

Evaluation of lexical resources and semantic networks on a corpus of mental associations

Laurianne Sitbon^{1,2} Patrice Bellot¹ Philippe Blache²

(1) Laboratoire d'Informatique d'Avignon - University of Avignon

(2) Laboratoire Parole et Langage - University of Provence

{laurianne.sitbon, patrice.bellot}@univ-avignon.fr, blache@lpl-aix.fr

Several cognitive processes and resources drive language production. A lack of phonological awareness may involve a reduction of the connections between phonological and mental lexicons. This can be caused by ageing or by various pathologies such as dyslexia. A known effect is the tip of the tongue : a word cannot be directly accessed but it can be recognized when presented into a list of words. We hypothesized that existing semantic networks can help to model the mental lexicon when a wide variety of semantic links are employed together. Previous work suggested the use of phonological distance (Zock, 2002) combined with several lexical resources (Reuer, 2004). This work is a feasibility study of a system combining 5 different semantic networks in order to provide a target word from a list of 10 proposals when the user provide 5 words that he mentally associates to the target word.

1 Evaluated resources

The study is based on 5 networks already available as themselves or constructed with available applications. We selected them in order to represent a wide range of cognitive approaches to mental association and also to represent a wide range of language levels. The mental association can be descriptive, paradigmatic or collocative. The language level is represented by 3 corpora : one journalistic, one literary and one generalist (extracted from the Web).

WordNet¹ is an available resource for paradigmatic relations between words. The French version of this semantic network is available inside EuroWordNet project. We used both synonymy and hierarchical relationships. We built a dynamic resource for descriptive relationships by extracting keywords of dictionary definitions available from <http://www.answers.com> dictionary. A collocation network has been built by computing mutual information of terms in a 20-words window on a corpus made of Le Monde journal extracts. Another collocation network has been built with Infomap tool applied on Corpatext² literature corpus. Infomap³ uses a vector space model approach on a term-document matrix to provide semantic associations. The third collocation associations are dynamically extracted for each word by computing the keywords of the 10 first documents provided by Google search engine when requesting the word.

¹<http://wordnet.princeton.edu>

²<http://www.lexique.org/public/corpatext.php>

³<http://infomap-nlp.sourceforge.net>

2 Evaluation corpus

The study is based on a corpus we constructed. It is made of 20 target words each associated with 5 other words by 50 users. The users filled the survey from their home computer through a web application. The target words were selected so that they belong to every evaluated network. According to Lexique 3 database⁴, their frequencies in films and books are low. All 50 users are adults with various education level and profession. Among the 250 associations provided for each target word, the average number of different ones is 76,8. This corpus is freely available on demand for research work and the language used is French.

3 Analysis

We evaluated the availability of each target word in the list of words obtained by relating each provided association with each lexical resource. Only the 100 first words provided by a resource are considered for each association. We count how many associations can lead to each target word. Each association leads to a list of at most 500 words (5 resources, 100 words each). If the target word is in this list, the associated word is called useful association (UA). If we consider only one resource, an UA occurs when the target word is in the list of 100 words provided by the resource for this association.

We first look at the marginality of each resource. This is the percentage of associations that become UA thanks to this resource and any other one. Between 27% and 72% of TOT provided by each resource are not provided by the other ones. 47% of TOT retrieved are provided only by the web resource. This value is between 3% and 13% for other resources. This reveals that combining various resources is useful because they provide useful information in different cases.

We next looked at the results provided by each target word. For each word, we numbered the percentage of provided associations (over 250) that are UA, in average and separately by resource. Then, we numbered the sets of 5 associations containing at least one and two UA, and the ones containing only UA. Finally, we numbered the sets of 5 associations where the first one provided by the user is UA. In average, 33% of provided associations are UA. 79% of sets of associations contain at least one UA, 49% contain at least two UA and 3% contain five UA. In 49% of sets the first association provided is UA. More precisely, it appears that for some target words, UA can be provided only by one resource. Some target words have almost only set of associations where the first one is UA, while some others have almost never a UA in first position. This confirm the very intuitive idea that some words are stronger linked together in world mind.

Lastly, we study users specificities. The objective is to know whether it is possible to select the most representative resources for one user, considering that each resource reflects more or less the cognition of this user. In order to highlight tendencies, we try to cluster users automatically. The K-means algorithm has been applied with the number of UA according to each resource for each target word. Each user is represented with 20 samples containing 5 parameters. This clustering highlights two classes of users, mainly separated by web resource efficiency criterion.

⁴<http://www.lexique.org>: a lexical database containing frequencies of words and lemmas in various corpora.

As a second approach, the Expectation-Maximization (EM) algorithm has been applied with the total number of UA the user provide for all target words according to each resource. Each user is represented with one sample containing 5 parameters. The algorithm provides three clusters. The first one distinguishes people more sensitive to paradigmatic and descriptive relations than others. The second one concerns people associating mainly with collocations (web and journalistic). The third one relates to people providing associations relating the target word in few ways whatever the resource is.

4 Using a combination of resources

The combination of the resources and the associations to reach a target word is achieved by measuring and weighting paths between each pair of associations. The principle is illustrated by Figure 1, with paths of size 1, 2 and 3. The score of a potential target word R is computed with the sum on equation 1. For each pair of associations a_i and a_j , we sum the inverse of the length of the minimal path $path(R, a_i, a_j)$ joining the associations through this potential target word. This can be weighted according to the maximal score $P_{res}(R, a_i, a_j)$ associated to the resource the potential target word can be provided by. A proposition list is composed of the N target words having the highest scores.

$$Score(R) = \sum_{i=0, j=0, i \neq j} P_{res}(R, a_i, a_j) \frac{1}{path(R, a_i, a_j)} \quad (1)$$

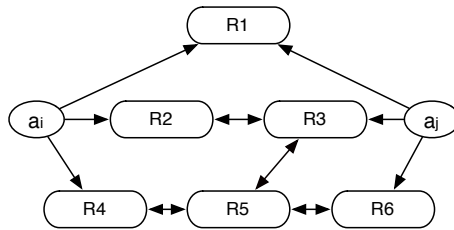


Figure 1: Affection of paths scores for various words between two associations.

The combination has been experienced on our data with paths of a maximum size of 1 item and an equal weight for each resource. This is the intersection between potential target words, where the score of a target word is the number of associations related to. We next evaluate the proportion of proposition lists containing the TOT, according to the size of the list (between 0 and 100 words). Previous results already show that the maximum reachable is 47 %. The first important result, according to the graph on figure 2, is that the TOT never appear in the 20 first proposals. The progression is linear until 35 % of TOT appearing in 100 words long proposition lists.

5 Discussion

The correlation between TOT and associations in our corpus can be consider not realistic, as long as they have not been registered in real TOT situations. However, they can be considered

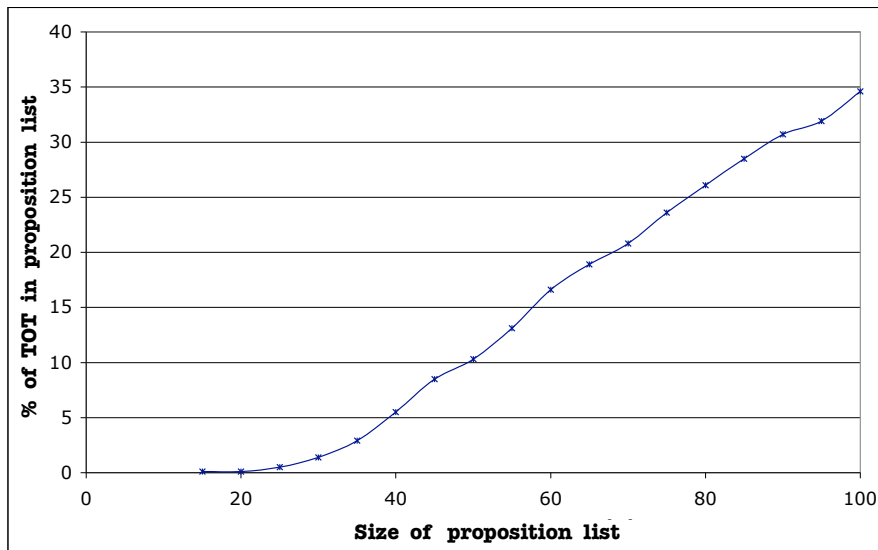


Figure 2: Rates of in N-first lists of proposals containing TOT.

as representative of users' mind. In that way they can be really useful for various evaluation of resources pretending to be representative of the mental lexicon. Our study with various resources shows that they tend to be complementary. Some are better representative of a group of users, but this is not clear how much the nature of the target word influence the user. The data can also be used in a psychological study, to determine whether some users tend to associate TOT visually or textually. As an example, the TOT "postman" has been associated 14 times with "dog". Another study could evaluate the efficiency of collocation algorithms when they are based on similar training corpus.

Future work will concentrate on a better weighting of the potential TOT. First, longer paths can improve the results. We hope that considering more steps will increase the scores of TOT in proposal lists. The complexity is exponential with each added step. Performance issues will be of interest for user interface. The implementation of users models in order to define resources' weight can also be of interest. Each potential TOT could be weighted according to global criterions, such as low frequency words, if psycholinguistic studies agrees on that point.

The idea of converging on TOT words by combining the related words to each association is mostly interesting for an audio application. Speech recognition of associated word would provide several weighted hypothesis. The potential TOT related to wrong recognized words would probably not be repeated in other associations' related words. Their scores should be low. The ambiguity raised by speech recognition will no longer be an issue.

References

- (Zock, 2002) M. Zock, 2002. Sorry, but what was your name again, or, how to overcome the tip-of-the tongue problem with the help of a computer ? Dans les actes de *COLING-Workshop on building and using semantic networks*, Taipei, Taiwan.
- (Reuer, 2004) V. Reuer, 2004. Language resources for a network-based dictionary. in *Workshop on Enhancing and Using Electronic Dictionaries ; COLING 2004*, 81–84.