Scientific Report Short Term Scientific Mission IS1305-21159

COST STSM Refer. Number: COST-STSM-IS1305-21159
Period: 2014-09-16 to 2014-09-28
COST Action: IS1305
STSM type: Regular (from France to Austria)
Guest/STSM applicant: Michael Zock, CNRS-LIF, Marseille, France
Host: Eveline Wandl,
Austrian Academy, 1010 Wien, Sonnenfelsgasse 19,
eveline.wandl-vogt@oeaw.ac.at

1 Purpose of the STSM

The goal of this *Scientific mission* was the definition of a roadmap to enhance existing electronic dictionaries to help authors (speakers/writers) overcome the tip-of-the-tongue problem (Brown & McNeill, 1996). Put differently, our ultimate goal is to help authors to find an elusive word.

Whenever we need a word, we look it up in the place where it is stored, the dictionary or the mental lexicon. Since it is unreasonable to perform search in the entire lexicon, I suggest to reduce this space in two steps. Given some input (seeds, i.e. words other than the target coming to the author's mind) the system considers all directly related words (step-1), which it organizes into a *categorial tree* allowing to present then the potential target words in a clustered and labeled form (step-2). In sum, I suggest to build (automatically) an association thesaurus and an automatic categorization method (categorial tree) to support navigation. Hence, rather than looking in a huge flat list, the user searches now only in the relevant part of the tree. This reduces not only the number of candidates among which to choose, but also the time needed to carry out search.

2 Description of the work carried out during the STSM

During my stay I have mainly worked on the roadmap mentioned here above (see below).

In addition I have given a talk at the Academy of Science (title : "Wheels for the mind of the language producer: microscopes, macroscopes, semantic maps and a good compass.") with the goal to show some of my work on language production.

In particular, I've tried to show how one can account for the fact that humans manage to produce language on the fly, i.e. spontaneously. I've tried to illustrate this for the building of syntactic structures, a problem theoretical linguists never even attempt, since it is outside of their scope. Dealing with competency, they describe structures and label them, but they never show how one gets from some meaning (concepts) to its corresponding form (language). Put differently, they do not show how one manages to map a conceptual structure on its linguistic counterpart.

In the second part of the talk I've presented a system designed to help people to learn rapidly to 'speak' a foreign language. The system is an open, i.e. a customizable web-based tool, allowing users to specify the words and structures they want to learn. It is generic and it supports the learning of typologically different european and oriental languages (Japanese,

Chinese). The last part of my talk was devoted to lexical access, the focus of this mission.

Next to this talk, I've tried to help establish links between Austrians colleagues working on related topics but not even knowing of their mutual existance. Indeed, I've known for many years the work of a brilliant Austrian researcher (Werner Winniwarter) who has done some work related to mine, and I've presented him to the colleagues of the Academy to see whether they could not do some work together. To this end I've organized a meeting where they've started to talk to each other and these contacts have continued beyond my stay in Vienna, possibly leading to a joint project.

3 Description of the main results obtained

I will quickly describe here the roadmap as it will be the backbone of my future work.

1.1 Background or problem : How to find the word that is eluding you?

One of the most vexing problems in speaking or writing is that one knows a given word, yet one fails to access it when needed. Suppose, you were looking for a word expressing the following ideas: 'superior dark coffee made of beans from Arabia', but could not retrieve the intended form 'mocha'. What will you do in a case like this? You know the meaning, you know how and when to use the word, and you even know its form, since you've used it some time ago, yet you simply cannot access it at the very moment of speaking or writing? Since dictionaries generally contain the target word, they are probably our best ally to help us find the form we are looking for. This being said, storage does not guarantee access. The very fact that a dictionary contains a word does not guarantee at all that we will also be able to find or locate it (Zock & Schwab, 2013; Tulving & Pearlstone, 1966).

Dictionary users typically pursue one of two goals (Humble, 2001): as *decoders* (reading, listening), they are generally interested in the meanings of a specific word, while as *encoders* (speakers, writer) they wish to find the form expressing an idea or a concept. This latter task is our goal. While most dictionaries satisfy the reader's needs, they do not always live up to the authors' expectations, helping them to find the elusive word. To be fair, one must admit though that great efforts have been made to improve the situation. In fact, there are quite a few onomasiological dictionaries. For example, Roget's Thesaurus (Roget, 1852), analogical dictionaries (Boissière, 1862, Robert 1993), Longman's Language Activator (Summers, 1993) various network-based dictionaries: WordNet (Fellbaum,1998; Miller et al., 1990), MindNet (Richardson et al., 1998), and Pathfinder (Schvaneveldt, 1989). There are also various collocation dictionaries (BBI, OECD), reverse dictionaries (Kahn, 1989; Edmonds, 1999) and OneLook which combines a dictionary, WordNet, and an encyclopedia, Wikipedia. A lot of progress has been made over the last few years, yet more can be done especially with respect to indexing (the organization of the data) and navigation. Given the possibilities modern computers offer with respect to storage and access, computational lexicography should probably jettison the distinctions between lexicon, encyclopedia, and thesaurus and unify them into a single resource.

2.2 Architecture and the roadmap

When experiencing word access problems we expect help from dictionaries, hoping to find the elusive term there. Unfortunately, so far there is still not yet a satisfying resource allowing authors (people being in the 'production mode': speakers/writers) to find easily and most of the time the resisting word. While WordNet or Roget's Thesaurus are helpful in some cases, more often than one might think, they are not. This is a problem we would like to overcome. Figure 1 displays in a nutshell our approach, word access being viewed (basically) as a two-step process: two for the user, and two for the resource builder. The task is basically finding a specific item (target word) within the lexicon. Put differently, the task is to reduce the entire set (all words contained in the lexicon) to one, the target. Since it is out of question to search in the entire lexicon, we suggest to reduce the search space in several steps, basically two.

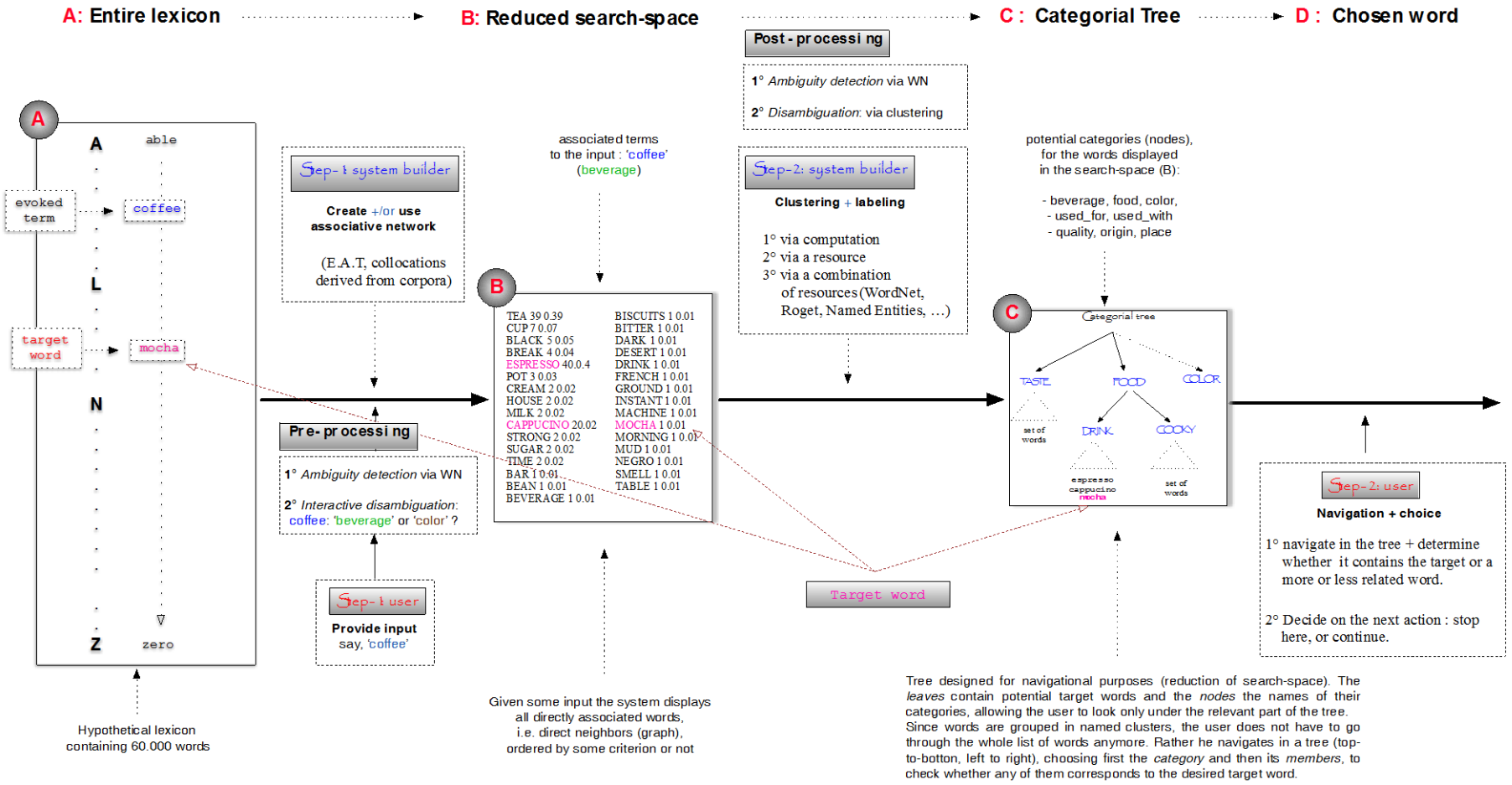

Fig. 1: Lexical access as a two-step process

4 References

Boissière, P. (1862). Dictionnaire analogique de la langue française: répertoire complet des mots par les idées et des idées par les mots. Paris. Auguste Boyer

Brown, R. & Mc Neill, D. (1966). The tip of the tongue phenomenon. Journal of Verbal Learning and Verbal Behavior, 5: 325-337

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

Fellbaum, C. (1998). WordNet: An Electronic Lexical Database and some of its Applications. MIT Press.

Humble, P. (2001). Dictionaries and Language Learners, Haag and Herchen.

Kahn, J. (1989). Reader's Digest Reverse Dictionary, *Reader's Digest*, London

Miller, G.A. (ed.) (1990): WordNet: An On-Line Lexical Database. International Journal of Lexicography, 3(4), 235-244.

Richardson, S., Dolan, W. & Vanderwende, L. (1998). Mindnet: Acquiring and structuring semantic information from text. In: ACL-COLING'98. Montréal: 1098-1102.

Robert, P., Rey A. & Rey-Debove, J. (1993). Dictionnaire alphabetique et analogique de la Langue Française. Le Robert, Paris.

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

Schvaneveldt, R. (ed.) (1989). Pathfinder Associative Networks: studies in knowledge organization. Ablex. Norwood, New Jersey, US.

Summers, D. (1993). Language Activator: the world's first production dictionary. Longman, London.

Tulving, E., & Pearlstone, Z. (1966). Availability versus accessibility of information in memory for words. Journal of Verbal Learning and Verbal Behavior, 5, 381-391

Zock, M. & 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, pp. 313-354

5 Future collaboration with the host institution (if applicable)

I have presented here the roadmap of a lexical resource whose task it to help authors to find the word they are looking for. More precisely, I have present here a framework for building a tool to support word access. To reach this goal several problems need be solved: creation of an association thesaurus to allow 'search space reduction', 'clustering the words retrieved in response to some input (available information)' and 'labeling the clusters' to ease navigation.

Before doing all this we need to define though a set of criteria that ought to be satisfied by the resources to be used. This will be one of the next steps, preceding the building or usage of existing resources.

I intend to continue collaboration with the colleagues in Vienna by possibly including colleagues from other countries and research institute. Indeed, I will spend three months next year at the university of Sapienza (Rome) to work with Roberto Navigli, the author of one of the best current resources, BabelNet.

6 Foreseen publications resulting from the STSM (if applicable)

The work here described shall naturally lead to the drafting of a proposal and in the long run possibly even to one or two publications.

7 Other comments (if any)

N/A