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

University of Stuttgart
Institute for Natural Language Processing

Samin Mahdizadeh Sani, Malak Rassem, Chris Jenkins,
Filip Miletić, Sabine Schulte im Walde

Introduction

The association between the meanings of noun compounds and the meanings of their constituents is not always the same.

- **climate change**: an alternation in climate patterns
- **snake oil:** false panacea

The diachronic evolution of noun compound meanings is an underexplored source of information.

- It is crucial for comprehending shifts in language usage.
- It is connected to changes in how similar linguistic structures are used within a language (Booij, 2019).

Can we use diachronic information to predict present-day compositionality?

Contributions

1. Comparing high-dimensional co-occurrence representations with sparser, semantically more elaborate topic model distributions to predict compositionality.

2. Examining the roles of prepositional compound paraphrases in prediction, (e.g., *climate change* ≈ *change **of** climate; change **in** climate;* etc.).

3. Providing a qualitative analysis of diachronic patterns for present-day low- vs. high-compositional compounds (e.g., *entrance hall* vs. *tennis elbow*).

Data

Gold Standard of Noun Compounds (Cordeiro et al., 2019)

- 210 English noun-noun compounds
- annotated by humans for the degrees of compositionality of the compounds

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

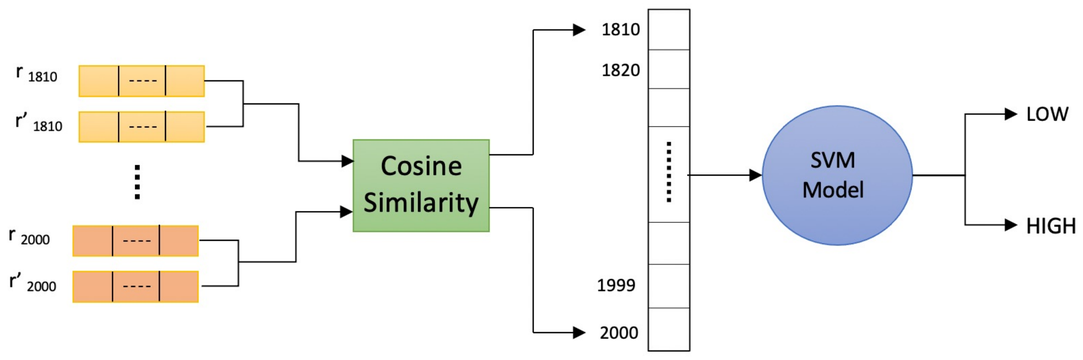
Diachronic corpus – CCOHA (Davies, 2012; Alatrash et al., 2020)

- a collection of texts spanning from the 1810s to the 2000s

Two levels of time granularity

- fine-grained: individual decades from the 1810s to the 2000s
- coarse-grained: six 30-year time slices starting from the 1830s

Method



Targets (created exclusively for each time slice):

- compounds (e.g., *climate change*)
- modifiers (e.g., *climate*); heads (e.g., *change*); both constituents
- prepositional compound paraphrases (e.g., *change in climate*)

Target Features & Binary Classification:

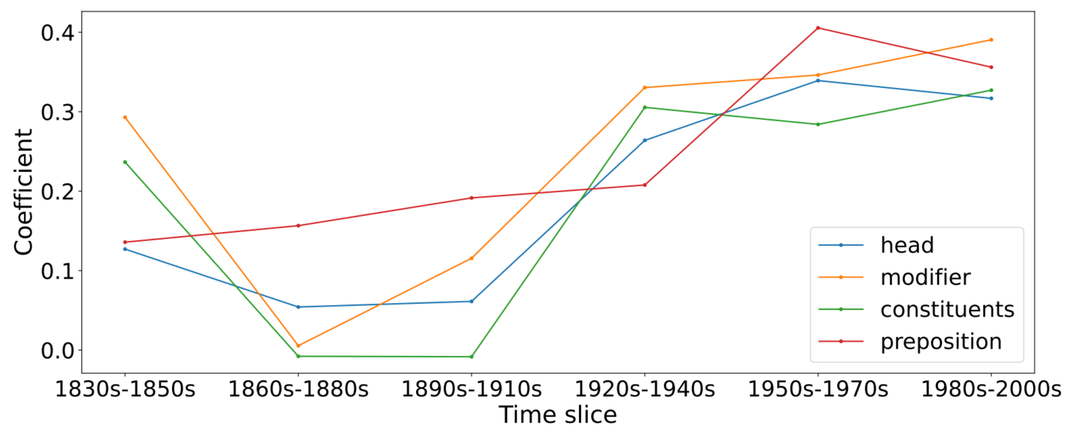
- represent target meanings via their contextual co-occurrences (window ±10); two models: direct co-occurrence vs. distribution over topic models
- create cosine relatedness vectors $\vec{v}(w_1,w_2) = \langle r_1, r_2, \ldots, r_n \rangle$ across *n* time slices
- apply a Support Vector Machine classifier to distinguish between the 60 least and the 60 most compositional compounds (low/high)

Results

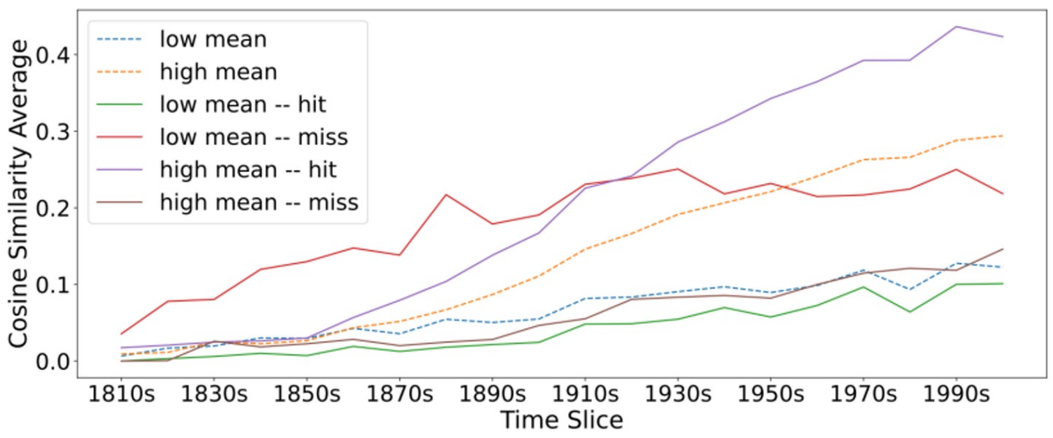
Features	Accuracy					
	compound		modifier		head	
	coocc.	topic	coocc.	topic	coocc.	topic
random	0.500	0.500	0.500	0.500	0.500	0.500
best last	**0.749**	0.683	**0.743**	0.645	0.645	0.615
diachronic cosine similarity for w_1–w_2						
compound–modifier	*0.741*	***0.745***	*0.703*	***0.706***	0.627	0.621
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

- **Strongest predictor** of present-day (non-)compositionality: compound–modifier vector similarities over time.
- **Prepositional paraphrases:** more reliable than compound–head similarities.
- **Constituents vectors (head+modifier):** Compound–constituent comparisons perform in between compound–modifier and compound–head comparisons. The **constituents do not seem to provide complementary information** regarding compound meaning.
- **Vector spaces:** The topic space **performs on par** with the co-occurrence space.
- Accuracy is rather high in all cases, confirming that **diachronic developments reveal distinctive patterns** related to present-day compositionality.

Qualitative Analysis



- Strongest correlations with human compositionality ratings are shown by compound–modifier and compound–preposition similarities in most time slices.



- Both co-occurrence and topic approaches capture rather similar information.
- As time progresses, the cosine similarities increase in both low- and high-compositional subsets.
- The increase is more noticeable for the high-compositional compounds.

Conclusion

- Diachronic features, based on temporal sequences of cosine similarities, capture distinctive patterns related to the compounds' present-day compositionality levels.
- Despite fewer dimensions in the topic models, the topic space performs on par with the co-occurrence space and captures rather similar information.
- As time progresses, differences between high- and low-compositional compounds become more pronounced.

References

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

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

[3] Reem Alatrash, Dominik Schlechtweg, Jonas Kuhn, and Sabine Schulte im Walde. 2020. CCOHA: Clean corpus of historical American English. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6958–6966, Marseille, France. European Language Resources Association.

Language Resources and Evaluation (LREC-COLING), May 2024