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*.

Lexical clustering blabla frequency blabla ((**diesel?**)) The emergence of those clusters is conditioned by frequency. Various studies have provided evidence (review in Diessel 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.

Recently, word embedding techniques have been used for various lexical tasks. Beekhuizen, Armstrong & Stevenson (2021): word2vec for distinguishing polysemes from homonyms?

Lee (2021) In combination with contextual word embeddings, clustering techniques have been used to detect homonymy.

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 Textual specificity

Two types of dispersion are used here. One is the deviation of proportions (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.

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.

4.2 Lexical specificity

5 Conclusion

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.

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.

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.

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.

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.

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*.

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/S0010027720302687.

Wasow, Thomas. 2015. Ambiguity avoidance is overrated. In Susanne Winkler (ed.), *Ambiguity*, 29–48. De Gruyter. https://doi.org/10.1515/9783110403589-003.

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.