Automatic analysis of word association data from the Evolex psycholinguistic tasks using computational lexical semantic similarity measures

Bruno Gaume[1], Ludovic Tanguy[1], Cécile Fabre[1], Lydia-Mai Ho-Dac[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, ludovic.tanguy, cecile.fabre, lydia-mai.ho-dac,`
`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
`lolla.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 : Université de Toulouse, Toulouse, France
`melanie.jucla@univ-tlse2.fr`

Abstract. This paper is the fruit of a multidisciplinary project gathering researchers in Psycholinguistics, Neuropsychology, Computer Science, Natural Language Processing and Linguistics. It proposes a new data-based inductive method for automatically characterising the relation between pairs of words collected in psycholinguistics experiments on lexical access. This method takes advantage of four complementary computational measures of semantic similarity. We compare these techniques by assessing their correlation with a manual categorisation of 559 distinct word pairs, and with the distribution of data produced by 30 test subjects. 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

Assessing and characterising lexical access is one of the main interests of psycholinguists. To build their experimental material, psycholinguists frequently use measures obtained from the analysis of large corpora (see for instance lexical frequency measures; New et al., 2004). In order to specifically tackle lexical semantic relations, word association tasks are very useful tools. In such a task, a participant has to say (or

1

write) a word in response to an auditory or written stimulus (e.g. "dog" as a response to "cat"). The variables typically analysed are latencies, error rate and the lexical frequency and length of the response. Of course, together with those quantitative data, a qualitative analysis of the grammatical and/or semantic relation between the stimulus and the answer helps addressing 1) how close two words can be in someone's mental lexicon, 2) the nearest neighbours of a specific word 3) whether this network is affected by age (Burke and Peters, 1986), gender, sociodemographic status and language pathologies (Péran et al., 2004). However two main problems arise. First, we lack norms about the typical answers produced by a large sample of participants so that we cannot reliably know whether a stimulus/response pair is more or less plausible for a large number of words (see for French norms Alario & Ferrand, 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). Second, a qualitative subject-by-subject and item-by-item analysis is time consuming and prone to interpretation. Such data can be obtained through the analysis of reference language data with Natural Language Processing (NLP) techniques and help psycholinguists to better and automatically analyse word association tasks.

In NLP, 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, inter alia). Nevertheless, even if the performance of these systems is sometimes 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 need for evaluation data when comparing and assessing these techniques (Hill et al. 2005, Baroni and Lenci 2015).

This paper proposes a step toward the satisfaction of both needs. We use the 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. The Evolex project (from which the data were collected) and the data collection process are both detailed in Section 2. Section 3 describes the manual annotation process and provides a linguistic analysis of the lexical relations in the dataset. We present the computational measures of semantic similarity in Section 4. Sections 5 contains the quantitative analyses and results while Section 6 presents a promising method able to characterise and cluster stimulus/response pairs.

2 Data collection: the Evolex protocol

2.1 Data collection

The Evolex Project is a multidisciplinary project gathering researchers in Psycholinguistics, Neuropsychology, Computer Science, NLP and Linguistics. Its main objective is threefold:

- to propose a new computerised tool to evaluate lexical access in population with or without language deficits;
- to complement and reinforce the neuropsychological characterisation of lexical access using a qualitative lexical analysis (and vice versa);
- to develop and train appropriated NLP tools to automatically measure and identify the lexical relations.

The Evolex protocol includes three different tasks to assess lexical access. The Verbal Fluency test is a common procedure that includes two semantic fluency tasks (Benton 1968) that consists in naming words belonging to the animal or fruit category and two phonemic fluency tasks (Newcombe 1969) that consist in naming words starting with the letters R or V. For the Picture Naming task: participants are shown a very explicit picture (e.g. *igloo*, *baby bottle*, *cat*) and have to vocalise the word depicted by the picture. The last task is the Word Association task. This paper focuses mainly on this task which consists in vocalising the first word coming to mind after listening to a simple word (e.g. *fruit*, *painting*, *igloo*).

The 60 items used as audio stimuli for the Word Association task were selected according to their grammatical category (nouns), number of syllables (the same amount 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). A fixed order was defined i.e. the same sequence of items is given to all participants. We chose this parameter so that the inter subject discrepancy in the answers could not be attributed to a simple list order effect. To maximise the reproducibility of the experiment, the audio stimuli were produced by a speech synthesis tool[1]. The task aims at collecting data on natural lexical organisation. By asking the participants to respond as quickly as possible, the experimenter avoids their use of possible strategies. Response times were not used in the study presented here.

One of the key innovations of the Evolex protocol is to propose a computer-assisted method for collecting and processing the data. The software includes a system that automatically recognises and analyses the vocal response. An Automatic Speech Recognition (ASR) tool transcribed the response and recorded the reaction time (i.e. the time period between the beginning of the stimulus and the beginning of the subject answer). A web interface allows the user to correct the ASR transcription.

This paper exploits a first data set of pairs of words collected from a pilot study conducted with 30 participants with no language disorders, native French speaker, aged between 15 and 58 (mean age 31 ± 13.06), with variable levels of education (from 10 to 20 years of schooling, mean 15.4 ± 2.97). The following instructions were given to participants: "*You will hear French common nouns. You will have to pronounce the*

---

[1] http://acapela-box.com/

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".

2.2 Data preprocessing: cleaning up and normalisation

We collected the 1800 individual responses (30 subjects, 60 stimuli). We grouped and filtered them according to the following criteria: the response must be a monolexical noun or a proper name, in its non-inflected form. In addition, we rejected two stimuli (and their associated responses) because of a phonological confusion induced by the speech synthesis system. We used a combination of automated post-processing and final manual verification and ended up with a total of 1544 individual validated responses, corresponding to 559 distinct stimulus-response word pairs.

3 Qualitative analysis of data

The 559 distinct stimulus-response pairs have been annotated by two judges. The tagset is composed of 12 tags, as illustrated in Table 1, and covers four types of relations. The first type of relations are classical lexical relations, as found for example in the WordNet database (Fellbaum 1998). They include synonyms, antonyms, generic-specific relations (hypernyms, hyponyms, co-hyponyms and instance) and part-whole relations (meronyms and holonyms). A second type of relation called *associated* holds between words that are semantically related in a broader way: they tend to appear in the same contexts (both textual and referential) because they are connected within the same class of objects or events (Morris and Hirst 2004). Syntagmatic relations concern words that tend to combine to form larger syntactic constituents (expressions, compounds, etc.). In the example given in Table 1, *fleur* (flower) and *peau* (skin) are not semantically related, but they are associated in the expression *à fleur de peau* (hypersensitive, thin-skinned). The fourth relation (*phonology*) refers to a phonological proximity between words, with no semantic connection. For a small proportion of the pairs, no specific relation could be identified (*none found*) or the connection seemed too far-fetched (as in the example in the last line of Table 1).

Independent double annotation has been performed and followed by adjudication. After this first step, 69 out the 559 annotated pairs received more than one tag because they could be part of several relations. Another stage of collective adjudication has been carried out to retain only the relation that was considered most prominent, on the basis of priority rules. In the resulting dataset, classical lexical relations altogether represent almost half of the pairs (49.5%), among which co-hyponyms (sisters of the same superordinate) stand out, although all classical relations other than antonyms are also well represented. Associated pairs make up more than one third of the set (36.1%). Syntagmatic relations form 8.8% of the pairs. The amount of phonological links is almost negligible (0.9%).

Relation	Example (stimulus / response)	# distinct pairs	%
antonym	*aube* (dawn) / *crépuscule* (dusk)	2	0.4%
associated	*balançoire* (swing) / *enfant* (child)	202	36.1%
co-hyponym	*balançoire* (swing) / *toboggan* (slide)	73	13.1%
holonym	*doigt* (finger) / *main* (hand)	29	5.2%
hypernym	*balançoire* (swing) / *jeu* (game)	52	9.3%
hyponym	*animal* (animal) / *chat* (cat)	45	8.1%
instance	*magicien* (wizard) / *Merlin* (Merlin)	6	1.1%
meronym	*balançoire* (swing) / *corde* (rope)	49	8.8%
phonology	*chapiteau* (circus tent) / *château* (castle)	5	0.9%
synonym	*canapé* (couch) / *sofa* (sofa)	21	3.8%
syntagmatic	*fleur* (flower) / *peau* (skin)	47	8.4%
none found	*perroquet* (parrot) / *placard* (closet)	28	5.0%

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

4 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 four techniques we tested differ in two ways. First, different resources were used: the first two make use of a large corpus of French, FrWaC (Baroni et al. 2009) which is a collection of Web pages from the *.fr* domain and consists of 2 billion words. The latter techniques are based on the *TLF* (*Trésor de la Langue Française*, see Dendien and Pierrel, 2003) dictionary from which we extracted the full text of all the definitions. Both resources have been POS-tagged and lemmatised with the *Talismane* toolkit (Urieli 2013). The second dimension on which these techniques differ is the fact that we consider either the first order similarity (meaning that words collocated in a corpus are similar, and that dictionary headwords are similar to the words appearing in their definition) or second order similarity (also known as distributional similarity), considering that words sharing first-order similar words show a possibly different degree of similarity. Each of these techniques is described in the following subsections.

4.1 Corpus-based first-order similarity (collocates)

The simplest measure of similarity is to consider collocation, i.e. the fact that some words appear frequently and systematically together. This corpus-based 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). It has also been shown to be correlated with words association data (Wettler et al. 2005).

We computed this similarity using Positive Pairwise Mutual Information, one of the most commonly used alternatives amongst collocation measures (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.

4.2 Corpus-based second-order similarity (word embeddings)

The second corpus-based measure relies on the principles of distributional semantics, which consider that words appearing in the same contexts have similar meanings. Second-order similarity can be computed in a number of ways (Baroni and Lenci 2010), but for a few years most of the work and research has focused on word embeddings. Word embeddings are a dense numeric representation of lexical items based on their distribution in a corpus. State-of-the-art methods for computing these embeddings rely on neural networks that are trained to predict words given context elements (or vice-versa), and are readily available. For this experiment, we used Word2vec (Mikolov et al. 2013), undoubtedly the most commonly used system and applied it to the same corpus used for first-order similarity i.e. FrWac. The following parameters were used: skipgram algorithm with negative sampling (rate 5), window size 5, 500 dimensions, subsampling rate 10^{-3}, 5 iterations, minimum frequency 100. As a result we obtained a dense matrix in which each word is represented by a numeric vector (of size 500). The cosine distance was then computed to measure the similarity between two words. Distributional semantics similarity measures are well known for capturing a wide spectrum of semantic relations (Baroni and Lenci 2011). This can be an issue for some tasks (Ferret, 2015) but was an asset in our case.

4.3 Dictionary-based first-order similarity (presence/absence in definitions)

The third technique, which uses a general-purpose dictionary for measuring first-order similarity, is based on a very simple principle: if a word appears in the definition of another word then the two words share a part of their semantic contents. We used this straightforward approach on the definitions extracted from the TLF dictionary without considering any explicit information that could be found in the dictionary such as cross-references. The only additional processing we applied was to symmetrise the relation. This similarity measure $S_{TLF}(x, y)$ is therefore binary: the similarity between x and y is 1 if x appears in the TLF's definition of y or vice-versa (or both), and 0 otherwise.

4.4 Dictionary-based second-order similarity (random walk across definitions)

For second-order similarity we used a random walk approach (Bollobas, 2002). This graph traversal technique is used to define a broader, more robust[2] measure of similarity between the nodes of a graph. We applied this technique to the undirected, unweighted G_{TLF} graph used in the first-order approach described in the previous subsection.

---

[2] We say "robust" in the sense that the similarities computed on graphs from different dictionaries are strongly related to one another, while this is not the case with first order methods (see Gaume et al., 2016).

Formally, this similarity measure is $P^t_{G_{TLF}}(x,y) \in [0,1]$, 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. This technique will therefore attribute a positive similarity score to two words whose definitions share words (the more words, and the more specific they are, the higher the value), or to two words appearing in the same definitions, and even to slightly more distant words in the original graphs. This method has proved to capture different kinds of semantic relation.

5 Quantitative analysis and results

As described in Sections 2 and 3, our dataset consists of 559 stimulus-response pairs of words, each with a hand-tagged semantic relation. In addition, we also know the response frequency, i.e. the number of subjects that gave the same response for a given stimulus as well as the four computed similarity values. We performed two kinds of analysis on this data.

First, we computed the correlation between the four similarity measures presented in Section 4 and the response frequency. We used the Spearman correlation coefficient over all pairs and obtained the scores presented in Table 2 below.

As can be seen, all correlation values are positive and statistically significant. The highest value is obtained for the dictionary-based second order similarity. For both resources, shifting from first to second order results in an increased correlation (up to 70% for Dictionary-based methods).

Similarity measure	Spearman's ρ	p-value
Corpus-based, 1st order	0.215	2.3e-07
Corpus-based, 2nd order	0.247	5.3e-09
Dictionary-based, 1st order	0.191	4.5e-06
Dictionary-based, 2nd order	0.325	1.7e-15

Table 2 : Spearman correlation between similarity measures and the response frequency.

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 containing the similarity values and response frequency for each pair, and then projected the categories on the reduced vector space. The main factor map is presented in Figure 1 below, representing 66% of the global variance).

Several elements can be learned from this analysis. It clearly shows that the two resources (corpus and dictionary) provide different aspects of lexical similarity, and that the shifting from first to second order preserves these differences. When looking at the categorised semantic relations (cf. Section 3), several phenomena can be identified. First, it appears that all similarity measures are negatively correlated to non-

classical relations. The *none* cases for which no semantic relation has been identified have low similarity values for all measures, and it is the same (to a lesser extent) for *phonology* and *instance* word pairs. Associated and syntagmatic relations appear in the centre of the factor map, indicating that no clear trend can be identified for these relations, although they are on the opposite side of the similarity vectors. This is somewhat surprising that even corpus-based first order similarity does not capture these cases. On the other (right) side of the map, we can find all classical semantic relations, although with varying correlations with the four similarity measures. It appears that dictionary-based methods capture the hypernymy relation more easily, while corpus-
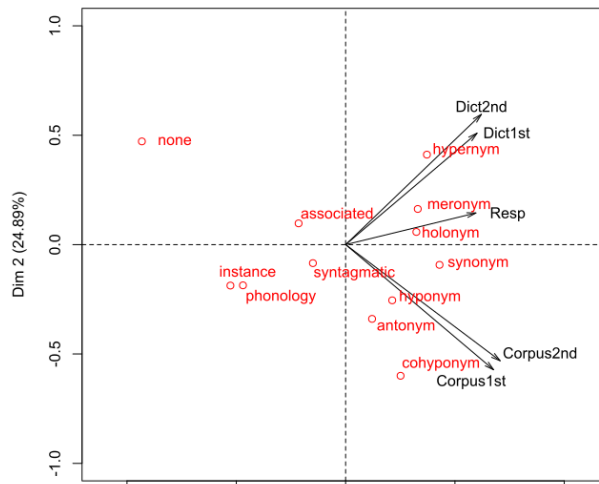
Figure 1: First factor map of PCA based on the 4 similarity measures and response frequency. Categorised relations (in red) are shown as additional variables

based methods favour co-hyponymy. Other relations are positively correlated with all measures, without a clear advantage for any of them. This indicates that automated measures can be useful for the detection of atypical responses. They will be tested against authenticated pathological responses in the near future.

In conclusion, we can see that the four tested methods manage to capture a significant part of the associations produced by subjects, with the more sophisticated (second order) measures showing a slightly higher correlation. The resources used for computing similarity have an influence on both the overall correlation, but more interestingly on the types of semantic relations between stimulus and response. However, none of these methods is particularly suited to identify non-paradigmatic associations.

6 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 responses 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 analysis.

Focusing on the stimuli *igloo*" and "*cat*, we extracted for each one the word embeddings of all responses (and 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 below. If the dimensions themselves cannot be interpreted, it appears that interesting clustering can be seen in the responses.

For *igloo*, we can see that all words related to the igloo's 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, this time 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 *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.
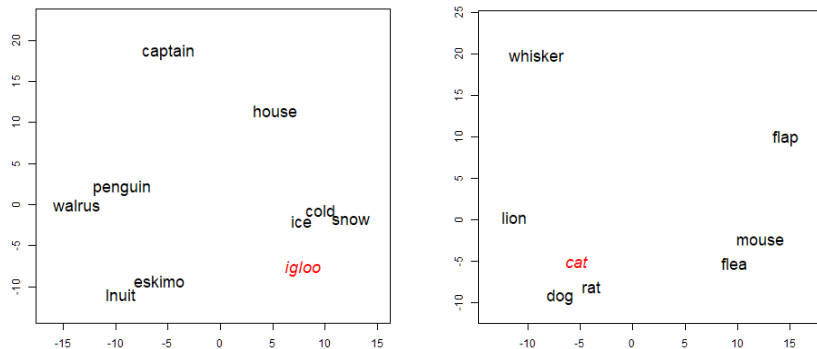
Figure 2: PCA maps of the responses to the stimuli (in red) "igloo" (left) and "cat" (right), based on word embeddings -- manual translation to English

7 Conclusion

In this paper, we described a series of experiments performed on the data collected from a word association task, in order to assess the possibility of using NLP techniques to automatically categorise the responses provided by non-pathological subjects. We manually tagged 559 different word pairs according to the lexical semantic relation between stimulus and response. We tested four different measures of similarity that

differ on the resource used (a general corpus and a dictionary), and the nature of similarity (first and second-order). We showed that dictionary-based second-order similarity provides a measure that has the highest correlation with the data in terms of number of subjects who produced the responses. We also showed that if all of these different measures have very low scores for non-semantically related pairs, and favour some of the more classical relations (hypernymy, synonymy and co-hyponymy), they cannot be used without further development to identify the other relations, and especially the non-paradigmatic associations. However, we also showed that these techniques can prove surprisingly efficient for the clustering of responses according to stimulus-specific semantic relations.

There are other factors that need to be taken into account in the near future. The reaction time of each response is known to be a significant information for categorising subjects, as are of course the generic characteristics of the subjects (age, education level, etc.). But the most important perspective for the work and methods presented here is of course the analysis of data collected from pathological subjects. We are confident that the similarity measures will be able to identify non-typical responses, but we will need further analysis in order to associate the non-typicality with specific pathologies or disorders.

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 (3): 209-226.

Baroni, M. and Lenci, A. 2010. Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics*, 36(4): 673–721.

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

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

Burke, D. M., & 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 Pierrel, JM. (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. In A. Lüdeling and M. Kytö (eds.), *Corpus Linguistics: An International Handbook* Vol.2. Berlin, New York, 1212-1248.

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

Ferret O. (2015) Typing Relations in Distributional Thesauri. In: Gala N., Rapp R., Bel-Enguix G. (eds) *Language Production, Cognition, and the Lexicon*. Springer.

Fellbaum, C. (1998) (ed). *WordNet: an Electronic Lexical Database*. MIT Press.

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.

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

Lenci, A. (2008). Distributional semantics in linguistic and cognitive research. *Italian Journal of Linguistics,* 20(1): 1-31.

Mikolov T., Chen K., Corrdao G. and Dean J. (2013). Efficient Estimation of Word Representations in Vector Space. In *ICLR Workshop*.

Morris, J., and Hirst, G. (2004). Non-classical lexical semantic relations. *Proceedings of the HLT-NAACL Workshop on Computational Lexical Semantics*, 46-51.

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

Newcombe, F. (1969). *Missle Wounds of the Brain. A Study of Psychological Deficits*. London: Oxford University Press.

Péran, P., Démonet, J. F., Pernet, C., & Cardebat, D. (2004). Verb and noun generation tasks in Huntington's disease. Movement Disorders, 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. University of Toulouse, France.

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).