The semantics of -*ee* and -*ation*: a distributional semantic approach

Viktoria Schneider

Ingo Plag's Birthday Celebrations

01/09/2022

Introduction: eventuality-related nominalizations

- What are eventuality-related nominalizations?

  (1) *employee*, *trainee*
  → participant reading

  (2) *Markham sets down the rules about park befoulment.* (Plag et al. 2018: 474)
  → whole eventuality reading

- Semantic representation provides eventualities and participants for word formation process (e.g., Plag et al. 2018, Kawaletz 2021)

- Research tends to focus on deverbal nominalizations (e.g., Barker 1998; Alexiadou 2010; Kawaletz & Plag 2015; Plag et al. 2018; Kawaletz 2021)

- Many nominalizing suffixes also attach to non-verbal bases (e.g., Plag 1999, 2004; Bauer et al. 2013)

Introduction: Distributional Semantics

- Distributional Semantics useful for analysis of nominalizations

  (e.g., Lapesa et al. 2018; Wauquier et al. 2018; Huyghe & Wauquier 2020)

- Difference in meaning = difference in distribution

- Word vector: computed by list of words in context of target word

- Distance between vectors = semantic similarity
  - High distance → unsimilar
  - Low distance → similar

- Measured in cosine similarity (other measures available)
  - Higher cosine similarity = higher similarity of semantics of words

(see, e.g., Lapesa et al. 2018)

Research questions

- How similar are the meanings of a derivative and its base word?
  - How similar are the meanings of **denominal** derivatives and their base words?
  - How similar are the meanings of **deverbal** derivatives and their base words?

- Which factors influence the cosine similarity between base and derivative?

- Do we find differences regarding different suffixes?

- Focus on -*ee* and -*ation*

Hypotheses

- Base and derivative similar
  - Eventive elements for word formation process already in base (e.g., Plag et al. 2018, Kawaletz 2021)

- Verbal bases more similar to their derivatives than nominal bases to their derivatives
  - Verbs clearly eventive (e.g., Van Valin & LaPolla 1997; Haspelmath 2001; Szabó 2015)
  - Word formation process more straightforward
  - Eventive elements easier identifiable for word formation process

Method

- FastText (Mikolov et al. 2018) *Common Crawl* subword model
  - 2 million pre-trained word vectors
  - Contains subword information to create new vectors based on *n*-grams

- Compare cosine similarity of denominal/deverbal derivatives and their nominal/verbal bases

Method

- Beta regression models to determine which factors influence the cosine similarity
  - Dependent variable: cosine similarity between base and derivative, range of (0,1)

Variables of interest	Expectation
Relative frequency of base/derivative	Higher relative frequency leads to higher segmentability (e.g., Hay & Baayen 2003) → higher cosine similarity
Word class of base	Verbal bases more similar to derivatives due to clearer eventuality
Polysemy of base	Higher polysemy of base leads to decrease of cosine similarity

biographee

debtee

The suffix -*ee*

46 denominal 312 deverbal

mentee

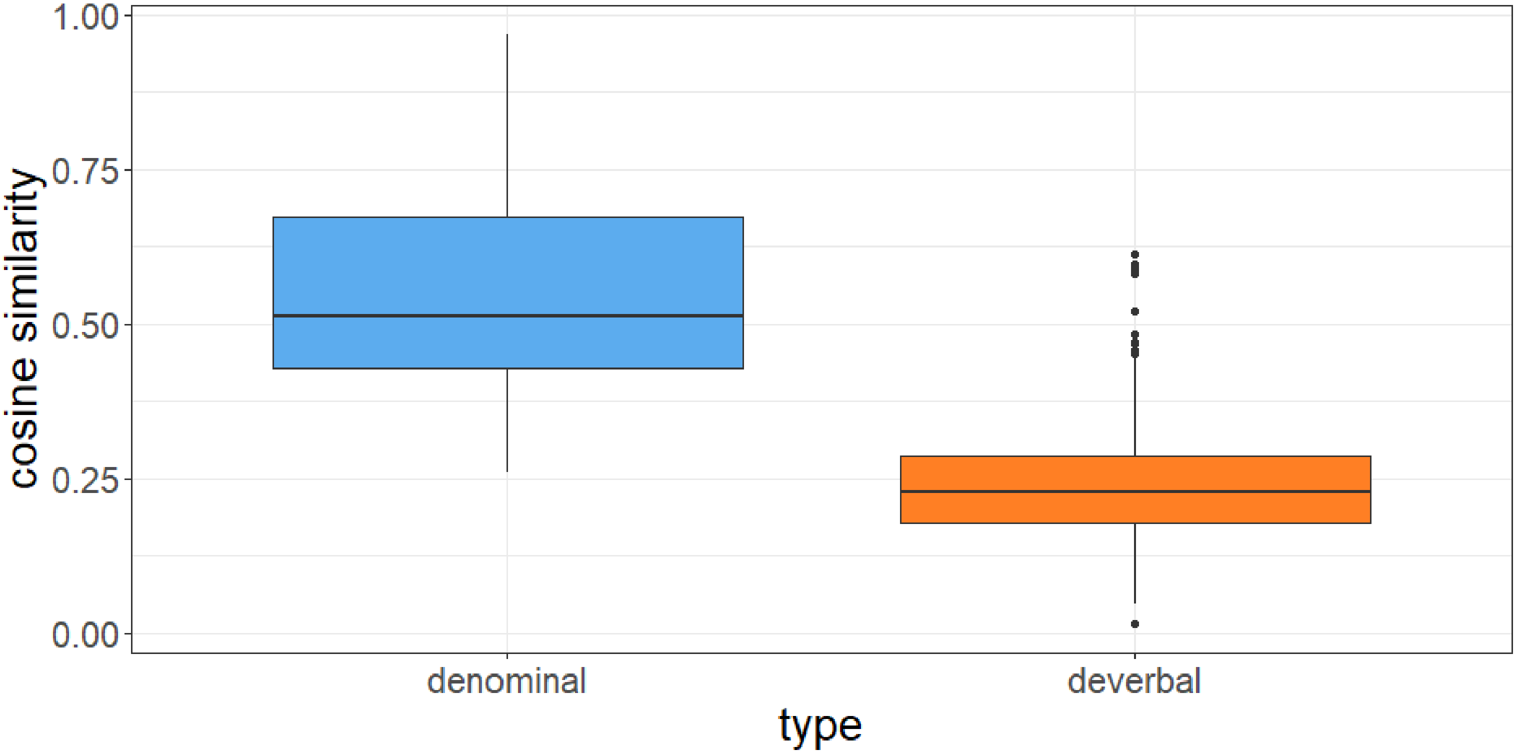
covenantee

tutee

Data from COCA (Davies 2008) and BNC (Davies 2004)

Similarity nominal and verbal bases and derivatives for -*ee*

- Cosine similarity of denominal derivatives and nominal bases higher than that of deverbal derivatives and verbal bases
- Contra expectation that deverbal derivatives more similar to verbal bases



Beta regression model -*ee*

- Polysemy of base
  - Not significant

- Relative frequency
  - Significant
  - Higher relative frequency decreases cosine similarity
  - Not expected

- Word class base
  - Significant
  - Cosine similarity decreases if base is a verb
  - Not expected

concertation

pixelation

The suffix -*ation*

67 denominal 72 deverbal

instrumentation

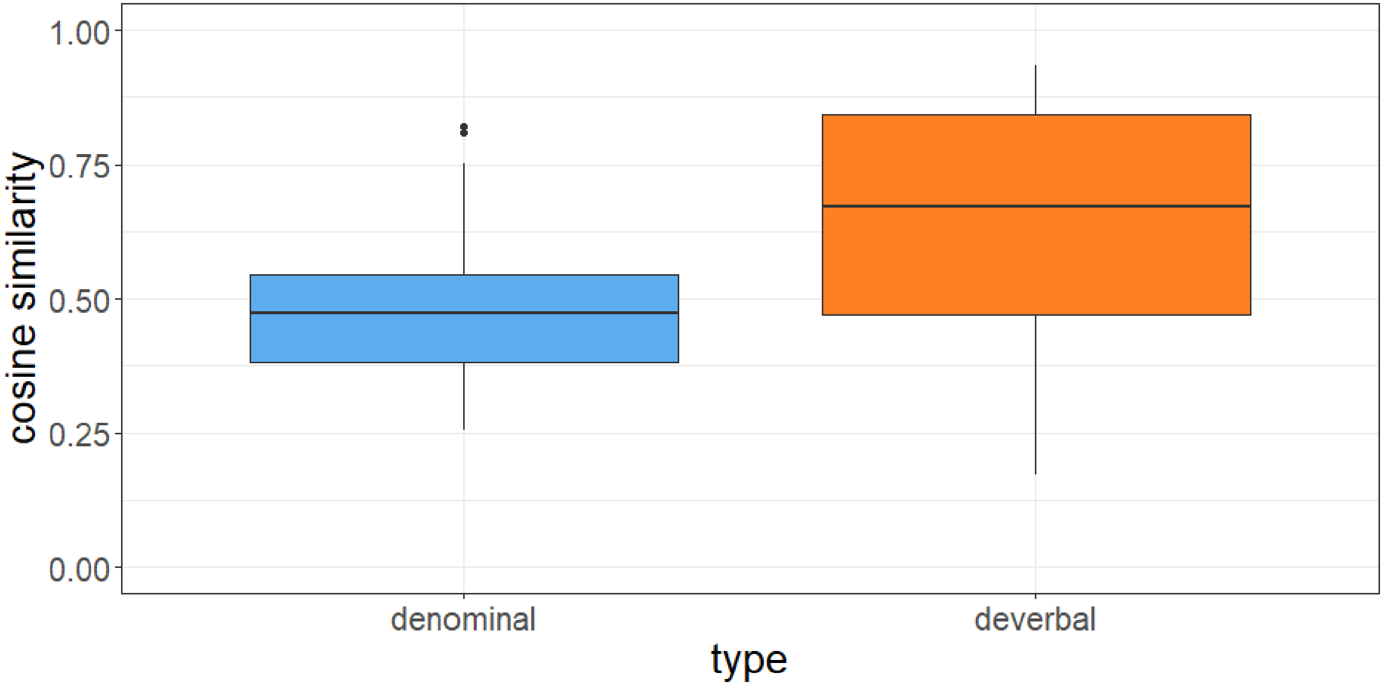
ozonation

impactation

Data from COCA (Davies 2008) and BNC (Davies 2004)

Similarity nominal and verbal bases and derivatives for -*ation*

- Denominal derivatives and nominal bases show lower cosine similarity than deverbal pairs → opposite picture than for -*ee*



Beta regression with principal component for -*ation*

- Correlations of relative frequency, base polysemy, word class

→ Principal Component (PC) Analysis to get rid of possible collinearity

- First principal component is retained for analysis as fulfills common criteria (e.g., O'Rourke et al. 2005; Baayen 2008; Schmitz et al. 2021)
  - Higher polysemy of base word decreases cosine similarity (expected)
  - Higher relative frequency decreases cosine similarity (unexpected)
  - Word class of base influences cosine similarity (verbal base higher cosine similarity, expected)

Differences -*ee* and -*ation*

- Cosine similarity
  - **Denominal** -*ee* derivatives more similar to nominal bases than **deverbal** derivatives to verbal bases
  - **Deverbal** -*ation* derivatives more similar to verbal bases than **denominal** derivatives to nominal bases

- Cosine similarity significantly influenced by
  - Relative frequency for both data sets (contra expectation)
  - Word class for both data sets (contra expectation for -*ee*, in line with expectation for -*ation*)
  - Polysemy of base for -*ation* data (in line with expectation)

References

- Alexiadou, Artemis. 2010. Nominalizations: A probe into the architecture of grammar part i: The nominalization puzzle. *Language and Linguistics Compass* 4(7). 496-511.
- Baayen, R. Harald. (2008). *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge University Press. doi:10.1017/CBO9780511801686.
- Barker, Chris. 1998. Episodic -ee in English: A thematic role constraint on new word formation. *Language* 74 (4), pp. 695-727.
- Bauer, Laurie, Rochelle Lieber & Ingo Plag. 2013. *The Oxford reference guide to English morphology*. Oxford: Oxford University Press.
- Davies, Mark. 2004. *British National Corpus* (from Oxford University Press). Available online at https://www.english-corpora.org/bnc/.
- Davies, Mark. 2008. *The Corpus of Contemporary American English: 400+ million words, 1990-present*. Available online at https://www.english-corpora.org/coca/.
- Haspelmath, Martin. 2001. Word classes and parts of speech. In Neil J. Smelser & Paul B. Baltes (eds.), *International encyclopedia of the social & behavioral sciences*, 16538-16545. Amsterdam: Elsevier.
- Hay, Jennifer & Harald Baayen. 2003. Phonotactics, parsing and productivity. *Italian Journal of Linguistics* 1. 99–130. doi:10.1.1.171.705.
- Kawaletz, Lea. 2021. *The semantics of english -ment nominalizations*. PhD Dissertation, Heinrich Heine-Universität Düsseldorf.
- Kawaletz, Lea & Ingo Plag. 2015. Predicting the semantics of English nominalizations: A frame-based analysis of –ment Suffixation. In: *Semantics of complex words*. Bauer, Laurie, Lívia Körtvélyessy, Pavol Štekauer (Eds.), pp. 289-319.
- Lapesa, Gabriella, Lea Kawaletz, Ingo Plag, Marios Andreou, Max Kisselew & Sebastian Padó. 2018. Disambiguation of newly derived nominalizations in context: A distributional semantics approach. *Word Structure* 11(3). 277–312. doi: 10.3366/word.2018.0131.
- Mikolov, Tomas, Edouard Grave, Piotr Bojanowski, Christian Puhrsch & Armand Joulin. 2018. *Advances in Pre-Training Distributed Word Representations*.
- O'Rourke, Norm; Hatcher, Larry; Stepanski, E. J. (2005). Using SAS for Univariete & Multivariate Statistics. SAS Institute Inc.
- Plag, Ingo. 1999. *Morphological productivity: Structural constraints in English derivation*. Berlin: Mouton de Gruyter.
- Plag, Ingo. 2004. Syntactic category information and the semantics of derivational morphological rules. *Folia Linguistica* 38(3-4). 193–225.
- Plag, Ingo, Marios Andreou & Lea Kawaletz. 2018. A frame-semantic approach to polysemy in affixation. In Olivier Bonami, Gilles Boyé, Georgette Dal, Hélène Giraudo & Fiammetta Namer (eds.), *The lexeme in descriptive and theoretical morphology*, 467–486. Berlin: Language Science Press.
- Schmitz, D., Plag, I., Baer-Henney, D., & Stein, S. D. (2021). Durational Differences of Word-Final /s/ Emerge From the Lexicon: Modelling Morpho-Phonetic Effects in Pseudowords With Linear Discriminative Learning. Frontiers in Psychology, 12, 2983.
- Szabó, Zoltán Gendler. 2015. Major parts of speech. *Erkenntnis* 80(S1). 3-29.
- Van Valin, Robert D. & Randy J. LaPolla. 1997. *Syntax: Structure, meaning and function*. Cambridge textbooks in linguistics. Cambridge: Cambridge Univ. Press.

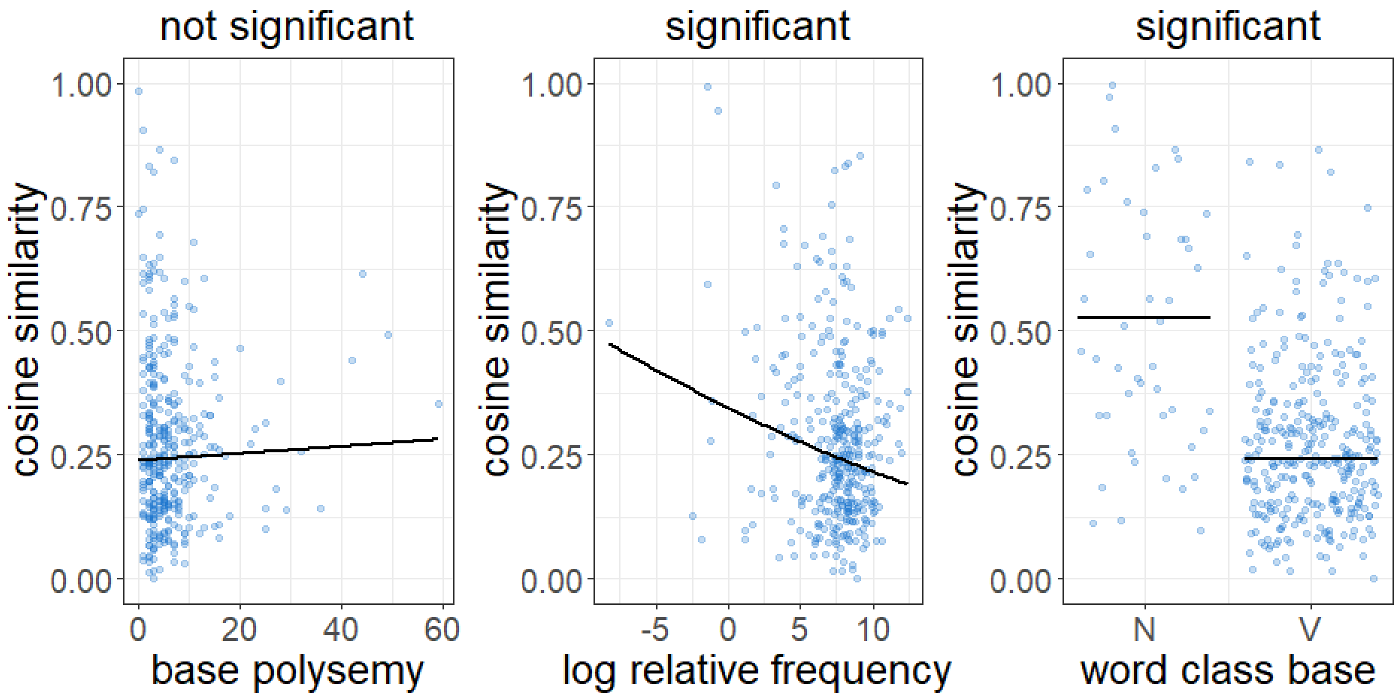
Thank you!

Beta regression model -*ee*

- Dependent variable cosine similarity range of (0,1)

Not expected
Higher relative frequency
→ lower cosine similarity

Not expected
Deverbal pairs expected
to be more similar

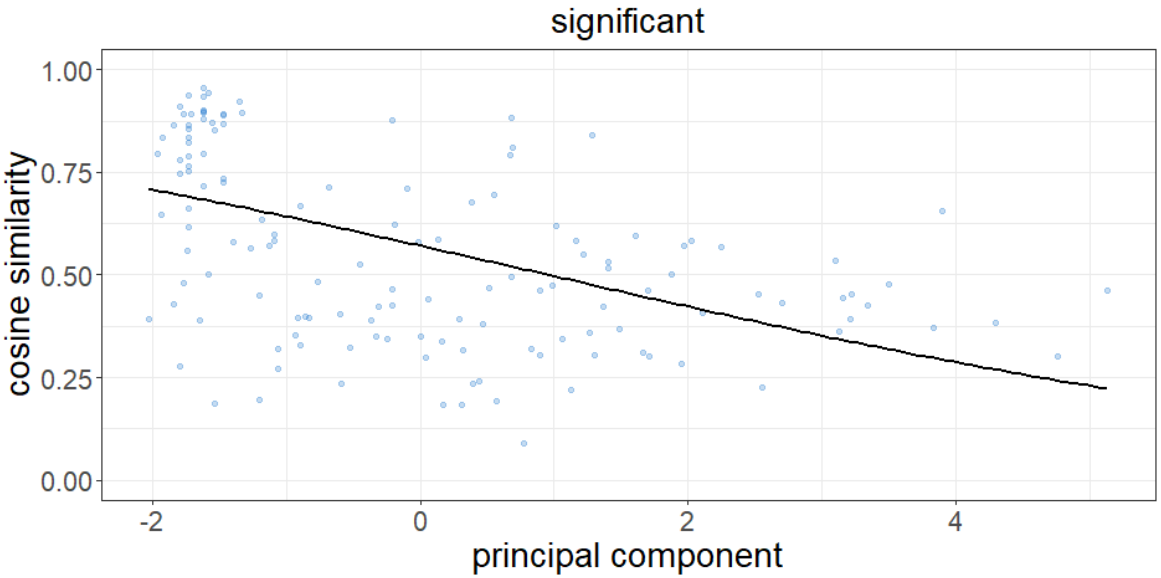


Beta regression with principal component for -*ation*

- Correlations of relative frequency, base polysemy, word class

→ Principal Component (PC) Analysis to get rid of possible collinearity

- First principal component is retained for analysis as fulfills common criteria (e.g., O'Rourke et al. 2005; Baayen 2008; Schmitz et al. 2021)
  - Higher polysemy of base word decreases cosine similarity (expected)
  - Higher relative frequency decreases cosine similarity (unexpected)
  - Word class of base influences cosine similarity (verbs higher cosine similarity, expected)



Beta regression with principal component for -*ation*

- Common criteria PCA
  - Eigenvalue higher than 1
  - Cumulative percentage explained higher than 80%
  - PC has to make sense in their loadings
    - Here it decreases cosine similarity

Vector space: Fasttext (Mikolov et al. 2018)

- *Common Crawl* subword model
  - 2 million pre-trained word vectors
  - Contains subword information to create new vectors based on *n*-grams

	#de	**deb**	**ebt**	**bte**	**tee**	**ee#**	**#fi**	**fis**	**ish**	**sh#**
debtee	1	1	1	1	1	1	0	0	0	0
fish	0	0	0	0	0	0	1	1	1	1

- Vectors for every word in the data set

Data set -*ee*

- Data from BYU Corpora (Davies 2004, 2008)

- Denominal: 46 (*debtee, biographee, covenantee*)
- Deverbal: 312 (*employee*, *devotee*, *appointee*)

Interim summary -*ee* data

- Denominal derivatives more similar to nominal bases than deverbal derivatives to verbal bases
- Word class of base significant: verbs lower cosine similarity compared to nouns (unexpected)
- Relative frequency decreases cosine similarity (unexpected)
- Polysemy of base not significant

Data set -*ation*

- Data from BYU Corpora (Davies 2004, 2008)

- Denominal: 67 (*concertation*, *instrumentation, ozonation*)
- Deverbal: 72 (*avocation*, *beneficiation, idolization*)

Open questions

- Why are the results for the similarity of denominal and deverbal derivatives and their bases different for the suffixes?
  - Difference due to ontology? (e.g. Van Valin & LaPolla 1997; Haspelmath 2001; Szabó 2015)
    - -*ee* creates participant readings → participants usually denoted by nouns
    - -*ation* refers to eventualities → eventualities usually denoted by verbs

Beta regression model – Principal component analysis (PCA)

- Problem
  - Correlations of relative frequency, base polysemy, word class
  - Collinearity in model

- Solution applied: PCA
  - Dimensionality of data reduced by transformation of problematic variables into principal components
  - Transformations lead to linear combinations of predictors
  - Resulting principal components are not correlated

- First principal component is retained for analysis as fulfills common criteria (e.g. O'Rourke et al. 2005; Baayen 2008; Schmitz et al. 2021)

Distributional Semantics

- Example suit

Suit.1: *The suit was in the closet, with the tie and the t-shirt*

Suit.2: *The lawyer filed a suit to the judge*

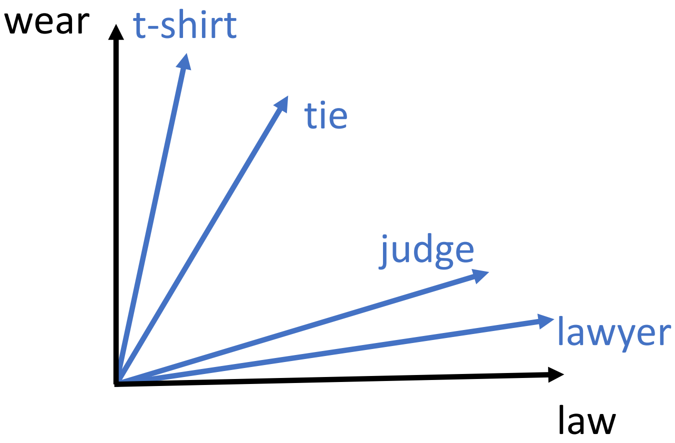
dimensions →

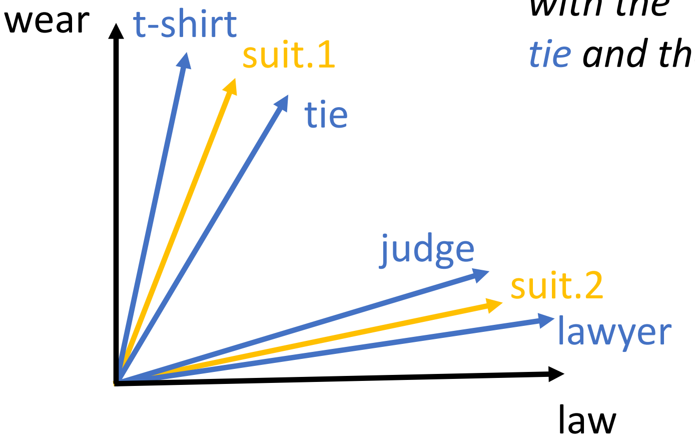
	tie	t-shirt	lawyer	judge
Suit.1	30	15	8	0
Suit.2	0	0	26	18

- Usually more dimensions: for example, 300 dimensions

Andreou et al. 2016

Distributional Semantics





Suit.1: *The suit was in the close with the*
tie and the t-shirt

Suit.2: *The lawyer filed a suit to the judge*

Andreou et al. 2016

Vector space – Fasttext (Mikolov et al. 2018)

- Database: pre-trained word vectors based on *Common Crawl* and *Wikipedia*

- Word vectors in 300 dimensions

- Problem: many denominal derivatives low frequency → not in pre-trained sets

- *Common Crawl* subword corpus
  - 2 million pre-trained word vectors
  - Contains subword information to create new vectors based on *n*-grams

- Vectors for every word in the data set

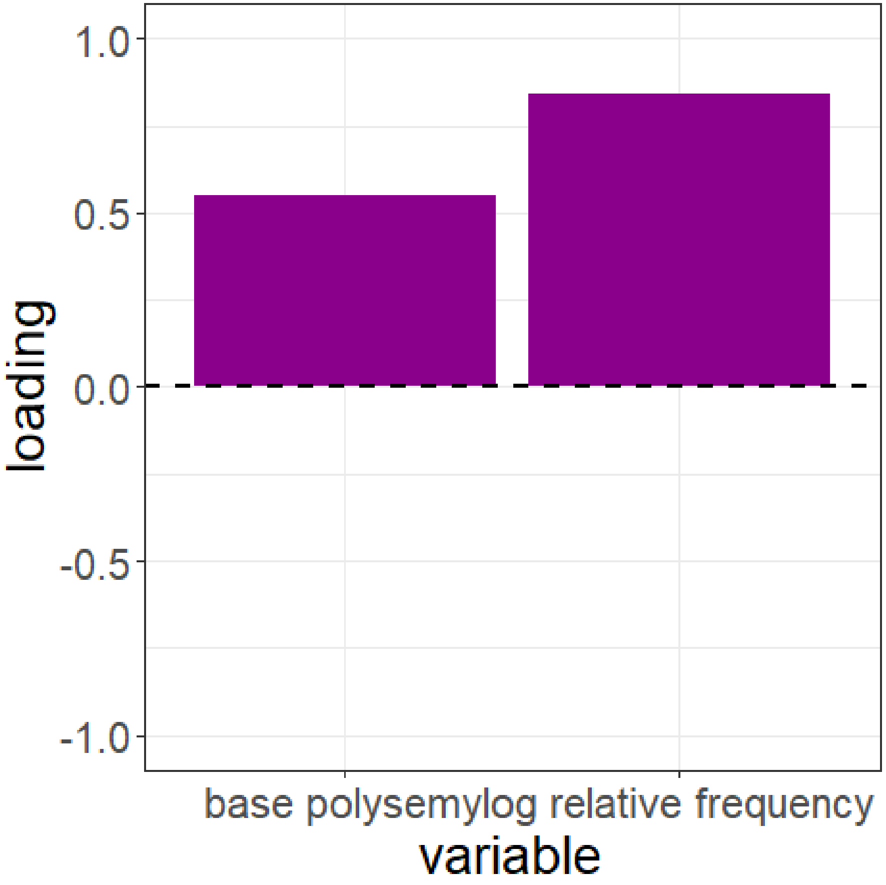
Method: Analysis

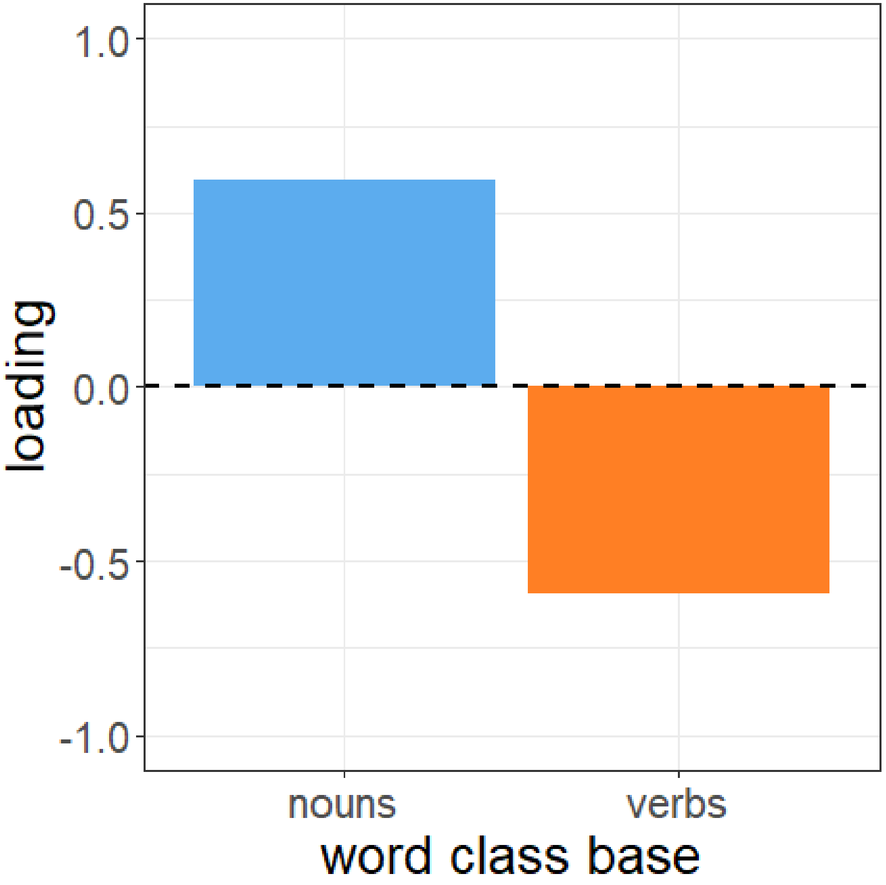
- Compare cosine similarity of denominal/deverbal derivatives and their nominal/verbal bases
- beta regression models to determine which factors influence the cosine similarity
  - Dependent variable: cosine similarity between base and derivative

Variables of interest	Expectation
Relative frequency of base/derivative	Higher relative frequency leads to higher segmentability → higher cosine similarity
Word class of base	Verbal bases more similar to derivatives due to clearer eventuality
Polysemy of base	Higher polysemy of base leads to decrease of cosine similarity

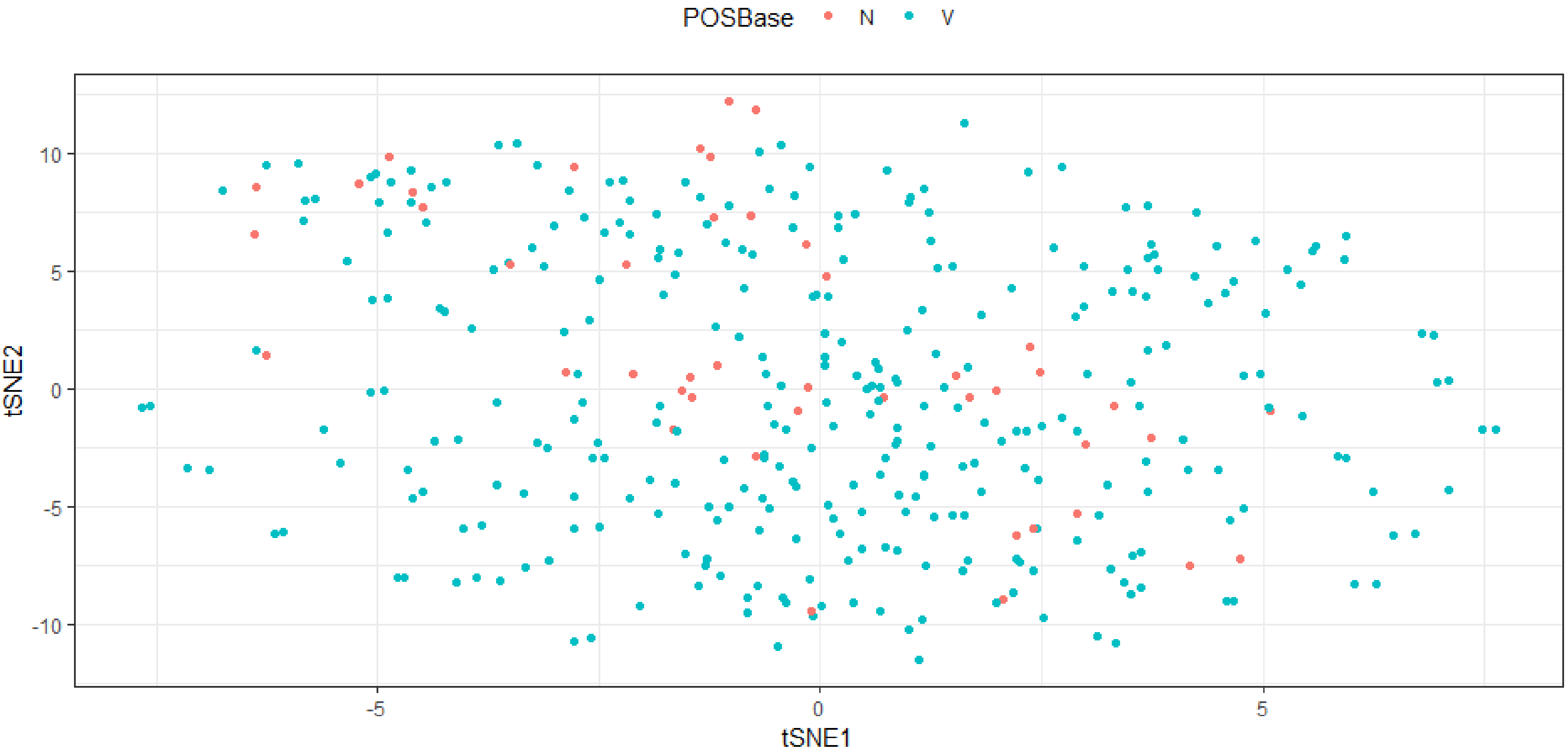
PCA

- Loading says how strong effect of variable in PC

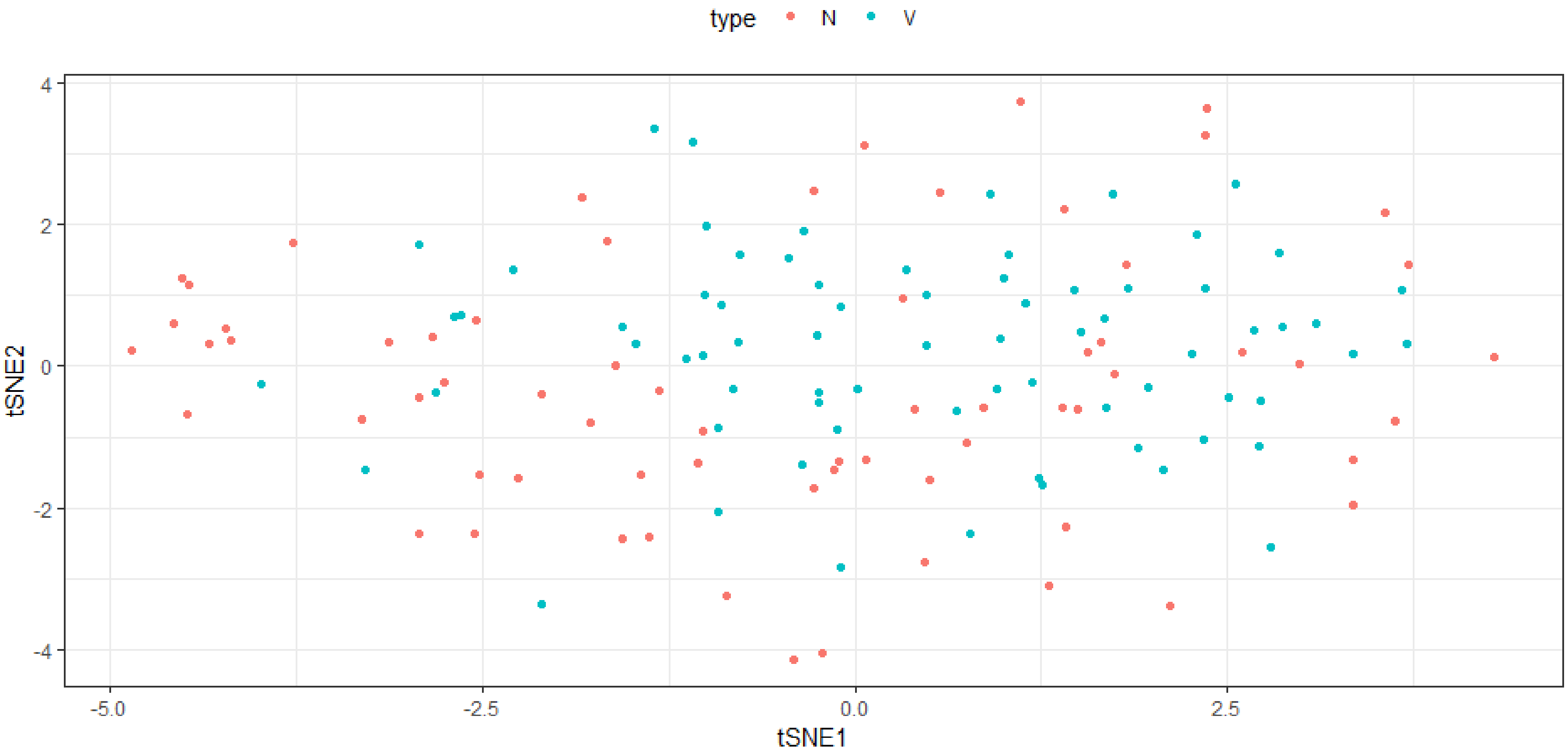




Appendix – t-SNE t-distributed stochastic neighbor embedding -*ee*



Appendix – t-SNE t-distributed stochastic neighbor embedding -*ation*


type
● N
● V
tSNE2
tSNE1