Modeling bilingual word associations as connected monolingual networks

Yevgen Matusevych, Amir Ardalan Kalantari Dehaghi and **Suzanne Stevenson**
Department of Computer Science, University of Toronto
yevgen@cs.toronto.edu
amirardalan.kalantaridehaghi@mail.utoronto.ca
suzanne@cs.toronto.edu

Abstract

Word associations are a common tool in research on the mental lexicon. Studies report that bilinguals produce different word associations in their non-native language than monolinguals, and propose at least three mechanisms responsible for this difference: bilinguals may rely on their native associations (through translation), on collocational patterns, and on the phonological similarity between words. In this paper, we first test the differences between monolingual and bilingual responses, showing that these differences are consistent and significant. Second, we present a computational model of bilingual word associations, implemented as a semantic network paired with a retrieval mechanism. Our model predicts bilingual word associations better than monolingual baselines, and translation is the main mechanism explaining its success, while collocational and phonological associations do not improve the model.

1 Introduction

In a free association task, participants are given a cue word (e.g., *apple*) and produce the first word that comes to their mind (e.g., *red* or *fruit*).[1] Free associations have been a common tool in the study of the mental lexicon because the observed pattern of associations can reflect the nature and strength of connections between words in semantic memory.

We focus on free associations as a means to better understand the structure and processing of the mental lexicon in bilinguals. Bilingual word associations have been studied for decades (see an overview in Meara, 2009). Despite a number of important findings, which we summarize in the following section, high-level conclusions about the association norms in bilinguals' non-native language are unclear – not only because of high variability in bilingual populations (DeKeyser, 2013), but also due to methodological factors (as explained by Boulton, 2003; Krzemińska-Adamek, 2014). Of specific concern for us is the lack of robust statistical analyses of the results. Many studies provide a selective qualitative analysis of the responses, and their findings can be inconsistent. In particular, it is unclear whether there are significant differences between native and non-native word associations (as compared, for example, to the instability of responses within a group of speakers over time).

We address this issue by providing a statistical analysis of the differences in English word association responses of Dutch[L1]–English[L2] bilinguals (collected by van Hell and de Groot, 1998) compared to English monolingual word association norms. After demonstrating a quantifiable difference between them, we then present the first computational model of bilingual word associations, which we use to investigate how the structure and processing of the bilingual lexicon could lead to the observed differences.

2 Related work

2.1 Non-native word associations

In general, non-native speakers' responses tend to differ from those of native speakers (e.g., Wolter, 2001; Zareva, 2007; Antón-Méndez and Gollan, 2010; Hui, 2011). Non-native speakers often produce responses that are translation equivalents of responses they would give in their native language (Meara, 1978) – in other words, L1 *mediates* their L2 responses (Nam, 2014). Such translations are produced more frequently when the cue word and

[1] In a so-called continued version of this task, participants give more than one response, but for consistency we always consider only the first response to each cue in this study.

its translation are cognates[2] (Taylor, 1976; van Hell and de Groot, 1998). Also, collocational responses (called 'syntagmatic'; e.g., *duty–free*, *opportunity–take*: Politzer, 1978; Riegel and Zivian, 1972) and phonological responses (*favor–flavor*: Meara, 1978; Namei, 2004) tend to be produced by non-native speakers more frequently than by monolinguals. Multiple examples of all these effects are well-documented, yet open questions remain regarding how systematic these differences are between bilinguals and monolinguals.

Van Hell and de Groot (1998, henceforth vHdG) carry out a free association experiment with Dutch–English bilinguals (i.e., native Dutch speakers who have been learning English). For us, their study is interesting in two respects. First, vHdG work with two similar groups of bilinguals and test one of the groups twice, which allows us to measure the consistency of responses between two groups of bilinguals, as well as within a single group. Second, large-scale monolingual association norms are available for both Dutch and English, which helps us both with our statistical analyses and in building a computational model. We use vHdG's data (1) to carry out a systematic comparison of monolingual and bilingual responses, and (2) to train and test a computational model that helps us predict whether the effects described above are systematic or not.

2.2 Existing computational models

Graph-based models (or semantic networks) have been widely used in research on semantic memory (see an overview by Beckage and Colunga, 2016). Despite their 'localist' approach in which a word is simply represented by a node (rather than using distributed representations), such models are a useful tool in the study of lexical access and acquisition. In particular, they have successfully replicated patterns of human verbal behavior in free word association (Enguix et al., 2014; Gruenenfelder et al., 2015), semantic fluency tasks (Abbott et al., 2015; Nematzadeh et al., 2016), lexical growth/acquisition (Stella et al., 2017; Bilson et al., 2015), assessment of semantic similarity (Jackson and Bolger, 2014; De Deyne et al., 2016), etc.

Naturally, a graph is only a static representation of the lexicon, although its structure presumably reflects lexical processing (Beckage and Colunga, 2016). To simulate the actual processing dynamics, various mechanisms have been proposed, such as spreading activation, random walk, entanglement, etc. (Galea et al., 2011; Zemla and Austerweil, 2017). In a spreading activation model, the activation starts at a given node and spreads across the graph over adjacent edges proportionally to edge weights (Anderson, 1983; Roelofs, 1992). Recently, De Deyne et al. (2016) used this approach on a free association graph to predict human similarity judgments for weakly-related concepts. We use a similar approach to model bilingual free associations in our computational model.

3 Data analysis

While vHdG explored various aspects of bilingual word associations, they did not compare the bilingual responses they collected to independent monolingual data. Here, we quantitatively compare vHdG's data against monolingual association norms, to see whether the non-native responses are indeed systematically different from those of native speakers. As vHdG argue, there is a lot of variability among bilinguals. Therefore, we need to compare the between-group differences (monolinguals vs. bilinguals) against within-group differences (two sets of bilinguals), to ensure that any between-group difference we find is due to more than the variation in responses among bilinguals.

3.1 Distance measures

Our goal is to compare two sets of responses to a particular cue word against each other. For this, we use two measures. The first is based on average precision, widely used in information retrieval. This measure treats one (unordered) set of responses as a gold standard and compares this set against another (ordered) set, considering only the top n responses. Because we are interested in measuring the *distance* between the two sets, we employ a complementary measure ρ to assess the distance between an unordered (shorter) set X and an ordered (longer) set Y:

$$\rho_n(X,Y) = 1 - \frac{\sum_{k=1}^{n} (P_k(X,Y) \times 1_k)}{|X \cap Y|} \quad (1)$$

where 1_k is an indicator function taking the value of 1 if $Y_k \in X$ and 0 otherwise, and P_k is the precision at k:

$$P_k(X,Y) = \frac{|X \cap Y_{1:k}|}{k} \quad (2)$$

[2] In literature on bilingualism, cognates are commonly defined as translations that have similar forms.

where $Y_{1:k}$ is the subset consisting of the first k responses in Y.

While average precision is frequently used in information retrieval, a shortcoming of this measure is that the order of responses in X does not matter. In practice, however, some of the responses can be several times more frequent than others. To account for this fact, we use a second measure, total variation distance υ, which considers two probability distributions X' and Y', associated with the likelihoods of responses in X and Y, respectively: e.g., $X' \sim \{\mathcal{L}(X_i), 1 \leq i \leq |X|\}$, where the likelihood is proportional to the response frequency in the human data (and later, to the association score in our model). The measure υ is then defined as:

$$\upsilon_n(X,Y) = \frac{1}{2} \sum_{r_i \in \{X \cup Y_{1:n}\}} \left| X'(r_i) - Y'_{1:n}(r_i) \right| \tag{3}$$

Sometimes response r_i does not appear in one of the lists; if, e.g., r_i is not in Y, we take $Y'(r_i) = 0$.

For both measures, we test two values of n: $n = 3$ to compare only the top three responses per cue in the data, and $n = |X|$ to compare the maximum possible number of responses per cue. Note that in the latter case, n varies per cue word, depending on the number of responses in X. We denote the respective measures as ρ_3, υ_3, and ρ_{max}, υ_{max}.

To focus on systematic differences between word associations and eliminate the noise from occasional responses and various word forms, in all the reported analysis we remove hapax legomena (responses that are only given by one participant) and lemmatize all the responses, using Frog (van den Bosch et al., 2007) for Dutch and NLTK WordNet lemmatizer (Bird et al., 2009) for English.

3.2 Same vs. different bilinguals

First, we test if our measures are sensitive enough to find expected differences between sets of free association responses. For this, we compare the difference in responses from two different sets of bilinguals to the difference in responses from a single set of bilinguals at two different times – i.e., we expect more variation in the two response sets in the former case than in the latter, in line with vHdG's results. We use their data, in which one group of bilinguals, B_1, performed the free association task twice ($B_{1\text{-}1}$ and $B_{1\text{-}2}$), while another group performed it only once (B_2). We

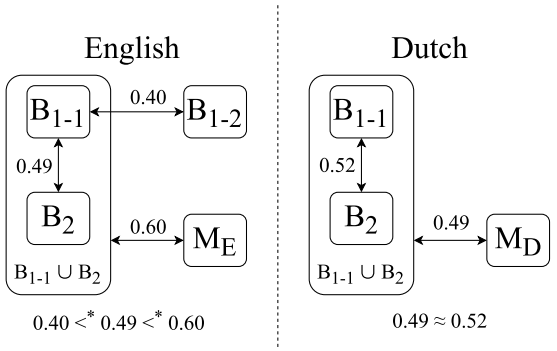

Figure 1: Distances (in terms of υ_3) between the responses given by different groups of participants.

then expect that $\rho_3(B_{1\text{-}1}, B_{1\text{-}2}) < \rho_3(B_{1\text{-}1}, B_2)$,[3] and the same for υ_3, ρ_{max}, and υ_{max}. We compute the ρ and υ values for responses given by vHdG's bilinguals to each of the 58 cue words.[4] Figure 1 (left panel) shows the distances in terms of υ_3 only (the differences in distances on the three other measures are *more* pronounced). We statistically compare the distances using Wilcoxon signed-rank test on pairwise differences per cue word. The results confirm our prediction on all measures: mean $\rho_3(B_{1\text{-}1}, B_{1\text{-}2}) = 0.35$ is less than mean $\rho_3(B_{1\text{-}1}, B_2) = 0.47$ ($p = .002$); for υ_3, the respective means are 0.40 and 0.49 ($p = .003$); for ρ_{max}, the means are 0.38 and 0.49 ($p = .004$); for υ_{max}, they are 0.35 and 0.46 ($p = .002$). The consistency of the observed differences across the four measures suggests that the same set of bilinguals gives more consistent responses across sessions than two different sets of bilinguals, and this effect cannot be explained by random variation. Ideally, we would carry out a similar analysis for monolingual speakers, but individual-level data for monolingual speakers is not available at the moment.

3.3 Bilinguals vs. English monolinguals

Given that our measures are sensitive to differences across response populations, we can now turn to our main goal of verifying differences in the responses of non-native speakers (that is, Dutch–English bilinguals tested in English) compared to native English speakers. We expect more consistency in the responses given by the two groups of bilinguals ($B_{1\text{-}1}$ vs. B_2), compared to bilinguals vs.

[3] Responses in the second session ($B_{1\text{-}2}$) may be biased, so we use $B_{1\text{-}1}$ in comparisons to B_2 here and to other sets below.

[4] For consistency, two cues that did not appear in English association norms were excluded from all analyses.

monolinguals ($B_{1\text{-}1} \cup B_2$ vs. M_E);[5] see Figure 1 (left panel). (For English monolingual responses M_E, we use the University of South Florida association norms: Nelson et al., 2004.) The results confirm our prediction: mean $\rho_3(B_{1\text{-}1}, B_2) = 0.47$ is less than mean $\rho_3(B_{1\text{-}1} \cup B_2, M_E) = 0.63$ ($p = .003$); for υ_3, the respective means are 0.49 and 0.60 ($p = .014$); for ρ_{max}, the means are 0.49 and 0.65 ($p = .002$); for υ_{max}, they are 0.46 and 0.60 ($p = .002$). In short, despite the high variation in bilinguals' responses, there is still significantly more consistency between groups of bilinguals than between monolinguals vs. bilinguals.

3.4 Bilinguals vs. Dutch monolinguals

Finally, we check whether the difference reported in the previous section is only observed in bilinguals' L2 (English), or is also found in their L1 (Dutch). Intuitively, we expect little difference between the responses of Dutch monolinguals and Dutch–English bilinguals tested in Dutch. In other words, there should be a similar degree of consistency in the responses given by, on the one hand, the two groups of bilinguals ($B_{1\text{-}1}$ vs. B_2), and on the other hand, by bilinguals vs. monolinguals ($B_{1\text{-}1} \cup B_2$ and M_D); see Figure 1 (right panel). (For Dutch monolingual responses M_D, we use the Dutch association norms from De Deyne et al., 2013, while the Dutch bilingual data is available from vHdG's experiment.) Statistical tests again confirm our predictions: mean $\rho_3(B_{1\text{-}1}, B_2) = 0.47$ is not different from mean $\rho_3(B_{1\text{-}1} \cup B_2, M) = 0.51$ ($p = .601$); for υ_3, the respective means are 0.52 and 0.49 ($p = .148$); for ρ_{max}, they are 0.50 and 0.56 ($p = .243$); for υ_{max}, they are 0.47 and 0.51 ($p = .625$).

To summarize our human data analyses, we have shown quantitatively that Dutch–English bilinguals give systematically different responses in English (their L2) from English monolinguals. While such a difference has long been observed, to our knowledge we are the first to statistically analyze this difference and show that it is greater than the inconsistency in responses across participants. Besides, this difference is specific to bilinguals' L2, as we did not observe it in bilinguals' L1 Dutch.

[5] We use $B_{1\text{-}1} \cup B_2$, as this combined data set provides more responses for the comparison; using $B_{1\text{-}1}$ or B_2 instead gives very similar results.

4 Computational model

We develop a computational model intended to investigate the difference found above between bilinguals and monolinguals in free association. Our hypothesis is that bilingual associations in L2 are influenced by their L1 through connections between the lexicons of their two languages. We create a bilingual Dutch–English semantic network as a weighted directed graph G with a set of nodes N, where N consists of cue and response words obtained from (monolingual) word association norms in the two languages: De Deyne et al. (2013) for Dutch and Nelson et al. (2004) for English.[6] We next describe the various types of edges connecting the nodes, and the spreading activation mechanism used as a retrieval mechanism.

4.1 Edge types and weights

Dutch and English associative edges, which connect nodes within the same language, effectively create two monolingual sub-networks.

L1 associative edges (DA) start at a Dutch cue word and end at all its Dutch responses, based on the monolingual Dutch association norms. The edge weights are proportional to conditional probabilities $p(response|cue)$ obtained from the norms.

L2 associative edges (EA) are created the same way, using the English association norms. The two resulting sub-networks are then connected to each other with two following types of edges.

Translation equivalent edges (TE) connect nodes that are translations of each other. Translations are obtained from two dictionaries: FreeDict[7] and dict.cc.[8] In many cases a node n has more than one translation (e.g., a and b). To determine which one is more frequent, we use OpenSubtitles,[9] a bilingual corpus of Dutch–English subtitles (Lison and Tiedemann, 2016). Word alignment was performed on a random sample of 50 million sentences using the method of Liang et al. (2006), and conditional probabilities of each Dutch–English and English–Dutch translation were extracted. If a and b are translations of node n, edges E_{na} and E_{nb} are weighted proportionally to the conditional probabilities $p(a|n)$ and $p(b|n)$.

Cognate edges (CG) are placed between translation equivalents that have similar orthographic

[6] All words were lemmatized, and hapax legomena and multiword responses were removed.

[7] http://freedict.org

[8] http://www.dict.cc

[9] http://www.opensubtitles.org

forms. Cognates are believed to enjoy a special status in bilinguals (van Hell and de Groot, 1998; Voga and Grainger, 2007). These edges are defined using a similarity measure S, which is complementary to the normalized Levenshtein distance (Ciobanu and Dinu, 2013). Given two words w_i and w_j, S is computed as:

$$S(w_i, w_j) = 1 - \frac{L(w_i, w_j)}{\max(|w_i|, |w_j|)} \quad (4)$$

where $L(w_i, w_j)$ is the Levenshtein distance between the words, and $|w|$ is the number of characters in w. We consider w_i and w_j to be cognates when they are translation equivalents in our dictionary, and $S(w_1, w_2) \geq 0.5$. This rather low threshold was chosen to capture cognates that are spelled differently due to morphological or etymological reasons, yet are similar in their pronunciation: *swell–zwellen*, *photography–fotografie*, etc.

Finally, we consider two extra types of edges, which connect English nodes to each other. As we mentioned earlier, there is some evidence that bilinguals tend to produce more orthographic and syntagmatic responses in their non-native language, and the following types of edges are intended to test whether this is a systematic effect.

Orthographic edges (OR) connect English words with similar spelling; they are weighted using the measure S defined above. We chose a higher threshold than for cognates, 0.75, to prevent the English network from becoming too dense. Here, for simplicity we assume that word spelling captures not only orthographic, but also phonological similarity between words, although in principle, phonological edges could be added as an independent type in the model.

Syntagmatic edges (SY) reflect collocations or pairs of words that frequently co-occur. Sometimes participants produce syntagmatic responses in the free association task, such as *duty–free*, *opportunity–take*, or *apple–red*. While our DA and EA edges capture such responses, there is some evidence that bilinguals produce more of these in their non-native language, so we add these SY edges. Specifically, we consider the most frequent bigrams and trigrams (one million each; from the Corpus of Contemporary American English: Davies, 2008), convert trigrams into skip-bigrams (*take _ opportunity*), and exclude stopwords (using the NLTK list: Bird et al., 2009) and words that do not appear in the English free association norms. For each

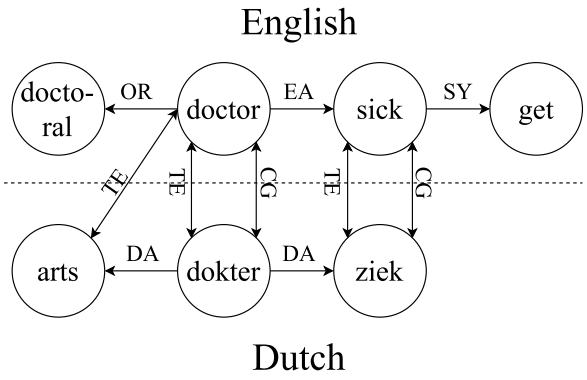

Figure 2: A part of the bilingual network.

pair of words, we compute their total number of co-occurrences in both bigrams and skip-bigrams, $F(w_1, w_2)$, and their total individual frequency, $F(w_1)$ and $F(w_2)$. Each weight for SY edge E_{ij} is set proportional to the respective conditional probability:

$$p(w_j|w_i) = \frac{F(w_i, w_j)}{F(w_i)} \quad (5)$$

Figure 2 shows a small part of the bilingual network with various types of edges.

4.2 Normalization of edge weights

We further weight each *type* of edges differently, to reflect their relative importance in the spreading activation process. These relative weights are the main parameters of our model. The model has six edge weight coefficients κ: κ_{DA}, κ_{EA}, κ_{TE}, κ_{CG}, κ_{OR}, and κ_{SY}, set as discussed in Section 5.2.

We normalize the edge weights of all outgoing edges of each node n to sum to 1, so that n passes on to its neighbors collectively the same amount of activation that it received. To do so, we first consider all outgoing edges of n a particular type – e.g., DA. We normalize the weights of all DA edges so that they sum to 1, and then multiply each weight by the respective coefficient, κ_{DA}. The same is done for all edge types. After that, we normalize the weights of *all* outgoing edges of n to sum to 1.

4.3 Retrieval algorithm

Given graph G with nodes N and edges E, the activation algorithm starts at a cue node n_{cue}, and activation spreads over edges to neighboring nodes, proportionally to the edge weights. This process is bounded in time by a parameter T, which is the upper limit of number of edges the activation can pass through. At the end, the model returns a ranked set of nodes (responses) $M = \{n_1, n_2, .., n_k\}$ and

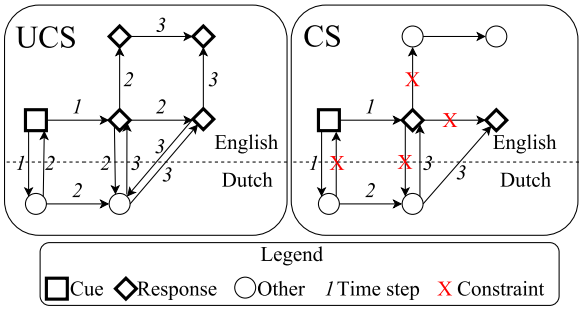

Figure 3: Spreading activation in the two models over a small part of bilingual network.

the respective likelihood value of each response, $\mathcal{L}(n_i)$:

$$\mathcal{L}(n_i) = \sum_{t=\{0..T\}} A^t(n_i) \quad (6)$$

where $A^t(n_i)$ is the activation score of n_i at time t:

$$A^t(n_i) = \sum_{n_j \in \{N \setminus n_i\}} A^{t-1}(n_j) w(E_{ji}) \quad (7)$$

where E_{ji} is the edge connecting n_j to n_i, and $w(E_{ji}) = 0$ if the two are not connected. Initially, $A^0(n_{cue}) = 1$; for all other nodes $A^0(n_i) = 0$.

5 Experimental setup

5.1 Task, models, and baselines

We test our model on the English free association task given to bilinguals in vHdG – i.e., Dutch–English speakers were given English cue words and asked to respond in English.

We consider two versions of spreading activation in the model, unconstrained and constrained (see Figure 3). In both versions, we set T – the maximum path length of spreading activation – to be 3, following the intuition that bilinguals may translate the English cue into Dutch (time $t = 1$), think of Dutch word associations ($t = 2$), and translate them back into English ($t = 3$).

In the **unconstrained version (UCS)** of the model, activation crosses *all* types of edges at each time step. Note that a T value of 3 enables activation to spread from the English to the Dutch subnetwork and back, but also allows activation to spread beyond the direct English associates. The next version of the model controls for this.

The **constrained version (CS)** simulates a bilingual who accesses direct English associates of the cue word, as well as English translations of direct Dutch associates of Dutch translations of the cue word. That is, they combine their direct English associations with direct Dutch associations. At $t = 1$, activation passes from the cue node to its English associates and to its Dutch translations, via EA and TE/CG edges, respectively. At time $t = 2$, activation passes only from the just-activated Dutch nodes via DA edges to their Dutch associates. Finally, at $t = 3$, activation passes only from the newly activated Dutch nodes (the associates of cue translations) via TE and CG edges back to English nodes. Conceptually, this version implements a speaker who relies on the word translation mechanism.

Because we have shown that human bilingual responses to the English free association task differ from those of monolinguals, we need to compare our model's performance to a monolingual (English) baseline. The **association norms baseline (BASE-AN)** corresponds to the English word association data set itself: i.e., we use EA edges only in the English subnetwork and set the maximum path length $T = 1$. An improvement over BASE-AN ensures that our model is producing a better match to bilingual data than simply outputting English monolingual associations. We also use a second monolingual baseline with the same subnetwork and edges; this **spreading activation baseline (BASE-SA)** instead uses $T = 3$, as in our model. This setting enables access to indirect English associations of the cue word (as in our model), but only through English connections (unlike our model). Comparing our model to BASE-SA indicates any improvement we see in our model is due to accessing the Dutch subnetwork (our theoretical claim) and not simply due to making indirect associations in English.

5.2 Model evaluation

In the test task, the model receives a set of cue words and generates multiple responses to each cue. Only English nodes can serve as responses, and their probabilities are normalized to sum to 1. The model responses are compared to human data using the measures defined in Section 3.1.

Our main goal is to test which types of edges systematically contribute to predicting bilinguals' (non-native) free word associations, and which do not. We have six parameters of the model related to edge weights (κ weights for the six types of edges) and a relatively low number of test items (58 cue words). To prevent overfitting, we perform

Table 1: Distances between model and human responses (averaged per cue word and per iteration). Best performance for each measure is in bold.

	Avg. score			
	ρ_3	υ_3	ρ_{max}	υ_{max}
BASE-AN	0.63	0.60	0.65	0.59
BASE-SA	0.63	0.61	0.66	0.61
UCS	0.63	0.60	0.63	0.58
CS	**0.59**	**0.57**	**0.61**	**0.56**

cross-validation on our data set, initially fitting only some of the κ parameters. Specifically, we first determine the best weights for the word association edges (κ_{DA} and κ_{EA}, which are essential for the task) and for the cross-language edges (κ_{TE} and κ_{CG}, which ensure that activation can pass from English to Dutch and back). We later test whether adding other edge types (SY and OR) improves the model.

For cross-validation, we use the Monte-Carlo method with $10,000$ iterations: in each iteration, the 58 cue words are randomly split into 48 training items and 10 test items. For each training subsample, we consider values $\{0, 1, 5, 10, 20, 25\}$ for each edge weight ($\kappa_{\text{DA}}, \kappa_{\text{EA}}, \kappa_{\text{TE}}, \kappa_{\text{CG}}$), run a grid search to find the best combination, and choose the four combinations (one per evaluation measure) which minimize the distance between the human and the model responses. These combinations are then evaluated on the respective test sub-sample.

6 Results

6.1 Testing the basic model

Table 1 provides average cross-validation scores for the two baselines and the two models. Recall that our scores are *distances* from human data, so lower values are better. We see that BASE-AN is a stronger baseline than BASE-SA. The UCS model shows little to no improvement over the baselines, and we only consider the CS model henceforth. The CS model shows a noticeable improvement over the stronger BASE-AN baseline, of 0.03–0.04 in terms of absolute distances, an improvement of 5%–6%.

Although the best combinations of edge weights of the CS model differ per iteration, one of them appears much more frequently than the others, over $12,000$ times: $(\kappa_{\text{DA}}, \kappa_{\text{EA}}, \kappa_{\text{TE}}, \kappa_{\text{CG}}) = (10, 5, 20, 25)$. To determine whether this combination makes significantly better predictions than the baselines, we test it on the full data set with responses to 58 cue words and run a series of Wilcoxon signed-rank tests (one per measure). The results show that the model (average scores $\rho_3 = 0.57$, $\upsilon_3 = 0.56$, $\rho_{max} = 0.60$, $\upsilon_{max} = 0.55$) is significantly better than both baselines on all measures, apart from υ_3 when compared to BASE-AN.

The comparisons to the baselines show that the CS model, but not the UCS model, predicts bilingual responses better than simply using monolingual responses, and it does so by using edges that link translations across English and Dutch.

6.2 Testing the model with extra edges

Here we see if adding the further two types of edges – OR and SY – improves the model predictions. We use the CS model with the best parameter combination, $(\kappa_{\text{DA}}, \kappa_{\text{EA}}, \kappa_{\text{TE}}, \kappa_{\text{CG}}) = (10, 5, 20, 25)$. Again, we cross-validate the model, this time running a grid search to find the best weights of the extra edges only, κ_{OR} and κ_{SY}. We look for the most frequent parameter combinations. The combination of the best CS model without the extra edges – that is, $(\kappa_{\text{OR}}, \kappa_{\text{SY}}) = (0, 0)$ – is about as frequent as a particular combination with syntagmatic edges – $(\kappa_{\text{OR}}, \kappa_{\text{SY}}) = (0, 1)$, and both of these perform the same on the full data set. Thus, OR and SY do not improve the model's performance overall. We return to this issue in the discussion.

Note that both for the UCS and CS models, we start by first fitting the κ values for associative edges and cross-language (translation and cognate) edges, because the literature generally agrees that L2 speakers use the translation mechanism at least to some extent (e.g., Meara, 2009). The other two mechanisms – collocations and form similarity – are tested as *additions* to the model. Effectively, this makes our basic CS model implement the learner relying on word associations (DA and EA edges) and translation equivalence (TE and CG edges), but not on collocation patterns or orthographic similarity between L2 words. One could also design a model without cross-language edges – that is, relying on L2 word associations (EA edges) together with collocations and/or orthography (OR and/or SY edges), which we do not present in this study for the lack of space.

6.3 Best model and error analysis

Here we look in detail at the best CS model and provide an error analysis. (For simplicity, we consider the model without SY edges.) This model weights direct monolingual associations more in

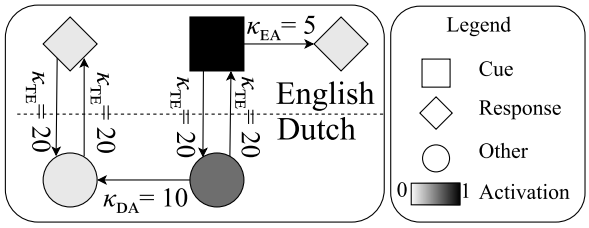

Figure 4: An illustration of the spreading activation in the best CS model (CG edges are not shown).

Dutch than in English: $\kappa_{DA} = 10$ vs. $\kappa_{EA} = 5$. Translation equivalents are also strongly connected to each other ($\kappa_{TE} = 20$), and cognates even more so ($\kappa_{CG} = 25$, which is *in addition* to the existing TE edge between them). This pattern of weights ensures that the translation operation is "cheap", and Dutch associates are readily activated; together these effectively make the contributions of English and Dutch associations similar in size. Figure 4 provides a toy example showing why this is the case. At the first step, a small share of the activation passes from the English cue to the English association, while the lion's share goes to the Dutch translation. At the second step, less than half of the activation at the Dutch translation proceeds to its associate; then in the third step, this activation is passed to the Dutch associate's translation.

This figure also shows why we cannot make conclusions about the contribution of a particular factor (e.g., translation equivalence, or the strength of English and Dutch word associations) based on the κ value of the corresponding edge type alone. Even though $\kappa_{TE} = 20$, $\kappa_{DA} = 10$, and $\kappa_{EA} = 5$, the contributions of native and non-native word associations to the final set of responses given by the model are similar, because English associations (the top right rhombus) are connected directly to the cue word, and the activation reaches them immediately upon the presentation of the cue, while Dutch associations (the top left rhombus) are further away from the cue word, and activation gets more dispersed as it passes through the network.

Table 2 shows the performance of the best model (vs. BASE-AN) for the best and worst cue words. For the majority of these cues the model is better than the baseline. For eight of these (*apple, block, bottle, chance, memory, season, shame, shoulder*), the improvement is consistent across the four measures. While the baseline relies on English word associations only, the model benefits from considering Dutch associations. This is because many

Table 2: Cue words for which the absolute difference between CS and BASE-AN is higher than 0.25 on at least one measure.

	Improvement over BASE-AN			
Cue	ρ_3	υ_3	ρ_{max}	υ_{max}
apple	0.56	0.41	0.30	0.32
block	0.11	0.21	0.33	0.24
bottle	0.56	0.40	0.23	0.15
chance	0.33	0.13	0.25	0.12
farm	0.00	0.06	0.60	0.27
flower	0.22	0.26	0.04	0.00
memory	0.17	0.34	0.03	0.16
season	0.83	0.55	0.63	0.52
shame	0.33	0.26	0.33	0.26
shoulder	0.67	0.34	0.29	0.20
attempt	−0.33	−0.03	−0.13	0.00
daughter	−0.33	−0.20	−0.33	−0.20
hospital	−0.33	−0.15	0.08	0.09
winter	0.00	−0.18	−0.25	−0.22

bilinguals' responses (e.g., *chance→possibility*, *shame→red*, *farm→farmer*) are missing in the monolingual data. In addition, some responses appear in the English monolingual data too, but are uncommon (e.g., *apple→pear, green*). In both cases, it is the translation edges that are responsible for the model's better performance.

Cue words on which the model is consistently worse than the baseline are *attempt, daughter*, and *winter*. For *hospital*, the model is only worse in predicting the top three responses. We find several reasons that may explain the model's errors.

Lack of data for some cues. The cue *attempt* is translated as *poging*, which activates a Dutch associate *probeersel* ['trial']. Because this word is not a cue in the Dutch association norms, all its activation is passed over its translation edges directly to *trial*, which yields relatively less activation for the more common response *suicide*.

Lack of word frequency information. For some cues (e.g., *hospital, winter*), the top human responses are words that are generally more frequent in English than are their Dutch translations (*nurse* vs. *verpleegster*, *spring* vs. *lente*).[10] In these cases, high frequency of English response words may lead speakers to rely more on English than on Dutch associations, which our model does not take

[10] As informed by relative word frequency information in English and Dutch subtitles (van Heuven et al., 2014; Brysbaert and New, 2009; Keuleers et al., 2010).

into account.

Language change. The data sets are not from the same time period (Dutch: 2010s; English: 1970s; bilingual: 1990s), so some responses that the model fails to reproduce may be attributed to language change: e.g., the response *duty→army* appears in the two older data sets, but not in the monolingual Dutch data, perhaps because conscription in the Netherlands was suspended in 1997.

7 Conclusion

We first showed that Dutch–English bilinguals in their L2 English give responses different from those of English monolinguals, but their L1 Dutch responses are not significantly different from those of Dutch monolinguals. While related observations have been reported in the literature (Wolter, 2001; Zareva, 2007; Antón-Méndez and Gollan, 2010; Hui, 2011, etc.), here we use a set of 58 cue words to demonstrate that this difference is consistent and is significantly larger than the difference between responses given by two groups of bilinguals.

Next, we presented a computational model based on a graph constructed from two monolingual word association data sets that were connected with additional cross-language edges. Our model predicts bilingual responses better than the monolingual baselines. The edge weights in the best model suggest that the contribution of L1 and L2 word associations is approximately equal in a group of Dutch–English bilinguals, and that translation equivalents (and cognates even more so) are strongly connected in the bilingual lexicon (in line with the findings on bilingual lexical access: e.g., Kroll et al., 2006; Dimitropoulou et al., 2011). Bilinguals may often translate L2 cues into L1, generate L1 associations, and translate them back into L1. In contrast, syntagmatic and orthographic responses that have been reported (e.g., Meara, 1978; Namei, 2004; Politzer, 1978) are not useful on the data set we used. Our results also suggest that it is not the case that bilinguals simply activate a broader cluster of L2 words and sample from those.

Van Hell and de Groot (1998) showed that bilinguals' responses might depend on the type of the cue word (e.g., noun–verb, abstract–concrete, cognate–non-cognate). As we intended to test how consistently various types of responses are produced across multiple cue words, we did not adjust the weights depending on the word type (except for cognates). Future research will consider enriching our network with such semantic and syntactic properties, as well as word frequency information. Another fruitful direction is to consider how to learn the association weights themselves, from textual and/or perceptual input (e.g., Griffiths et al., 2007; Gruenenfelder et al., 2015; Nematzadeh et al., 2016), rather than building them in from human norms; this would enable us to more realistically model the emergence of the bilingual lexicon.

Acknowledgments

We are grateful to Janet van Hell for sharing with us the word association data collected in her experiment with bilingual speakers.

References

Joshua T. Abbott, Joseph L. Austerweil, and Thomas L. Griffiths. 2015. Random walks on semantic networks can resemble optimal foraging. *Psychological Review* 122(3):558–569.

John R. Anderson. 1983. A spreading activation theory of memory. *Journal of Verbal Learning and Verbal Behavior* 22(3):261–295.

Inés Antón-Méndez and Tamar H. Gollan. 2010. Not just semantics: Strong frequency and weak cognate effects on semantic association in bilinguals. *Memory & Cognition* 38(6):723–739.

Nicole M. Beckage and Eliana Colunga. 2016. Language networks as models of cognition: Understanding cognition through language. In A. Mehler, A. Lücking, S. Banisch, P. Blanchard, and B. Job, editors, *Towards a theoretical framework for analyzing complex linguistic networks*, Springer, Berlin, Germany, pages 3–28.

Samuel Bilson, Hanako Yoshida, Crystal D. Tran, Elizabeth A. Woods, and Thomas T. Hills. 2015. Semantic facilitation in bilingual first language acquisition. *Cognition* 140:122–134.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: Analyzing text with the natural language toolkit*. O'Reilly, Sebastopol, CA.

Alex Boulton. 2003. Transfer and translation in L2 word associations: Comparing learner data across languages. In J.-C. Bertin, editor, *24th GERAS conference: Transfert(s)*. GERAS. https://hal.archives-ouvertes.fr/hal-00114289.

Marc Brysbaert and Boris New. 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.

Alina M. Ciobanu and Liviu P. Dinu. 2013. A dictionary-based approach for evaluating orthographic methods in cognates identification. In R. Mitkov, editor, *Proceedings of the International Conference Recent Advances in Natural Language Processing (RANLP 2013)*. Association for Computational Linguistics, pages 141–147. http://www.aclweb.org/anthology/R13-1019.

Mark Davies. 2008. The corpus of Contemporary American English (COCA): 520 million words, 1990–present. https://corpus.byu.edu/coca/.

Simon De Deyne, Daniel J. Navarro, Amy Perfors, and Gert Storms. 2016. Structure at every scale: A semantic network account of the similarities between unrelated concepts. *Journal of Experimental Psychology: General* 145(9):1228–1254.

Simon De Deyne, Daniel J. Navarro, and Gert Storms. 2013. Better explanations of lexical and semantic cognition using networks derived from continued rather than single-word associations. *Behavior Research Methods* 45(2):480–498.

Robert M. DeKeyser. 2013. Age effects in second language learning: Stepping stones toward better understanding. *Language Learning* 63(s1):52–67.

Maria Dimitropoulou, Jon A. Duñabeitia, and Manuel Carreiras. 2011. Masked translation priming effects with low proficient bilinguals. *Memory & Cognition* 39(2):260–275.

Gemma B. Enguix, Reinhard Rapp, and Michael Zock. 2014. How well can a corpus-derived co-occurrence network simulate human associative behavior? In A. Lenci, M. Padró, T. Poibeau, and A. Villavicencio, editors, *Proceedings of the 5th workshop on Cognitive Aspects of Computational Language Learning (CogACLL 2014)*, Association for Computational Linguistics, pages 43–48. http://www.aclweb.org/anthology/W14-0509.

David Galea, Peter Bruza, Kirsty Kitto, Douglas Nelson, and Cathy McEvoy. 2011. Modelling the activation of words in human memory: The spreading activation, spooky-activation-at-a-distance and the entanglement models compared. In D. Song, M. Melucci, I. Frommholz, P. Zhang, L. Wang, and S. Arafat, editors, *Quantum Interaction: 5th International Symposium, QI 2011, Revised Selected Papers*, Springer, Berlin, Germany, pages 149–160.

Thomas L. Griffiths, Mark Steyvers, and Joshua B. Tenenbaum. 2007. Topics in semantic representation. *Psychological Review* 114(2):211–244.

Thomas M. Gruenenfelder, Gabriel Recchia, Tim Rubin, and Michael N. Jones. 2015. Graph-theoretic properties of networks based on word association norms: Implications for models of lexical semantic memory. *Cognitive Science* 40(6):1460–1495.

Li Hui. 2011. An investigation into the L2 mental lexicon of Chinese English learners by means of word association. *Chinese Journal of Applied Linguistics* 34(1):62–76.

Alice F. Jackson and Donald J. Bolger. 2014. Using a high-dimensional graph of semantic space to model relationships among words. *Frontiers in Psychology* 5. https://doi.org/10.3389/fpsyg.2014.00385.

Emmanuel Keuleers, Marc Brysbaert, and Boris New. 2010. SUBTLEX-NL: A new measure for Dutch word frequency based on film subtitles. *Behavior Research Methods* 42(3):643–650.

Judith F. Kroll, Susan C. Bobb, and Zofia Wodniecka. 2006. Language selectivity is the exception, not the rule: Arguments against a fixed locus of language selection in bilingual speech. *Bilingualism: Language and Cognition* 9(2):119–135.

Małgorzata Krzemińska-Adamek. 2014. Word association patterns in a second/foreign language – What do they tell us about the L2 mental lexicon? *Lublin Studies in Modern Languages and Literature* 38(1):141–153.

Percy Liang, Ben Taskar, and Dan Klein. 2006. Alignment by agreement. In R. C. Moore, J. Bilmes, J. Chu-Carroll, and M. Sanderson, editors, *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*. Association for Computational Linguistics, pages 104–111. http://aclweb.org/anthology/N/N06/N06-1014.pdf.

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, and S. Piperidis, editors, *Proceedings of Tenth International Conference on Language Resources and Evaluation*. European Language Resources Association. http://www.lrec-conf.org/proceedings/lrec2016/pdf/947_Paper.pdf.

Paul Meara. 1978. Learners' word associations in French. *Interlanguage Studies Bulletin* 3(2):192–211.

Paul Meara. 2009. *Connected words: Word associations and second language vocabulary acquisition*. John Benjamins Publishing Company, Amsterdam, the Netherlands.

Hyunjeong Nam. 2014. L1 mediation in L2 lexical access. *The Journal of Modern British & American Language & Literature* 32(3):39–65.

Shidrokh Namei. 2004. Bilingual lexical development: A Persian–Swedish word association study. *International Journal of Applied Linguistics* 14(3):363–388.

Douglas L. Nelson, Cathy L. McEvoy, and Thomas A. Schreiber. 2004. The University of South Florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, & Computers* 36(3):402–407.

Aida Nematzadeh, Filip Miscevic, and Suzanne Stevenson. 2016. Simple search algorithms on semantic networks learned from language use. In A. Papafragou, D. Grodner, D. Mirman, and J. C. Trueswell, editors, *Proceedings of the 38th Annual Conference of the Cognitive Science Society*. Cognitive Science Society, Austin, TX, pages 1313–1318.

Robert L. Politzer. 1978. Paradigmatic and syntagmatic associations of first year French students. In V. Honsa and M. J. Hardman de Bautista, editors, *Papers in linguistics and child language: Ruth Hirsch Weir memorial volume*, Mouton, The Hague, the Netherlands, pages 203–210.

Klaus F. Riegel and Irina W. M. Zivian. 1972. A study of inter- and intralingual associations in English and German. *Language Learning* 22(1):51–63.

Ardi Roelofs. 1992. A spreading-activation theory of lemma retrieval in speaking. *Cognition* 42(1–3):107–142.

Massimo Stella, Nicole M. Beckage, and Markus Brede. 2017. Multiplex lexical networks reveal patterns in early word acquisition in children. *Scientific Reports* 7. https://doi.org/10.1038/srep46730.

Insup Taylor. 1976. Similarity between French and English words – a factor to be considered in bilingual language behavior? *Journal of Psycholinguistic Research* 5(1):85–94.

Antal van den Bosch, Bertjan Busser, Sander Canisius, and Walter Daelemans. 2007. An efficient memory-based morphosyntactic tagger and parser for Dutch. In Frank V. Eynde, Peter Dirix, Ineke Schuurman, and Vincent Vandeghinste, editors, *Selected papers of the 17th Computational Linguistics in the Netherlands Meeting*, LOT, Utrecht, the Netherlands, pages 99–114.

Janet G. van Hell and Annette M. B. de Groot. 1998. Conceptual representation in bilingual memory: Effects of concreteness and cognate status in word association. *Bilingualism: Language and Cognition* 1(3):193–211.

Walter J. B. van Heuven, Pawel Mandera, Emmanuel Keuleers, and Marc Brysbaert. 2014. SUBTLEX-UK: A new and improved word frequency database for British English. *The Quarterly Journal of Experimental Psychology* 67(6):1176–1190.

Madeleine Voga and Jonathan Grainger. 2007. Cognate status and cross-script translation priming. *Memory & Cognition* 35(5):938–952.

Brent Wolter. 2001. Comparing the L1 and L2 mental lexicon. *Studies in Second Language Acquisition* 23(1):41–69.

Alla Zareva. 2007. Structure of the second language mental lexicon: How does it compare to native speakers' lexical organization? *Second Language Research* 23(2):123–153.

Jeffrey C. Zemla and Joseph L. Austerweil. 2017. Modeling semantic fluency data as search on a semantic network. In G. Gunzelmann, A. Howes, T. Tenbrink, and E. Davelaar, editors, *Proceedings of the 39th Annual Meeting of the Cognitive Science Society*, Cognitive Science Society, Austin, TX, pages 3646–3651.