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

Viktoria Schneider

Abstract

This paper analyzes nominalizations with the suffixes *-ee* and *-ation* using a distributional semantic approach. Comparing denominal and deverbal nominalizations with the same suffix, a novel perspective on the eventuality of denominal nominalizations is given. The present study makes use of cosine similarities of word vectors of bases and their pertinent derivatives to gain insight into their underlying semantic similarity. The results of the present study show that nominalizations and their bases are overall similar to each other. This similarity is in some cases affected by relative frequency, base polysemy, and word class of the base.

1. Introduction

Nominalizations in English can have several different word classes as bases (see, e.g., Plag 2004; Bauer et al. 2013). Many nominalizations are eventuality-related as the examples in (1):

(1) a. *employee, trainee, tutee*
b. *Markham sets down the rules about park befoulment.* (from Plag et al. (2018))

The derivatives with the suffix *-ee* in (1) a. are participant readings (see, e.g., Barker 1998; Plag 2004; Schneider 2022). The nominalization with the suffix *-ment* in (1) b. refers to the whole eventuality denoted by the base verb. Due to this relation between eventualities and derivatives, research on such nominalizations mainly focuses on verbs (e.g., Barker 1998; Alexiadou 2010; Kawaletz & Plag 2015; Plag et al. 2018; Kawaletz 2021), because verbs are ontologically said to denote eventualities, whereas nouns are claimed to mostly do not (e.g., Van Valin & LaPolla 1997; Haspelmath 2001; Szabó 2015).

Recent studies on denominal derivatives showed that nouns as bases do also inherit the eventive material necessary for the derivational process rising an eventuality-related interpretation (Schneider 2022; Kotowski et al. 2022). This paper will perform a distributional semantic analysis to investigate whether the relatedness of derivative and base is also represented in their semantic similarities.

Distributional semantics has shown to be a useful approach for the analysis of derivatives (see, e.g., Lapesa et al. 2018; Wauquier et al. 2018; Huyghe & Wauquier 2020). Such a distributional approach looks at the distributions of words in contexts. If two words occur together in the same context often, they are assumed to be semantically similar. This similarity can be represented by so-called semantic word vectors. The following research questions are investigated in the present study:

> RQ1 As derivatives and their bases share semantic content, are they overall semantically similar?
> RQ2 Are the semantics of a deverbal derivative and its base more similar than the semantics of a denominal derivative and its base?
> RQ3 Do relative frequency, word class of the base, and the degree of polysemy of the base affect the similarity of base and derivative?

This study will investigate the cosine similarities between derivatives and bases to see whether the semantic connection of both is as close as assumed based on derivatives with the suffixes *-ee* and *-ation*. The remainder of this paper is structured as follows. First, the theoretical background on distributional semantics will be described in more detail. Section 3 describes the methodology and Section 4 presents the result of statistical analyses. Section 5 will discuss the findings and concludes this paper.

2. Background

The underlying hypothesis of distributional semantics is the so-called distributional hypothesis (see, e.g., Harris 1954). This hypothesis states that a difference in meaning is represented in a difference in distribution. Hence, if words occur together in different contexts the semantics of these words is not connected. On the other hand, if two words are often used together, their semantics is similar. This (dis)similarity can be captured in so-called word vectors. Each word is then represented by a string of numbers, a vector, which can then be compared to the vectors of other words. Usually, the first measure for such a comparison is their cosine similarity. A higher cosine similarity of word vectors expresses a higher semantic similarity of words' semantic, whereas a lower cosine similarity expresses a lower similarity (cf. Sitikhu et al. 2019; Huyghe & Wauquier 2020).

For example, a word like *suit* can have two different meanings. Both meanings can be distinguished by having a look at their distribution in linguistic examples. One meaning of *suit*.1 refers to a piece of clothing and another meaning of *suit*.2 refers to a legal document/concept. Example (2) shows two example sentences for the two readings:

(2) a. *Suit.1: The suit was in the closet, with the tie and the t-shirt.*
b. *Suit.2: The lawyer filed a suit to the judge.*
(from Lapesa et al. (2018))

Table 1 shows a toy example for the computation of word vectors for the two meanings of *suit*. *Suit*.1 occurs with the words *tie* and *t-shirt* clearly more often than with the words *lawyer* and *judge*. *Suit*.2 shows the opposite trend; it co-occurs more often with *lawyer* and *judge* than with *tie t-shirt*. This represents the different meanings of *suit*.

Table 1: General idea of word distributions for vector calculations.

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

In the toy example we have 4 words *tie*, *t-shirt*, *lawyer* and *judge* that we check for co-occurrence with the two meanings of the target word *suit*. This results in a word vector in four dimensions. For a better representation for the semantics of a word, a word vector with more dimensions is needed. Thus, a word vector is computed by the co-occurrence of the target word and all other words in the corpus and then reduced to 100 or more dimensions depending on the research questions, computational load, and method. 300 dimensions were used in this study as the vector spaced used is of 300 dimensions. The vector space and the methodology are described in section 3.2.

3. Methodology

3.1. Data

All derivatives used in this study were taken from the Corpus of Contemporary American English (COCA, Davies 2008) and the British National Corpus (BNC, Davies 2004). The data used in this study were coded according to the word class of the base. Due to the productive process of conversion in English, the decision of the word class of the base is not a trivial task. It was chosen to use a frequency criterion to determine the word class of the base. If a base occurs distinctively more frequent as a noun than a verb in COCA, the base was coded as a noun, if it was the other way around, the base was coded as a verb.

The data set for nominalizations with the suffix *-ee* contains 46 denominal and 312 deverbal derivatives. The data set for nominalizations with the suffix *-ation* contains 67 denominal and 72 deverbal derivatives. Examples for denominal and deverbal derivatives with both suffixes are shown in Example (3):

(3) denominal
a. *-ee: biographee, covenantee, debtee*
b. *-ation: concertation, instrumentation, ozonation*
deverbal
a. *-ee: appointee, devotee, employee*
b. *-ation: avocation, beneficiation, idolization*

3.2. Vectors

The vectors were created using fastText (Bojanowski et al. 2016), a python package which includes, but is not limited to, pre-trained vector spaces freely available (Mikolov et al. 2018). The word vectors are of 300 dimensions. For the computation of the word vectors for the derivatives and bases under investigation in this paper, the common crawl subword model was used. This model contains 2 million pre-trained word vectors and subword information based on 3-6 grams (Mikolov et al. 2018). Due to the inclusion of the subword information in the model, it is possible to compute new word vectors for words that are not in the pretrained set. This is extremely helpful for the recent study as many derivatives, especially the denominals, are often relatively low in frequency and therefore not found in pre-trained vector spaces. The word vectors for the derivatives and their bases in *-ee* and *-ation* were computed with a context window of [+/-5] and reduced to 300 dimensions. These vectors were then used for the analysis.

3.3. Analysis

In order to compare whether base and derivative are similar in meaning, the cosine similarities of their word vectors were computed in Python (Van Rossum & Drake 2009). The cosine similarities compare base and derivative with each other in the different types, denominal or deverbal, depending on the word class of the base. Beta regression models in R (R Core Team 2020) were used to find out which factors influence the similarity between base and derivative. Beta regression was chosen as the statistical tool of choice as the cosine similarities in this study were in the interval of [0,1]. There was no output with negative cosine similarity values. Thus, the dependent variable for the models is the cosine similarity of derivative and base.

The variables of interest, which were used as predictors in the model, are listed in Table 2[1]. These variables might have an influence on the cosine similarity between base and corresponding derivative and were thus included.

Table 2: Variables of interest and their expected effects

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

[1] For more information on segmentability, see, e.g., Hay & Baayen (2003). For more infor- mation on the ontology and eventuality of word classes, see, e.g., Van Valin & LaPolla (1997); Haspelmath (2001); Szabó (2015).

	derivatives due to clearer eventuality
BASE POLYSEMY	Higher polysemy of base leads to decrease of cosinesimilarity

Before the model was fitted, relative frequency was log-transformed following standard procedures to avoid issues of unreliable model estimates (Baayen 2008). The beta regression models were fitted using the betareg package (Cribari-Neto & Zeileis 2010). One model for the data with the suffix *-ee* and one model for the data with the suffix *-ation* were fitted, both with the following structure:

$$cosine\ similarity \sim base\ polysem + \log relative\ frequency + word\ class\ base \tag*{1}$$

4. Results

4.1. The suffix *-ee*

First, the results for the comparison of the denominal and deverbal derivatives with the suffix *-ee* will be described. Figure 1 shows the cosine similarities of bases and derivatives. The blue box on the left depicts the cosine similarities between denominal derivatives and bases. The orange box on the right describes the cosine similarities between the deverbal derivatives and bases. The denominal pairs have a median cosine similarity of about 0.5 and the deverbal pairs of 0.25. A higher cosine similarity corresponds to a higher similarity of compared vectors. Hence, the denominal derivatives and bases are clearly more similar to each other than the deverbal pairs. The difference between denominal and deverbal pairs is significant (Wilcoxon test, $p < 0.001$). This result is contra the expectation as the assumption is that verbs and their derivatives are more similar to each other as they operate on the same eventuality which is clearly denoted by the verb. Nouns, on the other hand, were expected to be less similar to their derivatives as they are not as straightforwardly eventive as verbs.

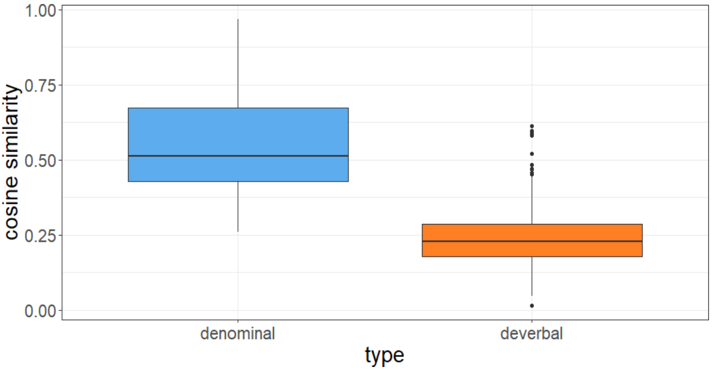

Figure 1: Cosine similarities for derivatives and bases. Blue are the cosine similarities of the denominal data, orange are the cosine similarities for the deverbal set.

Figure 2 shows the influence of the variables on all *-ee* derivatives in the beta regression model. No collinearity was observed. The polysemy of the base word does not show a significant effect. Relative frequency and the word class of the base, however, reach significance.

The higher the relative frequency is, the lower the cosine similarity becomes. This effect was not expected. A higher frequency of the base is expected to lead to a higher segmentability of the derivative and therefore for a higher similarity of base and derivative (cf. Hay & Baayen 2003).

The second predictor that reaches significance is word class of the base. If the base is a noun, the cosine similarity of base and derivative is higher than with a verbal base. This result is contra the expectation, as it was expected that deverbal derivatives are more closely connected to their bases as they operate on the same eventuality directly denoted by the base verb.

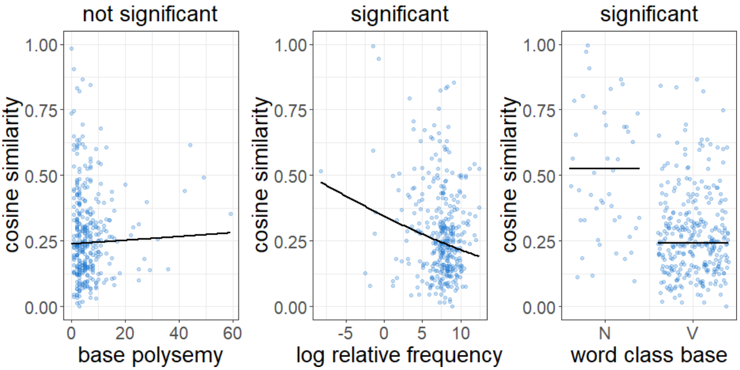

Figure 2: Beta regression model for all derivatives with the suffix *-ee*. The dependent variable cosine similarity has a value in the range of (0,1).

Summarizing, the results for the suffix *-ee* show that denominal derivatives are more similar in their meaning to their bases than deverbal derivatives and bases. This picture is contra the expectation as the assumption was that verbs are more closely connected to their derivatives as they directly denote eventualities. Influencing factors on the cosine similarity of base and derivative are the relative frequency of base and derivative and the word class of the base. The polysemy of the base does not show an influence although it was expected to do so.

4.2. The suffix *-ation*

The second suffix under investigation is *-ation*. A first look at the cosine similarities for denominal and deverbal base and derivative pairs in Figure 3 shows the opposite picture to *-ee*. The cosine similarity of the denominal pairs in blue are lower compared to the cosine similarities of deverbal derivatives and bases. The median of the denominal pairs is about 0.5 and for deverbal pairs around 0.7. The difference is significant (Wilcoxon test, p < 0.001).

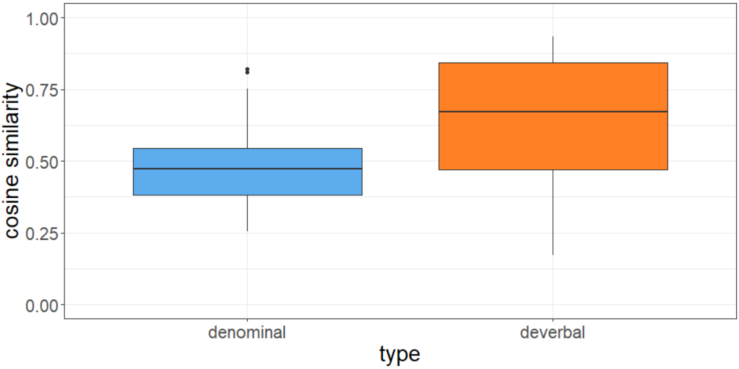

Figure 3: Cosine similarities for derivatives and bases. Blue are the cosine similarities of the denominal data, orange are the cosine similarities for the deverbal set.

Figure 4 summarizes the outcome of the beta regression model. All variables of interest influence the cosine similarity between bases and derivatives significantly. A higher polysemy of the base decreases the cosine similarity. This is expected as the derivative focuses on one reading of the base. If the base has more than one reading, the other readings are not similar to the semantics of the derivative and the similarity between base and derivative decreases. The higher relative frequency of the base also decreases the cosine similarity significantly. This is unexpected as a higher frequency of the base is said to lead to a higher segmentability of the derivative and should hence lead to a clearer connection between base and derivative (cf. Hay & Baayen 2003). The third variable influencing the cosine similarity of base and derivative in this model is the word class of the base. When the base is a noun the cosine similarity decreases and if the base is a verb the cosine similarity increases. This is expected as verbs and their derivatives were expected to have a closer relationship to each other as they operate on the same eventuality which is directly denoted by the verb.

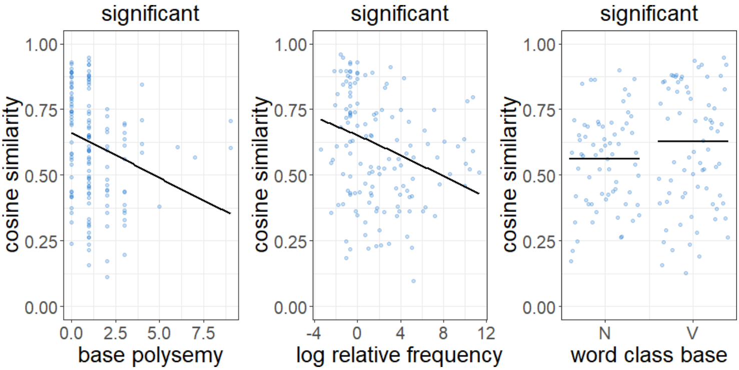

Figure 4: Beta regression model for all derivatives with the suffix *-ation*. The dependent variable cosine similarity has a value in the range of (0,1).

However, the analysis of the beta regression model poses a problem. The three variables of interest base polysemy, relative frequency, and word class of the base correlate with each other. This might bring collinearity into the model which might in turn lead to wrong results (Tomaschek et al. 2018).

The solution for the issue of collinearity in the model in this paper was to perform a principal component analysis (PCA). In a PCA, the dimensionality of all correlating variables is reduced by transforming the included variables into principal components (PC). The transformation leads to a linear combination of predictors and the resulting PCs are not correlated anymore. The first PC was retained for the analysis as it fulfills common criteria; Eigenvalue, cumulative percentage of variance explained and interpretability (for more information on PCA and criteria, see, e.g., O'Rourke et al. 2005; Baayen 2008; Schmitz et al. 2021). Figure 5 shows the loading of the PC which is an indicator of how strong the effect of which original variable in the PC is and which direction this effect points to. The loadings regarding base polysemy and relative frequency are both positive, while the loadings concerned with the word class of the base show a split: Loadings are positive for nouns, while they are negative for verbs. Thus, nouns go hand in hand with base polysemy and relative frequency, while higher values of relative frequency correspond to correspond to verbs as the word class of the base.

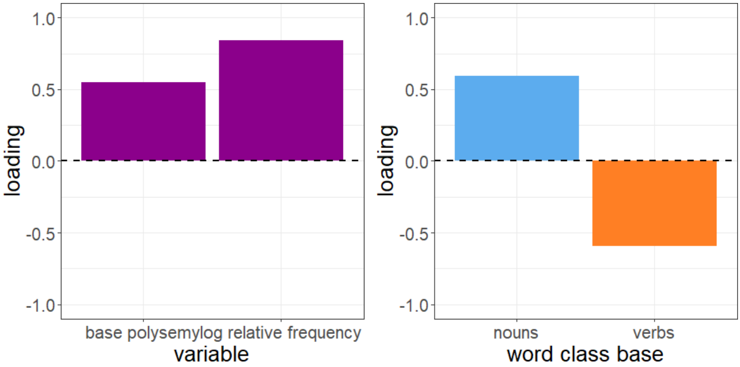

Figure 5: Loading of the retained principal component.

A second beta regression model was fitted with the PC as predictor:

$$cosine\ similarity \sim PC \qquad 2$$

Figure 6 shows that higher PC values come with significantly lower cosine similarities of bases and derivatives with *-ation*.

This is the same effect as was observed in the model with the correlations, and thus the potential issue of collinearity. Hence, a higher polysemy of the base and a higher relative frequency decrease the cosine similarity. Nouns as bases decrease the cosine similarity and verbs as bases increase the cosine similarity.

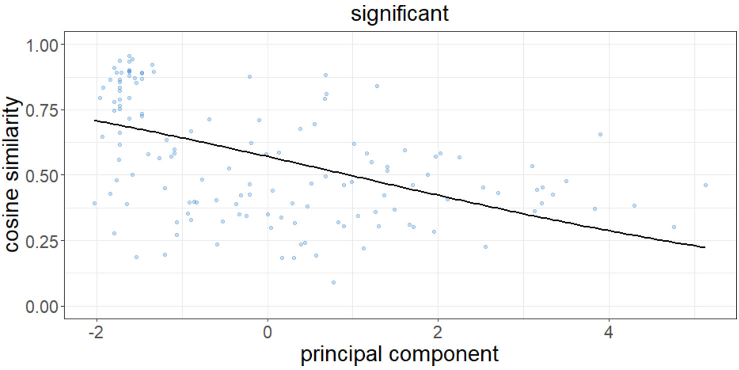

Figure 6: Effect of the principal component on the cosine similarity.

5. Discussion & Conclusion

The analysis showed that the variables relative frequency and word class of base have a significant influence on the cosine similarity of derivative and base in both data sets. Base polysemy reached significance only for the *-ation* data. The effect of the polysemy of the base goes in the expected direction as a base with more meanings is semantically more dissimilar to the derivative which picks out one specific reading of the base for its meaning. Relative frequency, on the other hand, decreases the cosine similarity in both data sets. This is an unexpected result as a higher frequency of the base should lead to a higher segmentability of the derivative which would strengthen the obvious connection between base and derivative. The effect of the word class of the base goes into different directions for both data sets.

The differences of the cosine similarities regarding the word class of the base in the two data sets might be explained by the suffix. Derivatives with the suffix *-ee* create a participant reading (see, e.g., Barker 1998; Plag 2004; Bauer et al. 2013; Plag 2018; Schneider 2022). Derivatives with *-ation*, on the other hand, describe processes (see, e.g., Bauer et al. 2013; Plag 2018). Assuming now that participants are usually represented as nouns and processes are usually denoted by verbs, the differences in similarities of the word classes of the base and their nominalizations are not unexpected. However, for both suffixes, verbs as bases are far more productive. This is caused by the ontology of verbs in general, as verbs usually denote eventualities with participants involved (for more on ontological categories, see, e.g., Van Valin & LaPolla 1997; Haspelmath 2001; Szabó 2015).

The fundamental assumption for this study is that bases and derivatives are similar in the first place. This is neither clearly confirmed nor rejected. The cosine similarity of base and derivative is influenced by several factors, namely the suffix, the word class of the base, the frequency of base and derivative, and partly by the polysemy of the base word. The assumption, based on ontological observations, that verbal bases are more similar to their derivatives as the word formation process is more straightforward due to the eventuality of verbs, is only true for derivatives with the suffix *-ation*; for *-ee*, the picture is reversed.

The present study raises further questions. First, does the distinction of the word class of bases play an important factor for the comprehension of derivatives? Second, how do cosine similarities of derivatives of further suffixes, for example, *-ment*, perform in comparison? Third, do semantic vectors computed by other approaches, for example, naive and linear discriminative learning, and their pertinent cosine similarities support the present results? These questions are subject to future research.

Acknowledgements

The author is grateful to Ingo Plag for the idea of implementing a distributional analysis. Many thanks to Dominic Schmitz for the help with computational issues in R and to the department of English Linguistics for helpful feedback.

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. doi:10.1111/j.1749-818X.2010.00209.x.

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). 695. doi:10.2307/417000.

Bauer, Laurie, Rochelle Lieber & Ingo Plag. 2013. *The Oxford reference guide to English morphology*. Oxford: Oxford Univ. Press 1st edn.

Bojanowski, Piotr, Edouard Grave, Armand Joulin & Tomas Mikolov. 2016. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.

Cribari-Neto, Francisco & Achim Zeileis. 2010. Beta regression in R. *Journal of Statistical Software* 34. 1–24. doi:10.18637/jss.v034.i02.

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 (COCA): One billion words, 1990-2019. https://www.English-corpora.org/coca/.

Harris, Zellig S. 1954. Distributional structure. *WORD* 10. 146–162. doi: 10.1080/00437956.1954.11659520.

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. doi:10.1016/B0-08-043076-7/02959-4.

Hay, Jennifer & Harald Baayen. 2003. Phonotactics, parsing and productivity. *Italian Journal of Linguistics* 1. 99–130. doi:10.1.1.171.705.

Huyghe, Richard & Marine Wauquier. 2020. What's in an agent? *Morphology* 30(3). 185–218. doi:10.1007/s11525-020-09366-2.

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 Laurie Bauer, Lívia Körtvélyessy & Pavol Štekauer (eds.), *Semantics of complex words*, 289–319. Dordrecht: Springer.

Kotowski, Sven, Viktoria Schneider & Lea Kawaletz. 2022. *Eventualities in nominalization semantics: The case of denominal -ment-formations*. Submitted.

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. In Nicoletta Calzolari (Conference chair), Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis & Takenobu Tokunaga (eds.), *Proceedings of the eleventh international* *conference on language resources and evaluation (lrec 2018)*, Miyazaki, Japan: European Language Resources Association (ELRA).

O'Rourke, N., L. Hatcher, E.J. Stepanski & I. SAS Institute. 2005. *A step-by-step approach to using SAS for univariate and multivariate statistics* Frommer's Complete Guides. Wiley.

Plag, Ingo. 2004. Syntactic category information and the semantics of derivational morphological rules. *Folia Linguistica* 38(3-4). 193–225. doi: 10.1515/flin.2004.38.3-4.193.

Plag, Ingo. 2018. *Word-formation in English* Cambridge textbooks in linguistics. Cambridge, United Kingdom and New York, NY: Cambridge University Press second edition.

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 SciencePress.

R Core Team. 2020. R: A language and environment for statistical computing. https://www.r-project.org/.

Schmitz, Dominic, Ingo Plag, Dinah Baer-Henney & Simon David Stein. 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. doi:10.3389/fpsyg.2021.680889.

Schneider, Viktoria. 2022. Eventualities in the semantics of denominal nominalizations. In Sven Kotowski & Ingo Plag (eds.), *The semantics of derivational morphology: Theory, methods, evidence*, de Gruyter. Accepted.

Sitikhu, Pinky, Kritish Pahi, Pujan Thapa & Subarna Shakya. 2019. A comparison of semantic similarity methods for maximum human interpretability. doi: 10.48550/ARXIV.1910.09129.

Szabó, Zoltán Gendler. 2015. Major parts of speech. *Erkenntnis* 80(S1). 3–29. doi: 10.1007/s10670-014-9658-1.

Tomaschek, Fabian, Peter Hendrix & R. Harald Baayen. 2018. Strategies for addressing collinearity in multivariate linguistic data. *Journal of Phonetics* 71. 249– 267. doi:10.1016/j.wocn.2018.09.004.

Van Rossum, Guido & Fred L. Drake. 2009. *Python 3 reference manual*. Scotts Valley, CA: CreateSpace.

Van Valin, Robert D. & Randy J. LaPolla. 1997. *Syntax: Structure, meaning and function* Cambridge textbooks in linguistics. Cambridge: Cambridge Univ. Press.

Wauquier, Marine, Cécile Fabre & Nabil Hathout. 2018. Différenciation sémantique de dérivés morphologiques à l'aide de critères distributionnels. *SHS Web of Conferences* 46. 08006. doi:10.1051/shsconf/20184608006.