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

Denis Paperno, Marco Marelli, Katya Tentori, Marco Baroni

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

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.

Keywords: word association, confirmation, probability judgment, linguistic corpora

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 et al., 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 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):

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

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 et al., 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 & 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 & 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 & Jones (2009) showed that PMI scores of word pairs are highly correlated to human semantic relatedness and synonymy judgments about

the same pairs. Pitler et al. (2010) found that PMI scores predict the correct bracketing of complex noun phrases (e.g. *retired* [*science teacher*] vs. [*retired science*] *teacher*). Pantel & 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 is an effective way to discover and characterize the different senses of a word. Bullinaria & 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 indentified 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, $P(h \mid e)$, and has its normative benchmark in Bayes theorem.[1] The second is known as confirmation, $c(h, e)$, and indicates the impact of e on the credibility of h. In particular, confirmation is positive iff $P(h \mid e) > P(h)$, negative iff $P(h \mid e) < P(h)$, and null otherwise. A plurality of alternative formal models, also known as "confirmation measures" (Fitelson, 1999), have been proposed by Bayesian epistemologists to quantify the degrees of confirmation, for example the r measure (Keynes, 1921), identical to PMI modulo the logarithm base:

$$r(h, e) = \ln \frac{P(h \mid e)}{P(h)} \tag{2}$$

Probability and confirmation are to some degree logically independent, even though both express statistical dependence between evidence and a hypothesis. Imagine drawing at random a student (X) from your university. You are provided only with one piece of evidence concerning the selected student: "X is a male" (e). In light of this information, do you think that it is more probable that "X likes shopping" (h_1) or that "X likes cigars" (h_2)? According to a recent survey run at UCL (Tentori et al., 2014), 61% of males versus 85% of female students like shopping, while 38% of males versus 12% of female students like cigars. This means that, at least as far as the UCL student population is concerned, h_1 is more probable than h_2 in light of e (i.e., $P(h_1 \mid e) = .61 > P(h_2 \mid e) = .38$). However h_2 is confirmed by e more than h_1 (i.e., $c(h_1, e) < c(h_2, e)$) because evidence e increases the probability of h_2 while it decreases that of h_1.

One can illustrate the same point with a linguistic example concerning the probability of words. Take a random sentence and consider the following two hypotheses: "the sentence contains the word *question*" (h_1) and "the sentence contains the word *manpower*" (h_2). Now assume you are given the evidence that "the sentence contains the word *resources*" (e). This evidence affects the two hypotheses at issue in different ways. The presence of *manpower* is highly confirmed by the presence of *resources*. Because of the meaningful association between these two words (manpower is a kind of resource), the probability of *manpower* increases by an order of magnitude compared to its baseline. On the other hand *resources* bears no particular relation to *question*. Nevertheless, because *question* is such a frequent word, its probability in the context of *resources* is still much higher (precisely, ten times higher) than the probability of *manpower* in the presence of *resources*. This example is another illustration of possible dissociation between confirmation and probability:

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

the presence of the word *resources* confirms the presence of the word *manpower* more than that of *question* (i.e., $c(h_1, e) < c(h_2, e)$), but the probability of the word *question* is still higher than that of *manpower* even in presence of the word *resources* (i.e., $P(h_1 \mid e) > P(h_2 \mid e)$).

Previous studies showed that confirmation judgments are, in general, rather reliable (Tentori et al., 2007; Mastropasqua et al., 2010) and more accurate and consistent over time than probability judgments (Tentori et al., 2014). Furthermore, (Lagnado & Shanks, 2002) find in a simulated medical diagnosis task that human probability judgments are based on confirmation (which they call 'predictiveness') rather than normative probability. Moreover, it seems that some typical biases reported in the probabilistic reasoning literature can be explained in terms of confirmation. A prominent example is provided by the conjunction fallacy (Tversky & Kahneman, 1982, 1983), which arises when the probability of two hypotheses occurring together (e.g, "Erik is blond and has blue eyes") is judged higher than the probability of one of them taken in isolation (e.g, "Erik is blond"). For a long time, it has been assumed (see, for example, Fantino et al., 1997; Costello, 2009; Yates & Carlson, 1986) that the conjunction fallacy rates are directly proportional to the perceived probability of the added conjunct ("Erik has blue eyes" in the example above). However, it has been recently shown (Crupi et al., 2008; Tentori et al., 2013) that the conjunction fallacy does not require the perceived probability of the added conjunct to be high, but it is really crucial that the added conjunct is perceived as inductively confirmed (by the available evidence and/or the other conjunct). This means that judgments (and errors) in a "pure" probabilistic task can be influenced by implicit assessments of confirmation.

Bayesian confirmation is applicable to reasoning about all kinds of probabilistic inferences and language should be no exception, as linguistic events — such as occurrence of certain words or constructions — can be easily seen as stochastic. In the three experiments of the current study, we test whether confirmation relations, as measured by corpus-based estimates of word association, affect participants' judgments of word probabilities. To the best of our knowledge, we are the first to pursue this question (for speakers' estimates of *single* word usage probabilities, see Alderson, 2007, and Balota et al., 2001). In particular, we compute and use the r measure of confirmation defined in Equation 2 above since it is equivalent up to multiplication by a constant (due to the different logarithmic basis) to the PMI score in Equation 1 above.[2]

Drawing an explicit connection between word association and confirmation opens a bridge between (corpus-based) linguistics and the psychology of reasoning that brings a variety of advantages for both fields. To start with, it is interesting to see whether phenomena which have been reported in the psychology of reasoning literature are also generalizable to the perceived probability of words. Furthermore, text corpora, a standard research tool in contemporary linguistics, can also be very useful in the study of reasoning, because they reflect statistical knowledge that has been implicitly developed over the years, rather than being either externally provided by the experimenter (as in traditional reasoning tasks employing abstract materials) or artificially acquired in training sessions (as in "natural sampling" experiments). From the linguistic point of view, our

[2]We choose r/PMI over alternative confirmation models for concreteness and due to the popularity of PMI in corpus studies, but we believe that our results do not depend on the choice of measure. We have also estimated word association using the l confirmation measure (Good, 1984), that ranked at the top in the empirical evaluation of confirmation measures of Tentori et al. (2007), and found that for our dataset of word co-occurrences l and r are almost identical, with Pearson correlation coefficient over .99. Other confirmation measures are similar to the z-score association measure, which in turn is known to be highly correlated with PMI for word co-occurrence data (Evert, 2005). These considerations allow us to use PMI/r as a representative of a variety of confirmation/word association measures; see Appendices A and B for examples of other confirmation and word association measures.

results offer a new perspective on the notion of statistical word association, which, as discussed above, is pervasive in language studies, and they show that phenomena that have been extensively studied in another branch of cognitive science under the confirmation rubric also hold of intuitions about word co-occurrence. This further supports the psychological validity of the word association measures used in linguistics, suggesting that they formalize some general principles that underlie human cognition.

2. General Method

We ran three experiments testing whether corpus-based association measures are predictive of human intuitions on word probabilities. The three experiments differ in the materials employed, but present a common experimental paradigm and procedure, which are described in the current section. The details of each experiment will then be presented in the relevant sections of the paper.

2.1. Participants

Participants were recruited from Amazon Mechanical Turk through the CrowdFlower platform (`www.crowdflower.com`), and were rewarded with $1 for every 100 judgments. Only participants from English-speaking countries were admitted, and they were requested to accept the task only if they were English native speakers. Crowdsourcing is a widespread method in economics and in the social sciences (Paolacci et al., 2010) and aims at fast collection of large amounts of data by exploiting online surveys. Recently it has been shown (Munro et al., 2010) that, as far as linguistic materials are tested, crowdsourcing provides as reliable data for psychological experiments as those collected in traditional, paper-and-pencil experiments.

2.2. Procedure

Participants were asked to perform a forced choice between two candidate target words (hence *targets*: $target_1$ and $target_2$) in the context of another word (hence *context*). Participants were asked to be as accurate as possible. We stressed that some comparisons could be more difficult than others, and that they would have informed us about this through confidence ratings. The following is a schematic example of a trial:

Imagine that someone comes up with a sentence containing the word
context (e.g., *alcohol*)

Which of the following is more likely?

☐ the sentence also contains the word $target_1$ (e.g., *fun*)

☐ the sentence also contains the word $target_2$ (e.g., *ingredient*)

After each question, we asked participants for how confident they were of their answer being correct on a 5-point scale ranging from "not confident at all" (1) to "completely confident" (5).

We adopted a forced-choice task, requiring participants to tell whether one target was more likely than another given the context, in order to avoid the more difficult challenge of providing a quantitative estimate of very small probability values.

2.3. Materials

Our experimental stimuli are based on word co-occurrence data from a huge text collection, comprised of the concatenation of ukWaC (`wacky.sslmit.unibo.it`), a mid-2009 dump of the English Wikipedia (`en.wikipedia.org`) and the British National Corpus (`www.natcorp.ox.ac.uk`), for a total of 2.8 billion words. The corpus has been tokenized, annotated with part-of-speech information and lemmatized with the TreeTagger software (Schmid, 1995). Statistics were extracted at the lemma level, that is, ignoring inflectional information; for example, word form *dogs* was counted as an instance of *dog*. The targets in our dataset come from the most frequent 8K nouns. The contexts are the 5,525 most frequent nouns, with the exclusion of the top 142 nouns. The rationale for the narrower frequency filtering of contexts was twofold. On the one hand, we wanted the contexts to be informative, as the experiments are designed to test the impact of context on the likelihood of targets. For this reason we excluded the most frequent nouns which tend to have quite general meanings and co-occur with words of all sorts. On the other hand, we aimed at excluding very rare words. For this purpose, we employed minimum frequency ranks of 5,525 for contexts and 8,000 for targets. Two distinct thresholds were used, since each context was presented with multiple targets. Therefore, if a participant failed to recognize a rare context, she would not be able to assess the conditional probability of any target presented with it. For targets this concern is less stringent than for contexts because even if one particular target is not recognized by the participant, conditional probability of other targets in the context can still be estimated.

In each of the three experiments, we selected a sample of context-target pairs in the corpus according to the criteria specific to the experiment. We then manually filtered out words that we judged to be outside of common English lexicon (such as *barangay* 'the smallest territorial unit in the Philippines'), contexts co-occurring with less than 50 targets, acronyms and very short words (of fewer than 3 characters, probably due to typographic errors), words that have salient homonyms of a different part of speech (e.g., *to break* vs. *a break*), and words that have spelling variants (*behavior* vs. *behaviour*); both homonymy and spelling variants could have led to inaccurate corpus-based probability estimates. After that, we performed a random selection of the desired number of stimuli.

Co-occurrence data collected from the corpora indicate how many times each target is attested in the same sentence with each context. From the co-occurrence data, for each target–context pair we calculated estimates of the conditional probability of target t in context c

$$P(t \mid c) = \frac{f(t,c)}{f(c)}$$

and confirmation:

$$C(t,c) = r(t,c) = \ln \frac{P(t \mid c)}{P(t)} = \ln \frac{f(t,c)N}{f(t)f(c)} \propto PMI(t,c)$$

where N is the corpus size in sentences, $f(t)$ and $f(c)$ are the number of sentences in which the target and the context occur, respectively, and $f(t,c)$ is the number of sentences that contain both the target and the context.

For each word pair, we calculate the percentile of their probability and confirmation values among all context-target pairs that co-occur at least once in the corpus. All zero values of estimated

probability (i.e., pairs of words never attested within the same sentence) were excluded from consideration.

As it is clear from the discussion above, our co-occurrence data were sentence-based. In principle, one could also use other criteria, such as counting words that occur within a certain distance limit from each other (i.e., *window-based* counts) or words linked by a certain syntactic relation (i.e., *syntactic* counts). For example, one could count targets that occur within 5 words from the context, or targets that syntactically modify the context. We chose sentence-based probabilities because we believe that the corresponding question ("Which word is more likely to occur within the same sentence as the word *alcohol*?") is more natural than the alternatives (e.g. "Which word is more likely to occur within five words on the right or five words on the left from the word *alcohol*?").

3. Experiment 1

3.1. Aim

Experiment 1 was meant to be the first investigation of the effects of the confirmation between words on their perceived probability. This effect was investigated in the most controlled setting possible, that is in target pairs with equal conditional probabilities, $P(target_1 \mid context) = P(target_2 \mid context)$. Under these conditions, if participants' probability judgments were only driven by $P(target \mid context)$, we would expect the choice rate for either target to be at chance level (50%). On the contrary, if confirmation relations affect probability judgments, we should find participants' choice to be influenced by $C(target,\ context)$.

3.2. Stimulus selection

The employed set of stimuli included 50 contexts. Every context was associated with 12 targets, 4 for each of 3 probability levels, as defined on the basis of $P(target \mid context)$: around 35^{th} (conditional probability: $mean = .00012; SD = .00009$), around 65^{th} (conditional probability: $mean = .00038; SD = .00011$), and around 95^{th} (conditional probability: $mean = .00469; SD = .00096$) probability percentile. No restrictions were imposed on $C(target,\ context)$. Each of the 4 targets associated with the same context and belonging to the same probability level was combined with the other 3 targets, leading to 6 context-target$_1$-target$_2$ tuples in which $P(target_1 \mid context)$=$P(target_2 \mid context)$. Each tuple constituted an experimental trial, for a total of 900 trials (50 contexts $\times$ 3 probability levels of the targets $\times$ 6 tuples).

An example of tuples with *addition* as common context is reported in Table 1. As can be seen, most of the differences in $C(target,\ context)$ are not easy to access on an intuitive basis. Figure 1 reports the overall distribution of the $C(target,\ context)$ values for each probability level.

Each trial was judged by 10 participants (for a total of 9,000 datapoints). A given participant could answer to no more than 120 trials. The order of targets in each pair was counterbalanced (i.e., 5 judgments were collected for the target$_1$-target$_2$ order, and 5 judgments for target$_2$-target$_1$ order).

3.3. Data analysis

Mixed-effects logit models were employed as the primary statistical tool (Baayen et al., 2008, Jaeger, 2008). We used as binary dependent variable which target was chosen in each pairwise comparison, coding target$_1$ as 0 and target$_2$ as 1. Therefore, the estimated parameters will indicate the change in the probability ratio of choosing target$_2$ over target$_1$.

$target_1$	$target_2$	$C(target_1,\ context)$	$C(target_2,\ context)$	Probability level
chore	waterfall	.711	-.202	35^{th}
chore	likeness	.711	.227	35^{th}
chore	spear	.711	.280	35^{th}
waterfall	likeness	-.202	.227	35^{th}
waterfall	spear	-.202	.280	35^{th}
likeness	spear	.227	.280	35^{th}
bench	judgement	.232	-.073	65^{th}
bench	signing	.232	.535	65^{th}
bench	bean	.232	.377	65^{th}
judgement	signing	-.073	.535	65^{th}
judgement	bean	-.073	.377	65^{th}
signing	bean	.535	.377	65^{th}
money	employee	.117	.778	95^{th}
money	job	.117	.134	95^{th}
money	strategy	.117	.559	95^{th}
employee	job	.778	.134	95^{th}
employee	strategy	.778	.559	95^{th}
job	strategy	.134	.559	95^{th}

Table 1: Target pairings for the context *addition* in Experiment 1. Each context-$target_1$-$target_2$ tuple corresponds to one experimental trial. E.g., participants were asked if, given the context *addition*, it would be more likely that the same sentence would also contain the target *chore* or the target *waterfall*.

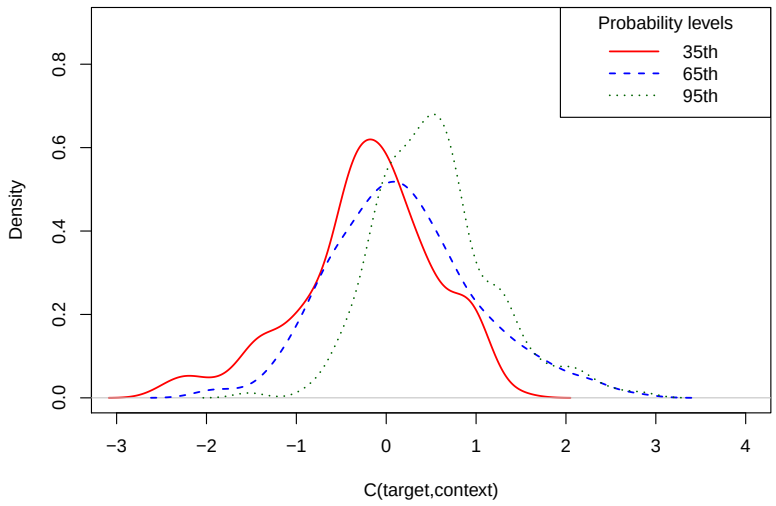

Figure 1: Distribution of the confirmation values at the three probability levels of Experiment 1.

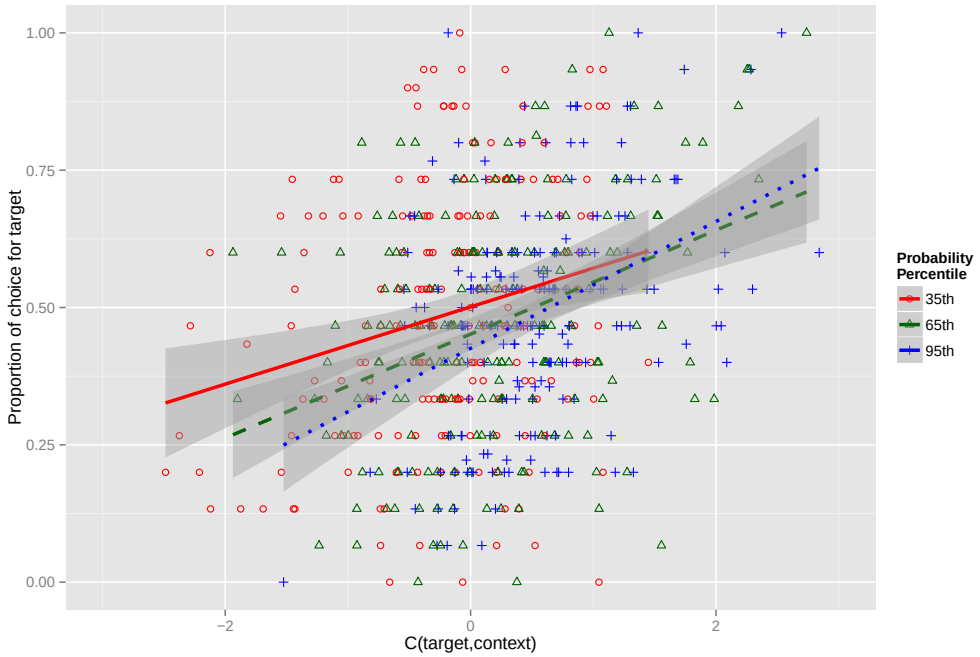

Figure 2: Association between $C(target,\ context)$ and probability of choosing the target at the different probability levels.

We introduced as predictor the difference in $C(target,\ context)$ between target_2 and target_1. The resulting measure, $\Delta C(target,\ context)$, will indicate whether $C(target,\ context)$ was in favor of target_2, and to what extent (with negative values indicating greater $C(target,\ context)$ values for target_1 and positive values indicating greater $C(target,\ context)$ for target_2). Therefore, we expect a positive association between $\Delta C(target,\ context)$ and the dependent variable. We also tested the effect of the probability level (sum-coded), as well as its interaction with $\Delta C(target,\ context)$. Non-independence of observations was dealt with by introducing a complex random structure, including participants and both targets as separate random intercepts. Low-confidence (≤ 2) judgments were excluded from the analysis (21% of the datapoints). Parameters that did not significantly improve the model fit were removed (effects were evaluated one by one on the basis of likelihood ratio tests). Once the final model was identified, atypical outliers were removed (employing 2.5 SD of the residual errors as criterion). The model was then refitted to ensure that the results did not depend on those few outliers. Removed outliers constituted less of 0.01% of the total datapoints.

3.4. Results

A total of 174 distinct participants performed experiment 1, each doing an average of 52 trials. Inter-rater agreement between participants was 77.4%. Figure 2 represents the association between each word $C(target,\ context)$ and its probability of being chosen by participants in the whole data set.

The results of our analyses, and the associated statistical tests, are reported in Table 2. There is a significant effect of $\Delta C(target,\ context)$ on the participants' choice: the higher $\Delta C(target,\ context)$, the more likely it is that target_2 will be chosen. When $\Delta C(target,\ context)$ is close to zero, that is, when $C(target,\ context)$ of target_1 and target_2 is very close, participants' choice is at chance level.

	Estimate	Std. Error	z-value	P
(Intercept)	-.0178	.0827	-0.22	.8301
$\Delta C(target,\ context)$	.6579	.0498	13.21	.0001
65^{th} probability	-.1322	.0862	-1.53	.1251
95^{th} probability	.0817	.0906	0.91	.3667
$\Delta C(target,\ context)$:$65^{\text{th}}$	-.0346	.0675	-0.51	.6082
$\Delta C(target,\ context)$:$95^{\text{th}}$	.1575	.0761	2.07	.0384

Table 2: Fixed effects in the final model of Experiment 1.

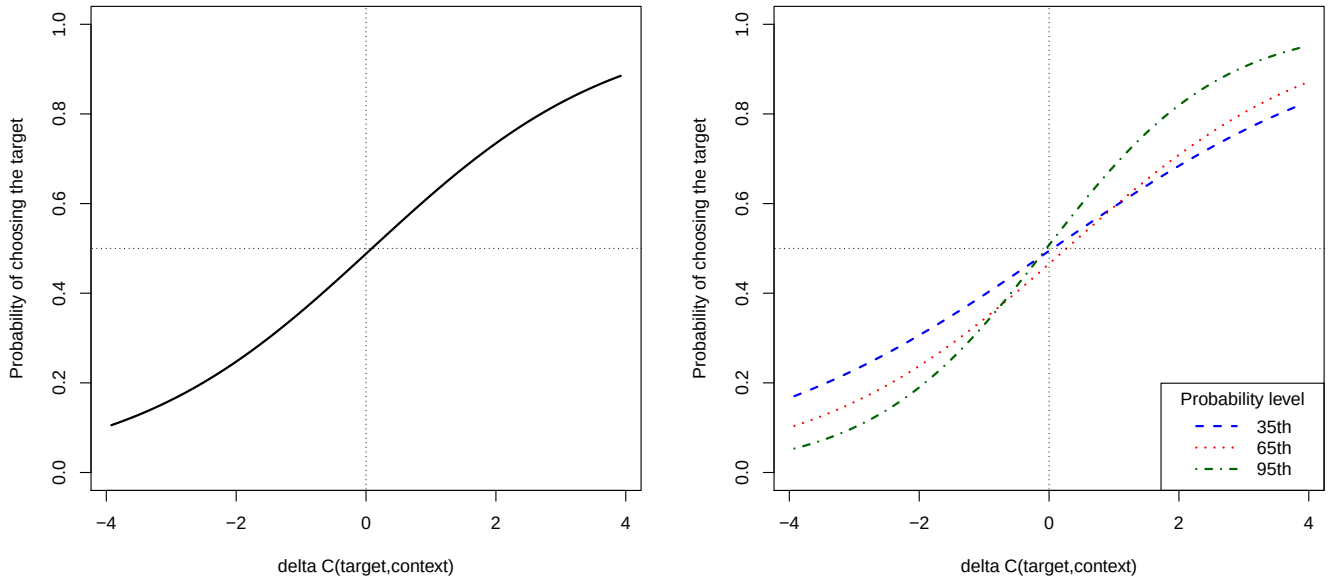

Figure 3: Graphical representation of the results of Experiment 1, as predicted by the statistical model: effects of $\Delta C(target,\ context)$ collapsed across (left panel) and differentiated by (right panel) probability levels.

Interaction terms with the probability levels were found to significantly improve model fit at $p = .1$ ($\chi^2(2) = 4.91$, $p = .0861$). In fact, when the two targets belong to the 95^{th} probability percentile, the effect is more pronounced as opposed to the grand mean.This is represented in the right panel of Figure 3: confirmation is always associated with the likelihood of choosing the target, but the slope is more pronounced at the 95^{th} percentile probability level. The left panel of Figure 3 reports the effect of $\Delta C(target,\ context)$ once the probability levels are collapsed. We obtained the same pattern of significant results when including low-confidence judgments. Employing random slopes, as proposed by Barr et al. (2013) for linear mixed models, did not lead to differences in the results of the significance tests, ensuring that the statistical model presented is not anticonservative because of its simplicity.

3.5. Discussion

Results of Experiment 1 support the initial hypothesis that confirmation of context on a target can modulate the perceived probability of the latter.

$target_1$	$target_2$	$C(target_1,\ context)$	$C(target_2,\ context)$	$P(target_1 \mid context)$	$P(target_2 \mid context)$
beer	translator	.005	1.127	.000535	.000499
beer	society	.005	-.102	.000535	.002568
beer	publishing	.005	1.196	.000535	.001462
beer	center	.005	-.066	.000535	.001231
translator	society	1.127	-.102	.000499	.002568
translator	publishing	1.127	1.196	.000499	.001462
translator	center	1.127	-.066	.000499	.001231
society	publishing	-.102	1.196	.002568	.001462
society	center	-.102	-.066	.002568	.001231
publishing	center	1.196	-.066	.001462	.001231

Table 3: Target pairings for the context *reporter* in Experiment 2.

We also observe a greater effect of confirmation at the highest probability level (95^{th} percentile). This interaction, although only weakly significant, might be tentatively explained as a familiarity effect: the more frequent a word, the more salient its confirmation relations.

4. Experiment 2

4.1. Aim

Experiment 2 is aimed at extending Experiment 1's investigation to a more general setting in which the two targets are not matched with respect to $P(target \mid context)$. Indeed, the effects of $C(target,\ context)$ reported in Experiment 1 could have been caused by the lack of any useful information to perform the task: it is possible that participants were exploiting confirmation only because probability values were identical. It is not obvious how much confirmation values would affect likelihood judgment in less restricted conditions with targets whose posterior probabilities are different. In order to address this concern, in Experiment 2 we employed a random sample of items from our corpus where probability and confirmation varied freely, and used both variables as predictors in a regression analysis on participants' choices.

4.2. Stimulus selection

We aimed at a sample covering most of the range of confirmation values and having a low correlation between confirmation and probability, in order to avoid collinearity issues that could have led to unreliable results in a regression approach. We found that the subset of all context-target pairs characterized by conditional probability $P(target \mid context)$ within the range between the 65^{th} and 95^{th} probability percentiles and confirmation $C(target,\ context)$ ranging from the 6^{th} to the 98^{th} percentile allowed us to fulfill the above mentioned requirements. Pearson correlation between these two independent variables in this subset is $r = .16$. We therefore randomly selected a set of stimuli consisting of 200 contexts, and for each of them, 5 targets satisfying the constraints on confirmation and probability levels.

Each of the 5 targets associated with the same context was combined with the other 4 targets, leading to 10 context-$target_1$-$target_2$ triples. Each triple constituted an experimental trial, with a total of 2,000 trials (200 contexts $\times$ 10 triples). Examples of trials for one particular context are reported in Table 3. Figure 4 reports the distribution of the two variables of interest in the experimental set.

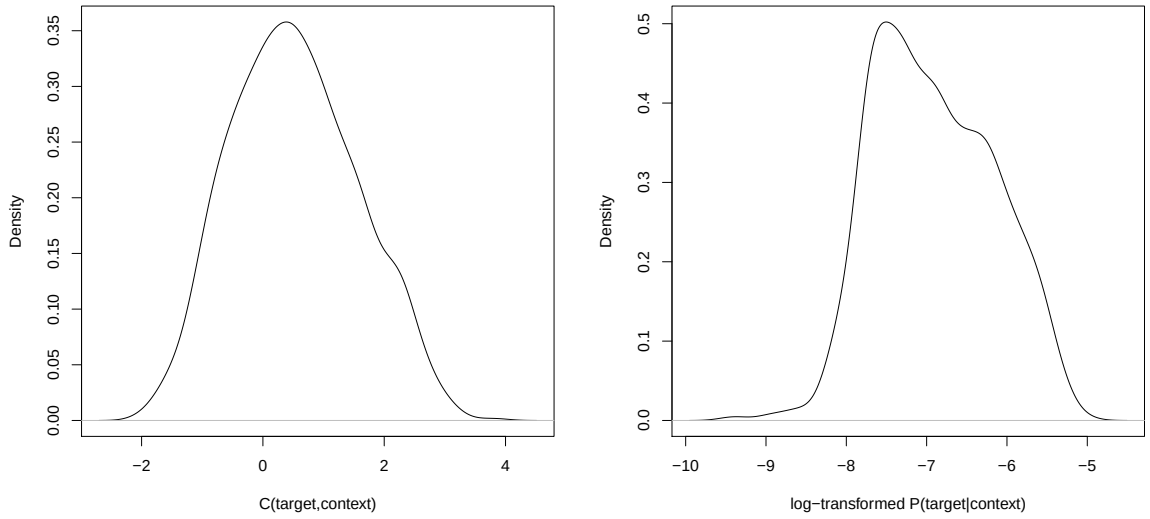

Figure 4: Distribution of the variables of interest in the dataset of Experiment 2.

Each trial was judged by 10 participants (for a total of 20,000 datapoints). A given participant could answer to no more than 120 trials. Target presentation order was counterbalanced in each pair (i.e., 5 judgments were collected for the target_1-target_2 order, and 5 judgments for target_2-target_1 order).

4.3. Data analysis

We followed the same approach as for the previous experiment. In this case, we also computed $\Delta P(\textit{target} \mid \textit{context})$ as the difference in (log-transformed)[3] $P(\textit{target} \mid \textit{context})$ between target_2 and target_1. Both $\Delta C(\textit{target}, \textit{context})$ and $\Delta P(\textit{target} \mid \textit{context})$ were included in the model as continuous predictors. Removed outliers constituted less than 0.01% of the total datapoints. We also excluded low-confidence (≤ 2) judgments (19% of the datapoints).

4.4. Results

A total of 257 distinct participants performed experiment 2, each answering to 78 trials on average. Inter-rater agreement between participants was 79.8%. Figure 5 represents the association between each word $C(\textit{target}, \textit{context})$ and its probability of being chosen by participants in the whole data set.

Table 4 reports the parameters included in the final model. Both $\Delta C(\textit{target}, \textit{context})$ and $\Delta P(\textit{target} \mid \textit{context})$ were found to significantly affect the likelihood of choosing target_2 over target_1. More specifically, both $\Delta C(\textit{target}, \textit{context})$ and $\Delta P(\textit{target} \mid \textit{context})$ increase the chances that target_2 is deemed most likely to occur in the same sentence as the context (see Figure 6). The interaction between the effects of the two variables ($\textit{estimate} = -.009$, $z = -0.49$, $p =$

[3]We applied a log-transformation in order to obtain a more Gaussian-shaped distribution of the probability predictor; however, the same results are obtained if we include raw probability values in the statistical model.

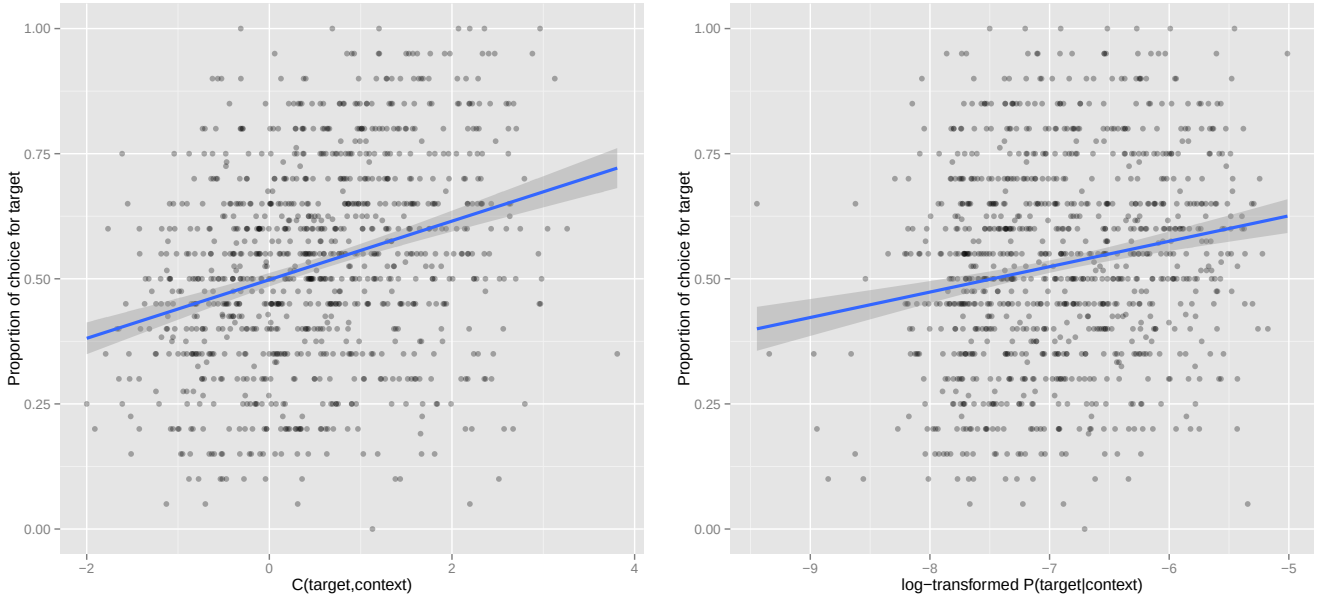

Figure 5: Association between $C(target,\ context)$ and probability of choosing the target (left panel), and between $P(target|\ context)$ and probability of choosing the target (right panel).

	Estimate	Std. Error	z-value	P
(Intercept)	.1022	.0755	1.35	.1761
$\Delta P(target \mid context)$	.4021	.0386	10.41	.0001
$\Delta C(target,\ context)$	.4117	.0291	14.18	.0001

Table 4: Fixed effects in the final model of Experiment 2.

.6181) was not included in the model, since it did not significantly contribute to the model fit ($\chi^2(1) = 0.23$, $p = .6291$). We obtained the same pattern of significant results when including low-confidence judgments. Again, including random slopes did not lead to differences in the results of the significance tests, ensuring that the statistical model presented is not anticonservative.

4.5. Discussion

The results of Experiment 2 are in line with and extend those of Experiment 1. Judgments of word co-occurrence are influenced by the relevant confirmation values even when the targets' difference in probability is in place to drive the choice.

5. Experiment 3

5.1. Aim

Experiment 3 is a follow-up, aimed at testing the effects described above under more controlled conditions, in order to rule out the possibility that the different distributions of the probability and confirmation predictors determined the effects reported in Experiment 2. We therefore adopted a 2×2 factorial design, explicitly dissociating (high vs. low) $C(target,\ context)$ and (high vs. low) $P(target \mid context)$ as predictors, and selected the stimuli so that the two variables are organized in strictly defined factors.

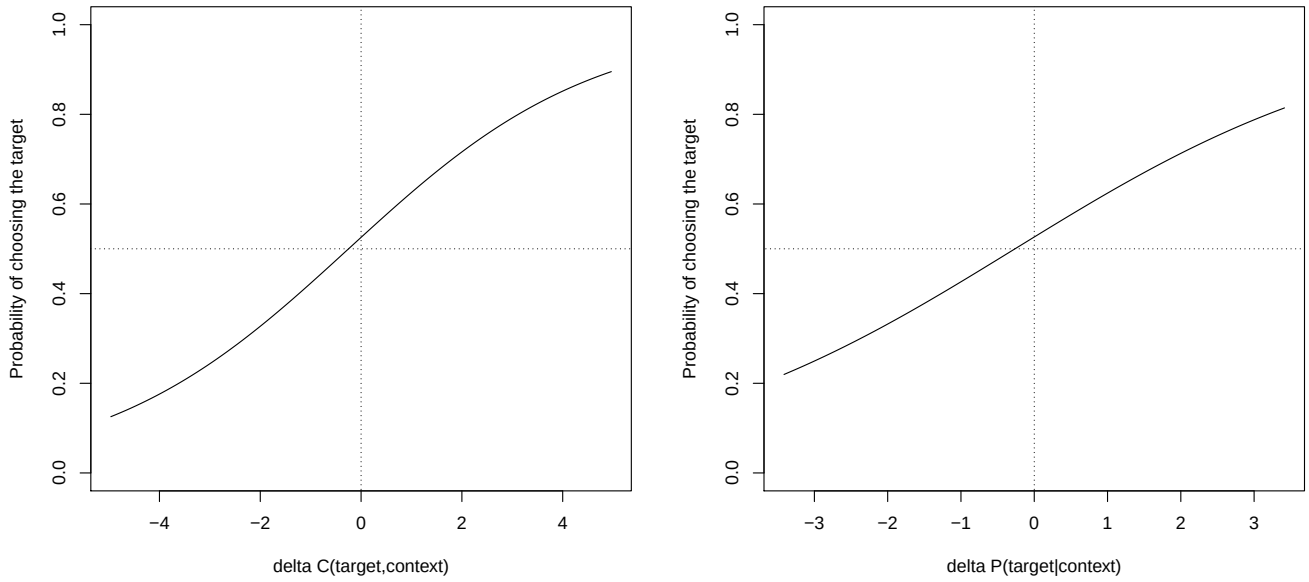

Figure 6: Main effects of $\Delta C(\textit{target},\ \textit{context})$ and $\Delta P(\textit{target} \mid \textit{context})$ in Experiment 2, as predicted by the statistical model.

target_1	target_2	$C(\textit{target}_1,\ \textit{context})$	$C(\textit{target}_2,\ \textit{context})$	$P(\textit{target}_1 \mid \textit{context})$	$P(\textit{target}_2 \mid \textit{context})$
archivist	witch	high	low	low	low
archivist	score	high	high	low	high
archivist	issue	high	low	low	high
witch	score	low	high	low	high
witch	issue	low	low	low	high
score	issue	high	low	high	high

Table 5: Target pairings for the context *songwriter* in Experiment 3.

5.2. Stimulus selection

The experimental set comprised 20 contexts, associated with a target for each of the following 4 conditions: high $C(\textit{target},\ \textit{context})$ and high $P(\textit{target} \mid \textit{context})$, high $C(\textit{target},\ \textit{context})$ and low $P(\textit{target} \mid \textit{context})$, low $C(\textit{target},\ \textit{context})$ and high $P(\textit{target} \mid \textit{context})$, low $C(\textit{target},\ \textit{context})$ and low $P(\textit{target} \mid \textit{context})$. We defined high and low levels of the variables of interest on a percentile basis: low levels were comprised between the 30^{th} and the 40^{th} percentile, high levels between the 80^{th} and the 90^{th} percentile. Each of the 4 targets associated with the same context was combined with the others 3 targets, leading to 6 context-target_1-target_2 tuples. Each tuple constituted an experimental trial, with a total of 120 trials (20 contexts $\times$ 6 tuples). Table 5 lists examples of targets for a given context.

Each target pair was judged by 20 participants (for a total of 2,400 datapoints). We did not impose a limit on the number of trials a participant could answer. Target presentation order was counterbalanced in each pair (i.e., 10 judgments were collected in the target_1-target_2 order, and 10 judgments in the target_2-target_1 order).

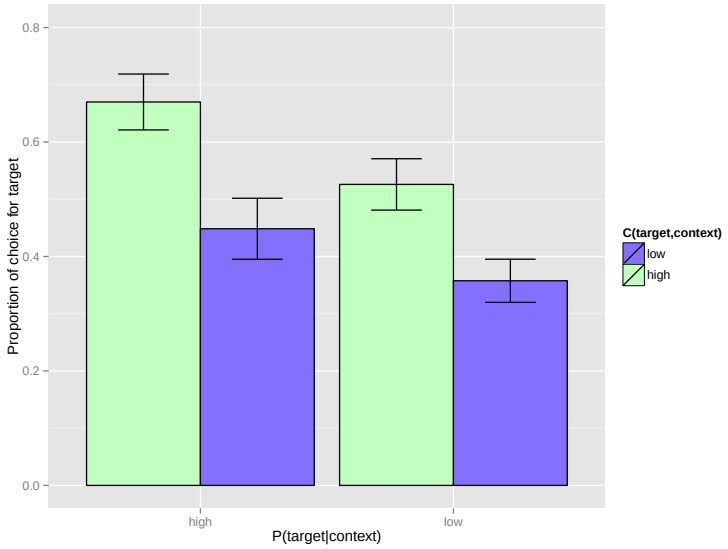

Figure 7: Proportion of a target being chosen given $C(target,\ context)$ and $P(target \mid context)$.

5.3. Data analysis

The same approach described for the previous experiments was followed in Experiment 3. However, in this case it was not possible to compute the predictor differences between target_2 and target_1, due to the factorial design adopted. Hence, both $P(target_2 \mid context)$ and $C(target_2,\ context)$ were included in the model as sum-coded predictors of the probability of choosing target_2 over target_1. The inclusion of random effects, along with the balanced design adopted (each target was presented with all the other targets), assures that the obtained results are net of the effect of contrasting targets_1. Removed outliers constituted less than 0.01% of the total datapoints. We also excluded low-confidence (≤ 2) judgments (26% of the datapoints).

5.4. Results

A total of 56 distinct participants performed experiment 3, each answering to an average of 43 trials. Inter-rater agreement between participants was 78.7%. Figure 7 reports the descriptive effects of $C(target,\ context)$ and $P(target \mid context)$ on participants' judgments when considering the whole data set: a target was more likely to be chosen when either variable was high.

Results of the mixed-effects analysis (Table 6) confirmed that both effects are significant; the interaction term ($estimate = 0.201$, $z = 1.253$, $p = .2103$), however, did not improve the goodness-of-fit of the model ($\chi^2(1) = 1.54$, $p = .2148$), suggesting that $C(target,\ context)$ and $P(target \mid context)$ have an additive effect on participants' choices. Again, including low-confidence judgments led to a consistent pattern of significant results and including random slopes did not lead to differences in the results of the significance tests.

5.5. Discussion

The results of Experiment 3 support, once more, the hypothesis that the intuitions about word occurrence depend not only on the probability values of interest, but are also influenced by confirmation relations. The effect emerges even when the variables are distributed across similar

	Estimate	Std. Error	z-value	P
(Intercept)	0.0305	.2651	0.12	.9083
$P(target \mid context)$	-0.6257	.1657	-3.77	.0002
$C(target,\ context)$	-0.3821	.1662	-2.29	.0215

Table 6: Fixed effects in the final model of Experiment 3.

ranges, indicating that the distribution mismatch between probability and confirmation cannot explain the results of Experiment 2.

6. Conclusion

This paper makes a novel contribution to two different research fields. We establish a relation between word association measures widely used in (corpus-based) linguistics and confirmation measures proposed by epistemologists and proved to be highly predictive of people's judgments in the psychology of reasoning. We believe that the results reported in this paper are only the first, most immediate empirical outcome of this connection.

In all three experiments, we found that confirmation consistently affects human judgment on word co-occurrence likelihood. Note that a similar effect has also been found in more traditional inductive probabilistic reasoning tasks (Lagnado & Shanks, 2002). Therefore, our findings suggest that a common probabilistic mechanism might be shared between language and other cognitive functions.

The notion of confirmation helps to explain the otherwise strange pattern that our experiments consistently show: for two words that have equal probabilities of occurrence in a given context, the less frequent one is perceived as more likely. Indeed, the less frequent word is in this case more confirmed by the context (has higher association with it).

Our results bring novel insights for both psychology and linguistics. Concerning the psychology of reasoning, we used a new kind of experimental stimuli to produce new data on human probability judgments. Under fully informed and rational reasoning, the posterior probability of a target word given a context word should entirely determine the response to our likelihood questions. However, as we saw in all three experiments, confirmation (measured as r/PMI) consistently affects likelihood responses as a distinct, independent predictor. That posterior probability does itself serve as a significant predictor of the response suggests that humans use an analog of Bayesian reasoning. However, human probabilistic reasoning is far from perfect, in a way that is predictable when taking into account the relevant confirmation values. It is noteworthy that confirmation, which is defined in terms of a combination of two or more probability values, is a factor that participates in human estimation of the very same probabilities we use to characterize it. In fact, the results of the experimental tests mentioned above (e.g., Tentori et al., 2007, 2014) suggest that confirmation judgments can be provided directly, without making the relevant probability values explicit. This gives us a reason to believe that the assessment of confirmation/association is a primitive kind of judgment not mediated by estimation of probability values. The fact that Bayesian models formally express confirmation as a function of some combination of probability values is only a historical coincidence, since the notion of probability was formalized much earlier than confirmation.

The effect of confirmation that we have found in this study is even more remarkable than those previously reported in the literature because in many of our stimuli the difference in confirmation between the two targets was not always salient from a psychological point of view (for example,

does the context *addition* confirm *signing* or *bean* more? Cf. Table 1 for the answer). The relevance of values of confirmation which might be too subtle for subjects to notice is a totally new empirical finding, and makes a concrete contribution to the literature on psychology of reasoning. It is also noteworthy that differences in tiny conditional probabilities, which in our study are all under .006, have a significant impact on people's performance.

Our study illustrates the usefulness of linguistic corpora as a source of probability and confirmation values that are based on a sample of text which is considered representative of people's linguistic experience. Probabilistic knowledge implicit in language data, as reflected in corpus statistics, can be exploited to study other reasoning-related tasks in more naturalistic settings. The study of probabilistic reasoning often relies on constructed scenarios where subjects may be provided with either explicit probability values, or implicit values acquired in an artificial setting. In contrast, reasoning on word probabilities relies on participants' previously acquired statistical knowledge, largely shared across speakers, which makes the experimental materials considerably less artificial.

Our work also contributes to the study of word association in linguistics. First, it is an intriguing finding that similar measures subtend both explicit judgments and the unconscious processes that lead to the observed effect of word association on many linguistic patterns. We therefore provide new independent behavioral evidence for the psychological validity of association measures that are commonly used to model phenomena such as collocation and semantic similarity. Second, we show that word association is not just a language-specific factor, but an instance of the more general notion of inductive confirmation. This further supports the plausibility of linguistic models based on word association by providing them with fundamental theoretical grounding.

Last but not least, our work provides further evidence that speakers are sensitive to very subtle statistical patterns present in corpora, validating corpora as a surrogate for linguistic experience of an individual. Indeed, it was not obvious from the start at all that corpus co-occurrence would affect, via either probability or confirmation, the participants' intuition about the likelihood of word co-occurrence. Linguistic corpora can be huge, one might argue, but they aren't necessarily representative of individual linguistic experience. Our study suggests that corpus statistics are indeed a reliable basis for predicting human performance. Note that in our experiments the corpus-derived confirmation (and probability) differences were relatively small (compare examples in Table 3), but the participants nonetheless proved sensitive to them.

In this paper, we explored intuitions on the probability of co-occurrence of single words within sentence boundaries; corpus statistics was collected accordingly. But the task doesn't necessarily have to be so simple. First, one could try going beyond sentence-based statistics, e.g., considering two words' co-occurrence within a certain distance from each other; alternatively, document-based co-occurrences could be considered, consistently with popular distributional models in psycholinguistics (Griffiths et al., 2007). Second, and perhaps more interestingly, one could consider probabilities not only of single words, such as *publishing*, but also phrases like *news publishing*. Contrasting a word with a phrase containing it could allow one to replicate, in a new domain, the conjunction fallacy (see p. 4) in which people judge the conjunction of two hypotheses to be more likely than one of them taken separately. For example, some respondents may perceive *news publishing* (occurrence of *publishing* and *news* as a modifier) to be more likely than *publishing* itself in the context of *reporter*. Third, one could test whether confirmation also plays a role in other experimental tasks in which conditional probabilities have proven to be useful predictors, such as semantic priming (Griffiths et al., 2007) or the various tasks for which surprisal (negative log of probability) has

been shown to be predictive, including reading times (Roark et al., 2009; Monsalve et al., 2012), garden-path effects (Hale, 2001), and ERP during reading (Frank et al., 2013). We leave exploring those possibilities to future research.

Author Note

Denis and Marco M. share first autorship, Katya and Marco B. share senior authorship. We thank Alessandro Lenci and Vincenzo Crupi for their feedback. We are also grateful to Klinton Bicknell, two anonymous reviewers, and the editor for their constructive comments. The third author acknowledges support by the MIUR grant PRIN 2010RP5RNM_006. The other authors acknowledge support by the ERC 2011 Starting Independent Research Grant n. 283554 (COMPOSES).

References

Alderson, C. (2007). Judging the frequency of English words. *Applied Linguistics*, *28*, 383–409.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*, 390–412. doi:10.1016/j.jml.2007.12.005.

Balota, D., Pilotti, M., & Cortese, M. (2001). Subjective frequency estimates for 2,938 monosyllabic words. *Memory and Cognition*, *29*, 639–47.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*, 255–78.

Bod, R., Hay, J., & Jannedy, S. (Eds.) (2003). *Probabilistic Linguistics*. Cambridge, MA: MIT Press.

Bullinaria, J., & Levy, J. (2007). Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior Research Methods*, *39*, 510–26.

Bullinaria, J., & Levy, J. (2012). Extracting semantic representations from word co-occurrence statistics: Stop-lists, stemming and SVD. *Behavior Research Methods*, *44*, 890–907.

Bybee, J. (2007). *Frequency of Use and the Organization of Language*. Oxford, UK: Oxford University Press.

Church, K., Gale, W. A., Hanks, P., & Hindle, D. (1991). Using statistics in lexical analysis. In *Lexical Acquisition: Using On-line Resources to Build a Lexicon* (pp. 115–64). Lawrence Erlbaum.

Church, K., & Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics*, *16*, 22–9.

Costello, F. J. (2009). How probability theory explains the conjunction fallacy. *Journal of Behavioral Decision Making*, (pp. 213–34). doi:10.1002/bdm.618.

Crupi, V., Fitelson, B., & Tentori, K. (2008). Probability, confirmation, and the conjunction fallacy. *Thinking & Reasoning*, *14*, 182–99.

Crupi, V., Tentori, K., & Gonzalez, M. (2007). On Bayesian measures of evidential support: Theoretical and empirical issues. *Philosophy of Science*, *74*, 229–52.

Dennis, S. F. (1965). The construction of a thesaurus automatically from a sample of text. In *Proceedings of the Symposium on Statistical Association Methods For Mechanized Documentation* (pp. 61–148). Washington, DC volume 269 of *National Bureau of Standards Miscellaneous Publication*.

Durrant, P. (2008). *High-frequency Collocations and Second Language Learning*. Dissertation University of Nottingham.

Eells, E. (1982). *Rational decision and causality*. Cambridge: Cambridge University Press.

Ellis, N., & Simpson-Vlach, R. (2009). Triangulating psycholinguistics, corpus linguistics, and education. *Corpus Linguistics and Linguistic Theory*, *5*, 61–78.

Evert, S. (2005). *The Statistics of Word Cooccurrences*. Ph.D dissertation Stuttgart University.

Evert, S. (2008). Corpora and collocations. In A. Lüdeling, & M. Kytö (Eds.), *Corpus linguistics: An international handbook* (pp. 1212–48). Berlin, Germany: Mouton de Gruyter.

Fantino, E., Kulik, J., Stolarz-Fantino, S., & Wright, W. (1997). The conjunction fallacy: A test of averaging hypotheses. *Psychonomic Bulletin & Review*, (pp. 96–101).

Fitelson, B. (1999). The plurality of Bayesian measures of confirmation and the problem of measure sensitivity. *Philosophy of Science*, (pp. 362–378).

Frank, S. L., Otten, L. J., Galli, G., & Vigliocco, G. (2013). Word surprisal predicts n400 amplitude during reading. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (pp. 878–83). Sofia, Bulgaria: Association for Computational Linguistics. URL: http://www.aclweb.org/anthology/P13-2152.

Good, I. J. (1984). The best explicatum for weight of evidence. *Journal of Statistical Computation and Simulation*, *19*, 294–9.

Griffiths, T., Steyvers, M., & Tenenbaum, J. (2007). Topics in semantic representation. *Psychological Review*, *114*, 211–44.

Hale, J. (2001). A probabilistic earley parser as a psycholinguistic model. In *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics on Language Technologies* NAACL '01 (pp. 1–8). Stroudsburg, PA, USA: Association for Computational Linguistics. URL: http://dx.doi.org/10.3115/1073336.1073357. doi:10.3115/1073336.1073357.

Jaeger, T. F. (2008). Categorical Data Analysis: Away from ANOVAs (transformation or not) and towards Logit Mixed Models. *Journal of Memory and Language*, *59*, 434–46. doi:10.1016/j.jml.2007.11.007.

Keynes, J. (1921). *A treatise on probability*. London: Macmillan.

Lagnado, D., & Shanks, D. (2002). Probability judgments in hierarchical learning: a conflict between predictiveness and coherence. *Cognition*, (pp. 81–112).

Lüdeling, A., & Kytö, M. (Eds.) (2008). *Corpus Linguistics: An International Handbook*. Berlin, Germany: Mouton de Gruyter.

Manning, C., & Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press.

Mastropasqua, T., Crupi, V., & Tentori, K. (2010). Broadening the study of inductive reasoning: Confirmation judgments with uncertain evidence. *Memory and Cognition*, *38*, 941–50. doi:10.3758/MC.38.7.941.

McDonald, S., & Ramscar, M. (2001). Testing the distributional hypothesis: The influence of context on judgements of semantic similarity. In *Proceedings of CogSci* (pp. 611–6).

Monsalve, I. F., Frank, S. L., & Vigliocco, G. (2012). Lexical surprisal as a general predictor of reading time. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics* EACL '12 (pp. 398–408). Stroudsburg, PA, USA: Association for Computational Linguistics. URL: http://dl.acm.org/citation.cfm?id=2380816.2380866.

Munro, R., Bethard, S., Kuperman, V., Lai, V. T., Melnick, R., Potts, C., Schnoebelen, T., & Tily, H. (2010). Crowdsourcing and language studies: the new generation of linguistic data. In *Proceedings of the NAACL HLT 2010* CSLDAMT '10 (pp. 122–30). Stroudsburg, PA, USA: Association for Computational Linguistics.

Pantel, P., & Lin, D. (2002). Discovering word senses from text. In *Proceedings of KDD* (pp. 613–9). Edmonton, Canada.

Paolacci, G., Chandler, J., & Panagiotis, G. I. (2010). Running experiments on amazon mechanical turk. *Judgment and Decision Making*, *5*, 411–9.

Pitler, E., Bergsma, S., Lin, D., & Church, K. (2010). Using web-scale n-grams to improve base NP parsing performance. In *Proceedings of COLING* (pp. 886–94). Beijing, China.

Recchia, G., & Jones, M. (2009). More data trumps smarter algorithms: Comparing pointwise mutual information with latent semantic analysis. *Behavior Research Methods*, *41*, 647–56.

Roark, B., Bachrach, A., Cardenas, C., & Pallier, C. (2009). Deriving lexical and syntactic expectation-based measures for psycholinguistic modeling via incremental top-down parsing. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1* EMNLP '09 (pp. 324–33). Stroudsburg, PA, USA: Association for Computational Linguistics. URL: http://dl.acm.org/citation.cfm?id=1699510.1699553.

Sag, I., Baldwin, T., Bond, F., Copestake, A., & Flickinger, D. (2002). Multiword expressions: A pain in the neck for NLP. In A. Gelbukh (Ed.), *Computational Linguistics and Intelligent Text Processing* (pp. 189–206). Berlin: Springer.

Schmid, H. (1995). Improvements in part-of-speech tagging with an application to German. In *Proceedings of the EACL-SIGDAT Workshop*. Dublin, Ireland.

Smadja, F., McKeown, K. R., & Hatzivassiloglou, V. (1996). Translating collocations for bilingual lexicons: A statistical approach. *Computational Linguistics*, *22*, 1–38.

Tentori, K., Chater, N., & Crupi, V. (2014). Judging the probability of a hypothesis versus the impact of evidenct: Which form of inductive inference is more reliable.

Tentori, K., Crupi, V., Bonini, N., & Osherson, D. (2007). Comparison of confirmation measures. *Cognition*, *103*, 107–19.

Tentori, K., Crupi, V., & Russo, S. (2013). On the determinants of the conjunction fallacy: Confirmation versus probability. *Journal of Experimental Psychology: General*, *142*, 235–55.

Tversky, A., & Kahneman, D. (1982). Judgments of and by representativeness. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 84–98). New York, NY: Cambridge University Press.

Tversky, A., & Kahneman, D. (1983). Extension versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, *90*, 293–315.

Yates, J. F., & Carlson, B. W. (1986). Conjunction errors: Evidence for multiple judgment procedures, including "signed summation". *Organizational Behavior and Human Decision Processes*, (pp. 230–53). doi:10.1016/0749-5978(86)90053-1.

Appendix A

Confirmation, $c(h, e)$, is a value that quantifies the epistemic impact of evidence e on the credibility of hypothesis h, such that $c(h, e) > 0$ iff $P(h \mid e) > P(h)$, $c(h, e) < 0$ iff $P(h \mid e) < P(h)$, and $c(h, e) = 0$ iff $P(h \mid e) = P(h)$.

A plurality of alternative formal models ("confirmation measures") have been proposed to quantify the degrees of confirmation. In all of the proposed models, $c(h, e)$ is formalized as a function of probabilities involving e and h, such as $P(h), P(h \mid e), P(e \mid h)$, etc. Some examples of confirmation models include:

$$r(h, e) = \ln \frac{P(h|e)}{P(h)} \qquad \text{(Keynes, 1921);}$$

$$l(h, e) = \ln \frac{P(e|h)}{P(e|\neg h)} \qquad \text{(Good, 1984);}$$

$$d(h, e) = P(h \mid e) - P(h) \qquad \text{(Eells, 1982);}$$

$$z(h, e) = \begin{cases} \frac{P(h|e)-P(h)}{P(\neg h)} & \text{iff } P(h \mid e) \geq P(h) \\ \frac{P(h|e)-P(h)}{P(h)} & \text{iff } P(h \mid e) < P(h) \end{cases} \qquad \text{(Crupi et al., 2007).}$$

Appendix B

Association measures are smooth functions from the frequency of co-occurrence of two words $f(w_1, w_2)$, their individual frequencies $f(w_1), f(w_2)$, and the size of the corpus sample N, upward monotone on co-occurrence frequency $f(w_1, w_2)$ and downward monotone on individual word frequencies $f(w_1), f(w_2)$ (Evert, 2005). Here are some examples of association measures:

$$\text{Dice} = \frac{2f(w_1,w_2)}{f(w_1)+f(w_2)} \qquad \text{(Smadja et al., 1996)}$$

$$\text{PMI} = \log_2 \frac{f(w_1,w_2)N}{f(w_1)f(w_2)}$$

(Church & Hanks, 1990)

$$\text{t\_score} = \frac{f(w_1,w_2)-E(w_1,w2))}{\sqrt{f(w_1,w_2)}} = \frac{f(w_1,w_2)-\frac{f(w_1)f(w2)}{N}}{\sqrt{f(w_1,w_2)}} \qquad \text{(Church et al., 1991)}$$

$$\text{z\_score} = \frac{f(w_1,w_2)-E(w_1,w2)}{\sqrt{E(w_1,w2)}} = \frac{f(w_1,w_2)-\frac{f(w_1)f(w2)}{N}}{\sqrt{\frac{f(w_1)f(w2)}{N}}} \qquad \text{(Dennis, 1965)}$$

where $E(w_1, w_2) = P(w_1)P(w_2) * N$ is the expected frequency of co-occurrence of w_1, w_2 under independence.

Many association measures including PMI, z-score, and t-score, satisfy the definition of a confirmation measure: they equal zero in case of independent occurrence of w_1, w_2, positive if w_1, w_2 confirm each other and negative otherwise. Notice that, once the corpus size N is fixed, such association measures can be interpreted as confirmation measures depending on prior and posterior probabilities of words, since $f(w_1) = P(w_1) \times N$ and $f(w_2) = P(w_2) \times N$, and finally,

$$f(w_1, w_2) = P(w_1, w_2) \times N = P(w_1 \mid w_2) \times P(w_2) \times N$$