Toward a computational multidimensional lexical similarity measure for modeling word association tasks in psycholinguistics

Bruno Gaume[1], **Lydia-Mai Ho-Dac**[1], **Ludovic Tanguy**[1], **Cécile Fabre**[1], **Bénédicte Pierrejean**[1], **Nabil Hathout**[1], **Jérôme Farinas**[2], **Julien Pinquier**[2] **Lola Danet**[3], **Patrice Péran**[4], **Xavier De Boissezon**[3], **Mélanie Jucla**[5]

1 CLLE-ERSS: CNRS and University of Toulouse, Toulouse, France
bruno.gaume, lydia-mai.ho-dac, ludovic.tanguy, cecile.fabre, benedicte.pierrejean, nabil.hathout}@univ-tlse2.fr

2 IRIT: University of Toulouse and CNRS, Toulouse, France
{jerome.farinas, julien.pinquier}@irit.fr

3 CHU de Toulouse & ToNIC: University of Toulouse, Inserm, UPS, Toulouse, France
lola.danet@inserm.fr, deboissezon.xavier@chu-toulouse.fr

4 ToNIC: University of Toulouse, Inserm, UPS, Toulouse, France
patrice.peran@inserm.fr

5 URI Octogone-Lordat: Universit de Toulouse, Toulouse, France
melanie.jucla@univ-tlse2.fr

Abstract

This paper presents the first results of a multidisciplinary project, the "Evolex" project, gathering researchers in Psycholinguistics, Neuropsychology, Computer Science, Natural Language Processing and Linguistics. The Evolex project aims at proposing a new data-based inductive method for automatically characterising the relation between pairs of french words collected in psycholinguistics experiments on lexical access. This method takes advantage of several complementary computational measures of semantic similarity. We show that some measures are more correlated than others with the frequency of lexical associations, and that they also differ in the way they capture different semantic relations. This allows us to consider building a multidimensional lexical similarity to automate the classification of lexical associations.

1 Introduction

The Evolex project[1] brings together researchers in Psycholinguistics and Natural Language Processing (NLP) and focuses on lexical access and lexical relations by pursuing a threefold objective: (1) to propose a new computerised tool for assessing lexical access in population with or without language deficits; (2) to complement and reinforce the neuropsychological characterisation of lexical access using both qualitative and quantitative analyses; (3) to develop and train appropriated NLP tools to automatically measure and identify lexical relations. From a neuropsychology's perspective, assessing and characterising lexical access involves answering basic questions such as: How close two words can be in someone's mental lexicon? What are the nearest neighbours of a specific word? Are there more or less "typical" relations between words and do age (Burke and Peters, 1986), gender, sociodemographic status and language deficits (Péran et al., 2004) have an impact on those relations? The traditional method for tackling such issues is to use word association tasks where a participant has to produce a word in response to a stimulus, i.e. a word that is read out loud or written (*e.g.* answering *dog* after hearing the stimulus *cat*). The variables typically analysed are latencies, error rate, length of the response and its lexical frequency obtained from the analysis of large corpora (see for instance lexical frequency measures (New et al., 2004)). There are two main problems with such a method. First, we lack benchmarks about the typical answers produced by a large sample of participants and thus cannot reliably know whether a stimulus/response pair is more or less plausible for a large number of words (see for French norms Ferrand and Alario (1998) based on 300 words for young adults, de La Haye (2003) based on 200 words for children and young adults and Tarrago et al. (2005) based on 150 words for elderly people). Secondly, a qualitative subject-by-subject and item-by-item analysis is time consuming and prone to subjective interpretation. An answer to these challenges

[1]Evolex.1 was funded by the FHU HoPES (Federation for Cognitive, Psychiatric and Sensory Disabilities) of the Toulouse University Hospital (CHU de Toulouse).

is to obtain such data through the analysis of reference language data with NLP techniques. The use of data-based inductive methods for automatically measuring the similarity between words is one of the key task in computational semantics. If the first methods were based on the collocation frequency of words in large corpora (Church and Hanks, 1990; Evert, 2009), newer techniques rely on the principles of distributional semantics (Lenci, 2008; Mikolov et al., 2013). Even if the performance of these systems is impressive for some specific tasks (analogy resolution, lexical substitution, etc.), they usually fail to provide a fine grained characterisation of the relation between two words. Current distributional semantic models tend to aggregate all the classical lexical relations (e.g. synonymy, hypo/hypernymy, meronymy) and to confuse relations between similar words (e.g. *couch* - *sofa*) and relations between associated words (e.g. *couch* - *nap*). There is also a need for evaluation data when comparing and assessing these techniques (Hill et al., 2015; Baroni and Lenci, 2010). This paper proposes a step toward the satisfaction of both needs. We use data gathered in psycholinguistics experiments to compare different similarity measures and at the same time investigate how using complementary computational semantic techniques can help characterising lexical relations between stimuli and responses provided by subjects in a word association task. Section 2 describes the Evolex protocol from which data was collected as well as the manual annotation of the lexical relations in the collected dataset. We present the computational measures of semantic similarity in Section 3. Section 4 contains the quantitative analyses and results.

2 Data collection process in Evolex and qualitative analysis of dataset

The Evolex protocol includes different tasks to assess lexical access: a semantic fluency test (Benton, 1968), a phonemic fluency test (Newcombe, 1969), a classical Picture Naming task and a Word Association task. In addition to these 4 tasks, participants undergo 5 Cognitive Assessment Tests (MoCA, reading aloud, Trail Making Test, Digit Span, Stroop). This paper focuses mainly on the Word Association task which consists in vocalising the first word coming to mind after listening to a simple item (e.g. *fruit*, *painting*, *igloo*). The items used as audio stimuli were selected according to their grammatical category (nouns), number of syllables (same number of occurrences of words of 1, 2 and 3 syllables) as well as their frequency in generic corpora (as given by the Lexique resource, (New et al., 2004)). This paper exploits a first dataset of pairs of words collected from a pilot study with 60 stimuli and conducted with 30 participants presenting no language disorders, that are native French speaker aged between 15 and 58 (mean age 31 $\pm$ 13.06), with various levels of education (from 10 to 20 years of schooling, mean 15.4 $\pm$ 2.97). The following instructions were given to participants: *You will hear French common nouns. You will have to pronounce the first word which comes to your mind related to the one you just heard as fast as possible. For instance, when you hear TABLE, you may answer CHAIR.*

co-hyponym:balanoire(swing)/toboggan(slide)	73(13.1%)
hypernym:balancoire(swing)/jeu(game)	52 (9.3%)
meronym:balancoire(swing)/corde(rope)	49 (8.8%)
hyponym:animal(animal)/chat(cat	45 (8.1%)
holonym:doigt(finger)/main(hand)	29 (5.2%)
synonym:canap(couch)/sofa(sofa)	21 (3.8%)
antonym:aube(dawn)/crpuscule(dusk)	2 (0.4%)
classical relations:	**271(48.5%)**
associated:balancoire(swing)/enfant(child)	202(36.1%)
syntagmatic:fleur(flower)/peau(skin)	47 (8.4%)
none found:perroquet(parrot)/placard(closet)	28 (5.0%)
instance:magicien(wizard)/Merlin(Merlin)	6 (1.1%)
phonology:chapiteau(circus tent)/chateau(castle)	5 (0.9%)
non classical relations:	**288(51.5%)**

Table 1: Breakdown of the semantic relations used to categorise the 559 distinct stimulus-response word pairs.

After cleaning up and normalising the 1800 (60×30) individual collected responses, we obtained 559 distinct stimulus-response pairs. Independent double annotation was performed and followed by adjudication. The tagset is composed of 12 tags including 7 classical relations. Table 1 gives the number and % of distinct pairs annotated according to these 12 relations.

3 Computational measures of semantic similarity

In this section we describe the different techniques used in order to compute the similarity measures that we apply to the stimulus-response word pairs collected from the Word Association task. The six techniques we tested differ according to (1) the linguistic resources they used and (2) the use of either a first or second order similarity. Three resources reflecting three points of view on language were distinguished: a large corpus, giving access to word usage; a dictionary, reflecting expert point

of view on word meaning; crowdsourced lexical resource resulting from a GWAP (Game With A Purpose) proposing a Word Association Task very similar to ours that offers the advantage of having access to many more participants. The corpus used is *FrWaC* (Baroni et al., 2009), a collection of Web pages from the .fr domain and consisting of 1.6 billion words. The dictionary is the *TLF* (Trésor de la Langue Française, see (Dendien and JM., 2003)). The crowdsourced lexical resource is part of the GWAP *JDM* (Jeux De Mots [2]) where players have to find as many words as possible and as fast as possible in response to a term displayed on the screen, according to several instructions involving different type of lexical relations (semantic association, synonymy, etc., see (Lafourcade, 2007)). The potential atypicality of answers is partially controlled by the the game where two anonymous and asynchronous players earn points each time they give the same answer. If an answer is rarely given by other players it gets more points. Several instructions are proposed including a Word Association task ("As-W" task) very similar to ours with the following instruction: *"You are being asked to enumerate terms most closely associated with the target word... What does this word make you think about?"*. The three resources have been POS-tagged and lemmatised with the Talismane toolkit (Urieli, 2013). The second dimension on which these techniques contrasts opposes 1st order similarity (cooccurrences or direct relation between two words in the dictionary or the lexical relation) to 2nd order similarity, also known as distributional similarity, considering that words sharing first-order similar words show a possibly different degree of similarity. 2nd order similarity measures require more complex algorithms such as word embeddings for processing corpus similarity and random walk approach (Bollobas, 2002) for dictionary and lexical resources. Each measure is described in the next subsections.

3.1 Corpus-based similarity

FrWaC.1st similarity considers collocation: two words are considered similar if they frequently and systematically collocate in the FrWaC corpus. This measure has a large number of uses in NLP and corpus linguistics, and is known to capture a large variety of semantic relations (Evert, 2009; Wettler et al., 2005). We computed this similarity using Positive Pairwise Mutual Information (Evert, 2009). Each word was considered using its POS-tag and lemma, and its collocations were extracted in a symmetrical rectangular (unweighted) window of 3 words in both directions.

FrWac.2nd similarity relies on the principle of distributional semantics, which considers that words appearing in the same contexts have similar meanings. 2nd-order similarity can be computed in a number of ways (Baroni and Lenci, 2010; Baroni et al., 2009), and for a few years most of the work and research has focused on word embeddings. For this experiment, we used Word2vec (Mikolov et al., 2013) on the same FrWac corpus to obtain a dense matrix in which each word is represented by a numeric vector. The cosine distance was then computed to measure the similarity between two words. In the absence of benchmark test sets for French (while many exist for English, including BLESS that can be used to tune a model for specific semantic relations (Baroni and Lenci, 2011), we relied on the default parameters[3].

3.2 Dictionary-based similarity

TLF.1st similarity is based on the principle that two words are considered similar if one appears in the definition of the other. We computed this similarity by building an undirected and unweighted graph with words as vertices (V) and relations between words as edges (E). The TLF.1st measure relies on the graph $G_{TLF} = (V_{TLF}, E_{TLF})$ where $\forall x, y \in V_{TLF}, \{x, y\} \in E_{TLF}$ iff x appears in the TLF's definition of y or vice-versa (or both). This similarity measure is therefore binary: the similarity between x and y is 1 if x and y are neighbors in G_{TLF} and 0 otherwise.

TLF.2nd similarity used a graph traversal technique. We adopted a random walk approach (Bollobas, 2002) that is known to provide a broader and more "robust" measure of similarity between the nodes of a graph (Gaume et al., 2016). By applying this technique to the G_{TLF} graph, TLF.2nd corresponds to $P^t_{G_{TLF}}(x, y) \in [0, 1]$ i.e. the probability of a walker crossing the links of G_{TLF}, starting on vertex x, to reach the vertex y, after t steps. In this study, the length of the random walks is $t = 3$.

[2] http://www.lirmm.fr/jeuxdemots/jdm-accueil.php

[3] Skipgram algorithm with negative sampling (rate 5), window size 5, 500 dimensions, subsampling rate $10-3$, 5 iterations, minimum frequency 100

Similarity measure	Spearman's ρ	p-value
FrWac.1st	0.25	2.06e-09
FrWaC.2nd	0.22	6.86e-08
TLF.1st	0.23	3.44e-08
TLF.2nd	0.38	8.48e-21
JDM.1st	0.47	2.30e-32
JDM.2nd	0.51	1.44e-38

Table 2: Spearman correlation.

3.3 Crowdsourced resource-based similarity

JDM.1st similarity also relies on graph techniques with the principle that words are more or less similar according to the number of pairs collected through the "As-W" task. We built a directed and weighted graph $G_{JDM} = (V_{JDM}, E_{JDM}, W_{JDM})$ where W_{JDM} are the weights of the links: $x \rightarrow y$ = the number of times the word y has been associated with x. The similarity between x and y is the weight of the link $x \rightarrow y$ in the graph G_{JDM}.

JDM.2nd similarity is computed by applying the technique used for TLF.2nd to the graph G_{JDM}, but where the probability of jumping in a step from a vertex x to a vertex y is then proportional to the weight of the edge $x \rightarrow y$ relative to the sum of the weights of the arcs coming out of x. As for TLF.2nd, the length of the random walks is $t = 3$.

4 Quantitative analysis and results

We performed two kinds of analysis on this data. First, we computed the correlation between the six similarity measures presented in Section 4 and the response frequency, i.e. the number of subjects that gave the same response for a given stimulus. We computed the Spearman correlation coefficient over all distinct pairs and obtained the scores presented in Table 2. We can see that all correlation values are positive and statistically significant. The highest value (0.51) is obtained with JDM.2nd. Using a random walk approach (2nd order) increases the Spearman correlation from 0.23 to 0.38, (up to 65%) for TLF-based methods and from 0.47 to 0.51 (up to 8%) for JDM-based methods. In order to get a more detailed view of the complementarity of these measures and to examine the behaviour of these measures regarding the semantic relations between stimulus and response, we performed a multidimensional analysis. We ran a standard Principal Component Analysis on the matrix with Stimulus/Response pairs (559) as rows and 19 columns i.e. 1 for pair frequency, 1 per similarity measure and 1 per tagged relation (e.g. synonymy, see Table 1) converted to a binary

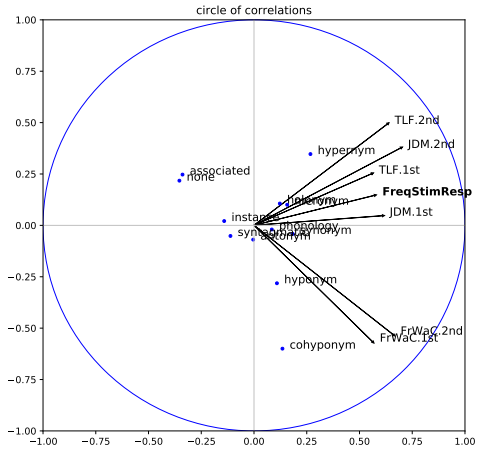

Figure 1: Circle of correlations in the first factor map of PCA.

value. The main factor map represents 33% of the global variance (see Fig. 1). Several elements can be learned from this analysis. It clearly shows that the three resources provide different aspects of lexical similarity, and that the shifting from 1st to 2nd order preserves these differences. When looking at the categorised semantic relations, several phenomena can be identified. First, it appears that all measures are positively correlated to classical semantic relations, although we observed some variation: measures based on lexical resources (TLF or JDM) capture the hypernymy relation more easily, while corpus-based similarity favour co-hyponymy. Other classical semantic relations are positively correlated with all measures, without a clear advantage for any of them. In contrast, all similarity measures are negatively correlated to non classical relations (none cases and associated word pairs). Instance, syntagmatic, antonym and phonology relations appear in the centre of the factor map, indicating that no clear trend can be identified for these relations. This is somewhat surprising that even corpus-based first order similarity (FrWac.1st) does not capture the pairs in syntagmatic relations.

5 Beyond semantic relations: clustering responses

Although the reliable identification of specific semantic relations between a stimulus and responses provided by the subjects is currently out of reach, some of the NLP techniques used to compute similarity can be used to provide a structure for the set of responses. This is especially the case for word embeddings, which are known to provide vector representation of words that are suitable for a number of semantic tasks. For example, we can use these representations to identify clusters of re-

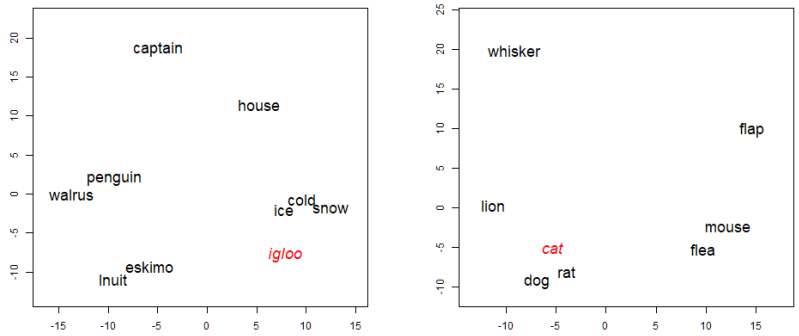

Figure 2: PCA maps of the responses to the stimuli (in red) igloo (left) and cat (right), based on word embeddings.

sponses based on their position in the vector space (vector space computed from the distribution of words in a corpus). We show here two examples of such an analysis. Focusing on the stimuli *igloo* and *cat*, we extracted the word embeddings for all responses (as well as the stimulus) and represented them in a two-dimensional space by the means of a PCA on the initial 500-dimension vectors. The results can be seen in Figure 2. While the dimensions themselves cannot be interpreted, it appears that interesting clustering can be identified in the responses. For *igloo*, we can see that all words related to an igloos typical climate and environment are gathered close to the stimulus (*cold*, *ice*, *snow*), while the prototypical inhabitants (*Eskimo*, *Inuit*) and fauna (*penguin*, *walrus*) are farther on the left. The hypernym *house* is located in another area, closer to the top. Another interesting case in this example is the presence of *captain* in the responses: it refers to a fictional character named Captain Igloo who used to appear in TV commercials for frozen fish sticks. Its position in the figure is understandably the most extremely afar from the stimulus. It is important to note that the semantic relations of most of the responses with this stimulus fall under the associated category, with the exception of the meronym *ice*, the hypernym *house* and the syntagmatically-related noun *captain*. However, it appears that word embeddings are able to separate them efficiently in relevant subsets. The results for *cat* are more self-explanatory, with the interesting case of *mouse* which is not considered as a close co-hyponym (as are *dog*, *rat* and *lion*) but more as an association because of the cat and mouse topoi.

6 Conclusion

This paper exploits a first dataset of pairs of words collected from the pilot study of the Evolex project. We proposed six techniques to compute lexical similarities of pairs of words. These six techniques are based on three kind of resources (large corpus, dictionary and crowdsourced lexical resource) with the computation of either first or second order similarity. First we computed the correlation between these six similarity measures and the response frequency. All correlation values are positive and statistically significant. The highest value (0.51) is obtained with JDM.2nd i.e. the method based on second order similarity using a short random walk approach over the crowdsourced lexical resource, collected with a protocol fairly similar to Evolex. From the experiments conducted, it appears that exceeding 0.51 might be challenging. This needs to be investigated with further experiments. Secondly, we show that the three resources provide different aspects of lexical similarity and that shifting from 1st to 2nd order preserves these differences. This conclusion will be very useful for the future of Evolex as a diagnostic tool in clinical studies. We are able to position each pair in a multidimensional space (one dimension by similarity) and to identify clusters of pairs with the final objective of defining region i.e. profiles for characterising an incoming answer to a stimulus. Such profiles may be then used for evaluating if a given phenomenon (context, age, sex, level of education, cognitive profile, language deficit, ...) favours the production of stimulus/response pairs positioned in a particular region of this multidimensional space, this can then help to identify the phenomenon as a hidden variable.

Other factors made available by the Evolex protocol have now to be taken into account, as for example the reaction time of each response and the results obtained by the participants to the other tasks of the Evolex protocol.

References

Baroni, M., Bernardini, S., Ferraresi, A., and Zanchetta, E. (2009). The wacky wide web: a collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43:209–226.

Baroni, M. and Lenci, A. (2010). Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics*, 36:673–721.

Baroni, M. and Lenci, A. (2011). How we blessed distributional semantic evaluation. In *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*, pages 1–10. Association for Computational Linguistics.

Benton, A. L. (1968). Differential behavioral effects in frontal lobe disease. *Neuropsychologia*, 6(1):53–60.

Bollobas, B. (2002). *Modern Graph Theory*. Springer-Verlag New York Inc.

Burke, D. M. and Peters, L. (1986). Word associations in old age: Evidence for consistency in semantic encoding during adulthood. *Psychology and Aging*, 1(4):283.

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

de La Haye, F. (2003). Normes d'associations verbales chez des enfants de 9, 10 et 11 ans et des adultes. *L'Année psychologique*, 103(1):109–130.

Dendien, J. and JM., P. (2003). Le trésor de la langue française informatisé : un exemple d'informatisation d'un dictionnaire de langue de référence. *TAL*, 44(2).

Evert, S. (2009). Corpora and collocations. *Corpus Linguistics: An International Handbook*, 2:1212–1248.

Ferrand, L. and Alario, F.-X. (1998). Normes d'associations verbales pour 366 noms d'objets concrets. *L'Année psychologique*, 98(4):659–709.

Gaume, B., Duvignau, K., Navarro, E., Desalle, Y., Cheung, H., Hsieh, S., Magistry, P., and Prévot, L. (2016). Skillex: a graph-based lexical score for measuring the semantic efficiency of used verbs by human subjects describing actions. *TAL*, 55, Numéro spécial sur Traitement Automatique des Langues et Sciences Cognitives(3):97 – 121.

Hill, F., Reichart, R., and Korhonen, A. (2015). Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41(4):665–695.

Lafourcade, M. (2007). Making People Play for Lexical Acquisition with the JeuxDeMots prototype. In *SNLP'07: 7th Int. Symposium on NLP*, Pattaya, Thailand.

Lenci, A. (2008). Distributional semantics in linguistic and cognitive research. *Italian journal of linguistics*, 20(1):1–31.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

New, B., Pallier, C., Brysbaert, M., and Ferrand, L. (2004). Lexique 2: A new french lexical database. *Behavior Research Methods, Instruments, & Computers*, 36(3):516–524.

Newcombe, F. (1969). *Missile wounds of the brain: A study of psychological deficits*. Oxford University Press.

Péran, P., Démonet, J.-F., Pernet, C., and Cardebat, D. (2004). Verb and noun generation tasks in huntington's disease. *Movement disorders: official journal of the Movement Disorder Society*, 19(5):565–571.

Tarrago, R., Martin, S., De La Haye, F., and Brouillet, D. (2005). Normes d'associations verbales chez des sujets âgés. *Revue Européenne de Psychologie Appliquée/European Review of Applied Psychology*, 55(4):245–253.

Urieli, A. (2013). *Robust French syntax analysis: reconciling statistical methods and linguistic knowledge in the Talismane toolkit*. PhD thesis, Université Toulouse le Mirail-Toulouse II.

Wettler, M., Rapp, R., and Sedlmeier, P. (2005). Free word associations correspond to contiguities between words in texts. *Journal of Quantitative Linguistics*, 12(2-3):111–122.