OLDER ENGLISH WORDS ARE MORE POLYSEMOUS

ALEKSANDRS BERDICEVSKIS

aleksandrs.berdicevskis@gu.se
Språkbanken (The Swedish Language Bank), University of Gothenburg
Gothenburg, Sweden

Word meanings change over time, usually following routes predicted by general cognitive principles. While significant advances in understanding lexical meaning change have been made, relatively few studies have focused on large-scale quantitative testing of the proposed meaning change laws. It has, for instance, been hypothesized that older words are on average more polysemous, since they have had more time to develop new meanings through meaning shifts. I perform a large-scale quantitative test of this hypothesis, extracting data for 16K English verbs, 45K adjectives and 102K nouns from the Oxford English Dictionary. I show that the hypothesis holds, but the correlation between age and polysemy depends on the word frequency, being stronger for the more frequent words.

1. Introduction

Studying semantic change can give us insights into language evolution, if we manage to understand cognitive processes that underly the change better (Hoefler, & Smith, 2008). An important type of semantic change is change in lexical meaning. Word meanings vary and change, usually following routes predicted by general cognitive principles, such as metaphor, metonymy, generalization and specialization (Nerlich, & Clarke, 2003).

While significant advances in understanding lexical meaning change have been made, relatively few studies have focused on large-scale quantitative testing of the proposed meaning change laws (but see, for example, Urban, 2011; Hamilton, Leskovec, & Jurafsky, 2016; Winter, Thompson, & Urban, 2014; Xu, Malt, & Srinivasan, 2017).

In this paper, I quantitatively test the assumption that older words are on average more polysemous (Lee, 1990). Since words become polysemous through meaning shifts, it is reasonable to expect that older words, which have had more time to develop additional meanings, would have done so.

While a plausible hypothesis, this is not necessarily true. Other factors might dwarf word's age and/or interact with it in complicated ways. Besides, meanings not only emerge, but also disappear, and, applying the same logic, one can predict that older words have had more chances to lose the existing meanings. Since the rates of the emergence and disappearance of lexical meanings are unknown, we cannot claim with certainty which of these diachronic process is dominant.

In other words, whether older words are more polysemous is an empirical question. I am aware of but two studies that address it empirically. Lee (1990) demonstrated that word age positively correlates with polysemy for two samples of 200 English nouns and one sample of 208 English adjectives. Flieller and Tournois (1994) studied a sample of 998 French nouns, and while, having other research questions, they did not focus on the relation between age and polysemy, they also report a positive correlation.

In this paper, I demonstrate positive correlation between word age and polysemy for three parts of speech (verbs, adjectives and nouns), not restricting myself to small samples, but using all words available in the Oxford English Dictionary (OED Online, 2019). The correlation coefficients I report can be used to quantify the average rate at which words develop new meanings.

2. Materials and methods

I browse the online edition of the OED,[1] extracting for every word its part of speech, number of separate meanings, date of entry and frequency.

I focus on three parts of speech: verbs, adjectives and nouns. Parts of speech may differ notably in their semantic behavior (and how lexicographers analyze its behavior), which is why I perform all comparisons only within parts of speech. For technical reasons, I ignore entries that ascribe two different parts of speech to a single lemma (e.g. *Aalenian*, *n.* and *adj.*). This, however, happens rarely: in most cases, if a word is polysemous across parts of speech, then each part of speech has its own entry (e.g. *iron* has separate entries as a noun, an adjective and a verb).

Homonyms (i.e. words that have the same graphical form, but are assigned to different entries, e.g. *abate*$_1$ 'to end' and *abate*$_2$ 'to seize') are treated as different words.

Entries marked as obsolete (by a cross † preceding the headword) are ignored.

For most entries, the OED provides the year when the word has first been attested in writing. While this, of course, is just an approximation to the real age of the word, it is as good as we can hope to get. Entries where no date is provided

[1] http://www.oed.com/, accessed April 2019

or where the information is considered unreliable (preceded by *ca* or *ante*, or represented as e.g. *17*..) are ignored, as are entries where *OE*, *ME* (resp. Old English, Middle English) etc. is provided instead of year. For date ranges like *1641-1642*, the year before the hyphen is treated as the date of entry. For early periods, the OED does not provide exact years (using instead notation like *OE*). However, automatic browsing results in small number of entries with suspiciously early dates (e.g. *170* or *688*). Manual check shows that most, if not all, entries with years earlier than 951 are due to errors at the OED website. For this reason, they are also ignored.

In order to establish how polysemous a word is I calculate a number of meanings listed within the entry. The OED distinguishes meanings at several levels: overarching meanings (labelled by Roman numerals), more specific meanings within each Roman-numeral meaning (labelled by Arabic numerals), submeanings within each Arabic-numeral meanings (labelled by small letters). I count the Arabic-numeral meanings, since they are closest to most traditional understandings of "different meanings of the same word". If there are no Arabic numerals within the entry, the word is considered to have a single meaning. Obsolete meanings, marked by a cross before the Arabic numeral, are ignored. If there is no cross, the meaning is *not* considered obsolete (and thus included in the analysis), even though there might be a note like *obsolete* or *archaic* within the definition. The reason is that the positioning and wording of such notes is not systematic and they cannot be reliably processed automatically. If all the Arabic-numeral meanings within the entry are obsolete, the word is ignored.

Frequency has been shown to be a major factor affecting polysemy (Hernández-Fernández et al., 2016; Fenk-Oczlon, & Fenk, 2010; Zipf, 1945). The OED entries do not contain exact frequency data, but they provide a frequency band the word belongs to, ranging from 1 (extremely rare) to 8 (very frequent).

It would have been better to use continuous frequency data rather than binned, but in order to obtain accurate frequency estimates a substantial amount of manual work is required (dealing with spelling variation, homonyms, morphological forms; comparing data from different corpora). Since this work has already been done by the OED editors when estimating frequency bands, I am relying on their data.

Some OED entries may differ from the principles that the automatic extraction described in this section relies upon, either due to different editorial policies in different periods or random errors and inconsistencies. This means there might be some noise in the data. Some entries containing obvious mistakes were manually removed, and spot checks did not reveal neither systematic biases nor random errors.

See supplementary materials for the scripts for processing the OED entries, the extracted data and the scripts for statistical analysis.

3. Results

Distribution of word counts per frequency bands and parts of speech is summarized in Table 1.[2]

Table 1. Distribution of word counts per frequency bands and parts of speech.

	Band 1	Band 2	Band 3	Band 4	Band 5	Band 6	Band 7	In total
Adjectives	5420	22998	9146	5345	1609	182	4	44704
Nouns	11013	43789	26922	15036	4555	883	72	102337
Verbs	1317	6317	3957	2961	1207	307	35	16101

For illustration purposes, the relation between age and polysemy for nouns from band 6 is represented on Figure 1.

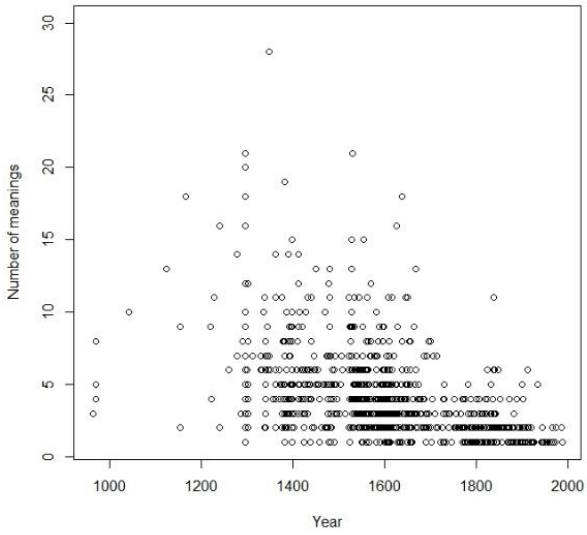

Figure 1. Number of meanings and year of entry for all nouns from frequency band 6.

[2] Interestingly, the distribution does not entirely follow the one that could be expected according to Zipf's law: there are always more words in band 2 than in band 1. It is probably explained by the fact that very infrequent words are less likely to get into a dictionary.

To estimate the effect of word age, I fit a Poisson regression model with number of meanings as the dependent variable, date of entry (YEAR) as a continuous predictor, part of speech (POS) as a categorical predictor (with adjectives as the reference level) and FREQUENCY BAND as a reverse-Helmert-coded categorical predictor. All two-way and three-way interactions are also included. To make the intercept more interpretable, it was set to the year 950 instead of 0 (the earliest words included in the analysis are dated 951).

The summary of the regression model are presented in Table 2. For brevity's sake, only the coefficients for YEAR, FREQUENCY BAND, POS and the two-way interactions between YEAR and the other two predictors are included (see the supplementary materials for the full summary of the model).

Table 2. Summary of the Poisson regression model: polysemy as predicted by year of entry, frequency and part of speech. Asterisk (*) marks significance at 0.05 level. See main text for more details.

Coefficient	**Estimate**	**SE**	**z-value**	**Pr(>|z|)**
(Intercept)	1.4e+00	1.1e-01	12.4	<0.001*
year	-1.0e-03	2.1e-04	-4.9	<0.001*
freq.band 2	1.9e-01	5.9e-02	3.2	0.002*
freq.band 3	2.3e-01	2.6e-02	8.7	<0.001*
freq.band 4	2.3e-01	1.7e-02	13.4	<0.001*
freq.band 5	2.1e-01	1.8e-02	11.9	<0.001*
freq.band 6	1.8e-01	3.5e-02	5.3	<0.001*
freq.band 7	2.7e-01	1.0e-01	2.6	0.008*
POS noun	1.7e-01	1.1e-01	1.5	0.130
POS verb	4.2e-01	1.2e-01	3.4	<0.001*
year × freq.band 2	-1.7e-04	6.8e-05	-2.5	0.013*
year × freq.band 3	-2.0e-04	3.1e-05	-6.5	<0.001*
year × freq.band 4	-2.0e-04	2.2e-05	-9.3	<0.001*
year × freq.band 5	-1.7e-04	2.5e-05	-6.9	<0.001*
year × freq.band 6	-1.1e-04	5.1e-05	-2.2	0.030*
year × freq.band 7	-1.1e-04	2.0e-04	-0.6	0.570
year × POS noun	-1.6e-04	2.1e-04	-0.8	0.451
year × POS verb	-3.6e-04	2.3e-04	-1.6	0.119

YEAR has a negative coefficient which is significantly different from zero, which means that older words do indeed have more meanings. FREQUENCY BANDS always have positive coefficients (reverse Helmert coding means that we are comparing words from band 2 with words from band 1, words from band 3 with words from bands 2 and 1, etc.). This reflects the well-established fact that more frequent words tend to be more polysemous (Hernández-Fernández et al., 2016; Fenk-Oczlon, & Fenk, 2010; Zipf, 1945). Verbs, according to the model, are significantly more polysemous than adjectives, while nouns are not.

All but one interactions between YEAR and FREQUENCY BAND have significant (but small) negative coefficients, which means the negative slope is steeper for higher bands. The only exception is band 7, probably due to the very small number of words in it. In other words, for more frequent words age matters more in terms of polysemy; the difference between older and newer words is larger. Interestingly, Lee (1990) does not observe an interaction effect between frequency and age in his data.

Among the coefficients that are not listed in Table 2, five are significant: the interaction between FREQUENCY BANDS 4, 5, 6 and POS verb (0.16, 0.06 and 0.09 respectively), between FREQUENCY BAND 5 and POS noun (0.05), between YEAR, FREQUENCY BAND 4 and POS verb (-8.7e-05); see supplementary materials for further details.

4. Discussion

One goal of the computational approaches to semantic change is to discover fundamental patterns of meaning evolution. Hamilton, Leskovec and Jurafsky (2016), for instance, provide evidence for *the law of conformity* (more frequent words have slower rate of semantic change) and *the law of innovation* (more polysemous words have higher rate of semantic change). This paper provides evidence for *the law of age*: older words are more polysemous.

The estimated rates of change, reported in Table 2, vary across parts of speech and words of different frequency. Apart from confirming that more frequent words are more polysemous, the results show that words from higher frequency bands develop new meanings at faster rates than words from lower bands, i.e. that the correlation between age and polysemy is stronger for frequent words. More detailed investigation using continuous frequency data would be required to understand the interaction between age and polysemy more precisely.

Quantification of semantic change enables us to test the existing qualitative theories about meaning: do the observed results fit with the theoretical predictions? Quantification also makes it possible to predict future changes or to reconstruct the earlier stages of the language.

Further research avenues can include:

- reproducing the study using corpus data instead of dictionary data (to estimate both the age of the word and the number of meanings, using automated sense-induction methods), although that would require large high-quality diachronic corpora;
- reproducing the study for other languages;
- quantifying the rate of disappearance of existing meanings;

- collecting more data about when new meanings appear (the year of the first known usage is provided in the OED for every meaning) in order to explore whether the trajectory is linear or has some other form;
- establishing semantic relations between older and newer meanings (is the new meaning the result of a metaphorical shift, or bleaching, or something else?). That would require either extensive manual annotation or high-quality automatic tools.

Supplementary materials

See https://github.com/AleksandrsBerdicevskis/polysemy.

References

Fenk-Oczlon, G., & Fenk, A. (2010). Frequency effects on the emergence of polysemy and homophony. *International Journal of Information Technologies and Knowledge, 4(2)*, 103–109.

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.

Hamilton, W., 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 Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, pp. 1489–1501.

Hernández-Fernández, A., Casas, B., Ferrer-i-Cancho, R., & Baixeries, J. (2016). Testing the Robustness of Laws of Polysemy and Brevity Versus Frequency. In: Pavel Král, Carlos Martín-Vide (eds). *Statistical Language and Speech Processing. SLSP 2016. Lecture Notes in Computer Science*, 9918, 19–29. Springer, Cham.

Hoefler, S., & Smith, A. D. (2008). Reanalysis vs. metaphor? What grammaticalisation can tell us about language evolution. *In The evolution of language: Proceedings of the 7th International Conference (EVOLANG7)*, pp. 163–170.

Lee, Christopher. (1990). Some hypotheses concerning the evolution of polysemous words. *Journal of Psycholinguistic Research, 19, 4*, 211–219.

Nerlich, B., & Clarke, D. (2003). Polysemy and flexibility: introduction and overview. In Brigitte Nerlich, David D. Clarke (eds). *Polysemy. Flexible Patterns of Meaning in Mind and Language*. Mouton de Gruyter, pp. 3–30.

OED Online. (2019). *Oxford English Dictionary Online.* Oxford University Press. Accessed April 2019.

Urban, M. (2011). Asymmetries in overt marking and directionality in semantic change. *Journal of Historical Linguistics*, 1, 3–47.

Winter, B., Thompson, G., & Urban, M. (2014). Cognitive factors motivating the evolution of word meanings: Evidence from corpora, behavioral data and encyclopedic network structure. In *Evolution of Language: Proceedings of the 10th International Conference (EVOLANG10)*, pp. 353–360.

Xu, Y., Malt, B. C., & Srinivasan, M. (2017). Evolution of word meanings through metaphorical mapping: Systematicity over the past millennium. *Cognitive Psychology*, 96, 41–53.

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