Towards an Onomasiological Study of Lexical Semantic Change through the Induction of Concepts

Bastien Liétard and **Mikaela Keller** and **Pascal Denis**
Univ. Lille, Inria, CNRS, Centrale Lille, UMR 9189 - CRIStAL, F-59000 Lille, France
first_name.last_name@inria.fr

Abstract

Lexical Semantic Change, the temporal evolution of the mapping between word forms and concepts, can be studied under two complementary perspectives: semasiology studies how given words change in meaning over time, while onomasiology focuses on how some concepts change in how they are lexically realized. For the most part, existing NLP studies have taken the semasiological (i.e. word-to-concept) view. In this paper, we describe a novel computational methodology that takes an onomasiological (i.e., concept-to-word) view of semantic change by directly inducing concepts from word occurrences at the different time stamps. We apply our methodology to a French diachronic corpus. We examine the quality of obtained concepts and showcase how the results of our methodology can <be used for the study of Lexical Semantic Change. We discuss its advantages and its early limitations.

1 Introduction

Lexical Semantic Change (LSC) is usually defined as the evolution of the meaning of words over time. In the last years, there has been an increasing number of computational approaches proposed to predict LSC between two periods (Schlechtweg et al., 2020; Zamora-Reina et al., 2022) or more (Kulkarni et al., 2015; Alsulaimani and Moreau, 2023). The most recent studies use contextualized word representations and compare how the representations from a later time period differ from those of an earlier period. While some of these approaches use aggregation of pairwise distances between representations of the two periods (Kutuzov and Giulianelli, 2020; Kutuzov et al., 2022), another range of work uses a clustering of a word's contextualized representations to distinguish its different senses, and compare sense inventories over time (Montariol et al., 2021; Laicher et al., 2021). However, this view of LSC is only focused on specific target words and their meanings: it considers change under an *semasiological* perspective. Another side of this two-faced problem is the *onomasiological* perspective, focused on changes in the way a given concept is expressed (Geeraerts et al., 2023).

While the semasiological perspective has been prevalent in recent NLP work on LSC, the onomasiological perspective is widespread in historical linguistics. And one can argue that this perspective has additional explanatory potential for uncovering and characterizing patterns of semantic change, as it takes a more systematic view of the lexicon. For instance, Traugott (1985) argues that the way we express abstract concepts usually borrows words from more concrete concepts; Georgakopoulos and Polis (2021) studied the mixed evolution of the naming of celestial objects and the naming of time-related concepts; Lehrer (1985) showed that animal metaphors of human traits (e.g. *snake* for *treacherous person*) often affect the whole naming of the animal over time. To the best of our knowledge, the only NLP work taking an onomasiological perspective is Franco et al. (2022), but it is limited in scope since they study the evolution of the lexical realizations of the concept DESTROY in Dutch.

One obvious obstacle for any large-scale onomasiological study of LSC is that it requires a concept inventory. In this paper, we propose a novel clustering-based approach that automatically induces concepts from the word occurrences in a diachronic corpus. In line with an onomasiological view, we propose to describe a concept as a set of lemmas that are used to express this concept in a corpus. Specifically, we use contextualized word vectors extracted from XLM-R to represent word occurrences. We rely on a two-step hierarchical clustering to learn the *concepts* from word occurrences at the different time periods. We obtain clusters of words that are supposed to represent concepts as well as a set of concepts that each lemma can refer to. We apply this methodology to a French corpus (the Presto Corpus, Blumenthal et al.

Proceedings of the 5th Workshop on Computational Approaches to Historical Language Change, pages 158–167
August 15, 2024 ©2024 Association for Computational Linguistics

2017), and discuss the quality of obtained clusters and the evolution of clusters and lemmas. Code for this study can be found at `https://github.com/blietard/towards-onomasio-semchange`.

2 Diachronic Concept Induction

We call *concept* the intended meaning behind the usage of a word, the mental representation associated with a word in a given context and abstracting over its denotation. In "*a bunch of people*" and "*a group of tourists*", "bunch" and "group" are synonymous and both denote the same concept. We call *naming* of a concept the set of lemmas used to refer to this concept. In this paper, we use the term "word" as a synonym of "lemma" to avoid repetitions.

Let W be a set of target lemmas. Let C be a corpus of texts spanning from time τ_{start} to time τ_{end}. The time span $[\tau_{start}, \tau_{end}]$ is divided into a set of M periods $T = \{t_1, \dots, t_M\}$. We call $o^w_{t,i}$ the i-th occurrence of the word $w \in W$ in the corpus in period t. We denote O^w_t the set of all $o^w_{t,i}$ for a given word w and given time period t.

2.1 Inducing concepts at a single period

Let us first consider a single time period t and only the corresponding occurrences. The goal is to automatically learn a clustering function that maps each word occurrence $o^w_{t,i}$ to a cluster c that represents a concept. Contrarily to Word Sense Induction that only regroups occurrences around the same word *sense*, our clustering aims to account for concepts shared across lemmas: all occurrences instantiating the same concept, whether they are of the same word or not, should be mapped to the same concept-cluster c. We propose to perform this concept clustering in two steps, a *lemma-centric* clustering and a *cross-lexicon* clustering.

In the first, lemma-centric clustering, an algorithm A_1 partitions each lemma w's occurrences to obtain a set of n_w clusters, as in Martinc et al. (2020), which we simply call *sense clusters*. The j-th sense of lemma w at time t group is represented with $s^w_{t,j}$. For each word w at time t we obtain the set $S^w_t = \{s^w_{t,1}, \dots, s^w_{t,n_w}\}$.

The second, cross-lexicon clustering aims at merging sense clusters containing occurrences with the same concept, and keep distinct sense clusters of occurrences with different meanings. In the representational space of all sense clusters of all lemmas ($\bigcup_{w \in W} S^w_t$), we apply another cluster algorithm A_2, obtaining clusters of sense clusters. The final obtained clusters are our *concept clusters*.

The mapping from occurrences to concepts is done by transitivity: if $s^w_{t,j}$ is clustered in a concept c, any occurrence $o^w_{t,i}$ clustered in the group represented by $s^w_{t,j}$ can be directly assigned to concept c. By extension, we say a concept cluster c contains a lemma w if one of the occurrences of w is assigned to c. Sense clusters of the same lemma w are said to be merged because their occurrences will appear in the same concept cluster in the end. Thus, when in our analysis we refer to the *senses* of a lemma and its degree of polysemy, we are only interested in the *concept-derived senses*, i.e. the set of the concept clusters that occurrences of a word are assigned to and not the intermediate sense clusters. A polysemous word is expected to be assigned to multiple clusters, while synonymous words are expected to be assigned to at least one common cluster.

2.2 Inducing concepts over time

For diachronic purposes, we need not only to consider concepts induced at one time t, but also to align concept clusters of different periods. Following existing work such as Kanjirangat et al. (2020), we propose to learn the clusterings merging all time periods, using all occurrences from C as a whole instead of learning clusterings for each time independently. Doing so, we can track the *evolution of a concept cluster* simply by looking at the occurrences from the different times that are assigned to this cluster. We can also track the *evolution of a lemma* by looking at the different clusters to which its occurrences are mapped over time. Not only does this allow to detect a semantic change, but it is also *characterizes* the type of change (revealing if the lemma gained and/or lost senses).

3 Experiments

We apply the proposed methodology (section 3.2) to an historical corpus. We discuss the quality of clusters in section 3.3, and conduct both semasiological and onomasiological studies in sections 3.4 and 3.5.

3.1 Diachronic Data

The Presto Corpus[1] is a French historical corpus of texts from 1500 to 1950 (Blumenthal et al., 2017).

[1] `http://presto.ens-lyon.fr/?page_id=584`

Most word occurrences are annotated with a Part-Of-Speech tag and the modern form of lemmas, allowing us to mostly ignore orthographic variations over time.[2] We use the freely available "Noyau" part which contains 53 documents. We focused this initial study study on Nouns. For statistical significance and because of the rather small size of the corpus, we selected the 623 most frequent noun lemmas across the overall time span to be our target words, tallying a total of 314k occurrences. In our analyses, we define 3 periods such that they all contain balanced portions of the data (33% of the target lemmas' occurrences): 1500-1699, 1700-1799 and 1800-1949. The 3 intervals share 498 out of the 623 selected lemmas. Unless stated otherwise (as in Section 3.4), our analyses are conducted using the full set of 623 lemmas. A discussion of our selection process and choice of periods can be found in Appendix A.3.

To decrease the impact of orthographic differences in old periods, we partially lemmatize sentences by replacing all nouns, verbs, adjectives and adverbs with their modern-form lemmas. While we acknowledge that these morphological replacements may bring a slight semantic deviation (e.g. singular instead of plural), we consider that overcoming orthographic discontinuities is a higher priority for the clustering to be as little as possible influenced by tokenization-based differences when using Contextualized Language Models to represent occurrences in a vector space.

French-English translations of examples used in this paper can be found in Appendix A.1.

3.2 Models and Algorithms

We use the XLM-R model (Conneau et al., 2020) (large) to get contextualized vector representations of occurrences of the 623 lemmas. For each lemma, we use Agglomerative Clustering with minimum linkage in place of algorithm A_1 to create (lemma-centric) clusters of occurrences, the sense clusters. As explained in Section 2.2, the clustering algorithm is applied on the *whole* set of occurrences for each lemma, regardless of time periods. Word embeddings contained in each cluster are averaged to obtain a single vector representation per sense cluster. The cross-lexicon clustering algorithm A_2 is applied to the set of sense cluster representatives of all words. In our experiments, A_2 is chosen to be Agglomerative Clustering with average linkage. In the end, occurrences are labeled with a concept cluster resulting from A_2 by transitivity from A_1, as described in 2.1.

We choose XLM-R because of its zero-shot cross-lingual transferability to French due to its multilingual training data. We extracted vectors from layers 14 to 17 (incl.), averaging over these layers to get the embeddings of the target word. Vectors of subwords were averaged if necessary to get a single vector. This choice of using intermediate/high layers of the model is motivated by the work of (Chronis and Erk, 2020) who found that lexical similarity (a core aspect of synonymy) was best represented in these layers. We also find these layers to produce qualitatively better clusters than last layers of the model (21 to 24).

In this diachronic data, there is no sense annotation to guide us in chosing the algorithm and hyperparameters. Therefore we relied on our expertise of French for accessing the quality of the obtained cluster in the most recent time period (1800-1949), similarly to analysis presented in Section 5. For a given combination of A_1 and A_2, we kept the set of hyperparameters that provides the highest number of concept clusters containing at least 2 but no more than 5 lemmas in the last time period. This upper limit of 5 was decided from preliminary observations that clusters containing more that 5 different lemmas almost always gathered lemmas that did not share a common concept but were linked by other non-semantic factors (for instance, words that shared a common subword token). In the absence of sense annotations to better evaluate the clusterings, this rule helps ensuring the recall of a maximum number of concept clusters, while avoiding clusters that are too large. For A_1, we tried K-means, Affinity Propagation and Agglomerative Clustering. For algorithm A_2, we tried Affinity Propagation and Agglomerative Clustering. To decide which algorithms to use, we kept the combination that produced the most plausible clusters of size 2 to 5 in this period, i.e. clusters containing actual (near-)synonyms. This process resulted in our preference for Agglomerative Clustering. More details on tried hyperparameter values in Appendix A.2.

Our double-clustering methodology using Agglomerative Clustering was benchmarked among other systems in a parallel study conducted in Liétard et al. (2024) on SemCor, a synchronic English corpus annotated with concepts from the original

[2] For the method to be applied on unannotated data, one could use a syntactic parser/lemmatizer with special rules for orthographic changes (e.g. VARD2, Baron and Rayson 2008).

Category	Total	Cluster size 2	3	4
Nb. of clusters	101	62	29	10
Synonyms	27%	32%	24%	0%
Near-synonyms	20%	15%	28%	30%
Lexical / topical	40%	42%	38%	40%
Invalid cluster	13%	11%	10%	30%

Table 1: Categorization of small induced concept-clusters in 1800-1949. Invalid clusters are those showing no semantic relation. Raw counts in Appendix A.6.

Princeton WordNet. It achieved the best results reaching a F_1 score of 0.60 and a precision of 0.80.

Improvement of the model selection criterion used in this initial study is left for future work.

3.3 Analysis of Induced Concepts

With the chosen clustering algorithms and corresponding hyperparameters, we obtain a total of 867 concept-clusters. In each period, 40% of the 867 clusters are not represented and another 40% are expressed with only a single lemma. In the 265 (31%) that are instantiated in *all* three periods, 54% of them also only contain a single lemma. This particular observation is in line with Clark (1993)'s principle of Conventionality : "*For certain meanings, there is a form that speakers expect to be used in the language community*". These distribution details can be found in Appendix A.5. We also found that while only 16% of concept clusters contain multiple lemmas, 46% of words have at least two senses:[3] *polysemy* is a more frequent phenomenon that *synonymy*. We also noticed that a small fraction of clusters (less than 7) are very large and gather lemmas not based on semantic similarity (e.g. based on a common subwords after being processed by the tokenizer (e.g. "autor**ité**", "postér**ité**")).

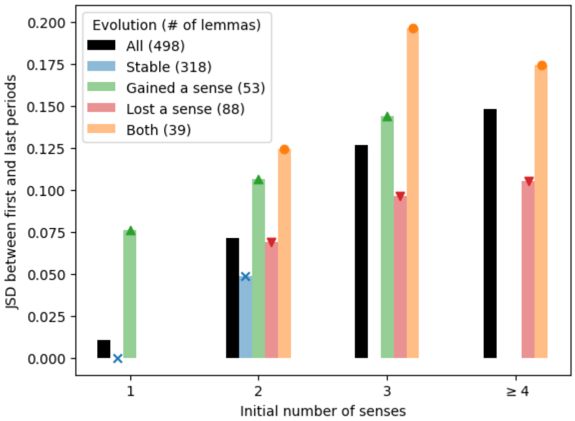

Figure 1: JSD and detected type of evolution of lemmas with respect to their initial number of senses. Missing points (no bar and no marker) indicate that no lemma in this category of evolution had this initial number of senses. Stable lemmas with 1 sense have a JSD of 0.

Concept Evolution	#Concepts
Expanded naming	27 (10%)
Shrinked naming	5 (2%)
Both	6 (2%)
Identical naming	227 (86%)

Table 2: Evolution of concept-clusters over time.

Clusters of smaller sizes are more reliable. Out of the 867 clusters, we manually evaluated the 101 concept-clusters of 2 to 4 lemmas in the last time interval, and the distribution of our annotations is displayed in Table 5. We focused only on the last time period because it is the closest to the current state of French. Only 10% of these small clusters are to be considered invalid. Around 30% are actual (cognitive) synonyms, and 20% are near-synonyms,[4] i.e. words that are not absolute synonyms but overlapping in meaning (e.g. "bourse", "fortune", "richesse", "trésor" denote an individual's wealth at different scales). The remainder exhibits a lexical (like hyper/hyponyms, antonyms, etc.) or another topical relation between words (e.g. "journée", "nuit", "soir"). This kind of clusters can be seen as partial semantic fields. Although they are not synonyms, we argue that they are still interesting for the study of LSC. For instance, the disappearance of a lemma from such a cluster could indicate a transfer of its unique semantic load to another word.

3.4 Evolution of Target Lemmas

In this section, we discuss the semantic evolution of target lemmas, i.e. a semasiological view. Here we only focus on the 498 (out of 623) lemmas that appear in every period and look at their concept-derived senses. We use these sense inventories to distinguish 4 categories of evolution a lemma can undergo. A lemma *gained a sense* if one of its senses in the last period is new compared to the first period. In the reverse scenario, we say the

[3] average polysemy: 2.28 senses per word ; average synonymy: 1.15 word per concepts

[4] using definitions of scales of synonymy provided by Stanojevic (2009).

lemma *lost a sense*. A lemma can also *both* gain and loose a sense between the first and the last time periods. A lemma is said to be *stable* if none of these cases apply.

Prior works like Giulianelli (2019) have proposed continuous measures of semantic change based on distribution in sense inventories. For each target word, we compute the Jensen Shannon Divergence (JSD) between the distribution of concept to which its occurrences are assigned in the first period and those of the last period. Both the categorical and the continuous approaches reflect LSC, the former allowing to charaterize the type of change and the latter accounting for the relative frequency of each sense.

Let us now consider the relation between LSC and polysemy. In Figure 1, we show these two measures of semantic change (evolution category and averaged JSD) with respect to the number of senses. 318 target lemmas out of 498 are *stable* in meaning and have lower JSD. Note that these stable lemmas have a very low number of senses (1 or 2). Conversely, lemmas with stronger semantic change are those with many senses. They are more prone to loose and/or gain a sense over time. We also find a significant ($p\text{-value} < 0.01$) positive correlation between the initial number of senses and JSD. This holds even if we only consider those with at east 2 initial senses. This observation echoes the Law of Innovation proposed by Hamilton et al. (2016) and studied by Luo and Xu (2018), stating that polysemy is positively correlated with semantic change.

3.5 Evolution of Induced Concepts

In this section, we adopt an onomasiological point-of-view. Let us focus on the 265 concepts that are instantiated at all time intervals. We are interested in the evolution of the naming of a given concept over time, i.e. changes in the set of lemmas appearing in the corresponding cluster between the first time interval (1500-1699) and the last one (1800-1949). Such a naming may have *expanded* (gained lemmas), *shrinked* (lost lemmas) or both, or neither and remained *identical*. The distribution of these cases is presented in Table 2

86% of induced clusters kept an identical naming, which we expected because we had no difficulty to understand the meaning of texts from 1550 in modern spelling.

In expanded-naming clusters, we find that it results in most cases from new lemmas *appearing* later in the corpus. We search their history in the TLFi, a reference dictionary for French[5], and find the introduction of the word in the corpus often coincides with a new meaning that is more general (less specific) than existing ones, and the cluster to which the new word is assigned indeed corresponds to this emerging sense. For instance, the word “tribu” appeared in the 1700-1799 interval in the corpus and is clustered with “peuple.” The TLFi indicates that it was in 1734 that “tribu” acquired its new meaning of “*social group based on ethnic kinship*”. Yet, we cannot verify that the introduction is caused by the new meaning. In other cases, the introduced word does not have existing senses and is newly created at the time of its appearence in the corpus (e.g. “incendie” (clustered with “feu”), only attested past 1600 in the TLFi).

In the case of shrinked-naming concepts, we can distinguish clusters in which a lemma disappeared from the corpus (e.g. “parquoi”, old alternative to “pourquoi” with which it was clustered at the begining) and clusters in which a lemma was removed from the cluster while still existing (e.g. “amitié”, no longer clustered with “amour”, as its use for romantic feelings became old-fashioned.)

4 Conclusion

In this paper, we proposed a new methodology for inducing concepts from word occurrences. We mapped each word to a set of concepts and each concept to a set of words at different time period. Using historical data in French, we made of proof-of-concept of this methodology and showed in an initial study that this approach allows to characterize the evolution of a word’s sense inventory, as well as those of a concept’s naming. This offers a promising direction and can lead to a better understanding of Lexical Semantic Change and its *systemic* aspects, enabling the investigation of both the semasiological and the onomasiological aspect of Lexical Semantic Change.

5 Limitations

Without access to sense-annotated diachronic data, we cannot evaluate with certainty the quality of induced concept-clusters. Therefore, while we conducted a qualitative evaluation on a portion of the clusters at the lemma level, we cannot evaluate the precision of the clustering at the occurrence level, neither whether we retrieved all actual concepts.

[5] http://atilf.atilf.fr/

To select the best set of hyperparameters, we choose to maximize the number of obtained clusters containing between 2 and 5 lemmas. We discarded the conventional use of statistical criterion such as Silhouette score because (i) this score puts an assumption on the shape/density of clusters and we don't believe it applies ; (ii) Martinc et al. (2020) already showed that Silhouette score was not satisfying when inducing senses for Lexical Semantic Change. Our criterion is inspired by the objective to retrieve a maximum number of concept and complete naming, and the observation that clusters of more than 5 lemmas are usually noisy and invalid. Without annotated data, we cannot ascertain how good this heuristic is. A future study could attempt to compare different heuristics to determine the most relevant to induce concepts.

Prior studies of LSC with word-sense clustering (Martinc et al., 2020; Kutuzov et al., 2022) found that clustering in raw vector spaces from Language Models sometimes find clusters of word *usages* instead of actual word *meanings*, which may happen in our lemma-centric clustering. We think the impact of this in the onomasiological setting is limited; this may explain the number of clusters actually corresponding to lexical/topical relations instead of actual (near-)synonymy. Improving the lemma-centric clustering to avoid this could increase the precision of obtained clusters in future studies.

The small size and the sparse nature of the corpus prevents detailed analysis and fine-grained results. Taking smaller time periods lead to very unbalanced number of lemmas/occurrences, and the 18th century is prominent compared to other.

The fact that a lemma is missing at a given period does not necessarily mean that it was not used at all at the time; it could be just an artefact of the small size of the corpus.

Our clustering approach appears to group together word tokenized in multiple subwords, without actual semantic relation between them. Further research could be made about these invalid clusters and how to parse them into plausible clusters.

Acknowledgments

We gratefully thank the anynymous reviewers for their insightful comments. This research was funded by Inria Exploratory Action COMANCHE.

References

Ashjan Alsulaimani and Erwan Moreau. 2023. Improving diachronic word sense induction with a nonparametric Bayesian method. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8908–8925, Toronto, Canada. Association for Computational Linguistics.

Alistair Baron and Paul Rayson. 2008. VARD2 : a tool for dealing with spelling variation in historical corpora. In *Proceedings of the Postgraduate Conference in Corpus Linguistics*. Aston University, Birmingham, UK.

Peter Blumenthal, Sascha Diwersy, Achille Falaise, Marie-Hélène Lay, Gilles Sourvay, and Denis Vigier. 2017. Presto, un corpus diachronique pour le français des xvie-xxe siècles. In *Actes de la 24eme conférence sur le Traitement Automatique des Langues Naturelles-TALN*, volume 17.

Gabriella Chronis and Katrin Erk. 2020. When is a bishop not like a rook? when it's like a rabbi! multi-prototype BERT embeddings for estimating semantic relationships. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 227–244, Online. Association for Computational Linguistics.

Eve V. Clark. 1993. Conventionality and contrast. In *The Lexicon in Acquisition*, Cambridge Studies in Linguistics, page 67–83. Cambridge University Press.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Karlien Franco, Mariana Montes, and Kris Heylen. 2022. Deconstructing destruction: A cognitive linguistics perspective on a computational analysis of diachronic change. In *Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change*, pages 23–32, Dublin, Ireland. Association for Computational Linguistics.

Dirk Geeraerts, Dirk Speelman, Kris Heylen, Mariana Montes, Stefano De Pascale, Karlien Franco, and Michael Lang. 2023. *Lexical variation and change*. Oxford University Press, London, England.

Thanasis Georgakopoulos and Stéphane Polis. 2021. Lexical diachronic semantic maps: Mapping the evolution of time-related lexemes. *Journal of Historical Linguistics*, 11(3):367–420. ISBN: 2210-2116 Publisher: John Benjamins Type: https://doi.org/10.1075/jhl.19018.geo.

Mario Giulianelli. 2019. *Lexical Semantic Change Analysis with Contextualised Word Representations*. University of Amsterdam - Institute for logic, Language and computation.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1501, Berlin, Germany. Association for Computational Linguistics.

Vani Kanjirangat, Sandra Mitrovic, Alessandro Antonucci, and Fabio Rinaldi. 2020. SST-BERT at SemEval-2020 task 1: Semantic shift tracing by clustering in BERT-based embedding spaces. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 214–221, Barcelona (online). International Committee for Computational Linguistics.

Vivek Kulkarni, Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2015. Statistically significant detection of linguistic change. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15, page 625–635, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Andrey Kutuzov and Mario Giulianelli. 2020. UiO-UvA at SemEval-2020 task 1: Contextualised embeddings for lexical semantic change detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 126–134, Barcelona (online). International Committee for Computational Linguistics.

Andrey Kutuzov, Erik Velldal, and Lilja Øvrelid. 2022. Contextualized embeddings for semantic change detection: Lessons learned. In *Northern European Journal of Language Technology, Volume 8*, Copenhagen, Denmark. Northern European Association of Language Technology.

Severin Laicher, Sinan Kurtyigit, Dominik Schlechtweg, Jonas Kuhn, and Sabine Schulte im Walde. 2021. Explaining and improving BERT performance on lexical semantic change detection. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 192–202, Online. Association for Computational Linguistics.

Adrienne Lehrer. 1985. *The influence of semantic fields on semantic change*, pages 283–296. De Gruyter Mouton, Berlin, New York.

Bastien Liétard, Pascal Denis, and Mikaella Keller. 2024. To word senses and beyond: Inducing concepts with contextualized language models. *Preprint*, arXiv:2406.20054.

Yiwei Luo and Yang Xu. 2018. Stability in the temporal dynamics of word meanings. In *CogSci*.

Matej Martinc, Syrielle Montariol, Elaine Zosa, and Lidia Pivovarova. 2020. Capturing evolution in word usage: Just add more clusters? In *Companion Proceedings of the Web Conference 2020*, WWW '20, page 343–349, New York, NY, USA. Association for Computing Machinery.

Syrielle Montariol, Matej Martinc, and Lidia Pivovarova. 2021. Scalable and interpretable semantic change detection. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4642–4652, Online. Association for Computational Linguistics.

Dominik Schlechtweg, Barbara McGillivray, Simon Hengchen, Haim Dubossarsky, and Nina Tahmasebi. 2020. SemEval-2020 task 1: Unsupervised lexical semantic change detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1–23, Barcelona (online). International Committee for Computational Linguistics.

Maja Stanojevic. 2009. Cognitive synonymy: A general overview. *Facta Universitatis, Series: Linguistics and Literature*, 7(2):193–200.

Elizabeth Closs Traugott. 1985. On regularity in semantic change. *Journal of Literary Semantics*, 14(3):155–173.

Frank D. Zamora-Reina, Felipe Bravo-Marquez, and Dominik Schlechtweg. 2022. LSCDiscovery: A shared task on semantic change discovery and detection in Spanish. In *Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change*, pages 149–164, Dublin, Ireland. Association for Computational Linguistics.

A Appendix

A.1 French-English translations

In this paper, we used a number of exemples in French, as our experimental data were in French. Translations can be found in Table 3.

A.2 Representations, Algorihtms, Hyperparameters

For k-means as A_1, we tried values of k between 2 and 10. For both A_1 and A_2, when using Agglomerative Clustering, we tried average, minimum and maximum linkage. We set a linkage threshold below which clusters are merged iteratively. Calling μ the average distance between occurrences of a considered set of occurrences, and σ the standard deviation, we set the value of this threshold to $\mu + n \times \sigma$, with n an hyperparameter. When using Agglomerative Clustering for A_1 on each set of occurrences of a lemma, n is shared across all lemmas but the linkage threshold is computed using each set of occurrences. As a result, we obtain a dynamic number of clusters that is more suited to

French	English
amitié	friendship, affection
amour	love
autorité	authority
bourse	purse
ennui	boredom
envie	envy
feu	fire
fortune	fortune, wealth
groupe	group
incendie	fire (vast and uncontrolled)
jour	day
journée	day, daytime
parquoi	old alternative for *pourquoi*
peuple	people
postérité	posterity
pourquoi	the *why*, an explaination
réseau	network
richesse	wealth
soir	evening
système	system
trésor	treasure
tribu	tribe

Table 3: French-English translations

each lemma. We tried value of n between -2 and +2.

A.3 Selection criterion

We find that the data are noisy, and a number of lemmas do not appear often. Indeed, the number of documents in the corpus is relatively small, leaving room for sparsity and discontinuity in the representations of lemmas. Therefore, we had to select a subset of them.

To this extent, we partition it into 50 years time spans. Doing so, we ensured that the number of documents was balanced between spans, and that we can control that selected lemmas are represented *frequently enough* and not sparsely across too large spans.

In order to mitigate the noise resulting from the sparse nature of the data, we apply the following selection criteria. We keep only lemmas :

- appearing in at least 3 consecutive spans,
- occurring at least 10 times in the overall corpus,
- at least 3 times in each spans where they are present,
- composed of a single word and of 3 characters at least.
- appearing in the first or the last span or both.

Doing so, we mitigate the risk for selected lemmas to be subject to unexplained discontinuity over time. The last criterion is applied because our analyses are conducted mainly by comparing early and late time periods.

After selection however, these 50 years long spans are not balanced enough in the corpus for fair analyses. See Appendix A.4.

A.4 Corpus description and choice of periods

Time span	#Doc.	$\lvert W \rvert$	#Occ.	Ratio
1500	6	484	12 014	24.8
1550	5	523	30 206	57.8
1600	6	537	35 202	65.6
1650	6	541	34 177	63.2
1700	4	547	10 729	19.6
1750	8	608	89 179	146.7
1800	6	614	46 778	76.2
1850	6	611	33 823	55.4
1900	6	599	22 146	37.0
Total	53	623	314 254	504.4

Table 4: Number of documents, of target words, of occurrences and ratio between occurrences and target words at the different spans (half centuries).

The number of documents, of selected target words and of their occurrences can be found in Table 4. Note that the number of occurrences is not uniform across the spans.

We remark here that the 18th century is an outlier. Its first half contains the lowest number of occurrences, but its second half is very big compared to any other span, containing around 28% of occurrences on its own.

Figure 2 shows that the number of target lemmas is not equally distributed over 50-years time spans, and that only a subset of them (425 out of 623) is actually appearing in all spans. The induction of concepts suffers a similar imbalance.

We posit three possible reasons for a lemma to be missing in a time span : (i) the lemma was not used in the language at the time, whether is appears later

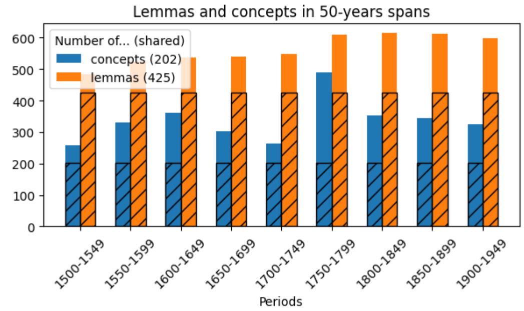

Figure 2: Number of lemmas and concepts in the different periods (half centuries). Hatched areas represent the 425 lemmas and 202 concepts appearing in all periods.

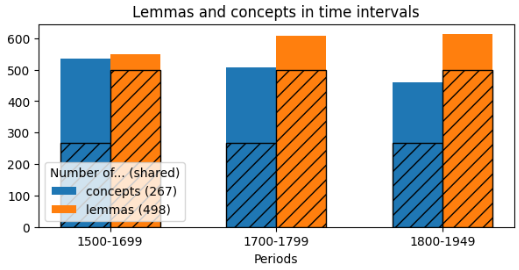

Figure 3: Number of lemmas and concepts in the different time intervals. Hatched areas represent the 498 lemmas and 265 concepts appearing in all periods.

or had already disappeared ; (ii) the lemma was used but is not represented in this span of corpus ; (iii) the lemma was used but rare, and as such does not appear in the corpus for this time span.

Similarly, we posit three possible reasons for concepts not to be instantiated: (i) some concepts may not exists in the language at some time spans (e.g. the concept of COMPUTER) ; (ii) this may be because the span is relatively short and the corpus is not uniformly distributed; (iii) our cluster induction may have failed to identify occurrences instantiating this concept. There is however no way for us to know which of these cases apply.

This leads us to consider larger time periods for our analyses : 1500-1699, 1700-1799 and 1800-1949. Such large periods would not be suitable for target words selection, as we need to ensure a word is *regularly* instantiated over time. At analysis time however, while large periods will prevent us to notice subtle or short-lived semantic changes, this balances the number of occurrences, of lemmas and of retrieved concepts (see Figure 3).

The length of considered periods for analysis has no influence on the actual clustering, as we apply the clustering algorithms on data from all periods.

		Cluster size		
Category	**Total**	**2**	**3**	**4**
Synonyms	27	20	7	0
Near-synonyms	20	9	8	3
Lexical/topical relation	41	26	11	4
Invalid cluster	13	7	3	3
Total	101	62	29	10

Table 5: Categorization of small induced concept-clusters in 1800-1949. Invalid clusters are those showing no semantic relation.

A.5 Distribution of concepts size over time

Figure 4 shows the distribution of concept sizes over time. At a given time, the concept size is the number of lemmas for which at least one occurrences is assigned to the concept.

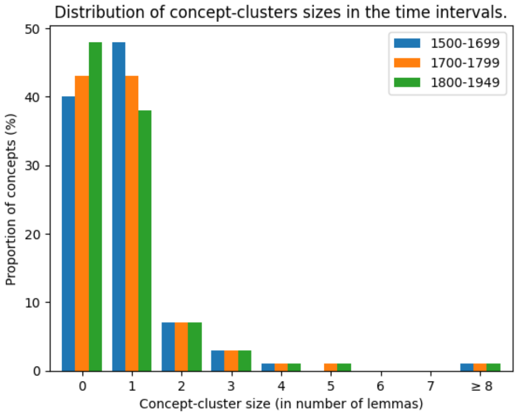

Figure 4: Distribution of the size of the 867 concept-clusters in the different time intervals. Size of 0 means that these concepts are not instantiated.

A.6 Qualitative analysis: raw counts

A.7 Evolution of the number of senses over time

Table 6 shows how the number of senses of lemma changes. Stable lemmas are those with a very low number of senses, while lemma that change have higher number of senses.

		Average number of senses		
Evolution of lemmas	**# Lemmas**	**1500-1699**	**1700-1799**	**1800-1949**
Lost a sense	88	2.84 ± 2.02	1.94 ± 1.44	1.32 ± 1.01
Gained a sense	53	1.15 ± 0.50	1.77 ± 1.12	2.30 ± 0.80
Both	39	4.95 ± 3.09	4.10 ± 2.60	4.18 ± 2.52
Stable	318	1.07 ± 0.25	1.18 ± 0.44	1.07 ± 0.25

Table 6: Evolution of the number of senses of target lemmas over time.