Running head: Computational models of semantic similarity

Explaining human performance in psycholinguistic tasks with models of semantic similarity based on prediction and counting: A review and empirical validation

Paweł Mandera*, Emmanuel Keuleers, and Marc Brysbaert

Department of Experimental Psychology, Ghent University, Belgium

Keywords: semantic models; distributional semantics; semantic priming; psycholinguistic resources

*Corresponding Author

Paweł Mandera

Ghent University, Department of Experimental Psychology

Henri Dunantlaan 2, room 150.025

9000 Ghent, Belgium

E-mail: pawel.mandera@ugent.be

Tel: +32 9 264 94 3

Abstract

Recent developments in distributional semantics (Mikolov et al., 2013) include a new class of prediction-based models that are trained on a text corpus and that measure semantic similarity between words. We discuss the relevance of these models for psycholinguistic theories and compare them to more traditional distributional semantic models. We compare the models' performances on a large dataset of semantic priming (Hutchison et al., 2013) and on a number of other tasks involving semantic processing and conclude that the prediction-based models usually offer a better fit to behavioral data. Theoretically, we argue that these models bridge the gap between traditional approaches to distributional semantics and psychologically plausible learning principles. As an aid to researchers, we release semantic vectors for English and Dutch for a range of models together with a convenient interface that can be used to extract a great number of semantic similarity measures.

Introduction

Distributional semantics is based on the idea that words with similar meanings are used in similar contexts (Harris, 1954). In this line of thinking, semantic relatedness can be measured by looking at the similarity between word co-occurrence patterns in text corpora. In psychology, this idea inspired a fruitful line of research starting with Lund and Burgess (1996) and Landauer and Dumais (1997). The goal of the present paper is to incorporate a new family of models recently introduced in computational linguistics and natural language processing research by Mikolov and colleagues (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013; Mikolov, Chen, Corrado, Dean, 2013) into psycholinguistics. In order to do so, we will discuss the theoretical foundation of these models and evaluate their performance on predicting behavioral data on psychologically relevant tasks.

Count and Predict Models

Although there are different approaches to distributional semantics, what they have in common is that they start from a text corpus and that they often represent words as numerical vectors in a multidimensional space. The relatedness between a pair of words is quantified by measuring the similarity between the vectors representing these words.

The original computational models of semantic information (arising from the psychological literature) were based on the idea that the number of co-occurrences of words in particular contexts formed the basis of the multidimensional space and that the vectors were obtained by applying a set of transformations to the count matrix. For instance, Latent Semantic Analysis (LSA; Landauer & Dumais, 1997) starts by counting how many times a word is observed within a document or a paragraph. The Hyperspace Analogue to Language

(HAL; Lund & Burgess, 1996) counted how many times words co-occurred in a relatively narrow sliding window, usually consisting of up to ten surrounding words. Because of the common counting step, following Baroni, Dinu & Kruszewski (2014) we will refer to this family of models as ***count models***.

In count models, the result of this first step is a word by context matrix. What usually follows is a series of transformations applied to the matrix. The transformations involve some kind of a weighting scheme, based on frequency-inverse document frequency, positive pointwise mutual information (PPMI), log-entropy, and/or a dimensionality reduction step (most commonly singular value decomposition; SVD). Sometimes the transformation is the defining component of the method, as is the case for LSA, which is based on SVD. In other cases, however, the transformations have been applied rather arbitrarily to the counts matrix based on empirical studies investigating which transformations optimized the performance on a set of tasks. For example, in its original formulation, the HAL model did not involve complex weighting schemes or dimensionality reduction steps, but later it was found that they improved the performance of the model (e.g., Bullinaria & Levy, 2007, 2012). Transformations are now often applied when training the models (e.g., Recchia & Louwerse, 2015; Mandera, Keuleers, & Brysbaert, 2015).

If we consider Marr's (1982) distinction between computational, algorithmic, and implementational levels of explanation, the count models are *only defined* at the *computational* level (Landauer & Dumais, p. 216): They consist of functions that map from a text corpus to a count matrix and from the count matrix to its transformed versions. Regarding the algorithmic level, Landauer and Dumais (1997) did not attribute any realism to the mechanisms performing the mapping. They only proposed that the counting step and its associated weighting scheme could be seen as a rough approximation of conditioning or

associative processes and that the dimensionality reduction step could be considered an approximation of a data reduction process performed by the brain. In other words, it cannot be assumed that the brain stores a perfect representation of word-context pairs or runs complex matrix decomposition algorithms in the same way as digital computers do.[1] In the case of HAL, even less was said about the psychological plausibility of the selected algorithms. Another problem is that count models require all the information to be present before the transformations are applied, whereas, in reality, learning in cognitive systems is incremental, not conditional on the simultaneous availability of all information.

In other words, although the count models, like all computational models, were very specific about which properties were extracted from the corpus to build the count matrix, and which mathematical functions were applied to the counts matrix in the transformation step, they made it much less clear how these computations could be performed by the human cognitive system.[2] This is surprising, given that the models originated in the psychological literature.

Unexpectedly, a recent family of models, which originated in computer science and natural language processing, may be more psychologically plausible than the count models. Mikolov and colleagues (2013a) argued that a relatively simple model based on a neural network (see Figure 1) can be surprisingly efficient at creating semantic spaces.

=== INSERT FIGURE 1 HERE ===

[1] It is known that dimension reduction can be performed by biological (e.g. Olshousen & Field, 1996) and artificial (Hinton & Salakhutdinov, 2006) neural networks. This fact is rarely mentioned when authors discuss various approaches to distributional semantics in the psycholinguistic literature.

[2] Although Landauer and Dumais (1997) discuss how the LSA algorithm could hypothetically be implemented in a neural network, this aspect is not reflected in their implementation of the model.

This family of models is built on the concept of prediction. Instead of explicitly representing the words and their context in a matrix, the model is based on a relatively narrow window (similar in size to the one often used in the HAL model) sliding through the corpus. By changing the weights of the network, the model learns to predict the current word given the context words (Continuous Bag of Words model; CBOW) or the context words given the current word (skip-gram model). Because of the predictive component in this family of models, again following Baroni et al. (2014), we will refer to these models as ***predict models***. As indicated above, there are two main types: the CBOW model and the skip-gram model.

Even though the predict models originated outside the context of psychological research and were not concerned with psychological plausibility, the simple underlying principle – implicitly learning how to predict one event (a word in a text corpus) from associated events–, is arguably much better grounded psychologically than constructing a count matrix and applying arbitrary transformations to it. The implicit learning principle is congruent with other biologically inspired models of associative learning (Rescorla & Wagner, 1972), given that they both learn on the basis of the deviation between the observed event and the predicted event (see Baayen, Milin, Filipovic Durdevic, Hendrix, and Marelli, 2011). An additional advantage of the model is that it is trained using a stochastic gradient descent, which in this case means that it can be trained incrementally with only one target-context pairing available for each update of the weights, and does not require all co-occurrence information to be present simultaneously as is the case with the count models.

To illustrate in what sense we consider the predict models to be psychologically plausible, we would like to compare them to the Rescorla-Wagner model – a classical learning model (for a review see Miller, Barnet, & Grahame, 1995), which has also been

successfully applied to psycholinguistics (Baayen et al., 2011). This model learns to associate cues with outcomes by being sequentially presented with training cases. For each training case, if there is a discrepancy between the outcomes predicted based on current association weights and the observed outcomes (lack of an expected outcome or presence of an unexpected outcome), the weights are updated using a simple learning rule.

Interestingly, the update rule of the Rescorla-Wagner model is known to be mathematically equivalent to the delta rule (Sutton and Barto, 1981), which describes stochastic gradient descent in a neural network composed of a single layer of connections and which was independently proposed outside of the context of psychological research (Widrow-Hoff, 1960). The same rule has been generalized to networks consisting of multiple layers of connections and non-linear activation functions as a backpropagation algorithm (Rumelhart, Hinton, & Williams, 1986) and is used to determine changes in connection weights in connectionist models. In other words, the Rescorla-Wagner model is just a special case of the backpropagation algorithm used with a stochastic gradient descent.

Similarly to the Rescorla-Wagner model, the learning mechanism which is used to train the predict models is also based on backpropagation with stochastic gradient descent. These models learn to minimize errors between the outcomes predicted on the basis of the cues and the observed outcomes by updating the weights of the connections between the nodes in the network when observing events in a text corpus. Here cues and outcomes correspond to target and context words in a sliding window, and each update of the weights is based on a predicted and observed pairing between the target word and its context. The learned semantic representation, which can be thought of as a pattern of activation of the hidden nodes for a word in an input layer, is learned as a by-product of learning to associate

contexts and target words. The model is usually trained in one pass over the corpus with the number of the training cases dependent on the size of the corpus.

In this sense, the predict models are trained using a similar technique as the Rescorla-Wagner learning rule, adapted for a network which includes a hidden layer and a non-linear activation function. It could be argued that introducing the hidden layer and non-linearity to the model make it conceptually more complex than the Rescorla-Wagner model.[3] However, it is clear that it may be impossible to represent more complex phenomena, such as semantics, in models as simple as the Rescorla-Wagner model. In the case of the predict models, the hidden layer is necessary to introduce a dimensionality reduction step (Olshousen & Field, 1996; Hinton & Salakhutdinoy, 2006) and a non-linear (softmax) activation function is necessary to transform activations of outcomes to probabilities. In fact it has been argued that using neural networks deeper than three layers may be necessary and justified to simulate and explain cognitive phenomena: deep neural networks have proven to be successful in a large variety of fields (for a review see Le Cun, Bengio, & Hinton, 2015) and hierarchal processing is also recognized as a fundamental principle of information processing in the human brain (Hinton, 2007). The need for recognizing deeper architectures as valid approaches to cognitive modeling has also been proposed in the psychological literature (Zorzi, Testolin, & Stoianov, 2013; Testolin, Stoianov, Sperduti, Zorzi, 2015).

[3] It is important to note that although a network with no hidden layers may be simpler conceptually, it does not necessarily mean that it is more parsimonious in terms of the number of parameters that need to be specified. For example, consider a network with 50,000 words as cues and the same number of outcomes. A fully-connected network with a single layer of connections, such as the Rescorla-Wagner model, would require 50,000 x 50,000 = 2.5 billion parameters (weights) to be specified, while introducing a hidden layer including 300 nodes drastically reduces this number to 2 x 50,000 x 300 = 30 million parameters (weights).

In addition to their potential theoretical appeal, the predict models were shown to offer a particularly good performance across a set of tasks and generally outperform the count models (Baroni et al., 2014; Mandera et al., 2015) or perform as well as the best tuned count model. On the other hand, it has also been argued that the superior performance of the predict models is largely due to using better tuned parameters as default for training these models than is the case for the count models (Levy, Goldberg, Dagan, & Ramat-Gan, 2015). Even if the performance of the predict models does not surpass that of the count models, they are generally much more compact in terms of how much computational resources they require, which is also of practical importance.

Although the predict models are built on a quite simple principle, it is not as obviously clear as in the case of the count models what, in mathematical terms, these models are computing (Goldberg & Levy, 2014). Interestingly, it has been argued that some of the predict models may implicitly perform a computation that is mathematically equivalent to the dimensionality reduction of a certain type of the count model. In particular, Levy and Goldberg (2014) argued that the skip-gram model is implicitly factorizing a PMI transformed count matrix shifted by a constant value. If this is the case, and the relationship between the two classes of models becomes well understood, this could create an interesting opportunity for psychologists by showing how mathematically well-defined operations (PPMI, SVD) can be realized on psychologically plausible systems (neural networks) to acquire semantic information.

Given the potential convergence of the predict and count models it becomes especially important to introduce the predict models to psycholinguistics. If the count models are well specified at Marr's (1982) computational level of explanation, the predict models could provide an algorithmic level explanation, bringing us closer to understanding how

semantic representations may emerge from incrementally updating the predictions about co-occurrences of events in the environment. Nevertheless, because to our knowledge this is the first time these models are discussed in a psycholinguistic context, in the current paper we did not focus on investigating the convergence of the two classes of models but chose to train different semantic spaces with typical parameter settings and details of the training procedures.

To advance our understanding of the new predict models (both CBOW and skip gram) and their relationship to the more traditional count models in a psychological context, we performed an evaluation of the three types of models against a set of psychologically relevant tasks. In order to gain a more complete picture of how these models perform we tried to explore their parameter space instead of limiting ourselves to a single set of parameters. In addition, we wanted to find out how much of what we have learned about count models can be generalized to the predict models.

Of course, the investigated implementations of the predict models are only loosely related to psychologically plausible principles (such as prediction). We do not claim that the investigated predict models represent a human capacity to learn semantics in a fully realistic way, but rather we argue that they should be investigated carefully because they may represent an interesting starting point for bridging the theoretical gap between the count models, various transformations applied as part of these models, and fundamental psychological principles.

Comparing Distributional Models of Semantics

There is a rich literature in which different approaches to distributional semantics have been evaluated. In general they form two types of investigations: Either various parameters and transformations within one approach are tested to find the most successful set of parameter settings (e.g. Bullinaria and Levy 2007, 2012), or different approaches are compared to each other to establish the best one (e.g. Baroni et al., 2014; Levy et al., 2015).

The evaluations are often based on a wide range of tasks. For example, Bullinaria and Levy (2007, 2012) compared the performance of a HAL-type count model on four tasks: The Test of English as a Foreign Language (TOEFL; Landauer & Dumais, 1997), distance comparison, semantic categorization (Patel et al., 1997; Battig & Montague, 1969), and syntactic categorization (Levy et al., 1998). The authors varied a number of factors such as the window size, the applied weighting scheme, whether dimensionality reduction was performed, whether or not the corpus was lemmatized (all inflected words replaced by their base forms), and so on. They found that the best results on their battery test were achieved by the models that used narrow windows, the PPMI weighting scheme, and a custom, SVD-based dimensionality reduction step. The lemmatization or use of stop-words did not improve the performance of the model.

Comparisons of different classes of models include a recent comparison of the predict approach to the traditional count model on a range of computational linguistic benchmark tasks: Baroni et al. (2014) compared the models using semantic relatedness (Rubenstein and Goodenough, 1965; Agirre et al., 2009, Bruni et al., 2014), synonym detection (TOEFL; similar to Landauer and Dumais, 1997), concept categorization (purity of clustering categorization, Almuhareb, 2006; Baroni et al., 2008; Baroni et al., 2010), selection preferences (noun-verb pairs, how similar are they as subject-verb or object-verb pairs,

Baroni and Lenci, 2010; Padó & Lapata, 2007; McRae et al., 1998), and analogy (Mikolov et al., 2013a) and found that the predict models had a superior performance on computational linguistic benchmark tasks and were more robust to varying parameter settings. Levy, Goldberg, Dagan, & Ramat-Gan (2015) show that although count models lack the robustness of predict models, they can work equally well with specific weighting schemes and dimensionality reduction procedures.

It is clear that the benchmark tasks from computational linguistics may not be the most relevant ones for issues related to human semantic processing and representation. For instance, a lot of attention has been devoted to how well various distributional semantic models perform on the TOEFL, which consists of choosing which of four response alternatives most closely matches a target word over 80 trials with increasing difficulty. Unless we want to model scholastic over-achievement, there is no a priori reason to believe that the model scoring best on this test is also the psychologically most plausible one. A simple psycholinguistic benchmark could consist of correctly predicting the proportion of alternatives chosen by participants. In this respect, the relatedness ratings or elicited associations tasks used in the computational linguistics benchmarks can also be considered valid benchmarks for psycholinguistics. However, evaluating computational models in psycholinguistics also involves comparing predictions about the time course associated with processing stimuli. The most frequently used task to study the time course of semantic processing in humans is semantic priming. This task consists of the presentation of a prime word followed by a target stimulus. Usually, the task involves either reading the target word out loud (naming) or deciding whether the stimulus is an existing word or a pseudoword (lexical decision). The task does not involve an explicit response about the semantic relationship between prime and target. However, it is assumed that the time it takes to name

the word out or to make a decision on its lexicality is decreased by the degree of semantic relatedness between the prime and the target. Therefore, in contrast to other benchmarks in which participants are asked to give explicit responses about semantic content, semantic priming is assumed to inform us about the implicit workings of semantic memory.

Predicting Semantic Priming with Distributional Models

The question of whether semantic similarity measures derived from distributional semantics models can predict semantic priming in human participants has been investigated in a number of psycholinguistic studies. In terms of the methodology employed these investigations can be divided in two classes. Some studies simply look at the stimuli across related and unrelated priming conditions and investigate whether there is a significant difference in semantic space derived similarity scores between these conditions. Other studies try to model the semantic priming at the item level by means of regression analysis.

The first class of studies is exemplified by Lund, Burgess, & Atchley (1995) who found that the HAL-derived similarity measures significantly differed for semantically related and unrelated conditions. A similar approach was taken by McDonald & Brew (2004) and Pado & Lapata (2007), who used distributional semantics models to model semantic priming data from Hodgson (1991). Jones, Kintch, & Mewhort (2006) compared the BEAGLE, HAL and LSA models on a wide range of priming tasks, and investigated differences in how well these methods mimicked the results of multiple priming studies.

The regression-based approach was already employed in Lund and Burgess (1996), who reported that relatedness measures derived from HAL significantly correlated with semantic priming data from an existing priming study (Chiarello, Burgess, Richards, & Pollock, 1990). A detailed examination of the factors modulating the size of the semantic

priming effect based on 300 pairs of words was conducted by Hutchison, Balota, Cortese, & Watson (2008). In a regression design the authors found no effect of the LSA score. However, it is worth noting that a large number of other predictors were entered in the analysis, including other semantic variables, such as forward and backward association strength from an association study by Nelson, McEvoy, & Schreiber (1998). Collinearity of these measures may have contributed to the fact that no significant effect of the LSA score was found. In addition, the null result does not prove that computational indices are unable to predict semantic priming, as the quality of the used semantic space may have been suboptimal.

Another item-level study was conducted recently by Günther, Dudschig and Kaup (2016) in German. In that study the authors carefully selected a set of items spanning the full range of LSA similarity scores computed on the basis of a relatively small corpus of blogs (about 5 million words). The authors found a small but significant effect of the LSA similarity scores on semantic priming. The critical difference between this study and the one conducted by Hutchison et al. (2008) was in how the authors analyzed the data: Hutchison et al. (2008) first subtracted RTs in the related condition from the RTs in the unrelated contition and then fitted regressions to the resulting difference. Günther et al. (2015) simply predicted the reaction times to the target words while including a set of other variables (including semantic similarity with the prime) as predictors. Difference scores between correlated variables are known to have a low reliability (Cronbach & Furby, 1970) and arguably reduced reliability may have contributed to lack of significant effect in the study by Hutchison et al. (2008).

Although the item-level, regression based approach has multiple advantages over factorial designs (Balota, Cortese, Sergent-Marshall, Spieler, & Yap 2004; Balota, Yap, Hutchison, Cortese, 2012), until recently it was difficult to conduct this type of analysis on a

sufficiently large number of items. Fortunately, due to the recent rise of megastudies (Keuleers & Balota, 2015), the situation is improving rapidly. Thanks to the semantic priming project (SPP) ran by Hutchison and colleagues (2013), we now have a much better opportunity to look at how much of the total variability in primed lexical decision times (LDT) and word naming times can be explained by semantic variables based on distributional semantics models. The advantage of this approach is that with enough data we can directly model RTs as a function of semantic similarity between the prime and the target, also including other critical predictors known to influence performance on psycholinguistic tasks. Because in a megastudy approach it is natural to focus on effect sizes more than on categorical decisions based on statistical significance, the method lends itself to comparing various semantic spaces by examining how much variance in RTs they account for.

Corpus Effects In Distributional Semantics

The performance of distributional semantics models in accounting for human data can be affected by the degree to which the training corpus of the model corresponds to the input human participants have been exposed to. Ideally, the model would be trained on exactly the same quality and size of data as participants of psycholinguistic experiments (typically first-year university students). Of course, this ideal can only be approximated. In particular, much of the language humans have been exposed to is spoken and can only be used for modeling purposes after a time-consuming transcription process. Instead, models are typically based on written language which is available in large quantities but is often less representative of typical language input.

It has been observed that frequency measures based on corpora of subtitles from popular films and television series outperform frequency measures based on much larger

corpora of various written sources. For instance, Brysbaert, Keuleers, and New (2011) showed that word frequency measures based on a corpus of 50 million words from subtitles predicted the lexical decision times of the English Lexicon Project (Balota et al., 2007) better than the Google frequencies based on a corpus of hundreds of billions words from books. A similar finding was reported by Brysbaert, Buchmeier, Conrad, Jacobs, Bölte, and Böhl (2011) for German. In particular, word frequencies derived from non-fiction, academic texts perform worse (Brysbaert, New, & Keuleers, 2012). On the other hand, Mandera, Keuleers, Wodniecka, & Brysbaert (2015) showed that a well-balanced corpus of written texts from various sources performed as well as subtitle-based frequencies in a Polish lexical decision task.

An interesting question in this respect is how important the corpus size is for distributional semantics vectors. Whereas a corpus of 50 million words may be enough for frequency measures of individual words, larger corpora are likely to be needed for semantic distance measures, as estimation of semantic vectors composed of hundreds of values may be a more demanding task than assigning a frequency to a word. Some evidence along these lines was reported by Recchia and Jones (2009), who observed that using a large corpus is more important than employing a more sophisticated learning algorithm. The two corpora they compared contained 6 million words versus 417 million words. On the other hand, De Deyne, Verheyen, & Storms (2015), based on a comparison between corpus samples of various sizes, conclude that corpus size is not critical for modeling mental representations.

In addition to the effect of corpus size, the language register tapped into by the corpus could also influence semantic distance measures based on distributional models. We will discuss this issue by comparing the performance of models based on subtitle corpora with the performance of models based on written materials. If subtitle corpora perform better than the

larger text corpora of written materials, this indicates that register is an important variable. In addition, if the concatenation of both corpora turns out to be inferior on some tasks, this is again an indication of the importance of the register captured by subtitle corpora.

Evaluating Semantic Spaces as Psycholinguistic Resources

The availability of the priming lexical decision and word naming megastudy data collected by Hutchison and colleagues (2013) makes a systematic comparison of various measures of semantic relatedness feasible and opportune. In addition to various distributional semantic models, semantic relatedness ratings can also originate from feature-based data (McRae, Cree, Seidenberg, & McNorgan, 2005), human association norms (Nelson et al., 1998), or semantic relatedness ratings (Juhasz, Lai, & Woodcock, in press). Although we will include these alternatives in our comparison, it should be noted that they have some important practical limitations: (1) they are defined only for a subset of words and (2) they do not exist in most languages that can be potentially of interest to psycholinguists.

To perform the evaluation, the logic of evaluating word frequency norms (Brysbaert and New, 2009; Keuleers et al., 2010) will be followed. In these evaluations, various word frequency norms are used to predict lexical decision and word naming RTs in order to identify the set of norms that accounts for the largest percentage of variance in the behavioral data (ideally together with other lexical variables that affect word processing times, such as word length and neighborhood density). An almost identical procedure can be applied to semantic spaces. A linear regression model can be fitted to the lexical decision and naming latencies of target words preceded by semantically related or unrelated primes. The variables known to influence word recognition (frequency, length, and similarity to other words) will be used as baseline predictors, to which the semantic distance between the prime and the

target derived from the various distributional semantics models will be added. This leads to the measurement of how much extra variance in behavioral data can be accounted for by adding relatedness measures from each distributional semantic model.

Although this approach can be informative of a model's absolute performance, it does not give an indication of the relative evidence in favor of each model. The approach based on comparing amount of variance explained is also biased towards more complex models when comparing them against the baseline (including more variables gives more explanatory power but may result in overfitting the training data). In order to overcome these limitations, we applied a regression technique based on Bayes factors (e.g. Wagenmakers, 2007) as described by Rouder and Morey (2013; see also Liang, Paulo, Molina, Clyde, & Berger, 2008). The Bayes factor is a measure of relative probability of the data under a pair of alternative models. This method also automatically incorporates a penalty for model complexity (Wagenmakers, 2007) and is flexible with respect to which models can be compared. For instance, it allows the comparison of non-nested models, which is difficult in a frequentist approach (Kass & Raftery, 1995). This property makes it possible to quantify the relative evidence in favor of models with predictors from various semantic spaces.

Although we consider the data from the semantic priming project as the most informative with respect to getting insight into the semantic system of typical participants in psychology experiments, we will also look at how well the various measures perform on a number of other tasks, and we will include some data from the Dutch language, to test for cross-language generalization. In addition, where possible we will compare the outcome of the new variables to those currently used by psycholinguists.

Initially, we intended to compare two count models (LSA-inspired and HAL-inspired) with two predict models (CBOW and skip-gram). However, when we tried to calculate the

LSA-type model on our corpora, it became clear that the number of documents (particularly in the UKWAC corpus) was too large to represent the term by document matrix in computer memory and perform SVD on that matrix. As a result, we had to use a non-standard, more scalable implementation of the SVD algorithm implemented in the Gensim toolkit (Rehurek & Sojka, 2010), which returned vectors that were not doing particularly well. Because it is not clear whether the bad performance of the LSA-type measure is due to the inferior performance of the LSA approach itself or to the algorithm, and because LSA-based measures in the past have done worse than HAL-based measures, we decided not to include the former in the analyses reported below. For the most important task (semantic priming), however, we do provide the LSA measures as provided by the Colorado website for comparison purposes.

Finally, to obtain a more nuanced view of how the models perform across different parameter settings we explored their parameter space. By doing so, we make sure that we give each model maximal opportunity and we can examine whether all models are similarly affected by, for instance, the size of the window around the target word or the number of dimensions included in the model.

Method

For each corpus the tokenization was done by extracting all the alphabetical strings. Following Bullinaria and Levy (2007, 2012) no lemmatization or exclusion of function words was used. To represent the degree to which two words are related according to the used semantic spaces we computed cosine distances between word vectors *u* and *v* according to the formula:

$$D_{cos}(u, v) = [1 - \frac{u \cdot v}{||u||||v||}]$$

In this formula $u \cdot v$ stands for a dot product between vectors u and v, and $||u||$ and $||v||$ for the length of the vector u and v respectively.

English

Text corpora

The corpora we used for creating the English semantic spaces were UKWAC (a corpus of about 2 billion words resulting from a web crawling program; Ferraresi, Zanchetta, Baroni, & Bernardini, 2008) and a corpus of about 385 million words compiled from film and television subtitles. More information about UKWAC can be found in Ferraresi et al. (2008).

The subtitle corpus was created based on 204,408 documents downloaded from the Open Subtitles website (http://opensubtitles.org) whose language was tagged as English by the contributors of that website. We first removed all subtitle related formatting. Next, to eliminate all documents that contained a large proportion of text in a language other than English, we calculated preliminary word frequencies based on all documents, and removed all documents in cases where the 30 most frequent words did not cover at least 30% of the total number of tokens in that subtitle file. Because many subtitles are available in multiple versions we implemented *duometer* [4], a tool for detecting near-duplicate text documents using the MinHash algorithm (Broder, 1997). The final version of the corpus contained 69,382 documents and 385 million tokens.

[4] We released *duometer* as an open-source project. The tool and its source code are available at: http://github.com/pmandera/duometer

We also combined the two corpora for the purpose of computing the semantic spaces. The combined corpus contained 2.33 billion tokens and 2.76 million documents.

Model training

We trained the (HAL-type) count model by sliding a symmetrical window through the corpus and counting how many times each pair of words co-occurred. We considered the 300,000 most frequent terms in the corpus as both target and context elements (Baroni et al., 2014). Next, we transformed the resulting word by word co-occurrence matrix using the positive pointwise mutual information (PPMI) scheme (Bullinaria & Levy, 2007). The transformation involved computing pointwise mutual information (Church & Hanks, 1990) for each pair of words *x* and *y* according to the formula:

$$PMI(x, y) = log_2 \frac{p(x, y)}{p(x)p(y)}$$

Where *p(x)* is the probability of the word *x* in the text corpus, *p(y)* is the probability of the word *y* in the text corpus and *p(x, y)* is the probability of the co-occurrence of the words *x* and *y*. In the final step, the values of the cells in the matrix for which the pointwise mutual information values were negative were substituted with 0, so that the matrix contained only non-negative values (hence *positive* pointwise mutual information).

We trained the CBOW and skip-gram models using Gensim (Rehurek & Sojka, 2010)[5], an implementation that is compatible with word2vec (Mikolov et al., 2013a) – the original implementation of the predict models. For these models, all word forms occurring minimally 5 times in the corpus were included. Each model was trained using 50, 100, 200,

[5] The toolkit is available at https://radimrehurek.com/gensim/

300 and 500 dimensions. We set the parameter k for negative sampling to 10 and the sub-sampling parameter to 1e-5. Sub-sampling is a method for mitigating the influence of the most frequent words (Mikolov et al., 2013a) by randomly removing words with a probability higher than a pre-specified threshold. Negative sampling is a computational optimization that avoids computing probabilities for all words in an output layer. In each learning case only a subset of words is considered.

An important parameter influencing the performance of count models (Bulinaria & Levy, 2007, 2012) is the size of the sliding window. We varied this parameter for the count and predict models in the range from 1 to 10 words before and after the target word[6] (i.e., the minimal window size of 1 included 3 words: the target word, one word before, and one word after).

Evaluation tasks

In order to keep vocabulary size constant across the count and predict models and across the three corpora used (subtitles, written texts, and their combination), we used only the subset of words that all semantic spaces had in common. We also wanted to compare our semantic spaces with the best performing space from Baroni et al. (2014; CBOW model with 400 dimensions, window size 5, negative sampling value 10, trained on the concatenation of the UKWAC, Wikipedia and the British National corpus including 2.8 billion words)[7]. Therefore, we further limited the vocabulary of the models to the intersection with the vocabulary of that dataset. The resulting semantic spaces contained 113,000 distinct words.

---

[6] The CBOW and skip-gram models limit the size of the window used on individual learning trials to a randomly chosen value in the range from 1 to the requested window size.

[7] Downloaded from: http://clic.cimec.unitn.it/composes/semantic-vectors.html

Semantic Priming – method

We used the data from the Semantic Priming Project (Hutchison et al., 2013), which contains lexical decision times and naming times to 1,661 target words preceded by four types of primes. Two prime types were semantically related to the target but differed in their association strength; the other two types were unrelated primes matched to the related primes in terms of word length and word frequency. The Semantic Priming Project contains two more variables of interest for our purpose. They are the semantic similarity measures derived from LSA (trained on the TASA corpus, 300 dimensions; Landauer & Dumais, 1997) and from BEAGLE (Jones et al., 2006). These numbers allow us to compare the newly calculated measures to the current state of the art in psycholinguistics. As the data were not available for all prime-target pairs, this further reduced the dataset. In the end 5,734 of the original 6,644 prime-target pairs remained.

For lexical decision (LDT) all non-word trials were excluded from the dataset and for both LDT and word naming we excluded all erroneous responses. We excluded all trials with RTs deviating more than 3 standard deviations from the mean and computed z-scores separately for each participant and each session. Finally, we averaged the z-scores for each prime-target pair and used the result as the dependent variable in our analyses.

Next, we fitted linear regression models with various predictors to evaluate the amount of variance in the standardized RTs that could be accounted for. First, we calculated a baseline model including log word frequency (SUBTLEX-US; Brysbaert & New, 2009), word length (number of letters), and orthographic neighborhood density (Coltheart, Davelaar, Jonasson, & Besner, 1977) of both the prime and the target (all variables as reported in the Semantic Priming Project dataset). Then, we fitted another linear regression model including the baseline predictors plus the measure of semantic distance between the prime and the

target provided by the semantic space we were investigating, and looked at how much extra variance the semantic similarity estimate explained. We used all pairs of stimuli across all conditions (both related and unrelated words).

Semantic priming – results

The baseline regression model including the logarithm of word frequency, length, and neighborhood density (all predictors included for both the prime and the target word) explained 38.9% of the variance in the lexical decision RTs and 31.2% of the variance in the word naming latencies (see Figure 2).

=== INSERT FIGURE 2. HERE ===

When the relatedness scores from the distributional semantics models were added as a predictor, the amount of variance explained increased for both tasks. The improvement was already highly significant for the relatedness measure based on the worst performing model. For LDT, this was the skip-gram model trained on the concatenation of the subtitle and the UKWAC corpus with dimensionality 500 and window size 1 [improvement relative to the baseline model: $F(1, 5729)=367.27$, $p < 0.001$)]; for word naming it was the skip-gram model trained on the UKWAC corpus with dimensionality 500 and window size 1 [$F(1, 5729)=99.778$, $p < 0.001$)].

As can be seen in Figure 2, the fit of the three semantic vectors depended on the task (LDT vs. naming), on the window size, and on the corpus taken into account. For each type

of a model, Table 1 shows its average performance across all tested parameters and the performance and parameters of the best model.

=== INSERT TABLE 1. HERE ===

Several interesting findings emerged from our analyses. First, in many cases the models trained on the subtitle corpus outperformed the models based on the UKWAC written corpus or the combination of the two corpora. This effect was particularly clear for the count (both in LDT and word naming) and the CBOW models (in LDT). The difference was less clear for the skip-gram models. In all cases, the addition of the 385 million words from the subtitle corpus to the 2.33 billion word corpus of written texts considerably improved performance.

A second remarkable observation is that the best models are quite comparable but have different window sizes. In particular, for the count model there is a steep decrease in performance with increasing window size above 3 which was not observed for the predict models. As a result, the optimal window size is larger for the predict models than for the count model.

Semantic priming – a comparison with the existing measures of semantic similarity

To further gauge the usefulness of the new semantic similarity measures, we compared the extra variance they explain to that explained by the currently used measures. The Semantic Priming Project database includes measures for LSA and BEAGLE. Currently,

if a distributional semantics model is used for the purpose of selecting experimental stimuli, psychologists tend to rely on the LSA space available through a web interface at the University of Colorado Boulder (http://lsa.colorado.edu/; Landauer & Dumais, 1997). This is understandable, as the semantic space was created to accompany a classic paper and because the resource has a practical interface which makes data extraction easy. Yet, given the recent developments in distributional semantics and the availability of much larger corpora than the one on which the LSA spaces were trained (most prominently the TASA corpus of about 11 million words), there is a need to reevaluate whether the LSA-based semantic spaces should remain the default choice for measuring semantic relatedness in psychological research.

The TASA-based LSA similarity scores explained 43.9% of the variance in lexical decision reaction times and 32.7% of the variance in naming. The BEAGLE scores explained 43.0% of the variance in lexical decision reaction times and 32.3% of the variance in word naming latencies.[8] All values are below those of the best performing CBOW model (45.5% in LDT and 33.2% in naming).

Our best models also compare well relative to the spaces trained by Baroni et al. (2014). The best performing semantic space of Baroni et al. (2014) explained 44.0% of the variance in lexical decision reaction times and 33.0% of the variance in word naming latencies.

To examine how much more variance could be explained by human word association norms (Nelson et al., 1998) and feature norms (McRae et al., 2005), we performed an analysis on the subsets of words that are included in these datasets.[9] We compared the

[8] BEAGLE scores based on cosine distances; the other measures performed worse.

[9] Similar analysis could in theory be run using the scores derived from the Simlex-999 and the Wordsim-353 ratings but the overlap with the semantic priming data was too small in these cases to allow a meaningful analysis.

semantic similarity indices based on the human data to those of the best count, CBOW and skip-gram spaces for the lexical decision task. There were 2,904 cue-target pairs that were simultaneously present in the priming data, the association norms and the vocabulary of our semantic spaces.

For this subset of Semantic Priming Project data, the baseline regression model (including logarithm of word frequency, length and neighborhood density of both the cue and the target) explained 38.9% of the variance in LDT and 31.2% in word naming. The model that additionally included human forward association strength explained 41.7% of the variance in lexical decision RTs and 32.7% of the total variance in word naming. The best performing count model (trained on the subtitle corpus, using window size 3) explained 42.3% of the variance in lexical decision RTs and 31.9% of the variance in word naming latencies. The best CBOW model (trained on the subtitle corpus; 300 dimensions; window size 6) accounted for 41.9% of the variance in LDT RTs and 32.0% in word naming latencies. The best skip-gram model (trained on the concatenation of the UKWAC and subtitle corpus; 200 dimensions; window 10) explained 41.0% of the variance in lexical decision and 32.1% of the variance in naming. As can be seen, all models performed very similarly and close to what can be achieved by human data. We would like to note, however, that it is harder to explain additional variance in RTs based on relatedness data, because the subset of the Semantic Priming Project that was used for this analysis contained only pairs of words generated as associates in the Nelson et al. (1998) database, which significantly reduced the range of relatedness values.

The intersection between the feature norms from McRae et al. (2005), the semantic priming data, and the vocabulary data of our datasets included 100 word pairs. The baseline model explained 37.0% of the variance in LDT RTs and 29.3% of the variance in word

naming latencies. Adding the relatedness scores computed as the cosine between the features vectors increased the percentage of variance accounted for by the model to 42.7% for LDT RTs and to 29.8% for word naming latencies. The amount of variance explained by the model in which we inserted the measures derived from the best performing count model was 54.6% for LDT RTs and 35.3% for word naming. In the case of the best CBOW model, the total explained variance amounted to 52.8% for lexical decision and 32.3% for naming. When the best performing skip-gram model word distance estimates were included in the model, it explained 52.3% of the variance in LDT RTs and 31.9% of the variance in word naming latencies. So, for this dataset, the semantic spaces actually outperformed the human data.

Semantic priming – Bayes factors analysis

To further gauge the importance of the semantic vectors, we calculated Bayes Factors. These inform us how much more likely one model is relative to another. For all Bayesian analyses reported in this paper we adopted an approach described by Rouder and Morey (2013; see also Liang, Paulo, Molina, Clyde, & Berger, 2008). We used default[10] mixture-of-variance priors on effect size. For both LDT and naming we first identified baseline models that included an optimal combination of lexical, non-semantic covariates. For both the prime and the target, we considered the following co-variates: log of word frequency, length, and orthographic neighborhood density.

A Bayes factor of 10 is assumed to be strong evidence for the superiority of a model and a Bayes Factor above 100 is considered as decisive evidence. As can be seen in Table 2,

[10] The default 'medium' setting for the rscaleCont argument in the *regressionBF* function in the R BayesFactor package, corresponding to the *r* scale = sqrt(2)/4. We also conducted a series of analyses with altered priors but this did not change the qualitative pattern of results, so we report only analyses conducted with default settings.

all Bayes Factors we calculated were far above these values. Moreover, we found evidence that combining some of the semantic vectors outperformed a model based on each of them separately, although it is not clear how useful such a combination would be for stimulus selection. Bayes Factor analysis also confirmed that models based on count and CBOW were decisively better than those based on word association norms and feature norms.

=== INSERT TABLE 2. HERE ===

When we compared the three types of models, we observed that in the analysis of LDT, the best performing count model (trained on the subtitle corpus with window size 3) did better than the best CBOW model (including 300 dimensions, trained on the subtitle corpora with window size 6; BF_{10}=521) and the best skip gram model (trained on the concatenation of the UKWAC and subtitle corpora, 200 dimensions, window size 10, BF_{10}=5.77 x 10^{24}). For naming, the best count model also outperformed the best CBOW model (BF_{01}=17.1) and even the best skip-gram model did better than CBOW (BF_{01}=21.2), although these differences were much smaller.

In summary, the Bayesian analysis showed overwhelming evidence in favor of including semantic relatedness measures derived from semantic spaces in both naming and LDT. Even the worst models did considerably better than the baseline model. There is some evidence that the best count model outperformed the best CBOW model, but this superiority is limited to a single window size. Finally, there is also some evidence that combining various relatedness measures may be advantageous. This suggests that different models may capture unique information that independently explains human performance in semantic priming.

Finally, it is clear that the distributional semantics models outperform the available human associations and feature norms in explaining human performance in semantic priming.

Word association norms – method

In order to evaluate how well the different models can predict human association data we used the dataset collected by Nelson, McEvoy and Schreiber (1998). This contains word associations for 5,019 stimulus words collected from over 6,000 participants. We limited the analysis to those associations that were present in all our semantic spaces, which resulted in a dataset of 70,461 different cue-response pairs (on average 14 associates per word).

To compare the word associations generated by humans to those generated by semantic spaces, we computed a metric based on the relative entropy between the probability distribution of the top 30 associates generated by the model and the associates generated by the human participants. This metric captures not only the probabilities for the words generated by humans but also evaluates whether the same words are generated by the semantic spaces.

To calculate the metric, the following steps were followed:

1. For each semantic space, we calculated the cosine distances between the cue word and all the other words, and selected the 30 words that were nearest to the cue word. A value of 30 corresponds to about twice the number of associates that are typically generated in human data. As such, it includes enough responses to be considered and does not deviate too much from the number of associates generated by humans.
2. Next, the similarity score for each associate was normalized by dividing it by the sum of all the similarity values for the cue. The same procedure was applied to the human

association data, with associate counts being converted to probabilities. If the semantic space did not include the associate that was present in the human data or vice versa, a value of 0 was assigned.

3. Next, an additive smoothing was applied to each distribution using a smoothing term of 1/n, in which n is the number of elements in the distribution.
4. The relative entropy between probability distribution P and another probability distribution Q was computed with the formula:

$$D_{KL}(P||Q) = \sum_i P(i) log_2 \frac{P(i)}{Q(i)}$$

5. Finally, the relative entropies were averaged across all cue words and the average relative entropy was used as the final score of a given semantic space. Note that a relative entropy measure is a measure of distance between probability distributions and, hence, the smaller the measure, the better the fit.

Word association norms - results

To compute a baseline for the performance of the models on the association norms, we used a set of semantic spaces with word vectors containing nothing but random values. The average relative entropy between the associations norms and 10 such randomly generated semantic spaces was 0.84 *(SD*=0.0001).

=== INSERT FIGURE 3. HERE ===

As can be seen in the upper panel of Figure 3, the models including semantic information did better than the baseline model (the line at the top of the graphs; remember than lower values are better here). The predict models did better than the HAL-type count model, with CBOW outperforming skip-gram. In addition, we again see the importance of including the subtitle corpus and of the window size. For the count model, the best performing model was trained on the subtitle corpus using window size 2 (relative entropy=0.70). For the CBOW model, the best performance was achieved by a model with 500 dimensions trained on the concatenation of the subtitle and the UKWAC corpora with window size 7, which had a relative entropy of 0.63. The best skip-gram model was trained on the same corpus, used the same window size, but had 300 dimensions and had relative entropy of 0.66. The average relative entropy for the measures derived from the count models was 0.73 (*SD*=0.02). For the CBOW models it was 0.69 (*SD*=0.03) and for the skip-gram model 0.71 (*SD*=0.03).

For comparison, the semantic space from Baroni et al. (2014) had a relative entropy of 0.68, which was better than the average of the models evaluated here but worse than the best of those models.

Similarity/Relatedness ratings - method

We used two datasets of human judgments of semantic similarity and relatedness to evaluate the correspondence between measures derived from semantic spaces and human semantic distance estimates.

Wordsim-353 (Agirre et al., 2009) is a dataset including 353 word pairs, with about 13 to 16 human judgments for each pair. For this dataset the annotation guidelines given to the judges did not distinguish between similarity and relatedness. However, the dataset was subsequently split into a subset of related words and a subset of similar words on the basis of two further raters' judgment about the nature of the relationship for each word pair.

The second set of human judgments is Simlex-999 (Hill, Reichart, & Korhonen, 2014), which contains similarity scores for 999 word pairs. What makes it different from Wordsim-353 is its clear distinction between similarity and relatedness. In the case of Simlex-999 participants were given very clear instruction to pay attention to the similarities between the words and not to their relatedness, so that word pairs such as *car* and *bike* received high similarity scores, whereas *car* and *petrol*, despite being strongly related, received low similarity scores.

To evaluate how well each semantic space reflects human judgments we computed Spearman correlations between the predictions of the models and the human ratings. When calculating the correlations we included only those pairs of words that were present in the combined lexicon of the semantic spaces.

Similarity/Relatedness ratings - results

As shown in Table 3, the correlations between semantic measures derived from the semantic spaces were higher for the sets of words from Wordsim-353 than for those from Simlex-999.

=== INSERT FIGURE 4. HERE ===

=== INSERT TABLE 3. HERE ===

As can be seen Figures 4, although the HAL-type count model did significantly worse than the CBOW and skip-gram models across the entire parameter set, this was particularly true for large window sizes. The best count model had a window size of one. The fit of the skip-gram model also tended to decrease with increasing window size, although there were differences between the three datasets tested. Performance of the CBOW model tended to be optimal for mid-range window sizes.

Importantly, there was little effect of window size for the CBOW models (except for the smallest sizes, which resulted in less good performance).

Interestingly, for this task, models trained on individual corpora tended to perform better than models trained on the combination of corpora.

TOEFL - method

TOEFL is a dataset of 80 multiple choice questions created by linguists to measure English vocabulary knowledge in non-native speakers. The task of the person taking the test is to decide which of four candidate words is most similar to the target word. Landauer and Dumais (1997) first used this task to evaluate a distributional semantics model.

In our evaluation, we consider that a model provides a correct answer to a TOEFL question when the correct candidate word has the smallest cosine distance to the target word in the semantic space compared to the other three candidate words. One point is awarded for that question in this case; zero points are given otherwise. When the target word or none of the four alternatives were present in the semantic space, we assigned a score of 0.25 to the item to simulate guessing.

TOEFL - results

The results are shown in the lower panel of Figure 3. The best count model (UKWAC corpus; window size 1) obtained a score of 83.7% on the TOEFL test. Average performance of the count models on this test was 61.2% (*SD*=9.76%).

The predict model with the highest score on TOEFL was a CBOW model with 500 dimensions and window size 1, trained on the concatenation of the UKWAC and the subtitle corpora (score=91.2%). The top skip-gram model was trained on the same corpus using the same window size but had 300 dimensions. On average, the CBOW models achieved a score of 73.4% (*SD*=10.9%) and the skip-gram models a score of 69.0% (*SD*=9.6%)

Models trained on the subtitle corpora performed worse on the TOEFL test than those trained on the UKWAC corpus or on the concatenation of both corpora, in line with the common sense prediction that the more we read the more rare words we know. Like before, the count model showed a strong decrease in precision with increasing window size. There was a decrease for the predict models as well, but it was less steep.

With a score of 87.5%, the semantic space from Baroni et al. (2014) surpassed the vast majority of our models on this task.

Dutch

Text corpus

We used the SONAR-500 text corpus (Oostdijk, Reynaert, Hoste, van den Heuvel, 2013) and a corpus of movie subtitles to train the distributional semantic models.

The SONAR-500 corpus is a 500 million words corpus of contemporary Dutch and includes a wide variety of text types. It is aimed at providing a balanced sample of standard Dutch based on textual materials from traditional sources such as books, magazines and newspapers, as well as Internet based sources (Wikipedia, websites, etc.).

Tokens from the SONAR-500 corpus were extracted using the FoLIa toolkit[11]. We found that the corpus contained a small number of duplicate documents. In order to remove them from the corpus we ran the MinHash duplicate detection using *duometer* within each category of texts in the corpus. The final version of the SONAR-500 corpus, after duplicate detection and applying our tokenization procedure included 406 million tokens (1.9 million documents).

In order to compile the subtitle corpus, we downloaded 52,209 subtitle files. The corpus was cleaned in the same way as the English subtitle corpus. The final Dutch subtitle corpus contained about 26,618 documents and 130 million tokens.

Finally, we combined the SONAR-500 corpus and the subtitle corpus. As the SONAR-500 corpus also includes movie subtitles, we only included documents from the subtitle corpus that did not have a duplicate in the SONAR-500 corpus. This resulted in a combined corpus of 530 million tokens (1.926 million documents).

Model training

We used the same procedure for training the semantic spaces as the one used for the English corpora. For the Dutch material, we only used the models with window sizes of 1, 2, 3, 5 and 10, because our experience with the evaluation of the English semantic spaces had shown that the results vary most between the initial values and the general trend in performance is similar at higher window sizes.

[11] http://proycon.github.io/folia/

When training the HAL-type count model, 300,000 types with the highest frequency were used as word and contexts. The PPMI weighting scheme was applied to the resulting co-occurrence matrix. The same parameter settings as for English were applied when training the predict models. However, we trained only models with 200 and 300 dimensions.

Evaluation Tasks

Semantic priming - method

Because there is no large, publicly available dataset of semantic priming in Dutch, our analysis was limited to two smaller datasets. The first one was based on a lexical decision experiment conducted by Heyman, Van Rensbergen, Storms, Hutchison and De Deyne (2015), which included 120 target words, each preceded by related and unrelated words. We used only words from the low memory load condition and for each prime-target pair we used the average reaction times for the two SOAs (1200 and 200 ms) used in the experiment. This resulted in a dataset of 240 prime-target pairs with associated RTs. For 236 of these pairs both the prime and the target were present in our semantic spaces and were included in further analyses.

The second dataset on which we based our analysis was collected by Drieghe and Brysbaert (2002). This dataset includes 21 target words with one semantically related prime word and two unrelated primes (one that was homophonic to the related prime and one that was completely unrelated). The small number of items in the second Dutch semantic priming dataset enabled only a very simple evaluation. In order to calculate how well each of the trained models fit the dataset we computed the distances between the primes and the targets for the related and the unrelated conditions, and we performed *t*-tests to verify whether the

distances in the unrelated conditions were larger than in the related condition, as is the case for the human reaction times.

Semantic priming - results

In the dataset from Heyman et al. (2015) the baseline model including log of word frequency and length for both the prime and the target explained only 4.8% of the variance in reaction times. The average performance of the models including various semantic predictors is presented in Table 4.

=== INSERT TABLE 4. HERE ===

The findings with the Dutch data are entirely compatible with what we observed in English. First, it is clear that the semantic vectors improved the percentage of variance accounted for. Even the worst performing model (count model based on the subtitle corpus trained with window size 10) did already 6% better (10.7% of the variance explained; contribution of this semantic predictor : $F(1, 230)=11.05$, $p = 0.001$). The increased performance of this model was confirmed in a Bayes Factor analysis ($BF_{10} = 110$).[12]

Second, also in line with the English data, the CBOW model explained the most variance in RTs (on average 19.1% of the variance explained), followed by the skip-gram model (17.7% of the variance explained), and the HAL-type count model (13.6% of the variance explained).

[12] As in the case of the English semantic priming data analysis, the Bayes factors are reported with reference to the optimal model that did not include semantic measures but was based on lexical variables only. This model included the logarithm of prime and target word frequency and was strongly supported relative to a model including intercept only (BF_{10}=29.01).

Third, the performance of the count model depended largely on the window size. The best performing count model had a window size of 2, was trained on a concatenation of the subtitle and SONAR corpora, and explained 16.2% of the variance in the reaction times. The best skip-gram model explained 20.7% of the variance. This model had 200 dimensions and was trained on the concatenation of the two corpora using window size 5. The best CBOW relatedness measures, which explained 22.4% of the variance in RTs, had 200 dimensions and were trained on the concatenation of the two corpora using window size 10. In the Bayes factor regression we found that the best model, overwhelmingly supported relative to the model based on lexical variables only (BF 10 =197,283,867), included the logarithm of prime and target word frequency in addition to the semantic relatedness measure from the best performing CBOW semantic space.

In a direct comparison of the relatedness measures derived from each type of models (count, CBOW and skip-gram), the Bayes factor analysis indicated a decisive advantage of the model including relatedness measures derived from the best CBOW model, relative to the model including the best count relatedness measures (BF_{10}=1682) and substantial evidence in favor of the CBOW relatedness measures relative to those derived from the skip-gram model (BF_{10}=8.4).

The dataset from Drieghe and Brysbaert (2002) contained a set of target words with one related prime and two unrelated primes. Because the dataset was too small to run analyses at the item level, we limited ourselves to *t*-tests. Table 4 gives the average similarity scores for the various models. It clearly shows that the semantic relatedness was larger in the related condition than in the unrelated conditions for all models. The situation was less convincing for the HAL-type count models. As in all previous analyses, the addition of the subtitle corpus considerably improved the predictive power of the models. For the best

model, the difference in semantic distance between the related and the unrelated primes had a standardized effect size of d = 1.4, which illustrates why the semantic vectors are such an important predictor for semantic priming studies.

Association norms - method

We used word association data from de Deyne and Storms (2008), who reported the associates most frequently given to 1,424 cue words. Like in the evaluation of the English data, we computed the average relative entropy between the probability distributions of the associates produced by our models and the human data.

Association norms - results

For the 1,424 cue words from de Deyne and Storms (2008), the baseline relative entropy score based on 10 randomly generated semantic vectors was 0.86 (*SD*=0.0005; lower is better).

The average relative entropy for the count models was 0.78 (*SD*=0.01). The best performing count model had a window size of 3 (trained on the SONAR-500 corpus), resulting in a relative entropy of 0.76.

The average relative entropy for the CBOW models was 0.79 (*SD*=0.03). The best performing model (relative entropy=0.74) was trained on the combined SONAR-500 and subtitle corpus, had 200 dimensions and a window of size 10.

The average relative entropy for the skip-gram models was 0.80 (*SD*=0.02) and the best performing model had the same parameters as the best performing CBOW model (relative entropy=0.75).

Influence of the window size

Our analyses indicated that the size of the window used to train the HAL-type count models is an extremely important parameter when training these models. At the same time, it has to be acknowledged that the count and the predict models use the window size parameter differently during training. While the typical count model considers full window size for each target word, the predict models per trial randomly choose a number between 1 and the requested window size and use that randomly chosen number as the window size for the training. This allows these models to utilize information about distant words but at the same time the average window size is reduced by half and the more distant words are included less often. To verify whether this aspect of the training can be responsible for the sharp drop in the performance of the count models that was not observed in the predict models we decided to train an additional set of count models using window sizes 1, 2, 3, 5, 7 and 10, on the English subtitle corpus and its concatenation with the UKWAC corpus. However, for this analysis we applied an analogous procedure of randomly choosing window size in each training step as is the case for the predict models.

As could be expected, we observed that using a randomized window size for training the count spaces decreased the speed at which performance of the spaces to predict semantic priming (Figure 2) data dropped with increasing window size. Nevertheless, the performance was still best at window size 3, even when a randomized window size was used. The improvement of using reduced window sizes was largest for the largest window sizes – for window size 10 the amount of explained variance increased by 0.7% (subtitle corpus) and 0.6% (concatenation of the corpora) in LDT and by 0.1% for naming (both subtitle corpus and the concatenation).

This analysis indicates that the random reduction of the window size attenuates the decreasing performance of the count models, making them more comparable to the predict models even for larger window sizes. However, the general trend of optimal performance with a window size of about 3 can still be observed.

Discussion

In this article we compared the performance of the recently proposed predict models of semantic similarity to the methods currently used in psycholinguistics by looking at how much variance the estimates explain in human performance data. In all cases, we saw an outcome that was at least equal to the existing measures and that was often superior to them. This was even true when we compared the measures based on semantic spaces to measures produced by human participants (e.g., word association norms or semantic features generated by participants), showing that the semantic vectors should be included in psycholinguistic research.

In line with previous findings (Baroni et al., 2014; Levy & Goldberg, 2014), the predict models were generally superior to the count models, although the best count models tended to come quite close to the predict models (and in a few cases even exceeded them). The most important variable for the count models was window size, as shown by Bullinaria and Levy (2007, 2012). A problem in this respect, unfortunately, is that the optimal window size seems to depend on the task. It equals 3 for semantic priming, 1 for semantic relatedness judgments, and 2 for the prediction of word associations. The performance rapidly drops for non-optimal window sizes, as shown in Figures 2-4. At the same time, our additional analysis indicated that applying the same procedure of randomly selecting window sizes, as done in

the predict models, may be a way to attenuate the decrease in performance for larger window sizes.

In contrast, the predict models are less influenced by window size. In addition, their performance generally increases with window size (certainly up to 5). Of these models, the CBOW models typically outperformed the skip-gram models and there are no indications in the data we looked at to prefer the latter over the former. In general, there was little gain when the dimensions of the CBOW model exceeded 300 (sometimes performance even started to decrease; this was particularly true for semantic priming and word associations).

Given the superior performance of the CBOW models, it is important to understand the mechanisms underlying them. As a practical example of the CBOW model, we discuss the model that had the best average performance for English and that we also recommend for general use in psycholingusitic research (see also the section on availability below). This model is trained on the combined UKWAC and subtitle corpus, has a window size of 6, and contains 300 dimensions. There are input and output nodes for each word form in the corpus encountered at least 5 times, leading to about 904 thousand input and output nodes. The dimensionality of the model is equal to the number of hidden nodes, which in this case is 300. The training of the model consists of the activation of the input nodes of the 6 words before the target word and the 6 words after the target word and predicting the activation of the output node corresponding to the target node. Over successive runs, the weights are adapted to improve performance. The semantic vector for a word consists of the 300 weights between the input node of a word and the hidden nodes after learning.

As shown in Figure 1, the CBOW model learns to predict the relationship between the target word and all words in the surrounding window simultaneously. In the HAL-type count model and in the skip-gram model, the relationship between the target word and each word in

the window is trained individually. As a metaphor, consider a paper with a long set of co-authors of which one has been removed. The task is to predict the missing author. The HAL-type count model and the skip-gram model can only predict the missing author based on the individual co-occurrence between each known co-author and their past co-authors, which could result in the predicted co-author being completely unrelated to the other co-authors on the paper. The CBOW model, on the other hand, would predict the missing author based on the simultaneous consideration of all other co-authors on the paper. The model would be more likely to predict a co-author who often writes together with all or part of the co-authors than someone who frequently co-authors with only one of them.

In light of the current findings, it is important to understand the differences between the discussed models in Marr's (1982) terms. The count model specifies a computational problem for the cognitive system (learning to associate semantically related words) and provides an abstract computational method for solving it using weighting schemes and dimensionality reduction. It has been argued (Levy & Goldberg, 2014) that the results of the skip-gram model can also be achieved by a certain type of a count model (PMI weighting shifted by a constant and dimensionality reduction steps) making the skip-gram model computationally equivalent to a count model. However, because the skip-gram model can be specified using prediction-based incremental learning principles in a neural network, it solves the computational problem posed by the count models in a way that is to a large extent psychologically plausible. Finally, although the CBOW model shares this algorithmic-level plausibility with the skip-gram model, CBOW cannot be reduced to a count model (Levy et al., 2015). Since the CBOW model compares favorably to the other investigated models it is an important task for future research to better understand this model at the computational level.

In this paper, we gave considerable attention to the type of corpus used to train a model. In computational linguistics, models are often found to perform best when trained on very large corpora (Banko & Brill, 2001) and this implies that register is second to size. Our data show that the large corpora typically used in computational linguistics are good for vocabulary tests, such as TOEFL but perform less well for psycholinguistic benchmarks such as semantic priming or word associate generation. On these tasks, corpora based on subtitles of films and television series perform better. When we consider what the TOEFL test requires, it is not surprising that training on very large corpora containing a large amount of specialist material is beneficial. Because TOEFL includes a large number of uncommon words, models trained on subtitle corpora can be expected to perform worse on this test. Indeed, we would expect a person reading the material included in the very large corpus to score quite highly on the TOEFL and we would be equally unsurprised if a person watching only films and television series would perform worse. In contrast, the impressive performance of the relatively small corpora of subtitles on the semantic priming and word association tasks is surprising. This implies that when it comes to accounting for human behavior it is important to train models on a corpus that has a register closer to what humans experience. Recall that the TOEFL benchmark is not about predicting how well humans do, but about scoring as highly as possible. Associations in the larger corpus better reflect the semantic system for someone who scores very well on the TOEFL, whereas associations based on the subtitle corpus reflect more of a central tendency. As an example, our reference CBOW model based on subtitles generates the following words as nearest semantic neighbors for *elephant: giraffe*, *tusk*, *zoo,* and *hippo.* On the other hand, the model trained on the combined UKWAC and subtitles corpus generates *howdah, tusked, rhinoceros,* and *mahout.* The first and second authors of this paper confess that they did not know what to make of two

of the latter associations until they learned that a *howdah* is a seat for riding on the back of an elephant and that a *mahout* is a professional elephant rider. The example clearly illustrates how the models based on the larger corpora score higher on the TOEFL. Future research could investigate whether the advantage of the larger corpora is still maintained when the actual human responses are the benchmark instead of the highest score.

On the basis of the current study, conclusions about the relationship between corpus register and size and human performance are risky because these variables were not independent of each other. Still, it seems possible to conclude that given that the subtitle corpora are smaller *and* in many cases perform better on predicting semantic priming, the register of the subtitles better represents the input of human participants. On the other hand, the question remains what precisely in the bigger corpora accounts for the worse performance. Even adding subtitle material to the large UKWAC corpus did not result in models that predicted semantic priming much better than the subtitle corpus alone (and a similar finding was observed in Dutch). An answer may be that the smaller subtitle corpus results in close semantic relationships that are shared by many participants, while the large corpus results in more specialized semantic relationships that are known by only a few participants. This additionally suggest that increasing the size of a subtitle corpus further may not necessarily result in better performance on a semantic priming task because more specialized semantic relationships could be developed at the expense of more universally shared ones. This point is given further weight by taking into account that corpora over a certain size stop being ecologically realistic.[13]

[13] Assuming a maximum reading rate of 300 words per minute (Carver, 1989; Lewandowski, Codding, Kleinmann, & Tucker, 2003), a person who has read 16 hours per day for 18 years, has come across 300*60*16*365.25*18 = 1.89 billion words at most.

Given the current set of results, we can unequivocally assert that distributional semantics can successfully explain semantic priming data, dispelling earlier claims (Hutchison et al., 2008). While Günther et al. (2015) found small effects for German, we obtain a strong and robust increase in the predictive power when the regression analysis includes semantic information derived from distributional semantics models. According to our analyses the predictions based on the semantic space models can match or exceed the ones based on human association datasets or feature norms. This is fortunate, because semantic similarity measures based on semantic spaces are available for many more words than human similarity or relatedness ratings and can be collected more easily for languages that do not yet have human ratings. In this regard, we should also point to recent advances in human data collection. For instance, collecting more than one response for each cue word in a word association task may lead to a more refined semantic network than the one we tested (De Deyne, Verheyen, Storms, 2015). It will be interesting to see how such a dataset compares to the semantic vectors we calculated.

Finally, it is of practical importance to mention that, at least for the semantic priming data, the pioneering LSA space available through a web-interface at the University of Colorado Boulder (1997) does not perform better than the reference semantic spaces we are releasing with the current paper. At the same time, it is surprising that the difference in performance is so small if we consider the size of the corpus (11 million words) on which the LSA space was based. The relative success of LSA based on the small TASA corpus suggests that books used in schools are another interesting source of input (arguably because it is a common denominator. These books and subjects have been read by most students).

Availability

A big obstacle to the widespread use of distributional semantics in psycholinguistics has been the gap between the producers and potential consumers of such spaces. Although several packages have been published that allow users to train various kinds of semantic spaces (e.g. S-Space, Jurgens & Stevens, 2010; DISSECT, Dinu, Pham, & Baroni, 2013; LMOSS, Recchia & Jones, 2009; HIDEX, Shaoul & Westburry, 2006), the large corpora and computational infrastructure as well as the technical know-how regarding training and evaluating semantic spaces is not available to many psycholinguists. Therefore, in order to encourage the exchange and use of semantic spaces trained by various research groups, we release a simple interface that can be used to measure relatedness between words on the basis of semantic spaces. Importantly, it can be used both as a standalone program and as a web-server that makes the semantic spaces available over the Internet. We believe that such an open-source contribution complements the existing ecosystem allowing researchers to train and explore semantic spaces (e.g. LSAfun; Günther, Dudschig, & Kaup, 2014). We encourage contribution from other researchers to the code base for our interface, which is hosted on a platform for sharing and collaborative development of programming projects.[14]

To make it as easy as possible for the authors of semantic spaces to work with our interface, two simple formats are used: the Character Separated Values (CSV) format and the matrix market format[15] that supports efficient representations of sparse matrices such as those created when training count models without dimensionality reduction.

We release a series of predict and count spaces for Dutch and English that were found to be consistently well performing in the present evaluations. Each of the spaces is released in

[14] The code is available at the address: http://crr.ugent.be/snaut/

[15] For more information about the matrix market format see: http://math.nist.gov/MatrixMarket/

a format compatible with our interface. The predict spaces can be also used with the LSAfun (Günther et al., 2014) interface.

In addition to the full semantic spaces for English and Dutch used for the present study we also make available smaller subspaces which may be very useful in many cases, as they can be explored using very limited computational resources. The smaller semantic spaces are based on two subset tokens from full space:

- a subset of the 150,000 most frequent words in each of the spaces
- a subset based on the lemmas found in the corpora

Information about how well each of the released semantic spaces performed on our evaluation tasks is shown in Tables 5 (for English) and 6 (for Dutch).

== INSERT TABLE 5. HERE ==

== INSERT TABLE 6. HERE ==

As semantic spaces can always be improved by finding superior methods or parameter settings, we know that the spaces that we trained can and will be outperformed by other spaces. Our interface fully encourages such developments.

Acknowledgement

This research was made possible by an Odysseus grant from the Government of Flanders.

References

Baayen, R. H., Milin, P., Filipovic Durdevic, D., Hendrix, P., & Marelli, M. (2011). An amorphous model for morphological processing in visual comprehension based on naive discriminative learning. *Psychological Review, 118,* 438-482.

Balota, D. A., Cortese, M. J., Sergent-Marshall, S. D., Spieler, D. H., & Yap, M. (2004). Visual Word Recognition of Single-Syllable Words. *Journal of Experimental Psychology: General*, *133*(2), 283–316. http://doi.org/10.1037/0096-3445.133.2.283

Balota, D. A., Yap, M. J., Hutchison, K. A., & Cortese, M. J. (2012). Megastudies: Large scale analysis of lexical processes. In J. S. Adelman (Ed.), *Visual Word Recognition Vol. 1: Models and Methods, Orthography and Phonology*. Hove, England: Psychology Press.

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.

Baroni, M., Dinu, G., & 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* (Vol. 1). Retrieved from http://clic.cimec.unitn.it/marco/publications/acl2014/baroni-etal-countpredict-acl2014.pdf

Banko, M., & Brill, E. (2001). Scaling to very very large corpora for natural language disambiguation. In *Proceedings of the 39th Annual Meeting on Association for*

Computational Linguistics (pp. 26–33). Association for Computational Linguistics. Retrieved from http://dl.acm.org/citation.cfm?id=1073017

Broder, A. Z. (1997). On the resemblance and containment of documents. In *Compression and Complexity of Sequences 1997. Proceedings* (pp. 21–29). IEEE. Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=666900

Brysbaert, M., Buchmeier, M., Conrad, M., Jacobs, A.M., Bölte, J., & Böhl, A. (2011). The word frequency effect: A review of recent developments and implications for the choice of frequency estimates in German. *Experimental Psychology, 58,* 412-424.

Brysbaert, M., Keuleers, E., & New, B. (2011). Assessing the usefulness of Google Books' word frequencies for psycholinguistic research on word processing. *Frontiers in Psychology, 2*:27.

Brysbaert, M., & New, B. (2009). Moving beyond Kučera 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. http://doi.org/10.3758/BRM.41.4.977

Brysbaert, M., New, B., & Keuleers, E. (2012). Adding Part-of-Speech information to the SUBTLEX-US word frequencies. *Behavior Research Methods, 44*, 991-997.

Bullinaria, J. A., & Levy, J. P. (2007). Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior Research Methods*, *39*(3), 510–526.

Bullinaria, J. A., & Levy, J. P. (2012). Extracting semantic representations from word co-occurrence statistics: stop-lists, stemming, and SVD. *Behavior Research Methods*, *44*(3), 890–907.

Carver, R. P. (1989). Silent reading rates in grade equivalents. *Journal of Literacy Research, 21*(2), 155-166.

Chiarello et al. (1990). Semantic and associative priming in the cerebral hemispheres: some words do, some words don't ... sometimes, some places. - PubMed - NCBI. Retrieved April 27, 2015, from http://www.ncbi.nlm.nih.gov/pubmed/2302547

Church, K. W., & Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics*, *16*(1), 22–29.

Coltheart, M., Davelaar, E., Jonasson, J., & Besner, D. (1977). Access to the internal lexicon. In S. Dornic (Ed.), *Attention and performance VI* (pp. 535–555). Hillsdale, NJ: Erlbaum.

Cronbach, L. J., & Furby, L. (1970). How we should measure "change": Or should we?. *Psychological Bulletin, 74*(1), 68-80.

De Deyne, S., Verheyen, S., & Storms, G. (2015). The role of corpus size and syntax in deriving lexico-semantic representations for a wide range of concepts. *The Quarterly Journal of Experimental Psychology*, *68*(8), 1643–1664. http://doi.org/10.1080/17470218.2014.994098

Dinu, G., Pham, N., Baroni, M. (2013). DISSECT: Distributional semantics composition toolkit. *Proceedings of the system demonstrations of ACL 2013 (51st Annual Meeting of the Association for Computational Linguistics)* (pp. 31–36). East Stroudsburg, PA, ACL.

Drieghe, D., & Brysbaert, M. (2002). Strategic effects in associative priming with words, homophones, and pseudohomophones. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*(5), 951–961. http://doi.org/10.1037//0278-7393.28.5.951

Ferraresi, A., Zanchetta, E., Baroni, M., & Bernardini, S. (2008). Introducing and evaluating ukWaC, a very large web-derived corpus of English. In *Proceedings of the 4th Web as Corpus Workshop (WAC-4) Can we beat Google* (pp. 47–54). Retrieved from http://www.researchgate.net/profile/Adam_Kilgarriff/publication/237127354_Proceedings_of_the_4_th_Web_as_Corpus_Workshop_(WAC4)_Can_we_beat_Google/links/00b7d5290647fbc33f000000.pdf#page=53

Günther, F., Dudschig, C., & Kaup, B. (2014). LSAfun - An R package for computations based on Latent Semantic Analysis. *Behavior Research Methods*. http://doi.org/10.3758/s13428-014-0529-0

Günther, F., Dudschig, C., & Kaup, B. (2016). Latent semantic analysis cosines as a cognitive similarity measure: Evidence from priming studies. *The Quarterly Journal of Experimental Psychology*, *69*(4), 626–653. http://doi.org/10.1080/17470218.2015.1038280

Harris, Z. (1954). Distributional structure. *Word, 10*(2-3), 1456–1162.

Heyman, T., Van Rensbergen, B., Storms, G., Hutchison, K. A., & De Deyne, S. (2015). The influence of working memory load on semantic priming. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(3), 911–920. http://doi.org/10.1037/xlm0000050

Hinton, G. E. (2007). Learning multiple layers of representation. *Trends in Cognitive Sciences*, *11*(10), 428–434. http://doi.org/10.1016/j.tics.2007.09.004

Hinton, G. E., & Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *Science*, *313*(5786), 504–507. http://doi.org/10.1126/science.1127647

Hutchison, K. A., Balota, D. A., Cortese, M. J., & Watson, J. M. (2008). Predicting semantic priming at the item level. *The Quarterly Journal of Experimental Psychology*, *61*(7), 1036–1066. http://doi.org/10.1080/17470210701438111

Hutchison, K. A., Balota, D. A., Neely, J. H., Cortese, M. J., Cohen-Shikora, E. R., Tse, C.-S., … Buchanan, E. (2013). The semantic priming project. *Behavior Research Methods*, *45*(4), 1099–1114. http://doi.org/10.3758/s13428-012-0304-z

Jones, M. N., Kintsch, W., & Mewhort, D. J. K. (2006). High-dimensional semantic space accounts of priming. *Journal of Memory and Language*, *55*(4), 534–552. http://doi.org/10.1016/j.jml.2006.07.003

Juhasz, B. J., Lai, Y.-H., & Woodcock, M. L. (2015). A database of 629 English compound words: ratings of familiarity, lexeme meaning dominance, semantic transparency, age of acquisition, imageability, and sensory experience. *Behavior Research Methods*, *47*(4), 1004–1019. http://doi.org/10.3758/s13428-014-0523-6

Jurgens, D., & Stevens, K. (2010). The S-Space package: an open source package for word space models. In *Proceedings of the ACL 2010 System Demonstrations* (pp. 30–35). Association for Computational Linguistics. Retrieved from http://dl.acm.org/citation.cfm?id=1858939

Kass, R.E., & Raftery, A.E. (1995). Bayes Factors. *Journal of the American Statistical Associatio,* 430 (90), 773-395.

Keuleers, E., & Balota, D. A. (2015). Megastudies, Crowdsourcing, and Large Datasets in Psycholinguistics: An Overview Of Recent Developments. *The Quarterly Journal of Experimental Psychology*, *68*(8), 1457–68. http://doi.org/10.1080/17470218.2015.1051065

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. http://doi.org/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*. http://doi.org/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. http://doi.org/10.3758/s13428-011-0118-4

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review; Psychological Review*, *104*(2), 211.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. http://doi.org/10.1038/nature14539

Levy, O., Goldberg, Y., Dagan, I., & Ramat-Gan, I. (2015). Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, *3*. Retrieved from http://u.cs.biu.ac.il/~nlp/wp-content/uploads/Improving-Distributional-Similarity-TACL-2015.pdf

Lewandowski, L. J., Codding, R. S., Kleinmann, A. E., & Tucker, K. L. (2003). Assessment of reading rate in postsecondary students. Journal of Psychoeducational Assessment, 21(2), 134-144.

Liang, F., Paulo, R., Molina, G., Clyde, M. A., & Berger, J. O. (2008). Mixtures of *g* Priors for Bayesian Variable Selection. *Journal of the American Statistical Association*, *103*(481), 410–423. http://doi.org/10.1198/016214507000001337

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers*, *28*(2), 203–208.

Mandera, P., Keuleers, E., & Brysbaert, M. (2015). How useful are corpus-based methods for extrapolating psycholinguistic variables? *The Quarterly Journal of Experimental Psychology*, 1–20. http://doi.org/10.1080/17470218.2014.988735

Mandera, P., Keuleers, E., Wodniecka, Z., & Brysbaert, M. (2014). Subtlex-pl: subtitle-based word frequency estimates for Polish. *Behavior Research Methods*. http://doi.org/10.3758/s13428-014-0489-4

Marr, D. (1982), *Vision: A Computational Approach*, San Francisco: Freeman & Co.

McDonald, S., & Brew, C. (2004). A Distributional Model of Semantic Context Effects in Lexical Processing. In D. Scott, W. Daelemans, & M. A. Walker (Eds.), *ACL* (pp. 17–24). ACL. Retrieved from http://dblp.uni-trier.de/db/conf/acl/acl2004.html#McDonaldB04

McRae, K., Cree, G. S., Seidenberg, M. S., & McNorgan, C. (2005). Semantic feature production norms for a large set of living and nonliving things. *Behavior Research Methods*, *37*(4), 547–559.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013a). Efficient Estimation of Word Representations in Vector Space. *arXiv:1301.3781 [cs]*. Retrieved from http://arxiv.org/abs/1301.3781

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems* (pp. 3111–3119).

Miller, R.R., Barnet, R.C., Grahame, N. J. (1995). Assessment of the Rescorla-Wagner Model. *Psychological Bulletin, 117* (3), 363-386.

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (1998). The University of South Florida word association, rhyme, and word fragment norms. http://w3.usf.edu/FreeAssociation/.

New, B., Brysbaert, M., Veronis, J., & Pallier, C. (2007). The use of film subtitles to estimate word frequencies. *Applied Psycholinguistics*, *28*(04). http://doi.org/10.1017/S014271640707035X

Olshausen, B. A., & Field, D. J. (1996). Emergence of Simple-Cell Receptive Field Properties by Learning a Sparse Code for Natural Images. *Nature*, *381*, 607–609.

Padó, S., & Lapata, M. (2007). Dependency-based construction of semantic space models. *Computational Linguistics*, *33*(2), 161–199.

Patel et al. (1997). ESRLTC.ps. Retrieved April 27, 2015, from https://www.cs.bham.ac.uk/~jxb/PUBS/ESRLTC.ps

Recchia, G., & Jones, M. N. (2009). More data trumps smarter algorithms: Comparing pointwise mutual information with latent semantic analysis. *Behavior Research Methods*, *41*(3), 647–656. http://doi.org/10.3758/BRM.41.3.647

Recchia, G., & Louwerse, M. M. (2015). Reproducing affective norms with lexical co-occurrence statistics: Predicting valence, arousal, and dominance. *The Quarterly Journal of Experimental Psychology*, *68*(8), 1584–1598. http://doi.org/10.1080/17470218.2014.941296

Rehuřek, R., & Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks* (pp. 45–50). Valletta, Malta: ELRA.

Rouder, J. N., & Morey, R. D. (2012). Default Bayes Factors for Model Selection in Regression. *Multivariate Behavioral Research*, *47*(6), 877–903. http://doi.org/10.1080/00273171.2012.734737

Rumelhart, D. E., Hinton, G. E., & Williams, R. (1986). Learning representations by back-propagating errors. *Nature, 323*, 533–536. doi: 10.1038/323533a0

Shaoul, C., & Westbury, C. (2006). Word frequency effects in high-dimensional co-occurrence models: A new approach. *Behavior Research Methods*, *38*(2), 190–195.

Sutton, R. S., & Barto, A. G. (1981). Toward a modern theory of adaptive networks: expectation and prediction. *Psychological Review*, *88*(2), 135.

Testolin, A., Stoianov, I., Sperduti, A., & Zorzi, M. (2015). Learning Orthographic Structure With Sequential Generative Neural Networks. *Cognitive Science*. Advance online publication. http://doi.org/10.1111/cogs.12258

Zorzi, M., Testolin, A., & Stoianov, I. P. (2013). Modeling language and cognition with deep unsupervised learning: a tutorial overview. *Frontiers in Psychology*, *4*. http://doi.org/10.3389/fpsyg.2013.00515

Wagenmakers, E.-J. (2007). A practical solution to the pervasive provlems of p values. *Psychonomic Bulletin & Review*, 14 (5), 779-804.

Widrow, G., & Hoff, M. E. (1960). *Adaptive switching circuits.* Institute of Radio Engineers, Western Electronic Show and Convention, Convention Record, Part 4, 96-104.

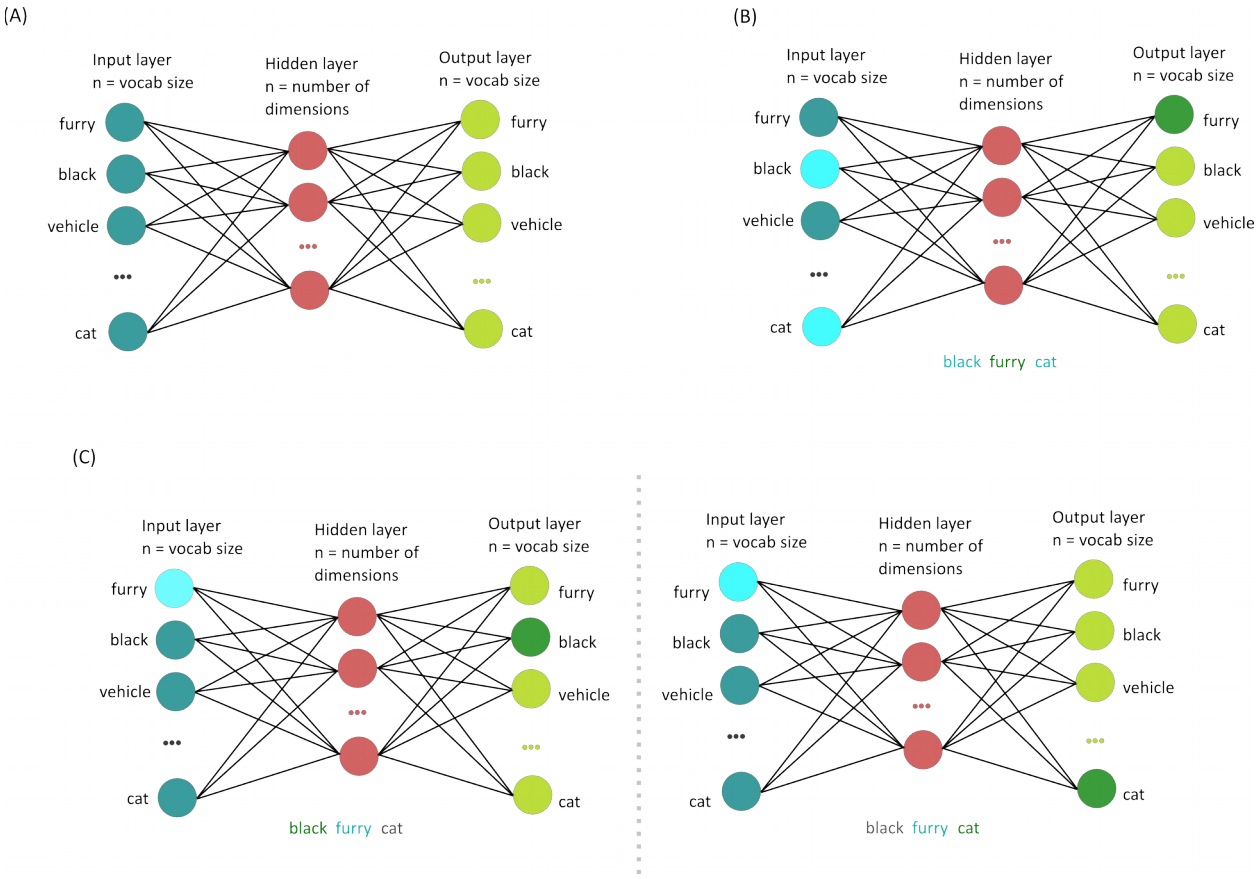

Figure 1. Both the CBOW and the skip-gram models are simple neural networks (*a*) consisting of an input, a hidden and an output layer. In the input and the output layers each node corresponds to a word. So, the number of nodes in these layers is equal to the total number of entries in the lexicon of the model. The number of nodes in the hidden layer is a parameter of the model. The training is performed by sliding a window through a corpus and adjusting the weights to better fit the training examples. When the model encounters a window including a phrase *black furry cat,* the CBOW model (*b*) represents the middle word *furry* by an activation of the corresponding node in the output layer and all context words (*black* and *cat*) are simultaneously activated in the input layer. Next, the weights are adjusted based on the prediction error. In the case of the skip-gram model (*c*) the association between each of the context words (*black* and *cat*) is predicted by the target word (*furry*) in a separate

learning step. When training is finished, the weights between the nodes and the input layer and the hidden nodes are exported as the resulting word vectors.

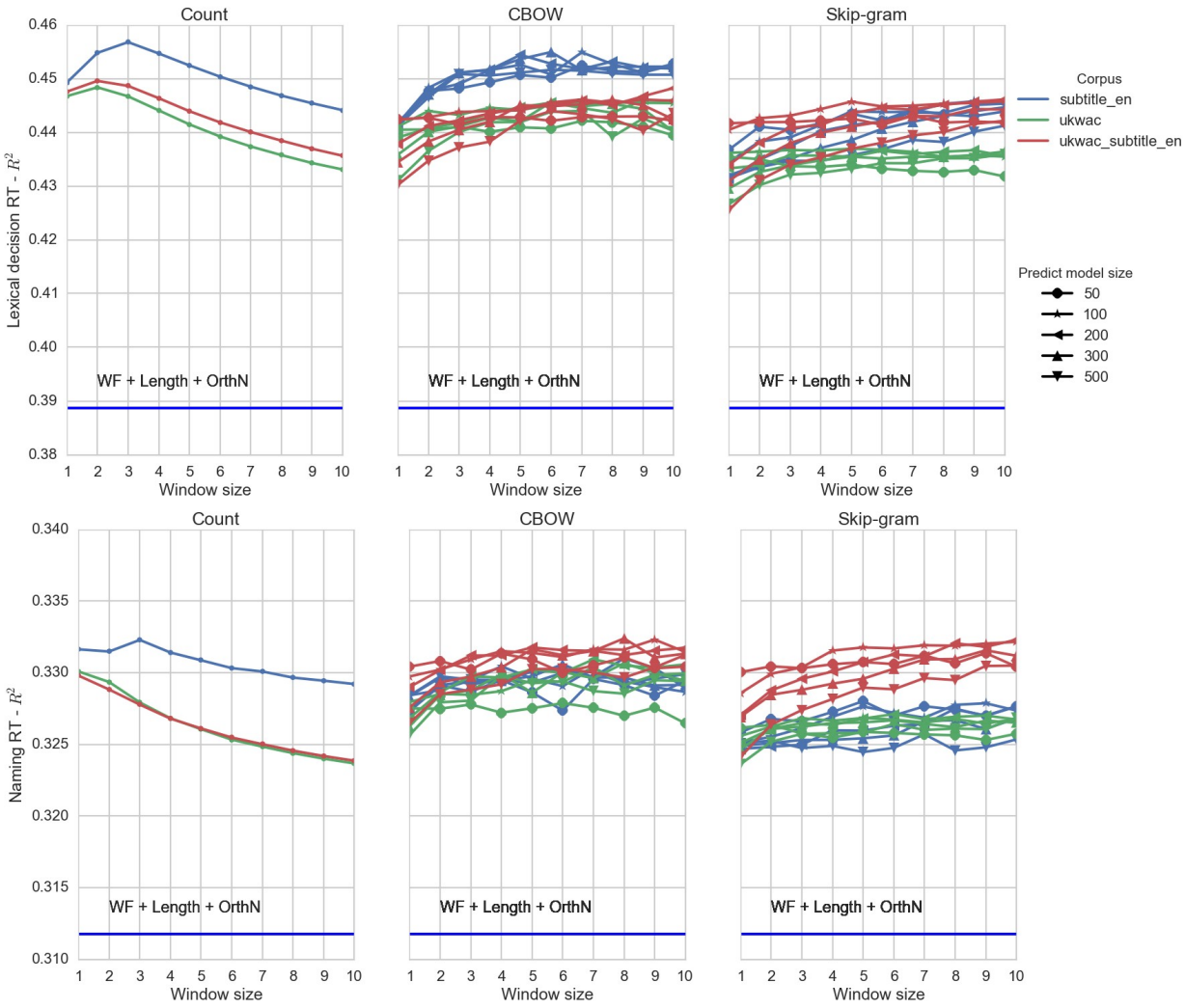

Figure 2. Performance of the three types of models on the Semantic Priming Project (Hutchison et al., 2013) dataset. The straight blue lines indicate the performance of the baseline model which did not include semantic predictors. Although the best count model in the LDT tasks performs slightly better than the best predict model (CBOW), its performance decreases rapidly with increasing window size. For naming, the predict models generally provide a better fit to the behavioral data. The models trained on the subtitle corpora or on the

concatenation of the subtitle corpus and the UKWAC corpus perform particularly well on these tasks.

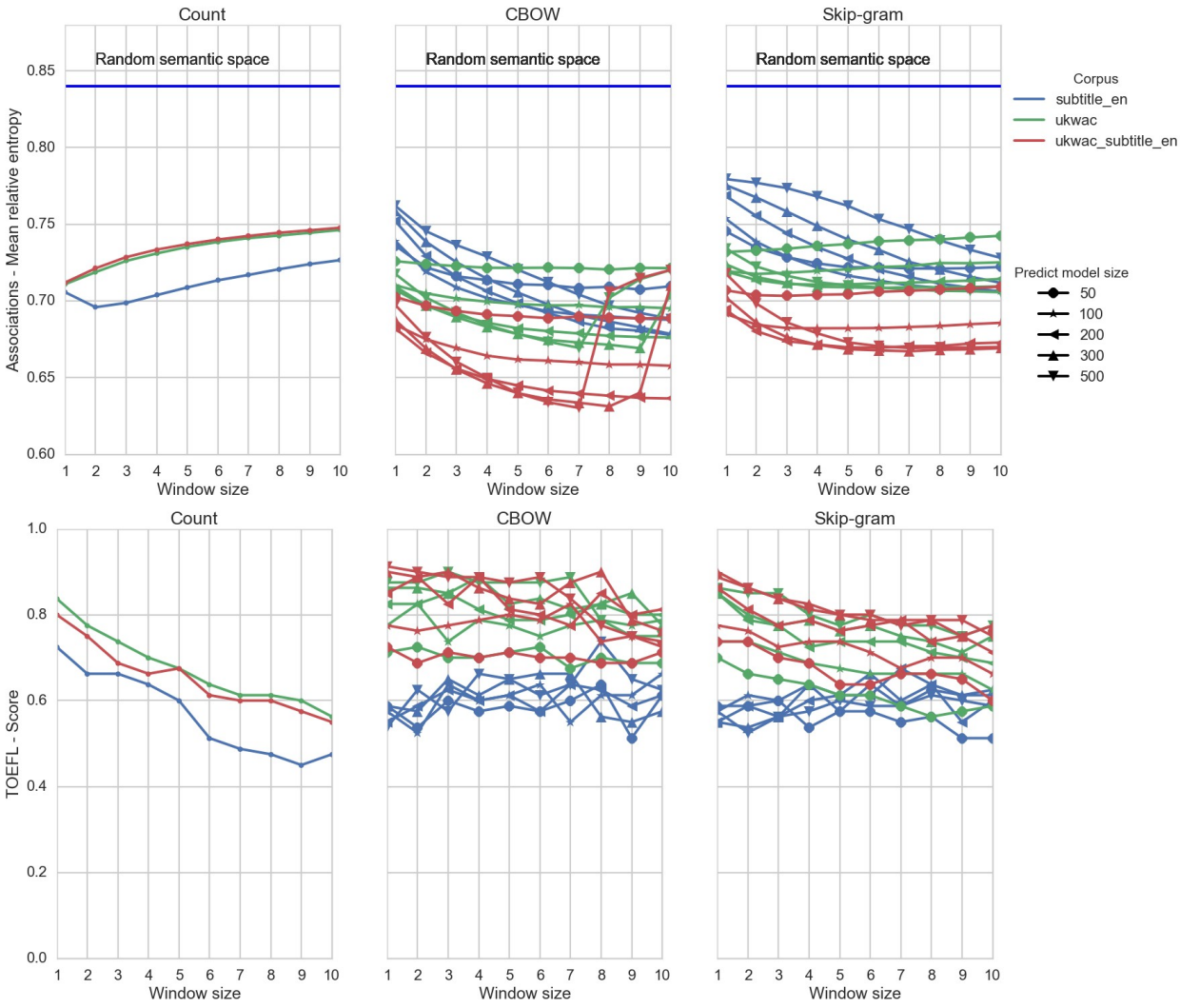

Figure 3. Performance of the three types of models on the association norms dataset (upper panels) and on TOEFL (lower panels). The predict models generally outperform the count models. Models trained on a subtitle corpus perform worse than the models trained on the UKWAC corpus or a concatenation of the two corpora. Note that for the association norms lower entropy is better.

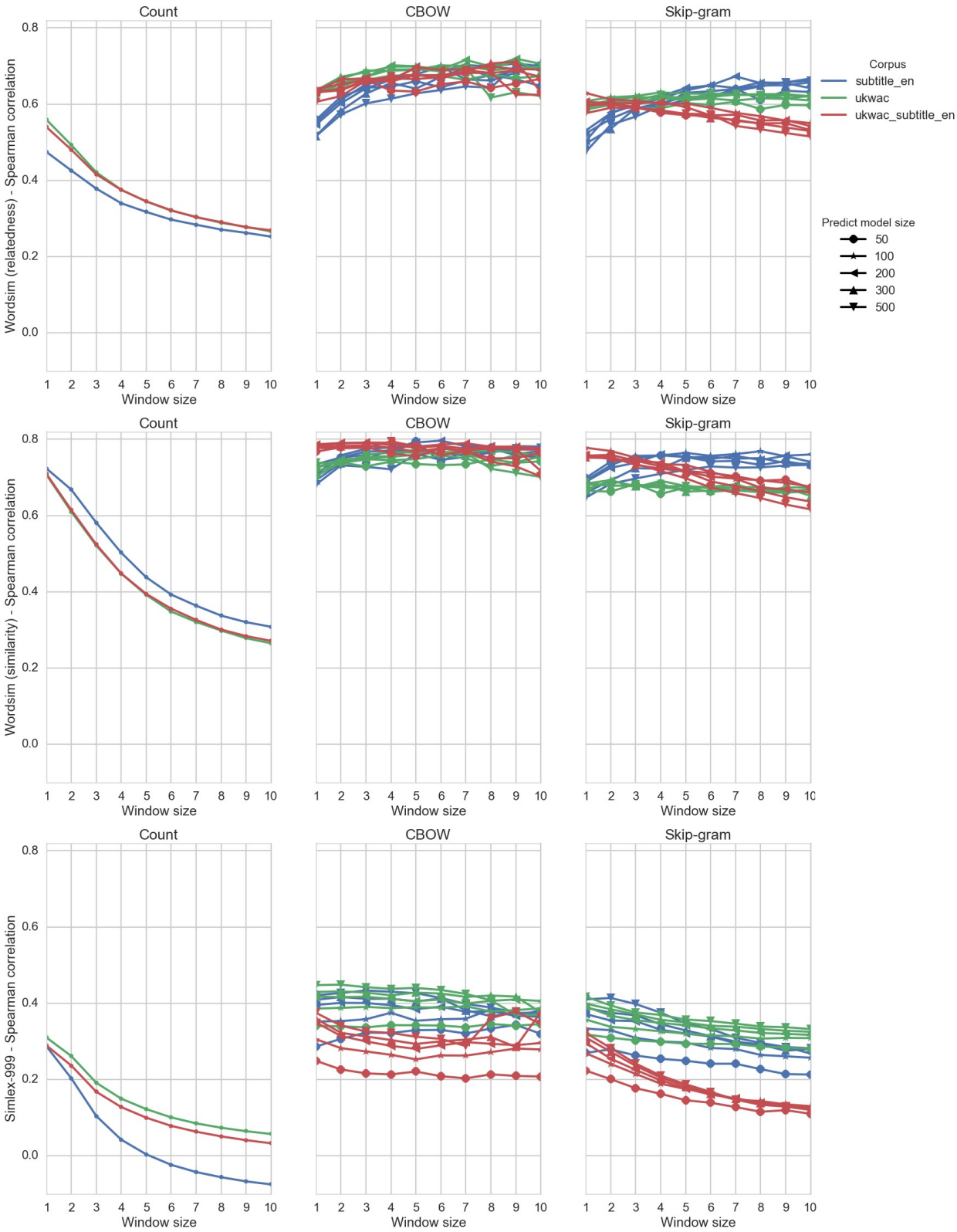
Count
CBOW
Skip-gram
Wordsim (relatedness) - Spearman correlation
Wordsim (similarity) - Spearman correlation
Simlex-999 - Spearman correlation
Window size
Corpus
subtitle_en
ukwac
ukwac_subtitle_en
Predict model size
50
100
200
300
500

Figure 4. Performance of the three types of models on the similarity and relatedness ratings datasets (absolute values of correlations). There is a robust advantage of the predict models.

Table 1. Results of the the analysis of the English semantic priming data. The table shows the average percentage of variance explained in lexical decision reaction times and naming latencies for different classes of models across all training procedures (the second column), the associated standard deviations (the third column) and performance (the fourth column), and the details of the training procedure (the fifth column) for the best model of each type.

	All models		Best model	
	Average R^2	**R^2 SD**	**R^2**	**Model training**
LDT				
count	44.5%	0.6%	45.7%	subtitle, window 3
CBOW	44.5%	0.5%	45.5%	subtitle, window 6, dim. 300
skip-gram	43.9%	0.5%	44.6%	UKWAC + subtitle, window 10, dim. 200
Naming				
count	32.8%	0.3%	33.2%	subtitle, window 3
CBOW	33.0%	0.1%	33.2%	UKWAC + subtitle, window 8, dim. 300
skip-gram	32.7%	0.2%	33.2%	UKWAC + subtitle, window 10, dim. 200

Table 2. Results of the Bayes factor analysis of the English semantic priming data. Bayes factors for the baseline models are reported with reference to the intercept-only model and, for the remaining models, with reference to the baseline model. In the baseline models only lexical variables but no semantic distance measures were considered. The worst and best relatedness measures included in the Bayesian analyses were selected separately for each task based on the R^2 in the previous analyses.

Model type	**Variables in the selected model**	**Bayes Factor**
LDT		
baseline (lexical only)	$WF_{target} + len_{target} + ON_{target}$	$BF_{10} = 2.15 \times 10^{605}$
lexical + worst relatedness	$WF_{target} + WF_{prime} + len_{target} + ON_{target} + rel_{worst}$	$BF_{1baseline} = 1.24 \times 10^{74}$
lexical + best relatedness	$WF_{target} + WF_{prime} + len_{target} + ON_{target} + rel_{best}$	$BF_{1baseline} = 2.10 \times 10^{144}$
lexical + multiple relatedness	$WF_{target} + WF_{prime} + len_{target} + ON_{target} + rel_{BEAGLE} + rel_{CBOW} + rel_{count}$	$BF_{1baseline} = 4.79 \times 10^{161}$
Naming		
baseline (lexical only)	$WF_{target} + WF_{prime} + len_{target} + len_{prime}$	$BF_{10} = 5.99 \times 10^{457}$
lexical + worst relatedness	$WF_{target} + WF_{prime} + len_{target} + len_{prime} + rel_{worst}$	$BF_{1baseline} = 1.72 \times 10^{20}$
lexical + best relatedness	$WF_{target} + WF_{prime} + len_{target} + len_{prime} + rel_{best}$	$BF_{1baseline} = 2.34 \times 10^{36}$
lexical + multiple relatedness	$WF_{target} + len_{target} + len_{prime} + rel_{CBOW} + rel_{count}$	$BF_{1baseline} = 5.50 \times 10^{41}$

Note. WF_{target} = log10 of the target word frequency; WF_{prime} = log10 of the prime word frequency; len_{target} = number of letters in the target word; len_{prime} = number of letters in the prime word; ON_{target} = orthographic neighborhood density of the target word; rel_{worst} = the worst relatedness measure; rel_{best} = the best relatedness measure; rel_{CBOW} = the best CBOW relatedness measure; rel_{count} = the best count measure; rel_{BEAGLE} = the relatedness measure based on BEAGLE. For the sake of computational efficiency, in the analyses including multiple relatedness measures we removed orthographic neighborhood density of the prime from the set of considered predictors.

Table 3. Results of the analysis of the English similarity and relatedness ratings. The table shows the average correlation between similarity and relatedness ratings and semantic relatedness measures derived from different classes of models across all training procedures (the second column), the associated standard deviations (the third column) and performance (the fourth column), and the details of the training procedure (the fifth column) for the best model of each type.

	All models		Best model	
	Average r	r SD	r	Model training
Wordsim-353 relatedness (n=238)				
count	-0.35	0.09	-0.55	UKWAC, window 1
CBOW	-0.66	0.04	-0.72	UKWAC, window 9, dim. 200
skip-gram	-0.59	0.04	-0.67	subtitle, window 7, dim. 200
Wordsim-353 similarity (n=196)				
count	-0.43	0.15	-0.72	subtitle, window 1
CBOW	-0.76	0.02	-0.8	subtitle, window 6, dim. 200
skip-gram	-0.7	0.04	-0.78	UKWAC + subtitle, window 1, dim. 100
Simlex-999 (n=998)				
count	-0.1	0.11	-0.31	UKWAC, window 1
CBOW	-0.35	0.06	-0.45	UKWAC, window 2, dim. 500
skip-gram	-0.27	0.08	-0.42	UKWAC, window 1, dim. 500

Table 4. Average results obtained from different classes of models for the words in different conditions in the two Dutch semantic priming experiments (Heyman et al., 2015; Drieghe & Brysbaert, 2002). The first column lists the corpora on which the models were trained. The second column shows the different types of models. The HAL-like count models using window sizes smaller and larger or equal to 5 are shown separately. Window sizes mattered less for the predict models so they are all reported together. The next column reports the percentage of variance explained in the dataset from Heyman et al. (2015). The following three columns display average effect sizes of comparisons between various conditions in the dataset from Drieghe & Brysbaert (2002). The last three columns report the mean and standard deviation of the semantic distances between cues and targets in each of the conditions. All statistics are averaged over all parameter settings used to train the models.

		Heyman et al.	Drieghe et al.					
			Cohen's d			Average distance		
Corpus	Model	R^2	Related vs Control 1	Related vs Control 2	Control 1 vs Control 2	Related	Control 1	Control 2
SONAR-500	HAL w. < 5	.148	.92	.98	.16	.91 (SD=.05)	.95 (SD=.02)	.95 (SD=.02)
	HAL w. >= 5	.119	.95	1	.15	.91 (SD=.05)	.95 (SD=.02)	.95 (SD=.02)
	CBOW	.191	.73	.73	-.03	.92 (SD=.06)	.95 (SD=.01)	.96 (SD=.01)
	skip-gram	.183	.56	.82	.45	.81 (SD=.09)	.85 (SD=.05)	.87 (SD=.04)
SONAR-500 + subtitle-nl	HAL w. < 5	.152	.55	.83	.48	.8 (SD=.1)	.84 (SD=.05)	.86 (SD=.04)
	HAL w. >= 5	.125	.43	.62	.4	.84 (SD=.09)	.87 (SD=.04)	.89 (SD=.03)
	CBOW	.207	1.34	1.25	-.12	.63 (SD=.16)	.85 (SD=.08)	.84 (SD=.1)
	skip-gram	.194	1.44	1.38	-.08	.57 (SD=.15)	.81 (SD=.08)	.81 (SD=.1)
subtitle-nl	HAL w. < 5	.140	1.36	1.15	-.38	.47 (SD=.19)	.77 (SD=.12)	.72 (SD=.16)
	HAL w. >= 5	.117	1.35	1.23	-.24	.48 (SD=.14)	.68 (SD=.08)	.66 (SD=.09)
	CBOW	.172	1.42	1.37	-.11	.38 (SD=.11)	.58 (SD=.08)	.57 (SD=.09)
	skip-gram	.153	1.34	1.09	-.34	.31 (SD=.13)	.52 (SD=.11)	.48 (SD=.14)

Table 5. Performance of the released English semantic spaces on the evaluation tasks.

		Semantic priming project												
			Lexical decision		Naming		Associations relative entropy	Simlex-999		Wordsim-353 relatedness		Wordsim-353 similarity		TOEFL
subset	model	N	R^2 baseline	R^2 model	R^2 baseline	R^2 model		N	r	N	r	N	r	score
lemmas	subtitle, CBOW, dim. 300, window 6	5311	.399	.465	.319	.337	.696	999	-.414	236	-.672	196	-.765	.559
top 150000	subtitle, CBOW, dim. 300, window 6	5738	.389	.455	.312	.331	.698	998	-.412	238	-.671	196	-.765	.663
full	subtitle, CBOW, dim. 300, window 6	5738	.389	.455	.312	.331	.698	999	-.414	238	-.671	196	-.765	.663
lemmas	subtitle, count, window 3	5311	.399	.471	.319	.339	.696	999	-.106	236	-.382	196	-.581	.494
top 150000	subtitle, count, window 3	5738	.389	.457	.312	.332	.699	998	-.104	238	-.378	196	-.581	.663
top 300000	subtitle, count, window 3	5738	.389	.457	.312	.332	.699	999	-.106	238	-.378	196	-.581	.659
lemmas	UKWAC + subtitle, CBOW, dim. 300, window 6	5311	.399	.454	.319	.338	.633	999	-.301	236	-.673	196	-.776	.666
top 150000	UKWAC + subtitle, CBOW, dim. 300, window 6	5738	.389	.445	.312	.331	.636	998	-.3	238	-.676	196	-.776	.834
full	UKWAC + subtitle, CBOW, dim. 300, window 6	5738	.389	.445	.312	.331	.636	999	-.301	238	-.676	196	-.776	.853
lemmas	UKWAC + subtitle, count, window 1	5311	.399	.458	.319	.336	.708	998	-.289	236	-.54	196	-.71	.628
top 150000	UKWAC + subtitle, count, window 1	5738	.389	.448	.312	.330	.712	998	-.289	238	-.54	196	-.71	.809
top 300000	UKWAC + subtitle, count, window 1	5738	.389	.448	.312	.330	.712	998	-.289	238	-.54	196	-.71	.828

Table 6. Performance of the released Dutch semantic spaces on the evaluation tasks. For the evaluation based on data from Heyman et al. (2015) all datasets included 236 prime-target pairs and the baseline model based on lexical predictors explained 6.44% of the variance in RTs. The only exception consisted of models based on the 150,000 most frequent words which included 264 prime-target pairs (baseline model explained variance: 6.22%). All models based on Drieghe et al. 2015 included 63 pairs of words.

		Heyman et al.				Drieghe et al.			
			Cohen's d			Average distance			Associations relative entropy
subset	model	R^2	Related vs Control 1	Related vs Control 2	Control 1 vs Control 2	Related	Control 1	Control 2	
lemmas	SONAR-500, count, window 3	.143	.832	.976	.334	.883 (SD=.062)	.925 (SD=.027)	.934 (SD=.024)	.766
top 150000	SONAR-500, count, window 3	.143	.832	.976	.334	.883 (SD=.062)	.925 (SD=.027)	.934 (SD=.024)	.764
top 300000	SONAR-500, count, window 3	.143	.832	.976	.334	.883 (SD=.062)	.925 (SD=.027)	.934 (SD=.024)	.765
lemmas	SONAR-500 + subtitle, count, window 2	.162	.991	1.03	.156	.905 (SD=.050)	.947 (SD=.017)	.950 (SD=.020)	.774
top 150000	SONAR-500 + subtitle, count, window 2	.162	.991	1.03	.156	.905 (SD=.050)	.947 (SD=.017)	.950 (SD=.020)	.772
top 300000	SONAR-500 + subtitle, count, window 2	.162	.991	1.03	.156	.905 (SD=.050)	.947 (SD=.017)	.950 (SD=.020)	.773
lemmas	SONAR-500 + subtitle, CBOW, dim. 200, window 10	.224	1.532	1.542	.044	.633 (SD=.149)	.904 (SD=.069)	.907 (SD=.068)	.745
top 150000	SONAR-500 + subtitle, CBOW, dim. 200, window 10	.222	1.532	1.542	.044	.633 (SD=.149)	.904 (SD=.069)	.907 (SD=.068)	.739
full	SONAR-500 + subtitle, CBOW, dim. 200, window 10	.224	1.532	1.542	.044	.633 (SD=.149)	.904 (SD=.069)	.907 (SD=.068)	.743