Multi-word Measures: Modeling Semantic Change in Compound Nouns

Chris Jenkins and **Filip Miletić** and **Sabine Schulte im Walde**
Institute for Natural Language Processing, University of Stuttgart
{christopher.jenkins,filip.miletic,schulte}@ims.uni-stuttgart.de

Abstract

Compound words (e.g. *shower thought*) provide a multifaceted challenge for diachronic models of semantic change. Datasets describing noun compound semantics tend to describe only the predominant sense of a compound, which is limiting, especially in diachronic settings where senses may shift over time. We create a novel dataset of relatedness judgements of noun compounds in English and German, the first to capture diachronic meaning changes for multi-word expressions without prematurely condensing individual senses into an aggregate value. Furthermore, we introduce a novel, sense-targeting approach for noun compounds that evaluates two contrasting vector representations in their ability to cluster example sentence pairs. Our clustering approach targets both noun compounds and their constituent parts, to model the interdependence of these terms over time. We calculate time-delineated distributions of these clusters and compare them against measures of semantic change aggregated from the human relatedness annotations.

1 Introduction

Novel uses of language (coining new senses of words and phrases) are readily available at all times, but not necessarily 'sticky' to the point that they see widespread adoption, making the new sense apparent in a sample of text. In addressing this challenge, models of language change are aided by comparing terms that may have changed against reference points assumed not to have changed. Compound nouns in English and German provide an exciting test ground of hypothesized changing and reference terms for computational models of semantic change, due to their sustained productivity in both languages and diverse sets of possible relationships between the meanings of the constituent words of a compound, and the compound as a whole.

We compare changes and continuities in noun-compound semantics from a historical perspective, comparing the extent to which a compound like *gold mine* might differ in its usage over time compared with both its constituent words (*gold*, *mine*) and with the use of further compounds that share a constituent (e.g. *silver mine*, which lacks the broader metaphorical sense that *gold mine* has – see the first row of Table 1). In this way, we go beyond existing work in lexical semantic change detection that focuses on the change of individual words without considering any mutual dependencies with other words.

Unrelated	"He never returned to his architectural **gold mine** of the prewar period."	"Next, there was a copra plantation up North, then a bogus **gold mine** in Colombia, a charter business,..."
Related	"One morning they saw a workman standing on the **brick wall**, who looked about him as if quite at leisure..."	"Most of the blast from a car bomb outside the French embassy was absorbed by a thick **brick wall**."

Table 1: Two pairs of sentences and their average annotation on the relatedness of the compound in bold.

Given that we do not know in advance with how many senses a compound of interest was used across ≈200 years, we use clustering methods[1] to find groups of contexts that are similar. We employ contextual word embeddings to represent the target compounds, because they allow each target to have multiple representations in the same space. This lets us compare the contexts of compound nouns across time periods, and to predict whether the distribution of each target differs sufficiently to constitute an overall change (via broadening or narrowing the available set of senses). Importantly, the development of these contexts is not considered in isolation, but rather in terms of changes in the

[1] Code available at https://gitlab.com/cjenk/sem-change-clustering/

Findings of the Association for Computational Linguistics: ACL 2025, pages 10850–10864
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

contexts of the compound's constituent nouns, or in terms of sets of noun compounds that share either the same modifier or head constituent.

While most prior evaluations of semantic change models only use aggregate quantitative ratings of target words, we propose a multi-faceted assessment. We evaluate our clustering approaches in terms of (1) accuracy against pairwise gold-standard example sentences that we annotated for semantic relatedness, (2) with respect to multiple measures of semantic change, which we obtained by comparing average relatedness ratings across time periods, and (3) by inducing individual senses by clustering the graph of annotated examples.

We provide the following contributions: (1) A novel dataset of relatedness ratings for English and German noun compounds across diachronic contexts. (2) Expansion of the lexical semantic change and semantic-relatedness clustering tasks to complex lexical items (i.e., noun compounds). (3) Contrasting evaluation methods using in-context semantic relatedness ratings, and measures of semantic change in aggregate.

2 Prior Work

The precise definition of 'compound' can be difficult to arrive at (e.g. via constraints of structural integrity), especially cross-linguistically. We take after Bauer (2017) to view compounds as 'word-like' (functioning as a whole like a word), while being composed of more than one 'word' whose contributions to the overall meaning we can analyze separately.

Computational Modeling of Noun Compounds Modeling the diachronic development of compound nouns builds on the well-established synchronic task of predicting the degree of compositionality for a compound i.e., the extent to which its components contribute to its overall meaning (Baroni et al., 2014; Reddy et al., 2011; Schulte im Walde et al., 2016a; Cordeiro et al., 2019; Alipoor and Schulte im Walde, 2020; Miletic and Schulte im Walde, 2023), typically using vector space representations created from the distribution of terms from a corpus. Bringing the task of compositionality prediction to a diachronic setting, Dhar et al. (2019) use SVD to jointly train Word2Vec embeddings for noun compounds and their constituents, comparing them using cosine similarity across different segments of a diachronic corpus in order to estimate the degree of compositionality at different points in time. Other approaches include Mahdizadeh Sani et al. (2024), who applied topic models to co-occurrence vectors of noun compounds and their constituents, using similarity comparisons at different time slices to make binary high/low compositionality predictions. To our knowledge there is no diachronic extension to the gold-standard compositionality ratings that exist for present-day English and German, nor is there a general semantic change dataset for English and German noun-compounds.

Lexical Semantic Change (LSC) Detection A general formulation of the LSC task as given in SemEval-2020 Task 1 (Schlechtweg et al., 2020) is to identify words that have gained or lost senses between two time-separated subcorpora, contrasting with words that were stable. The task was defined with binary and graded variants. Although the LSC task is ultimately to predict that a word's meaning has changed, some means of representing polysemy may help to detect the expansion or contraction of possible senses for a word, potentially coexisting with other, stable senses. To give a contrasting example: the newer meaning of *computer* (machine) has superseded *computer* (occupation), while *rock* (stone) coexists with the newer *rock* (music genre).

Corpus-based LSC detection techniques make use of the assumption that a change in a term's meaning can be apprehended through changes in the context that the term appears in. Essentially, this is an articulation of the distributional hypothesis (Harris, 1954)[2], which can be interpreted either weakly, that the meaning of a term correlates with its contextual use, or more strongly, that meaning *is* exactly defined by the set of contexts that the term is employed in (Arseniev-Koehler, 2021). We tend toward the former, correlational interpretation of the distributional hypothesis, which better fits our focus on the *dynamic* nature of semantics.

Diachronic Representations Since many existing approaches to LSC detection operate in terms of aggregate meaning changes, static representations (e.g. Word2Vec; Mikolov et al., 2013) have been widely utilized. When used in LSC detection, static representations are created separately using separate time-delineated corpora, and subsequently the two vector spaces are aligned to enable direct comparisons between the two representations

[2] Stated even more strongly by Firth: "The use of the word 'meaning' is subject to the general rule that each word when used in a new context is a new word." (Firth, 1957, p.190)

of each word of interest (Hamilton et al., 2016; Schlechtweg et al., 2019). The nature of static representations collapses all senses in which a given token is used into a single representation, making it difficult to interpret *how* a word's meaning may have changed, only that some change is apparent in aggregate. By allowing for the same word or sequence of words to have multiple vector representations, contextualized representations obviate the need to align separate vector spaces in order to compare uses from separate (sub)corpora, and have the potential to represent multiple senses of a word in the same embedding space. The larger parameter space of contextual models (prototypically BERT; Devlin et al., 2019) necessitates a corresponding compression of the number of unique tokens in the model's vocabulary through the use of sub-word tokenization algorithms like SentencePiece (Kudo and Richardson, 2018). Work such as Giulianelli et al. (2020), Martinc et al. (2020), and Kanjirangat et al. (2020) shows that contextualized embeddings can be clustered to obtain predictions of LSC. However, they are much more resource-intensive than static embeddings to use, as noted by Montariol et al. (2021).

Word-in-Context and Word-Sense Induction The Word-in-Context dataset (Pilehvar and Camacho-Collados, 2019) was developed to expand typical semantic understanding benchmarks deliberately towards words that require more supporting context to be understood. The English Word-in-Context test set contains very few noun compounds, always in closed form (e.g. *pocketbook*), and is derived from synchronic lexical resources like WordNet (Miller, 1995), limiting its comparability with our present investigation. The German dataset from the cross-lingual Word-in-Context dataset (Raganato et al., 2020), does however contain many noun compounds. This discrepancy may be only a consequence of the orthographic differences between English and German, and not an explicit decision to exclude or include complex nouns.

Periti and Tahmasebi (2024), in their survey on the use of contextualized word embeddings for the LSC task, note that many approaches skip directly from semantic proximity judgments (like those of the Word-in-Context datasets) directly to the quantification of meaning shifts over time, while skipping over an intermediate task of Word Sense Induction — the explicit clustering of in-context word uses into separate senses.

3 Corpora and Annotated Data

In this section we introduce the diachronic corpora that our clustering models operate over, as well as the annotated gold-standard data that we produced to evaluate the models.

3.1 Diachronic Corpora

For both English and German, we use diachronic corpora spanning roughly 200 years. For English we use the Cleaned Corpus of Historical American English (**CCOHA**) (Davies, 2012; Alatrash et al., 2020). It contains texts dating from 1810 to the 2010s, balanced by decade as well as genre (newspapers, magazines, fiction/non-fiction books).

For German, we use the Deutsches Textarchiv (**DTA**) (Berlin-Brandenburgische Akademie der Wissenschaften, 2022). The DTA is a reference corpus of the German language, containing texts from 1472 to 1969, with a focus on the 17th through 19th centuries. It is curated[3] to balance between fiction, non-fiction, and scientific writing, but not necessarily the relative amount of text per year.

We use data from two eras in either corpus, as shown in Table 2. Each 'early' and 'late' era has a buffer period between it, to help us to strike a greater contrast between the eras, since all cross-era comparisons are at least 70 years apart.

3.2 Annotated Data

We introduce our sets of 19 English and 43 German **target compounds** (Table 4, full German set in Appendix D), which were derived from existing compositionality datasets, and describe our procedure for annotating these compounds' meanings in the diachronic corpora.

3.2.1 Noun Compounds and Compositionality

We used two sets of ratings of compositionality, i.e., the degree to which a constituent part of a compound contributes to the overall meaning of the compound, one for each language: German (868 total entries) (Schulte im Walde et al., 2016b) and English (280 total entries) (Cordeiro et al., 2019). These ratings were collected using the annotator's *overall* understanding of the terms, either without any explicit context for the German compounds, or interpolated from three example sentences for the

[3] https://www.deutschestextarchiv.de/doku/ueberblick

Corpus	Early Era	Tokens	Late Era	Tokens
CCOHA	1830–1859	51M	1980–2009	92M
DTA	1700–1759	33M	1870–1909	36M

Table 2: Descriptive statistics for diachronic corpora.

Corpus	Targ.	Ex. Pairs	Ratings (skipped)	Avg. ρ Agr.
CCOHA	19	393	2702 (111)	0.33
DTA	43	899	5144 (308)	0.28

Table 3: Annotated sentence pairs and IAA.

English compounds. Numerical ratings are given for each compound with respect to its modifier (for these languages, the first) or head (second) constituent. Higher values indicate a stronger relationship between the meaning of the constituent and the compound as a whole. Examples of highly compositional compounds (with respect to the head constituent) include *wedding day* (head: *day*), and *Seewasser* (en: seawater - lit: 'sea' + 'water', head: *Wasser*), while *nest egg* (financial savings) and *Zeughaus* (en: armory - lit: 'stuff' + 'house') are examples of non-compositional compounds with respect to their modifier.

3.2.2 Compound Relatedness Judgments

In order to evaluate our system in its ability to distinguish between multiple senses of the same term, we collected diachronic in-context ratings of compounds' meanings — see Table 1 for example sentence pairs. To collect these ratings, we used an annotation schema inspired by the DURel dataset (Schlechtweg et al., 2018)[4]. Starting from the compositionality datasets described above, we selected 19 English compounds and 43 German compounds based on a minimum frequency threshold applied to each era (10 for English and 20 for German[5]). Each use of a compound was presented to annotators including the context of the previous and subsequent sentence from the CCOHA or DTA corpora. They were asked to rate the relatedness of the use of the compounds on a scale from 1 (unrelated) to 4 (identical), with an option to skip the example pair if there was not sufficient evidence to make a decision. Sentence pairs were randomly sampled to form three groups: pairs of sentences occurring within either the early or the late era of the diachronic corpora, or sentence pairs traversing the two eras. Following Schlechtweg et al. (2018), we calculate the mean relatedness rating for each target within each era, as well as for the set of pairs annotated across eras (Table 4). Lower mean ratings within either era point to a greater degree of polysemy at that time. The difference between the late era mean and early era mean ($\Delta Late$) serves as a measure of the change in use-relatedness over time, which could be attributed to new senses being used (negative values), or their falling out of use (positive values), while higher values of the average relatedness rating of pairs that are compared across both eras ($\mu(Compare)$) indicate a weaker

Target	μ(E)	μ(L)	μ(C)	ΔL	JSD
field work	2.06	2.92	2.42	0.86	0.00
ins. company	2.77	3.52	3.21	0.76	0.13
nest egg	2.26	2.98	2.53	0.72	0.22
silver spoon	2.17	2.89	2.30	0.72	0.27
winter solstice	3.21	3.61	3.26	0.40	0.00
bank account	3.00	3.38	3.28	0.38	0.00
wedding day	3.52	3.64	3.6	0.12	0.00
brick wall	2.55	2.54	2.96	-0.01	0.21
fairy tale	2.61	2.36	2.73	-0.25	0.15
mother tongue	3.39	2.92	3.33	-0.47	0.00
love song	3.18	2.61	2.78	-0.57	0.24
balance sheet	3.38	2.76	2.96	-0.62	0.03
foot soldier	3.20	2.54	2.79	-0.66	0.16
pocket book	3.11	2.25	1.68	-0.86	0.36
market place	3.04	2.14	2.30	-0.90	0.38
gold mine	2.88	1.94	2.49	-0.93	0.33
elbow room	2.71	1.77	2.84	-0.94	0.24
ground floor	3.57	2.59	3.33	-0.99	0.18
calendar month	3.81	2.62	2.73	-1.19	0.00
Ruhestand	2.43	3.58	2.27	1.16	0.28
Rechtsstreit	2.63	3.74	3.51	1.10	0.11
Uhrwerk	2.48	3.44	2.21	0.97	0.16
Triebwerk	2.22	3.14	1.88	0.92	0.41
...	...	...	...	...	
Mauerwerk	3.11	3.16	3.22	0.05	0.03
Sonnenstrahl	2.88	2.9	3.18	0.02	0.07
Heerführer	3.37	3.36	3.57	-0.01	0.00
Sonnenlicht	3.01	2.96	3.04	-0.05	0.03
Bergwerk	3.33	3.24	2.99	-0.08	0.03
...	...	...	...	...	
Windspiel	3.08	2.43	2.81	-0.65	0.17
Sonnenblume	3.34	2.70	3.23	-0.65	0.11
Salzwasser	3.44	2.67	2.86	-0.76	0.21
Feldzug	3.53	2.46	3.42	-1.07	0.13

Table 4: List of compounds (top: all English targets; bottom: sample German targets), and average relatedness ratings (1: unrelated; 4: identical) for each era, across the two eras, and JSD across eras from clustered annotation graph. Sorted by ΔLate.

[4] We used the platform Phitag: `https://phitag.ims.uni-stuttgart.de/`

[5] Different thresholds were used because the English targets occurred less frequently than the German targets

change between the two eras.

Participants were recruited through Prolific[6] and were paid an average of £8.57/hour for their work. In total, 58 English-speaking and 107 German-speaking participants were recruited, who annotated a total of 393 English sentence pairs and 899 German sentence pairs (see Table 3 and Appendix D, including translations).

4 Clustering Pipeline

We cluster our target compounds with the goal of matching the semantic relatedness judgments. We use the k-Means algorithm[7] to cluster context-sensitive vector representations of our target terms in order to group instances with related meanings together, and to separate unrelated meanings. The value of k was estimated experimentally for each run, by using silhouette scores (Rousseeuw, 1987; Kanjirangat et al., 2020) to select a value of k between 4 and 32[8]. The vector representation and the set of related terms that are clustered along with the target compounds both vary across experiments.

4.1 Vector Representations

We experiment with two different vector representations (and their joint, concatenated use). The first representation is derived from the BERT models. As an alternative approach, we also experiment with second-order random-indexing, because we expect this representation to be more isotropic than the BERT-derived representations (Ethayarajh, 2019), potentially making them easier to cluster.

BERT Representations We domain-adapt a monolingual BERT model for German and English on the corresponding full diachronic corpus. See appendix B.1 for settings. For each target, a BERT embedding is created for each time the target appears in the relevant corpus, by feeding that sentence through the domain-adapted BERT model, and taking the first four layers of hidden states for each of the (potentially several) tokens corresponding to a particular span where the the target occurs, and averaging these together. We refer to Miletic and Schulte im Walde (2023) on noun-compound compositionality using BERT to inform our selection of layers.

Second-Order Random-Indexing We use a second, simpler vector representation (random indexing), derived from Basile et al. (2015). A base vocabulary of representations is first created by taking each vocabulary item from the BERT model's tokenizer and assigning it a sparse vector of length $1,000$, with a random assignment of ten 1s and ten -1s, with the remaining dimensions set to zero. This is represented in the third row (rand. vecs) of Figure 1. Next, the set of all *first-order* context words is gathered, by searching for each target compound's uses in the corpus, and adding the words found in a bi-directional window of 5 space-delimited (lemmatized) tokens (1st row of Figure 1). These first-order context lemmas are filtered using a method derived from (Schütze, 1998).[9] Representations are created for each first-order context lemma by again searching the full corpus to find the same 5-lemma window on either side of the first-order context term (2nd row of Figure 1), and summing the random-indexing vectors (the base vocabulary of vector representations) corresponding to each one,[10] and then averaging the vector by the number of occurrences of the term in the corpus. Turning back to the target compounds: each is ultimately represented by averaging the representations of each of its (first order) context words (in a 5-lemma window), making the representation an aggregation of the second-order contexts that these first-order contexts occur in (flowing in the reverse direction of the arrows in Figure 1). In rare cases where the target has no available first-order contexts, its representation is formed by averaging the random-indexing vectors corresponding to the target compound itself.

Combining Representations When both BERT and second-order random-indexing representations are used together, the vectors are simply concatenated. In all cases, the feature vector for each example is normalized.

4.2 Clustering Modes

Instances from both eras of the corpus are clustered together; the year that any given instance is drawn from is only used for post-hoc inference

[6] https://www.prolific.com/

[7] https://scikit-learn.org

[8] Early experiments indicated a general decrease in performance for higher values of k.

[9] We first filter the context set using a per-language stopword list from NLTK (Bird and Loper, 2004), then select the $1,500$ contexts most dependent on the presence of a target compound, using a χ^2 based criterion, and another $1,500$ lemmas selected by (highest) frequency in the first-order contexts.

[10] Note that these are subdivided using the granularity of the BERT SentencePiece tokenizer.

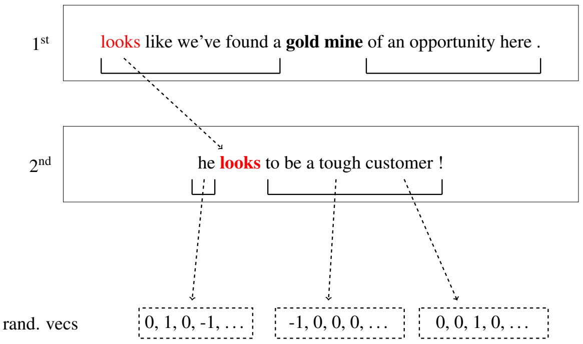
1st
looks like we've found a gold mine of an opportunity here .
2nd
he looks to be a tough customer !
rand. vecs
0, 1, 0, -1, ...
-1, 0, 0, 0, ...
0, 0, 1, 0, ...

Figure 1: Construction of second-order random-index vectors. First-order contexts of target 'gold mine' are identified. Contexts of those contexts are represented by a vocabulary of random sparse vectors.

$$\text{gold mine} + \begin{cases} \emptyset \\ \text{gold} \times 4 \\ \text{mine} \times 4 \\ \textbf{gold}\text{ chain} \times 1, \textbf{gold}\text{ dust} \times 1, \ldots \\ \text{iron }\textbf{mine} \times 1, \text{silver }\textbf{mine} \times 1, \ldots \end{cases}$$

Figure 2: Possible sampling options of related terms per instance of target *gold mine*, depending on clustering mode.

after clustering has been completed. All of our configurations include examples from the full set of target compounds, but vary in the inclusion of additional items.

The **compounds-only** setting does not include any additional items, which serves as a baseline. In the **compound-constituent** setting, we take advantage of the variable relationship that compounds' potentially numerous senses have with the meaning of their constituents (e.g. the relationship between *silver spoon* and *spoon* differs between the metaphorical and literal use of the compound), and cluster all target compounds and their constituents together. Head constituents and modifier constituents are included in separate runs.

In the **constituent family sets** (**CFS**) setting, we cluster all our target compounds together with other compound nouns that share a single constituent with a target compound, e.g. *gold mine, copper mine, coal mine* share the head constituent *mine*. As with **compound-constituent** clustering, separate runs are performed for related compounds that share modifier constituents, and those that share head constituents. The CFS compounds were obtained by searching the corpora for all noun compounds that contained the corresponding constituent, filtered to include only sequences of exactly two nouns, not preceded or followed by another noun. Some cleaning of the resulting list of potential compounds was performed to remove words that are not noun compounds, e.g. *Beispiel* is a combination of the preposition *bei* with the noun *Spiel* (game), and is thus rejected. In total, 103 shared-constituent compounds were gathered for English, and 640 for German. Although we do not have relatedness ratings for these sets of related compounds, we are interested in using them as an alternative foil to the compounds' constituents, to contrast between senses like the metaphorical meaning of *gold mine*: source of abundance, or financial opportunity, which is not shared by other *mine* compounds.

Sampling Procedure Due to time complexity constraints for clustering large numbers of examples, and also frequency asymmetry between many compounds and their much more frequent constituents, we use the following sampling procedure to prevent the target compounds from being overwhelmed by non-target examples, and to keep the total number of examples manageable.

We target a maximum number of instances, 15k. In order to be able to evaluate the clustering performance over pairs of sentences that were annotated, we first include all such pairs as a (nearly) uniform initial sample. Further uses are sampled until the maximum is reached, by first randomly selecting an era (weighted by total tokens in that sub-corpus), and then randomly selecting a target compound (weighted by the target's frequency in that era). Every time a compound target, e.g. *gold mine*, is sampled in this way, one of the following types of additional uses is sampled (see Figure 2): modifier constituent (*gold*), head constituent (*mine*), compounds from the target's modifier family set (one each: *gold chain, gold dust, ...*), compounds from the target's head constituent family set (one each: *iron mine, silver mine, ...*). As the constituent words tend to be more numerous, they are sampled at a ratio of four constituents for each target compound.

Experimental Configurations We experimentally altered the following parameters and settings in our clustering runs: the vector representation: either BERT vectors, second-order random-indexing vectors, or both representations concatenated together. Additionally, the following five settings were used to include additional targets which were clustered alongside the target compounds: no additional targets (targets$_{compounds}$), head constituents (targets$_{head-const}$), modifier constituents (targets$_{mod-const}$), compounds in the same head constituent family set(targets$_{head-cfs}$), or compounds in the same modifier constituent family set (targets$_{mod-cfs}$).

5 Evaluation

5.1 Evaluating via Relatedness Ratings

We compare clustering runs against pairs of sentences rated for the relatedness of the meanings of target compounds. This involves two simplifying assumptions to enable direct comparisons: binarizing the 1-4 relatedness rating given by the annotators, and simplifying the representational space of a clustering run by binarizing the outcome of clustering: either a pair of examples is in the same cluster, or not. We use accuracy scores to treat true positives with the same weight as true negatives (it is just as important that unrelated examples are in different clusters, as it is that related examples are clustered together).

A secondary perspective on the annotated sentences was achieved by using the WUG (Word Usage Graph) tool (Schlechtweg et al., 2021; Schlechtweg, 2023; Schlechtweg et al., 2024) to cluster the annotated sentence pairs such that pairs that were mutually rated as similar tend to form clusters, and pairs rated maximally dissimilar are not clustered together, using a variation of correlation clustering (Bansal et al., 2004). This method gives us a view of the desired structure of clusters. See Figure 3 in Appendix D for a clustering example.

We evaluate our system clusters against these annotation clusters using the v-Measure (Rosenberg and Hirschberg, 2007), the harmonic mean of homogeneity (extent to which clusters contain only one class) and completeness (extent to which all members of a gold class are clustered together) – analogous to the f-measure which combines the contributions of precision and recall.

5.2 Evaluating Semantic Change Measures

We compare several methods for measuring semantic change in the aggregate for each of our target compounds (internal measures) against *aggregate* measures of change derived from our dataset of relatedness ratings (external measures).

5.2.1 Internal Change Measures

PRT Our first approach to comparing representations between eras is to use the cosine distance between *prototype* (PRT) representations for each era, where the prototypes are obtained by averaging the representations for each target's instances in the era. This serves as a measure of semantic change (Giulianelli et al., 2020; Martinc et al., 2020) with relatedness ranging from 0 (identical) to 1 (orthogonal), i.e., relatedness decreases with greater distance.

APD Following Giulianelli et al. (2020), we also compute the average pairwise distance (APD) of all individual representations of compounds or constituents paired across the two eras, using cosine distances. This measure has the same range from 0–1 as the PRT measure, and lower distances indicate greater relatedness between the two eras.

Jensen-Shannon Divergence We can characterize the cluster membership of each target as a probability distribution across k clusters. This allows us to use the Jensen-Shannon divergence (JSD) as a measure of the difference between two distributions, yielding a value between 0 (identical distributions) and 1 (maximally unrelated distributions). Our intuition is that a term whose sense distribution does not change over time should have its instances in each era clustered into a similar distribution of clusters, or vice versa, and that this aggregate measure of the divergence between two distributions reflects this (dis)continuity.

We compute the JSD between the cluster distributions for each target's instances in time t_{early} and in time t_{late}. We refer to this as $JSD(t_{early}, t_{late})$.

5.2.2 External Change Measures

The above internal change measures are compared against the delta later ($\Delta(L)$) and average compare ($\mu(C)$) per target ratings introduced in section 3.2.2. We collapse the two kinds of change (innovative and reductive) represented by the $\Delta(L)$ into one by taking its absolute value. We also compare against the compositionality ratings, taking them as an indirect measure of change, by relying on the hypothesis that noun compounds tend toward non-compositionality over time (Bybee, 2015). In this way, we expect to see a negative correlation between increased compositionality and greater semantic change. Finally, for each target, we calculate the JSD for the distribution of clustered annotated uses for the early and late time periods, reproducing the graded change measure from SemEval 2020 Task 1 (Schlechtweg et al., 2020). All external change measures other than compositionality are shown in Table 4.

6 Results and Discussion

6.1 Evaluation via Relatedness Ratings

In Tables 5 and 6, we report the minimum, maximum, and average scores across experimental settings, holding either the type of vector representation or target set constant.

Pairwise Accuracy We evaluate our models' performance over the set of annotated sentence pairs (Table 5). For both languages, runs using the second-order random-indexing representations alone scored better than the BERT representations, but this difference was more pronounced for the German experiments. The concatenation of the two representations (Vecs$_{both}$) produced very similar results as using BERT representations alone. Across both languages, configurations not involving any additional targets (targets$_{compounds}$) were either the best performing (English) or near-best (German). For English, the best configurations using additional targets included the head-CFS, and for German the modifier constituents.

Comparing System and Annotated Clusters For the English experiments in Table 6, all three types of representations had the same maximum V-measure score of 0.39, while configurations that used BERT vectors had a slightly higher average V-measure. Configurations that included additional clustering targets scored higher than clustering compounds alone, with the best results (0.39) obtained with the head-CFS configurations.

The German experiments in Table 6 show a large contrast between the configurations using BERT vectors and those without, the former scoring about twice as high. V-measure scores according to items included in clustering were largely similar, with the exception of mod-CFS, which were 0.10 worse on average (at 0.19) than the other configurations.

We found a strong, highly-significant negative correlation ($\rho = -0.84$) between the accuracy and V-measure scores for the German experiments. No significant correlation was found for the English experiments. This discrepancy confirms our interest in contrastive evaluation criteria, and warrants further investigation, e.g., by exhaustively annotating all example pairs for a single target, to see if accuracy and V-measure scores would converge. The much larger performance gap between 2nd-order random-indexing and BERT vectors for German we attribute to sub-word tokenization issues arising from the German compounds' greater lengths.

6.2 Correlation with Change Measures

Taking the highest performing configurations in terms of accuracy and V-measure (separately) for each language, we evaluate the clustering results from each configuration for how well they correlate with measures of semantic change. We operationalize change in this way to capture distinctions between targets that are *more* stable or *more* different, recognizing that there are many possible differences between our diachronic subcorpora external to the semantic change or stability of our target compounds. In Table 7, we report the Spearman's ρ correlation between four external measures of change derived from the relatedness ratings, and three internal measures.

All internal measures have the same relatedness polarity. The expected correlation direction with the external measures is: positive for $|\Delta(L)|$ and JSD_{anno}, negative for $\mu(C)$ and the compositionality ratings.

Best Configurations by Accuracy Table 7 reports correlations obtained from the results of the two best performing configurations, ranked by pairwise accuracy on the annotated example pairs, one for each language. Both configurations used 2nd-order random-indexing representations, and had a low best-k of 4, while they differed in that the German configuration included head constituent family sets, while the best English run clustered the target compounds alone – as such, it is not evaluated against compositionality ratings, since these depend on a particular constituent.

All of the internal change measures exhibited a significant correlation with the $\mu(Compare)$ external change measure in English, where the strongest correlation of $\rho = -0.64$ was observed with the APD metric. APD, however, was not significantly correlated with the English $|\Delta(Late)|$ measure, where both the PRT measure and the JSD of target compounds across the two eras did show small but significant correlations.

For the German configuration, the strongest significant correlations between internal and external change measures were weaker ($\rho = 0.37$ and

$\rho = 0.35$), between the PRT and the $|\Delta(Late)|$ and compositionality ratings, respectively. A similar correlation of $\rho = 0.32$ was found between the APD and compositionality ratings, however, the polarity of the correlations with compositionality ratings are contrary to our expectations. This may be due to the nearly 200 year gap between the end of the DTA data and the compositionality ratings, underscoring the need for in-context compositionality annotations.

Best Configurations by V-Measure Regarding the highest V-measure scores in Table 7 comparing output clusters with the clustered graph of relatedness annotations, only two pairs of internal and external measures of change were found to have a significant Spearman's ρ correlation, both involving the English configuration: PRT and $\mu(C)$ and APD and $\mu(C)$. The magnitude of these correlations was stronger than their equivalent from the best English system in terms of pairwise accuracy, but the significance level was lower.

Sense Aggregation The measures more sensitive to separate senses (JSD of system clusters, JSD of clustered annotations) were generally worse than internal measures of change that aggregated the representations. The small correlation seen with the best English system in terms of accuracy's $JSD_{(e,l)}$ measure against $\mu(C)$ and $|(\Delta(L))|$ motivates future work to improve the representations in terms of dis-aggregated senses.

	en			de		
Setting	Min.	Max.	Avg.	Min.	Max.	Avg.
Vecs$_{BERT}$	0.40	0.44	0.42	0.28	0.31	0.29
Vecs$_{2nd}$	0.44	0.50	**0.47**	0.45	0.55	**0.51**
Vecs$_{both}$	0.40	0.47	0.43	0.27	0.30	0.28
targets$_{compounds}$	0.43	0.50	**0.46**	0.27	0.54	0.37
targets$_{head-const}$	0.42	0.46	0.44	0.28	0.47	0.35
targets$_{mod-const}$	0.40	0.49	0.43	0.30	0.55	**0.39**
targets$_{head-cfs}$	0.44	0.47	0.45	0.27	0.55	0.37
targets$_{mod-cfs}$	0.40	0.44	0.42	0.29	0.45	0.34

Table 5: Accuracy scores for clustering of annotated sentence pairs.

	en			de		
Setting	Min.	Max.	Avg.	Min.	Max.	Avg.
Vecs$_{BERT}$	0.26	0.39	**0.34**	0.21	0.38	0.33
Vecs$_{2nd}$	0.22	0.39	0.28	0.14	0.21	0.17
Vecs$_{both}$	0.26	0.39	0.33	0.23	0.39	**0.35**
targets$_{compounds}$	0.22	0.32	0.28	0.17	0.39	0.31
targets$_{head-const}$	0.27	0.36	0.33	0.21	0.38	**0.32**
targets$_{mod-const}$	0.22	0.37	0.32	0.17	0.37	0.29
targets$_{head-cfs}$	0.39	0.39	**0.39**	0.16	0.38	0.30
targets$_{mod-cfs}$	0.26	0.29	0.27	0.14	0.23	0.19

Table 6: Mean V-measure scores for each target, comparing system and annotated clusters.

	$\mu(C)$		$\|\Delta(L)\|$		Compos.		JSD_{anno}	
	acc	vm	acc	vm	acc	vm	acc	vm
English								
PRT	-0.54^{*}	-0.70^{*}	0.47^{*}	0.08		-0.44	0.17	0.55
APD	-0.64^{**}	-0.70^{*}	0.32	0.23		-0.55	0.45	0.65
$JSD_{e,l}$	-0.50^{*}	-0.63	0.51^{*}	-0.05		-0.14	0.22	0.41
German								
PRT	0.08	0.07	0.37^{*}	0.22	0.35^{*}		0.03	0.05
APD	0.11	0.19	0.26	0.19	0.32^{*}		-0.05	-0.01
$JSD_{e,l}$	-0.12	-0.01	0.18	0.14	0.20		-0.11	0.15

Table 7: Spearman's ρ correlations between internal (rows) and external (columns) semantic change measures. Best configs per language in terms of accuracy (acc) and V-measure (vm):
Accuracy: English: best k:4, $_{2nd-order}$,$_{compounds}$. German: best k:4, $_{2nd-order}$,$_{head-cfs}$;
V-measure: English: best k:8, $_{2nd-order}$;$_{head-cfs}$. German: best k:30, $_{both}$,$_{compounds}$
$*: p < 0.05$, $**: p < 0.01$

6.3 Conclusion

In this paper we introduced a novel, sense-targeting clustering approach and a novel dataset of relatedness judgments of English and German noun-compounds, evaluating contrasting vector representations without prematurely collapsing senses into prototypes, using a range of contrastive evaluation criteria. We find reason to continue exploring the use of simpler vector representations in data-limited historical settings, due to their higher performance in terms of pairwise accuracy, but evaluation in terms of V-measure caution against over-interpreting this result. The inclusion of constituents or constituent-family sets in clustering did not result in a decisive improvement over clustering with compounds alone, however, both the most accurate German configuration and the English configuration with the highest average V-measure included head-CFS, suggesting that further refinement of target selection and sampling may improve this approach. All internal measures of semantic change were correlated with English external change measures, but they largely failed to align for German, in spite of both languages' best configurations having similar accuracy scores. Overall, we recommend relying on $|\Delta(L)|$ as an external measure of change, and find prototype and average pair-wise distances to be reliable default options for internal measures.

Limitations

The CCOHA does not contain any orthographic normalization layer, like the DTA does (see 3.1), so it is possible that we fail to consider some uses of our target items that are merely spelled in a non-standard way. That the CCOHA data is relatively modern, and only sourced from American-English texts should reduce (but not completely eliminate) the effects of purely orthographic variation on our analysis.

Acknowledgements

Our work was supported by the DFG Research Grant SCHU 2580/5-1 (*Computational Models of the Emergence and Diachronic Change of Multi-Word Expression Meanings*).

References

Reem Alatrash, Dominik Schlechtweg, Jonas Kuhn, and Sabine Schulte im Walde. 2020. CCOHA: Clean corpus of historical American English. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6958–6966, Marseille, France. European Language Resources Association.

Pegah Alipoor and Sabine Schulte im Walde. 2020. Variants of vector space reductions for predicting the compositionality of English noun compounds. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4379–4387, Marseille, France. European Language Resources Association.

Alina Arseniev-Koehler. 2021. Theoretical foundations and limits of word embeddings: what types of meaning can they capture? *CoRR*, abs/2107.10413.

Nikhil Bansal, Avrim Blum, and Shuchi Chawla. 2004. Correlation clustering. *Mach. Learn.*, 56(1–3):89–113.

Marco Baroni, Raffaella Bernardi, and Roberto Zamparelli. 2014. Frege in space: A program for composition distributional semantics. *Linguistic Issues in Language Technology*, 9.

Pierpaolo Basile, Annalina Caputo, and Giovanni Semeraro. 2015. Temporal random indexing: A system for analysing word meaning over time. *Italian Journal of Computational Linguistics*, 1:55–68.

Laurie Bauer. 2017. *Compounds and Compounding*. Cambridge Studies in Linguistics. Cambridge University Press.

Berlin-Brandenburgische Akademie der Wissenschaften. 2022. Deutsches Textarchiv. Grundlage für ein Referenzkorpus der neuhochdeutschen Sprache. https://www.deutschestextarchiv.de/.

Steven Bird and Edward Loper. 2004. NLTK: The natural language toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 214–217, Barcelona, Spain. Association for Computational Linguistics.

Joan Bybee. 2015. *Language Change*. Cambridge Textbooks in Linguistics. Cambridge University Press.

Silvio Cordeiro, Aline Villavicencio, Marco Idiart, and Carlos Ramisch. 2019. Unsupervised compositionality prediction of nominal compounds. *Computational Linguistics*, 45(1):1–57.

Mark Davies. 2012. Expanding horizons in historical linguistics with the 400-million word corpus of historical american english. *Corpora*, 7:121–157.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Prajit Dhar, Janis Pagel, and Lonneke van der Plas. 2019. Measuring the compositionality of noun-noun compounds over time. In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, pages 234–239, Florence, Italy. Association for Computational Linguistics.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 55–65, Hong Kong, China. Association for Computational Linguistics.

J. Firth. 1957. *Papers in Linguistics, 1934-1951, by J.r. Firth*. Oxford University Press.

Mario Giulianelli, Marco Del Tredici, and Raquel Fernández. 2020. Analysing lexical semantic change with contextualised word representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3960–3973, Online. Association for Computational Linguistics.

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.

Zellig S Harris. 1954. Distributional structure. *Word*, 10(2-3):146–162.

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.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing.

Samin Mahdizadeh Sani, Malak Rassem, Chris W. Jenkins, Filip Miletić, and Sabine Schulte im Walde. 2024. What can diachronic contexts and topics tell us about the present-day compositionality of English noun compounds? In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 17449–17458, Torino, Italia. ELRA and ICCL.

Matej Martinc, Petra Kralj Novak, and Senja Pollak. 2020. Leveraging contextual embeddings for detecting diachronic semantic shift. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4811–4819, Marseille, France. European Language Resources Association.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space.

Filip Miletic and Sabine Schulte im Walde. 2023. A systematic search for compound semantics in pretrained BERT architectures. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1499–1512, Dubrovnik, Croatia. Association for Computational Linguistics.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

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.

Francesco Periti and Nina Tahmasebi. 2024. A systematic comparison of contextualized word embeddings for lexical semantic change. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4262–4282, Mexico City, Mexico. Association for Computational Linguistics.

Mohammad Taher Pilehvar and Jose Camacho-Collados. 2019. WiC: the word-in-context dataset for evaluating context-sensitive meaning representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1267–1273, Minneapolis, Minnesota. Association for Computational Linguistics.

Alessandro Raganato, Tommaso Pasini, Jose Camacho-Collados, and Mohammad Taher Pilehvar. 2020. XL-WiC: A multilingual benchmark for evaluating semantic contextualization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7193–7206, Online. Association for Computational Linguistics.

Siva Reddy, Diana McCarthy, and Suresh Manandhar. 2011. An empirical study on compositionality in compound nouns. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 210–218, Chiang Mai, Thailand. Asian Federation of Natural Language Processing.

Andrew Rosenberg and Julia Hirschberg. 2007. V-measure: A conditional entropy-based external cluster evaluation measure. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 410–420, Prague, Czech Republic. Association for Computational Linguistics.

Peter J. Rousseeuw. 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65.

Dominik Schlechtweg. 2023. *Human and Computational Measurement of Lexical Semantic Change*. Ph.D. thesis, University of Stuttgart, Stuttgart, Germany.

Dominik Schlechtweg, Pierluigi Cassotti, Bill Noble, David Alfter, Sabine Schulte Im Walde, and Nina Tahmasebi. 2024. More DWUGs: Extending and evaluating word usage graph datasets in multiple languages. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 14379–14393, Miami, Florida, USA. Association for Computational Linguistics.

Dominik Schlechtweg, Anna Hätty, Marco Del Tredici, and Sabine Schulte im Walde. 2019. A wind of change: Detecting and evaluating lexical semantic change across times and domains. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 732–746, Florence, Italy. Association for Computational Linguistics.

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.

Dominik Schlechtweg, Sabine Schulte im Walde, and Stefanie Eckmann. 2018. Diachronic usage relatedness (DURel): A framework for the annotation of lexical semantic change. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 169–174, New Orleans, Louisiana. Association for Computational Linguistics.

Dominik Schlechtweg, Nina Tahmasebi, Simon Hengchen, Haim Dubossarsky, and Barbara McGillivray. 2021. DWUG: A large resource of diachronic word usage graphs in four languages. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7079–7091, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Sabine Schulte im Walde, Anna Hätty, and Stefan Bott. 2016a. The role of modifier and head properties in predicting the compositionality of English and German noun-noun compounds: A vector-space perspective. In *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*, pages 148–158, Berlin, Germany. Association for Computational Linguistics.

Sabine Schulte im Walde, Anna Hätty, Stefan Bott, and Nana Khvtisavrishvili. 2016b. GhoSt-NN: A representative gold standard of German noun-noun compounds. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2285–2292, Portorož, Slovenia. European Language Resources Association (ELRA).

Hinrich Schütze. 1998. Automatic word sense discrimination. *Computational Linguistics*, 24(1):97–123.

A Corpora Details

A.1 CCOHA

Since the dataset includes texts from the more recent past, its curators redact 10 tokens after every 200 tokens, to comply with copyright restrictions.

A.2 DTA

We use orthographic modernization provided in the DTA to avoid string-mapping errors (e.g. mapping uses of ‘ſ’ to the modern ‘s’).

B Configurations

B.1 Domain Adaptation

For both languages, we domain-adapted the base models for 3 epochs over the respective corpora. We used a learning rate of 5e-5, masking 15% of tokens, 768-dimensional hidden layer, vocabulary of 30k, with a maximum sequence length of 128 tokens. We used a Nvidia GeForce RTX A6000 GPU, running for approximately 130 hours (English) and 43 hours (German).

We use `bert-base-german-cased` (`https://www.deepset.ai/german-bert`) and `bert-base-uncased` (Devlin et al., 2019) (`https://huggingface.co/bert-base-uncased`) for German and English, respectively. Each model is domain-adapted on the full diachronic corpus using lemmatized text, to reduce the set of possible string representations of our target items (due to e.g. inflections for plural use).

B.2 Packages, versions

We used the following Python (version 3.10.4) packages (`name`, version):
`nltk`: 3.7, `numpy`: 1.23.3, `scikit-learn`: 1.2.2, `scipy`: 1.10.1, `spacy`: 3.5.0, `torch`: 1.13.0, `transformers`: 4.24.0.

C Instructions for Annotators

C.1 English

You will be presented with a series of examples of the same target word used in two different contexts. Each target word is highlighted in green, and the sentence containing the target word is bordered by ☞ ☜ icons. The task is to rate each pair of contexts for how similar the meanings of the green-highlighted words are (in these uses):

- from 1 (unrelated)
- to 4 (identical)

There is a separate rating for situations where you can’t decide – this could happen if there is not enough context to tell what is meant by an example.

C.2 German (translated from English)

Sie werden eine Reihe von Beispielen sehen, in denen dasselbe Zielwort in zwei verschiedenen Kontexten verwendet wird. Jedes Zielwort ist grün hervorgehoben, und der Satz, der das Zielwort enthält, ist mit ☞ ☜ -Symbolen umrandet. Die Aufgabe ist, jedes Paar von Kontexten nach der Ähnlichlichkeit der Bedeutungen der grün hervorgehobenen Wörter zu bewerten (in diesen Verwendungen):

- von 1 (kein Bezug)
- bis 4 (identisch)

Es gibt eine gesonderte Bewertung für Situationen, in denen man sich nicht entscheiden kann – dies kann passieren, wenn nicht genügend Kontext vorhanden ist, um zu erkennen, was mit einem Beispiel gemeint ist.

D Annotated Examples

Sentence pairs to be annotated were randomly sampled from the corpora. Each pair was rated by 5-10 annotators; 10 English annotators’ and 34 German annotators’ ratings were excluded due to their failure across a set of three hand-written quality-control questions or (pairwise, averaged) inner-annotator agreement lower than Spearman’s $\rho = 0.1$. The average inner-annotator agreement of $\rho = 0.33$ (English) and $\rho = 0.28$ (German) should be considered in light of the overall quantity of the annotators, and their lack of professional linguistic training, when compared with e.g. the annotators of DURel (Schlechtweg et al., 2018).

Target	μ(Early)	μ(Late)	μ(Compare)	ΔLate	$JSD(_{e,l})$
Ruhestand (*retirement*)	2.43	3.58	2.27	1.16	0.28
Rechtsstreit (*legal dispute*)	2.63	3.74	3.51	1.10	0.11
Uhrwerk (*clockwork*)	2.48	3.44	2.21	0.97	0.16
Triebwerk (*engine*)	2.22	3.14	1.88	0.92	0.41
Murmeltier (*marmot*)	2.21	3.06	2.58	0.85	0.32
Eisenwerk (*iron works*)	2.84	3.48	2.06	0.65	0.27
Trauerspiel (*tragedy*)	2.64	3.28	2.55	0.64	0.19
Stückwerk (*piece work*)	2.31	2.70	2.53	0.39	0.38
Streitsache (*litigation*)	3.03	3.42	3.30	0.39	0.00
Zeughaus (*armory*)	2.58	2.94	2.73	0.37	0.21
Gesichtszug (*facial expression*)	3.11	3.45	3.40	0.34	0.11
Feuerwerk (*firework*)	2.86	3.20	2.45	0.34	0.24
Kartenspiel (*card game*)	3.03	3.34	3.22	0.31	0.03
Wortspiel (*wordplay*)	3.19	3.42	3.15	0.23	0.03
Schauspiel (*play (theater)*)	2.52	2.75	2.39	0.23	0.27
Hausstand (*household*)	2.46	2.68	2.23	0.22	0.47
Grundfläche (*footprint (floor)*)	3.21	3.4	3.14	0.20	0.10
Meerwasser (*seawater*)	3.14	3.31	3.25	0.17	0.11
Eifersucht (*jealousy*)	3.15	3.32	2.97	0.16	0.20
Sündenfall (*lapse (religious)*)	2.94	3.06	2.93	0.11	0.07
Sonnenuhr (*sundial*)	3.58	3.65	3.24	0.08	0.03
Tagewerk (*day's work*)	2.35	2.41	2.52	0.06	0.28
Mauerwerk (*masonry*)	3.11	3.16	3.22	0.05	0.03
Sonnenstrahl (*sunbeam*)	2.88	2.9	3.18	0.02	0.07
Heerführer (*general (military)*)	3.37	3.36	3.57	-0.01	0.00
Sonnenlicht (*sunlight*)	3.01	2.96	3.04	-0.05	0.03
Bergwerk (*mine*)	3.33	3.24	2.99	-0.08	0.03
Ziegenbock (*male goat*)	2.42	2.32	2.73	-0.10	0.27
Sonnenschein (*sunshine*)	3.28	3.18	3.02	-0.10	0.17
Kreuzzug (*crusade*)	3.21	3.10	2.94	-0.11	0.17
Brunnenwasser (*well water*)	3.31	3.20	3.46	-0.12	0.07
Seewasser (*seawater*)	3.23	3.09	2.95	-0.14	0.20
Zitronensaft (*lemon juice*)	3.91	3.75	3.78	-0.16	0.11
Stockwerk (*story / floor*)	3.38	3.17	2.99	-0.21	0.07
Kinderspiel (*child's play / something easy*)	2.35	2.05	2.15	-0.30	0.36
Wunderwerk (*marvel, miracle*)	2.87	2.54	2.41	-0.33	0.38
Bildhauer (*sculptor*)	3.50	3.15	3.55	-0.36	0.00
Leinöl (*flax seed oil*)	3.59	3.19	3.44	-0.39	0.00
Rehbock (*male deer*)	3.08	2.61	2.81	-0.47	0.14
Windspiel (*wind chimes*)	3.08	2.43	2.81	-0.65	0.17
Sonnenblume (*sunflower*)	3.34	2.70	3.23	-0.65	0.11
Salzwasser (*salt water*)	3.44	2.67	2.86	-0.76	0.21
Feldzug (*campaign*)	3.53	2.46	3.42	-1.07	0.13

Table 8: Full list of German target compounds, and average relatedness ratings (1: unrelated; 4: identical) for each era, across the two eras, and JSD across eras from clustered annotation graph. Sorted by ΔLate. Reference English translations only for predominant sense.

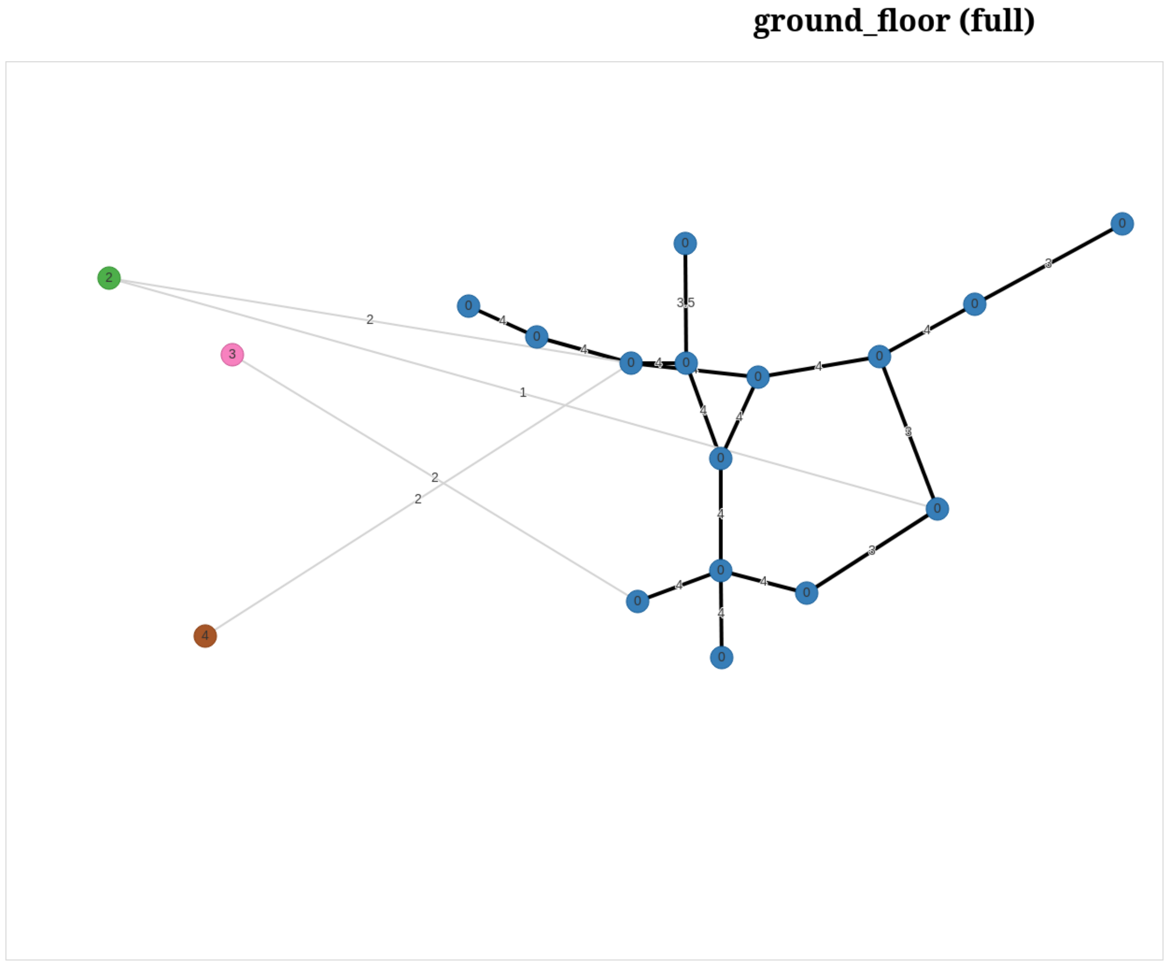

Figure 3: Example Correlation Clustering of annotation graph by WUG tool. Each node represents an example sentence where 'ground floor' was used. Numbers in each node are cluster labels and numbers on edges are aggregated relatedness values.

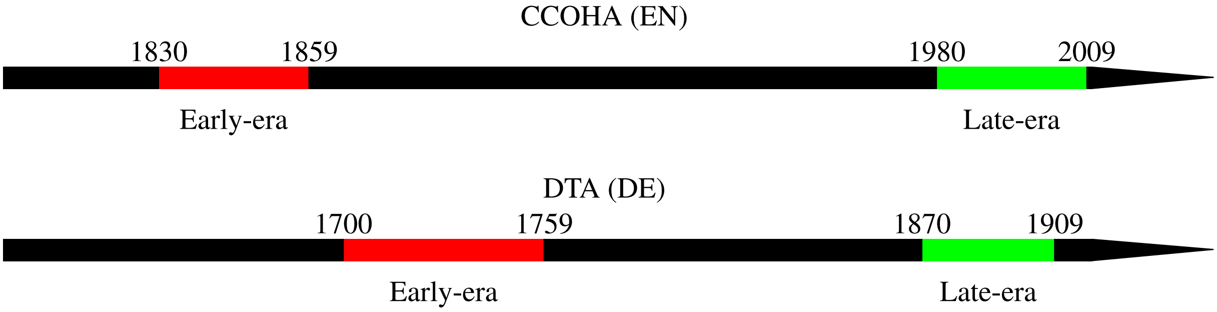

Figure 4: Timeline with full range and selected early and late eras for each corpus.