

Overview

The goal of our project is to develop a model which captures speakers' intuition regarding constructions and their extensibility. To achieve this, we aim to:



- i. characterize constructions with regards to their usage, as attested in linguistic corpora, and
- ii. predict speakers' evaluation of newly coined instantiations of these constructions.

Research questions:

- ⚙️ How does the diversity profile of a construction affect its extensibility?
- ⚙️ What determines whether coinages are instances of constructional productivity or creativity?

Case study: Hebrew possessive constructions

Two competing constructions:

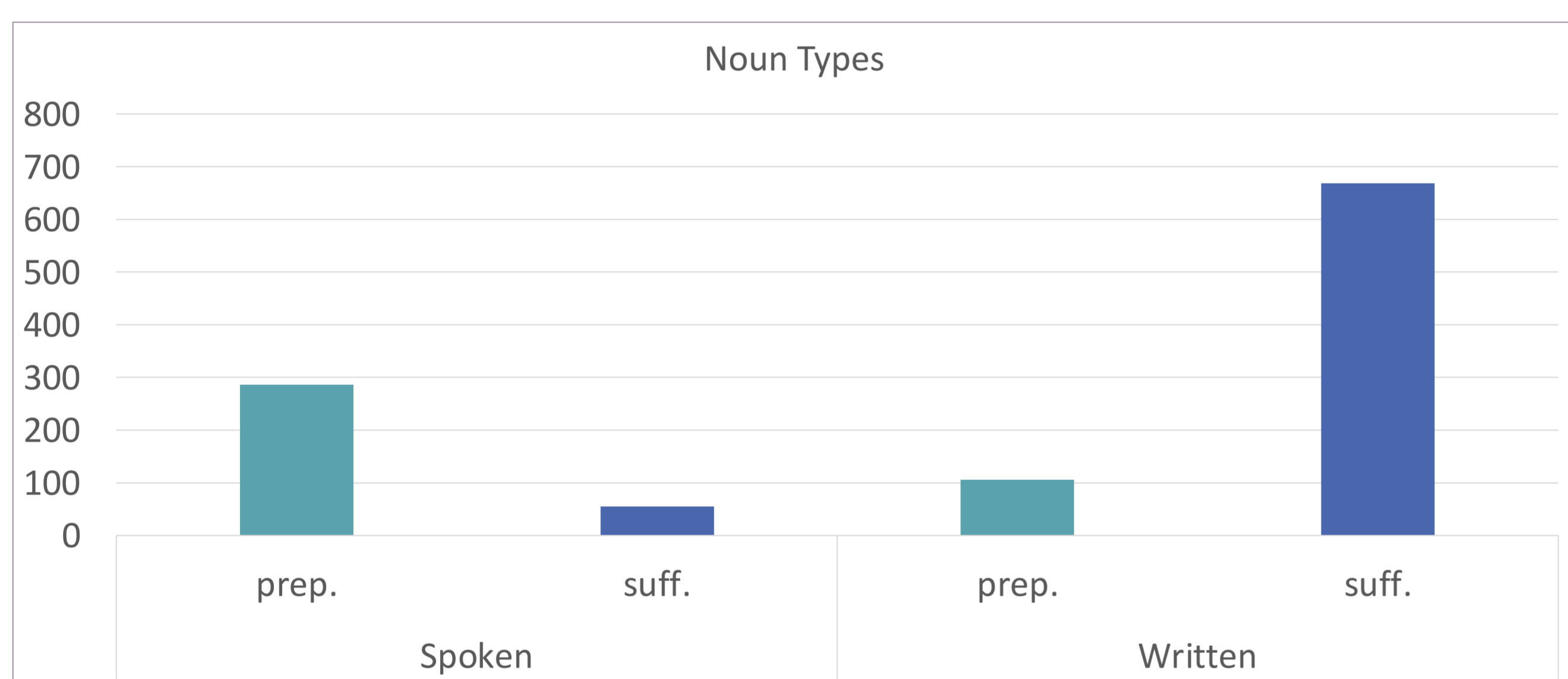
- | | | |
|--------------------------------------------------------------------|-------------------------------------------------------------------------------------|---------------------------------------|
| <p>(1) ha-jem jel-i
the-name of-POSS.1s
'my name'</p> |  | Prepositional
construction |
| <p>(2) jm-i
name-POSS.1s
'my name'</p> |  | Suffixed
construction |

Two genres:

- ⚙️ **Spoken:** The Corpus of Spoken Israeli Hebrew (Izre'el et al., 2002) & The Haifa Corpus of Spoken Hebrew (Maschler et al., 2021)
- ⚙️ **Written:** IAHLTwiki, a UD-treebank of Wikipedia entries (Zeldes et al., 2022)

Diversity

Variety: The number of types into which items can be classified

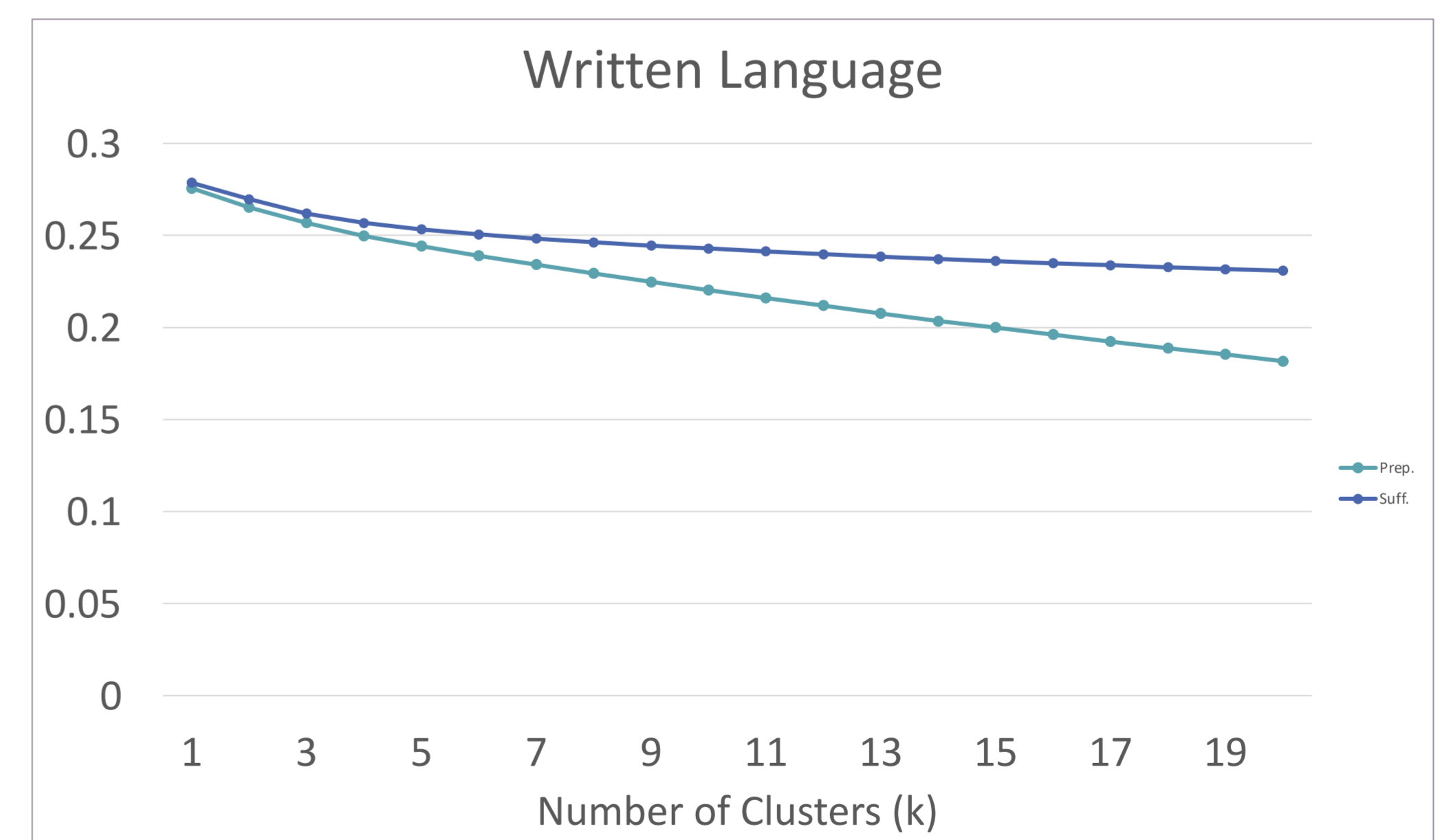
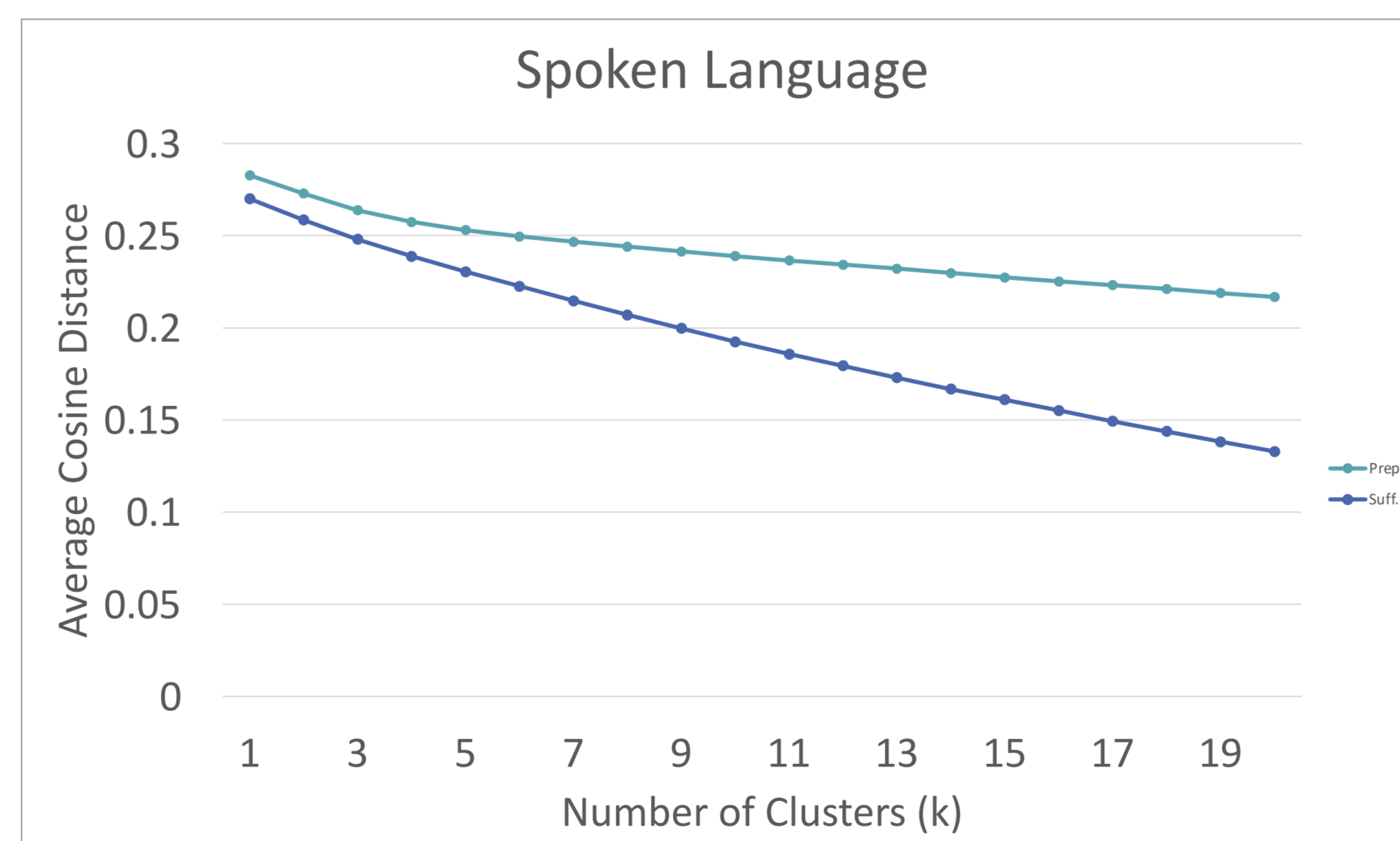


Balance: The uniformity of the type – item distribution.

	Spoken		Written	
	prep.	suff.	prep.	suff.
Tokens	653	172	118	2,096
Types	286	55	105	668
Type-token ratio	0.438	0.320	0.890	0.319
#hapax legomena	222	34	92	351
Potential productivity	0.340	0.198	0.780	0.170

Disparity: The degree to which types differ from each other within a category.

We used AlephBERT to represent the semantics of the types and calculated the disparity of each category by running a *k*-means algorithm 100 times for each $k = [1..20]$ and calculating the **average cosine distance** between each type and the prototype (centroid) of its cluster.



Next steps:

- ⚙️ We will design & run experiments testing the relative effect of a construction's diversity dimensions on its extensibility.
- ⚙️ Our hypothesis is that the demand for semantic similarity of a coinage to the attested uses will vary across the constructions as a function of their diversity.

Selected references

- Erb, I. 2022. *From synchrony and diachrony and back: the case of Hebrew pronominal possessives*. MA thesis, Tel-Aviv University.
- Izre'el, S., B. Hary & G. Rahav. 2002. Designing CoSIH: The corpus of Spoken Israeli Hebrew. *International Journal of Corpus Linguistics* 6(2). 171–197.
- Maschler, Y., H. Polak-Yitzhaki, S. Fishman, C. Miller Shapiro, N. Goretsky, G. Aghion & O. Fofliger. 2021. *The Haifa Corpus of Spoken Hebrew*. <https://sites.google.com/humanities.haifa.ac.il/corpus>
- Seker, A., E. Bandel, D. Bareket, I. Brusilovsky, R. G., and R. Tsarfaty. 2022. AlephBERT: Language Model Pre-training and Evaluation from Sub-Word to Sentence Level. In *Proceedings of ACL.2022*
- Zeldes, A., N. Howell, N. Ordan & Y. Ben Moshe. 2022. A secondwave of UD Hebrew treebanking and cross-domain parsing. In *Proceedings of EMNLP 2022*.