Proceedings of the Fourteenth Workshop on Semantic Evaluation*. Barcelona (online): International Committee for Computational Linguistics, Dec. 2020, pp. 126–134. DOI: 10.18653/v1/2020.semeval-1.14. URL: https://aclanthology.org/2020.semeval-1.14.

[114] Andrey Kutuzov and Lidia Pivovarova. “RuShiftEval public data”. Version 3656c1f. In: (2021). URL: https://github.com/akutuzov/rushifteval_public.

[115] Andrey Kutuzov, Lidia Pivovarova, and others. “RuShiftEval: a shared task on semantic shift detection for Russian”. In: *Computational linguistics and intellectual technologies: Papers from the annual conference Dialogue.* Redkollegija sbornika.

[116] Andrey Kutuzov, Erik Velldal, and Lilja Øvrelid. "Contextualized embeddings for semantic change detection: Lessons learned". In: *Northern European Journal of Language Technology, Volume 8*. 2022.

[117] Andrey Kutuzov, Erik Velldal, and Lilja Øvrelid. "Tracing armed conflicts with diachronic word embedding models". In: *Proceedings of the Events and Stories in the News Workshop@ACL 2017*. Ed. by Tommaso Caselli et al. Vancouver, Canada: Association for Computational Linguistics, Aug. 2017, pp. 31–36. DOI: `10.18653/v1/w17-2705`. URL: `https://doi.org/10.18653/v1/w17-2705`.

[118] Andrey Kutuzov et al. "Diachronic word embeddings and semantic shifts: a survey". In: *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018*. Ed. by Emily M. Bender, Leon Derczynski, and Pierre Isabelle. Santa Fe, New Mexico, USA: Association for Computational Linguistics, Aug. 2018, pp. 1384–1397. URL: `https://www.aclweb.org/anthology/C18-1117/`.

[119] Andrey Kutuzov et al. "NorDiaChange". Version 9cdd80f. In: (2021). URL: `https://github.com/ltgoslo/nor_dia_change`.

[120] Andrey Kutuzov et al. "NorDiaChange: Diachronic Semantic Change Dataset for Norwegian". In: *Proceedings of the Thirteenth Language Resources and Evaluation Conference, LREC 2022, Marseille, France, 20-25 June 2022*. Ed. by Nicoletta Calzolari et al. European Language Resources Association, 2022, pp. 2563–2572. URL: `https://aclanthology.org/2022.lrec-1.274`.

[121] Caterina Lacerra et al. "ALaSca: an Automated approach for Large-Scale Lexical Substitution". In: *Proceedings of the 30th International Joint Conference on Artificial Intelligence*. International Joint Conferences on Artificial Intelligence, 2021. DOI: `10.24963/ijcai.2021/528`.

[122] Severin Laicher et al. "CL-IMS @ DIACR-Ita: Volente o Nolente: BERT does not outperform SGNS on Semantic Change Detection". In: *Proceedings of the 7th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2020)*. Ed. by Valerio Basile et al. Online: CEUR.org, 2020.

[123] Severin Laicher et al. “Explaining and Improving BERT Performance on Lexical Semantic Change Detection”. In: *Proc. of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop, EACL 2021, Online, April 19-23, 2021*. Ed. by Ionut-Teodor Sorodoc et al. Association for Computational Linguistics, 2021, pp. 192–202.
DOI: `10.18653/v1/2021.eacl-srw.25`. URL: `https://doi.org/10.18653/v1/2021.eacl-srw.25`.

[124] Tarutuulia Laine and Greg Watson. “Linguistic sexism in The Times-A diachronic study”. In: *International Journal of English Linguistics* 4.3 (2014), p. 1.

[125] Zhenzhong Lan et al. “ALBERT: A Lite BERT for Self-supervised Learning of Language Representations”. In: *Proc. of ICL*. OpenReview.net, 2020.

[126] Thomas K Landauer, Peter W Foltz, and Darrell Laham. “An introduction to latent semantic analysis”. In: *Discourse Processes* 25.2 (1998), pp. 259–284.

[127] Roger Lass. “Phonology and morphology”. In: *The Cambridge history of the English language* 2 (1992), pp. 1066–1476.

[128] Bandy X Lee et al. “Transforming our world: implementing the 2030 agenda through sustainable development goal indicators”. In: *Journal of public health policy* 37.1 (2016), pp. 13–31.

[129] Yoav Levine et al. “SenseBERT: Driving Some Sense into BERT”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, July 2020, pp. 4656–4667.
DOI: `10.18653/v1/2020.acl-main.423`. URL: `https://aclanthology.org/2020.acl-main.423`.

[130] Omer Levy and Yoav Goldberg. “Neural Word Embedding as Implicit Matrix Factorization”. In: *Proc of NeurIPS*. Ed. by Z. Ghahramani et al. Vol. 27. Curran Associates, Inc., 2014.

[131] Charlton T. Lewis. *An Elementary Latin Dictionary*. New York, Cincinnati, and Chicago: American Book Company, 1890.

[132] Charlton T. Lewis and Charles Short. *A Latin Dictionary, Founded on Andrews' edition of Freund's Latin dictionary revised, enlarged, and in great part rewritten by Charlton T. Lewis, Ph.D. and Charles Short*. Oxford: Clarendon Press, 1879.

[133] Mike Lewis et al. "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension". In: *Proc. of ACL*. Ed. by Dan Jurafsky et al. Association for Computational Linguistics, 2020, pp. 7871–7880.

[134] Jen-Yi Lin et al. "The Structure of Polysemy : A Study of Multi-sense Words Based on WordNet". In: *Proceedings of the 16th Pacific Asia Conference on Language, Information and Computation*. Jeju, Korea: The Korean Society for Language and Information, Jan. 2001, pp. 320–329.
DOI: `http://hdl.handle.net/2065/12239`. URL: `https://aclanthology.org/Y02-1031`.

[135] Ruixi Lin and Hwee Tou Ng. "Does BERT Know that the IS-A Relation Is Transitive?" In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 94–99.
DOI: `10.18653/v1/2022.acl-short.11`. URL: `https://aclanthology.org/2022.acl-short.11`.

[136] Pengfei Liu et al. "Pre-Train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing". In: *ACM Computing Surveys* 55.9 (Jan. 2023).
ISSN: 0360-0300.

[137] Qianchu Liu et al. "AM2iCo: Evaluating Word Meaning in Context across Low-Resource Languages with Adversarial Examples". In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*. Ed. by Marie-Francine Moens et al. Association for Computational Linguistics, 2021, pp. 7151–7162.
DOI: `10.18653/v1/2021.emnlp-main.571`. URL: `https://doi.org/10.18653/v1/2021.emnlp-main.571`.

[138] Yinhan Liu et al. "RoBERTa: A Robustly Optimized BERT Pretraining Approach". In: *arxiv* abs/1907.11692 (2019).

[139] Isabelle Lorge and Janet Pierrehumbert. *Not wacky vs. definitely wacky: A study of scalar adverbs in pretrained language models*. 2023. arXiv: 2305.16426 [cs.CL].

[140] Verena Lyding et al. "The paisa'corpus of italian web texts". In: *9th Web as Corpus Workshop (WaC-9)@ EACL 2014*. EACL (European chapter of the Association for Computational Linguistics). 2014, pp. 36–43.

[141] Enrique Manjavacas and Lauren Fonteyn. "Adapting vs. Pre-training Language Models for Historical Languages". In: *Journal of Data Mining and Digital Humanities* NLP4DH (June 2022). DOI: 10.46298/jdmdh.9152. URL: https://hal.inria.fr/hal-03592137.

[142] Gianna Marcato and Eva-Maria Thüne. "Gender and female visibility in Italian". In: *Gender across languages: The linguistic representation of women and men* 2 (2002), pp. 187–217.

[143] Jani Marjanen et al. "Clustering Ideological Terms in Historical Newspaper Data with Diachronic Word Embeddings". In: *5th International Workshop on Computational History, HistoInformatics@TPDL 2019, Oslo, Norway, September 12, 2019*. Ed. by Melvin Wevers et al. Vol. 2461. CEUR Workshop Proceedings. CEUR-WS.org, 2019, pp. 21–29. URL: https://ceur-ws.org/Vol-2461/paper_4.pdf.

[144] Federico Martelli et al. "SemEval-2021 Task 2: Multilingual and Cross-lingual Word-in-Context Disambiguation (MCL-WiC)". In: *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*. Online: Association for Computational Linguistics, Aug. 2021, pp. 24–36. DOI: 10.18653/v1/2021.semeval-1.3. URL: https://aclanthology.org/2021.semeval-1.3.

[145] Matej Martinc, Petra Kralj Novak, and Senja Pollak. "Leveraging Contextual Embeddings for Detecting Diachronic Semantic Shift". English. In: *Proceedings of the Twelfth Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, May 2020, pp. 4811–4819. ISBN: 979-10-95546-34-4. URL: https://aclanthology.org/2020.lrec-1.592.

[146] Matej Martinc et al. "Capturing Evolution in Word Usage: Just Add More Clusters?" In: *Proc. of the Web Conference 2020*. New York, NY, USA: Association for Computing Machinery (ACM), 2020, 343–349.

[147] John P. McCrae, Dennis Spohr, and Philipp Cimiano. "Linking Lexical Resources and Ontologies on the Semantic Web with Lemon". In: *The Semantic Web: Research and Applications - 8th Extended Semantic Web Conference, ESWC 2011, Heraklion, Crete, Greece, May 29-June 2, 2011, Proceedings, Part I*. Ed. by Grigoris Antoniou et al. Vol. 6643. Lecture Notes in Computer Science. Springer, 2011, pp. 245–259. DOI: `10.1007/978-3-642-21034-1\_17`. URL: `https://doi.org/10.1007/978-3-642-21034-1\_17`.

[148] Barbara McGillivray. *Dataset: Latin lexical semantic annotation*. Figshare. DOI: https://doi.org/10.18742/16974823.v1. 2021.

[149] Barbara McGillivray and Gard B. Jenset. "Quantifying the quantitative (re-)turn in historical linguistics". In: *Humanities and Social Sciences Communications* 10.37 (2023). DOI: `https://doi.org/10.1057/s41599-023-01531-2`.

[150] Barbara McGillivray and Adam Kilgarriff. "Tools for historical corpus research, and a corpus of Latin". In: *New Methods in Historical Corpus Linguistics*. Ed. by Paul Bennett et al. Tübingen: Narr, 2013, pp. 247–257.

[151] Barbara McGillivray et al. "A new corpus annotation framework for Latin diachronic lexical semantics". English. In: *Journal of Latin Linguistics* 21.1 (July 2022), pp. 47–105. DOI: `https://doi.org/10.1515/joll-2022-2007`.

[152] Barbara McGillivray et al. "DWUG LA: Diachronic Word Usage Graphs for Latin". In: (Aug. 2021). DOI: `10.5281/zenodo.5255228`. URL: `https://doi.org/10.5281/zenodo.5255228`.

[153] Barbara McGillivray et al. "Graph Databases for Diachronic Language Data Modelling". In: *Proceedings of Language, Data and Knowledge 2023 (LDK 2023)*. 2023.

[154] Barbara McGillivray et al. "Using Graph Databases for Historical Language Data: Challenges and Opportunities". In: *Proceedings of the 19th Italian Research Conference on Digital Libraries, Bari, Italy, February 23-24, 2023*. Ed. by Alessia Bardi et al. CEUR Workshop Proceedings. CEUR-WS.org, 2023.

[155] Jean-Baptiste Michel et al. "Quantitative analysis of culture using millions of digitized books". In: *science* 331.6014 (2011), pp. 176–182.

[156] Timothee Mickus et al. “What do you mean, BERT?” In: *Proceedings of the Society for Computation in Linguistics 2020*. New York, New York: Association for Computational Linguistics, Jan. 2020, pp. 279–290. URL: https://aclanthology.org/2020.scil-1.35.

[157] Tomás Mikolov et al. “Efficient Estimation of Word Representations in Vector Space”. In: *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*. Ed. by Yoshua Bengio and Yann LeCun. 2013. URL: http://arxiv.org/abs/1301.3781.

[158] George A. Miller. “WORDNET: a Lexical Database for English”. In: *Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, USA, February 23-26, 1992*. Morgan Kaufmann, 1992. URL: https://aclanthology.org/H92-1116/.

[159] George A. Miller et al. “A Semantic Concordance”. In: *Human Language Technology: Proc. of a Workshop Held at Plainsboro, New Jersey, USA, March 21-24, 1993*. Morgan Kaufmann, 1993. URL: https://aclanthology.org/H93-1061/.

[160] Stefano Minozzi. “Latin WordNet, una rete di conoscenza semantica per il latino e alcune ipotesi di utilizzo nel campo dell’Information Retrieval”. In: *Strumenti digitali e collaborativi per le Scienze dell’Antichita*. Ed. by Paolo Mastandrea. Venezia: Università Ca’ Foscari, 2017, pp. 123–134.

[161] Stefano Montanelli and Francesco Periti. “A Survey on Contextualised Semantic Shift Detection”. In: *arXiv preprint arXiv:2304.01666* (2023).

[162] R. Navigli and S. Ponzetto. “BabelNet: The Automatic Construction, Evaluation and Application of a Wide-Coverage Multilingual Semantic Network”. In: *Artificial Intelligence* 193 (2012), pp. 217–250.

[163] Roberto Navigli. “Word Sense Disambiguation: A Survey”. In: *ACM Comput. Surv.* 41.2 (Feb. 2009).
ISSN: 0360-0300.
DOI: 10.1145/1459352.1459355. URL: https://doi.org/10.1145/1459352.1459355.

[164] Debora Nozza. “Exposing the limits of Zero-shot Cross-lingual Hate Speech Detection”. In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint*

*Conference on Natural Language Processing (Volume 2: Short Papers)*. Online: Association for Computational Linguistics, Aug. 2021, pp. 907–914.
DOI: 10.18653/v1/2021.acl-short.114. URL: https://aclanthology.org/2021.acl-short.114.

[165] Robert Parker et al. "English Gigaword fifth edition, 2011". In: *Linguistic Data Consortium, Philadelphia, PA, USA* (2011).

[166] Marco Passarotti et al. "Interlinking through Lemmas. The Lexical Collection of the LiLa Knowledge Base of Linguistic Resources for Latin". In: *Studi e Saggi Linguistici* 58 (1 2020). Ed. by Marco Passarotti.
DOI: 10.4454/ssl.v58i1.277.

[167] Hermann Paul. *Prinzipien der sprachgeschichte*. Vol. 6. Walter de Gruyter, 2010.

[168] Paolo Pedinotti and Alessandro Lenci. "Don't Invite BERT to Drink a Bottle: Modeling the Interpretation of Metonymies Using BERT and Distributional Representations". In: *Proceedings of the 28th International Conference on Computational Linguistics*. Barcelona, Spain (Online): International Committee on Computational Linguistics, Dec. 2020, pp. 6831–6837.
DOI: 10.18653/v1/2020.coling-main.602. URL: https://aclanthology.org/2020.coling-main.602.

[169] Francesco Periti and Haim Dubossarsky. "The Time-Embedding Travelers at WiC-ITA". In: *Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), Parma, Italy, September 7th-8th, 2023*. Ed. by Mirko Lai et al. Vol. 3473. CEUR Workshop Proceedings. CEUR-WS.org, 2023.
URL: https://ceur-ws.org/Vol-3473/paper47.pdf.

[170] Francesco Periti et al. "What is Done is Done: an Incremental Approach to Semantic Shift Detection". In: *Proc. of LChange*. Dublin, Ireland: ACL, May 2022, pp. 33–43.

[171] Emanuele Pianta, Luisa Bentivogli, and Christian Girardi. "MultiWordNet: developing an aligned multilingual database". In: *First international conference on global WordNet*. 2002, pp. 293–302.

[172] Mohammad Taher Pilehvar and José Camacho-Collados. "WiC: the Word-in-Context Dataset for Evaluating Context-Sensitive Meaning Representations". In: *Proceedings of the 2019 Conference of the North*

*American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*. Ed. by Jill Burstein, Christy Doran, and Thamar Solorio. Association for Computational Linguistics, 2019, pp. 1267–1273.
DOI: 10.18653/v1/n19-1128. URL: https://doi.org/10.18653/v1/n19-1128.

[173] Harm Pinkster. *Sintassi e semantica latina*. Rosenberg & Sellier, 1991.

[174] Ondrej Prazák, Pavel Pribán, and Stephen Taylor. "UWB @ DIACR-Ita: Lexical Semantic Change Detection with CCA and Orthogonal Transformation". In: *Proceedings of the Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020)*. Ed. by Valerio Basile et al. Vol. 2765. CEUR Workshop Proceedings. Online event: CEUR-WS.org, Dec. 2020. URL: http://ceur-ws.org/Vol-2765/paper110.pdf.

[175] Ondrej Prazák et al. "UWB at SemEval-2020 Task 1: Lexical Semantic Change Detection". In: *Proceedings of the Fourteenth Workshop on Semantic Evaluation, SemEval@COLING2020*. Ed. by Aurélie Herbelot et al. Barcelona (online): International Committee for Computational Linguistics, Dec. 2020, pp. 246–254. URL: https://www.aclweb.org/anthology/2020.semeval-1.30/.

[176] William H. Press. *Numerical recipes in C++: the art of scientific computing, 2nd Edition (C++ ed., print. is corrected to software version 2.10)*. Cambridge University Press, 2002.
ISBN: 0521750334. URL: https://www.worldcat.org/oclc/48241370.

[177] Behrang QasemiZadeh and Laura Kallmeyer. "Random positive-only projections: PPMI-enabled incremental semantic space construction". In: **SEM 2016 - 5th Joint Conference on Lexical and Computational Semantics, Proceedings*. 2016, pp. 189–198.
ISBN: 9781941643921.
DOI: 10.18653/v1/s16-2024. URL: https://www.aclweb.org/anthology/S16-2024/.

[178] Maxim Rachinskiy and Nikolay Arefyev. "Zeroshot Crosslingual Transfer of a Gloss Language Model for Semantic Change Detection". In: *Computational Linguistics and Intellectual Technologies - Papers from*

*the Annual International Conference "Dialogue" 2021*. Vol. 2021-June. Section: 20. 2021.

[179] Colin Raffel et al. “Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer”. In: *J. Mach. Learn. Res.* 21 (2020), 140:1–140:67.

[180] Colin Raffel et al. “Exploring the limits of transfer learning with a unified text-to-text transformer”. In: *The Journal of Machine Learning Research* 21.1 (2020), pp. 5485–5551.

[181] Alessandro Raganato et al. “XL-WiC: A Multilingual Benchmark for Evaluating Semantic Contextualization”. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*. Ed. by Bonnie Webber et al. Association for Computational Linguistics, 2020, pp. 7193–7206. DOI: `10.18653/v1/2020.emnlp-main.584`. URL: `https://doi.org/10.18653/v1/2020.emnlp-main.584`.

[182] Abhilasha Ravichander et al. “On the Systematicity of Probing Contextualized Word Representations: The Case of Hypernymy in BERT”. In: *Proceedings of the Ninth Joint Conference on Lexical and Computational Semantics*. Barcelona, Spain (Online): Association for Computational Linguistics, Dec. 2020, pp. 88–102. URL: `https://aclanthology.org/2020.starsem-1.10`.

[183] Nils Reimers and Iryna Gurevych. “Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks”. 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)*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 3982–3992. DOI: `10.18653/v1/D19-1410`. URL: `https://aclanthology.org/D19-1410`.

[184] Karl Christian Reisig. *Professor K. Reisig's Vorlesungen über lateinische Sprachwissenschaft*. Verlag der Lehnhold'schen Buchhandlung, 1839.

[185] Don Ringe and Joseph F Eska. *Historical linguistics: Toward a twenty-first century reintegration*. Cambridge University Press, 2013.

[186] Ian Robinson, Jim Webber, and Emil Eifrem. *Graph Databases. New Opportunities for Connected Data*. 2nd. O'Reilly Media, 2015.

[187] Julia Rodina and Andrey Kutuzov. “RuSemShift: a dataset of historical lexical semantic change in Russian”. In: *Proceedings of the 28th International Conference on Computational Linguistics*. Barcelona, Spain (Online): International Committee on Computational Linguistics, Dec. 2020, pp. 1037–1047.
DOI: 10.18653/v1/2020.coling-main.90. URL: https://aclanthology.org/2020.coling-main.90.

[188] M.A. Rodriguez and P. Neubauer. “Constructions from dots and lines”. In: *Bul. Am. Soc.Info. Sci. Tech.* 36 (6 2010), 35–41.

[189] P. Roelli. *Latin as the Language of Science and Learning*. De Gruyter, 2021.

[190] Guy D. Rosin, Ido Guy, and Kira Radinsky. “Time Masking for Temporal Language Models”. In: *WSDM ’22: The Fifteenth ACM International Conference on Web Search and Data Mining, Virtual Event / Tempe, AZ, USA, February 21 - 25, 2022*. Ed. by K. Selcuk Candan et al. ACM, 2022, pp. 833–841.
DOI: 10.1145/3488560.3498529. URL: https://doi.org/10.1145/3488560.3498529.

[191] Guy D. Rosin and Kira Radinsky. “Temporal Attention for Language Models”. In: *Findings of the Association for Computational Linguistics (NAACL 2022)*. Seattle, United States: Association for Computational Linguistics (ACL), July 2022, pp. 1498–1508.

[192] Maja R. Rudolph and David M. Blei. “Dynamic Embeddings for Language Evolution”. In: *Proceedings of the 2018 World Wide Web Conference on World Wide Web, WWW 2018*. Ed. by Pierre-Antoine Champin et al. Lyon, France: ACM, Apr. 2018, pp. 1003–1011.
DOI: 10.1145/3178876.3185999. URL: https://doi.org/10.1145/3178876.3185999.

[193] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. “Learning representations by back-propagating errors”. In: *nature* 323.6088 (1986), pp. 533–536.

[194] Pavel Rychlỳ. “A Lexicographer-Friendly Association Score”. In: *RASLAN 2008 Recent Advances in Slavonic Natural Language Processing* (2008), p. 6.

[195] Alma Sabatini. “Occupational titles in Italian: Changing the sexist usage”. In: *Sprachwandel und feministische Sprachpolitik: Internationale Perspektiven*. Springer, 1985, pp. 64–75.

[196] Magnus Sahlgren. “An introduction to random indexing”. In: *Proceedings of the Methods and Applications of Semantic Indexing Workshop at the 7th International conference on Terminology and Knowledge Engineering, TKE 2005*. Copenhagen, Denmark, Aug. 2005.

[197] Flora Sakketou et al. “Investigating User Radicalization: A Novel Dataset for Identifying Fine-Grained Temporal Shifts in Opinion”. In: *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, June 2022, pp. 3798–3808. URL: `https://aclanthology.org/2022.lrec-1.405`.

[198] Victor Sanh et al. “DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter”. In: *arxiv* (2019).

[199] Sascha Schimke, Claus Vielhauer, and Jana Dittmann. “Using adapted levenshtein distance for on-line signature authentication”. In: *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.* Vol. 2. IEEE. 2004, pp. 931–934.

[200] Dominik Schlechtweg, Sabine Schulte im Walde, and Stefanie Eckmann. “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, NAACL-HLT, Volume 2 (Short Papers)*. Ed. by Marilyn A. Walker, Heng Ji, and Amanda Stent. New Orleans, Louisiana, USA: Association for Computational Linguistics, June 2018, pp. 169–174.
DOI: `10.18653/v1/n18-2027`. URL: `https://doi.org/10.18653/v1/n18-2027`.

[201] Dominik Schlechtweg et al. “A Wind of Change: Detecting and Evaluating Lexical Semantic Change across Times and Domains”. In: *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Volume 1: Long Papers*. Ed. by Anna Korhonen, David R. Traum, and Lluís Màrquez. Florence, Italy: Association for Computational Linguistics, July 2019, pp. 732–746.
DOI: `10.18653/v1/p19-1072`. URL: `https://doi.org/10.18653/v1/p19-1072`.

[202] Dominik Schlechtweg et al. “DWUG: A large Resource of Diachronic Word Usage Graphs in Four Languages”. In: *Annual Conference of the North American Chapter of the Association for Computational Linguistics,*

*(NAACL-HLT 2021)*. Mexico City, Mexico: Association for Computational Linguistics, 2021.

[203] Dominik Schlechtweg et al. "DWUG DE: Diachronic Word Usage Graphs for German". Version 2.3.0. In: (Dec. 2022). DOI: `10.5281/zenodo.7441645`. URL: `https://doi.org/10.5281/zenodo.7441645`.

[204] Dominik Schlechtweg et al. "DWUG EN: Diachronic Word Usage Graphs for English". Version 2.0.1. In: (Dec. 2022). DOI: `10.5281/zenodo.7387261`. URL: `https://doi.org/10.5281/zenodo.7387261`.

[205] Dominik Schlechtweg et al. "SemEval-2020 Task 1: Unsupervised Lexical Semantic Change Detection". In: *Proceedings of the Fourteenth Workshop on Semantic Evaluation, SemEval@COLING2020*. Ed. by Aurélie Herbelot et al. Barcelona (online): International Committee for Computational Linguistics, Dec. 2020, pp. 1–23. URL: `https://www.aclweb.org/anthology/2020.semeval-1.1/`.

[206] Bess Schrader. *What's the Difference Between an Ontology and a Knowledge Graph? (White Paper)*. Tech. rep. Enterprise Knowledge, (consulted September 8, 2021). URL: `https://enterprise-knowledge.com/whats-the-difference-between-an-ontology-and-a-knowledge-graph/`.

[207] Rico Sennrich, Barry Haddow, and Alexandra Birch. "Neural Machine Translation of Rare Words with Subword Units". In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Katrin Erk and Noah A. Smith. Berlin, Germany: Association for Computational Linguistics, Aug. 2016, pp. 1715–1725. DOI: `10.18653/v1/P16-1162`. URL: `https://aclanthology.org/P16-1162`.

[208] Sofia Serrano and Noah A. Smith. "Is Attention Interpretable?" In: *Proc. of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, July 2019, pp. 2931–2951. DOI: `10.18653/v1/P19-1282`. URL: `https://aclanthology.org/P19-1282`.

[209] Philippa Shoemark et al. "Room to Glo: A systematic comparison of semantic change detection approaches with word embeddings". In: *Proc. of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics. 2020, pp. 66–76.

[210] Melanie Siegel and Francis Bond. "OdeNet: Compiling a German-WordNet from other Resources". In: *Proceedings of the 11th Global Wordnet Conference*. University of South Africa (UNISA): Global Wordnet Association, Jan. 2021, pp. 192–198. URL: `https://aclanthology.org/2021.gwc-1.22`.

[211] Borsci Simone, Boscarol Maurizio, Cornero Alessandra, et al. *Il Protocollo eGLU 2.1. Il Protocollo eGLU-M. Come realizzare test di usabilità semplificati per i siti web ei servizi online delle PA. Glossario dell'usabilità*. 2015.

[212] Gustaf Stern. "Meaning and change of meaning; with special reference to the English language." In: (1931).

[213] Milan Straka. "UDPipe 2.0 Prototype at CoNLL 2018 UD Shared Task". In: *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*. Brussels, Belgium: Association for Computational Linguistics, 2018, pp. 197–207. DOI: `10.18653/v1/K18-2020`. URL: `https://www.aclweb.org/anthology/K18-2020`.

[214] Milan Straka, Jan Hajic, and Jana Straková. "UDPipe: Trainable Pipeline for Processing CoNLL-U Files Performing Tokenization, Morphological Analysis, POS Tagging and Parsing". In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016*. Ed. by Nicoletta Calzolari et al. Portorož,Slovenia: European Language Resources Association (ELRA), May 2016. URL: `http://www.lrec-conf.org/proceedings/lrec2016/summaries/873.html`.

[215] Morris Swadesh. "Salish internal relationships". In: *International Journal of American Linguistics* 16.4 (1950), pp. 157–167.

[216] Nina Tahmasebi, Lars Borin, and Adam Jatowt. "Survey of Computational Approaches to Lexical Semantic Change". In: *1st International Workshop on Computational Approaches to Historical Language Change*

*2019* (2018). arXiv: `1811.06278`. URL: `http://arxiv.org/abs/1811.06278`.

[217] Nina Tahmasebi and Thomas Risse. "Finding IndividualWord Sense Changes and their Delay in Appearance". In: *International Conference Recent Advances in Natural Language Processing*. Assoc. for Computational Linguistics Bulgaria, Nov. 2017, pp. 741–749. DOI: `10.26615/978-954-452-049-6_095`.

[218] Nina Tahmasebi et al. "DWUG SV: Diachronic Word Usage Graphs for Swedish". Version 2.0.1. In: (Dec. 2022). DOI: `10.5281/zenodo.7389506`. URL: `https://doi.org/10.5281/zenodo.7389506`.

[219] Xuri Tang. "A state-of-the-art of semantic change computation". In: *Natural Language Engineering* 24.5 (Sept. 2018), pp. 649–676. ISSN: 14698110. DOI: `10.1017/S1351324918000220`.

[220] Wayne A Taylor. *Change-point analysis: a powerful new tool for detecting changes*.

[221] Deutsches Textarchiv. *Grundlage für ein Referenzkorpus der neuhochdeutschen Sprache. Herausgegeben von der Berlin-Brandenburgischen Akademie der Wissenschaften*. `http://www.deutschestextarchiv.de/`. 2017.

[222] Thesaurusbüro München Internationale Thesaurus-Kommission, ed. *Thesaurus linguae latinae*. Berlin: Mouton de Gruyter, 1900–.

[223] Elizabeth Closs Traugott. "Semantic change: Bleaching, strengthening, narrowing, extension". In: *Encyclopedia of Language and Linguistics*. Elsevier, 2006.

[224] Elizabeth Closs Traugott and Richard B. Dasher. *Regularity in semantic change*. Cambridge: Cambridge University Press, 2001.

[225] Rocco Tripodi et al. "Tracing Antisemitic Language Through Diachronic Embedding Projections: France 1789-1914". In: *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*. Florence, Italy: Association for Computational Linguistics, Aug. 2019, pp. 115–125. DOI: `10.18653/v1/W19-4715`. URL: `https://aclanthology.org/W19-4715`.

[226] Adam Tsakalidis et al. "Mining the UK Web Archive for Semantic Change Detection". In: *Proceedings of the International Conference on Recent Advances in Natural Language Processing, RANLP 2019, Varna, Bulgaria, September 2-4, 2019*. Ed. by Ruslan Mitkov and Galia Angelova. INCOMA Ltd., 2019, pp. 1212–1221. DOI: 10.26615/978-954-452-056-4_139. URL: https://doi.org/10.26615/978-954-452-056-4_139.

[227] S. Ullmann. *The Principles of Semantics*. Glasgow University publications. Jackson, 1957. URL: https://books.google.it/books?id=YOZYAAAAMAAJ.

[228] Ashish Vaswani et al. "Attention is All you Need". In: *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*. Ed. by Isabelle Guyon et al. 2017, pp. 5998–6008. URL: https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

[229] Denny Vrandečić and Markus Krötzsch. "Wikidata: a free collaborative knowledgebase". In: *Communications of the ACM* 57.10 (2014). Publisher: ACM New York, NY, USA, pp. 78–85.

[230] Ekaterina Vylomova, Sean Murphy, and Nicholas Haslam. "Evaluation of Semantic Change of Harm-Related Concepts in Psychology". In: *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*. Florence, Italy: Association for Computational Linguistics, Aug. 2019, pp. 29–34. DOI: 10.18653/v1/W19-4704. URL: https://aclanthology.org/W19-4704.

[231] Jonas Wallat et al. "Probing BERT for Ranking Abilities". In: *Advances in Information Retrieval*. Ed. by Jaap Kamps et al. Cham: Springer Nature Switzerland, 2023, pp. 255–273. ISBN: 978-3-031-28238-6.

[232] Benyou Wang et al. "On Position Embeddings in BERT". In: *International Conference on Learning Representations*. 2021. URL: https://openreview.net/forum?id=onxoVA9FxMw.

[233] Anna Wegmann, Florian Lemmerich, and Markus Strohmaier. "Detecting Different Forms of Semantic shift in Word Embeddings via Paradigmatic and Syntagmatic Association Changes". In: *Proc. of ISWC*. Springer. Athens, Greece, Nov. 2020, pp. 619–635.

[234] Melvin Wevers. "Using Word Embeddings to Examine Gender Bias in Dutch Newspapers, 1950-1990". In: *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*. Florence, Italy: Association for Computational Linguistics, Aug. 2019, pp. 92–97.
DOI: 10.18653/v1/W19-4712. URL: https://aclanthology.org/W19-4712.

[235] Shuyi Xie et al. "PALI at SemEval-2021 Task 2: Fine-Tune XLM-RoBERTa for Word in Context Disambiguation". In: *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*. Online: Association for Computational Linguistics, Aug. 2021, pp. 713–718.
DOI: 10.18653/v1/2021.semeval-1.93. URL: https://aclanthology.org/2021.semeval-1.93.

[236] Zijun Yao et al. "Dynamic Word Embeddings for Evolving Semantic Discovery". In: *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining, WSDM 2018*. Ed. by Yi Chang et al. Marina Del Rey, CA, USA: ACM, Feb. 2018, pp. 673–681.
DOI: 10.1145/3159652.3159703. URL: https://doi.org/10.1145/3159652.3159703.

[237] David Yenicelik, Florian Schmidt, and Yannic Kilcher. "How does BERT Capture Semantics? A Closer Look at Polysemous Words". In: *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*. Online: Association for Computational Linguistics, Nov. 2020, pp. 156–162.
DOI: 10.18653/v1/2020.blackboxnlp-1.15. URL: https://aclanthology.org/2020.blackboxnlp-1.15.

[238] Frank D. Zamora-Reina, Felipe Bravo-Marquez, and Dominik Schlechtweg. "DWUG ES: Diachronic Word Usage Graphs for Spanish". Version 4.0.0 (full). In: (Apr. 2022).
DOI: 10.5281/zenodo.6433667. URL: https://doi.org/10.5281/zenodo.6433667.

[239] Frank D. Zamora-Reina, Felipe Bravo-Marquez, and Dominik Schlechtweg. "LSCDiscovery: A Shared Task on Semantic Change Discovery and Detection in Spanish". In: *Proc. of the Workshop on Computational Approaches to Historical Language Change (LChange)*. Dublin,

Ireland: Association for Computational Linguistics (ACL), May 2022, pp. 149–164.

[240] Berliner Zeitung. *Diachronic newspaper corpus published by Staatsbibliothek zu Berlin*. http://zefys.staatsbibliothek-berlin.de/index.php?id=155. 2018.

[241] Mengjie Zhao et al. "Quantifying the Contextualization of Word Representations with Semantic Class Probing". In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. Online: Association for Computational Linguistics, Nov. 2020, pp. 1219–1234. DOI: 10.18653/v1/2020.findings-emnlp.109. URL: https://aclanthology.org/2020.findings-emnlp.109.

[242] Fan Zhou and Chengtai Cao. "Overcoming Catastrophic Forgetting in Graph Neural Networks with Experience Replay". In: *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*. AAAI Press, 2021, pp. 4714–4722. DOI: 10.1609/aaai.v35i5.16602. URL: https://doi.org/10.1609/aaai.v35i5.16602.

[243] Jinan Zhou and Jiaxin Li. "TemporalTeller at SemEval-2020 Task 1: Unsupervised Lexical Semantic Change Detection with Temporal Referencing". In: *Proceedings of the Fourteenth Workshop on Semantic Evaluation, SemEval@COLING2020*. Ed. by Aurélie Herbelot et al. Barcelona (online): International Committee for Computational Linguistics, Dec. 2020, pp. 222–231. URL: https://www.aclweb.org/anthology/2020.semeval-1.27/.