An evaluation of the Google Books ngrams for psycholinguistic research

Emmanuel Keuleers[*], *Marc Brysbaert*[*], *Boris New*[‡]

Google recently released ngram frequencies based on Google Books, a massive collection of digitized book volumes published between 1550 and 2008 (Michel et al., 2011). As an indication of the size: the Google Books corpus is claimed to cover about 4 % of all books ever published, and for English, the corpus represents about 361 billion word tokens.

The Google books database is a potential goldmine for psycholinguistic research, but care should be taken not to overestimate its value, based only on size or reputation of the publishers.

We present a critical analysis of the Google Books 1-grams (word frequencies, page counts and document counts for billions of words) for English and French, evaluating their use for psycholinguistic research. Firstly, we examine how the Google Books 1-grams over the years predict lexical decision reaction times and accuracies (measured by R Squared) from various larger and smaller lexical decision and naming megastudies, such as the English Lexicon project (Balota et al., 2007) and the French lexicon project (Ferrand et al., 2010). This analysis reveals several anomalies, such as a sudden drop in R Squared in the middle of the 20th century (Figure 3.1), and the rather strange finding that word frequencies from the year 1800 are better predictors of naming latencies than

[*]Ghent University
[‡]CNRS and University Paris Descartes

word frequencies from 2008 (Figure 3.2). Then, we compare the R Squared values obtained with the Google Book frequencies to those of the SUBTLEX film subtitle frequencies for English (Brysbaert & New, 2009) and French (New, Brysbaert, Veronis, & Pallier, 2007), which have been proven excellent predictors of behavioral task measures. The analyses show that the R Squared values obtained with Google Book frequencies rarely match the R Squared value obtained with these smaller databases. Finally, we establish the 'age' of different current and less current word frequency measures used in psycholinguistics by observing in which years they reach peak correlations with the Google Books frequencies.

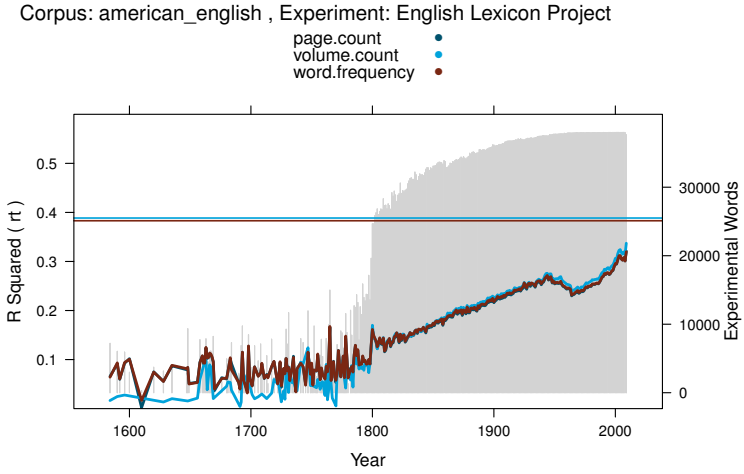

Figure 3.1: Percentages of variance explained by the Google American English ngrams in the RT data of the English Lexicon Project as a function of the years in which the books were published. The three lines indicate different values reported by Google: the number of occurrences of the word, the number of pages on which the word occurs, and the number of books in which the word appears. The light grey bars indicate the number of words from the English Lexicon Project found in the Google books over the various years (ordinate to the right). The red horizontal line indicates the percentage of variance explained by SUBTLEX-US word frequency; the blue horizontal line indicates the percentage of variance explained by the number of SUBTLEX-US films in which the word appears. RT data based on words with accuracy > .66.

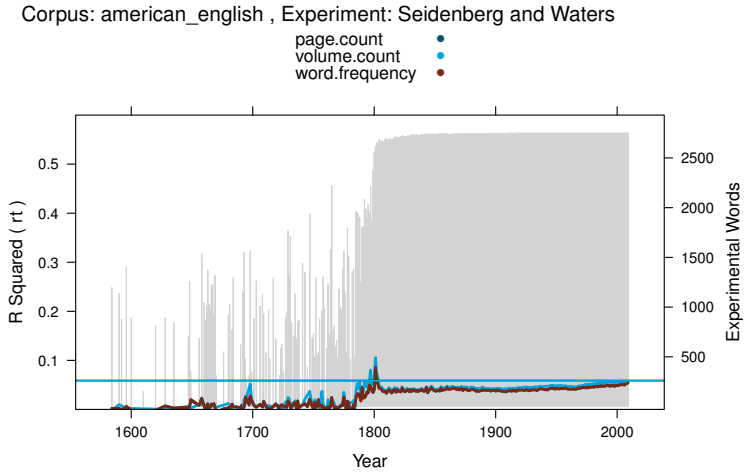

Figure 3.2: Percentages of variance explained by the Google American English ngrams in the naming latencies of the Seidenberg & Waters (1989) word naming study. Horizontal lines: Percentages of variance explained by SUBTLEX- US.

Contact: Emmanuel Keuleers <emmanuel.keuleers@ugent.be>

References

Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. A., Kessler, B., Loftis, B., et al. (2007). The English Lexicon Project. *Behavior Research Methods*, *39*(3), 445–459.

Brysbaert, M., & New, B. (2009). Moving beyond Kucera and Francis: a critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, *41*(4), 977–990.

Ferrand, L., New, B., Brysbaert, M., Keuleers, E., Bonin, P., Meot, A., et al. (2010). The French Lexicon Project: Lexical decision data for 38,840 French words and 38,840 pseudowords. *Behavior Research Methods*, *42*(2), 488–496.

Michel, J. B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Google Books Team, et al. (2011). Quantitative analysis of culture using millions of digitized books. *Science*, *331*, 176–182.

New, B., Brysbaert, M., Veronis, J., & Pallier, C. (2007). The use of film subtitles to estimate word frequencies. *Applied Psycholinguistics*, *28*(4), 661–677.

Seidenberg, M. S., & Waters, G. S. (1989). Word recognition and naming: a mega study. *Bulletin of the Psychonomic Society*, *27*, 489.