Semantic Context Visualization to Promote Vocabulary Learning

Caroline Barrière
(Caroline.Barriere@nrc-cnrc.gc.ca)
Conseil National de Recherche du Canada
Claude St-Jacques
(Claude.St-Jacques@nrc-cnrc.gc.ca)
Conseil National de Recherche du Canada

Traditional access through the alphabetically organized macrostructure (words defined) of the dictionary was convenient in a printed form. Online versions of dictionaries lead us to rethink our access approach and to exploit, as suggested by Humblé, the increased value of the dictionary in computer assisted language instruction. In a self-learning environment, a situation favored by today's wide-spread access to computers, an online dictionary is a valuable resource for reading comprehension. However, learners can quickly become discouraged if the information about a word searched for is buried among too much other information, e.g. the many definitions listed for highly polysemous words. Our current research suggests a method for providing specific guidance to a user to ease his access to information and promote vocabulary learning during his dictionary searches.

We present a tool, *REFLEX*, built on a mathematical model of a fuzzy logic search engine. We suggest that the microstructure (the content of the entries) of a dictionary be considered as a corpus. From this corpus, using fuzzy operators, we can calculate a similarity matrix, called a fuzzy pseudo-thesaurus (Miyamoto), expressing the degree of association between each pair of lemma (base forms of words) found in the corpus. This similarity calculation is based on the tendency of two words to co-occur within sentences. The fuzzy pseudo-thesaurus is pre-calculated and used at the search (query) time (Klir and Yuan).

Presently, the tool is for English learners of French and uses a learner`s dictionary called *DAFLES* (*Dictionnaire de l'Apprenant du Français Langue Seconde et Étrangère*) (Verlinde et al.) merging all its defining sentences to build a corpus. Portability to languages other than French would require work at the pre-processing stage, for example to tokenize the sentences (split them into word units) and lemmatize the words (e.g. lemma should be found for diverse forms of nouns and verbs). Portability to other types of dictionaries would require an understanding of the different types of information (definitions, examples, notes) contained before merging them into a corpus. Each type of information might be of different value to a learner, and then could be given more or less weight in our model.

REFLEX provides a graphical visualization of a word's semantic context. Figure 1 shows the portion of the pseudo-thesaurus relating to the word "scientifique" (scientific) as automatically built from *DAFLES*. The surrounding words (*chercheur* – researcher, *étudier* – study, *observation* – observation, *théorique* – theoretical, *rechercher* – research, etc.), express words related to *scientifique*, their graphical distance being proportional to their calculated distance.

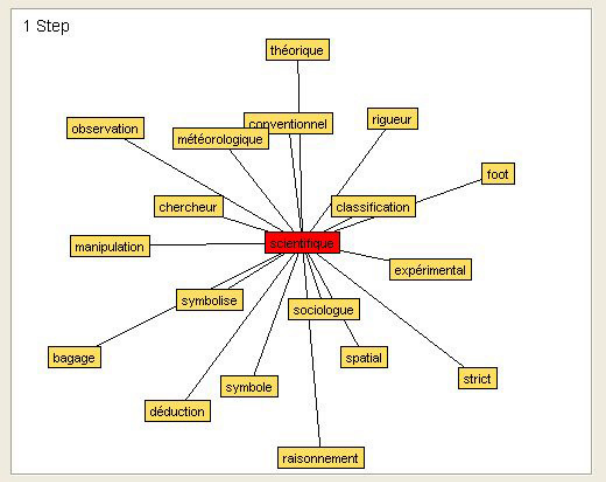

Figure 1: Semantic context for scientifique

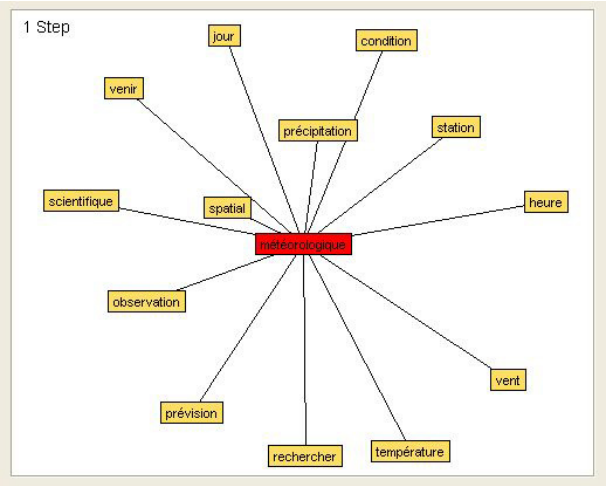

Figure 2: Semantic context for météorologique (meteorological)

Semantic contexts give a graphical representation similar to semantic maps which are widely used in language teaching to favor vocabulary acquisition (Brown and Perry; Carrell; Crow and Quigley). Semantic maps are usually provided in classroom settings and are manually constructed by instructors interactively with the students. Here semantic contexts are produced in

real-time, and *REFLEX* provides navigation capabilities for the user to easily move to other portions of the pseudo-thesaurus. For example, a click of a mouse on the word *météorologique* within the semantic context of Figure 1 brings the learner to the semantic context shown in Figure 2.

The implicit relations shown through the arcs linking the concepts are not limited to paradigmatic relations as found in *WordNet* (Fellbaum) or *MindNet* (Richardson et al.)[1] For example, the association between *observation* and *researcher* in Figure 1 could not be labeled with a paradigmatic relation. The complex relation, *typical activity*, is closer to a lexical function as defined by Mel'cuk.

REFLEX also serves as a guide for dictionary searches by making use of information found in the context of occurrence of the unknown word. It leads the learner not necessarily to the entry corresponding to that unknown word, but to any information from dictionary entries likely to help the understanding of the word within that particular context. This provides a clear filter for polysemous words and helps the learner identify among all the possible definitions and examples, the ones most relevant to their current reading situation.

Let us take for example the French polysemous word *culture*, which could relate to *cultivate* (flowers or vegetables) or *culture* (social knowledge). The following sentence is taken from a text "L'école dans les deux langues" shown in a reading comprehension software (*DidaLect*, Duquette et al.): "*En étudiant la langue et la culture de l'autre, Anglo-Saxons et Latinos apprennent en même temps à se connaître et à se respecter.*" .

Assuming the word culture is unknown to a learner, she launches a search. The context window of occurrence of that word is analyzed to obtain a query vector made up of the searched word and its neighbours (*culture, étudier, langue, apprendre, connaître, respecter*)[2]. This query vector is expanded on via the fuzzy pseudo-thesaurus to include related words, each with a weight corresponding to their similarity to the original query vector (*culture 1.0, étudier 1.0, langue 1.0, apprendre 1.0, connaître 1.0, respecter 1.0, didactique 0.1, pédagogique 0.09, éducatif 0.09, créatif 0.08, race 0.08, impoli 0.06, curiosité 0.06, brassage 0.06, fermier 0.06, formateur 0.06*)[3]. The enlarged vector then represents an extended context of a word, and is used in an information retrieval task to extract interesting definitions and examples pertaining to that context from the corpus of sentences. These sentences are presented to the user in decreasing order of relevance, as shown in Table 1 (relevance is given in first column), and come from various polysemous entries.

Rel.	Sentence	Entry	Polysemy of that entry
0.6	*Lorsqu'une personne étrangère s'assimile (à une société, une culture), elle s'intègre dans cette société, dans cette culture, elle adopte…*	*assimiler*	3 other meanings (confound, understand, group)
0.4	*Le brassage d'idées, de plusieurs choses, personnes, cultures ou races est le mélange ou la combinaison d'idées, de plusieurs choses, personnes…*	*brassage*	1 other meaning (beer brewing)
0.4	*Lorsqu'un élève, un étudiant révise un cours, il étudie, il parcourt à nouveau un cours qu'il a déjà appris.*	*réviser*	2 other meanings (reviewing a text inspect a vehicle)
0.4	*Lorsque quelque chose appartient à un endroit, à une période, à une culture, à quelque chose, cette chose est caractéristique de cet endroit, …*	*appartenir*	5 other meanings (belong to a group, own something, have to do something, have responsibility of something, part of a machine)

Table 1 – Relevant sentences for the understanding of the word culture in context

Using the extended context of a word, combined with knowledge of word association as pre-calculated in the pseudo-thesaurus, *REFLEX* gathers the distributed information from the dictionary to help the understanding of a word (often polysemous) as expressed in a specific reading context.

Yet, free browsing is not necessarily a bad thing for learners. In fact, researchers in second language learning (Aston) debate on the usefulness of corpora in which learners could be left to wander among sentences by themselves. On one hand, free browsing mode encourages autonomous discovery of different or new senses of words, but on the other hand, too much discovery within a reading comprehension task might take the learner too far from the original goal. Furthermore, both activities (free or guided exploration) are more or less appropriate depending on the learner`s profile (efficient or less efficient learner).

Although intended to guide, *REFLEX* is easily adaptable toward a discovery mode by simply removing the contextual cues from the query vector. The flexibility of *REFLEX* can also be seen in the presentation of the semantic maps. For efficient learners, larger maps can be shown, such as presented in Figure 1, but for less efficient learners, maps can be restricted to only contain the few closest words.

In conclusion, *REFLEX* shows much potential for online dictionary exploration. It is constructed on sound mathematical principles and has potential for adaptability to different learning purposes and types of learners. As our next step, we will integrate *REFLEX* as a module within *DidaLect*, a reading comprehension software, with focus on adaptability parameters. Investigation into automatic labeling of associations is an ambitious longer term research goal. Finally, we hope for future explorations with other dictionaries and other languages.

---

1. Paradigmatic relations, such as hyperonymy, meronymy, synonymy, are of great value in language learning and do provide an explicit way for learners to organize the new vocabulary in their own knowledge stores. The purpose of semantic contexts is complementary to this organization by providing a semantic field around a word.
2. Lemmatized forms of content common nouns are taken. We do not use proper nouns (they are not in the dictionary) or function words.
3. We only show values larger than 0.06, but this threshold could be adjusted.

Bibliography

Aston, G. *Learning with Corpora.* Houston: Athelstan, 2003.

Brown, T.S., and F.L. Perry Jr. "A comparison of three learning strategies for ESL vocabulary acquisition." *TESOL Quarterly* 25 (1991): 655-670.

Carrell, P.L. "Content and Formal schemata in ESL reading." *TESOL Quarterly* 21 (1987): 461-481.

Crow, J.T., and J.R Quigley. "A semantic field approach to passive vocabulary acquisition for reading comprehension." *TESOL Quarterly* 19 (1985): 497-513.

Duquette, L., A. Desrochers, and S. Szpakowicz. "Adaptive Courseware for Reading Comprehension in French as a Second Language : The Challenges of Multidisciplinarity in CALL." *Proceedings of the eleventh International CALL conference, University of Antwerp.* Antwerp, 5-7 September 2004. 85-92.

Fellbaum, C. *WordNet: An Electronic Lexical Database.* Cambridge, Mass: MIT Press, 1998.

Humblé, P. *Dictionaries and Language Learners.* Frankfurt am Main, Germany: Haag + Herchen Verlag GmbH., 2001.

Klir, G.J., and B. Yuan. *Fuzzy Sets and Fuzzy Logic.* Upper Saddle River, NJ: Prentice Hall, 1990.

Mel'cuk, I. "Semantic description of lexical units in an explanatory combinatorial dictionary: Basic principles and heuristic criteria." *International Journal of Lexicography* 1.3 (1988): 165-188.

Miyamoto, S. *Fuzzy Sets in Information Retrieval and Cluster Analysis.* Dordrecht, Netherlands: Kluwer Academic Publishers, 1990.

Richardson, S.D., W.B. Dolan, and L. Vanderwende. "MindNet: Acquiring and Structuring Semantic Information from Text." *Proceedings of the ACL'98.* Montreal, 1998. 1098-1102.

St-Jacques, C., and C. Barrière. "L'inférence dictionnairique : de la créativité poétique à celle du raisonnement flou." *Cahiers de lexicologie* 85 (2004): 129-155.

Verlinde, S., , and GRELEP (Groupe de Recherche en Lexicographie Pédagogique). *Dafles (Dictionnaire d'apprentissage du français langue étrangère ou seconde).* . Accessed 2005-04-11. <http://www.kuleuven.ac.be/dafles/acces.php?id=>