

Constructing Reversible Lexical Databases

A dictionary tool developed by SERC under the auspices of the CLVV

EUROMAP case study by Lisanne Teunissen, Nederlandse Taalunie (May 2003)

CLVV

Commissie voor Lexicografische Vertaalvoorzieningen

<http://www.taalunie.org/ /werk/woordenschat.html#CLVV>

c/o Algemeen Secretariaat of the Dutch Language Union
Postbus 10595, 2501 HN The Hague,
The Netherlands
tel.: +31 70 346 95 48 - fax: +31 70 365 98 18
clvv@ntu.nl

key contacts:

Willy Martin, president
Annemieke Hoorntje, secretary

Size:

- 5 Committee members
- 1 secretary
- approx. 18 lexicographical teams at Dutch and Flemish universities and at Dutch departments abroad

founded in 1993,
initial budget: 2 million Euro (1993-1996)

independent intergovernmental body;
the Dutch Language Union is legally and financially responsible

SERC

Software Engineering Research Centre

<http://www.serc.nl/>

Postbus 424, 3500 AK Utrecht,
The Netherlands
tel.: +31 30 230 89 66 - fax: +31 30 230 89 99
info@serc.nl

key contacts:

Harald Vogt, principal advisor
Mark van Elswijk, principal advisor

Size:

- 4 management team members
- 10 (principal) advisors
- 2 PhD students

founded in 1987 within the SPIN programme,
independent since 1992

now part of CIBIT Consultants|Educators,
a service provider in the field of knowledge management,
portals and ICT architecture

CLVV – the need for an adequate lexicographical tool

The CLVV (Commissie voor Lexicografische Vertaalvoorzieningen = Committee for Interlingual Resources) is an intergovernmental body of lexical experts set up in 1993 by the Ministry of Education and Science of both Flanders and the Netherlands in order to improve and stimulate the production of bilingual dictionaries and lexical databases with Dutch as a source or target language. The Committee has launched several lexicographical projects which are, commercially speaking, non-viable, yet of great social, economical, cultural and political relevance. Translation dictionaries and learner's dictionaries have been published for Swedish-Dutch vv, Italian-Dutch vv, Arabic-Dutch vv, Hungarian-Dutch, Dutch-Czech, and Russian-Dutch. Many more are under construction, such as Polish-Dutch vv, Danish-Dutch vv, Greek-Dutch vv, Portuguese-Dutch vv, Korean-Dutch, and Dutch-Indonesian.

“However, not only is it the Committee's task to have concrete products realized, but also to see to it that, if needed, adequate lexicographical tools and infrastructure are provided for,” says Willy Martin, president of the CLVV. Soon after its launch, the Committee observed that its lexicographical teams were in need of a generic and powerful editing tool. The majority of the teams are responsible for the dictionaries in both directions of a language pair, but as it turned out, no off-the-shelf solution was available for their preferred method of working. The Committee decided to have a prototype built of an editor with importing and exporting facilities, and – which would make it more innovative than existing editors – the power to accurately reverse lexical databases.

SERC – the solution provider

The CLVV approached SERC (Software Engineering Research Centre), a consultancy firm based in Utrecht and specialized in software engineering, software quality and system architecture. SERC was founded in 1987 as part of SPIN (Stimuleringsprojectteam Informaticaonderzoek), with the goal of increasing the level of knowledge in the area of software engineering in the Netherlands. This both publicly and privately funded programme came to an end in 1992. SERC decided to continue its activities as an independent, self-supporting company selling research and consultancy. Harald Vogt, one of the company's principal advisors, proudly says: “We have an outstanding reputation based on high-quality and current knowledge, and many satisfied customers.” SERC recently became part of CIBIT Consultants|Educators, a medium-sized consultancy firm with around seventy employees.

SERC offers its services in the following specific areas:

- Capitalizing on legacy systems
- Tailored knowledge transfer

- Software development at a mature level
- Making an informed choice for quality software
- Selecting software and tools
- System architecture review

The request by the CLVV was to analyse the lexicographical process within the bilingual projects and to develop an editor fitted within the area of software development. The SERC software developers translated the ideas and wishes of the Committee and the lexicographical teams into a working prototype. Based on the experiences, this prototype was further developed into a fully functional, very powerful editor.

OMBI – the solution

The editor developed by SERC is called OMBI (Omkeerbare Bilinguale Lexicale Databanken = Reversible Bilingual Lexical Databases). While the editing function is taking in translations from language X to language Y, OMBI simultaneously stores the reversed counterparts from Y to X. The result is a non-directional bilingual database, from which dictionaries in both directions can be automatically derived. This may seem a trivial process, but in fact it is not as the tool does not merely state that if word form x is a translation of word form y, then word form y is a translation of word form x. Only rarely is translation a straightforward symmetrical relation between word forms. Martin explains: “It is not *words* that are translated into other words, but rather *words in a specific meaning*. The English word *horse* is a translation of the Dutch word *paard*, but only in the meaning of the latter as ‘certain animal’, not in its meaning ‘certain chesspiece’.”

The advantages of using a tool like OMBI are obvious: while making an X→Y dictionary, a greater part of the Y→X dictionary is already created, reducing the amount of work drastically. Martin: “After a testing period in which language pairs such as Dutch-Estonian and English-Portuguese were used, the reduction of labour was estimated to be at least one third of the total workload (for the two databases). On the other hand, it became clear that, in order to work with OMBI properly, one needs to have a fairly thorough bilingual competence.”

Future plans

“Although OMBI has been developed with Dutch and Flemish government money this does not imply that it can only be used in projects with Dutch as a source and/or target language,” Martin points out. SERC developer Van Elswijk adds, however, that although the editor is in principle language-independent, a ‘new’ language sometimes calls for adaptations. For example, the tool had to be adjusted thoroughly for use by the Arabic team, to cater for the (right-to-left) Arabic script and the specific morphological characteristics of the language. Other new languages may impose other, unforeseen demands on OMBI.

The editor will, together with all the lexical databases, be inherited by the Nederlandse Taalunie when the CLVV comes to an end later this year. The Dutch Language Union is now in the process of setting up an HLT Agency for the management, maintenance and distribution of these and other Dutch language resources produced with government money. This agency will combine the infrastructures required for different resources to reduce the costs for both the tangible (equipment, data, software, licences, etc.) and the intangible infrastructure (experts, personnel etc.). It will ensure optimal visibility and accessibility by functioning as a one-stop-shop supplier, and allow the Dutch language area to act as a powerful partner in international HLT and related projects. Through the HLT Agency, OMBI will become available for anyone who wants to build bilingual dictionaries, commercial and non-commercial users the like (though possibly under different conditions).

It is to be expected that the full-fledged editor with all its options and parameters may be a little too complex, however, for use by individuals outside the CLVV framework. The HLT Agency will be able to offer basic technical support, but probably not to the same extent as has been provided by the CLVV. Therefore, it is now being investigated whether a simpler version of the editor (“OMBI Light”) can be implemented, one that is self-explaining and easier to use for standard purposes.

References

For more information, see:

- W. Martin & A. Tamm (1996). “OMBI: An editor for constructing reversible lexical databases”. In M. Gellerstam, J. Järborg, S.-G. Malmgren, K. Norén, L. Rogström & C. Røjder Papmehl (eds.) *Euralex '96. Proceedings I-II*. Göteborg University.
- W. Martin, U. Heid, I. Schuurman, J. Beeken & G. Laureys (1998). *On the Construction of Bilingual Dictionaries. Feasibility Study carried out by order of the European Commission*, The Hague/Stuttgart.