Analyzing and modeling free word associations

Yevgen Matusevych
Department of Computer Science
University of Toronto
yevgen@cs.toronto.edu

Suzanne Stevenson
Department of Computer Science
University of Toronto
suzanne@cs.toronto.edu

Abstract

Human free association (FA) norms are believed to reflect the strength of links between words in the lexicon of an average speaker. Large-scale FA norms are commonly used as a data source both in psycholinguistics and in computational modeling. However, few studies aim to analyze FA norms themselves, and it is not known what are the most important factors that guide speakers' lexical choices in the FA task. Here, we first provide a statistical analysis of a large-scale data set of English FA norms. Second, we argue that such analysis can inform existing computational models of semantic memory, and present a case study with the topic model to support this claim. Based on our analysis, we provide the topic model with dictionary-based knowledge about word synonymy/antonymy, and demonstrate that the resulting model predicts human FA responses better than the topic model without this information.

Keywords: free association, semantic memory, statistical modeling, topic model, latent Dirichlet allocation.

Introduction

In a free association (henceforth FA) task, speakers are exposed to a cue word and produce the first response word that comes to their mind (e.g., *smile→happy, award→trophy*). A collection of responses to each cue given by a large group of speakers, together with the frequency of each response, constitutes *free association norms* (e.g., Nelson, McEvoy, & Schreiber, 2004). Such norms are believed to reflect the strength of the links between words in the lexicon of an average speaker, and studies in psycholinguistics often rely on this information to explain various cognitive processes related to lexical semantic memory, such as semantic priming, lexical retrieval, etc. (e.g., Hutchison, 2003; Carpenter, 2009).

In computational modeling research, FA norms are commonly used as ground truth data to evaluate models of semantic memory that induce semantic relations between words from a text corpus: LSA, BEAGLE, LDA topic model, word2vec, etc. (e.g., Nematzadeh, Meylan, & Griffiths, 2017; Gruenenfelder et al., 2016; Griffiths, Steyvers, & Tenenbaum, 2007). However, a full understanding of the successes and failures of such models on FA norms is lacking, because we we are still far from a complete understanding of FA norms: that is, which properties of the cue and/or the response word lead speakers to produce a particular response (but see Schulte im Walde et al., 2008; Clark, 1970). We believe this lack of understanding may be one of the reasons explaining the generally low fit between the human FA norms and the predictions made by existing models of semantic memory.

Our contribution in this study is two-fold: First, we provide a detailed quantitative analysis of a set of FA norms, considering a wide range of psycholinguistically motivated variables. This analysis highlights the important properties of the cue and/or the response word that influence speakers' lexical choices in this task. Second, we demonstrate how the results of this analysis can be used to improve existing models of human semantic memory, with a case study on the widely-used LDA topic model.

FA norms analysis

We work with the University of South Florida (USF) FA norms (Nelson et al., 2004), the largest English FA norms currently available. Before turning to our analysis of the variables that affect speakers' choices of responses in this task, we briefly review related analysis of such norms.

Related work

Nelson, Dyrdal, & Goodmon (2005) undertake a quantitative analysis to reveal predictive factors underlying the USF norms. Almost all of the factors they consider are inspired by a network representation of the FA norms themselves, in which the nodes are cue words, c, and response words, r, and directed edges are weighted with the probability of a response given a cue, $p(r|c)$. In a multiple regression, $p(r|c)$ is predicted by variables such as mediated association strength (the sum of probabilities of c's associates that link r to c within two associative steps), backward strength (the probability that r evokes c in FA norms), etc. As these predictor variables are computed based on the FA network, the same FA norms are effectively used for deriving both the response variable and the explanatory variables in regression. This type of analysis helps us understand the internal structure of the FA network, but does not relate this network to external factors that may explain why speakers choose certain responses over others.

To consider such external factors, other studies (Aitchison, 1987; Clark, 1970)) report on a number of patterns observed in FA norms, which reflect various types of relations between cue and response words. To briefly summarize these patterns, speakers are likely to produce responses that (1) are semantically related to the cue word (*umbrella–rain, touch–hand*); (2) can substitute the cue in some contexts – that is, introduce minimal semantic or morphological contrast with the cue (e.g., antonyms: *long–short*, synonyms: *hungry–starved*, members of the same morphological paradigm: *was–were, mine–yours*); (3) often co-occur with the cue in word combinations (*young–boy*), collocations (*get–along*), or idioms (*ham–eggs*); and/or (4) are similar to the cue in their orthographic or phonological form (*favor–flavor*). More recently, Schulte im Walde et al. (2008) provided a detailed analysis of this type for German FA norms, while several studies (e.g., Gruenenfelder et al., 2016, Chaudhari et al., 2011, Peirsman & Geeraerts, 2009) proposed and evaluated a number of formal models that predict FA norms based on cue–response co-occurrence in a corpus.

Finally, the FA task draws on lexical retrieval, the speed of which may depend on a word's frequency or age of acquisition (Juhasz, 2005), concreteness (Kroll & Merves, 1986), and semantic ambiguity (Eddington & Tokowicz, 2015). These properties of a response word (independently of a cue) may affect the probability of its production in the FA task.

To our knowledge, no studies have considered a wide range of variables known to affect speakers' lexical choices within a single statistical model predicting human FA norms. Such a model must consider, on the one hand, variables characterizing the relation between the cue and the response, and on the other hand, variables characterizing the response independently of the cue. In addition, it should ideally control for the independent characteristics of the cue words, as they may modulate the effects of the other two groups of variables. Our goal here is to fit such a statistical model to the FA norms.

Predictor variables

We use a multiple regression analysis to predict the probability, $p(r|c)$, of a response r given a cue c, by a number of external variables that have been shown to affect lexical access in general or FA choices in particular. We consider three groups of variables that capture: (1) relations between c and r, (2) independent characteristics of r, and (3) independent characteristics of c (control variables).

1. Characteristics of cue–response pair. These include 4 subgroups, corresponding to the 4 general patterns observed in the human FA data and enumerated in the previous section.

1a. Semantic relatedness can be captured by word co-occurrence in a broad context. If two words frequently co-occur in the same discourse unit – conversation, document, etc. – a link may develop between these words in memory. We use the TASA corpus ($\sim$ 15M words in $37,653$ educational documents: Landauer & Dumais, 1997), because it consists of well-defined discourse units (documents) and is commonly used for training models of semantic memory. We compute document-context pointwise mutual information based on the frequency of c and r's occurrence and the frequency of their co-occurrence in **d**ocuments: $\mathrm{PMI}_d(c,r) = \log_2 \frac{p(c,r)}{p(c)p(r)} = \log_2 \frac{N\,\mathrm{freq}(c,r)}{\mathrm{freq}(c)\,\mathrm{freq}(r)}$, where N is the total number of tokens in the corpus.

1b. Context substitutability, which reflects paradigmatic relations between words, can be operationalized in multiple ways (see, e.g., Van Rensbergen et al., 2015). We use (1) a binary variable of whether or not r is a synonym/antonym of c (using Thesaurus.com[1]); (2) a binary variable of whether c and r belong to the same part of speech (using the most probable POS tag assigned by spaCy[2]); (3) an absolute real-valued difference between the estimated concreteness scores of c and r (as provided in Brysbaert et al., 2014), indicating how well the words match in their level of concreteness.

1c. Word combinations can be captured by measures of c and r's co-occurrence in a narrow context window. A narrow window requires a larger corpus than TASA, and we use the English part ($\sim$ 1B tokens) of OpenSubtitles-2018,[3] a corpus of movie subtitles (Lison & Tiedemann, 2016), to compute a **w**ord-context measure $\mathrm{PMI}_w(c,r)$ based on the number of c and r's co-occurrences in bigrams and 1-skip-bigrams. Unlike our document-context measure $\mathrm{PMI}_d(c,r)$, intended to capture semantic relatedness, $\mathrm{PMI}_w(c,r)$ shows how likely c and r occur in a combination (or syntagmatic relation).

1d. Orthographic similarity is computed as one minus the normalized Levenshtein distance between c and r.

2. Characteristics of response word. These variables include: (a) r's log-frequency, extracted from OpenSubtitles-2018; (b) r's estimated age of acquisition (as provided in Kuperman et al., 2012); (c) r's number of meanings and (d) senses, extracted from the Wordsmyth online dictionary;[4] (e) r's estimated concreteness value (Brysbaert et al., 2014).

3. Characteristics of cue word. These are control variables, introduced to account for the fact that the effects of the variables above may be modulated by c's own characteristics. We consider the same set of 5 variables as in group (2) for r.

Analyses and results

Following the setup commonly adopted in studies on modeling semantic memory (e.g., Gruenenfelder et al., 2016; Griffiths et al., 2007), we only consider the top five human responses to each cue in the human FA norms. Also, we discard responses for which not all the variables are available. We fit a mixed-effects regression to the resulting data (4513 cues with $20,951$ responses), using the predictor variables, their two-way interactions, and a random intercept per cue.[5]

The regression results in Table 1 suggest that the three corpus-based variables – freq (r) as well as the word-context PMI_w (reflecting c and r's ability to combine) and the document-context PMI_d (reflecting c and r's semantic relatedness) – are the best independent predictors of $p(r|c)$. That is, responses that are frequent overall, or frequently co-occur with the cue word, are likely to be produced.

Other important predictors include the age of acquisition of r (words acquired earlier are preferred), synonymy/antonymy relations between c and r (responses that are synonyms or antonyms of the cue are preferred), and the difference in their concreteness scores (responses that match the cue in the degree of their concreteness are preferred). Other pre-

[1] `http://www.thesaurus.com`

[2] `https://spacy.io`

[3] `http://www.opensubtitles.org`

[4] `http://www.blairarmstrong.net/tools/excel_wordsmyth_words_nummeaning_numsenses_partsofspeechfreq.zip`

[5] In all models, the values of each predictor were (1) divided by its standard deviation to position all predictors on the same scale, and (2) centered around 0, to reduce multicollinearity. We control for variance inflation (VI) by removing interactions with VI factor ≥ 3 until VI < 3 for all predictors.

Table 1: Mixed-effects regression fitted to human FA data. Non-significant main predictors, small interactions ($|\beta| < 0.05$), and control variables (properties of c) are not shown.

Predictor	β	*SE*	*p*
$\mathrm{PMI}_w(c,r)$	0.19	0.01	$<.001^{***}$
freq (r)	0.14	0.01	$<.001^{***}$
$\mathrm{PMI}_d(c,r)$	0.13	0.01	$<.001^{***}$
syn/ant (c,r)	0.11	0.01	$<.001^{***}$
age (r)	−0.11	0.01	$<.001^{***}$
Δ concr (c,r)	−0.09	0.01	$<.001^{***}$
POS match (c,r)	0.04	0.01	$<.001^{***}$
senses (r)	−0.03	0.01	$<.001^{***}$
concr (r)	0.03	0.01	$<.001^{***}$
orth. sim (c,r)	0.01	0.01	$.050^{*}$
syn/ant (c,r) × freq (r)	0.06	0.01	$<.001^{***}$
$\mathrm{PMI}_d(c,r)$ × age (r)	−0.05	0.01	$<.001^{***}$

Goodness of fit: the fixed effects (as well as the fixed and random effects together) explain 13.3% of the data variance.

dictors have less explanatory power in the regression: there is a small preference for response words which have the same POS as the cue, which are less ambiguous (have fewer different senses), more concrete, and orthographically more similar to c. Finally, there are significant interactions terms: first, responses with high corpus frequency are even more likely to be produced if they are also synonyms/antonyms of the cue, and second, the positive effect of c and r's semantic relatedness (PMI_d) is reduced if r is acquired late in life.

The presented analysis gives us useful insights about the factors underlying human lexical choices in the FA task, which is our first goal in this study. But considering our second goal, related to models of semantic memory, we need to be able to compare the role of the factors in human FA norms to their role in the responses generated by a model. To do so, we need a way to measure the *relative importance* of each factor, so that their importance could be compared across the two data sets. One way to do so is to rank the predictors based on the values of their β-coefficients in the regression, but the presence of interactions terms makes this ranking biased: it is difficult to interpret to what extent the presence of a particular predictor in the regression affects the overall fit to the data.

Instead of relying on the β-coefficients, we follow the existing practice in using regression by random forests, a method known for its ability to implicitly capture interactions between predictors (e.g., Grömping, 2009). We fit a random forest with 1000 trees to the human FA data, using the same set of main predictors (no interactions), and compute the relative importance of each predictor by considering an increase in model's error when the data for that predictor is randomly permuted (Breiman, 2001; Liaw & Wiener, 2002).

The results (see Figure 1) are largely compatible with the predictors' relative effect sizes (β-coefficients) in the mixed-effects regression, with a few differences in the exact order.

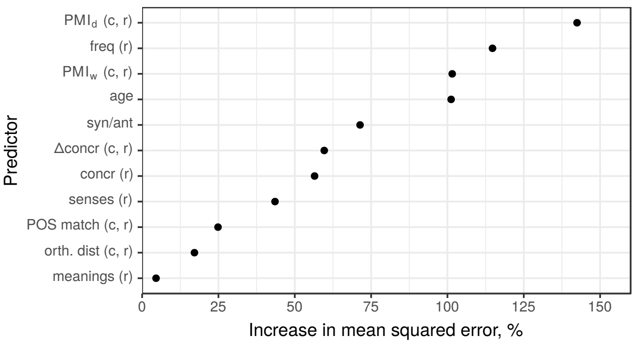

Figure 1: Relative importance of predictors in a random forest regression fitted to the human FA norms. Larger increase in error is associated with higher importance.

The document-context $PMI_d(c,r)$ is the most important predictor here, followed by three predictors of similar importance: r's frequency, r's age of acquisition, and the word-context $PMI_w(c,r)$. These are followed by three more predictors of similar importance: synonymy/antonymy relations between c and r, the difference in the degree of their concreteness, and r's own concreteness, while the other variables are less important.

To summarize, we have determined which factors are likely to drive human responses in the FA task, and next we test whether the same factors can explain the responses generated by a computational model of semantic memory.

Modeling FA norms with LDA topic model

Several studies (Nematzadeh et al., 2017; Gruenenfelder et al., 2016; Griffiths et al., 2007) compare a number of computational models of semantic memory, including LSA, BEAGLE, word2vec, GloVe, and the LDA topic model, in their ability to predict human FA norms. The results suggest that the topic model outperforms LSA and BEAGLE, and is at least as good as word2vec and GloVe when trained on the same amount of data. An advantage of the topic model over the latter two is its ability to capture asymmetric associative relations between words often observed in human FA norms: e.g., $p(penguin|bird) \ll p(bird|penguin)$. This motivates our use of the topic model in this study.

The LDA topic model is a generative model that takes a corpus (a collection of individual documents) as input and finds a set of topics, so that each document can be defined as a mixture of such topics. Each topic is a probability distribution over words, which makes it possible to compute a conditional probability of one word given another word, and to use this probability as an equivalent of the $p(r|c)$ in human FA norms (Griffiths et al., 2007). We use this model to generate a data set of responses and compare it to the human FA norms, both in terms of actual responses (as in earlier studies) and in terms of relative importance of predictors (as in the previous section). This latter comparison provides intuition about the strengths and weaknesses of the model.

Table 2: Mixed-effects regression fitted to the topic model data. Non-significant main predictors, small interactions ($|\beta| < 0.05$), and control variables are not shown.

Predictor	β	*SE*	*p*
PMI$_d$ (c, r)	0.22	<0.01	<.001***
PMI$_w$ (c, r)	0.18	<0.01	<.001***
freq (r)	0.14	0.01	<.001***
syn/ant (c, r)	0.05	<0.01	<.001***
Δ concr (c, r)	−0.05	<0.01	<.001***
orth. sim (c,r)	0.04	<0.01	<.001***
POS match (c, r)	0.03	<0.01	<.001***
concr (r)	0.02	<0.01	<.001***
meanings (r)	−0.01	<0.01	.002**
senses (r)	−0.03	0.01	.013*
syn/ant (c, r) × freq (r)	0.06	0.01	<.001***
PMI$_d$ (c, r) × age (r)	−0.05	<0.01	<.001***

Goodness of fit: the fixed effects alone explain 25.3% of the data variance, while the full model explains 42.3%.

Analyzing predictions of the topic model

Closely following procedure for training and testing adopted in the previous studies (Nematzadeh et al., 2017; Gruenenfelder et al., 2016; Griffiths et al., 2007),[6] we train an LDA topic model on the TASA corpus (Landauer & Dumais, 1997), use the trained model to generate, for each cue present in the FA norms, all potential responses and their probabilities, and consider all $p(r|c)$ values such that r appears in FA norms (either as a response to c or as a cue word).

We evaluate the model by comparing its predictions to the human FA norms, using two measures employed previously. (1) Following Griffiths et al. (2007), we compute $M = \{M_i, 1 \leq i \leq 5\}$, a set of median ranks, as follows: for each cue c, we find the rank of r_i – the i^{th} human response to c – in the list of the model's predicted responses to c. M_i is then the median of this set of ranks across all c. (2) Following Gruenenfelder et al. (2016), we compute P, the percentage of top 5 human responses that appear among the model's top 8 predictions, averaged over all cue words. Our replication yields results very similar to those reported previously:

$$\begin{aligned} M &= \{20, 57, 100, 125, 165\} \\ P &= 24\% \end{aligned} \tag{1}$$

That is, the median rank M_1 of the first human response in the model's predictions is 20, etc., and on average, 24% of the top 5 human responses are in the model's top 8 predictions.

While the two evaluation measures quantify the match between human FA norms and the model's predictions, our goal here is to further discover which high-level behavioral patterns the model can and cannot reproduce. For this, we apply the same regression analyses as in the previous section to the $p(r|c)$ values generated by the topic model. The topic model relies on a word's frequency and on the frequency of word co-occurrences in a document context (Griffiths et al., 2007), so that we expect freq(r) and $PMI_d(c,r)$ to be important predictors of the model's responses. At the same time, it is known that word co-occurrences in a corpus can implicitly encode various types of information (e.g., Louwerse & Zwaan, 2009), which the topic model may detect. In this case, other variables than freq(r) and $PMI_d(c,r)$ may also appear as important predictors of the model's $p(r|c)$ values. We fit the same two types of regressions (mixed-effects and random forests) to the predictions of the topic model.[7] The results (Table 2 and Figure 2A) show that freq(r) and $PMI_d(c,r)$ are, as expected, main predictors of the model's responses, and, as indicated by the relative importance plot, have disproportionately higher importance in this data set than in human FA norms. At the same time, the topic model certainly captures some word-context co-occurrence information: $PMI_w(c,r)$ appears as the third most important predictor in the plot (second in mixed-effects regression).

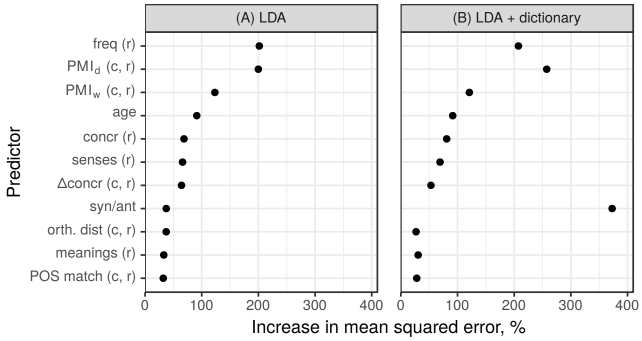

Figure 2: Relative importance of predictors in a random forest regression fitted to the data generated by (a) the topic model; (b) the dictionary-informed topic model.

In contrast, the synonymy/antonymy relations between c and r do not explain the model's responses as well as the human FA norms. This can be explained by the fact that the topic model treats each document as a bag-of-words and has no means to determine whether the two words can substitute each other in a phrase. So, for example, *sun* and *solar* might be as similar in the model as *sun* and *star* if these three words always occur in the same documents.

To summarize, our analysis reveals that, while the topic model is able to extract some high-level semantic information from textual data beyond word co-occurrence, its responses yield fewer synonyms or antonyms of the cue than humans do. We propose that the model's predictions can be improved by providing the model with this type of information.

[6] $T = 1700$ topics, $\alpha = 50/T$, $\beta = 0.01$, 3 sampling chains (MCMC) with 1600 iterations each; taking a sample every 100 iterations after 800 and averaging word similarities over all samples.

[7] Because some cue words are missing in the TASA corpus, the model data includes fewer cues than the human data: 4222. For consistency with our P measure, we consider the model's top 8 responses to each cue (and filter out the responses for which not all predictor variables are available): in total $30,070$.

Table 3: Mixed-effects regression fitted to the data of the dictionary-informed topic model. Non-significant main predictors, small interactions ($|\beta| < 0.05$), and control variables are not shown.

Predictor	β	*SE*	*p*
syn/ant (c, r)	0.42	<0.01	$<.001^{***}$
PMI$_d$ (c, r)	0.26	0.01	$<.001^{***}$
PMI$_w$ (c, r)	0.24	0.01	$<.001^{***}$
freq (r)	0.22	0.01	$<.001^{***}$
orth. sim (c,r)	0.04	<0.01	$<.001^{***}$
concr (r)	0.04	0.01	$<.001^{**}$
Δ concr (c, r)	−0.04	<0.01	$<.001^{***}$
POS match (c, r)	0.04	<0.01	$<.001^{***}$
syn/ant (c, r) × freq (r)	0.12	0.01	$<.001^{***}$
PMI$_w$ (c, r) × freq (r)	0.10	0.01	$<.001^{***}$
syn/ant (c, r) × PMI$_w$ (c, r)	0.08	<0.01	$<.001^{***}$
PMI$_d$ (c, r) × age (r)	−0.10	0.01	$<.001^{***}$
syn/ant × PMI$_d$ (c, r)	0.05	<0.01	$<.001^{***}$
PMI$_w$ × concr (c, r)	0.05	<0.01	$<.001^{***}$

Goodness of fit: the fixed effects alone explain 35.5% of the data variance, while the full model explains 44.0%.

Improving predictions of the topic model

We use the same setup as in the previous section to extract $p(r|c)$ values generated by the topic model, but then additionally upweight $p(r|c)$ by a constant k if r appears as a synonym or antonym of c in a dictionary (Thesaurus.com).[8] To ensure that our manipulation achieves the desired effect in the model, we again fit the same two types of regression to the $p(r|c)$ values generated by this enhanced model. The results in Table 3 and Figure 2(B) show that the synonymy/antonymy variable in this model is much higher in importance, as expected. The relative importance values of the other predictors are similar to those in the original topic model, although the value for PMI$_d$ is higher here, probably due to the interaction of this predictor with the synonymy/antonymy (see Table3).

We now need to test whether adding the synonymy/antonymy information actually improves the fit of the model's predictions to the human FA norms. We again apply the two evaluation measures, and we see substantial improvement (cf. Eqn. (1) showing the scores of the original model):

$$\begin{aligned} M &= \{12, 37, 67, 95, 123\} \\ P &= 28\% \end{aligned} \tag{2}$$

Here, the median rank of the first human response in the model's predictions is 12, vs. 20 in the original topic model. Also, P shows an error reduction of over 15% from the original model. Wilcoxon signed-rank tests with cue words as individual items show that for each M_i (where $1 \leq i \leq 5$) as well as for P, the improvement of the topic model provided with dictionary information over the original topic model is statistically significant: all $p < .001$ (Bonferroni-corrected).

In total, for 21% of cue words, the predictions of the dictionary-informed model match the human data better in terms of P measure; examples of corresponding cue–response pairs include *abstract→concrete, shoe→boot*, and *good→great*. At the same time, for 5% of cues, the dictionary-informed model is worse – cases when the correct predictions, such as *milk→drink, newspaper→article*, or *deep→sea*, are replaced in the model's top responses by synonyms/antonyms of the cue that humans do not produce.

Discussion

We provided the first statistical analysis of human FA norms that considers a wide range of external variables known to explain cognitive processes in human semantic memory. We also used the results of our analysis to reveal a weak point of the LDA topic model in modeling the human norms, and to guide the addition of information to the model that improves its fit to human data.

Our results show that corpus-based (distributional) variables are the most important predictors of human word choices in the FA task. These variables include co-occurrence of a cue and a response word in a document (reflecting semantic relatedness of words) and in a narrow context window (reflecting their syntagmatic relations), as well as the independent corpus frequency of candidate response words (presumably reflecting their cognitive entrenchment). Another important factor is the response word's age of acquisition, which explains some variance over and above the distributional variables, a result consistent with the independent effect of age observed in word and picture identification tasks (e.g., Juhasz, 2005). The final factors predictive of human FA responses – the synonymy/antonymy relations between a cue and a response word and the match in their degree of concreteness – have to do with the ability of the two words to substitute each other in phrases. In linguistic terms, speakers rely on, among other factors, so-called paradigmatic relations between candidate responses and the cue word.

Although we considered a wide range of variables, there are certainly other corpus-based measures of interest (e.g., ΔP, dispersion, PPMI: see an overview by Gries & Ellis, 2015) and important psycholinguistic variables (e.g., words' valence and arousal: Van Rensbergen et al., 2015). Also, each variable is likely to interact with the part of speech of the cue word (cf. Schulte im Walde et al., 2008). Future work needs to take these factors into account.

Our analysis of a particular model of semantic memory – the LDA topic model – shows that one of its weaknesses is the inability to capture paradigmatic relations between words, such as synonymy/antonymy. Rubin et al. (2014) show that such an ability naturally arises in a model if it either relies on a word-by-word matrix, or performs a singular value decomposition (SVD) with dimensionality reduction on a word-by-document matrix. The topic model instead constructs a word-by-document matrix and employs LDA for dimensionality reduction, and appears to not induce reliable paradigmatic re-

[8] We set $k = 10$, but any value $k > 1$ has a similar effect.

lations between words. This suggests that the differences between SVD and LDA (as dimensionality reduction methods), such as the amount of data variance encoded in top-*N* SVD dimensions vs. top-*N* LDA topics, may have implications for the model's ability to encode paradigmatic relations between words; this issue should also be addressed in future research.

In our case study, we showed that the fit between the predictions generated by the topic model and the human FA norms can be improved by providing the model with simple dictionary-based information capturing one type of paradigmatic (substitutability) relations between words – synonymy/antonymy. Based on this result, we propose that the topic model (as a model of semantic memory) can be improved by incorporating into its inference algorithm a mechanism that would capture such paradigmatic relations, in a manner similar to the integration of topical and syntactic information in the model of Griffiths et al. (2005). The success of our simple manipulation in this study – providing the model with readily-available dictionary information – demonstrates how a comparative analysis of high-level patterns in the human free association data vs. model predictions can inform existing models of semantic memory.

Acknowledgments: Supported by NSERC RGPIN-2017-06506. We thank Amir Ardalan Kalantari Dehaghi for his technical help with some of the predictor variables and the topic model.

References

Aitchison, J. (1987). *Words in the mind: An introduction to the mental lexicon*. Blackwell.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5–32.

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, *46*, 904–911.

Carpenter, S. K. (2009). Cue strength as a moderator of the testing effect: The benefits of elaborative retrieval. *JEP: Learning, Memory, and Cognition*, *35*, 1563–1569.

Chaudhari, D. L., Damani, O. P., & Laxman, S. (2011). Lexical co-occurrence, statistical significance, and word association. In *Proc. EMNLP–2011.* ACL.

Clark, H. H. (1970). Word associations and linguistic theory. *New Horizons in Linguistics*, *1*, 271–286.

Eddington, C. M., & Tokowicz, N. (2015). How meaning similarity influences ambiguous word processing: The current state of the literature. *Psych. Bulletin & Review*, *22*, 13–37.

Gries, S. T., & Ellis, N. C. (2015). Statistical measures for usage-based linguistics. *Language Learning*, *65*, 228–255.

Griffiths, T. L., Steyvers, M., Blei, D. M., & Tenenbaum, J. B. (2005). Integrating topics and syntax. In *Proc. NIPS–17.*

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological Review*, *114*, 211–244.

Grömping, U. (2009). Variable importance assessment in regression: Linear regression versus random forest. *The American Statistician*, *63*, 308–319.

Gruenenfelder, T. M., Recchia, G., Rubin, T., & Jones, M. N. (2016). Graph-theoretic properties of networks based on word association norms: Implications for models of lexical semantic memory. *Cognitive Science*, *40*, 1460–1495.

Hutchison, K. A. (2003). Is semantic priming due to association strength or feature overlap? A microanalytic review. *Psych. Bulletin & Review*, *10*, 785–813.

Juhasz, B. J. (2005). Age-of-acquisition effects in word and picture identification. *Psych. Bulletin*, *131*, 684–712.

Kroll, J. F., & Merves, J. S. (1986). Lexical access for concrete and abstract words. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *12*, 92–107.

Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods*, *44*, 978–990.

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*, *104*, 211–240.

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R News*, *2*, 18–22.

Lison, P., & Tiedemann, J. (2016). OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proc. LREC–2016.* ELRA.

Louwerse, M. M., & Zwaan, R. A. (2009). Language encodes geographical information. *Cognitive Science*, *33*, 51–73.

Nelson, D. L., Dyrdal, G. M., & Goodmon, L. B. (2005). What is preexisting strength? Predicting free association probabilities, similarity ratings, and cued recall probabilities. *Psych. Bulletin & Review*, *12*, 711–719.

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

Nematzadeh, A., Meylan, S. C., & Griffiths, T. L. (2017). Evaluating vector-space models of word representation, or, the unreasonable effectiveness of counting words near other words. In *Proc. CogSci–39.*

Peirsman, Y., & Geeraerts, D. (2009). Predicting strong associations on the basis of corpus data. In *Proc. EACL–12.*

Rubin, T. N., Kievit-Kylar, B., Willits, J. A., & Jones, M. N. (2014). Organizing the space and behavior of semantic models. In *Proc. CogSci–36.*

Schulte im Walde, S., Melinger, A., Roth, M., & Weber, A. (2008). An empirical characterisation of response types in German association norms. *Research on Language and Computation*, *6*, 205–238.

Van Rensbergen, B., Storms, G., & De Deyne, S. (2015). Examining assortativity in the mental lexicon: Evidence from word associations. *Psych. Bulletin & Review*, *22*(6), 1717–1724.