Ambiguity and Word Classes Lol!

Alexander Rauhut
Institut für Englische Philologie
Freie Universität Berlin

2021-07-15

Contents

1 Introduction

Ambiguity is pervasive in language. blablabla

Conversion is a productive process that inevitably produces homophones. Additionally, noun-verb and verb-noun conversions are doubly homophonous with inflectional affixes in *-s*.

This study will investigate potential large scale effects that the homonymy of *-s* has.

The emergence of lexical clusters is conditioned by frequency. Various studies have provided evidence (review in Diessel 2016)

In addition to the shared form, there is extensive conversion in the English language, allowing verbs to be readily used as nouns and vice-versa. However, the function of verbal *-s* and nominal *-s* is quite different, and the immediate grammatical context can be expected to be enough for disambiguation.

In section 2, I will review some of the literature on ambiguity and ambiguity avoidance and gather information from other fields of linguistics. Section 3 is an explorative overview over some ambiguous uses of *-s*. The approach taken is using POS ambiguity tags and ranked association values (log-likelihood ratio/LLR) in the same spirit of collostruction analysis (Stefanowitsch & Gries 2003). In addition, I correlated the resulting association values with cosine similarities between the base forms and affixed forms taken from a simple GloVe embedding (Pennington, Socher & Manning 2014; Selivanov, Bickel & Wang 2020) trained on the BNC. The result shows that larger differences between base and affixed form lead to significantly fewer ambiguous tags.

Finally, in section 4, I will present a series of GAMLSS models (Generalized Additive Models for Location, Scale and Shape (Stasinopoulos, Rigby & Bastiani 2018; Rigby & Stasinopoulos 2005)) that attempt to model the influence of verb-noun or noun-verb conversion on the distribution of *-s* and *-ed* suffixation while controlling for textual dispersion, regularity of occurrence and contextual flexibility.

In fact, it is difficult to come up with genuine examples of structural ambiguity caused by the homonymy of *-s*.

The study will conclude with a series of generalized additive models in order to explore effects of conversion on the morphological composition of lexemes. The data was preprocessed using the *Corpus Workbench* (Evert & Hardie 2011), all analyses were coded in *R* 4.1.0 (R Core Team 2021; Dowle & Srinivasan 2021), figures were created with *ggplot2* (Wickham 2016).

2 Avoiding Ambiguity

formal equivalence homonymy and polysemy

that deletion is not permitted in non-extraposed subject clauses Wasow (2015)

(1) This paper demonstrates that the differences in duration of English final S as a function of the morphological function it expresses.

2.1 Brainstorm

Hypothesis: third person singular -s is avoided in noun-verb conversion take a polysemous word and gather identical n-grams

2.2 Literature

Studies like Plag, Homann & Kunter (2017) call into question whether there is perfect homonymy between different types of word final *-s*. The study found significant length differences, and hypothesizes that Yung Song et al. (2013) did not find any such effect, however, albeit on a more lexically restricted study.

In a more recent study, Tomaschek et al. (2021) show that length of final *-s* can be modelled as having a discriminatory function depending on the lexical and phonological context. In other words, the duration

was found to decrease with increasing contextual ambiguity. “Energy is not invested in a signal that creates confusion instead of clarity.” (Tomaschek et al. 2021: 154) This points to some degree of ambiguity sensitivity, even though it is not entirely clear what it means for ambiguity avoidance.

Systematic phonetic differences potentially contribute to the disambiguation

Wasow (2015) surveys a variety of studies ... and concludes that ambiguity avoidance has not been shown to be as common as expected considering the pervasiveness of ambiguities in language. Grice underestimated the pervasiveness of ambiguity. Provides an attempt at a taximony. - word order freezing as ambiguity avoidance - that deletion only in unambiguous syntactic context - German free syntax but default SVO reading. Wasow does not call it default.

Piantadosi, Tily & Gibson (2012)

Trott & Bergen (2020): Linguistic forms are expected to show a certain degree of ambiguity in order to provide an efficient system. Simulations suggest that ambiguity caused by homophony is more common in natural languages than expected, even when taking into account. Homophones are smoothed out in lexically neighboring areas.

3 Considerations

Textual co-occurrence can only serve as an upper bound. Prosodic markers in spoken language can further disambiguate most types of ambiguity. The same is true for more general types of background knowledge.

4 Methodology

4.1 Data

Lemmas that had *-s* forms that only occurred in totally fixed contexts, therefore scoring alpha (Hapax) values of 0 were excluded as outliers. Exclusion of those observations significantly improved distributional properties of the models. On the lower frequency bands this is due to a substantial decrease of noise as it is difficult to estimate the proportion of *-s* only based on a few occurrences of *-s*. For the higher frequency bands, it is in many cases equal to the exclusion of mostly names and idiomatic expressions. This is conceptually unproblematic since multi-word structures should be treated as individual units, and as a different category of lexeme.

4.2 Textual specificity

In order to measure the textual specificity of lexemes, and penalize forms with very specific contextually bound uses, two types of dispersion are used. One is the deviation of proportions across texts (DP.norm) (Gries 2008), more specifically the normalized version (cf. Lijffijt & Gries 2012). DP.norm is a corpus-part-based measure bound between 0 and 1. Values close to 1 indicate a high deviation, therefore a low dispersion. It can be interpreted as a measure of how evenly tokens are spread over the corpus parts.

It cannot account for short bursts of occurrences (example?), therefore, word growth dispersion (DWG) (Zimmermann 2020) is used as well, which is based on distances between occurrences. DWG is a measure of how regularly a token occurs across an entire corpus occurs across an entire corpus.

If ambiguity is regularly avoided, it should be expected that with growing word class ambiguity, the use of *-s* becomes more context-dependent. Considering a given dispersion or “clumpiness” of contexts in which a lexeme is likely to occur, the existence of avoidance-contexts should cause a penalty to dispersion and make the distribution clumpier.

Both measures were designed to measure dispersion or commonness of lexical items as extension to the most commonly used plain frequencies, however they highlight different aspects both conceptually and empirically (cf. Gries 2021 on using multiple measures of dispersion and association).

Problem: The contexts in which ambiguity is avoided/not avoided might be rather evenly dispersed themselves. This could mask potential clumpiness of *-s* occurrences.

It is difficult to identify avoidance contexts formally. Lexical and syntactical correlates of the avoided structure might be avoided as a side effect, and cause fuzzier overall differences in structure.

4.3 Lexical specificity

The ratio of hapaxes (henceforth α_1 Evert 2005: 130) on either side of a given lexeme was used as a simple measure of how fixed the immediate lexico-grammatical context is.

4.4 Model

While controlling for clumpiness with dwg did not substantially change the effect of lol, it did improve the fit of the model

why gam: (Wood et al. 2016; Wood 2017) Semi-parametric Both extremely high and extremely low values of dwg suggest special values. Most of the variables used in the model are not expected to have a monotonic relationship with the proportion of *-s* occurrences. Very badly dispersed lexemes are overrepresented in terms of frequency, and an extremely even dispersion suggests uses typical of function words. The same is true for productivity or contextual fixedness. Both extremely high and extremely low values are untypical.

dp did not improve the model significantly and was dropped also due to correlation with dwg.

Predicting counts of *-s* directly with poisson or negative binomial regression leads to problematic model properties such as highly skewed residuals. Human perception has been found to be more sensitive to proportional changes of stimuli rather than absolute ones (cf. Kromer 2003). With that in mind, the independent variable to be predicted was set to the proportion of occurrences of *-s*. As such the possible values can range from 0 to 1 inclusively. Therefore, the distribution chosen to be fitted was a zero-one-inflated beta distribution (Ospina & Ferrari 2012; Rigby et al. 2019). To allow for the necessary flexibility in the distribution, the final models were created with the R package *gamlss* (Rigby & Stasinopoulos 2005) using the smoothing splines provided by *mgcv* (Wood 2017).

Strongly kurtotic, high amount of extreme values, which is not unexpected for language data.

5 Conclusion

Word embeddings and transformer networks like BERT may provide an interesting route for further studies in ambiguity (e.g. Beekhuizen, Armstrong & Stevenson 2021; Du, Qi & Sun 2019; Wiedemann et al. 2019). Clustering techniques can be used to detect homonymy (Lee 2021), which could be used in corpus annotation as a replacement of classical lemmatization. If robust and carefully applied, this could potentially lead to decreases in noise. Those recent successes in practical application are promising for the use in more descriptive application. They can provide another angle on co-occurrence statistics, and were only sparsely used in this study since there has been little systematic application in corpus linguistics. Include more contextual and "world knowledge" information, such as images (e.g. Kottur et al. 2016; Shahmohammadi, Lensch & Baayen 2021). Recent advances in context embedding might provide tools to further test where real ambiguity exists and where it affects the system of language.

There are no significant differences in spread and frequency of *-s* in lexical forms that occur as verb-noun or noun-verb conversions. (???)

6 Bibliography

Beekhuizen, Barend, Blair C. Armstrong & Suzanne Stevenson. 2021. Probing lexical ambiguity: Word vectors encode number and relatedness of senses. *Cognitive Science* 45(5). e12943. https://doi.org/10.1111/cogs.12943. https://onlinelibrary.wiley.com/doi/abs/10.1111/cogs.12943.

Diessel, Holger. 2016. Frequency and lexical specificity in grammar: A critical review. In Heike Behrens & Stefan Pfänder (eds.), *Experience counts: Frequency effects in language*, 209–238. De Gruyter. https://doi.org/10.1515/9783110346916-009.

Dowle, Matt & Arun Srinivasan. 2021. *Data.table: Extension of 'data.frame'*. https://CRAN.R-project.org/package=data.table.

Du, Jiaju, Fanchao Qi & Maosong Sun. 2019. Using BERT for word sense disambiguation. *CoRR* abs/1909.08358. http://arxiv.org/abs/1909.08358.

Evert, Stefan. 2005. *The statistics of word cooccurrences: Word pairs and collocations*. Friedrich-Alexander-Universität Erlangen-Nürnberg PhD thesis. http://www.collocations.de/phd.html.

Evert, Stefan & Andrew Hardie. 2011. Twenty-first century corpus workbench: Updating a query architecture for the new millennium. University of Birmingham.

Gries, Stefan. 2021. A new approach to (key) keywords analysis: Using frequency, and now also dispersion. *Research in Corpus Linguistics* 9(2). 1–33. https://doi.org/10.32714/ricl.09.02.02.

Gries, Stefan Th. 2008. Dispersions and adjusted frequencies in corpora. *International journal of corpus linguistics*. John Benjamins 13(4). 403–437.

Kottur, Satwik, Ramakrishna Vedantam, José MF Moura & Devi Parikh. 2016. Visual word2vec (vis-w2v): Learning visually grounded word embeddings using abstract scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4985–4994.

Kromer, Victor. 2003. A usage measure based on psychophysical relations. *Journal of Quantitative Linguistics*. Taylor & Francis 10(2). 177–186.

Lee, Younghoon. 2021. Systematic homonym detection and replacement based on contextual word embedding. *Neural Processing Letters* 53(1). 17–36. https://doi.org/10.1007/s11063-020-10376-8.

Lijffijt, Jefrey & Stefan Th Gries. 2012. Correction to stefan th. Gries'"dispersions and adjusted frequencies in corpora" international journal of corpus linguistics 13: 4 (2008), 403-437. Citeseer.

Ospina, Raydonal & Silvia L. P. Ferrari. 2012. A general class of zero-or-one inflated beta regression models. *Computational Statistics & Data Analysis* 56(6). 1609–1623. https://doi.org/10.1016/j.csda.2011.10.005. https://www.sciencedirect.com/science/article/pii/S0167947311003628.

Pennington, Jeffrey, Richard Socher & Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 1532–1543.

Piantadosi, Steven T., Harry Tily & Edward Gibson. 2012. The communicative function of ambiguity in language. *Cognition* 122(3). 280–291. https://doi.org/10.1016/j.cognition.2011.10.004. https://www.sciencedirect.com/science/article/pii/S0010027711002496.

Plag, Ingo, Julia Homann & Gero Kunter. 2017. Homophony and morphology: The acoustics of word-final s in english. *Journal of Linguistics*. Cambridge University Press 53(1). 181–216.

R Core Team. 2021. *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Rigby, R. A. & D. M. Stasinopoulos. 2005. Generalized additive models for location, scale and shape, (with discussion). *Applied Statistics* 54.3. 507–554.

Rigby, Robert A, Mikis D Stasinopoulos, Gillian Z Heller & Fernanda De Bastiani. 2019. *Distributions for modeling location, scale, and shape: Using GAMLSS in r*. Chapman; Hall/CRC.

Selivanov, Dmitriy, Manuel Bickel & Qing Wang. 2020. *text2vec: Modern text mining framework for r*. https://CRAN.R-project.org/package=text2vec.

Shahmohammadi, Hassan, Hendrik Lensch & R Harald Baayen. 2021. Learning zero-shot multifaceted visually grounded word embeddings via multi-task training. *arXiv preprint arXiv:2104.07500*.

Stasinopoulos, Mikis D, Robert A Rigby & Fernanda De Bastiani. 2018. GAMLSS: A distributional regression approach. *Statistical Modelling*. SAGE Publications Sage India: New Delhi, India 18(3-4). 248–273.

Stefanowitsch, Anatol & Stefan Th Gries. 2003. Collostructions: Investigating the interaction of words and constructions. *International journal of corpus linguistics*. John Benjamins 8(2). 209–243.

Tomaschek, Fabian, Ingo Plag, Mirjam Ernestus & R. Harald Baayen. 2021. Phonetic effects of morphology and context: Modeling the duration of word-final s in english with naïve discriminative learning. *Journal of Linguistics*. Cambridge University Press 57(1). 123–161. https://doi.org/10.1017/S0022226719000203.

Trott, Sean & Benjamin Bergen. 2020. Why do human languages have homophones? *Cognition* 205. 104449. https://doi.org/10.1016/j.cognition.2020.104449. https://www.sciencedirect.com/science/article/pii/S0

010027720302687.
Wasow, Thomas. 2015. Ambiguity avoidance is overrated. In Susanne Winkler (ed.), *Ambiguity*, 29–48. De Gruyter. https://doi.org/10.1515/9783110403589-003.
Wickham, Hadley. 2016. *ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.
Wiedemann, Gregor, Steffen Remus, Avi Chawla & Chris Biemann. 2019. Does BERT make any sense? Interpretable word sense disambiguation with contextualized embeddings. *arXiv preprint arXiv:1909.10430*.
Wood, S. N. 2017. *Generalized additive models: An introduction with r*. 2nd edn. Chapman; Hall/CRC.
Wood, S. N., N., Pya & B. S"afken. 2016. Smoothing parameter and model selection for general smooth models (with discussion). *Journal of the American Statistical Association* 111. 1548–1575.
Yung Song, Jae, Katherine Demuth, Karen Evans & Stefanie Shattuck-Hufnagel. 2013. Durational cues to fricative codas in 2-year-olds' american english: Voicing and morphemic factors. *The Journal of the Acoustical Society of America*. Acoustical Society of America 133(5). 2931–2946.
Zimmermann, Richard. 2020. Word growth dispersion — a single corpus part measure of lexical dispersion. Paper presented at ICAME41 Heidelberg. https://www.youtube.com/watch?v=k8etOvRcF4c.