

Local associations and semantic ties in overt and masked semantic priming

Andrea Nadalini
International School
for Advanced Studies
Trieste, Italy
anadalini
@sissa.it

Marco Marelli
Bicocca University
Milan, Italy
marco.marelli
@unimib.it

Roberto Bottini
Center for mind/brain
sciences
Trento, Italy
roberto.bottini@
unitn.it

Davide Crepaldi
International School
for Advanced Studies
Trieste, Italy
dcrepaldi
@sissa.it

Abstract

English. Distributional semantic models (DSM) are widely used in psycholinguistic research to automatically assess the degree of semantic relatedness between words. Model estimates strongly correlate with human similarity judgements and offer a tool to successfully predict a wide range of language-related phenomena. In the present study, we compare the state-of-art model with pointwise mutual information (PMI), a measure of local association between words based on their surface cooccurrence. In particular, we test how the two indexes perform on a dataset of semantic priming data, showing how PMI outperforms DSM in the fit to the behavioral data. According to our result, what has been traditionally thought of as semantic effects may mostly rely on local associations based on word co-occurrence.

Italiano. *I modelli semantici distribuzionali sono ampiamente utilizzati in psicolinguistica per quantificare il grado di similarità tra parole. Tali stime sono in linea con i corrispettivi giudizi umani, e offrono uno strumento per modellare un'ampia gamma di fenomeni relativi al linguaggio. Nel presente studio, confrontiamo il modello con la pointwise mutual information (PMI), una misura di associazione locale tra parole basata sulla loro cooccorrenza. In particolare, abbiamo testato i due indici su un set di dati di priming semantico, mostrando come la PMI riesca a spiegare meglio i dati com-*

portamentali. Alla luce di tali risultati, ciò che è stato tradizionalmente considerato come effetto semantico potrebbe basarsi principalmente su associazioni locali di co-occorrenza lessicale.

1 Introduction

Over the past two decades, computational semantics has made a lot of progress in the strive for developing techniques that are able to provide human-like estimates of the semantic relatedness between lexical items. Distributional Semantic Models (DSM; Baroni and Lenci, 2010) assume that it is possible to represent lexical meaning based on statistical analyses of the way words are used in large text corpora. Words are modeled as vectors and populate a high-dimensional space where similar words tend to cluster together. Meaning relatedness between two words corresponds to the proximity of their vectors; for example, one can approximate relatedness as the cosine of the angle formed by two word-vectors:

$$\cos\theta = \frac{\vec{a} \cdot \vec{b}}{\|\vec{a}\| \cdot \|\vec{b}\|}$$

DSMs have been proposed as a psychologically plausible models of semantic memory, with particular emphasis on how meaning representations are achieved and structured (e.g. LSA, Landauer and Dumais, 1997; HAL, Lund and Burgess, 1996). So, they can be pitted against human behavior, in search for psychological validation of this modeling. For example, the model's estimates have been used to make reliable predictions about the processing time associated with the stimuli (Baroni et al., 2014; Mandera et al., 2017).

The technique most commonly used to explore semantic processing is the priming paradigm (McNamara, 2005), according to which the recognition of a given word (the *target*) is easier if preceded by a related word (the *prime*; e.g., cat–dog). Interestingly, facilitation can be observed both when the prime word is fully visible and when it is kept outside of participants’ awareness through visual masking (Forster and Davis, 1984; de Wit and Kinoshita, 2015). In this technique, the prime stimulus is displayed shortly, embedded between a forward and a backward string (Figure 1).

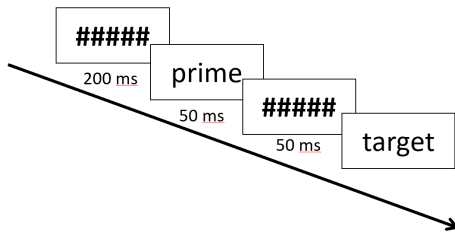


Figure 1: exemplar trial in a masked priming experiment. The prime stimulus is briefly presented (≤ 50 ms), between the two masks, before the onset of the target stimulus.

Beside words’ distribution, one can be interested in the local association strength between lexical items, starting from the assumption that two words that are often used close to each other, tend to become associated. Yet, a given pair may be often attested only because the two components are in turn highly frequent. Therefore, raw frequency counts are often transformed into some kinds of association measure which can determine if the pair is attested above chance (Evert, 2008). A common method is to compute pointwise mutual information (PMI) between two words, according to the formula:

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

where $p(w_1, w_2)$ corresponds to the probability of the word pair, while $p(w_1)$ and $p(w_2)$ to the individual probabilities of the two components (Church and Hanks, 1990).

PMI has been used to model a wide range of psycholinguistics phenomena, from similarity judgements (Recchia and Jones, 2009) to reading speed (Ellis and Simpson-Vlach, 2009). Moreover, PMI has also been shown to successfully generalize to non-linguistic fields as epistemology and psychology of reasoning (Tentori et al., 2014). On the other hand, PMI has the limit of

over-estimating the importance of rare items (Manning and Schütze, 1999).

Despite many DSMs use measures of local association between words like PMI to build contingency matrices, the information conveyed by two similar word-vectors is different from the information conveyed by two highly recurrent words. Cosine similarity is based on “higher order” co-occurrences: two words are similar in the way they are used together with all the other words in the vocabulary. Local measures as PMI instead rely only on the effective co-presence of two given words. Two synonyms like the words *car* and *automobile* are not likely to often appear close to each other in a given text, still they represent the same referent, and therefore expected to be used in similar contexts.

Based on these considerations, PMI and DSMs can be pitted against human behavior, in search for psychological validation of this modeling. In particular, we tested how PMI and cosine proximity predicts priming in a set of data encompassing different prime visibility conditions (masked vs unmasked) and prime durations (33, 50, 200, 1200 ms).

2 Our Study

2.1 Material

All the stimuli used in the current study were Italian words. 50 words referring to animals and 50 words referring to tools were used as target stimuli. Each word in this list was paired with three words from the same category, resulting in 300 unique prime-target couples which were divided into three rotations. We add to each rotation 100 additional filler trials which will not be included in the analysis step. More precisely, we used abstract word as target stimuli, paired with animals and tool primes different from those presented in the experimental trials. In this way we ensured that the response to the target was not predictable by the presence of the prime.

Relatedness estimates were obtained by looking at the stimuli distribution across the ItWac corpus, a linguistic database of nearly 2 billion words built through web crawling (Baroni et al., 2009). We downloaded the lemmatized and part-of-speech annotated corpus, freely provided by the authors. All characters were set to lowercase, and special characters were removed together with a list of stop-words.

PMI between the word pairs was computed based on frequency counts gained by sliding a 5-words window along ItWac. Cosine proximity between word vectors was obtained training a word2vec model (Mikolov et al., 2013) on the same corpus. Model’s parameters were set according to the WEISS model (Marelli, 2017). All words attested at least 100 times were included in the model, which was trained using the continuous-bag-of-words architecture, a 5-word window and 200 dimensions. The parameter k for negative sampling was set to 10, and the sub-sampling parameter to 10^{-5} .

Correlations between semantic and lexical variables are shown in Table 1.

	Target length	Target frequency	PMI	cosine
Target length	1			
Target frequency	-.211	1		
PMI	.091	-.205	1	
cosine	.147	-.059	.541	1

Table 1: Correlations between lexical and semantic indexes in our stimulus set.

2.2 Methods

Participants: Overall, 246 volunteers were recruited for the current study, and were assigned to the different prime timing conditions. All subjects were native Italian speakers, with normal or corrected-to-normal vision and no history of neurological or learning diseases.

Apparatus: All stimuli were displayed on a 25’’ monitor with a refresh rate of 120 Hz, using MatLab Psychtoolbox. The words and the masks were presented in Arial font 32, in white color against a black background.

Procedure: Participants were engaged in a classic YES/NO task, requiring them to classify the stimuli as members of either the animal or the tool category, according to the instructions. YES-response were always provided with the dominant hand.

Each unique prime-target pair was presented only once to each participant. Experimental sessions included a total of 200 trials, which were divided into two blocks. In one block, subjects were asked to press the yes-button if the target word referred to an animal, while in the other

block they were asked to press the yes-button if the target word referred to a tool. The order of the two blocks was counterbalanced across subjects. 10 practice and 2 warm-up trials were presented before each block. Participants could take a short break halfway through each block.

Each trial began with a 750 ms fixation-cross (+). Prime duration was varied across experiments: 33, 50, 200 and 1200 ms respectively. In the former two conditions, prime visibility was prevented through forward and backward visual masks. Finally, the target word was left on the screen until a response was provided.

Prime visibility task. In the experiments with the masked primes, participants were not informed about their presence. This was only revealed after the relevant session, when participants were invited to take part into a prime visibility task requiring them to spot the presence of the letter “n” within the masked word. After the first two examples, where prime duration was increased to 150 ms to ensure visibility, 10 practice and 80 experimental trials were displayed. Prime visibility was quantified through a d-prime analysis carried out on each participant (Green and Swets, 1966).

2.3 Results

Response times (RT) were analyzed on accurate, yes-response trials only. RT were inverse transformed to approximate a normal distribution and employed as a dependent variable in linear mixed-effects regression models. This analysis allows us to control for all the covariates that may have affected the performance, such as trial position in the randomized list, rotation, RT and accuracy on the preceding trial, the response required in the preceding trial, frequency and length of the target. All these variables, together with the two semantic indexes (*PMI* and *cosine proximity*), were entered in the model as fixed effects, while participants and items were considered as random intercepts. Model selection was implemented stepwise, progressively removing those variables whose contribution to goodness of fit was not significant.

In the masked priming data, neither PMI nor cosine proximity were reliable predictors by themselves ($p=.298$ and $p=.206$, respectively). However, both indexes interacted with prime visibility as tracked by participants’ d-prime ($F_{pmi*d_t}(1, 9750)= 13.74, p<.001$; $F_{cos*d_t}(1,$

9745)= 13.24, $p < .001$). As illustrated in Figure 1, the more each participant could see the prime word, the higher the priming effect she displayed.

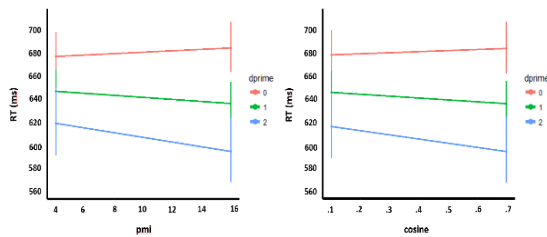


Figure 1. Interaction between d' and prime–target association. Both PMI (left) and cosine proximity (right) effects become stronger as prime visibility (d') increases. Error bars refer to 95% C.I.

In the overt priming data, both PMI and cosine proximity yield a significant main effect (50ms presentation time: $F_{pmi}(1,9769) = 10.36$, $p = .001$; $F_{cos}(1, 9769) = 8.602$, $p = .0058$), but only PMI significantly predicts priming when both indexes are entered into the model ($F_{pmi}(1,9769) = 10.36$, $p = .001$; $F_{cos}(1,9769) = 0.60$, $p = .489$). Results were very consistent across conditions and showed the same pattern when prime presentation time was 200ms or 1200ms (see Figure 2).

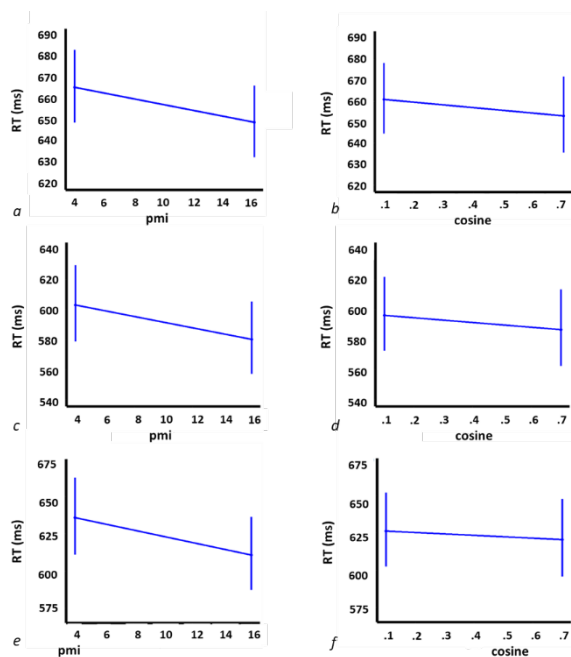


Figure 2. Significant effect of PMI (left) and non-significant effect of cosine proximity (right) across prime presentation times (50ms, 200ms, 1200ms on the first, second and third row respectively). Error bars refer to 95% C.I.

Conclusion

Thanks to the help of computational methods, we provided new insights on the nature of the processing that supports semantic priming. Overall, effects seem to be primarily driven by local word associations as tracked by Pointwise Mutual Information—when semantic priming emerged, PMI effects were consistently stronger and more solid than those related to DSM estimates. This would be in line with previous literature suggesting that the behavior of the human cognitive system may be effectively described by Information Theory principles. For example, Paperno and colleagues (Paperno et al., 2014) showed that PMI is a significant predictor of human judgements of word co-occurrence.

The results from masked priming offer another important insight—some kind of prime visibility may be required for semantic/associative priming to emerge. Other studies have shown genuine semantic effects with subliminally presented stimuli (Bottini et al., 2016). However, they typically used words from small/closed classes (e.g., spatial words, planet names). Conversely, we drew stimuli across the lexicon, and sampled from very large category such as animals and tools; this may point to an effect of target predictability. In general, our data cast some doubts on a wide-across-the-lexicon processing of semantic information outside of awareness.

References

- Baroni M., S. Bernardini, A. Ferraresi and E. Zanchetta. (2009). The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora. *Language Resources and Evaluation* 43 (3): 209-226.
- Baroni, M., Dinu, G., and Kruszewski, G. (2014). Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (Vol. 1, pp. 238-247)
- Bottini, R., Bucur, M., and Crepaldi, D. (2016). The nature of semantic priming by subliminal spatial words: Embodied or disembodied?. *Journal of Experimental Psychology: General*, 145(9), 1160.
- de Wit, B., and Kinoshita, S. (2015). The masked semantic priming effect is task dependent: Reconsidering the automatic spreading activation process. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 41(4), 1062.
- Ellis, N. C., and Simpson-Vlach, R. (2009). Formulaic language in native speakers: Triangulating psy-

- cholingistics, corpus linguistics, and education. *Corpus Linguistics and Linguistic Theory*, 5(1), 61-78.
- Evert, S. (2008). Corpora and collocations. *Corpus linguistics. An international handbook*, 2, 223-233.
- Forster, K. I., and Davis, C. (1984). Repetition priming and frequency attenuation in lexical access. *Journal of experimental psychology: Learning, Memory, and Cognition*, 10(4), 680.
- Green D.M. and Swets J.A. (1966). *Signal detection theory and psychophysics*. Wiley New York.
- Landauer, T. K., and Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, 104(2), 211.
- Lund, K., and Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior research methods, instruments, and computers*, 28(2), 203-208.
- Mandera, P., Keuleers, E., and Brysbaert, M. (2017). Explaining human performance in psycholinguistic tasks with models of semantic similarity based on prediction and counting: A review and empirical validation. *Journal of Memory and Language*, 92, 57-78.
- Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing*. MIT press.
- Marelli, M. (2017). Word-Embeddings Italian Semantic Spaces: A semantic model for psycholinguistic research. *Psihologija*, 50(4), 503-520.
- McNamara, T. P. (2005). *Semantic priming: Perspectives from memory and word recognition*. Psychology Press.
- Mikolov, P., Chen, K., Corrado, G. S. Dean, J. (2013). Efficient estimation of word representations in vector space. Available from ArXiv:1301.3781.
- Paperno, D., Marelli, M., Tentori, K., and Baroni, M. (2014). Corpus-based estimates of word association predict biases in judgment of word co-occurrence likelihood. *Cognitive psychology*, 74, 66-83.
- Recchia, G., and Jones, M. N. (2009). More data trumps smarter algorithms: Comparing pointwise mutual information with latent semantic analysis. *Behavior research methods*, 41(3), 647-656.
- Rohaut, B. and Naccache, L. (2018), What are the boundaries of unconscious semantic cognition? *Eur J Neurosci*. . doi:10.1111/ejn.13930.
- Tentori, K., Chater, N., and Crupi, V. (2016). Judging the Probability of Hypotheses Versus the Impact of Evidence: Which Form of Inductive Inference Is