Lexical access via a simple co-occurrence network

Trouver les mots dans un simple réseau de co-occurrences

Gemma Bel-Enguix[1] Michael Zock
CNRS-LIF, UMR 7279, Aix Marseille Université, Marseille
gemma.belenguix@gmail.com, michael.zock@lif.univ-mrs.fr

RÉSUMÉ

Au cours des deux dernières décennies des psychologues et des linguistes informaticiens ont essayé de modéliser l'accès lexical en construisant des simulations ou des ressources. Cependant, parmi ces chercheurs, pratiquement personne n'a vraiment cherché à améliorer la navigation dans des 'dictionnaires électroniques destinés aux producteurs de langue'. Pourtant, beaucoup de travaux ont été consacrés à l'étude du phénomène du *mot sur le bout de la langue* et à la construction de réseaux lexicaux. Par ailleurs, vu les progrès réalisés en neurosciences et dans le domaine des réseaux complexes, on pourrait être tenté de construire un simulacre du dictionnaire mental, ou, à défaut une ressource destinée aux producteurs de langue (écrivains, conférenciers). Nous sommes restreints en construisant un réseau de co-occurrences à partir des résumés de Wikipedia, le but étant de vérifier jusqu'où l'on pouvait pousser une telle ressource pour trouver un mot, sachant que la ressource ne contient pas de liens sémantiques, car le réseau est construit de manière automatique et à partir de textes non-annotés.

ABSTRACT

During the last two decades psychologists and computational linguists have attempted to tackle the problem of word access via computational resources, yet hardly none of them has seriously tried to support 'interactive' word finding. Yet, a lot of work has been done to understand the causes of the *tip-of-the-tongue problem* (TOT). Given the progress made in neuroscience, corpus linguistics, and graph theory (complex graphs), one may be tempted to emulate the mental lexicon, or to build a resource likely to help authors (speakers, writers) to overcome word-finding problems. Our goal here is much more limited. We try to identify good hints for finding a target word. To this end we have built a co-occurrence network on the basis of Wikipedia abstracts. Since the network is built automatically and from raw data, i.e. non-annotated text, it does not reveal the kind of relationship holding between the nodes. Despite this shortcoming we tried to see whether we can find a given word, or, to identify what is a good clue word.

MOTS-CLÉS: accès lexical, anomie, mot sur le bout de la langue, réseaux lexicaux
KEYWORDS: lexical access, anomia, tip of the tongue (TOT), lexical networks

1 Introduction

Lexical choice is an obligatory step in language production. During this stage, the

[1] This work has been supported by the European Commission under a Marie Curie Fellowship.

author (speaker or writer) has to select a word expressing the concept or idea he/she has in mind. Of course, before choosing a word, one must have accessed a set of words from which to choose. While writers may use an external resource (dictionary) in case of word finding problems, speakers always rely on the internal or mental lexicon (human brain) which is known for its remarkable organisation. It is still a matter of debate where and in what form words are stored in the brain, yet, there is a general belief concerning dictionaries, namely: the bigger (the more entries), the better. While making sense from a practical point of view, this statement may nevertheless be misleading. Storage does not imply accessibility. This is well known via the 'tip of the tongue'-problem (TOT, Brown & McNeill, 1996; Brown, 1991)[2], but this holds also for electronic resources. For example,, variations of the input (query) or variations concerning the principle underlying the building of the resource may affect considerably the success of finding a given target word (Zock & Schwab, 2013). While authors need dictionaries, the latter are only truly useful if the words they contain are easily accessible. To allow for this we need good indexes (Zock & Schwab, 2013).

Lexical access has been widely studied and modelled by psychologists (Dell, 1986; Levelt et al. 1999). However, none of this work addresses the problem of word finding via an electronic resource. The work done by computational lexicographers is generally based on the readers' needs: words are listed alphabetically, and little if any provision is made to allow for conceptual input. Indeed, what kind of information (query, conceptual input) should a user give if the target words are 'avatar', 'tiara' or 'eschatology'? While there are many kinds of dictionaries, only very few of them are really helpful for the writer or speaker. Still, great efforts have been made to improve the situation. In fact, there are quite a few *onomasiological* dictionaries, like *Roget's Thesaurus* (Roget, 1852), and various network-based dictionaries, with *WordNet* (Fellbaum, 1998; Miller et al., 1990) being the best known. There are also various collocation dictionaries (BBI, OECD), reverse dictionaries (Edmonds, 1999, or Wordsmyth, www.wordsmyth.net) and *OneLook*, which combines a dictionary (*WordNet*) and an encyclopedia (*Wikipedia*). Finally, there is *MEDAL* (Rundell and Fox, 2002), a thesaurus produced with the help of Kilgariff's *Sketch Engine* (Kilgarriff et al., 2004).

Despite its shortcomings, of all these proposals WordNet (WN) clearly stands out. While being built manually, it embodies a number of features known from the mental lexicon: the lexicon is a multidimensional network whose nodes (words) are linked via various kinds of relations. WNs have been built for many languages (http://www.globalwordnet.org), and the initial resource has been adapted and improved, to yield eXtendedWN, (Mihalcea et Moldovan, 2001) an application able to support a great number of tasks in NLP.

Other networks have been built differently. For example, JeuxDeMot (JdM, Lafourcade, 2007) was built via a huge community (crowdsourcing) playing games. The approach is similar to other web-based resources, like Open Mind Word Expert (Mihalcea et

[2] The TOT-problem consists in the fact that an author knows a word, but is occasionally unable to access it. Typically, he has activated most of the target's features, but fails to retrieve some of the crucial, final, sound related fragments. This is why the speaker has the impression that the word has nearly made it, but not quite. The word is stuck on the tip of the tongue.

Chklovski, 2003) and SemKey (Marchetti *et al.*, 2007). JdM is coupled with AKI (Joubert et Lafourcade, 2012), which is supposed to allow for word access. To what extend this is truly so remains an empirical question, despite the fact that the initial results look quite promising (Joubert et al. 2011).

Zock et al. (2010) propose an association-based index to support interactive lexical access for language producers. To this end they suggest to build a matrix on the basis of co-occurrences. Put differently, they try to capture word associations and the links holding between them. This approach seems attractive as the network is built automatically, corpus-based, computer-supported, and the resource allows for graph-based analysis (relative distance, clustering effects, etc.). However, this work is also confronted with some unsolved problems like disambiguation of the input (query, clue), explicitation of the link type and clustering of the output (the answers given in response to a query). Usability will be hampered as long as all this cannot be done automatically. We try to tackle a similar problem, but we do not address interactive search, but only automatic access. More precisely, we try to address the tip-of-the-tongue problem by using a graph-kind of approach. Overall, the following ideas underlie this work:

a) usage of an non-annotated source, containing a large number of words;
b) structuring of the lexicon in the simplest way possible, i.e. by relying only graph theory and statistics;
c) exclusive usage of co-occurrences for building the graph. Semantic relations are ignored at this stage;
d) exclusive reliance on automatic processing (hence, no manual annotations);
e) conception of very simple graph search algorithms.

The approach is extremely simple. We use co-occurrences because it is a straightforward way to structure words on the basis of weights, i.e. numerical values. We do not claim any cognitive relevance other than statistics, which seem nevertheless to work when modelling language production (Levelt et al. 1999).

2 Co-occurrence network

Our goal is to build a co-occurrence graph able to achieve similar results to the ones of annotated systems. To achieve this goal, we decided to start with a large, non-annotated corpus: the entire set of Wikipedia's abstracts, i.e. almost 4 million documents. To build the graph our system runs through a pipeline of five modules:

1. document cleaning (deletion of stop-words);
2. parsing of the abstracts and extraction of 'Nouns' and 'Adjectives';
3. lemmatisation of word forms to avoid duplicates (horse, horses);
4. computation of the (un-directed) graph's nodes. Links are created between direct neighbours;
5. computation of the edges' weights. The weight of an edge is equal to the number of its occurrences. We only use absolute values.

Performing the above described operations yields a graph of 1.595.133 (different) nodes, of which nearly half (48%, i.e. 765.081) are happaxes, that is, terms occurring only once within the source. In order to understand the reason for this, one must take

into account the nature of the resource, and the nature of the words used.

Since our source is an encyclopaedia, it contains an unusually high number of terms related to science, history, peoples' names or names of geographical locations, concepts from other languages... Concerning the extracted words, it should be noted that only nouns and adjectives were used. The deletion of verbs and adverbs is motivated by practical considerations: decreasing the size of the network alleviates processing. Put differently, our choice has been made only for this specific experiment. We wanted to focus only on nouns and adjectives, maintaining them even if their weights are very low. Stop words have been also eliminated, but for a different reason. They are hardly ever used as 'clues', and using them nevertheless may bias the results. Finally, we get a weighted list of nodes.

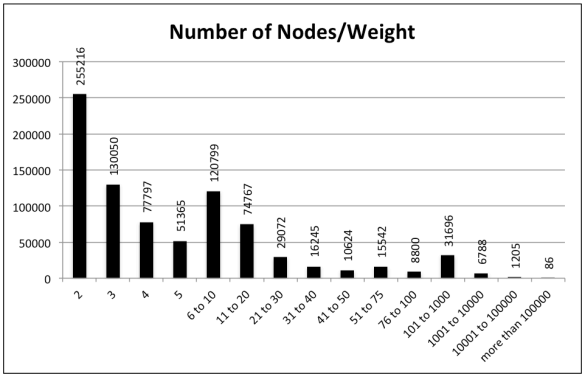


FIGURE 1 – Weights of the nodes of the graph

Figure 1 shows the distribution of frequencies. The weight of most nodes is below 10, speaking in absolute terms. Yet, 86 words are solid hubs with more than 100000 occurrences. Here are the 20 nodes with the greatest weight:

[(State, 502915), (Born, 424243), (New, 349236), (County, 348655), (District, 344620), (First, 339583), (American, 330643), (United, 320260), (School, 280589), (Village, 277337), (City, 276718), (Album, 272357), (Film, 260753), (National, 251727), (Family, 247912), (University, 239137), (Year, 238700), (South, 236760), (Part, 231373), (Football, 224046)]

Note, that the weight of more than 2/3 of edges is 1, the weight of the remaining third is >1, the proportion being 69/31. Moreover, there is only one edge with a value greater than 100000, *state-united*, i.e. 'United States', the most frequently mentioned co-occurrence in the Wikipedia abstracts. The weight of the following edges exceeds 30000:

[(State United, 152347), (High School, 70053), (New York, 66052), (War World, 59523), (Administrative District, 58113), (Census Population, 55299), (District Gmina, 51501), (Administrative Village, 46922), (New Zealand, 44320), (Football League, 42994), (Kingdom United, 39798), (Album Studio, 36887), (Olympics Summer, 34421), (Ii War, 33800), (Railway Station, 33723), (Capital Regional, '3300)]

The distribution of edges' weights is shown in Figure 2. Data whose value is 1 are omitted in the figure.

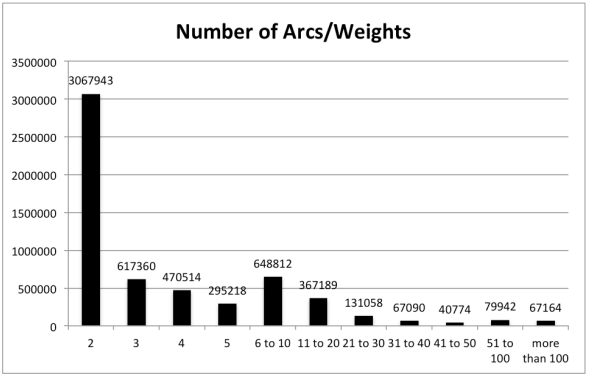


FIGURE 2 – Weights of the edges of the graph

3 Search algorithm

The search of the target word $\mathcal{T}$ in a graph $\mathcal{G}$, is done via some clues, say c_1, c_2, c_3, (mouse, Prolog, Java, in figure 3) which act as inputs. $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ stands for the graph, with $\mathcal{V}$ expressing the set of vertices and $\mathcal{E}$ the set of edges. The clues c_1, c_2, $c_3 \in \mathcal{V}$. N(i) expresses the neighbourhood of a node ($i \in \mathcal{V}$) and is defined as 'every $j \in \mathcal{V} \mid e_{i,j} \in \mathcal{E}$. The search algorithm is as follows:

- Define the neighbourhood of c_1, c_2, c_3, $N(c_1)$, $N(c_2)$, $N(c_3)$;
- Get the set of nodes $V_T = N(c_1) \cap N(c_2) \cap N(c_3)$ and consider $V_c = \{c_1, c_2, c_3\}$ to be the set of nodes representing the clues. We define a subgraph of $\mathcal{G}$, G_T, that is a complete bipartite graph, where every element of V_T is connected to every element of V_c;
- Rank the nodes of V_T according to their strength (s) in G_T. For every v in V_T, $s_v = 1/3\ (w(vc_1) + w(vc_2) + w(vc_2))$.

4 Performance

Taking random examples, the system's capacity to find words is remarkably good, provided that all the clues are from the same domain. Otherwise performance may degrade: compare (a1, b1) and b2. In the first two cases the target appears on top of the list, whereas in b-2 the target word gets demoted to the 13th position. Being from a different domain, the clue 'India' impedes performance. On the other hand, widening the clues' semantic scope has as a positive effect, see c1, c2.

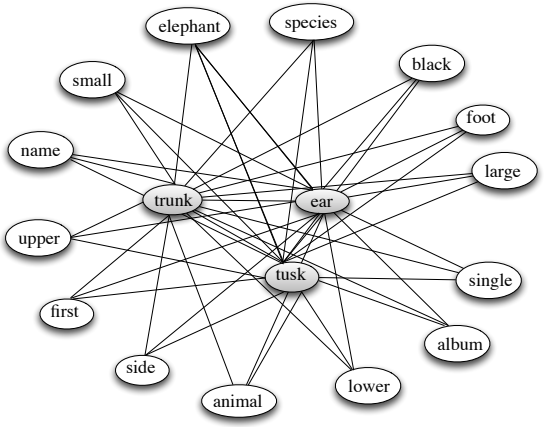
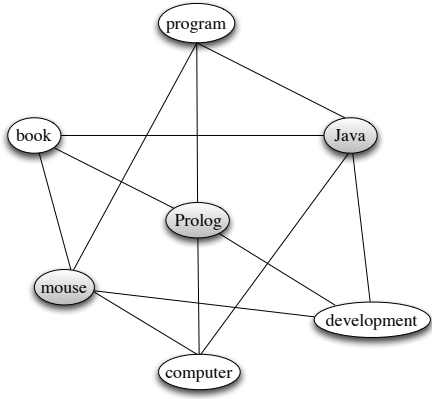


FIGURE 3A – Graph G_T for (tusk, trunk, ear)



3B –G_T for (mouse, Prolog, Java)

a) Target: '*hand*':

– The clues 'finger', 'wrist', 'glove', yield 9 hits, displaying the target in the first position: 1 (**hand**, 153); 2 (right, 29); 3 (arm, 25); 4 (part, 24); 5 (first, 21); 6 (side, 18); 7 (worn, 17); 8 (person, 12); 9 (game, 8).

b) Target: '*elephant*':

1. By entering the words 'tusk', 'trunk', 'ear' (figure 3a), we get a list of 14 items of which the first 10 are as follows: 1 (**elephant**), 51; 2 (upper), 28; 3 (species, 28); 4 (single, 25); 5 (lower, 24); 6 (small, 23); 7 (album, 22); 8 (large, 19); 9 (name, 18); 10 (side, 17).

2. If we provide 'tusk', 'trunk', 'India', we get the target in the 13th position, right after 'first', 'year', 'country', 'name', 'member', 'species', 'born', 'family', 'small', 'large', 'long', 'upper', '**elephant**'.

c) Target: '*computer*':

1. The clues 'mouse', 'keyboard', 'screen' produce a large number of hits. The program displays only the first fifty. 1 (player, 600); 2 (**computer**, 264); 3 (first, 192); 4 (appearance, 191); 5 (name, 178); 6 (album, 99); 7 (small, 90); 8 (role, 89); 9 (music, 89); 10 (band, 82).

2. The clues 'mouse', 'Prolog' and 'Java' (figure 3-b) produce only four hits: 1 (program, 58); 2 (**computer**, 47); 3 (development, 31); 4 (book, 16)

The given examples could make us believe that the program works quite well. While being true, this is not always the case. For example, when we tried the examples used by (Zock et Schwab, 2011), namely, 'wine', 'harvest', 'grape', the system was unable to find the target word 'vintage'. On the other hand, by changing slightly the input, providing 'vintage', 'harvest', and 'grape', we did get '*wine*' in the first position and with a very strong score (735). This suggests both a conclusion and a question: (a) the

algorithm is not yet good enough, since it works in some cases, but not in others; (b) since some terms are definitely better triggers or cues than others, we may wonder what are good cue words, and to this end we could use this resource in order to answer this question empirically. This is a possibility we are currently exploring.

5 Conclusions

Experiments done with the resource built on the basis of the co-occurrences extracted from Wikipedia shows that it allows for accessing words. It also shows, if ever necessary, that not all words are equally good as inputs. This being so, we could use this resource as a workbenchto find out empirically which words, or which specific kind of words are good inputs for a given target word.

While there is little doubt that Wikipedia is a quite useful source, it does also have its shortcomings. For example, it does not contain episodic knowledge (information concerning current events, anecdotes,...), hence, it may be good to consider other types of texts containing more common words (authentic exchanges between people).

Concerning the system's performance one may conclude that it is quite good, but we should bear in mind that we dealt with automatic access and not interactive word finding. While the number of hits is (within limits) of little importance in the former case —(computers will find quickly a word even in a huge list, say, a list of 3000 tokens),— it becomes a critical issue in the latter case. This is why typing the links, or clustering the output is an important component for supporting interactive word search. This being said, getting a clearer picture concerning clues may still be of interest for those interested in designing tools to support word access.

References

BROWN, A. (1991). A review of the *tip of the tongue* experience. *Psychological Bulletin*, 10, pages 204-223

BROWN, R. et MC NEILL, D. (1966). The tip of the tongue phenomenon. *Journal of Verbal Learning and Verbal Behavior*, 5, pages 325-337

DELL, G. (1986). A spreading-activation theory of retrieval in sentence production. *Psychological Review*, 93, 283-321.

EDMONDS, D. (ed.) (1999). The Oxford Reverse Dictionary, *Oxford University Press*, Oxford, 1999.

FELLBAUM, C. (éd.) (1998). WordNet: An Electronic Lexical Database and some of its Applications. Cambridge, MA: MIT Press.

JOUBERT, A., LAFOURCADE, M. (2012). A new dynamic approach for lexical networks evaluation. In Choukri et al. (eds.), Proceedings LREC'12 (*Eight International Conference on Language Resources and Evaluation*), Istanbul, Turkey, European Language Resources Association (ELRA).

JOUBERT, A. LAFOURCADE, M., SCHWAB, D. et ZOCK, M. (2011). Évaluation et consolidation d'un réseau lexical via un outil pour retrouver le mot sur le bout de la langue. Actes de

la 18ème conférence sur le Traitement Automatique des Langues Naturelles (TALN), Montpellier, pp. 295-306

KILGARRIFF, A., RYCHLY, R., SMRZ, P. et TUGWELL, D. (2004). *The Sketch Engine.* Proceedings of the 11th Euralex International Congress. Lorient, France, pages 105-116

LAFOURCADE, M. (2007). Making people play for lexical acquisition. In *Proceedings SNLP 2007* (*7th Symposium on Natural Language Processing*), Pattaya, Thailande.

LEVELT, W., ROELOFS, A. et MEYER, A. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences*, 22, pages 1-75.

MARCHETTI, A., TESCONI, M., RONZANO, F., ROSELLA, M. et MINUTOLI, S. (2007). SEMKEY. A semantic collaborative tagging system. *In Proceedings of WWW2007*, Banf, Canada.

MIHALCEA, R. et MOLDOVAN, D. (2001). Extended wordnet: progress report. *In NAACL 2001 (Workshop on WordNet and Other Lexical Resources)*, Pittsburgh, USA.

MIHALCEA, R. et CHKLOVSKI, T. (2003). Open Mind Word Expert: Creating large annotated data collections with web userp's help. In *LINC 2003* (*Proceedings of the EACL 2003 Workshop on Linguistically Annotated Corpora*), Budapest.

MILLER, G., BECKWITH, R., FELLBAUM, C., GROSS, D., MILLER, K. (1990). Introduction to WordNet: an on-line lexical database. *International Journal of Lexicography* 3, pages 235-244.

PENNACCHIOTTI, M., PANTEL, P. (2006). A bootstrapping algorithm for automatically harvesting semantic relations. In Proceedings of ICoS (*Inference in Computational Semantis*), Boxton, England, pages 87-96

RUNDELL, M. et FOX, G. (eds.) (2002). Macmillan English Dictionary for Advanced Learners (MEDAL). Oxford

ROGET, P. (1852). *Thesaurus of English Words and Phrases*. Longman, London

TURNEY, P.D. (2006). Similarity of semantic relations. *Computational Linguistics* 32, pages 379-416

ZOCK, M., FERRET, O., SCHWAB, D. (2010). Deliberate word access: an intuition, a roadmap and some preliminary empirical results. *Int J Speech Technol* 13, pages 201-218

ZOCK, M. et SCHWAB, D. (2011). Storage does not garantee access: The problem of organizing and accessing words in a speaker's lexicon. *Journal of Cognitive Scien*ce 12, pages 233-259

ZOCK, M. et SCHWAB, D. (2013) L'index, une ressource vitale pour guider les auteurs à trouver le mot bloqué sur le bout de la langue. In Gala, N. et M. Zock (éds). Ressources lexicales : construction et utilisation. Lingvisticae Investigationes, John Benjamins, Amsterdam, The Netherlands