Evaluation Of Lexical Resources And Semantic Networks On A Corpus Of Mental Associations

Laurianne Sitbon[1,2] **Patrice Bellot**[1] **Philippe Blache**[2]

(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

Abstract

When a user cannot find a word, he may think of semantically related words that could be used into an automatic process to help him. This paper presents an evaluation of lexical resources and semantic networks for modelling mental associations. A corpus of associations has been constructed for its evaluation. It is composed of 20 low frequency target words each associated 5 times by 20 users. In the experiments we look for the target word in propositions made from the associated words thanks to 5 different resources. The results show that even if each resource has a usefull specificity, the global recall is low. An experiment to extract common semantic features of several associations showed that we cannot expect to see the target word below a rank of 20 propositions.

1. Introduction

The construction of semantic resources has been driven in several ways, considering formal relationships between words or associative relationships. Some of these resources claim to model the mental lexicon in various manners. A resource capable of completely model the mental lexicon would not only be a treasure of psychologists, but could also provide an assistance in language production. 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 (Snowling, 2000). 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 (Brown and McNeill, 1966). According to current connexionist models of language production (Dell et al., 1999), verbal information is organised in mind into layers of units : semantic features, words and phonemes. We hypothesized that existing semantic networks can help to model the mental lexicon when a wide variety of semantic links are employed together. Previous works suggested the use of phonological distance (Zock, 2002) combined with several lexical resources (Reuer, 2004). The aim of the work presented here is to validate this hypothesis in a context more general than the tip of the tongue while focusing on similar situations.We propose 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 provides 5 words that he mentally associates to the target word. We first evaluate the ability of each semantic network to provide a target word from provided associations independently and we investigate the correlation between users and semantic networks. Finally we evaluate a simple combination approach.

2. 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 a wide range of linguistic levels. The association can be descriptive, paradigmatic or syntagmatic. The linguistic level is represented by 3 corpora : one journalistic, one literary and one generalist (extracted from the Web).

We built a dynamic resource for descriptive relationships by extracting keywords from dictionary definitions extracted from `http://www.answers.com` dictionary (*def*).

WordNet[1] 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 (*ewn*).

The syntagmatic relations are represented here by various approaches of statistical collocation. 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 according to (Church and Hanks, 1990) approach (*coocc*). It has been used by (Ferret and Zock, 2006) for similar experiments.

We have built another collocation network with Infomap tool applied on Corpatext[2] literature corpus (*lsa*). Infomap[3] uses a vector space model approach on a term-document matrix to provide semantic associations. The vector space model is combined with latent semantic analysis (LSA) principle (Deerwester et al., 1990) in order to reduce the lexical space into a concept space based on word cooccurrences. (Landauer et al., 1998) demonstrates that LSA leads to a good representativity of the mental lexicon.

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 (*web*). Keywords are extracted according to their *tf.idf* score based on their frequency in the

[1] `http://wordnet.princeton.edu`

[2] `http://www.lexique.org/public/corpatext.php`

[3] `http://infomap-nlp.sourceforge.net`

document (*tf*) and their inverse document frequency (*idf*), the number of documents they appear in inside a corpus of French newspapers.

resource	associations
def	Mardi Gras, cinder, Italian
ewn	blast, funfair, clearance, sight, show
lsa	opera, actor, balls, Venice, fairyland
coocc	party, city, street, mask, music
web	Martinique, Nice, Lent, Rio, multicolor

Table 1: Translation of examples of words associated to Carnival according to each ressource. Resources are key words of definitions (def), EuroWordNet (ewn), a semantic map (lsa), a cooccurrences network (cooc) and key words of web results (web).

3. Evaluation corpus

The study is based on a corpus we constructed. It is made of 20 target words (TW) each associated with 5 other words (associations) by 50 users. We assume that words targeted in tip of the tongue phenomenon are low frequency words as shown by (Burke et al., 1991). The target words were selected so that they belong to every evaluated resource. Table 2 contains the 20 chosen target words and their associated english translations and their frequencies in a french movie dialogs corpus and in a french books database, according to Lexique 3 database[4]. These frequencies are low. We assume that the association that can be made during a tip of the tongue phenomenon are similar to standard association that can be made from a known word. This is a strong assumption that has been made here for convenience. The users filled the survey from their home computer through a web application. All 50 users are adults with various education levels and profession. They were specifically instructed to avoid deflections while proposing words associated only to the target and not to the preceeding association. Table 7 contains an extract of the corpus with all association sets for the target word *facteur* (factor, postman).
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.

4. Single words Analysis

We evaluated the availability of each target word in the list of words obtained by relating each provided association with each lexical resource, as illustrated on Figure 1. 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.

[4] http://www.lexique.org

TW	Translation(s)	f. movie	f. book
veau	veal	6.20	16.96
baril	barrel. cask	4.22	3.04
casino	casino	17.41	10.81
entretien	care. interview	17.71	27.77
cambriolage	burglary. breaking in	9.34	2.77
Java	Java	0.72	2.30
menhir	menhir	0.18	0.68
quiche	quiche	0.66	0.68
kleenex	facial tissue	2.11	2.91
lion	lion. leo	17.05	33.04
brioche	brioche. paunch	7.29	7.09
facteur	factor. postman	11.27	14.32
festin	party. banquet	5.12	5.68
chaussette	sock	14.58	22.84
massif	massif. massive	8.13	22.30
virtuel	virtual	2.53	2.16
rugby	rugby	1.87	3.11
landau	carriage	1.20	4.59
snowboard	snowboard	N/A	N/A
oie	goose	5.90	9.32

Table 2: Chosen words becoming TW in the corpus and their possible translations in English and their frequency in a film dialog database (f. movie) and in a book database (f. book).

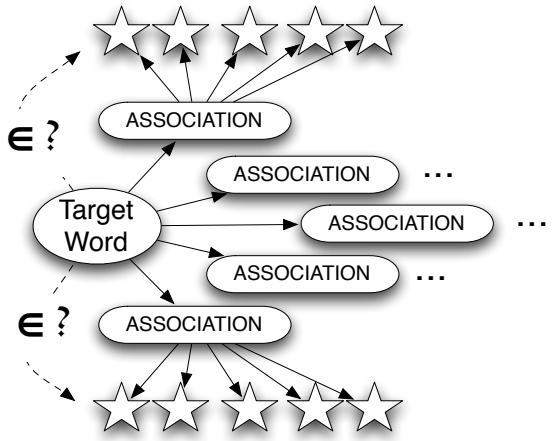

Figure 1: Illustration of the way of evaluating resources : an associated word in relation with the TW thanks to a particular resource is called an UA for this resource.

4.1. General results

First we 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. The average values are presented in Table 3. These results show that the dynamic resource based on the web is the most effective, whilst the recall rate is only around 20%. The *total* value expresses the recall for any resource (the percentage of target words that are UA for at least one resource). The fact that this value is 1.5 times the best recall rate suggests that each resource may bring its own specificity and that combining all resources

can significantly improve the retrieval of the target word. While looking at individual target word results, it appears that for some target words, UA can be provided only by one resource, for all 50 sets of associations provided by the users.

Then, we numbered the sets of 5 associations provided by the users that contain at least one and two UA, and the ones only containing UA. We also numbered the sets of 5 associations where the first one provided by the user is an UA. The average values for all association sets are shown in Table 4. 79% of sets of associations contain at least one UA, 49% contain at least two UA and 3% contain five UA. In 49.5% of sets the first association provided is UA. When we considered individual target words results, it appeared that 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.

def	ewn	lsa	coocc	web	total
5.22	6.52	3.54	11.78	20.56	33.2

Table 3: Global recall of each resource : percentage of associations becomming UA thanks to each resource. Resources are key words of definitions (def), EuroWordNet (ewn), a semantic map (lsa), a cooccurrences network (cooc) and keywords of web results (web).

% of TW with N UA			% of TW
N > 0	N > 1	N = 5	with 1st UA :
79	49	3	49.5

Table 4: Percentage of target words/association sets for which at least 1 association is UA (N>0), at least 2 associations are UA (N>1), all 5 associations are UA (N=5) and first proposed association is UA.

4.2. Overlap between resources

Next we look at the distinctiveness of each resource. Distinctiveness can be represented by the unique recall rate which is the percentage of associations that become UA based on this resource alone. Unique recall rates of each resource are given in Table 5 as a percentage of the UA found in the recall of this resource that are unique (%resource) along with the portion of the global recall rate (that is the number of UA thanks to any resource) that these unique UAs represent (%global). The unique recall is represented on Figure 2 with the gray zone, considering only one other resource R2. According to results provided in Table 5, between 23% and 70% of target words provided by each resource are not provided by the other ones. 44% of target words retrieved are provided only by the web resource. This value is between 2% and 16% for other resources. This reveals that combining various resources is useful because they provide useful information in different cases.

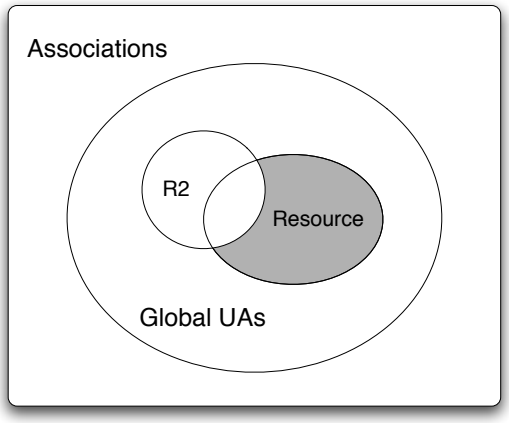

Figure 2: Ilustration of the unique recall rate.

relations	def	ewn	syntagmatic		
			lsa	coocc	web
%resource	40%	33%	23%	47%	70%
%global	6%	6%	2%	16%	44%

Table 5: Unique recall rates of each resource (percentage of associations that become UA thanks to this resource and any other one), expressed regarding the global recall of this resource (the number of UA thanks to this resource) (%resource) , or regarding the global recall (the number of UA thanks to any resource) (%global). Resources are key words of definitions (def), EuroWordNet (ewn), a semantic map (lsa), a cooccurrences network (cooc) and key words of web results (web).

4.3. Users profiles

Lastly, we study users' specificities. The purpose is to know whether it is possible to select the most representative resources for a given 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.

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. Hence each user is represented with one sample containing 5 parameters. The algorithm provides three clusters detailed in Table 6. The first one concerns people associating mainly with collocations (web and journalistic). The second one distinguishes people more sensitive to paradigmatic and descriptive relations than others.The third one relates to people providing associations relating the target word in few ways whatever the resource is.

	Class 1		Class 2		Class 3	
	Avg.	S.D.	Avg.	S.D.	Avg.	S.D.
web	20.6	3.1	25.2	2.3	13.6	2.6
def	4.6	1.4	6.9	2.2	4.3	1.6
ewn	5.9	1.9	9.4	1.6	3.7	1.3
lsa	3.7	1.1	4.5	1.4	1.8	1
coocc	11.6	1.8	13.4	2.9	9.7	2
Inst	48%		32%		20%	

Table 6: Classes provided by Expectation-Maximization algorithm with associated mean values and standard deviation for each parameter. The parameters are the percentage of UA of each user for each resource. The repartition of the instance affected to each class is also provided.

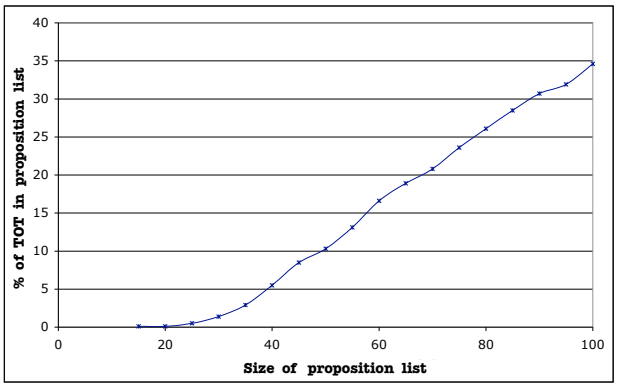

Figure 4: Rates of in N-first lists of proposals containing TW.

5. Using a combination of resources

The combination of the resources and the associations to reach a target word (TW) is achieved by measuring and weighting paths between each pair of associations. The principle is illustrated by Figure 3, with paths of size 1, 2 and 3. The score of a potential target word W 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 $path(W, a_i, a_j)$ joining the associations through this potential target word. This can be weighted according to the maximal score $P_r(W, a_i, a_j)$ 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.

$$Score(W) = \sum_{i \neq j} P_r(W, a_i, a_j) \frac{1}{path(W, a_i, a_j)} \qquad (1)$$

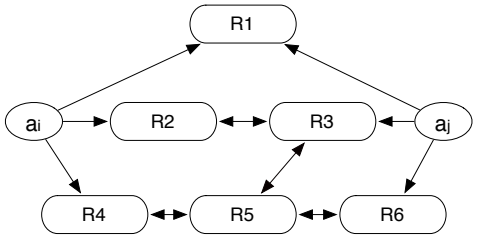

Figure 3: Affectation of paths scores for various words between two associations.

The combination has been experimented on our data with paths of a maximum size of 1 word 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. Next we evaluate the proportion of proposition lists containing the TW, according to the size of the list (between 0 and 100 words). Previous results already showed that the maximum reachable is 47 %. The first important result, according to the graph on Figure 4, is that the TW never appears in the 20 first proposals. The maximum reachable rate is 35 % of TW when considering 100 words long propostition lists.

6. Discussion

The correlation between TW and associations in our corpus can be considered not realistic, as long as they have not been registered in real TW situations. However, they can be considered as representative of user's mind. In that way they can be really useful for various evaluations 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 TW visually or textually. As an example, the TW "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. The main disadvantage of collocation approaches evaluated here is to discard secondary meanings. (Ji et al., 2003) propose an approach based on *contextonyms* that distinguishes different senses of a word by creating classes of associated terms. We will experiment such approaches in a future work.

Future work will concentrate on a better weighting of the potential TW. First, longer paths can improve the results. We hope that considering more steps will increase the scores of TW 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 TW could be weighted according to global criterions, such as low frequency words, if psycholinguistic studies agree on that point.

7. References

R. Brown and D. McNeill. 1966. The tip of the tongue phenomenon. *Journal of verbal learning and verbal behaviour*, 5:325–337.

D. M. Burke, D. G. MacKay, J. S. Worthley, and E. Wade. 1991. On the tip of the tongue : what causes word finding failures in young and older adults ? *Journal of Memory and Language*, 30:542–579.

Kenneth Church and Patrick Hanks. 1990. Word association norms, mutual information , and lexicography. *Computational Linguistics*, 16(1):177–210.

S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407.

G. Dell, F. Chang, and Z. Griffin. 1999. Connectionist models of language production : lexical access and grammatical encoding. *Cognitive Science*, 23:517–542.

Olivier Ferret and Michael Zock. 2006. Enhancing electronic dictionaries with an index based on associations. In *Coling/ACL joint conference*, pages 281–288, Sydney, Australia.

Hyungsuk Ji, Sabine Ploux, and Eric Wehrli. 2003. Lexical knowledge representation with contextonyms. In *MT Summit IX*, New Orleans, USA.

Thomas Landauer, P. W. Foltz, and D. Laham. 1998. Introduction to latent semantic analysis. *Discourse Processes*, 25:259–284.

Veit Reuer. 2004. Language resources for a network-based dictionary. In *Workshop on Enhancing and Using Electronic Dictionaries ; following COLING 2004*, pages 81–84.

M. J. Snowling. 2000. *Dyslexia*. Blackwell.

Michael 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 ? In *COLING-Workshop on building and using semantic networks*, pages 1–6, Taipei, Taiwan.

poste, lettre, courrier, vélo, casquette
multiplier, poste, vélo, casquette, sacoche
postier, lettre, colis, courrier, bicyclette
lettre, courrier, mobylette, poste, colis
poste, courrier, cyclomoteur, fonctionnaire, colis
lettre, colis, chien, vélomoteur, poste
courrier, lettre, timbre, boîte à lettre, correspondance
sonnette, lettre, poste, saint ex, boîte aux lettres
courrier, bleu, chien, poste, livraison
courrier, enveloppe, poste, jaune, lettre
lettre, cheval, courrier, besancenot, vélo
courrier, facture, nouvelles, boîte aux lettres, matin
lettre, vélo, jaune, timbre, livrer
courrier, boîte à lettre, chien, timbre, mobylette
film, lettre, timbre, poste, mobylette
lettre, message, facture, recommandé, ronde
cheval, besancenot, vélo, tati, casquette
lettre, colis, poste, envoyer, recevoir
lettre, aventure, vélo, recommandé, poste
poste, mathématiques, courrier, chien, vélo
poste, chien, cocu, vélo, courrier
courrier, boîte aux lettres, poste, chien, vélo
courrier, jaune, 4L, colis, lettres
courrier, vélo, lettre, poste, explicatif
poste, courrier, recommandé, nouvelle, mobylette
fils, poste, courrier, garfield, vélo
sonne, camionnette, rhésus, fils, timbres
poste, courrier, vélo, chien, recommande
sonne, courrier, poste, x, élément
casquette, bleu, bicyclette, sacoche, rue
poste, lettre, puissance, alpha, produit
courrier, poste, besace, mobylette, morsure chien
poste, courrier, recommandé, livreur, vélo
lettre, chien, vélo, courrier, jaune
Poste, Courrier, Jaune, Nouvelle, recommandé
courrier, vélo, poste, bonne nouvelle, facture
courrier, jaune, chien, multiplication, bleu
lettre, courrier, poste, vélo, colis
vélo, jaune, boîte aux lettres, attente, nouvelles
lettre, poste, bicyclette, ennui, ronde
lettre, nouvelles, surprise, chien, vélo
courrier, poste, chien, timbre, recommandé
poste, lettre, vélo, colis, postier
lettre, Poste, facture, nouvelle, vélo
poste, lettre, courrier, colis, retard
PTT, distribution, vélo, costume, lettre
poste, lettre, colis, vélo, casquette
poste, matin, lettre, vélo, jaune

Table 7: Associations sets proposed by all users for the target word facteur, presented in the same order as the users proposed them.