Classifying Noun Compounds for Present-Day Compositionality: Contributions of Diachronic Frequency and Productivity Patterns

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

Abstract

We investigate the diachronic evolution of the frequency and productivity of English noun compounds and their constituents relative to their degree of compositionality. We focus on 185 compounds with human compositionality ratings and a range of quantitative information from a large diachronic corpus. We cast our task as binary classification, and show that both diachronic frequency and productivity are useful in determining the present-day degree of compositionality of English noun compounds.

1 Introduction

Multiword expressions such as noun compounds (e.g. *flea market*) are semantically idiosyncratic to some degree, i.e. the meaning of the full expression is not entirely (or even not at all) predictable from the meanings of its constituents (Sag et al., 2002; Baldwin and Kim, 2010). While noun compounds have been extensively explored across research disciplines from synchronic perspectives, this paper provides a novel diachronic approach to predict their present-day compositionality.

More specifically, we investigate the diachronic evolution of the frequency and productivity of English noun compounds and their constituents relative to their degree of compositionality. Our analysis relies on an established gold standard dataset with human compositionality ratings, and a diachronic corpus of English covering approximately two centuries. We hypothesize that distinct frequency and productivity patterns of diachronic evolution can be observed for compounds whose degree of compositionality is high (such as *maple tree*, *prison guard*, *climate change*) vs. low (such as *flea market*, *night owl*, *melting pot*). We cast our task as a binary classification problem, and show that both diachronic frequency and productivity provide useful information in determining the present-day degree of compositionality of English noun compounds.

2 Related work

Existing computational studies have examined noun compounds from a range of perspectives. Common approaches include predicting the meaning of the whole compound (Mitchell and Lapata, 2008; Dima et al., 2019), the semantic relations between a compound's constituents (Girju et al., 2005; Ó Séaghdha, 2007; Dima et al., 2014), and the compound's degree of compositionality, usually framed as an unsupervised ranking task relying on static (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) or contextualized word embeddings (Garcia et al., 2021a,b; Miletic and Schulte im Walde, 2023). A small subset of previous work has also taken into account the distinct linguistic roles and empirical characteristics of compound constituents, showing that compositionality prediction is affected by properties such as frequency, productivity, and ambiguity (Schulte im Walde et al., 2013, 2016; Alipoor and Schulte im Walde, 2020; Miletic and Schulte im Walde, 2023; Schulte im Walde, 2023). However, all of the cited studies adopt a synchronic perspective. As to our knowledge, only two previous approaches applied a diachronic perspective: Dhar et al. (2019) and Dhar and van der Plas (2019) exploited the Google n-gram corpus and information-theoretic as well as cosine distance measures to predict the compositionality of the compounds in Reddy et al. (2011), and to detect novel compounds, respectively.

In this paper, we provide a novel diachronic approach motivated from a linguistic perspective: we expect the present-day degree of compositionality to differ for high- vs. low-frequent compounds and for compounds with high- vs. low-frequent constituents (Lee, 1990; Hamilton et al., 2016, i.a.), as well as for compounds with high- vs. low-productive constituents (Jurafsky et al., 2001;

Proceedings of the 19th Conference on Natural Language Processing (KONVENS 2023), pages 40–51
September 18–22, 2023.

Hilpert, 2015, i.a.). We further compare the diachronic features against the use of present-day linguistic properties so as to assess the scope of compositionality information recovered through our diachronic approach.

3 Data

3.1 Gold standard of noun compounds

We use the collection of English noun compounds introduced by Cordeiro et al. (2019). It includes an initial set of 90 compounds created by Reddy et al. (2011)[1] and a further 190 compounds annotated by Cordeiro and colleagues using the same rating procedure.[2] Of these, we retain a total of 210 compounds for which both constituents are tagged as nouns in the dataset.

Human annotators were asked to provide compositionality ratings in terms of literality, on a scale from 0 (not at all literal) to 5 (very literal). They provided scores for the interpretation of the whole compound (e.g. *crash course*), as well as for the use of the modifier (*crash*) and the head (*course*) within it. Sample compounds and their ratings are shown in Table 1.

	Compositionality rating		
Compound	Modifier	Head	Compound
guinea pig	0.47 ± 0.72	0.47 ± 0.72	0.24 ± 0.56
flea market	0.38 ± 0.81	4.71 ± 0.84	1.52 ± 1.13
pain killer	4.71 ± 0.64	1.33 ± 1.11	2.05 ± 1.36
health insurance	4.53 ± 0.88	4.83 ± 0.58	4.40 ± 1.17

Table 1: Sample gold standard compounds with compositionality ratings (mean and standard deviation).

3.2 Corpus

As diachronic corpus data for the modeled noun compounds, we rely on the clean version of the Corpus of Historical American English (CCOHA) (Davies, 2012; Alatrash et al., 2020). It contains >400 million words, and ranges from 1810 to 2010. For present-day data, we use ENCOW (Schäfer and Bildhauer, 2012; Schäfer, 2015), a large web corpus that contains ≈9.5 billion tokens. Both corpora are lemmatized, tagged and parsed.

3.3 Empirical diachronic properties

We retrieve the following empirical diachronic properties per decade for our target compounds and their constituents:

- The *frequencies* of the gold standard compounds and their constituents.
- The *productivities* of the constituents of the gold standard compounds, i.e. the number of compounds a constituent appears in: morphological family size (de Jong et al., 2002).

For the latter, we consider a construction to be a relevant (candidate) compound if it is tagged as a sequence of two nouns, neither preceded nor followed by a noun.

4 Experimental setup

To assess whether highly compositional compounds and their constituents exhibit distinct patterns of diachronic evolution of productivity and frequency, we divide the 185 compounds that occur in at least one timeslice in CCOHA into different classes of compositionality. We do that for three types of compositionality ratings: on the level of the whole compound, the modifier, and the head. We cast our task as binary classification of the extremes with maximally different targets regarding their levels of compositionality, thus enforcing a clear picture of distinctiveness.

More specifically, we obtain balanced classes of the 62 least and most compositional compounds, modifiers, and heads ($\frac{1}{3}$ of the targets within each class, leaving out 61 mid-scale items). The compositionality ranges for the sets of least/most compositional compounds are $[0.18, 1.61]$ and $[4.20, 5.00]$, respectively. For the least/most compositional modifiers, the compositionality ranges are $[0.14, 1.76]$ and $[4.56, 5.00]$. For the least/most compositional heads, they are $[0.00, 2.79]$ and $[4.50, 5.00]$.

We conduct experiments for two levels of granularity of timeslices, in order to assess whether temporally finer-grained patterns provide more information related to present-day compositionality, with the potential trade-off of increasing sparsity. In the setup with finer-grained timeslices, we consider decades from the 1830s to the 2000s; in the coarser-grained setup, we combine these decades into 30-year timeslices. Since the sub-corpora of the two earliest decades, the 1810s and 1820s, are considerably smaller than the subsequent ones, we disregard those. Table 2 provides a summary of the sizes of our timeslices in millions of tokens.

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

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

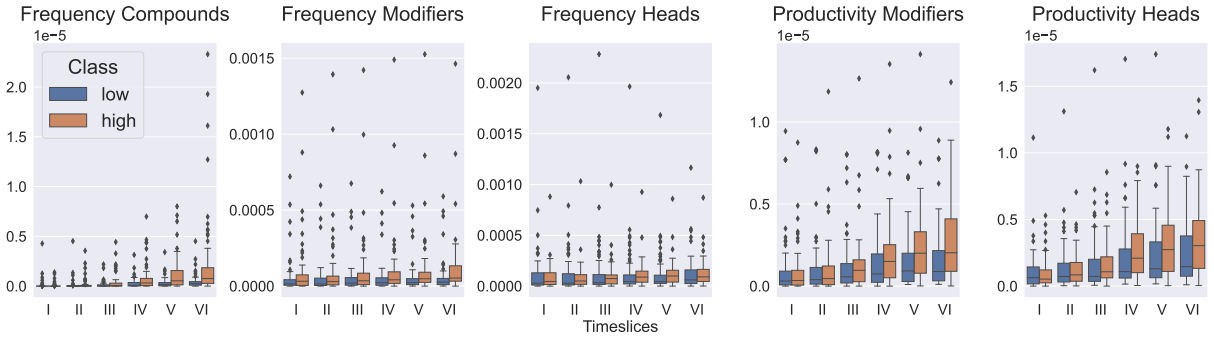

Figure 1: Development of the properties over time per class for the compound compositionality experiment. Timeslices: I: 1830s-1850s, II: 1860s-1880s, III: 1890s-1910s, IV: 1920s-1940s, V: 1950s-1970s, VI: 1980s-2000s.

Timeslice	1830s	1840s	1850s	1860s	1870s	1880s
Total$_{fine}$	16.7	19.4	20.0	20.6	22.6	24.4
Total$_{coarse}$	56.1			67.6		
Timeslice	1890s	1900s	1910s	1920s	1930s	1940s
Total$_{fine}$	24.6	26.7	27.7	31.2	30.1	29.9
Total$_{coarse}$	79.0			91.2		
Timeslice	1950s	1960s	1970s	1980s	1990s	2000s
Total$_{fine}$	30.3	29.6	29.4	31.3	34.6	36.5
Total$_{coarse}$	89.3			100.4		

Table 2: Timeslice sizes for the fine- and coarse-grained timeslices in million tokens.

We assess whether the two compositionality classes for the compounds and for the modifier and head constituents, respectively, have distinct patterns of diachronic evolution in terms of five empirical properties: the compound frequency F_C; the frequency F_M and productivity P_M of the modifier; and the frequency F_H and productivity P_H of the head. For each of the properties, we construct feature vectors $V = [v_1, v_2, \ldots, v_n]$ containing the retrieved values of the respective property across n timeslices. To account for differences in the corpus sizes of the timeslices, each retrieved property value is normalized by the total number of tokens in the respective timeslice. In configurations where we use multiple properties, their feature vectors are concatenated.

Figure 1 outlines the development of each of the empirical properties over the coarse timeslices, for the respective two classes defined for compound-level compositionality; see Appendix C for properties across constituent classes. Appendix E shows to which degree the properties correlate with each other across timeslices. We report Spearman's rank-order correlation coefficient ρ. In most cases the properties do not correlate at all, or just moderately. We find strong correlations only between the frequency and productivity of a constituent within the same timeslice, with an average $\rho = 0.77$ for modifiers and 0.88 for heads.

We conduct experiments using each of the properties individually, using the combination of the frequency of both constituents and the productivity of both constituents (F_{MH} and P_{MH}), the combination of all frequency measures F_{CMH}, and the combination of all features $F_{CMH}P_{MH}$. Other permutations in the following are denoted by combinations of the contained properties (e.g. $P_M F_H$).

In all experimental settings, we use a support vector machine (SVM) as the classifier. To account for data sparsity and overfitting in our results, we evaluate with repeated k-fold cross-validation, using 8 repetitions with different permutations of the compound data and 4 folds per repetition.

Even though our focus is on diachronic evolution, we also compare our approach against a standard static approach, using only synchronic information from (i) the last CCOHA timeslice of either granularity and (ii) present-day information retrieved from ENCOW. For each of the five empirical properties, we order the targets in descending order by that property and assign the positive label[3] to the first N compounds. More specifically, we collect results for all potential class splits, moving from $N = 0$, i.e. no compound is assigned to the positive class, to $N = 124$, i.e. all compounds are assigned to the positive class.

5 Results

The results of our classification experiments are shown in Table 3, which focuses on individual properties as features, as well as combinations of frequency measures, productivity measures, and all five collected properties as features. It further

[3]We refer to the class of highly compositional compounds as the positive class.

	Accuracy					
	Compound		Modifier		Head	
Features	coarse	fine	coarse	fine	coarse	fine
Random	0.500	0.500	0.500	0.500	0.500	0.500
Best last	0.694	0.702	0.710	0.702	0.669	0.637
Best ENCOW	**0.782**		**0.831**		**0.669**	
F_C	***0.663***	***0.665***	0.595	0.600	***0.631***	***0.633***
F_M	0.585	0.597	***0.649***	***0.629***	0.457	0.455
F_H	0.649	0.647	0.519	0.523	0.627	0.617
F_{MH}	0.637	0.643	0.605	0.624	0.592	0.595
F_{CMH}	0.654	0.644	0.594	0.620	0.570	0.576
P_M	0.629	0.626	0.632	0.606	0.457	0.448
P_H	0.571	0.564	0.502	0.472	0.554	0.550
P_{MH}	0.612	0.597	0.610	0.607	0.538	0.518
$F_{CMH}P_{MH}$	0.619	0.634	0.590	0.608	0.568	0.574

Table 3: Classification results for the three experiments per property used as features. We report accuracy for coarse- and fine-grained time slices, as well as the best last coarse- and fine-grained timeslices and the best ENCOW setting. Bold values are the best overall, and bold italic values are the best diachronic settings.

reports the best results for each static synchronic setting. In the coming discussion, we also reference additional combinations of features that are relevant for specific setups. We provide the full results of all permutations of features for each of the experiments in Appendix D. Regarding the static synchronic approach, the effect of positive class size on the compound compositionality experiment is shown in Appendix B.

Overall, we find that all diachronic properties are informative for compound compositionality and that the properties of a given constituent are informative for the compositionality of that constituent (e.g. P_H for head compositionality). The results for combinations of properties indicate that they are informative if they include an informative property. In most cases, however, results for combinations are below those for included properties.

Across the target properties, the best settings of all static synchronic approaches outperform our diachronic setup. This is not especially surprising: our aim is to predict the present-day degree of compositionality, and (near-)present-day data is likely better suited to this task. Moreover, the best synchronic results are systematically obtained using ENCOW data, which is ≈100 times larger than the last coarse CCOHA slice; this suggests that the diachronic approach is hindered by data sparsity. Nevertheless, its performance is well above chance, which confirms that diachronic developments capture distinct patterns with respect to present-day compositionality. Since this issue is the main focus of our work, we limit the remaining discussion of results to our diachronic experiments.

Compound compositionality. All configurations of properties from both granularities of timeslices significantly outperform the random choice baseline ($p < 0.001$).[4] Amongst the individual properties and main combinations summarized in Table 3, F_C performs best, followed by F_{CMH}; this applies both to the fine-grained and the coarse-grained setup. Combinations of properties tend to perform similarly to the most informative property in them. A noteworthy exception is $F_C P_M$, which obtains the best overall result with an accuracy of 0.675 for the coarse-grained and 0.702 for the fine-grained timeslice setup. We hypothesize that this is due to the information of both properties being complementary, as indicated by a weak correlation (average $\rho = 0.30$ per timeslice). Regarding modifier properties, P_M is more informative than F_M. This is flipped for head properties F_H and P_H.

The results do not differ significantly between timeslice granularities, changing in the range of $\pm 1.5\%$. This indicates that the diachronic development of properties retrieved from coarse-grained timeslices is as informative as their counterparts from finer-grained timeslices. We hypothesize that this may be due to two potential reasons. (i) We observe in our data that considerable change in the properties either happens fairly quickly or slowly over time (cf. Section 6). Both are captured to a similar extent in both timeslice granularities. (ii) Despite providing more detailed information, the fine-grained developments may be more susceptible to sparsity and ultimately may not be more informative than the coarse-grained timeslices.

Modifier compositionality. In contrast to the compound compositionality experiment, F_M appears to be most informative for the compositionality of the modifier, followed by P_M. Similarly to the first experiment, combinations of properties are fairly informative, but less so than the most informative property in them. With their results not differing significantly from the random baseline, both F_H and P_H appear to be uninformative for predicting the modifier compositionality class. As expected, we observe that the properties of the modifier are relevant for modifier compositionality, while the properties of the head are not. Similarly to the compound compositionality experiment, results generally do not differ significantly between timeslice granularities for settings with results well above the random baseline. There is a significant

[4] All significance tests were done using the chi-square test.

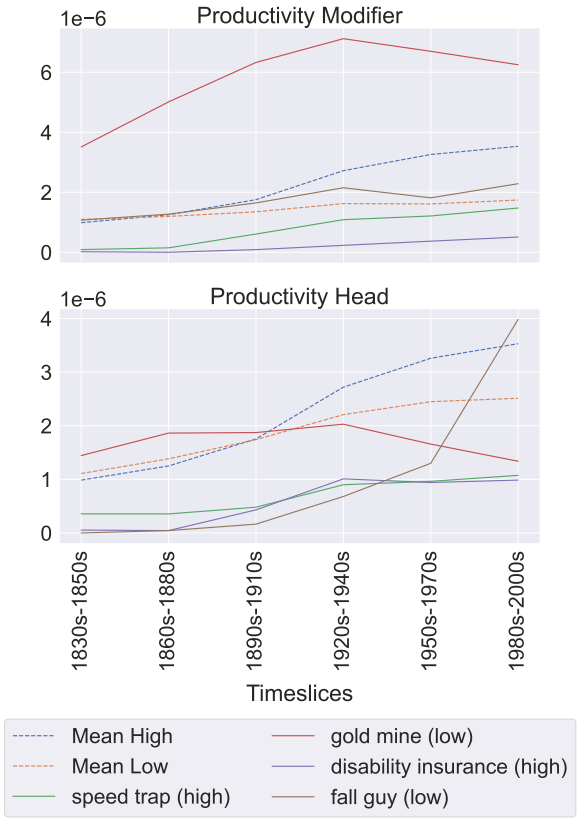

Figure 2: Diachronic development of the productivity of the modifier and of the head for sample compounds. For each example, the compositionality class is indicated in parentheses. Dashed lines indicate the means for the two classes.

difference for the settings P_M and F_{CMH}, but the trends point in opposite directions without an immediately apparent explanation.

Head compositionality. Similarly to the compound compositionality experiment, F_C is the most informative feature. This is in line with the dominant linguistic role of the head in compound structure. The remaining results are overall comparable to the modifier compositionality experiment but flipped: F_H is more informative than P_H, and the results of F_M and P_M are worse than the random baseline. Results do not differ significantly between timeslice granularities.

6 Qualitative analysis

To further assess when the patterns of diachronic development are informative for the classification of present-day compositionality, we inspect where the models fail. We find that, over all the runs, across features and experiments, low-compositionality compounds are misclassified more often than highly compositional ones.

We look more closely into examples from both classes that are misclassified in over 80% of runs in the compound compositionality experiment. Some misclassified compounds of either class exhibit a diachronic evolution profile that clearly differs from the mean trend for their class. For instance, the trend in P_H for *fall guy* (low compositionality) is more similar to the overall trend of the high compositionality class, with a steep increase in later timeslices, while we observe the inverse for *speed trap* (high compositionality), see Figure 2. This, however, does not appear to be the only issue at stake, since profiles of misclassified instances also differ within a class, e.g. for *fall guy* and *gold mine*.

On a more general level, frequently misclassified compounds from both classes exhibit similar patterns in similar ranges for most properties, for instance *speed trap* and *gold mine* (cf. Figure 2 for productivity and Appendix A for frequency evolution). Since the means of both classes are similar to one another across properties, we hypothesize that patterns close to the means or below may be too similar across classes to be informative.

7 Conclusion

We presented experiments aimed at classifying English noun compounds in terms of their present-day degree of compositionality. We proposed a novel diachronic approach, relying on the evolution of frequency and productivity patterns for compounds and their constituents. Both types of features are informative, with our single best diachronic classifier combining the strongest individual variants of frequency and productivity features. The highest performance overall is obtained by a synchronic method based on a much larger present-day corpus, but our diachronic approach is still indicative of distinct compound development profiles relative to their degree of compositionality. This overall demonstrates the relevance of diachronic data in modeling noun compounds, thereby confirming the potential of this under-researched area.

Acknowledgments

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

Limitations

Our experiments were limited to two quantitative properties – frequency and productivity – used to analyze noun compounds in a single language, English. This has potential implications for the generalizability of our results. From a linguistic standpoint, compound properties vary widely across languages. For instance, where English has productive patterns combining two nouns, often in an open (space-separated) compound, German has closed compounds; Romance languages widely rely on N-Prep-N patterns; the structure in many Slavic languages involves patterns of nominal declension; and so forth. The most useful diachronic information for compositionality prediction may vary across these cases. Future work may also investigate the diachronic evolution of other compound properties, such as the degree of ambiguity of the constituents or the semantic relations between them.

Ethical considerations

We do not believe that the research presented in this paper raises ethical concerns. We analyzed the diachronic evolution of a specific type of linguistic structure in English, based on standard aggregate estimates of word usage derived from a large corpus. No personally identifiable or otherwise sensitive information was targeted by our modeling approach. Previously created datasets were used in line with their intended use and licenses.

We acknowledge the fact that the corpus we used contains documents written in American English over the last two centuries. It therefore likely captures biases mirroring the societal inequalities typical of the time in which those texts were produced. However, we do not expect general quantitative properties of a small subset of the vocabulary – on which we relied – to be significantly affected by any potential biases.

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.

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

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

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

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).

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.

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

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

Martin Hilpert. 2015. From *hand-carved* to *computer-based*: Noun-participle compounding and the upward strengthening hypothesis. *Cognitive Linguistics*, 26(1):1–36.

Daniel Jurafsky, Alan Bell, Michelle Gregory, and William D. Raymond. 2001. Probabilistic relations between words: Evidence from reduction in lexical production. In Joan Bybee and Paul Hopper, editors, *Frequency and the Emergence of Linguistic Structure*, Typological Studies in Language, pages 229–254. John Benjamins, Amsterdam / Philadelphia.

Christopher J. Lee. 1990. Some hypotheses concerning the evolution of polysemous words. *Journal of Psycholinguistic Research*, 19(4):211–219.

Filip Miletic 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.

Jeff Mitchell and Mirella Lapata. 2008. Vector-based models of semantic composition. In *Proceedings of ACL-08: HLT*, pages 236–244, Columbus, Ohio. Association for Computational Linguistics.

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

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

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. In *Proceedings of the Conference on Intelligent Text Processing and Computational Linguistics*, Mexico City, Mexico.

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.

Roland Schäfer. 2015. Processing and querying large web corpora with the COW14 architecture. In *Proceedings of Challenges in the Management of Large Corpora 3 (CMLC-3)*, Lancaster.

Roland Schäfer and Felix Bildhauer. 2012. Building large corpora from the web using a new efficient tool chain. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012)*, pages 486–493, Istanbul, Turkey. European Language Resources Association.

Sabine Schulte im Walde. 2023. Collecting and investigating features of compositionality ratings. In Voula Giouli and Verginica Barbu Mititelu, editors, *Multiword Expressions in Lexical Resources. Linguistic, Lexicographic and Computational Perspectives*, Phraseology and Multiword Expressions. Language Science Press, Berlin, Germany.

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.

A Frequency for target examples over time

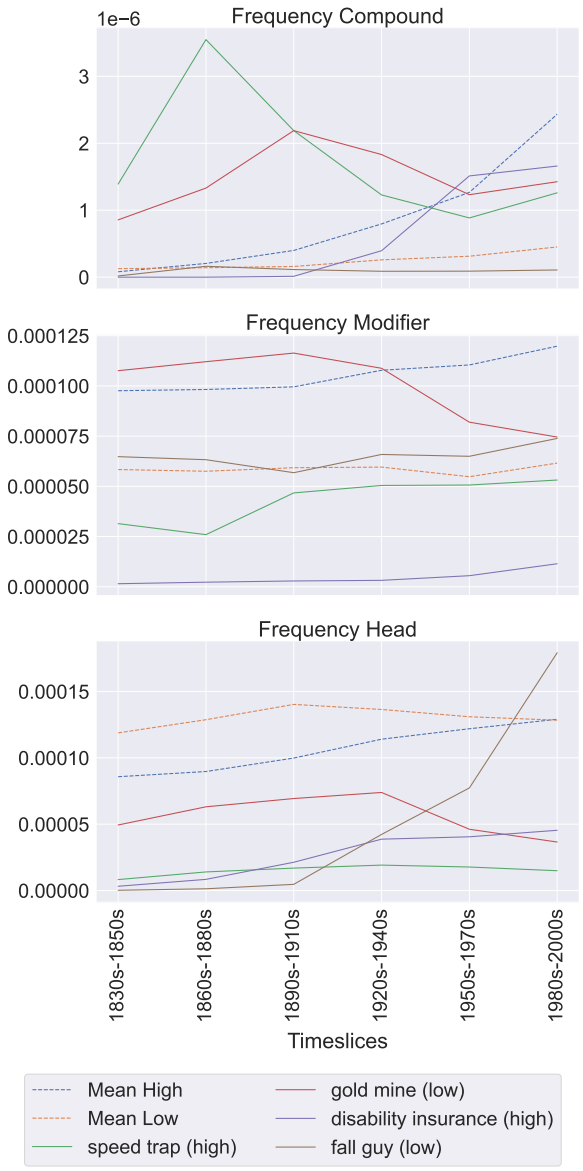

Figure 3: Frequency over time for examples. The class of an example is indicated in parentheses.

B Effect of varying set size in synchronic experiments

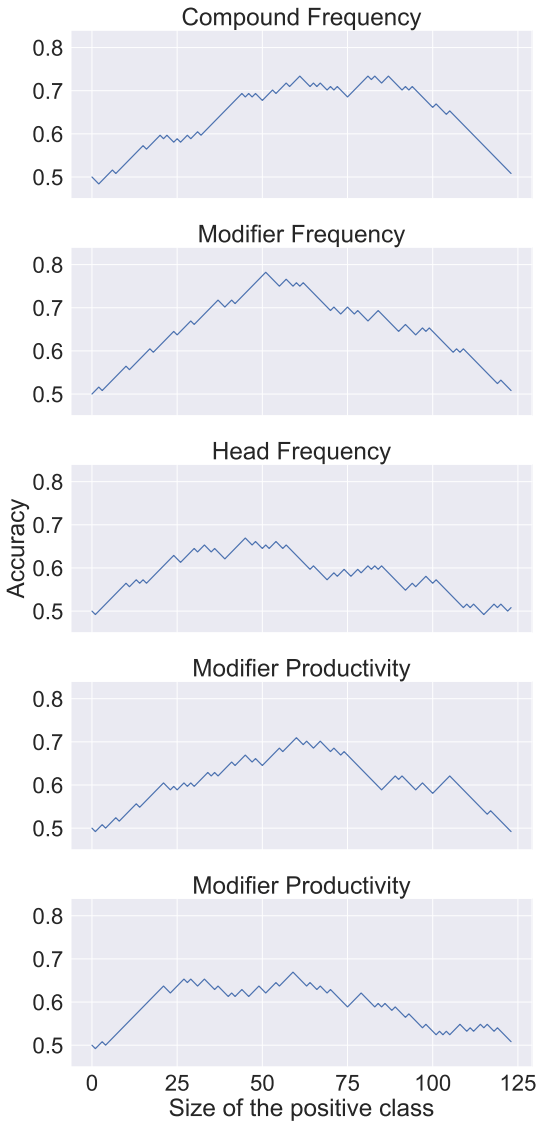

Figure 4: Static compound experiment results per positive class size with ENCOW data.

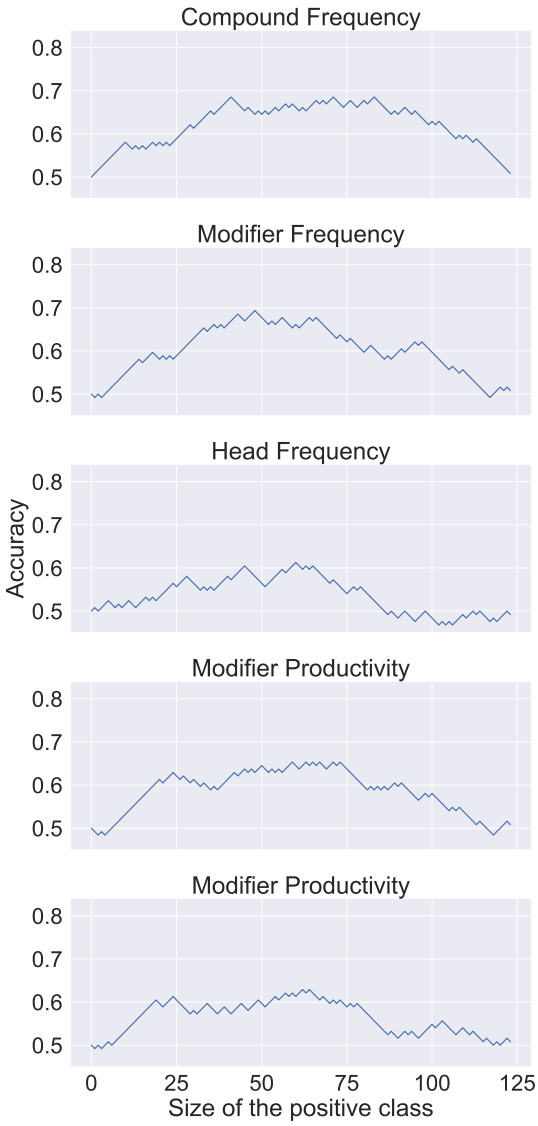

Figure 5: Static compound experiment results per positive class size with the last coarse timeslice.

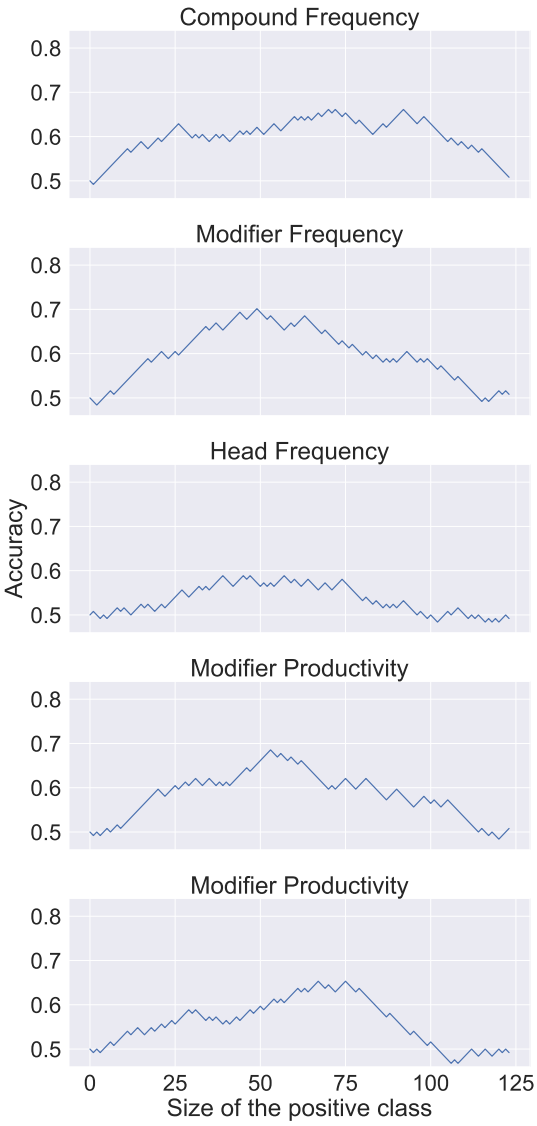

Figure 6: Static compound experiment results per positive class size with the last fine timeslice.

C Development of properties over time

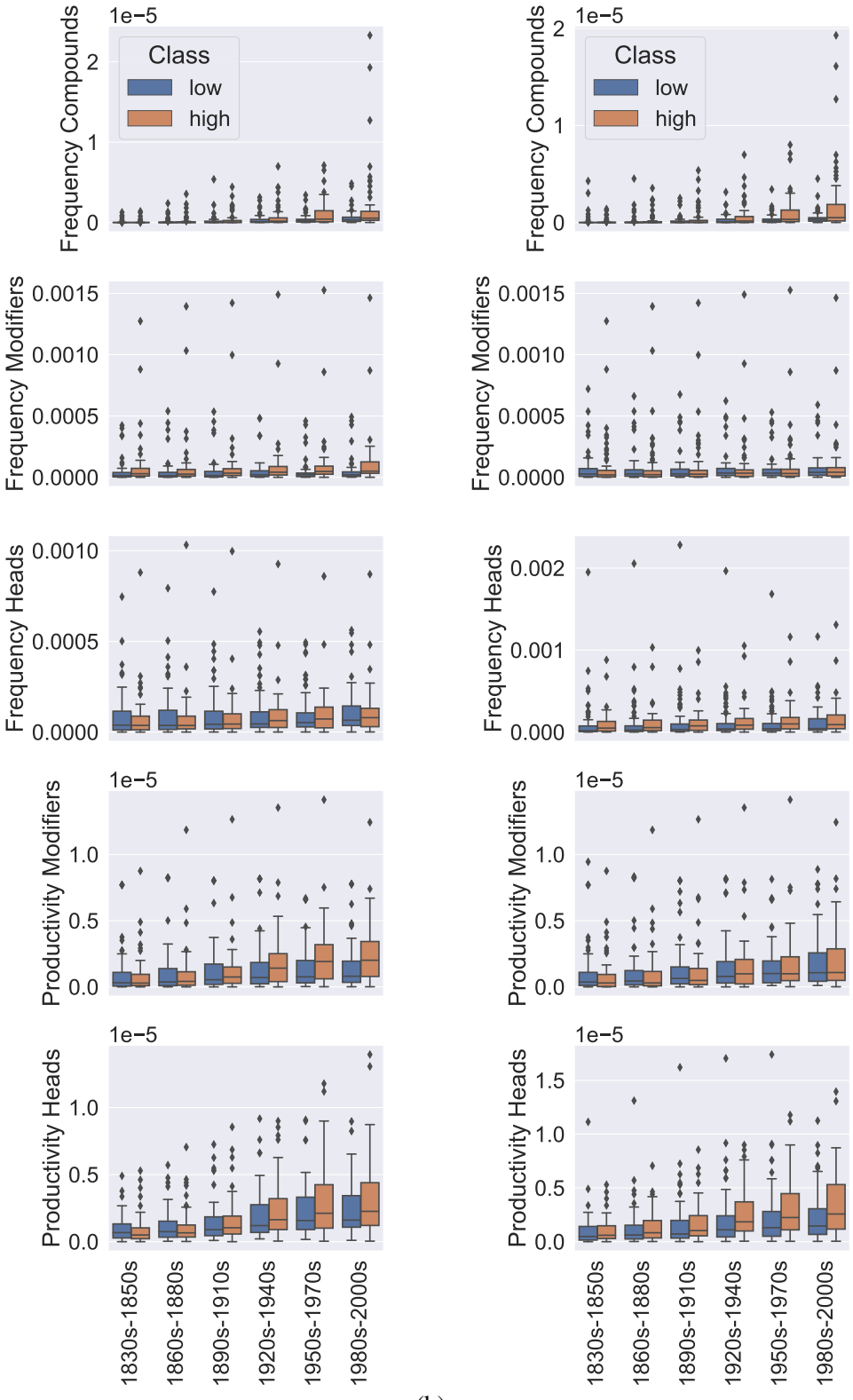

Figure 7: (a) Development of the properties over time per class for the modifier compositionality experiment. (b) Development of the properties over time per class for the head compositionality experiment.

D Full results

	Accuracy					
	Compound		Modifier		Head	
Features	coarse	fine	coarse	fine	coarse	fine
F_C	0.663	0.665	0.595	0.600	0.631	0.633
F_M	0.585	0.597	0.649	0.629	0.457	0.455
F_H	0.649	0.647	0.519	0.523	0.627	0.617
F_{MH}	0.637	0.643	0.605	0.624	0.592	0.595
F_{CMH}	0.629	0.635	0.594	0.620	0.570	0.576
P_M	0.629	0.626	0.632	0.606	0.457	0.448
P_H	0.571	0.564	0.502	0.472	0.555	0.550
P_{MH}	0.612	0.597	0.610	0.607	0.538	0.518
$F_M P_M$	0.579	0.590	0.638	0.634	0.461	0.457
$F_M P_H$	0.579	0.589	0.639	0.635	0.459	0.456
$F_M P_{MH}$	0.578	0.589	0.637	0.635	0.458	0.455
$F_C P_M$	0.675	0.702	0.651	0.652	0.554	0.558
$F_C P_H$	0.630	0.626	0.575	0.579	0.614	0.615
$F_C P_{MH}$	0.662	0.650	0.614	0.621	0.588	0.590
$F_H P_M$	0.654	0.644	0.500	0.509	0.620	0.613
$F_H P_H$	0.654	0.644	0.504	0.510	0.620	0.614
$F_H P_{MH}$	0.654	0.639	0.497	0.510	0.620	0.612
$F_{MH} P_M$	0.629	0.636	0.595	0.618	0.572	0.577
$F_{MH} P_H$	0.630	0.635	0.594	0.618	0.573	0.577
$F_{MH} P_{MH}$	0.624	0.636	0.588	0.614	0.569	0.570
F_{CM}	0.579	0.590	0.637	0.635	0.458	0.457
$F_{CM} P_M$	0.578	0.590	0.637	0.634	0.458	0.455
$F_{CM} P_H$	0.577	0.589	0.638	0.635	0.457	0.456
$F_{CM} P_{MH}$	0.580	0.590	0.638	0.634	0.457	0.452
F_{CH}	0.651	0.644	0.504	0.509	0.620	0.613
$F_{CH} P_M$	0.654	0.638	0.494	0.510	0.619	0.612
$F_{CH} P_H$	0.655	0.638	0.500	0.509	0.620	0.612
$F_{CH} P_{MH}$	0.654	0.641	0.499	0.501	0.619	0.610
$F_{CMH} P_M$	0.622	0.636	0.589	0.614	0.567	0.570
$F_{CMH} P_H$	0.623	0.637	0.588	0.614	0.568	0.571
$F_{CMH} P_{MH}$	0.619	0.634	0.590	0.608	0.568	0.574

Table 4: Full classification results for the three experiments per property used as features. We report accuracy for properties retrieved from coarse- and fine-grained time slices.

	Accuracy														
	Compound					Modifier					Head				
Features	coarse	fine	c. last	f. last	ENC.	coarse	fine	c. last	f. last	ENC.	coarse	fine	c. last	f. last	ENC.
F_C	0.663	0.665	0.686	0.662	**0.734**	0.595	0.600	0.629	0.637	**0.702**	0.631	0.633	0.669	0.637	**0.669**
F_M	0.585	0.597	0.694	0.702	**0.782**	0.649	0.629	0.710	0.702	**0.831**	0.457	0.455	0.548	0.573	**0.605**
F_H	0.649	0.647	0.613	0.589	**0.669**	0.519	0.523	0.565	0.573	**0.605**	0.627	0.617	0.645	0.621	**0.661**
P_M	0.629	0.626	0.653	0.685	**0.710**	0.632	0.606	0.653	0.661	**0.710**	0.457	0.448	0.540	0.556	**0.573**
P_H	0.571	0.564	0.629	0.653	**0.669**	0.502	0.472	0.556	**0.629**	0.613	0.554	0.550	**0.637**	**0.637**	**0.637**

Table 5: Results per feature including synchronic/last timeslices results. For the experiment using the last timeslices/synchronic data, we report the best result across positive class sizes. Best result per feature and compositionality setting is bolded. Abbreviations: *c. last* = last coarse timeslice, *f. last* = last fine timeslice, *ENC.* = ENCOW.

E Correlations between properties over time

Timeslice	P_M-P_H	P_M-F_M	P_M-F_H	P_M-F_C	P_H-F_M	P_H-F_H	P_H-F_C	F_M-F_H	F_M-F_C	F_H-F_C
1830s-1850s	-0.06	0.79	-0.11	0.34	-0.03	0.89	0.35	0.32	0.18	0.32
1860s-1880s	-0.06	0.78	-0.11	0.44	-0.07	0.90	0.32	0.29	0.31	0.29
1890s-1910s	-0.12	0.79	-0.11	0.41	-0.05	0.90	0.33	0.28	0.36	0.28
1920s-1940s	-0.14	0.80	-0.14	0.37	-0.06	0.90	0.26	0.26	0.30	0.26
1950s-1970s	-0.12	0.78	-0.14	0.27	-0.01	0.89	0.28	0.26	0.26	0.26
1980s-2000s	-0.06	0.79	-0.13	0.19	0.04	0.87	0.26	0.21	0.26	0.21

Table 6: Correlations between properties per coarse-grained timeslice.

Timeslice	P_M-P_H	P_M-F_M	P_M-F_H	P_M-F_C	P_H-F_M	P_H-F_H	P_H-F_C	F_M-F_H	F_M-F_C	F_H-F_C
1830s	-0.03	0.76	-0.09	0.35	-0.03	0.87	0.31	0.27	0.16	0.27
1840s	-0.02	0.75	-0.09	0.33	-0.04	0.87	0.33	0.27	0.18	0.27
1850s	-0.06	0.76	-0.12	0.31	-0.03	0.88	0.29	0.24	0.15	0.24
1860s	-0.02	0.77	-0.10	0.36	-0.05	0.88	0.28	0.24	0.21	0.24
1870s	-0.09	0.77	-0.13	0.36	-0.06	0.89	0.35	0.28	0.24	0.28
1880s	-0.03	0.77	-0.08	0.40	-0.04	0.88	0.35	0.30	0.25	0.30
1890s	-0.08	0.77	-0.09	0.41	-0.03	0.90	0.33	0.28	0.29	0.28
1900s	-0.10	0.77	-0.11	0.30	-0.03	0.90	0.32	0.27	0.26	0.27
1910s	-0.13	0.78	-0.12	0.41	-0.05	0.87	0.31	0.24	0.33	0.24
1920s	-0.13	0.77	-0.14	0.36	-0.06	0.90	0.28	0.21	0.28	0.21
1930s	-0.13	0.81	-0.13	0.39	-0.05	0.89	0.31	0.29	0.31	0.29
1940s	-0.12	0.78	-0.12	0.33	-0.02	0.89	0.27	0.26	0.26	0.26
1950s	-0.13	0.77	-0.14	0.27	-0.04	0.89	0.29	0.26	0.22	0.26
1960s	-0.10	0.77	-0.12	0.29	-0.01	0.88	0.24	0.23	0.27	0.23
1970s	-0.10	0.77	-0.13	0.26	0.00	0.88	0.26	0.25	0.24	0.25
1980s	-0.07	0.77	-0.13	0.21	0.02	0.87	0.31	0.27	0.24	0.27
1990s	-0.06	0.78	-0.11	0.26	0.04	0.87	0.28	0.23	0.29	0.23
2000s	-0.07	0.79	-0.12	0.17	0.02	0.87	0.20	0.16	0.24	0.16

Table 7: Correlations between properties per fine-grained timeslice.