The inevitability and pervasiveness of semantic change poses challenging questions for theories of linguistic and cultural evolution: Why do the meanings of some words change faster than others? How does frequency of usage influence such changes in meaning?

Finding quantitative answers to these questions requires new methods that can go beyond the case-studies of a few words (often followed over widely different time-periods) that are our most common diachronic data (Bréal, 1897; Ullmann, 1962; Blank, 1999; Hopper and Traugott, 2003; Traugott and Dasher, 2001).

One promising avenue is the use of distributional methods, in which words are embedded in vector spaces according to their co-occurrence relationships (Bullinaria and Levy, 2007; Turney and Pantel, 2010), and the embeddings of words are then compared across time-periods. This new direction has been effectively demonstrated in a number of case-studies (Sagi et al., 2011; Wijaya and Yeniterzi, 2011; Gulordava and Baroni, 2011; Jatowt and Duh, 2014) and used to perform large-scale linguistic change-point detection (Kulkarni et al., 2014) as well as to test a few specific hypotheses, such as whether English synonyms tend to change meaning in similar ways (Xu and Kemp, 2015). However, these works employ widely different embedding approaches and test their approaches only on English.

In this work, we (i) systematically compare several state-of-the-art distributional approaches on the task of quantifying semantic change, (ii) test these approaches on four languages, and (iii) show that these approaches can be used to reveal quantitative laws of semantic evolution.

We begin by analyzing how different distributional measures give rise to different notions of semantic change. We show that certain measures are more sensitive to meaning-changes in nominal domain, while other measures are more sensitive to changes in verbs, adverbs, and adjectives. This distinction aligns with previous research on historical semantic change, where it has been noted that nominals are more likely to undergo changes due to irregular sociocultural shifts while the other major parts-of-speech participate in more regular processes of semantic change, such as subjectification or grammaticalization (Traugott and Dasher, 2001).

We also compare the quality of these different distributional measures, and different embedding approaches (raw PPMI, word2vec, and SVD), using a novel set of diachronic benchmarks. These benchmarks test whether the methods detect attested examples of semantic change and also evaluate the quality of changes that they automatically discover. These experiments reveal that low-dimensional embeddings provide higher-quality estimates of semantic change, compared to high-dimensional raw PPMI vectors.

Finally, we build off these insights and show...
that distributional approaches can be used to uncover statistical laws of semantic change. We consider two related, central, and unanswered questions in semantic change.

One is the role of frequency. Frequency plays a key role in other linguistic changes, associated sometimes with faster change—sound changes like lenition occur in more frequent words—and sometimes with slower change—high frequency words are more resistant to morphological regularization (Bybee, 2007; Pagel et al., 2007; Lieberman et al., 2007). Our analysis reveals that frequency has a strong stabilizing effect on word semantics: across languages, rates of semantic change were found to be proportional to a small negative power of word-frequency.

Another unanswered question is the relationship between semantic change and polysemy. Words gain senses over time as they semantically drift (Bréal, 1897; Wilkins, 1993; Hopper and Traugott, 2003), and polysemous words\(^1\) occur in more diverse contexts, affecting lexical access speed (Adelman et al., 2006) and rates of L2 learning (Crossley et al., 2010). But we don’t know whether the diverse contextual use of polysemous words makes them more or less likely to undergo change (Geeraerts, 1997; Winter et al., 2014; Xu et al., 2015). Furthermore, polysemy is strongly correlated with frequency—high frequency words have more senses (Zipf, 1945; İlgen and Karaöglan, 2007)—so understanding how polysemy relates to semantic change requires controlling for word frequency. We show that, after controlling for frequency, polysemy exhibits a strong positive association with rates of semantic change.

Overall, these two factors—frequency and polysemy—are found to explain between 48% and 88% of the variance\(^2\) in rates of semantic change (across languages). This remarkable degree of explanatory power indicates that frequency and polysemy are perhaps the two most crucial linguistic factors that explain rates of semantic change over time.

Our analysis highlight a new role for frequency and polysemy in language change and the importance of distributional models in diachronic research.

\(^1\)We use ‘polysemy’ here to refer to related senses as well as rarer cases of accidental homonymy.

\(^2\)Marginal \(R^2\) (Nakagawa and Schielzeth, 2013).

References


Bahar İlgen and Bahar Karaöglan. 2007. Investigation of Zipf’s ‘law-of-meaning’ on Turkish corpora. In *22nd International Symposium on Computer and Information Sciences*, pages 1–6. IEEE.