Obsolescence: a missing piece in the puzzle of polysemy and time

Vivian G. Li[*]

[*]liguo.vivian@gmail.com
[*]Department of Linguistics, Yale University, New Haven, USA

Previous studies reported a positive correlation between a word's age and the number of senses it has. The underlying mechanism has been attributed to the idea that semantic changes inevitably occur over time (i.e., words do not stop changing in meaning even when it already has several senses). However, evidence from those studies falls short of fully supporting what the purported mechanism entails, as those studies focused exclusively on non-obsolete senses in non-obsolete words. A comprehensive test of the predictions of the diachronic mechanism requires evidence that the reported positive correlation extends to obsolete words, and that the correlation holds when obsolete senses are also counted. This study provides that missing piece for the puzzle. Examining over 36,000 English verbs, we show that longer lifespan correlates with a greater number of changes/senses in obsolete as well as non-obsolete words, though the effect is stronger in non-obsolete words.

1. Introduction

As a language evolves, the meanings of its words change, which often involves gaining or losing senses. While there have been extensive discussions on the nature and representation of polysemy (Robins, 1967; Geeraerts, 1993; Norvig & Lakoff, 1987), relatively less is known about its diachronic development, e.g. do words inevitably undergo more changes over time?

This question has been addressed in some studies, but results remain, in our opinion, inconclusive. Lee (1990) hypothesized that words that have entered the language early would be more polysemous than words that have entered the language more recently. The logic being that *older words have more time to undergo changes than younger words*. Analyzing three sets of about 200 randomly sampled English nouns and adjectives, Lee showed that year of entry indeed negatively correlated with the number of (non-obsolete) senses a word has. Flieller and Tournois (1994) similarly reported a significant correlation between year of entry and the status of being polysemous (as opposed to monosemous) in a sample of 998 French words consisting of nouns, verbs, adjectives and adverbs. Berdicevskis (2020) extended Lee (1990)'s analysis to all non-obsolete nouns, adjectives and verbs documented in the Oxford English Dictionary, and concurred with Lee (1990)'s conclusions.

We consider the evidence from Lee (1990) and subsequent studies inconclusive because it falls short of fully supporting what the purported mechanism entails. The underlying diachronic mechanism (i.e., time begets changes) amounts to claiming that "all words would undergo more *changes* (or, more *sense-gaining events*) over time," which is a prediction about obsolete as well as non-obsolete words, and is a prediction about the total number of changes (or the number of sense-gaining events) rather than the net amount of senses that survived till the present day. Lee (1990) and subsequent studies focused exclusively on non-obsolete senses in non-obsolete words, thus leaving their findings susceptible of survivorship bias.

A comprehensive test of the predictions of the diachronic mechanism therefore requires evidence that the reported positive correlation between a word's age and the number of senses/changes extends to obsolete words, and that the correlation holds when obsolete senses are counted. If, for instance, obsolete words are generally monosemous irrespective of the length of their lifespans, then the positive correlation would hold only for non-obsolete words; consequently, the correlation should not be taken as a diachronic generalization about polysemy. Similarly, if words that entered the language recently appear less polysemous (i.e., possess fewer non-obsolete senses) than words that entered early because they have disproportionally many obsolete senses, then the correlation is likely a corollary of older words having a higher proportion of non-obsolete senses than younger words do, rather than a diachronic regularity of polysemy.

In what follows, we provide the missing piece in the puzzle of polysemy and time, by extending Lee (1990) and Berdicevskis (2020)'s analyses to obsolete words and obsolete senses.

2. Method

Our primary source of data is verbs from the online version of the Oxford English Dictionary (OED). We focus on verbs because verbs tend to show more changes than nouns or adjectives (Dubossarsky, Weinshall, & Grossman, 2016).

OED organizes word senses into hierarchical groups, such that a word may have several major sense categories, and under each major sense category, there are sub-senses. Most of the subsenses are dated. We counted the dated senses as a proxy for the number of senses, on the assumption that if the senses could be dated (potentially differently), that in itself is evidence that those senses are distinct (see e.g. Hamilton, Leskovec, and Jurafsky (2016) on the difficulty in counting senses). The number of changes a word has gone through is counted as the sum of 1) the number of non-obsolete senses and 2) twice the number of obsolete senses, to take both gain and loss of senses into account.

Lifespan of a word is calculated as the span between the earliest and the latest year of use among all senses. If any of a word's senses are still in use, the latest year would be 2023. If the earliest year of use is "Old English" or "late Old English", we use 800 and 1025 as a proxy, as they are roughly the midpoint of those respective periods (Baugh & Cable, 1993; Sweet, 1990). If the earliest year of use is annotated with "*a*(nte)", "*c*(irca)", or ".." (e.g. *17..*), only the numeric value would be used for calculating the lifespan (e.g. *a1700, c1700, 17..* would all be treated as *1700*). If a word's lifespan thus calculated comes out as 1 (meaning the earliest and latest year of use are the same), and the earliest year of use contained annotations (e.g., "Old English"), the word would be excluded from analysis, since the estimated lifespan would likely deviate substantially from the actual value.

For nonobsolete words, we also extracted their frequency band value (this information is unavailable for obsolete words),[1] which reflects the words' overall frequency in written English from 1970 to the present. The frequency bands denote nine levels of frequencies (on a logarithmic scale),[2] with 8 being the most frequent ($>$ 1000 per million words), and 0 being the least frequent.

Where appropriate, we ran parallel analyses on a set of non-obsolete verbs, nouns, and adjectives extracted from the same source in 2019 (Berdicevskis, 2020).

3. Results

Both obsolete and nonobsolete verbs show an inverse relationship between the number of words with n changes/senses and *n* (Fig. 1). More than half of the words (14416 out of 26480 nonobsolete verbs; 7864 out of 9808 obsolete verbs) are monosemous. Correspondingly, those words have only minimal number of changes: one change (sense gain) for non-obsolete words, two changes (gain and loss) for obsolete words. The most polysemous non-obsolete verb (*set*) has undergone 335 changes, with a total of 262 senses, whereas the most polysemous obsolete verb (*yknow*) has undergone 38 changes, with 19 senses in total.

To examine the effect of lifespan and potential differences between obsolete and nonobsolete words (Fig. 2), we fit a Poisson regression model on the entire dataset, with the *number of changes* as the dependent variable, and *lifespan*, word's *state* (obsolete vs. nonobsolete), and their *interaction* as independent variables. Results (Table 1) revealed that a word is more likely to have a high number of changes when it has a longer lifespan (for nonobsolete words: $\beta_{lifespan} = 0.0037$; for obsolete words: $\beta_{lifespan} = 0.0013$) , although the effect is weaker in obsolete words than in nonobsolete words ($\beta_{lifeSpan \times state} = -0.0024, p < 0.0001$).

[1]OED distinguishes obsolete and non-obsolete words. A word is obsolete if all its senses are obsolete, "this usually means that no evidence for the term can be found in modern English" (Oxford University Press, n.d.b).

[2]Although the description on the official site (https://www.oed.com/information/understanding-entries/frequency/) states that there are eight levels of frequency band, some non-obsolete words are marked with 0 ("Frequency: 0 out of 8" in the html code). We therefore distinguish nine levels of frequencies.

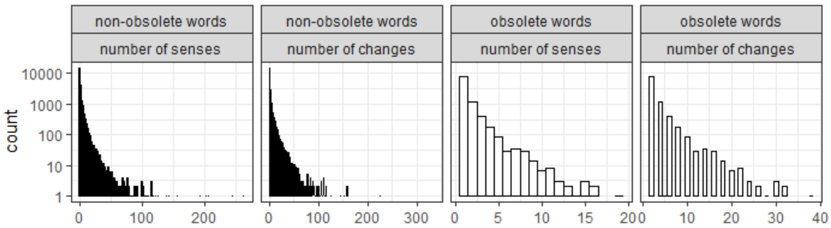

Figure 1. Histograms for the distribution of non-obsolete and obsolete words by the number of changes/senses they have. Y axis is on logarithmic scale.

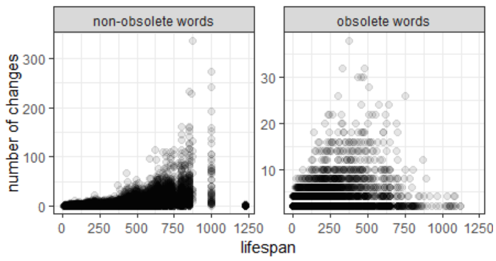

Figure 2. Distribution of lifespan by the number of changes, in non-obsolete and obsolete words. Each grey circle represents a data point (i.e. a word). Dark regions indicate overlapping data points.

Table 1. Coefficients from the regression model on the number of changes.

Predictor	Coefficient	SE	z value	p value
(Intercept)	-0.1240	0.0066	-18.65	< 0.0001
lifespan	0.0037	0.00001	338.66	< 0.0001
state=obsolete	0.9615	0.0102	94.42	< 0.0001
lifespan $\times$ state=obsolete	-0.0024	0.00003	-78.42	< 0.0001

A Poisson regression model using *lifespan*, *word state* (obsolete vs. nonobsolete), and their *interaction* to predict the total number of senses showed a similar pattern. Results echo the findings on the number of changes: a word is more likely to have more senses when it has a longer lifespan (for nonobsolete words: $\beta_{lifespan} = 0.0034$, for obsolete words: $\beta_{lifespan} = 0.0013$, although the effect is weaker in obsolete words than in nonobsolete words ($\beta_{lifespan \times state} = -0.0021, p < 0.0001$).

To check if the addition of frequency alters the results, we fit a Poisson regression model on non-obsolete verbs alone, with the *number of changes* as the dependent variable, and *lifespan*, *frequency-band*, and their *interaction* as independent variables. Since frequency band values reflect words' frequency differences on a logarithmic scale, we treat them as a continuous variable (c.f. Berdicevskis, 2020). Lifespan ($\beta = 0.0012, p < 0.0001$) remains a significant positive predictor for the number of changes, so are frequency band ($\beta = 0.2638, p < 0.0001$) and their interaction ($\beta_{lifespan \times freqBand} = 0.0003, p < 0.0001$). These results indicate that among non-obsolete verbs, a word is more likely to have a high number of changes if it is of a long lifespan or high frequency; the effect of lifespan is enhanced in high frequency words in comparison to low frequency words (or alternatively, the effect of frequency is more prominent among words with longer lifespan than among words of shorter lifespan). A Poisson regression model for the total number of senses in nonobsolete verbs showed the same pattern.

To gauge if our findings would generalize to other parts of speech, we fit Poisson regression models to the dataset used in Berdicevskis (2020). We regressed the number of changes/senses (total_nmeanings) against year of entry (with the earliest year, 950, re-set as 0, to make the intercept more interpretable, as in Berdicevskis, 2020), frequency band, part of speech, and all two-way and three-way interactions. The effects of year and frequency are consistent with our findings based on verbs alone (see Fig. 3 for an illustration of $year(lifespan) \times frequency$ interaction).

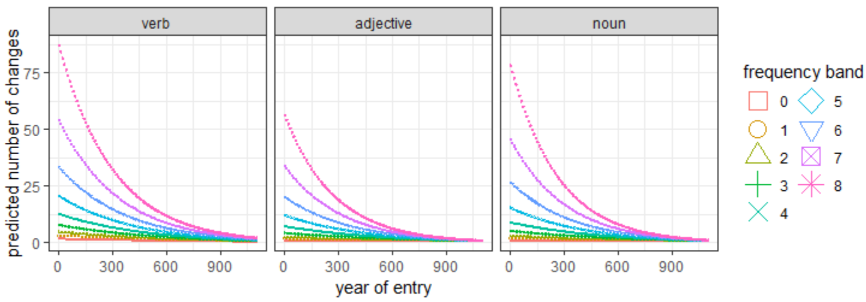

Figure 3. Predicted number of changes from the model based on Berdicevskis (2020)'s data.

4. Discussion

In this study we explored the relation between time and meaning change. Our analysis improves over previous ones in that ours incorporated obsolete words and obsolete senses, which are necessary components for assessing diachronic generalizations but have long been overlooked. We showed that longer lifespan is in general positively correlated with more changes in meaning (and the total number of senses), for non-obsolete and obsolete words alike, although the effect of lifespan is stronger in non-obsolete words than in obsolete words. Our finding complements Berdicevskis (2020), Lee (1990) and others and strengthens the claim that *words undergo more changes (or become more polysemous) over time*.

Our analyses also offer valuable insights to word obsolescence and the distribution of polysemy. Firstly, it has been claimed that "polysemy is pervasive" (Falkum & Vicente, 2015) and that "virtually every word is polysemous to some extent" (Vicente & Falkum, 2017), but such claims have not been empirically examined. Our results suggest that polysemy may not be as pervasive as expected, or rather, a large portion of the lexicon may be reasonably considered monosemous (for instance, as having one core meaning liable to contextual variation rather than having several more disparate senses). Our analyses are based on the OED ; future research could explore the empirical distribution of monosemy vs. polysemy with alternative approaches to track word/sense obsolescence and quantify polysemy. Secondly, in addition to the difference in the magnitude of lifespan effects, we observed a sharp contrast between obsolete and non-obsolete words in terms of the maximum number of senses (19 vs. 262). It is not clear if 19 represents an upper bound for the number of senses obsolete words could have. Future studies could delve into the nature of word obsolescence, and delineate how obsolete and non-obsolete words differ in other respects. Thirdly, we identified a Zipfian-like (Zipf, 1932) relation between the number of words with n senses and n. Subsequent investigations could survey the distribution of word senses in other languages, and look into the origins of this pattern.

In our analyses of obsolete words, we did not have their frequency information, and consequently excluded frequency as a predictor. The finding of significant lifespan effects in models without frequency indicates that lifespan may be a general effect in words of all frequencies. For comparison, we included frequency-band in models for non-obsolete words alone. Although frequency has been shown to correlate with semantic change and polysemy (Hamilton et al., 2016; Zipf, 1945), recent studies argued that frequency is not a causal factor for semantic change (Dubossarsky et al., 2016; Dubossarsky, Weinshall, & Grossman, 2017; Keidar, Opedal, Jin, & Sachan, 2022) and that change in meaning precedes changes in frequency (Feltgen, Fagard, & Nadal, 2017). Corroborating Berdicevskis (2020), frequency-band in our models correlated significantly with the number of changes/senses, and its interaction with lifespan was significant. We hypothesize that frequency, if it is not a causal factor, may have functioned as a proxy for some covert factor, e.g. prototypicality of senses (Dubossarsky et al., 2016). Note that in those models, the effect of lifespan was not subsumed by frequency, which strengthens the validity of lifespan as an independent predictor for the number of changes/senses. Irrespective of how frequency effects are interpreted, the significant interaction between lifespan and frequency implies, at least, that subclasses of words may exhibit diverse diachronic trajectories. Future studies may explore further how subclasses of a part of speech could vary systematically in their diachronic behaviour.

References

Baugh, A., & Cable, T. (1993). *A history of the English language.* Routledge.

Berdicevskis, A. (2020). Older English words are more polysemous. In A. Ravignani, C. Barbieri, M. Flaherty, Y. Jadoul, E. Lattenkamp, H. Little, M. Martins, K. Mudd, & T. Verhoef (Eds.), *The Evolution of Language: Proceedings of the 13th International Conference (Evolang13)* (p. 14-21).

Dubossarsky, H., Weinshall, D., & Grossman, E. (2016). Verbs change more than nouns: a bottom-up computational approach to semantic change. *Lingue e linguaggio*, *15*(1), 7–28.

Dubossarsky, H., Weinshall, D., & Grossman, E. (2017). Outta control: Laws of semantic change and inherent biases in word representation models. In *Proceedings of the 2017 conference on empirical methods in natural language processing* (pp. 1136–1145).

Falkum, I. L., & Vicente, A. (2015). Polysemy: Current perspectives and approaches. *Lingua*, *157*, 1-16.

Feltgen, Q., Fagard, B., & Nadal, J.-P. (2017). Frequency patterns of semantic change: corpus-based evidence of a near-critical dynamics in language change. *Royal Society open science*, *4*(11), 170830.

Flieller, A., & Tournois, J. (1994). Imagery value, subjective and objective frequency, date of entry into the language, and degree of polysemy in a sample of 998 French words. *International Journal of Psychology*, *29*(4), 471–509.

Geeraerts, D. (1993). Vagueness's puzzles, polysemy's vagaries. *Cognitive Linguistics*, *4*, 223-272.

Hamilton, W. L., Leskovec, J., & Jurafsky, D. (2016). Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistic* (p. 1489-1501). Stroudsburg, PA: Assoc. Comput. Linguist.

Keidar, D., Opedal, A., Jin, Z., & Sachan, M. (2022). Slangvolution: A causal analysis of semantic change and frequency dynamics in slang. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (p. 1422–1442). Dublin, Ireland: Association for Computational Linguistics.

Lee, C. J. (1990). Some hypotheses concerning the evolution of polysemous words. *Journal of Psycholinguistic Research*, *19*, 211–219.

Norvig, P., & Lakoff, G. (1987). Taking: A study in lexical network theory. In J. Aske, N. Beery, L. A. Michaelis, & H. Filip (Eds.), *Annual meeting of the Berkeley Linguistics Society* (Vol. 13, pp. 195–206). Berkeley: Berkeley Linguistics Society.

Oxford University Press. (n.d.a). *Oxford English Dictionary.* `https://www.oed.com/`. (Accessed: 2023-09-30)

Oxford University Press. (n.d.b). *OED terminology.* `https:`

//www.oed.com/information/understanding-entries/oed-terminology. (Accessed: 2024-01-08)

Robins, R. H. (1967). *A short history of linguistics.* Bloomington: Indianna University Press.

Sweet, H. (1990). *A new English grammar: Logical and historical. part 1: Introduction, phonology and accidence* (Vol. 1). Oxford At The Clarendon Press.

Vicente, A., & Falkum, I. L. (2017). Polysemy. In *Oxford research encyclopedia of linguistics.* Retrieved 11 Oct. 2023, from https://oxfordre.com/linguistics/view/10.1093/acrefore/9780199384655.001.0001/acrefore-9780199384655-e-325.

Zipf, G. K. (1932). *Selected studies of the principle of relative frequency in language.* Harvard university press.

Zipf, G. K. (1945). The meaning-frequency relationship of words. *The Journal of general psychology*, *33*(2), 251–256.