Cognitive Psychology 74 (2014) 66–83

Contents lists available at ScienceDirect

Cognitive Psychology

journal homepage: www.elsevier.com/locate/cogpsych

Corpus-based estimates of word association predict biases in judgment of word co-occurrence likelihood

Denis Paperno *[,1], Marco Marelli [1], Katya Tentori [1], Marco Baroni [1]

Center for Mind/Brain Sciences, University of Trento, Italy

ARTICLE INFO

Article history:
Accepted 17 July 2014
Available online 25 August 2014

Keywords:
Word association
Confirmation
Probability judgment
Linguistic corpora

ABSTRACT

This paper draws a connection between statistical word association measures used in linguistics and confirmation measures from epistemology. Having theoretically established the connection, we replicate, in the new context of the judgments of word co-occurrence, an intriguing finding from the psychology of reasoning, namely that confirmation values affect intuitions about likelihood. We show that the effect, despite being based in this case on very subtle statistical insights about thousands of words, is stable across three different experimental settings. Our theoretical and empirical results suggest that factors affecting traditional reasoning tasks are also at play when linguistic knowledge is probed, and they provide further evidence for the importance of confirmation in a new domain.

1. Introduction

It has long been observed that the linguistic competence of native speakers is affected by the statistical distribution of linguistic units in natural speech (Bod, Hay, & Jannedy, 2003; Bybee, 2007). Unfortunately, it is impossible to reconstruct the whole linguistic experience of a single speaker. However, a vast literature has shown that *corpora*, that is, very large collections of texts (millions or billions of words) produced in natural communicative situations provide reasonable estimates of the statistical patterns encountered in the experience of an average speaker, and thus

* Corresponding author.

[1] Denis and Marco M. share first autorship, Katya and Marco B. share senior authorship.

http://dx.doi.org/10.1016/j.cogpsych.2014.07.001
0010-0285/

can be successfully used in empirical models of language (Lüdeling & Kytö, 2008; Manning & Schütze, 1999).

One of the most robust generalizations emerging from corpus-based studies is that many linguistic phenomena are not only influenced by the absolute frequency of co-occurrence of words (or other linguistic units), but also by their degree of statistical association. A variety of *association measures*, meant to quantify to what degree two words tend to occur together beyond chance in a given corpus, have thus been proposed and used (Evert, 2005). Among them, the oldest and still most widely used is *Pointwise Mutual Information* (*PMI*; Church & Hanks, 1990):

$$\mathrm{PMI} = \log_2 \frac{P(w_1, w_2)}{P(w_1)P(w_2)} = \log_2 \frac{P(w_1|w_2)}{P(w_1)} = \log_2 \frac{P(w_2|w_1)}{P(w_2)} = \log_2 \frac{f(w_1, w_2)}{E(w_1, w_2)} \quad (1)$$

defined in terms of probabilities P of word occurrence, where the last expression shows how PMI is computed when standard maximum likelihood estimates of probabilities are assumed: $f(w_1, w_2)$ is the absolute co-occurrence frequency of words w_1, w_2; $f(w_{1|2|})$ is the absolute frequency of word $w_{1|2|}$; and $E(w_1, w_2) = P(w_1)P(w_2) \times N = f(w_1)f(w_2)/N$ (for N the sample size, e.g., the number of words in the source corpus) is the expected frequency of co-occurrence of w_1, w_2 under the hypothesis of independence. For two words to have high PMI, it is not sufficient nor necessary to co-occur frequently in absolute terms (this is also true for most other association measures). Rather, the observed co-occurrence count of the two words must be higher than what is expected by chance given their independent frequencies. In other words, a relatively low absolute co-occurrence frequency might lead to high association if the words of interest are very rare, whereas high co-occurrence frequency does not imply strong association for very frequent words.

The original interest in PMI and other association measures stemmed from the empirical observation that they can predict the degree to which a word pair behaves, linguistically, as a single unit (a *multi-word expression*, such as *Hong Kong* or *red herring*: see, e.g., Church & Hanks, 1990; Evert, 2008; Sag, Baldwin, Bond, Copestake, & Flickinger, 2002). But in the last few decades PMI and association in general have been shown to play a fundamental role in modeling a much wider variety of linguistic and psycholinguistic phenomena. To cite just a few examples, Ellis and Simpson-Vlach (2009) found that PMI scores significantly predict acceptability intuitions about word sequences, the speed in starting to articulate the sequence when reading it aloud, and the speed in producing the last word in a sequence after reading those preceding it. Durrant (2008) found that PMI is a good predictor of free association and the degree of priming of modifier-noun pairs. McDonald and Ramscar (2001) (who used an association measure closely related to PMI, called the Log Odds-Ratio measure) showed that similarity judgments about marginally familiar or nonce words and target terms were predicted by the presence of words with high target-term association in the context of the rated words (e.g., *samovar* was judged more similar to *kettle* if presented in a context containing other words with high statistical association to *kettle*). Recchia and Jones (2009) showed that PMI scores of word pairs are highly correlated to human semantic relatedness and synonymy judgments about the same pairs. Pitler, Bergsma, Lin, and Church (2010) found that PMI scores predict the correct bracketing of complex noun phrases (e.g. *retired* [*science teacher*] vs. [*retired science*] *teacher*). Pantel and Lin (2002) clustered words based on their profile of PMI association with a set of context terms, and found that measuring the similarity of the PMI distribution of a single word to multiple clusters of words is an effective way to discover and characterize the different senses of a word. Bullinaria and Levy (2007, 2012) found that vectors recording PMI-based co-occurrence profiles of words perform best, among many competitors, in a variety of tasks such as semantic categorization of words or identifying synonyms.

Interestingly, the notion of *association* from corpus-based linguistics is immediately related to the notion of *confirmation* as developed, completely independently, in epistemology and the psychology of reasoning. Two distinct notions have been identified which are relevant in describing an inductive inference from evidence e to hypothesis h. The first is the posterior probability of h in light of e, $P(h|e)$, and has its normative benchmark in Bayes theorem.[2] The second is known as confirmation,

[2] The *posterior* probability of h after seeing evidence e is of course the *conditional* probability of h given e : $P(h|e)$. We will thus refer to this same quantity indifferently as posterior or conditional probability.