Modeling the Evolution of English Noun Compounds with Feature-Rich Diachronic Compositionality Prediction

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

Abstract

We analyze the evolution of English noun compounds, which we represent as vectors of time-specific values. We implement a wide array of methods to create a rich set of features, using them to classify compounds for present-day compositionality and to assess the informativeness of the corresponding linguistic patterns. Our best results use BERT – reflecting the similarity of compounds and sentence contexts – and we further capture relevant and complementary information across approaches. Leveraging these feature differences, we find that the development of low-compositional meanings is reflected by a parallel drop in compositionality and sustained semantic change. The same distinction is echoed in transformer processing: compositionality estimates require far less contextualization than semantic change estimates.

1 Introduction

Noun compounds are composed of at least two constituents, whose relatedness to the overall meaning ranges from compositional (e.g., *love song*) to idiomatic (e.g., *glass ceiling*). Given their ubiquitous nature and downstream challenges for NLP systems, compounds have been extensively modeled in present-day language (for overviews, see e.g. Baldwin and Kim, 2010; Miletić and Schulte im Walde, 2024). But what happened in the evolution of *glass ceiling* that triggered its idiomatic usage denoting societal barriers to success? And does this process differ from the development of *love song*, which still straightforwardly refers to a song about love? With only limited empirical evidence available to date, these questions remain largely open in historical linguistics. Moreover, since compounding is a highly productive process, understanding how it unfolds is also relevant in applications such as adapting NLP systems to evolving language use.

Empirical research directly addressing these issues is limited to a small number of studies suggesting that the evolution of features such as frequency, productivity, and time-specific cooccurrences is predictive of present-day compositionality (Dhar et al., 2019; Maurer et al., 2023; Mahdizadeh Sani et al., 2024). But none of these studies systematically compared linguistic features of different complexity; used the more versatile neural meaning representations; or directly estimated degrees of lexical semantic change (cf. Tahmasebi et al., 2021), a broader phenomenon closely connected to changes in compositionality (Bybee, 2015). As a result, we still lack a comprehensive understanding of diachronic compositionality evolution.

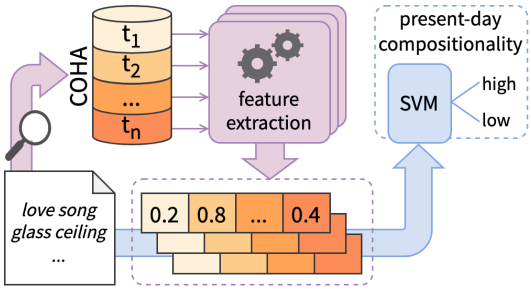

Figure 1: We analyze the temporal evolution of noun compound compositionality. Given a compound's uses from COHA and a feature extraction method, we represent the compound as a vector of time-specific values. These vectors are used to classify the target compounds for present-day compositionality and thereby assess the informativeness of the linguistic features we compare.

Our paper aims to robustly identify empirical linguistic factors explaining different meaning patterns in noun compound evolution. We also examine the effect of common modeling choices on the quality of our linguistically motivated feature set. We pose the following research questions:

RQ1 Which diachronic properties are predictive of present-day compositionality?

RQ2 What is the relationship between compositionality evolution and semantic change?

RQ3 How robust are different features to changes in data settings and modeling strategies?

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 20071–20092

Figure 1 presents a high-level overview of our approach centered on binary classification of present-day compositionality. Our primary objective is not to achieve high classification accuracy, but rather to assess the predictive power of the *evolution* of different types of linguistic information. We use various modeling approaches to define a rich feature set which is linguistically motivated and highly variable in complexity, going from raw corpus information to diverse strategies of deriving semantic knowledge from BERT-family models. We test over 7,000 constellations of experimental settings.

Our contributions are as follows. (1) An in-depth analysis of diachronic linguistic features for compositionality prediction, highlighting the strong informativeness of context-specified BERT information (0.849) but also the complementarity of simpler approaches. (2) A direct comparison of compositionality and semantic change estimates, with novel empirical insights into their joint role in the development of low-compositional meanings. (3) An analysis of feature robustness, notably linking compositionality and semantic change to diverse degrees of contextualization in transformer models.[1]

2 Related Work

Computational models of noun compound compositionality operationalize the relatedness of the constituents to the overall compound meaning. This is usually done by comparing their respective representations in word embedding models (Reddy et al., 2011; Schulte im Walde et al., 2013, 2016; Salehi et al., 2014, 2015; Cordeiro et al., 2019; Alipoor and Schulte im Walde, 2020). A more recent line of work uses noun compound compositionality to examine the linguistic knowledge in pretrained language models (Shwartz and Dagan, 2019; Garcia et al., 2021a,b; Dankers et al., 2022; Miletić and Schulte im Walde, 2023; Buijtelaar and Pezzelle, 2023), generally finding compositionality information to be encoded but also strongly localized.

Diachronic linguistic research proposes specific trajectories of change in compound meanings (e.g., from compositional to non-compositional; Bybee, 2015), but very few computational studies model this phenomenon. Dhar and van der Plas (2019) classify novel noun compounds for plausibility based on diachronic features. Dhar et al. (2019) use time-specific cooccurrences to predict present-day compositionality in a regression task. Maurer et al. (2023) and Mahdizadeh Sani et al. (2024) frame the task as binary classification and respectively use diachronic frequency and productivity features, and time-specific cooccurrence and topic models. All these studies find that diachronic evolution of features is predictive of present-day compositionality, but the extent of this trend is method-dependent. This points to the need for a controlled evaluation of different, and more recent, methods.

More general research on semantic change detection has proposed a variety of methods (Gulordava and Baroni, 2011; Hamilton et al., 2016a,b; Schlechtweg et al., 2019; Giulianelli et al., 2020; Gonen et al., 2020; Rosin et al., 2022; Cassotti et al., 2023, i.a.), but they have not been applied to multiword expressions such as noun compounds. Moreover, several transformer models have been pretrained on historical data (Hosseini et al., 2021; Manjavacas and Fonteyn, 2022; Schweter et al., 2022), making them potentially useful for semantic change tasks; they are yet to be evaluated at scale.

3 Data

As corpus data, we use the clean version of the Corpus of Historical American English (CCOHA; Davies, 2012; Alatrash et al., 2020), a genre-balanced collection of texts from 1810 to 2009 with ≈430m words. The first two decades are discarded due to limited size. We use the lemmatized, POS-tagged version. Further, we assess the informativeness of different available amounts of data by defining two types of time slices: **fine-grained** (one decade: 1830–1839, 1840–1849, etc.) and **coarse-grained** (three decades: 1830–1859, 1860-1889, etc.). A higher temporal resolution should reflect language changes more precisely; on the other hand, some methods are sensitive to limited data, so aggregating periods could be more robust. See Appendix A for sizes of individual time slices.

We use gold standard compositionality ratings for noun compounds by Reddy et al. (2011) and Cordeiro et al. (2019). Each item has three ratings: compound-level compositionality, meaning contribution of the modifier, and meaning contribution of the head (examples in Table 1). Raters gave type-level literality judgments from 0 (not at all literal) to 5 (very literal). We start from the 210 noun–noun compounds in the datasets, and retain the 166 which appear in our corpus at least once in at least two successive time slices of both granularities.

[1]The code used for our analyses is available at https://github.com/FilipMiletic/CompoundEvolution/

Item	Modifier			Head			Compound		
time difference	4.8	±0.5	HI	4.9	±0.2	HI	4.9	±0.2	HI
health care	4.7	±0.5	HI	3.2	±1.9	LO	4.7	±0.4	HI
loan shark	4.8	±0.4	HI	0.4	±0.7	LO	0.9	±1.1	LO
silver lining	0.2	±0.4	LO	0.2	±0.4	LO	0.2	±0.4	LO

Table 1: Sample items with compositionality information on the level of the modifier, head, and compound: mean gold standard ratings, standard deviations, and binary labels (HI: high comp., LO: low comp.).

4 Experimental Setup

We frame compositionality prediction as a binary classification problem. For each of three types of compositionality ratings (on the level of the compound, the modifier, and the head), we rank the compounds by that rating, retaining the first 60 and the last 60 as the low-compositional and high-compositional extremes (cf. sample targets in Table 1 and further details in Appendix B). This provides a balanced dataset that also avoids issues with average ratings in the mid-range, which often reflect strong annotator disagreement rather than the actual mid-range property of the scale (Pollock, 2018; Knupleš et al., 2023). The same approach is used in prior work on compositionality (Shwartz and Dagan, 2019; Maurer et al., 2023; Mahdizadeh Sani et al., 2024).

To assess the predictive power of the *evolution* of different linguistic properties, we represent each compound with a diachronic feature vector whose dimensions correspond to time slices. Given a feature, each dimension of the diachronic vector corresponds to that feature's value in one time slice. We define a linguistically motivated set of features which describe distinct and interpretable aspects of a compound's usage at a given point in time. We start from direct features (§4.1), i.e., empirical corpus information quantifying compound and constituent rate of use. On a more complex level, we induce a wide range of derived features (§4.2) from high-dimensional meaning representations. They capture semantically richer patterns of compound and constituent relatedness: based on lexical and topical patterns; in and out of sentence context; within and across time periods.[2]

The diachronic feature vectors are used as input to a support vector machine, implemented using the `SVC` class with default parameters from `scikit-learn` (Pedregosa et al., 2011). We train binary classifiers for each of the three types of compositionality ratings, yielding compound, modifier, and head predictions. To minimize overfitting, we use repeated k-fold cross-validation with 5 folds and 10 repetitions, and report mean accuracy across those runs. Any missing values are imputed with the mean value for the dimension in question.

4.1 Direct Features

On the simplest level, we characterize a compound's evolution with straightforward **dispersion measures**, and directly populate diachronic feature vectors with these time-specific values. We compute the **frequency** of all compounds and their constituents; and the **productivity** of all constituents formulated as morphological family size (de Jong et al., 2002), i.e., the number of unique compound-types in which a constituent appears in that specific role (modifier or head). We calculate productivity by counting all compound candidates containing a constituent; these are heuristically defined as a sequence of two nouns, neither preceded nor followed by a noun. For example, in the time slice corresponding to the 2000s, the compound *beauty sleep* has a raw modifier productivity of 103, reflecting the fact that *beauty* is used as the modifier in 103 distinct compounds in that time slice (*beauty product*, *beauty salon*, etc.). Both frequency and productivity are calculated for each time slice and then normalized by the respective slice's total size.

4.2 Derived Features

On a more complex level, we derive feature values from high-dimensional meaning representations which capture richer patterns of compound usage. We first discuss the meaning representation models we implement (§4.2.1) and then describe the process of deriving feature values from them (§4.2.2).

4.2.1 Meaning Representation Models

We implement (i) static representations – cooccurrence, word2vec, and topic models – to identify broad type-level tendencies of compound usage; and (ii) contextualized representations from transformer models, to better reflect finer-grained phenomena such as polysemy.

Cooccurrence models. For each time slice, we represent a target word's meaning as a vector of cooccurrence counts within a 10-word symmetrical window. We only use content words (nouns, verbs,

[2]Further approaches could yield stronger performance (e.g., a classifier directly on top of transformer embeddings) but without capturing compound usage in an interpretable way comparable to our other features. Such approaches therefore fall outside our focus on interpretable linguistic insights.

adjectives, adverbs) as vector dimensions. We also create a more restricted variant to compute distributional neighbors (cf. below): as potential neighbors, we only use nouns; as context words, we exclude the 50 most frequent ones and retain only those that appear at least 500 times throughout CCOHA. For comparisons of vectors across two time slices, we use the intersection of two models' context words as their dimensions.

word2vec. As an alternative neural approach, we train time-specific word2vec models (Mikolov et al., 2013) using skip-gram with negative sampling and setting window size to 10, vector dimensions to 100, minimum frequency to 1, and other parameters to default values. We train the models using the `gensim` library (Řehůřek and Sojka, 2010). For each time slice, we train the model on all CCOHA data from that time slice.[3] When comparing vectors from two time slices, we first align the models to a shared vector space using Orthogonal Procrustes, which corresponds to finding the best rotational alignment between two matrices so as to minimize the distance between their vectors. We use the implementation proposed by Hamilton et al. (2016b) and independently align all pairs of successive time slices.

Topic models. We represent topics using the stochastic block model (Peixoto, 2019), which automatically determines the number of topics in a hierarchical manner. It defines a graph whose nodes correspond to target words, and relies on identifying community structures within it. We use the first three levels of hierarchy, with respectively 3088, 83, and 9 topics for coarse-grained time slices; and 2164, 103, and 20 topics for fine-grained time slices. Target words are represented as vectors whose dimensions correspond to topics, and values correspond to a topic's probability of being represented for the target.

Transformer models. To deploy a transformer model, we collect all occurrences of a compound from a given time slice and feed them into the model, one sentence at a time. We retain embeddings for each token in the sequence at every layer. We use 9 pretrained models of different complexity and pretraining data, starting with **BERT** (Devlin et al., 2019) and **RoBERTa** (Liu et al., 2019) as well-established base architectures. We explore two computationally efficient alternatives: **DistilBERT** (Sanh et al., 2019), trained using knowledge distillation from BERT; and **ALBERT** (Lan et al., 2020), based on a simplified architecture which shares parameters across layers. Given our focus on noun compounds, we use **SpanBERT** (Joshi et al., 2020) as it was trained on predicting spans of multiple tokens. And since we are working with diachronic data, we use models optimized for historical language: **hmBERT** (Schweter et al., 2022), a multilingual model whose English pretraining data contains books from 1510 to 1900; **histLM** (Hosseini et al., 2021), pretrained on English books from 1760 to 1900; **MacBERTh** (Manjavacas and Fonteyn, 2022), pretrained on various corpora (including COHA) from 1450 to 1950; and **MacBERTh$_{\text{WSD}}$**, finetuned on word sense disambiguation. See App. C for implementation details.

[3] Due to randomness, we train five word2vec models per time slice and report average values.

4.2.2 Deriving Feature Values from Models

Given a meaning representation model described above, for each compound we derive diachronic feature values through pairwise comparisons of vectors representing different linguistic structures of interest in a given time slice. We now describe the types of target vectors on which we rely; the temporal settings in which we compare them (within a time slice or across a pair of successive time slices); and the similarity functions applied in the pairwise comparisons. A summary is shown in Figure 2.

Target vectors. We experiment with different target vectors, i.e., representations of different linguistic structures of interest taken from one of the meaning representation models described above. We consider the following types of target vectors: **`comp`**, corresponding to the compound (e.g., *climate change*); **`modif`**, corresponding to the modifier (e.g., *climate*); **`head`**, corresponding to the head (e.g., *change*); and only in transformer models, **`cont`**, corresponding to the sentential context of a compound, i.e. the average of all tokens in a sentence except for the compound, `[CLS]` and `[SEP]`. Note that in static models, `comp` is a dedicated vector learned from the preprocessed (underscore-joined) occurrences of a compound; in transformer models, we are bound by the pretrained tokenizer so we calculate `comp` by averaging `modif` and `head`.

Temporal settings. Pairwise comparisons of chosen target vectors are performed in two temporal settings: within a time slice, which corresponds to standard approaches for predicting the degrees

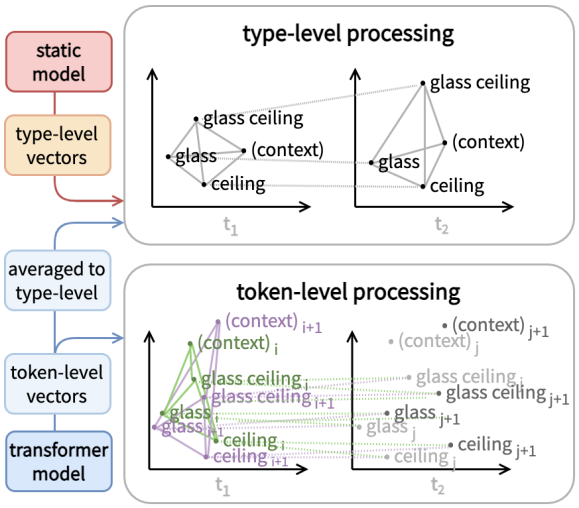

Figure 2: Illustration of different strategies to obtain derived feature values using within-time comparisons (solid lines) and across-time comparisons (dotted lines). **Type-level processing** is applied to (i) vectors from static models; (ii) averaged vectors from transformer models. Type-level context embedding is only used with transformer models. **Token-level processing** is applied to token-level vectors from transformer models. We illustrate it with two occurrences per time period, and for clarity only show within-time comparisons in t_1.

of compositionality of noun compounds (Cordeiro et al., 2019; Miletić and Schulte im Walde, 2023); and across time slices, which corresponds to standard approaches for predicting degrees of semantic change over time (Schlechtweg et al., 2019; Giulianelli et al., 2020). For **within-time estimates**, we use all pairwise combinations of target vectors: `comp`–`modif`, `comp`–`head`, `modif`–`head` and, only in transformer models, `comp`–`cont`, `head`–`cont`, and `modif`–`cont`. We use the directly obtained values, as well as three well-established composition functions (Reddy et al., 2011) to combine modifier and head information. Given pairwise estimates for (`modif`, `comp`) and (`head`, `comp`), the composite measures are: **ADD**, the sum of the two estimates; **MULT**, the product of the two estimates; **COMB**, the sum of **ADD** and **MULT**. We also use `cont` instead of `comp`. For **across-time estimates**, given a pair of successive time slices t_1 and t_2, we compare the representations of a given target – limited to `comp`, `modif`, or `head` – from t_1 and t_2.

Similarity functions. Pairwise comparisons of target vectors rely on different similarity functions. As the default, we compute the **pairwise cosine score** over two target vectors. This is straightforward for static models since they provide only one vector for each target word. With transformer models, we encode individual examples of a target word, yielding a different vector for each occurrence. We then need to aggregate this information, which we do using two approaches. In **type-level** processing, we first average over all contextualized vectors for a given target type (`comp`, `modif` etc.) and then compute the cosine for pairs of aggregated target vectors; this last operation is directly parallel to that for static models. In **token-level** processing, we compute the cosine directly for contextualized vectors and then average those cosine scores. For within-time estimates, we consider each sentence individually and compute the cosine for all pairs of target vectors in that sentence (`comp`–`modif`, `comp`–`head` etc.). For across-time estimates, we consider all contextualized vectors for a given type of target vector (e.g., `comp` occurring in sentence $1, 2, ..., n$) and compute cosine scores for all pairs of those vectors which come from different time slices.

For cooccurrence and word2vec models, we also compute **semantic neighborhood estimates** since using a broader set of words may alleviate instability due to data sparsity. In across-time settings, this further obviates the need for noise-prone model alignment (cf. Dubossarsky et al., 2017). We adopt two measures proposed for semantic change detection, but generalize them to comparisons of any two targets, within and across time. Given k nearest neighbors of two words, $\mathbf{NN_{SHARE}}$ is the proportion of overlapping neighbors in the two sets (Gonen et al., 2020). $\mathbf{NN_{COS}}$ is obtained by taking the union of the neighbors; defining a vector for each target populated with the cosine scores for that target and each neighbor in the union; and taking the cosine over the two vectors (Hamilton et al., 2016a). We use $k \in \{10, 20, 50, 100, 200, 500, 1000\}$ following Gonen et al. (2020).

5 Results and Discussion

5.1 Informativeness of Different Features

We begin by addressing **RQ1** and identifying diachronic features which are most predictive of present-day compositionality based on classification performance. We compare features across different approaches, constituents, and target items.

Performance across approaches. We first compare approaches based on their single best classification result (Table 2; for details on corresponding implementations, see Appendix D). We interpret these results as the least conservative estimates of

Family	Approach	Accuracy Comp	Modif	Head
Dispersion	frequency	**0.631**	0.608	**0.619**
	productivity	0.622	**0.629**	0.566
Static representations	cooccurrences	0.767	0.743	0.763
	word2vec	**0.844**	**0.871**	**0.776**
	topic model	0.746	0.793	0.693
Transformers (general)	BERT	**0.849**	0.748	**0.792**
	RoBERTa	0.726	0.720	0.724
	ALBERT	0.847	**0.774**	0.760
	DistilBERT	0.793	0.742	0.769
	SpanBERT	0.740	0.688	0.751
Transformers (historical)	histLM	**0.716**	**0.704**	**0.688**
	hmBERT	0.613	0.650	0.642
	MacBERTh	0.682	0.692	0.660
	MacBERTh-WSD	0.668	0.692	0.660
Random		0.500	0.500	0.500

Table 2: Best results across method families and individual approaches (bold: best in a family; shading: best overall). Accuracy is reported for compositionality ratings on the level of the compound, modifier, and head.

	Accuracy Compound		Modifier		Head	
Mean acc.	0.555	±0.080	0.542	±0.074	0.537	±0.075
Alt. info Δ						
Compound			–0.001	±0.056	0.007	±0.056
Modifier	–0.006	±0.058			0.001	±0.073
Head	–0.024	±0.056	–0.029	±0.078		

Table 3: Top: *predictability* of different types of compositionality ratings, measured as mean accuracy across all implementations. Bottom: *relative informativeness* of features specific to different compound structures, measured as mean change in accuracy when substituting a classification setting containing features specific to the predicted type of compositionality rating (on the level of the compound, modifier, or head; columns) with those corresponding to other parts of the compound (rows). Shading: Wilcoxon signed-rank test $p < 0.05$.

the informativeness of each approach, i.e., best-case results which may be subject to variability under different implementations (explored below).

All methods reach performance well above the random baseline, confirming that **diachronic features from all our approaches are predictive of present-day compositionality, but to variable extents**. The lowest-performing features are, perhaps unsurprisingly, the simplest. Frequency and productivity yield accuracy in the range of 0.6, comparable to the setup by Maurer et al. (2023); and static count-based representations (cooccurrence and topic models) are in the range of 0.7–0.8, similarly to Mahdizadeh Sani et al. (2024).

The approaches we introduce yield a further accuracy increase of $\approx$ 0.1 points. The strongest values are 0.849 for compound predictions; 0.871 for modifier predictions; and 0.792 for head predictions. They are mostly obtained by BERT-based approaches, suggesting that present-day compositionality is best captured by representations which directly reflect the use of target expressions in sentence context. But this result does not entail that classification accuracy is monotonic with respect to modeling complexity. In fact, decontextualized representations from word2vec – a considerably simpler, shallow neural architecture – perform on par with or, in some settings, better than BERT. Moreover, transformer models other than BERT are generally weaker; one exception is ALBERT, with results competitive or superior to BERT's (especially for modifier predictions) despite a simpler architecture.

Surprisingly, transformer models aligned with our target expressions (SpanBERT) and domain (historical pretraining data) do not fare well. Their performance is in the 0.6–0.7 range, placing them above dispersion-based features but – strikingly – below even count-based vector representations. While we expected to benefit from the broader temporal coverage in the historical models' pertaining data, based on these results we hypothesize that they go too far back in time relative to our corpus.

Performance across constituents. We now turn to differences across types of compositionality ratings (for the compound as a whole, modifier, or head), which we take to reflect the role of compound structures in compositionality evolution. As an indication of *predictability* of different scores, we note that the strongest mean accuracy is for compound-level predictions, followed by the modifier and then the head (Table 3, top). Looking at prior studies on (a subset of) the same gold standard, this trend parallels most diachronic approaches (Maurer et al., 2023; Mahdizadeh Sani et al., 2024; but *contra* Dhar et al., 2019) and contrasts synchronic ones (Schulte im Walde et al., 2016; Miletić and Schulte im Walde, 2023), which report overall better results for head information.

We further analyze the *relative informativeness* of features which target the whole compound, the modifier, or the head. To do so, we calculate the change in accuracy when substituting a feature specific to the predicted type of compositionality with another one (e.g., for modifier compositionality

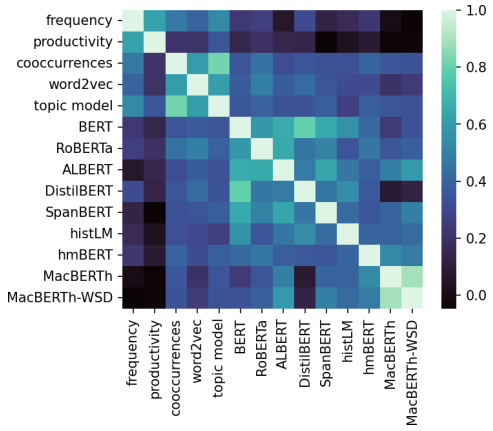

Figure 3: Spearman's correlation between approaches based on item-level proportion of misclassifications.

prediction, we compare the use of modifier frequencies vs. head frequencies, modifier–compound vs. head–compound embeddings, etc.), while keeping all other experimental settings fixed. Table 3 (bottom panel) shows the mean changes in accuracy across such pairs of experimental configurations.

Compound features are the most informative: they yield the best results for compound-level compositionality, and are non-detrimental or even beneficial for constituent compositionality. Modifier features are slightly less informative, as they somewhat penalize compound compositionality but not other prediction scores. Head features are the least informative, on average yielding clear drops in performance for both compound and modifier compositionality. While we should not overstate the mean differences in accuracy, together with score predictability (as defined above) they suggest that **present-day compositionality is most closely reflected by the evolution of compound and modifier properties, and less directly by changes in head properties**.

Performance across target items. We now assess if individual approaches better predict the compositionality of some *subset* of target items. For each compound, we calculate the proportion of classification runs in which it was misclassified, aggregating over all implementations of a given approach. We then use these compound-level rates of misclassification to compute Spearman's correlation between all pairs of approaches. Figure 3 shows that approaches from the same family (dispersion, static representations, general transformers, and historical transformers) are more correlated with one another than with approaches from other families. Indeed, mean correlations for approaches from the same family are systematically higher than their mean correlations with any other family. This indicates that **different diachronic features capture qualitatively different aspects of relevant information** rather than being more or less informative along a single dimension.

5.2 Compositionality vs. Semantic Change

Zooming into qualitative differences, we now answer **RQ2** by contrasting the informativeness of compositionality vs. semantic change estimates.

Within- vs. across-time estimates. The overall mean accuracy for within-time estimates (0.546 ± 0.080) is statistically significantly higher than for across-time estimates (0.538 ± 0.062).[4] We further break down this trend across approaches and types of compositionality ratings (Appendix E). Although across-time estimates obtain better results for a subset of approaches (ALBERT and some historical transformers), most cases align with the aggregate trend. This indicates that the process of **semantic change explains some aspects of compositionality evolution, but is not the only factor**.

Evolution of features. To understand how these multiple factors interact, we inspect the evolution of a subset of features (Figure 4). Over time, high-compositional compounds increase in frequency and in relatedness of compounds to constituents and to sentential context (within-time estimates). A given compound's time-specific representations are never perfectly related, i.e., its meaning changes across decades, but the rate of that change is smaller (across-time estimates). Looking at individual compounds, *love song* has an atypical frequency pattern, but shows expectedly high and stable compositionality estimates and low semantic change estimates. By contrast, *bank account* has less stable compositionality estimates – possibly reflecting cultural changes captured by context – but follows class trends for frequency and semantic change.

Low-compositional compounds diverge from these patterns in a variety of ways: their frequency also increases over time, but far more slowly; the relatedness between compound and constituent meanings decreases; and the relatedness between compounds and sentences is relatively stable. They exhibit systematically lower relatedness across decades, i.e., their meaning changes at a higher

[4] Mann-Whitney–U test ($p < 0.001$).

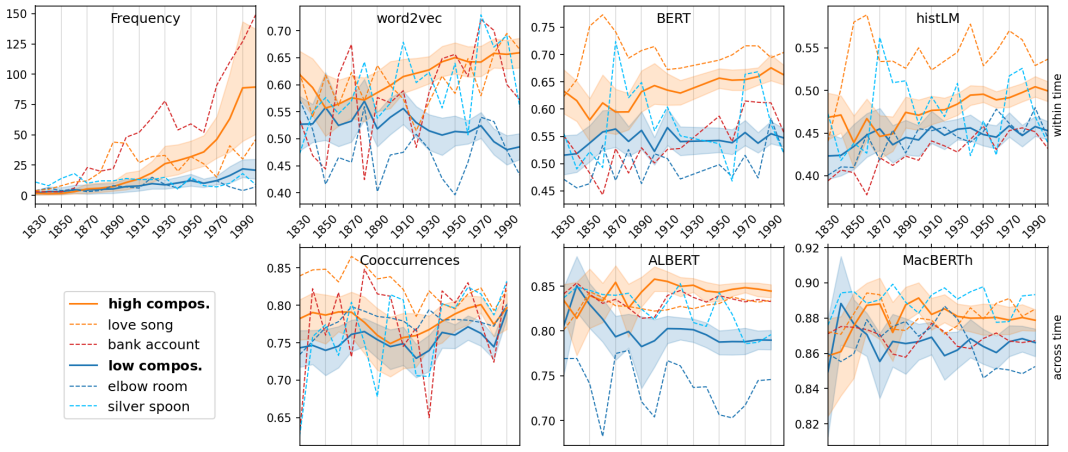

Figure 4: Evolution of features for high-compositional vs. low-compositional compounds (class-level mean with 95% CI and sample compounds). Top: within-time (compositionality) estimates; bottom: across-time (semantic change) estimates. Since all estimates use vector similarity, low across-time values indicate high semantic change. The plot shows best-performing implementations for each approach family in the fine-grained setting (see App. D).

rate. The example of *elbow room* closely mirrors these class-level trends. It is contrasted by *silver spoon*, whose estimates exhibit strong variance and overlap with the high compositionality ranges. This may be due to the compositional meaning still competing with the (now prevalent) non-compositional interpretation. These results overall indicate that **the development of low-compositional meanings is reflected by a decrease in compositionality paralleled by a sustained rate of semantic change**. This empirical finding is consistent with the theoretical claim that compounds first emerge with compositional interpretations and only later develop non-compositional ones (Bybee, 2015).

5.3 Feature Robustness

Turning to **RQ3**, we now assess whether experimental settings affect diachronic features derived from different approaches and, by extension, our understanding of the underlying linguistic patterns.

Diachronic data. We compare diachronic features against a synchronic setup using the last time slice. Diachronic information yields a mean performance improvement (0.006 ± 0.050) which is limited, but positive and statistically significant across approaches (Appendix F). This indicates that **diachronic information tends to be helpful in compositionality classification**. We also compare fine-grained (10-year) and coarse-grained (30-year) time slices. Mean accuracy difference favors coarse-grained data (0.005 ± 0.040), but its polarity and significance vary across approaches, suggesting a lack of clear trend.

love song (1890-1910)	song_nn, madrigal_nn, melody_nn, ditty_nn, lullaby_nn
love (1890-1910)	affection_nn, lover_nn, lovingness_nn, pure-minded_jj, passion_nn
song (1890-1910)	sing_vv, melody_nn, love_song_nn, madrigal_nn, ditty_nn
love song (1980-2000)	ballad_nn, song_nn, lyric_nn, bluesy_jj, ditty_nn
love (1980-2000)	unconsummated_jj, unrequited_jj, lover_nn, joy_nn, passion_nn
song (1980-2000)	tune_nn, sing_vv, ballad_nn, lyric_nn, love_song_nn

Table 4: Top 5 neighbors for *love song* (word2vec).

Static representations. Features derived from static representations (cooccurrences, word2vec, topic model) yield consistently strong predictions using pairwise cosine scores. For cooccurrences and word2vec, we also use neighborhood-based estimates NN_{COS} and NN_{SHARE}. They provide competitive or single best results, but are subject to variability across approaches (NN_{COS} strongly penalizes cooccurrence models) and temporal settings (stronger contribution in across-time settings). We conclude that **on the type level, compositionality evolution is most robustly reflected by directly measured cooccurrence similarities, but is also captured by even a small number of nearest neighbors**. The informativeness of neighborhood information is qualitatively illustrated in Table 4, which shows different patterns for *love song*: persistent within-time compositionality (cf. neighbor overlap with *song*) and slight semantic change over time (cf. neighbor differences between the periods). See Appendix G for further quantitative results.

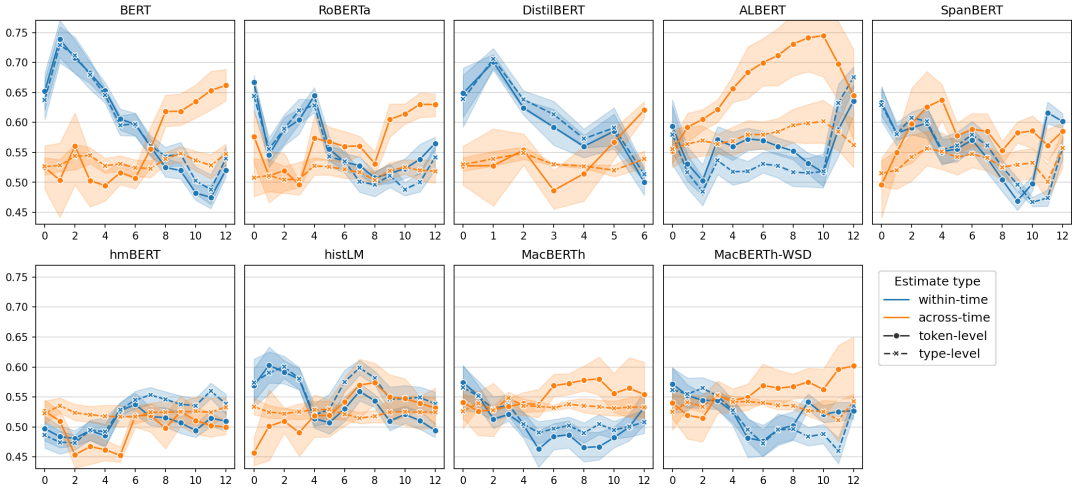

Figure 5: Accuracy for transformer models (mean and 95% CI). X-axis shows the layer used for representations.

Transformer models. The features derived from transformer models are distinguished by the layer and type-level vs. token-level processing (cf. Figure 5 for compound predictions; Appendix H for modifier and head predictions). Within-time estimates generally perform best in the initial layers, whereas across-time estimates relatively steadily increase in performance towards the final layers. Put differently, **on the token level, compositionality evolution is best captured by weakly contextualized within-time information and strongly contextualized across-time information**. As further illustration, consider the following examples:

(1) The cabin in which we dined is below [...] and one hundred and eighty persons can sit down at once and each have **elbow room** sufficient for all the purposes of figuring with the knife and fork (1907)

(2) The Piedmont Park [...] will have a lot more **elbow room** in a few years if the group that oversees the park can pull off an ambitious expansion plan (2007)

We hypothesize that contextualization penalizes within-time estimates – which compare *different target structures* within *the same sentence* – by reducing the distinctiveness of those target structures (e.g., *elbow room* vs. remaining tokens in a given sentence). By contrast, contextualization may benefit across-time estimates – which compare *the same target structure* across *different sentences* – by better capturing the meaning differences reflected by the surrounding context and in that way disambiguating the target structure's usages (e.g., comparing *elbow room* in the two time periods).

That said, historical models benefit less from increased contextualization in across-time settings. A potential reason is that pretraining on historical data produces representations which directly reflect historical usages, whereas pretraining on contemporary data might require contextual information to make the representations useful for semantic change. For further analyses, see Appendix H.

In terms of broader insights, our results support the tentative consensus that compositionality information is more recoverable in lower layers (Miletić and Schulte im Walde, 2024). The interaction we posit between contextualization and within-time vs. across-time estimates is compatible with the encoding of type-level semantics in lower layers (Vulić et al., 2020) and senses in higher layers (Coenen et al., 2019). We also replicate results for token-level vs. type-level processing, confirming their limited effect on compositionality prediction (Miletić and Schulte im Walde, 2023) and benefits from token-level approaches on semantic change prediction (Laicher et al., 2021). Our results come from a single experiment and as such constitute a more stringent replication of diverse prior findings.

6 Conclusion

We analyzed the evolution of English noun compounds using binary classification of present-day compositionality. We implemented a sweeping, linguistically motivated set of diachronic features, and analyzed their informativeness with respect to compositionality. While all our approaches can predict compositionality well above the random baseline, their informativeness is highly variable and complementary in nature. Further, the development of low-compositional meanings is reflected by a parallel drop in compositionality and sustained semantic change, a distinction also reflected in degrees of contextualization of transformer representations.

Limitations

This study evaluated methods to represent noun compound meanings over time and, based on those representations, predict their present-day compositionality. While we aimed for a comprehensive setup providing insights into different model mechanisms, this could be further expanded in different ways. With respect to transformer models, our focus was on differences in architectures and pretraining data, but another important factor is model size; architectures with different numbers of parameters could therefore be compared.

We also considered using instruction-tuned autoregressive models. In preliminary experiments, we prompted LLMs to generate compositionality and semantic change scores given input sentences from our corpus. We were unable to identify a reasonably robust prompting setup, suggesting the need to strongly optimize the models' instruction-following for the target tasks. We consider that this prerequisite falls outside the scope of the present paper and we reserve it for future work. Recall moreover that we aim to contrast linguistically interpretable types of information. Our setup already includes a variety of BERT-derived features, which are linguistically closely related to LLM-derived features (i.e., they both reflect compound usage in sentence context). We therefore do not consider LLMs as central to the empirical focus of our work.

We used a relatively small gold standard dataset (we retained 120 instances for classification), which is however well-established on the task of compositionality prediction, and we are unaware of larger comparable datasets. Moreover, the compositionality ratings it contains are limited to a synchronic present-day perspective by virtue of them being produced by contemporary speakers. For obvious reasons, we cannot know with certainty how speakers from earlier stages of language use (say, 100 years ago) would have perceived the compositionality of the same target items.

From a different perspective, our experiments are limited to a single type of multiword expressions in one European language, English. The conclusions that we draw must be seen within this specific context. Expanding our approach to include different types of expressions (e.g. particle verbs, idioms, or light verb constructions) and different languages would provide more robust insights in model behaviors as well as the underlying linguistic mechanisms.

Acknowledgments

The research presented here was supported by 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 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 4379–4387, Marseille, France. European Language Resources Association.

Timothy Baldwin and Su Nam Kim. 2010. Multiword expressions. In Nitin Indurkhya and Fred J. Damerau, editors, *Handbook of Natural Language Processing*, pages 267–292. CRC Press, Boca Raton, USA.

Lars Buijtelaar and Sandro Pezzelle. 2023. A psycholinguistic analysis of BERT's representations of compounds. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2230–2241, Dubrovnik, Croatia. Association for Computational Linguistics.

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

Pierluigi Cassotti, Lucia Siciliani, Marco DeGemmis, Giovanni Semeraro, and Pierpaolo Basile. 2023. XL-LEXEME: WiC pretrained model for cross-lingual LEXical sEMantic changE. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1577–1585, Toronto, Canada. Association for Computational Linguistics.

Andy Coenen, Emily Reif, Ann Yuan, Been Kim, Adam Pearce, Fernanda Viégas, and Martin Wattenberg. 2019. Visualizing and measuring the geometry of BERT. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates.

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

Verna Dankers, Christopher Lucas, and Ivan Titov. 2022. Can transformer be too compositional? Analysing idiom processing in neural machine translation. In

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 3608–3626, Dublin, Ireland. Association for Computational Linguistics.

Mark Davies. 2012. Expanding horizons in historical linguistics with the 400-million word Corpus of Historical American English. *Corpora*, 7(2):121–157.

Nivja H. de Jong, Laurie B. Feldman, Robert Schreuder, Matthew Pastizzo, and R.Harald Baayen. 2002. The processing and representation of Dutch and English compounds: Peripheral morphological and central orthographic effects. *Brain and Language*, 81(1):555–567.

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.

Prajit Dhar and Lonneke van der Plas. 2019. Learning to predict novel noun-noun compounds. In *Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019)*, pages 30–39, Florence, Italy. Association for Computational Linguistics.

Haim Dubossarsky, Daphna Weinshall, and Eitan Grossman. 2017. Outta control: Laws of semantic change and inherent biases in word representation models. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1136–1145, Copenhagen, Denmark. Association for Computational Linguistics.

Marcos Garcia, Tiago Kramer Vieira, Carolina Scarton, Marco Idiart, and Aline Villavicencio. 2021a. Assessing the representations of idiomaticity in vector models with a noun compound dataset labeled at type and token levels. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2730–2741, Online. Association for Computational Linguistics.

Marcos Garcia, Tiago Kramer Vieira, Carolina Scarton, Marco Idiart, and Aline Villavicencio. 2021b. Probing for idiomaticity in vector space models. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3551–3564, Online. Association for Computational Linguistics.

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.

Hila Gonen, Ganesh Jawahar, Djamé Seddah, and Yoav Goldberg. 2020. Simple, interpretable and stable method for detecting words with usage change across corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 538–555, Online. Association for Computational Linguistics.

Kristina Gulordava and Marco Baroni. 2011. A distributional similarity approach to the detection of semantic change in the Google Books Ngram corpus. In *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*, pages 67–71, Edinburgh, UK. Association for Computational Linguistics.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016a. Cultural shift or linguistic drift? Comparing two computational measures of semantic change. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2116–2121, Austin, Texas. Association for Computational Linguistics.

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

Kasra Hosseini, Kaspar Beelen, Giovanni Colavizza, and Mariona Coll Ardanuy. 2021. Neural language models for nineteenth-century English. *Journal of Open Humanities Data*, 7:22.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. SpanBERT: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Urban Knupleš, Diego Frassinelli, and Sabine Schulte im Walde. 2023. Investigating the nature of disagreements on mid-scale ratings: A case study on the abstractness-concreteness continuum. In *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, pages 70–86, Singapore. Association for Computational Linguistics.

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.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A lite BERT for self-supervised learning of language representations. In *Proceedings of the Eighth International Conference on Learning Representations (ICLR 2020)*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv:1907.11692*.

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.

Enrique Manjavacas and Lauren Fonteyn. 2022. Adapting vs. pre-training language models for historical languages. *Journal of Data Mining & Digital Humanities*.

Maximilian Maurer, Chris Jenkins, Filip Miletic, and Sabine Schulte im Walde. 2023. Classifying noun compounds for present-day compositionality: Contributions of diachronic frequency and productivity patterns. In *Proceedings of the 19th Conference on Natural Language Processing*, Ingolstadt, Germany.

Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, ICLR 2013, Workshop Track Proceedings*, Scottsdale, Arizona, USA.

Filip Miletić and Sabine Schulte im Walde. 2024. Semantics of multiword expressions in transformer-based models: A survey. *Transactions of the Association for Computational Linguistics*, 11:593–612.

Filip Miletić and Sabine Schulte im Walde. 2023. A systematic search for compound semantics in pre-trained 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.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Tiago P. Peixoto. 2019. Bayesian stochastic blockmodeling. In Patrick Doreian, Vladimir Batagelj, and Anuska Ferligoj, editors, *Advances in Network Clustering and Blockmodeling*, chapter 11. Wiley Online Library.

Lewis Pollock. 2018. Statistical and methodological problems with concreteness and other semantic variables: A list memory experiment case study. *Behavior Research Methods*, 50(3):1198–1216.

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.

Radim Řehůřek and Petr Sojka. 2010. Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA.

Guy D. Rosin, Ido Guy, and Kira Radinsky. 2022. Time masking for temporal language models. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, WSDM '22, page 833–841, New York, NY, USA. Association for Computing Machinery.

Bahar Salehi, Paul Cook, and Timothy Baldwin. 2014. Using distributional similarity of multi-way translations to predict multiword expression compositionality. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 472–481, Gothenburg, Sweden. Association for Computational Linguistics.

Bahar Salehi, Paul Cook, and Timothy Baldwin. 2015. A word embedding approach to predicting the compositionality of multiword expressions. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 977–983, Denver, Colorado. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. In *5th Workshop on Energy Efficient Machine Learning and Cognitive Computing – NeurIPS 2019*.

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.

Sabine Schulte im Walde, Anna Hätty, and Stefan Bott. 2016. 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, Stefan Müller, and Stefan Roller. 2013. Exploring vector space models to predict the compositionality of German noun-noun compounds. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*, pages 255–265, Atlanta, Georgia, USA. Association for Computational Linguistics.

Stefan Schweter, Luisa März, Katharina Schmid, and Erion Çano. 2022. hmBERT: Historical multilingual language models for named entity recognition. In *Proceedings of the Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum*, volume 3180 of *CEUR Workshop Proceedings*, pages 1109–1129.

Vered Shwartz and Ido Dagan. 2019. Still a pain in the neck: Evaluating text representations on lexical composition. *Transactions of the Association for Computational Linguistics*, 7:403–419.

Nina Tahmasebi, Lars Borin, and Adam Jatowt. 2021. Survey of computational approaches to lexical semantic change. In Nina Tahmasebi, Lars Borin, Adam Jatowt, Yang Xu, and Simon Hengchen, editors, *Computational Approaches to Semantic Change*, pages 1–91. Language Science Press, Berlin.

Ivan Vulić, Edoardo Maria Ponti, Robert Litschko, Goran Glavaš, and Anna Korhonen. 2020. Probing pretrained language models for lexical semantics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7222–7240, Online. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

A Data Details

Data distribution across time. The distribution of data from CCOHA across individual time slices is presented in Table 5.

Decade	Time slice size Fine	Coarse
1830s	15.3	
1840s	17.7	51.3
1850s	18.3	
1860s	18.8	
1870s	20.6	61.6
1880s	22.2	
1890s	22.4	
1900s	24.3	71.7
1910s	25.1	
1920s	28.2	
1930s	27.1	82.1
1940s	26.9	
1950s	27.3	
1960s	26.6	80.2
1970s	26.4	
1980s	28.0	
1990s	31.0	91.8
2000s	32.7	

Table 5: Corpus size (in millions of tokens) for fine-grained and coarse-grained time slices.

Dataset use and licenses. Our use of datasets is in line with their intended use for research, as described below in more detail. While it is conceivable that the corpus data may contain offensive or otherwise sensitive content, we use it for an aggregate analysis of lexical semantics on the level of noun compounds and only obtain numerical system outputs. None of the instances in the gold standard dataset are offensive.

COHA: We acquired COHA with an academic multi-user license prior to this study. For copyright compliance, corpus creators use the @ symbol to redact 10 tokens after every 832,200 tokens.

Compositionality ratings: The gold standard data by Reddy et al. (2011)[5] and Cordeiro et al. (2019)[6] is publicly available and is not associated with a specific license.

[5] http://www.dianamccarthy.co.uk/downloads.html

[6] https://pageperso.lis-lab.fr/carlos.ramisch/?page=downloads/compounds

B Target Selection

Following the general overview in Section 3, we now elaborate on the process of target selection from the gold standard compositionality datasets by Reddy et al. (2011) and Cordeiro et al. (2019).

Frequency threshold. We start from the 210 noun–noun compounds in the datasets and retain those appearing in CCOHA at least once in at least two successive time slices of both granularities. We exclude the following 44 items: *agony aunt*, *armchair critic*, *baby blues*, *backroom boy*, *blame game*, *call centre*, *carpet bombing*, *cash cow*, *cheat sheet*, *cloud nine*, *contact lenses*, *copy cat*, *couch potato*, *cutting edge*, *diamond wedding*, *end user*, *eye candy*, *fine line*, *head teacher*, *honey trap*, *information age*, *injury time*, *insane asylum*, *insider trading*, *job fair*, *labour union*, *mailing list*, *music journalist*, *number crunching*, *panda car*, *pecking order*, *rat run*, *sacred cow*, *search engine*, *shrinking violet*, *sitting duck*, *smoking gun*, *snail mail*, *spinning jenny*, *stag night*, *top dog*, *video game*, *web site*, *zebra crossing*. The excluded items have lower average compositionality ratings than the retained items: on the level of the compound (2.4 ± 1.4 vs. 2.9 ± 1.6), the modifier (2.8 ± 1.6 vs. 3.0 ± 1.8), and the head (2.7 ± 1.8 vs. 3.3 ± 1.6).

Binary labels. We define binary compositionality labels for the 166 targets retained after frequency filtering. For each compositionality rating (on the level of the compound, the modifier, and the head), we sort the items based on that rating. We retain the lowest 60 as low-compositionality items, the highest 60 as high-compositionality items, and discard those in the middle. Their distribution across the rating scale is shown in Figure 6. Although the low-compositional class encompasses a broader range of values (largely due to the frequency skew noted above), the two classes remain clearly distinct, as also shown by sample items in Table 6. Importantly, this approach enables us to replicate the classification setup from prior work (Maurer et al., 2023; Mahdizadeh Sani et al., 2024) with a comparable number of targets and also limit the potentially detrimental effect of mid-range items with high standard deviation.

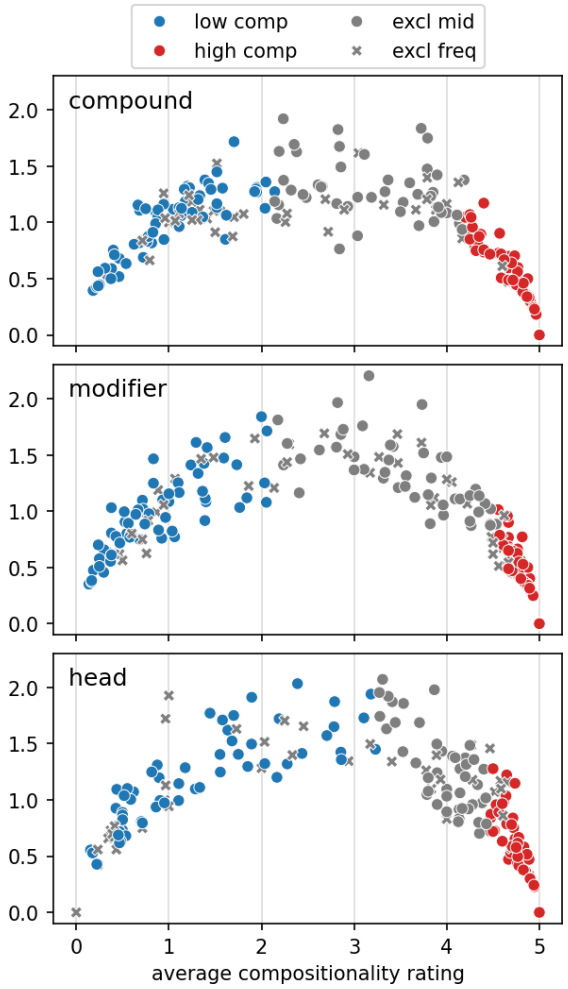

Figure 6: Distribution of gold standard targets based on their average compositionality ratings (x-axis) and standard deviation across annotators (y-axis). **Subplots:** ratings on the level of the whole compound, the modifier, and the head. **Categories:** low compositionality class (blue); high compositionality class (red); items excluded due to insufficient frequency (gray crosses); items excluded as part of the mid-range (gray dots).

Compound targets			**Modifier targets**			**Head targets**		
basket case	0.2	±0.4	**basket** case	0.1	±0.4	elbow **grease**	0.2	±0.6
silver lining	0.2	±0.4	**kangaroo** court	0.2	±0.4	sex **bomb**	0.2	±0.5
guinea pig	0.2	±0.6	**silver** lining	0.2	±0.4	silver **lining**	0.2	±0.4
sugar daddy	0.2	±0.4	**crocodile** tear	0.2	±0.5	loan **shark**	0.4	±0.7
nut case	0.2	±0.4	**sugar** daddy	0.2	±0.7	elbow **room**	0.4	±0.9
gravy train	0.3	±0.6	**nut** case	0.2	±0.6	nut **case**	0.4	±1.1
flower child	0.4	±0.5	**rat** race	0.2	±0.5	gravy **train**	0.4	±0.8
elbow room	0.4	±0.6	**gravy** train	0.3	±0.5	word **painting**	0.5	±0.7
goose egg	0.4	±0.8	**goose** egg	0.3	±0.7	think **tank**	0.5	±0.6
sex bomb	0.4	±0.7	**snake** oil	0.4	±0.6	guinea **pig**	0.5	±0.7
...			...			...		
cotton candy	1.6	±0.8	**silver** screen	1.4	±1.6	market **place**	2.4	±2.0
lip service	1.6	±1.1	**silver** spoon	1.6	±1.5	brain **teaser**	2.4	±1.4
firing line	1.7	±1.7	**firing** line	1.6	±1.7	cocoa **butter**	2.7	±1.6
night owl	1.9	±1.3	**polo** shirt	1.7	±1.4	cable **car**	2.8	±1.6
fairy tale	1.9	±1.3	**fairy** tale	1.8	±1.0	banana **republic**	2.8	±1.9
con artist	1.9	±1.3	**pocket** book	1.8	±1.1	jet **lag**	2.9	±1.4
foot soldier	1.9	±1.3	**china** clay	2.0	±1.8	rush **hour**	2.9	±1.4
think tank	2.0	±1.1	**lip** service	2.0	±1.3	life **belt**	3.1	±1.7
pain killer	2.0	±1.4	**beauty** sleep	2.1	±1.1	health **care**	3.2	±1.9
crash course	2.1	±1.3	**tennis** elbow	2.1	±1.7	silver **screen**	3.2	±1.4
car park	4.2	±1.0	**bow** tie	4.5	±0.9	crime **rate**	4.5	±0.8
rice paper	4.2	±1.1	**calendar** month	4.6	±1.0	peace **conference**	4.5	±0.9
speed trap	4.2	±0.9	**travel** guide	4.6	±0.8	graveyard **shift**	4.5	±0.7
traffic jam	4.3	±1.0	**pillow** slip	4.6	±0.7	world **conference**	4.5	±1.3
health check	4.3	±0.9	**computer** expert	4.7	±0.7	chain **reaction**	4.5	±0.7
skin tone	4.3	±1.0	**radio** station	4.7	±1.0	research **project**	4.5	±1.0
incubation period	4.3	±0.8	**ground** floor	4.7	±0.7	street **girl**	4.6	±0.9
football season	4.3	±0.8	**credit** card	4.7	±0.5	mail **service**	4.6	±0.6
phone book	4.3	±0.7	**public** service	4.7	±0.7	prison **term**	4.6	±0.6
food market	4.3	±0.8	**fair** play	4.7	±0.7	subway **system**	4.6	±0.6
...			...			...		
engine room	4.9	±0.3	**mail** service	4.9	±0.3	academy **award**	4.9	±0.2
time difference	4.9	±0.2	**speed** limit	4.9	±0.2	cocktail **dress**	5.0	±0.0
winter solstice	4.9	±0.2	**computer** program	4.9	±0.2	grandfather **clock**	5.0	±0.0
climate change	5.0	±0.2	**health** check	5.0	±0.0	graduate **student**	5.0	±0.0
insurance company	5.0	±0.0	**crime** rate	5.0	±0.0	engine **room**	5.0	±0.0
music festival	5.0	±0.0	**insurance** company	5.0	±0.0	lime **tree**	5.0	±0.0
prison term	5.0	±0.0	**music** festival	5.0	±0.0	polo **shirt**	5.0	±0.0
prison guard	5.0	±0.0	**wedding** anniversary	5.0	±0.0	milk **tooth**	5.0	±0.0
wedding anniversary	5.0	±0.0	**wedding** day	5.0	±0.0	wedding **anniversary**	5.0	±0.0
wedding day	5.0	±0.0	**winter** solstice	5.0	±0.0	tear **gas**	5.0	±0.0

Table 6: Targets retained for prediction of compositionality on the level of the compound, the modifier, and the head. Top: low-compositionality class; bottom: high-compositionality class. The targets are sorted by average compositionality rating. For space reasons, we omit 40 intermediate targets for each class.

C Transformer Models and Computational Infrastructure

Architectures and infrastructure. We use transformer model implementations from HuggingFace Transformers (Wolf et al., 2020). We provide an overview of deployed pretrained models in Table 7, together with their number of parameters and the corresponding identifier on HuggingFace Hub. We use base, uncased, monolingual English models, except for hmBERT (only available as cased and multilingual) and SpanBERT (only available as cased). All models have 12 hidden layers, except for DistilBERT which has 6. We ran experiments on a single Nvidia GeForce RTX A6000 GPU with 48 GB of memory. Inference requires around one hour per model, for a total runtime of ≈10 hours.

Model	Params	HuggingFace Hub ID
ALBERT	12 M	`albert/albert-base-v2`
BERT	109 M	`google-bert/bert-base-uncased`
DistilBERT	66 M	`distilbert/distilbert-base-uncased`
histLM	109 M	`Livingwithmachines/bert_1760_1900`
hmBERT	111 M	`dbmdz/bert-base-historic-multilingual-cased`
MacBERTh	109 M	`emanjavacas/MacBERTh`
$\text{MacBERTh}_{\text{WSD}}$	109 M	`emanjavacas/MacBERTh-metric-wsd`
RoBERTa	125 M	`FacebookAI/roberta-base`
SpanBERT	108 M	`SpanBERT/spanbert-base-cased`

Table 7: Summary of used pretrained models

Further implementation details. As input sequences, we used compound examples extracted from the lemmatized, POS-tagged version of CCOHA, but they were stripped of POS tags before being fed into the models. While the models are pretrained on non-lemmatized text, we opted for lemmatized input because it can benefit predictions on tasks such as semantic change detection (Laicher et al., 2021); it also allowed us to maintain direct comparisons with the remaining approaches. We set the maximum size of input sequences to 256 tokens and truncate them otherwise. The pretrained tokenizer splits out-of-vocabulary tokens into subword fragments; when this occurs for a token of interest, we average over the subwords. If it occurs when computing the embedding of the full compound, we take the micro-average of all tokens produced for the modifier and the head.

We do not fine-tune the models principally because we are comparing models pretrained on different types of data, whose effect we do not wish to obscure. More generally, our measures rely on meaning differences in context, which off-the-shelf models capture both for compositionality (Miletić and Schulte im Walde, 2023) and semantic change (Laicher et al., 2021).

D Best Performing Configurations

Details on individual best-performing configurations are presented in Table 8.

E Comparison of Within-Time and Across-Time Estimates

The distribution of accuracy values obtained using within-time vs. across-time estimates is plotted in Figure 7.

F Additional Results for Diachronic Data Settings

We provide further analyses on the effect of temporal granularity (fine-grained vs. coarse-grained time slices); and using the full range of diachronic information vs. only the most recent time slice (diachronic vs. static approach). We plot relative accuracy differences in Figure 8 and examine their distribution across approaches.

Regarding temporal granularity, positive values indicate a better performance of the fine-grained approach. While the median difference tends to hover around 0, we also observe model-specific behaviors. The extreme median values are the ones for ALBERT (0.017), indicating a preference for fine-grained information; and for word2vec (-0.014), suggesting a preference for coarse-grained information. The latter trend aligns with the sensitivity of word2vec to the amount of training data, which increases approximately three-fold in the coarse-grained setup. Beyond the central tendency, most methods show clear outliers, with the highest absolute differences in accuracy of ≈ 0.2. This indicates that specific *combinations* of experimental settings are strongly affected by granularity.

As for the diachronic vs. static approaches, positive values indicate a better performance of the diachronic setting. The strongest median effect is shown by the frequency-based approach (0.027), followed by word2vec and topic models (0.009 for both); as before, we note clear outliers across the methods. Importantly, although median values are overall low, they are all positive (with the sole exception of MacBERTh-WSD). Compared to the disparate results for granularity, this trend points to a more consistent effect and confirms the general potential of diachronic information.

Approach	Dims	Pred.	Granularity	Similarity	Estimate	Target vectors		Param.	Acc.
Best implementations overall for each approach									
frequency	1	comp	coarse	direct	within	comp			0.631
		modif	fine	direct	within	modif			0.608
		head	coarse	direct	within	comp			0.619
productivity	1	comp	fine	direct	within	modif			0.622
		modif	fine	direct	within	modif			0.629
		head	fine	direct	within	head			0.566
cooccurrences	from 233k to 489k	comp	coarse	nn-share	within	comp	COMB	k10	0.767
		modif	coarse	nn-share	within	comp	modif	k10	0.743
		head	coarse	nn-share	within	comp	ADD	k10	0.763
word2vec	100	comp	coarse	cos	within	comp	modif		0.844
		modif	coarse	nn-cos	within	comp	modif	k500	0.871
		head	coarse	nn-share	within	comp	ADD	k500	0.776
topic model	coarse 9 fine 20	comp	coarse	cos	within	comp	modif	h2	0.746
		modif	coarse	cos	within	comp	modif	h2	0.793
		head	fine	cos	within	comp	MULT	h2	0.693
BERT	768	comp	fine	cos-token	within	cont	COMB	l1	0.849
		modif	fine	cos-type	within	cont	ADD	l2	0.748
		head	fine	cos-type	within	cont	COMB	l1	0.792
RoBERTa	768	comp	coarse	cos-type	within	cont	MULT	l0	0.726
		modif	coarse	cos-token	within	comp	MULT	l0	0.720
		head	coarse	cos-token	within	head	cont	l0	0.724
ALBERT	4096	comp	fine	cos-token	across	modif		l10	0.847
		modif	coarse	cos-token	across	modif		l10	0.774
		head	fine	cos-token	across	modif		l8	0.760
DistilBERT	768	comp	coarse	cos-token	within	cont	MULT	l0	0.793
		modif	coarse	cos-token	within	cont	MULT	l0	0.743
		head	fine	cos-type	within	cont	ADD	l1	0.769
SpanBERT	768	comp	coarse	cos-token	within	cont	COMB	l0	0.740
		modif	coarse	cos-token	across	head		l3	0.688
		head	coarse	cos-type	within	head	cont	l0	0.751
histLM	768	comp	fine	cos-type	within	comp	cont	l1	0.716
		modif	fine	cos-token	within	modif	cont	l1	0.704
		head	fine	cos-type	within	comp	cont	l1	0.688
hmBERT	768	comp	fine	cos-type	within	head	cont	l4	0.613
		modif	fine	cos-type	within	cont	MULT	l12	0.650
		head	coarse	cos-type	within	modif	cont	l7	0.642
MacBERTh	768	comp	fine	cos-token	across	modif		l12	0.682
		modif	fine	cos-token	within	cont	MULT	l0	0.692
		head	coarse	cos-type	within	cont	MULT	l0	0.660
MacBERTh$_{\text{WSD}}$	768	comp	coarse	cos-token	across	modif		l12	0.668
		modif	fine	cos-token	within	cont	MULT	l0	0.692
		head	coarse	cos-type	within	cont	MULT	l0	0.660
Best implementations for compound-level predictions in fine-grained setting (plotted in Figure 4)									
frequency	1	comp	fine	direct	within	comp			0.627
word2vec	100	comp	fine	cos	within	comp	modif		0.827
BERT	768	comp	fine	cos-token	within	cont	COMB	l1	0.849
histLM	768	comp	fine	cos-type	within	comp	cont	l1	0.716
cooccurrences	26k–240k	comp	fine	nn-share	across	modif		k1000	0.667
ALBERT	4096	comp	fine	cos-token	across	modif		l10	0.847
MacBERTh	768	comp	fine	cos-token	across	modif		l12	0.682

Table 8: Details on best-performing implementations across approaches and predicted scores (compositionality on the level of the compound, head, and modifier). Dims: number of dimensions of the corresponding meaning representations (before deriving feature values used for diachronic compositionality classification); note that for cooccurrence models it varies depending on granularity and time slice. Param: additional parameter depending on the approach (k: number of nearest neighbors; h: hierarchy level for stochastic block model; l: transformer layer).

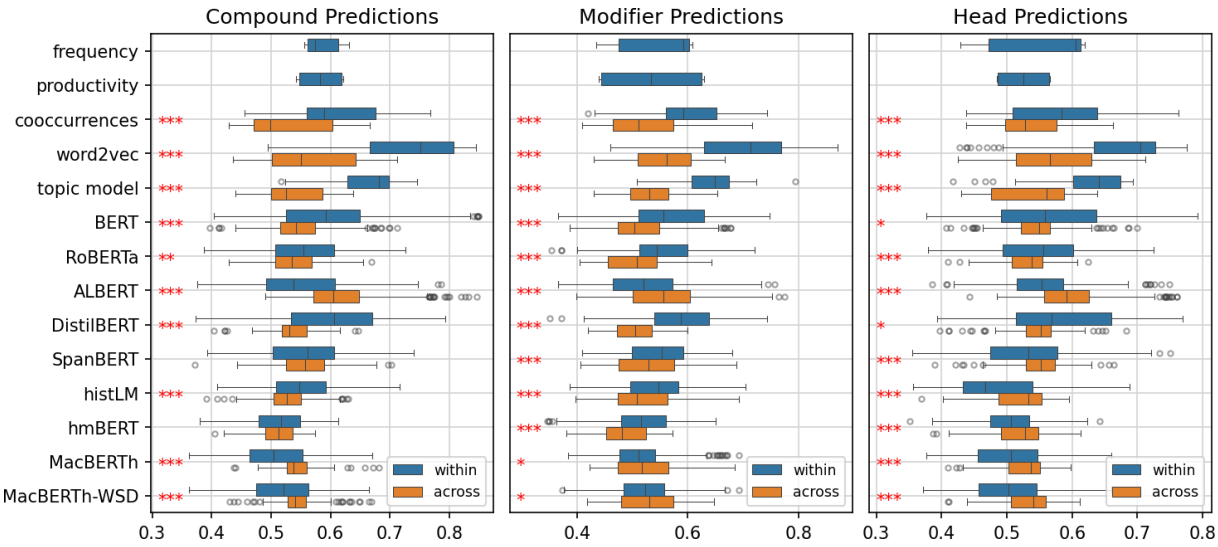

Figure 7: Distributions of accuracy values for within-time vs. across-time estimates, across the three predicted types of compositionality scores. Statistical significance markers report the Mann-Whitney–U test: $p < 0.001$ (***), $p < 0.01$ (**), $p < 0.05$ (*). Frequency and productivity are only implemented as within-time estimates, but are provided for comparison.

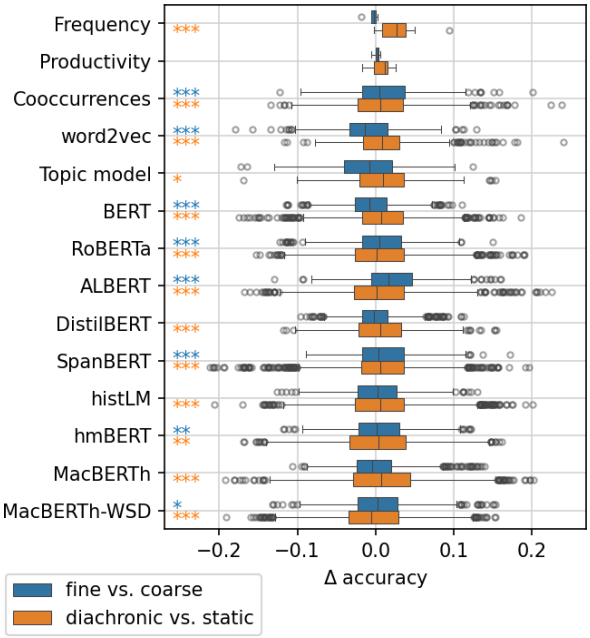

Figure 8: Differences in accuracy depending on data settings. Granularity: positive values indicate better performance of fine-grained time slices (acc_{fine}–$\text{acc}_{\text{coarse}}$). Time span: positive values indicate better performance of the diachronic setting ($\text{acc}_{\text{diachr.}}$–$\text{acc}_{\text{static}}$). Statistical significance markers report the Wilcoxon signed-rank test for the corresponding distributions of accuracy values: $p < 0.001$ (***), $p < 0.01$ (**), $p < 0.05$ (*).

G Additional Results for Static Representations

We implemented three approaches based on static representations: the SBM topic model, count-based cooccurrence vectors, and word2vec. The results are summarized in Figure 9.

On the broadest level, the predictions based on pairwise cosine scores taken over meaning vectors (gray boxes) show that topic models and cooccurrence vectors have a similar range of performance, and are clearly outperformed by word2vec. Regarding topic-based representations, recall that we use different hierarchy levels induced by the stochastic block model (§4.2.1); the number of topics in successive levels differs by roughly an order of magnitude. Each stepwise move from a lower to a higher hierarchy level reduces mean accuracy by around 0.010 points. However, the single best result for all three prediction targets is obtained by the highest hierarchy level (i.e., the lowest number of topics), suggesting that the corresponding representational information is highly informative in some parameter combinations, but not very robust given its outlier status (Table 9).

For cooccurrence and word2vec models, we also use the neighborhood-based estimates NN_{COS} and NN_{SHARE}. The best of the two for a given model is competitive with or outperforms cosine-based predictions. However, there is variability across approaches: NN_{COS} strongly penalizes cooccurrence models, whereas it is the better of the two

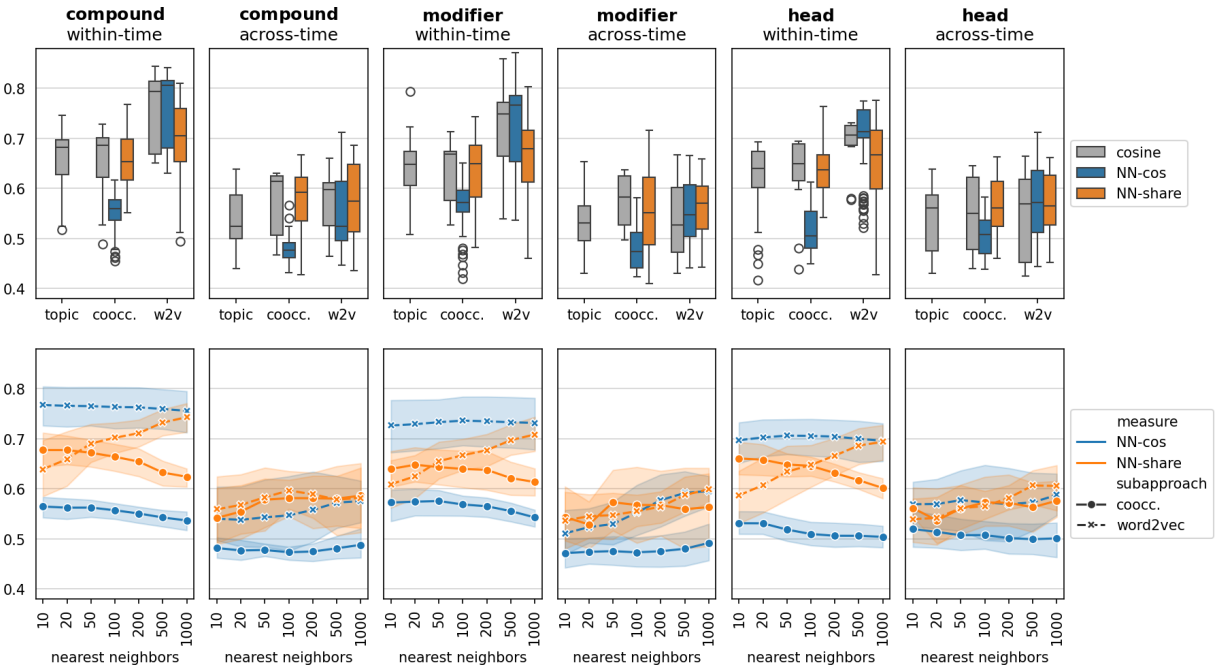

Figure 9: Accuracy for topic models, count-based cooccurrences and word2vec. Top: distribution for all estimates. Bottom: neighbor-based estimates (mean accuracy and 95% confidence interval) depending on the number of nearest neighbors.

pred. score	level	mean	min	max
compound	0	**0.630**	0.459	0.717
	1	0.619	**0.485**	0.710
	2	0.611	0.440	**0.746**
modifier	0	**0.618**	**0.502**	0.718
	1	0.605	0.441	0.723
	2	0.593	0.431	**0.793**
head	0	**0.600**	**0.450**	0.686
	1	0.593	0.430	0.685
	2	0.581	0.418	**0.693**

Table 9: Accuracy values for the topic model broken down by the types of compositionality ratings and the level of hierarchy in the stochastic block model.

estimates for word2vec. There are also differences depending on temporal settings: in comparison to cosine-based estimates, neighborhood estimates are particularly useful in across-time comparisons, possibly reflecting a greater degree of representational stability compared to the standard cosine approach which requires model alignment.

As for the number of nearest neighbors k, within-time performance tends to decline with an increase in k, except for word2vec's NN_{SHARE}. In contrast, across-time performance marginally but steadily increases for NN_{COS} and peaks around $k = 100$ for NN_{SHARE}. Put differently, within-time estimates, which directly correspond to compositionality predictions, benefit from a word's most immediate semantic neighborhood. Across-time estimates, which measure semantic change, are more accurate with a broader range of points of comparison.

H Additional Results for Transformers

Average performance across layers. Full results for compositionality prediction using transformer-based models are provided in Figure 10 for modifier predictions; and in Figure 11 for head predictions.

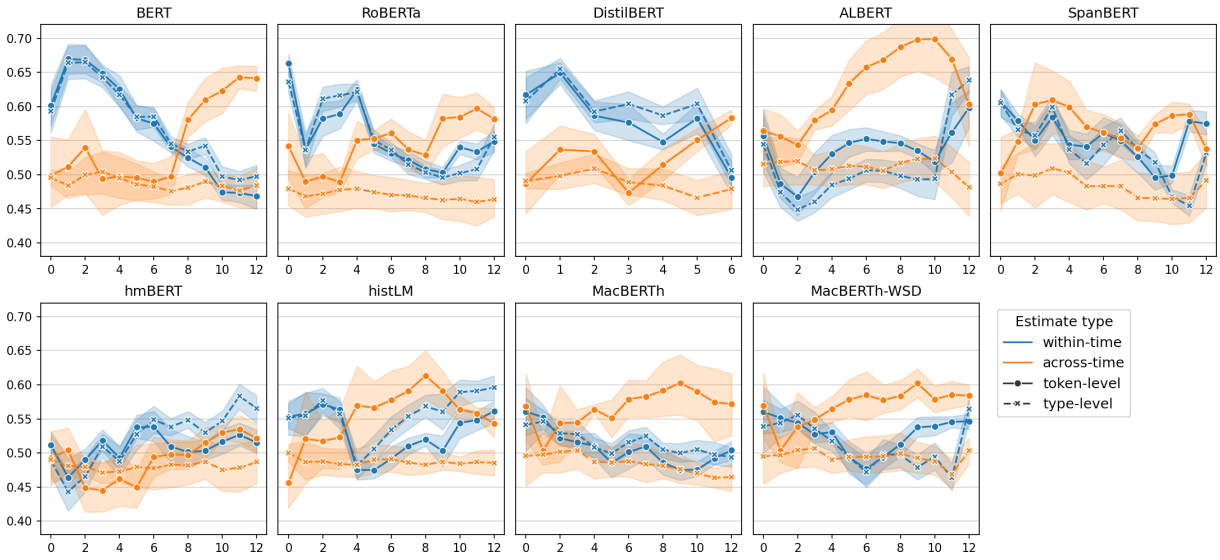

Figure 10: Modifier compositionality prediction.

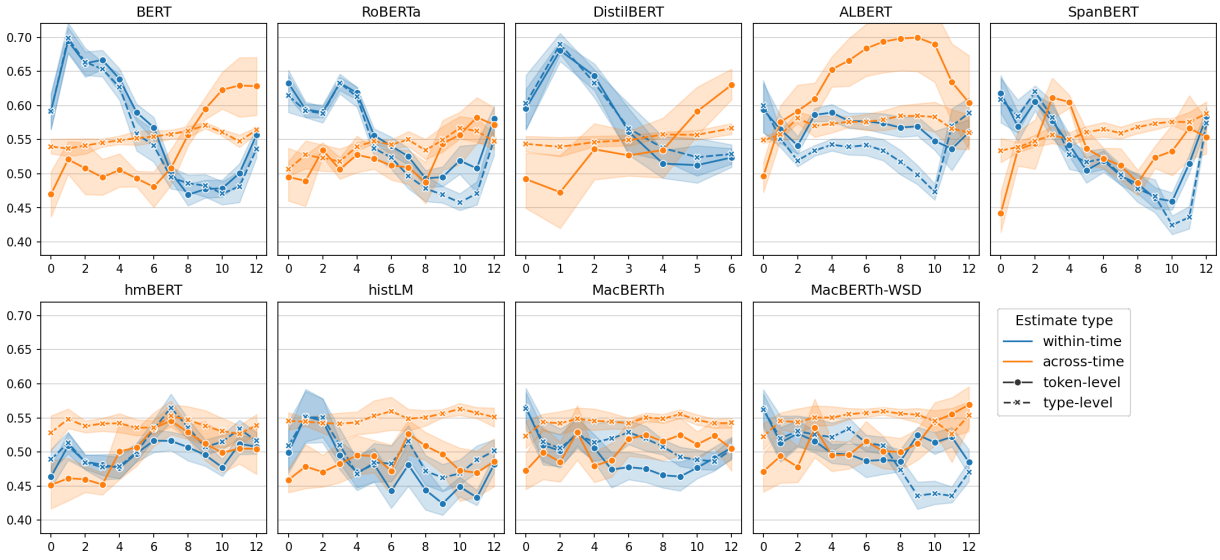

Figure 11: Head compositionality prediction.

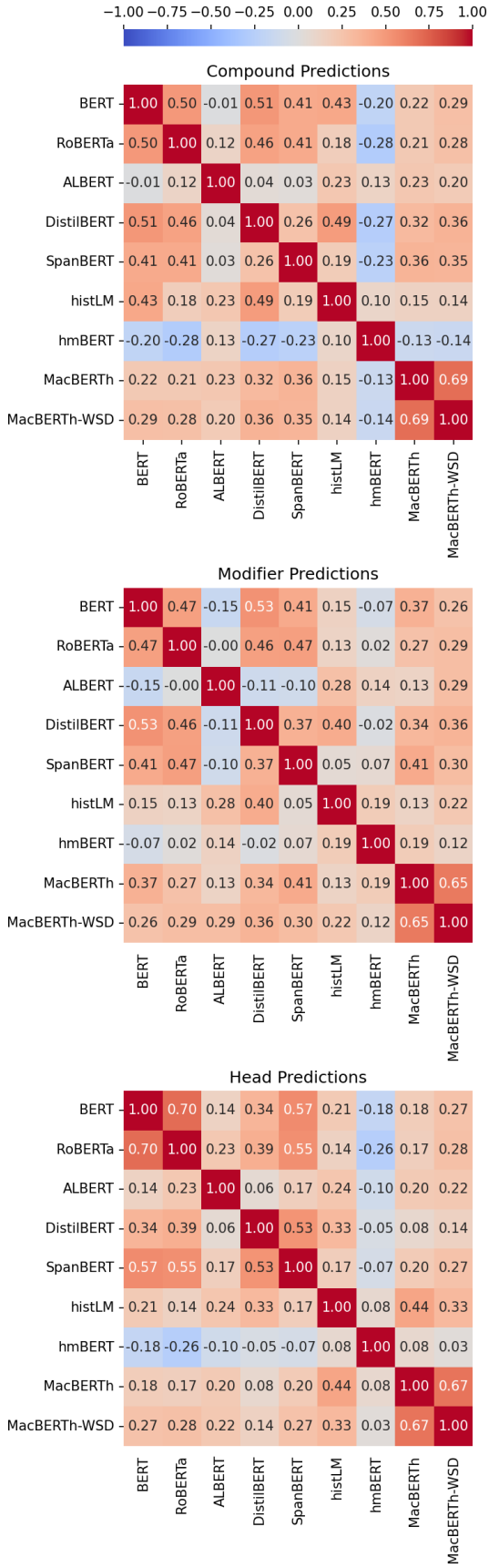

Figure 12: Correlation matrix (Spearman's ρ) comparing the performance of different transformer models across their individual implementations.

Across-model patterns of feature robustness. Moving beyond raw performance ranges (§5.3), we analyze *the similarity of different models* regarding their robustness to *constellations of experimental parameters*. We compute Spearman's correlation coefficient across pairs of models, comparing the prediction accuracy two models obtain while keeping other parameters unchanged (Figure 12).

Most BERT variants trained on present-day data are moderately correlated to one another (BERT, RoBERTa, DistilBERT, SpanBERT). This likely reflects the use of the same pretraining data mix. One exception to this trend is ALBERT, which was also pretrained on the same data, but presents a clearly distinct performance trend which aligns more closely to historical models. This pattern is indicative of a stronger effect of the architecture changes introduced with ALBERT (in particular, a strong optimization focus with parameter sharing across layers) compared to other BERT variants.

Historical BERT models exhibit weaker correlations to other models on a general level. Strikingly, though, they tend to be more strongly correlated to general-purpose models rather than to other historical models. The specifics of this trend vary across types of compositionality ratings (on the level of the compound, modifier, or head), likely due to distinct roles of compound structures in their evolution (cf. §5.1). Furthermore, hmBERT is the only model with a clear tendency towards negative correlations with other models, which may be due to it being only available as a cased and multilingual model (and not uncased and English-only, like the other models we deployed). Such effects should be taken into consideration by future work relying on off-the-shelf historical language models.

More generally, while this analysis highlights different subgroups of models, it also yields pairwise correlations which are only weak to moderate (average $\rho \approx 0.2$). This indicates that different models are sensitive to rather different sets of parameter settings, which we further explore below.

Ablation study. We examine *the robustness of each transformer model* with respect to *individual experimental parameters* through an ablation study. For each model and each type of compositionality rating (on the level of the compound, the modifier, and the head), we take as reference point the best performing experimental configuration. We then assess the drop in accuracy when replacing the value of one experimental setting at a time – layers, target embeddings, type-level/token-level processing, and temporal granularity – with other potential values for that setting, and keeping all other settings unchanged. We plot the results in Figure 13.

On a general level, different models vary noticeably with respect to their sensitivity to experimental parameters. Most remarkably, BERT obtains the single best result in our experiments, but it is also the least robust to changes in parameters, with potential drops in accuracy close to 0.4 points. The second strongest transformer model, ALBERT, exhibits a similar trend. This tendency is contrasted by models with weaker top results which are however less sensitive to parameter changes. For example, RoBERTa and DistilBERT lose up to ≈ 0.3 accuracy points with specific layer choices.

Regarding the relative effect of different parameters, the choice of layers and target embeddings clearly plays a decisive role for model performance on our task. Values alternative to the best-performing choice yield an average accuracy decrease of $\approx 0.1 - 0.2$ depending on the model. By contrast, the choice of type-level vs. token-level processing has a comparatively limited effect in within-time settings. But its role is much more pronounced – similar to that of layers – in across-time settings, where higher results are near-systematically obtained using token-level processing. Finally, the difference between fine-grained and coarse-grained time slices is overall limited, in general yielding a difference < 0.1 accuracy.

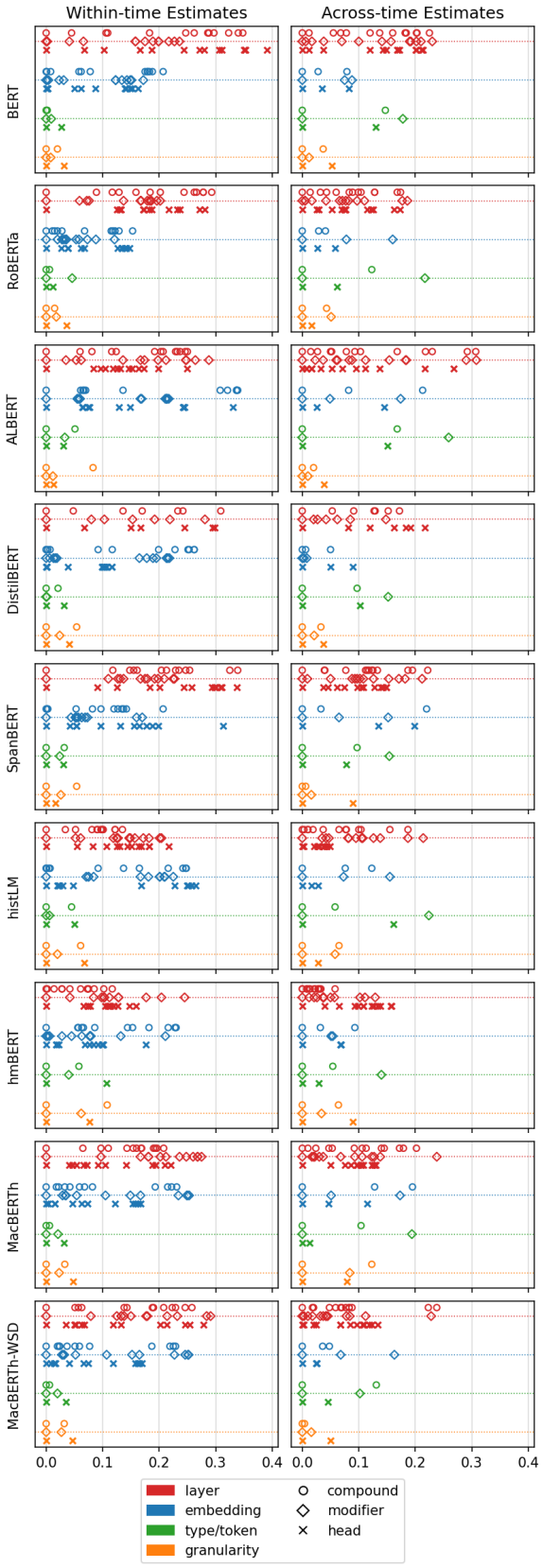

Figure 13: Drop in accuracy (x-axis) compared to the best-performing setting when manipulating the values of one experimental parameter at a time.