Of crowds and corpora: A marriage of measures

Emmanuel Keuleers, Paweł Mandera, Michaël Stevens, & Marc Brysbaert
Department of Experimental Psychology
Ghent University
Henri Dunantlaan 2, 9000 Gent, Belgium
{emmanuel.keuleers, pawel.mandera,
michael.stevens, marc.brysbaert}@ugent.be

Abstract

We discuss the relationship between a word's corpus frequency and its prevalence –the proportion of people who know the word– and show that they are complementary measures. We show that adding word prevalence as a predictor of lexical decision reaction time in the Dutch lexicon project increases explained variance by more than 10%. In addition, we show that, for the same dataset, word prevalence is the best independent predictor of word processing time.

1 Introduction

Word frequency is one of the most important measures in the cognitive study of word processing, both theoretically and methodologically. Its contribution in explaining behavioural measures such as reaction time is so large that researchers take great care in collecting large and reliable corpora and in applying the best possible word frequency estimates in their research.

1.1 Where the corpus is weak the crowd is strong

A drawback of frequency counts is that, regardless of corpus size, lower counts are unreliable. As an example, consider asking a random sample of 100 people whether they know each of the word types that occur just once in a large corpus. Although frequency for all these types is equal, the number of judges knowing each word will vary from zero to one hundred and, as the judges are language users, words known to many of them may be considered to occur more often in language than words which are known by fewer of them. Following this reasoning, the estimate of *the number of language users who know a word*, or *word prevalence* may give a better indication of occurrence than corpus frequency counts.

1.2 Where the corpus is strong the crowd is weak

On the other hand, consider presenting the same random sample of people with words from the language's core vocabulary. Since these words will be known to all of the judges, *prevalence* will be singularly high and uninformative. In this case corpus counts should be a much better estimate of occurrence.

2 Testing the prevalence measure

To test the complementarity of prevalence and frequency as measures of occurrence, we used prevalence norms for Dutch collected through a lexical decision task presented as an online vocabulary test (Keuleers, Stevens, Mandera, & Brysbaert, in press). Each participants saw 100 stimuli (about 70 words and 30 nonwords) selected randomly from a list of 54,319 words and 21,734 nonwords. In the current analysis, we used the data of 190,771 participants who indicated that they were living in Belgium, giving us about 250 observations per word. The score for a word obtained by fitting a Rasch model –a mathematical model simultaneously ranking participants by ability and test-items by difficulty– to the data was considered an operationalization of its prevalence.

In Vito Pirrelli, Claudia Marzi, Marcello Ferro (eds.): *Word Structure and Word Usage*. Proceedings of the NetWordS Final Conference, Pisa, March 30-April 1, 2015, published at http://ceur-ws.org

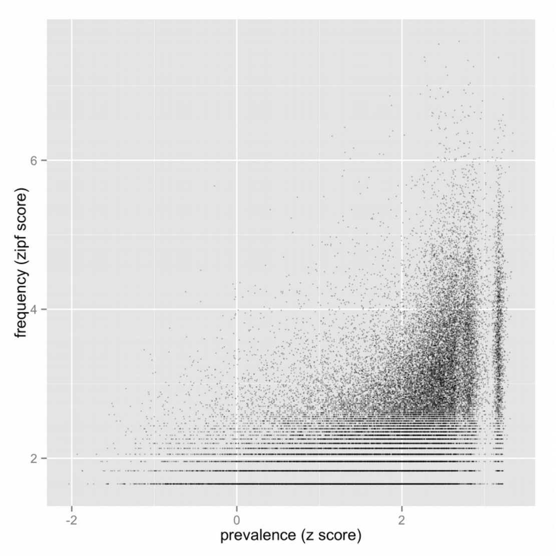

Figure 1: The relationship between frequency and prevalence. Word frequency is displayed as Zipf-score (log frequency per billion words; Van Heuven et al., 2014).

Figure 1 shows the complementary relation between the SUBTLEX-NL word frequencies (based on 42 million word corpus of film and television subtitles; see Keuleers, Brysbaert, & New, 2012) and the prevalence measure obtained from the online vocabulary test. Higher z-scores indicate more prevalent words. The dark lines at the bottom half of the plot indicate words with singularly low frequencies over a large range of prevalence. The elongated cluster at the right side of the plot shows words with nearly full prevalence over large frequency ranges.

In addition, we investigated the relationship between prevalence and other typical measures of word frequency. Table 1 gives an overview of these correlations.

	Frequency	Prevalence	OLD 20	Length
Frequency	1.00	0.35	-0.34	-0.37
Prevalence	0.35	1.00	0.00	0.07
OLD20	-0.34	0.00	1.00	0.74
Length	-0.37	0.07	0.74	1.00
Contextual Diversity	0.98	0.36	-0.34	-0.35

Table 1: Correlations between main predictors of Lexical Decision RT in the Dutch Lexicon ProjectTable 1 shows that the correlation between prevalence and frequency was relatively low (.34), giving further evidence that prevalence is distinct from word frequency and contextual diversity –a word's document count– which correlates very highly with word frequency.

Finally, we used the data from the 7,885 items in the Dutch Lexicon Project (Keuleers et al., 2010) for which both frequency and prevalence were available to examine the contributions of Dutch corpus word frequency (SUBTLEX-NL, Keuleers et al., 2010) and word prevalence on average reaction times.

In single variable analyses, log word frequency explained about 36.13% of the variance in reaction times and prevalence explained about 33.03% of the variance in reaction times.

This was also made clear when both measures were considered in the same analysis, where both measures jointly explained 51.37 % of the variance in reaction times. The unique contributions to explained variance (eta-squared) were 27.39% for frequency and 23.87% for prevalence. In further analyses, we found that including the quadratic trend of word frequency and contextual diversity did not substantially alter this pattern of results.

3 Conclusion

The results show that, next to word frequency, prevalence is by far the most important independent contributor to visual word recognition times, suggesting that prevalence should be included in any analysis where word corpus frequency is considered to be relevant. However, several questions remain open. First, what is the influence of corpus size on the relation between corpus word frequency and prevalence and on the contribution of prevalence to lexical processing? Second, how well does prevalence perform on others tasks and in other languages? Finally, does the effect of prevalence on word processing truly lie in a better measurement of word occurrence or does it partly reflect an independent property associated with the learnability of a word?

Acknowledgments

The text of this abstract is an early summary of findings from a larger study reported in the Quarterly Journal of Experimental Psychology as *Word knowledge in the crowd: Measuring vocabulary size and word prevalence in a massive online experiment.* (Keuleers, E., Stevens, M., Mandera, P., & Brysbaert, M., in press).

References

Balota, D. A., Yap, M. J., Hutchison, K. A., Cortese, M. J., Kessler, B., Loftis, B., … Treiman, R. (2007). The English lexicon project. *Behavior Research Methods*, 39(3), 445–459.

Keuleers, E., Brysbaert, M., & New, B. (2010). SUBTLEX-NL: A new measure for Dutch word frequency based on film subtitles. *Behavior Research Methods, 42*(3), 643–650. doi:10.3758/BRM.42.3.643

Keuleers, E., Diependaele, K., & Brysbaert, M. (2010). Practice Effects in Large-Scale Visual Word Recognition Studies: A Lexical Decision Study on 14,000 Dutch Mono- and Disyllabic Words and Nonwords. *Frontiers in Psychology, 1*. doi:10.3389/fpsyg.2010.00174

Keuleers, E., Lacey, P., Rastle, K., & Brysbaert, M. (2011). The British Lexicon Project: Lexical decision data for 28,730 monosyllabic and disyllabic English words. *Behavior Research Methods, 44*(1), 287–304. doi:10.3758/s13428-011-0118-4

Keuleers, E., Stevens, M., Mandera, P., & Brysbaert, M. (in press). Word knowledge in the crowd: Measuring vocabulary size and word prevalence in a massive online experiment. *Quarterly Journal of Experimental Psychology*. doi:10.1080/17470218.2015.1022560