What Can Diachronic Contexts and Topics Tell Us About the Present-Day Compositionality of English Noun Compounds?

Samin Mahdizadeh Sani[1] **Malak Rassem**[2] **Chris Jenkins**[2]
Filip Miletić[2] **Sabine Schulte im Walde**[2]
[1]University of Tehran, Iran
[2]Institute for Natural Language Processing, University of Stuttgart, Germany
[1]samin.mehdizadeh@ut.ac.ir
[2]{malak.rassem, christopher.jenkins, filip.miletic, schulte}@ims.uni-stuttgart.de

Abstract

Predicting the compositionality of noun compounds such as *climate change* and *tennis elbow* is a vital component in natural language understanding. While most previous computational methods that automatically determine the semantic relatedness between compounds and their constituents have applied a synchronic perspective, the current study investigates what diachronic changes in contexts and semantic topics of compounds and constituents reveal about the compounds' present-day degrees of compositionality. We define a binary classification task that utilizes two diachronic vector spaces based on contextual co-occurrences and semantic topics, and demonstrate that diachronic changes in cosine similarities – measured over context or topic distributions – uncover patterns that distinguish between compounds with low and high present-day compositionality. Despite fewer dimensions in the topic models, the topic space performs on par with the co-occurrence space and captures rather similar information. Temporal similarities between compounds and modifiers as well as between compounds and their prepositional paraphrases predict the compounds' present-day compositionality with accuracy $>$0.7.

Keywords: noun compounds, compositionality, vector spaces, topic models, diachronic models

1. Introduction

One fascinating aspect of language is its compositional structure, which allows humans to create novel words and fosters creativity in compounding (Costello and Keane, 2000; Benczes, 2011) and other multi-word expressions. While understanding the meaning of multiword expressions is an easy task for native speakers, it is quite challenging for computational models to determine the meaning of a complex expression with regard to its individual parts. For example, compare the meaning of the English noun compound *climate change*, where the meaning may be obtained from the meanings of its constituents, to *snake oil* 'false panacea', where the meaning of the compound cannot be derived from either *snake* or *oil*. Building computational models to understand the meaning of complex structures is therefore a crucial component for a wide range of NLP applications. An underexplored source of information in this area is the diachronic evolution of noun compound meanings, a process connected to changes in the use of functionally similar linguistic structures (Booij, 2019) and facilitated by human speakers' ability to interpret novel compound meanings in a systematic way (Wisniewski, 1996).

Previous studies have investigated compositionality by utilizing composite functions to mathematically combine constituent vectors (Mitchell and Lapata, 2008; Coecke et al., 2010; Baroni et al., 2014; Yee and Kalita, 2016; Dima et al., 2019), or by comparing the representations of compounds and their constituents (Reddy et al., 2011; Schulte im Walde et al., 2013, 2016b; Salehi et al., 2014, 2015; Cordeiro et al., 2019; Alipoor and Schulte im Walde, 2020; Miletic and Schulte im Walde, 2023). These studies follow a synchronic approach, and up to date only a few papers have looked into what diachronic information reveals about compositionality: Dhar et al. (2019) and Dhar and van der Plas (2019) employ the Google n-grams corpus[1] and explore the effects of time span length and frequency cut-off on cosine similarity and information-theoretic measures of compositionality over time. Maurer et al. (2023) examine how diachronic changes in frequency and productivity of compounds and constituents influence the compounds' degrees of compositionality. Moreover, while previous studies have explored the role of heads and modifiers in compound compositionality (Girju et al., 2005; Ó Séaghdha, 2007; Dima et al., 2014; Schulte im Walde et al., 2016a), no work has looked into the role of prepositional variants of compounds (e.g., *climate change* is a *change in/of climate*) in synchronic or diachronic models of compositionality.

This paper investigates what diachronic changes in representations and in the semantic relatedness between compounds and constituents reveal about the compounds' present-day compositionality. We

[1]https://commondatastorage.googleapis.com/books/syntactic-ngrams/index.html

LREC-COLING 2024, pages 17449–17458

ensure comparability with our prior work in Maurer et al. (2023) by modeling a well-established set of English noun compounds in a binary classification setup, but we assess a different range of feature variants in compound and constituent representations. Our contribution is three-fold: We first compare high-dimensional co-occurrence representations against sparser but semantically more elaborate topic model distributions, when predicting degrees of compositionality. Second, we examine the role of prepositional compound paraphrases in contrast to the individual roles of modifier and head constituents in the prediction. Finally, we provide a qualitative analysis of differences in diachronic patterns for present-day low- vs. high-compositional compounds.

2. Data

In this section we present our target compounds and the diachronic corpus we use.

2.1. Gold Standard of Noun Compounds

We focus on English noun compounds where both the head and the modifier are nouns. As gold standard, we use the dataset of open English noun compounds created by Cordeiro et al. (2019), including 90 compounds from Reddy et al. (2011) along with further 190 compounds added by Cordeiro and colleagues. This dataset was annotated for the degrees of compositionality of the compounds, rated on a scale from 0 (not literal at all) to 5 (very literal). The rating was based on the semantic contribution of the modifier/head to the compound's meaning, as well as the compositionality of the whole compound phrase. We retain a subset of 210 compounds, where both constituents are nouns. Table 1 shows sample compounds and their ratings.

Compound	Compositionality Rating		
	modifier	head	compound
climate change	4.90±0.30	4.83±0.38	4.97±0.18
entrance hall	4.87±0.35	4.13±0.91	4.40±0.74
tennis elbow	2.06±1.71	4.29±1.36	2.35±1.69
crocodile tears	0.19±0.47	3.79±1.05	1.25±1.09

Table 1: Examples of compounds and their compositionality ratings (mean ± standard deviation).

2.2. Corpus

For diachronic information, we utilize the clean version of the Corpus of Historical American English (CCOHA) (Davies, 2012; Alatrash et al., 2020), a collection of texts spanning from the 1810s to the 2000s. It contains 483.6 million tokens across time slices, with the potential to provide insights into language changes. Additionally, we leverage two levels of time granularity: fine-grained and coarse-grained time slices. In the fine-grained setup, we consider individual decades from the 1810s to the 2000s, while for the coarse-grained one we combine these into six 30-year time slices starting from the 1830s.

3. Experimental Setup

Our goal is to examine the diachronic development of compound meanings and their constituents. More specifically, we focus on changes in compound meanings over time in relation to their present-day degrees of compositionality. For this, we represent the meanings of our compound targets (section 3.1) using two types of vector spaces, based on co-occurrences (section 3.2) and topic models (section 3.3), and determine cosine similarities used as relatedness features in a binary classification to differentiate between low- and high-compositional compounds (section 3.4).

3.1. Targets

We generate a range of individual vectors for compounds (e.g., *climate change*), modifiers (e.g., *climate*), heads (e.g., *change*), and prepositional compound paraphrases of their constituent combinations (e.g., *change **in** climate*, *change **of** climate*, etc.). Prepositional variants are a novel feature for this task which we use to examine one type of compound paraphrases in a controlled manner. Each of the vectors is specific for a fine or coarse time slice, e.g., one vector for *climate change* in the 1830s, one for representing it in the 1840s, etc.

3.2. Co-occurence Space

To construct standard co-occurrence vectors, we leverage both fine-grained and coarse-grained time slices. We apply a window size of ±10 words across the lemmatized and POS-tagged CCOHA corpus and count co-occurrences of target and context words, while restricting context words to content words. We do not apply any frequency threshold for context words or co-occurrences in order to avoid exacerbating data sparsity.

3.3. Topic Modeling Space

Based on the co-occurrence vectors, we create topics by applying stochastic block models (SBMs) (Peixoto, 2019), community networks which aim to avoid both overfitting and underfitting. This approach generates a graph where nodes represent compounds and their co-occurrences, with specific edge densities connecting subsets of nodes. A key advantage of this method is the fact that it creates topics in a multi-level hierarchical approach, with the number of topics determined automatically depending on the hierarchy level (Gerlach et al., 2018). We utilize levels 1, 2, and 3, resulting in 2164, 103, and 20 topics, respectively. Subsequently, each target is represented by a k-dimensional vector, where k stands for the number of topics, and the value within each dimension indicates the probability of the corresponding topic being represented in the target's vector.

3.4. Binary Classification Task

To examine the diachronic patterns in compounds and to relate them to their present-day compositionality, we define a binary classification task with the objective of distinguishing between low- and high-compositional compounds. For that, we obtain balanced classes of the 60 least and the 60 most compositional compounds, compound-modifier and compound-head pairs, i.e., the extreme two sets in compositionality, while ignoring those in the mid ranges. We expect the distinction between extremely (non-)compositional compound sets to provide a clear picture of differences in compound properties (see Schulte im Walde et al. (2016a); Alipoor and Schulte im Walde (2020); Schulte im Walde and Frassinelli (2022); Maurer et al. (2023) for similar extreme-based classification constellations). Table 2 shows the ranges of the mean ratings for the 3×2 extreme sets. We then apply a support vector machine (SVM) as the binary classifier between the low- vs. high-compositional, and present accuracy results.

	Compositionality Range	
	low	high
compound	[0.18, 1.52]	[4.26, 5.00]
modifier	[0.00, 2.80]	[4.60, 5.00]
head	[0.14, 1.41]	[4.65, 5.00]

Table 2: Ranges of mean ratings for compound phrases and compound-constituent combinations.

As diachronic feature input for the classifier we create feature vectors $v(w_1, w_2) = [v_1, v_2, ..., v_n]$, where v_i represents the cosine similarity score between the co-occurrence/topic vectors of target words w_1 and w_2 (e.g., a compound and its modifier; a compound and a prepositional paraphrase; etc.) in the i^{th} time slice. As target vectors we use those of the compounds, the modifiers, and the heads, as well as a "constituent" vector (summing the modifier and head vectors for a specific compound), a "preposition" vector (summing over the vectors of a compound's prepositional paraphrases), and a "combined" vector (summing over a compound and all its prepositional paraphrase vectors).

While our focus is on diachronic evolution, we also carry out a comparison with a standard static approach. For this purpose, we train our SVM classifier using the last time slice of either granularity: the 2000s in the fine-grained setup, and 1980s-2000s in the coarse-grained setup. Additionally, repeated k-fold cross-validation is employed in all of our experimental scenarios to address concerns of data sparsity and overfitting, using 20 repetitions with different permutations of the data and 4 folds per repetition.

4. Results

Table 3 presents the results for predicting present-day compound compositionality across variants of diachronic similarities between w_1 and w_2 vectors, while relying on fine-grained time slices, which include the overall best results. Results for coarse-grained time slices as well as for alternative numbers of topics are provided in Appendix A. Regarding our experiments on the last time slice, we report the best individual results out of all comparisons of w_1 and w_2 vectors; full results can be found in Appendix B.

Most of the results are clearly above the baseline, with best accuracy scores per column between 0.667 and 0.745. The highest accuracy in our diachronic setup for both spaces is reached when predicting compound compositionality using compound–modifier vector similarity, i.e., the temporal development of vector-space similarity between compounds and their modifiers represents the strongest predictor of the compounds' present-day (non-)compositionality. The lowest results are obtained when we predict head or modifier compositionality employing the opposite constituents (e.g., using the compound–head similarity to predict the modifier's contribution to the compound's meaning), while employing the same constituent is rather successful (e.g., using the compound–modifier vector to predict the modifier's contribution to the compound's meaning). Among constituents and also prepositional paraphrases, the modifiers are generally the most reliable in similarity-based predictions for compound compositionality (also in the "combined" setup), while the heads are the worst; similarities between compounds and prepositional paraphrases show in-between results, i.e., they are still more reliable than compound–head similarities. Interestingly, in almost all cases the compound–constituent comparisons reach accuracy scores between compound–modifier and compound–head comparisons; we conclude that the comparisons do not benefit from knowing about both constituents, i.e., the constituents do not seem to provide complementary information regarding compound meaning.

Comparing the vector spaces, the results of the

Features	Accuracy					
	compound		modifier		head	
	coocc.	topic	coocc.	topic	coocc.	topic
random	0.500	0.500	0.500	0.500	0.500	0.500
best last	**0.749**	0.683	**0.743**	0.645	0.645	0.615
diachronic cosine similarity for w_1–w_2						
compound–modifier	*0.741*	***0.745***	*0.703*	***0.706***	0.627	0.621
compound–head	0.673	0.697	0.585	0.590	0.678	0.666
compound–constituents	0.710	0.701	0.626	0.635	0.658	***0.667***
compound–preposition	0.710	0.716	0.650	0.666	0.653	0.650
combined–modifier	0.733	0.683	0.695	0.669	0.631	0.540
combined–head	0.704	0.617	0.609	0.504	***0.695***	0.637
combined–constituents	0.721	0.703	0.633	0.630	0.666	0.666

Table 3: Classification results for co-occurrence and topic (k = 2164) spaces relying on fine-grained time slices, and for our static approach. Bold values represent the overall best results per column, while italic values indicate the best diachronic setting per column. The overall top three values in the diachronic setup are highlighted in green.

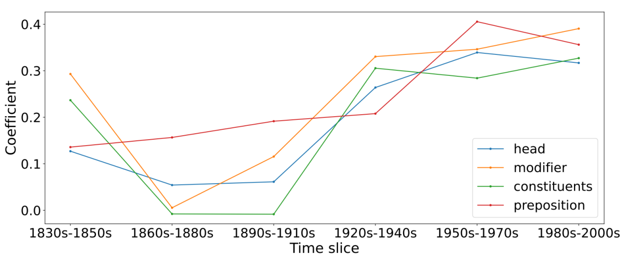

Figure 1: Correlations between cosine similarity scores and human compositionality ratings, for compound targets.

topic modeling approach vary depending on the number of topics k predetermined by the algorithm (cf. section 3.3). They are comparable to the co-occurrence approach only when $k = 2164$, and they deteriorate with a reduction in the number of topics. Although an explanation is not immediately apparent, the topic-based approach is competitive in the setup involving the “compound” vector, but it clearly lags behind the co-occurrence approach when relying on the “combined” vector (obtained by summing the vector of the compound and those of its prepositional variants). It is also interesting to note that the number of vector dimensions is significantly lower in the topic than in the co-occurrence spaces (cf. Appendix C); these dimensions are arguably the ones that capture the most relevant information, as indicated by overall stability in performance.

When comparing the use of the static approach with the diachronic one, we observe that the static approach outperforms the diachronic setup in most of our co-occurrence experiments. This indicates – somewhat unsurprisingly – that most aspects of present-day compositionality are easier to predict from co-occurrence patterns that are the closest to the present. However, the reverse occurs when employing the topic modeling approach, meaning that the diachronic setting is more effective within this framework. Nonetheless, this finding coupled with the fact that accuracy is rather high in all cases described above, confirms that diachronic developments reveal distinctive patterns related to present-day compositionality.

5. Qualitative Analysis

Figure 1 presents correlations between cosine similarities over time in relation to human compositionality ratings. As can be seen, the strongest correlations are associated with compound–modifier and compound–preposition similarities in most time slices. This aligns with our classification results, where the accuracies for compound–modifier and compound–preposition ranked among the highest.

We also examine where the model fails: Figure 2 plots the mean values of the diachronic cosine similarity features in cases where the model makes accurate predictions (hits) or inaccurate ones (misses), in order to determine whether misclassified examples deviate from the overall pattern observed in low-/high-compositional subsets. As the plot illustrates, in both the low- and high-compositional subsets the correctly classified examples display patterns quite similar to the overall mean of their corresponding subset. On the

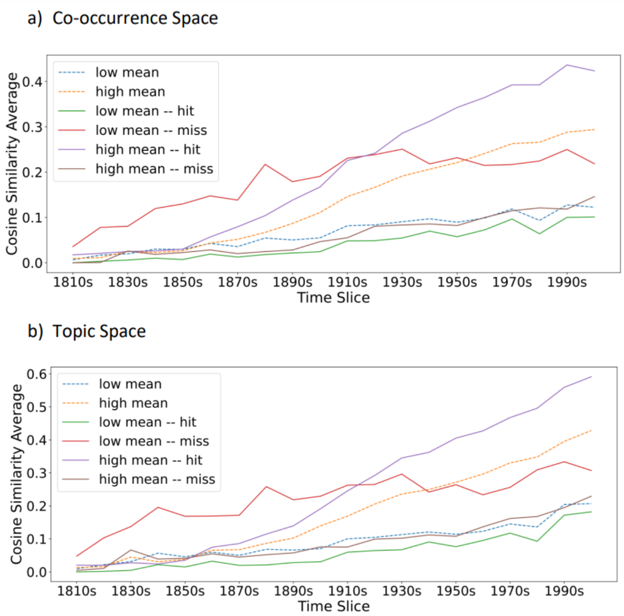

Figure 2: Averages of diachronic cosine similarity features for the low- and high-compositionality subsets, along with their respective means for misclassified and accurately classified examples.

other hand, misclassified examples tend to move towards the pattern of the opposite class, leading to erroneous predictions made by the classifier. Particularly for the misclassified examples within the high-compositional compounds, their diachronic developments closely mirror the trend represented by the average of low-compositional subsets. This observation explains why our classifier does not perform well on high-compositional subsets.

Overall, we also observe that as time progresses, the cosine similarities increase in both groups. However, the increase is more noticeable for the high-compositional compounds, and as we move forward that distinction between the two classes becomes more evident. This observation offers a further explanation for the high accuracies achieved by our static approach, which only uses data from the last time slice.

To investigate whether the two vector spaces lead to misclassifications of the same or different target items, we also calculate how strongly the misclassified items overlap. For the co-occurrence space, the proportion of misclassified items shared with the topic approach is 0.895; vice versa, for the topic space with 2,164 topics, it is 0.919. In other words, both approaches struggle with a similar set of items, indicating that they capture very similar information.

6. Discussion and Conclusion

This study presented a series of binary classification experiments to investigate the connection between the evolution of English noun compound meanings over time and their present-day degrees of compositionality. We proposed two vector spaces in a diachronic approach, with both co-occurrence and topic modeling spaces performing comparably well with accuracy >0.7. As the number of dimensions is significantly lower in the topic modeling approach, the topic dimensions are arguably well-suited to our semantic task. We also observed a stronger contribution of modifiers (and prepositional paraphrases) than heads in their distributional indication of compound compositionality.

Our results thus suggest that diachronic features, based on temporal sequences of cosine similarities, capture distinctive patterns related to the compounds' present-day compositionality levels. Additionally, a qualitative analysis reveals that as we move forward in time, differences between high- and low-compositional compounds become more pronounced. These observations more generally provide a better understanding of the way in which compound meanings evolve over time.

As previously noted, our setup is similar to a prior study (Maurer et al., 2023) which leverages diachronic frequency and productivity of compounds to classify them for present-day compositionality. The key results of that paper align with our findings, confirming that when using diachronic features, predicting the compositionality of compounds is slightly easier than predicting the meaning contributions of heads or modifiers. From a diachronic linguistic standpoint, we observe the same pattern in the temporal development of compounds: a progressive increase in the differences between low- and high-compositional compounds, as well as more pronounced changes in high-compositional compounds compared to the low-compositional ones over the time span captured by our data. From a methodological standpoint, the best results for predicting compound compositionality are obtained using compound–head comparisons in the prior study, and compound–modifier comparisons in our experiments. This points to a different nature of the compositionality information captured by the evolution of frequency and productivity vs. vector space similarities. More generally, the best classification accuracy in the diachronic setting is higher in the current approach (0.74) than in our previous study (0.66). This finding confirms the utility of more complex representational information capturing co-occurrence and topic patterns, as opposed to more directly accessible quantitative information such as frequency and productivity.

7. Acknowledgements

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*).

8. Bibliographical 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.

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

Réka Benczes. 2011. Blending and creativity in metaphorical compounds: A diachronic investigation. In Hans-Jörg Schmid and Sandra Handl, editors, *Windows to the Mind: Metaphor, Metonymy and Conceptual Blending*, pages 247–268. Mouton de Gruyter.

Geert Booij. 2019. Compounds and multi-word expressions in Dutch. In Barbara Schlücker, editor, *Complex Lexical Units*, pages 95–126. De Gruyter, Berlin, Boston.

Bob Coecke, Mehrnoosh Sadrzadeh, and Stephen Clark. 2010. Mathematical foundations for a compositional distributional model of meaning. *arXiv preprint arXiv:1003.4394*.

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

Fintan J. Costello and Mark T. Keane. 2000. Efficient creativity: Constraint-guided conceptual combination. *Cognitive Science*, 24(2):299–349.

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

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.

Corina Dima, Daniël de Kok, Neele Witte, and Erhard Hinrichs. 2019. No word is an island – A transformation weighting model for semantic composition. *Transactions of the Association for Computational Linguistics*, 7:437–451.

Corina Dima, Verena Henrich, Erhard Hinrichs, and Christina Hoppermann. 2014. How to tell a Schneemann from a Milchmann: An annotation scheme for compound-internal relations. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1194–1201, Reykjavik, Iceland. European Language Resources Association (ELRA).

Martin Gerlach, Tiago P. Peixoto, and Eduardo G. Altmann. 2018. A network approach to topic models. *Science Advances*, 4(7):eaaq1360.

Roxana Girju, Dan Moldovan, Marta Tatu, and Daniel Antohe. 2005. On the semantics of noun compounds. *Computer Speech Language*, 19(4):479–496. Special issue on Multiword Expression.

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.

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.

Jeff Mitchell and Mirella Lapata. 2008. Vector-based models of semantic composition. In *Proceedings of the Association for Computational Linguistics - Human Language Technologies*, pages 236–244, Columbus, Ohio. Association for Computational Linguistics.

Diarmuid Ó Séaghdha. 2007. Annotating and learning compound noun semantics. In *Proceedings of the Association for Computational Linguistics 2007 Student Research Workshop*, pages 73–78, Prague, Czech Republic. Association for Computational Linguistics.

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.

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.

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.

Sabine Schulte im Walde and Diego Frassinelli. 2022. Distributional measures of abstraction. *Frontiers in Artificial Intelligence: Language and Computation 4:796756. Alessandro Lenci and Sebastian Pado (topic editors): "Perspectives for Natural Language Processing between AI, Linguistics and Cognitive Science"*.

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).

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.

Edward J. Wisniewski. 1996. Construal and similarity in conceptual combination. *Journal of Memory and Language*, 35(3):434–453.

Kyra Yee and Jugal Kalita. 2016. Composition of compound nouns using distributional semantics. In *Proceedings of the 13th International Conference on Natural Language Processing*, pages 20–29, Varanasi, India. NLP Association of India.

A. Diachronic Approach Results

Features	Accuracy					
	compound		modifier		head	
	coarse	fine	coarse	fine	coarse	fine
compound - modifier	0.718	**0.741**	**0.670**	**0.703**	0.572	0.627
compound - head	0.687	0.673	0.581	0.585	0.671	0.678
compound - constituents	0.682	0.710	0.615	0.626	0.601	0.658
compound - preposition	0.719	0.710	**0.670**	0.650	0.639	0.653
combined - modifier	**0.725**	**0.733**	0.669	0.695	0.590	0.631
combined - head	0.686	0.704	0.600	0.609	**0.684**	**0.695**
combined - constituents	0.699	0.721	0.613	0.633	0.595	0.666

Table 4: Full classification results for the co-occurrence space. Bold values show the best diachronic settings among different combinations in each setup. The top three values are highlighted in green.

Features	Accuracy		
	compound	modifier	head
k=2164			
compound - modifier	**0.745**	**0.706**	0.621
compound - head	0.697	0.590	0.666
compound - constituents	0.701	0.635	**0.667**
compound - preposition	0.716	0.666	0.650
combined - modifier	0.683	0.669	0.540
combined - head	0.617	0.504	0.637
combined - constituents	0.703	0.630	0.666
k=103			
compound - modifier	0.675	0.636	0.583
compound - head	0.631	0.549	0.593
compound - constituents	0.627	0.585	0.585
compound - preposition	**0.721**	**0.668**	**0.628**
combined - modifier	0.625	0.590	0.531
combined - head	0.587	0.526	0.560
combined - constituents	0.589	0.524	0.547
k=20			
compound - modifier	0.546	0.513	0.503
compound - head	0.546	0.498	0.510
compound - constituents	0.551	0.516	0.557
compound - preposition	**0.697**	**0.645**	**0.617**
combined - modifier	0.540	0.508	0.479
combined - head	0.543	0.500	0.494
combined - constituents	0.532	0.497	0.501

Table 5: Full classification results for the topic space. The reported accuracy scores are for fine-grained time slices. Bold values show the best diachronic settings among different combinations in each setup. The top three values are highlighted in green.

B. Static Approach Results

Features	Accuracy					
	compound		modifier		head	
	coarse	fine	coarse	fine	coarse	fine
compound - modifier	0.748	0.731	0.690	0.738	0.587	0.588
compound - head	0.712	0.639	0.568	0.507	0.686	0.635
compound - constituents	0.651	0.622	0.549	0.518	0.584	0.557
compound - preposition	0.722	0.682	0.666	0.627	0.638	0.605
combined - modifier	**0.769**	**0.749**	**0.702**	**0.743**	0.602	0.579
combined - head	0.720	0.639	0.587	0.524	**0.693**	**0.645**
combined - constituents	0.673	0.639	0.589	0.549	0.582	0.566

Table 6: Full classification results for the co-occurrence space using the last time slice. Bold values are the best results. The top three values are highlighted in green.

Features	Accuracy		
	compound	modifier	head
k=2164			
compound - modifier	**0.683**	**0.645**	0.572
compound - head	0.645	0.538	**0.615**
compound - constituents	0.596	0.546	0.487
compound - preposition	0.648	0.621	0.611
combined - modifier	0.543	0.503	0.509
combined - head	0.605	0.634	0.567
combined - constituents	0.591	0.550	0.478
k=103			
compound - modifier	0.668	**0.663**	0.580
compound - head	0.650	0.590	**0.602**
compound - constituents	0.615	0.575	0.577
compound - preposition	**0.702**	0.623	**0.602**
combined - modifier	0.560	0.613	0.459
combined - head	0.570	0.579	0.474
combined - constituents	0.501	0.542	0.417
k=20			
compound - modifier	0.671	**0.645**	**0.635**
compound - head	0.567	0.502	0.596
compound - constituents	0.621	0.580	0.634
compound - preposition	**0.696**	0.620	0.604
combined - modifier	0.559	0.527	0.557
combined - head	0.442	0.427	0.508
combined - constituents	0.548	0.452	0.579

Table 7: Full classification results for the topic space using the last time slice. The reported accuracy scores are for fine-grained time slices. Bold values are the best results. The top three values are highlighted in green.

C. Co-occurrence vector sizes across time slices

Time slice	1810s	1820s	1830s	1840s	1850s
Vector size	26,037	67,562	107,551	11,6345	117,912
Time slice	1860s	1870s	1880s	1890s	1900s
Vector size	130,504	127,363	133,251	135,107	154,318
Time slice	1910s	1920s	1930s	1940s	1950s
Vector size	145,804	180,780	207,433	206,424	219,877
Time slice	1960s	1970s	1980s	1990s	2000s
Vector size	218,822	226,366	216,653	211,838	240,127

Table 8: Vector sizes for fine-grained setting.

Time slice	1830s-1850s	1860s-1880s	1890s-1910s
Vector size	233,034	265,967	298,455
Time slice	1920s-1940s	1950s-1970s	1980s-2000s
Vector size	433,741	492,554	484,757

Table 9: Vector sizes for coarse-grained setting.