# Turning a Bilingual Dictionary into a Lexical-Semantic Database

**Thierry Fontenelle**
(European Commission Translation Service, Luxembourg)

Tübingen: Max Niemeyer Verlag in cooperation with the Dictionary Society of North America and the European Association for Lexicography (Lexicographica series maior, edited by Sture Allén et al., volume 79), 1997, xvi+328 pp; paperbound, ISBN 3-484-30979-2, DM 158.00, SFR 141.00, OES 1153.00

*Reviewed by*
*Christine Thielen*

In the last 15 years of computational linguistic research, a lot of work has focused on the investigation of the reusability of lexical resources (machine-readable dictionaries and corpora) because it has been realized that the development of large-scale NLP systems heavily depends on a huge and powerful lexicon.

*Turning a Bilingual Dictionary into a Lexical-Semantic Database*, based on the Ph.D. thesis of Thierry Fontenelle (University of Liège), is a new attempt to reuse an existing machine-readable bilingual dictionary, the Collins-Robert English-French dictionary (CR), this time to construct a huge lexical-semantic database (70,000 pairs of collocates). It describes in detail what work was necessary to make such a lexical resource processible for further (manual) enrichment and at last accessible for various applications.

The reader learns something about the "technical" conversion of the typesetting tape into a flexible relational database (dBaseIII+) and some of the problems with the typographic inconsistencies of a printed dictionary that could not be resolved completely automatically (Chapter 6). At this stage of the work it is already apparent how useful such a database could be, providing access to the dictionary through different starting points, not only by the headword. Now it is possible to generate a monofunctional view of a multifunctional dictionary (designed for encoding and decoding), as was shown in the COMPASS project (Feldweg and Breidt 1996).

The main emphasis is on the lexicographical work that was required for the systematic enrichment of the database with lexical-semantic information. Pairs of collocates are related to each other syntagmatically and paradigmatically. Fontenelle prefers the descriptive power of Mel'čuk's (1988) lexical functions to Pustejovsky's (1995) qualia structure, since it is not clear how the latter can be adapted to modeling verbs. Fontenelle does not withhold the disadvantages of Mel'čuk's lexical functions (LFs) from the reader. For one thing, because of the limited number of LFs, he suggests a few additional ones (Chapter 8). For example, he proposes the PART function, which defines a part-whole relationship and which relates to world knowledge and is therefore not covered by Mel'čuk's purely lexical Meaning⇔Text Theory, but is very useful in an information retrieval context. Also, the assignment of lexical functions to a given pair of collocates was done manually, to avoid inconsistencies. Only a small number of lexical functions could have been detected automatically by the analysis of defining formulas (Chapter 7). The lexicographer uses a menu-driven interface to the database

and chooses among a predetermined part-of-speech-dependent list of possible LFs. This does not solve the problem of ambiguous lexical functions, due to the sometimes vague definitions. Fontenelle provides some linguistic tests and morphological clues (Chapter 9), but they cover only a few cases. This is clearly a deficiency (of which the author is aware), but it is unsurprising, given the magnitude of the task. This is the first attempt to apply such a theory to a really huge amount of general-language bilingual data and it shows the descriptive power of lexical functions. To give a better impression of the conversion process, here are the CR entries and the respective representation with the lexical function SING (single unit) from page 115:

**mouthful** *n [food]* bouchée

**gulp 1** *n* **(b)** *(mouthful) [food]* bouchée, goulée

**portion** *n* (...) *(of food: helping)* portion

**ration** *n (allowance: of food, goods etc)* ration

→

Sing(food) = gulp, mouthful, morsel, portion, ration...

As applications of the database, Fontenelle mentions its potential value as a testbed for linguistic theories, e.g., to extract linking information about ergative verbs (Chapter 12), and its possible use in a computer-aided language learning system to teach collocations to language learners (Chapter 14). Astonishingly, he does not say very much about the usability of the database for large-scale NLP systems, perhaps because the data has to be exported to the representation formalism used by the NLP system and because there is no lexical inheritance incorporated in the database. But in his general conclusions (Chapter 15), Fontenelle writes about attempts to link the CR database to WordNet (Fellbaum 1998), the huge English electronic thesaurus; together with the inheritance mechanisms of WordNet, the database could be very useful to information retrieval systems, for instance.

In summary, the book is very well worth reading for someone who is interested in recent advances in lexical semantics or in the more application-oriented area of computational lexicography. Perhaps Fontenelle's work may be a further step towards bridging the gap between linguists and lexicographers. The structure of the book is well thought-out and gives the reader the possibility of choosing between the more theoretical parts, with a good introduction to Pustejovsky's Generative Lexicon and Mel'čuk's Meaning⇔Text Theory, and the possible applications of the bilingual database. Each chapter closes with a short conclusion section very useful to get the main points and to prevent the reader from losing the thread. At any rate, with the construction of this database, Fontenelle and his colleagues have provided a valuable bilingual lexicon that could contribute greatly to the performance of an NLP system.

**References**

Feldweg, Helmut and Elisabeth Breidt. 1996. COMPASS. An intelligent dictionary system for reading text in a foreign language. In Kiefer, F. and G. Kiss, editors, *Papers in Computational Lexicography, COMPLEX '96*, Budapest, pages 53–62.

Fellbaum, Christiane (editor). 1998. *WordNet: An electronic lexical database*. The MIT Press, Cambridge, MA.

Mel'čuk, Igor. 1988. Semantic description of lexical units in an explanatory combinatorial dictionary: Basic principles and heuristic criteria. *International Journal of Lexicography*, 1(3): 165–188.

Pustejovsky, James. 1995. *The Generative Lexicon*. The MIT Press, Cambridge, MA.

*Christine Thielen* was a researcher at the University of Tübingen and at the German Research Center for Artificial Intelligence, Saarbrücken, where she was occupied with the construction and

maintenance of various lexical databases. In the COMPASS project she worked on the conversion of the Collins German-English dictionary. The resulting dictionary database was used for further enrichment and lookup procedures and was integrated in an intelligent dictionary system for reading text in a foreign language. Thielen's address is: Hindemithweg 14, D–69245 Bammental, Germany; e:mail: Christine.Thielen@+-online.de.