One database, many dictionaries — varying co(n)text with the dictionary application TshwaneLex

Gilles-Maurice de Schryver
Ghent University & TshwaneDJe HLT
gillesmaurice.deschryver@{UGent.be, tshwanedje.com}

David Joffe
TshwaneDJe HLT
david.joffe@tshwanedje.com

Abstract

This paper provides background information for a software demonstration of TshwaneLex, during which the actual use of the application is illustrated in real time. The focus of the demonstration is on two main aspects, together with a related aspect in each case, of particular interest to the ASIALEX 2005 conference. These are full Unicode support and customisable sorting on the one hand, and advanced DTD (Document Type Definition) aspects and Linked View mode on the other. Together, they provide the backbone for the claim that a single TshwaneLex database successfully provides for multiple dictionaries.

The dictionary compilation software TshwaneLex

TshwaneDJe HLT has been producing the dictionary compilation software TshwaneLex since 2002. In addition to most of the eleven South African National Lexicography Units who are currently using TshwaneLex – both for the compilation itself of their (monolingual) dictionaries, as well as for the presentation of their results on the Web – this software is now also used at a number of widely respected dictionary publishing houses such as Oxford University Press, Macmillan and Van Dale Lexicografie. Government-sponsored research centres, such as the Royal National Academy of Medicine in Spain, have also begun to build their latest reference databases around TshwaneLex. Copies of TshwaneLex have furthermore also been acquired by a variety of dictionary teams worldwide, who are compiling dictionaries for amongst others Lingála, Cilubà and Kiswahili (all spoken in Africa), Welsh, Irish and Estonian (all lesser-known European languages), Bai and Chinese (both spoken in China), Motu (an Austronesian language used in Papua New-Guinea), and Inezeño Chumash (a Native-American language from the US). Each of those languages needs its own script, and each of those projects needs its own dictionary grammar, both of which TshwaneLex provides for.

A general introduction to TshwaneLex, with a focus on a selection of lexicographic underpinnings, may be found in Joffe & De Schryver (2004), while an example of an online application that revolves around TshwaneLex has been described in De Schryver & Joffe (2004). As pointed out in those publications, TshwaneLex contains numerous unique and highly developed lexicographic features. For example:

- An advanced cross-reference system not only shows related (incoming and outgoing) cross-references of the current lemma, but also automatically updates target homonym and sense numbers when these change.
- A filter function not only allows the user to work with a subset of lemmas in the dictionary based on specified criteria, a dictionary text search function further enables complex search queries on that filtered section using Unicode regular expressions.
- A compare/merge feature visually displays differences between database versions, and allows changes to be selectively merged into the main database.

In addition to paper, dictionaries can be published on the Web with the online dictionary module, which features a sophisticated query logging system. The localisable user interface allows users to browse the dictionary in their own language, and their preferred language may further be used to dynamically customise the language of the meta-language within returned articles. This feature is also extended to the electronic dictionary module.

Full Unicode support

Unicode, the international character set standard, is supported throughout TshwaneLex, and on all levels in the dictionary database. This allows not only the ability to enter data from virtually any language, but also even the simultaneous utilisation of both Asian and Latin characters in any attribute field in the database. For languages such as Chinese, Japanese or Korean, or say Arabic or Hebrew, data can be entered directly into TshwaneLex using any of the Input Method Editors (IMEs) available in Windows 2000 or XP. See Figure 1, which shows a screenshot of an elementary bilingual English-Chinese dictionary.

Completely customisable sorting

The default sorting method supported by TshwaneLex is a configurable four-pass table-based sorting system based on the ISO 14651 standard. The four different passes are used for various characteristics that may take precedence over one another, viz. the so-called ‘base alphabet’, diacritics, uppercase/lowercase differences, and so-called ‘ignorable’ characters (typically non-alphabetic characters such as spaces and punctuation marks). This is shown in Figure 2, where the sorting tables have been configured for the Estonian alphabet.

TshwaneLex automatically takes care of the sorting of lemmas, thus freeing the lexicographer from having to do so. However, many different methods of sorting exist, and often many even for the same language, thus the question arises as to how to support any possible sorting method that may be desired. To solve this, TshwaneLex includes an extendibility mechanism whereby users can create plug-ins to add support for new sorting methods. As a result, any sorting system (e.g. by radical/stroke count or by pinyin romanised form for Chinese) may be used.

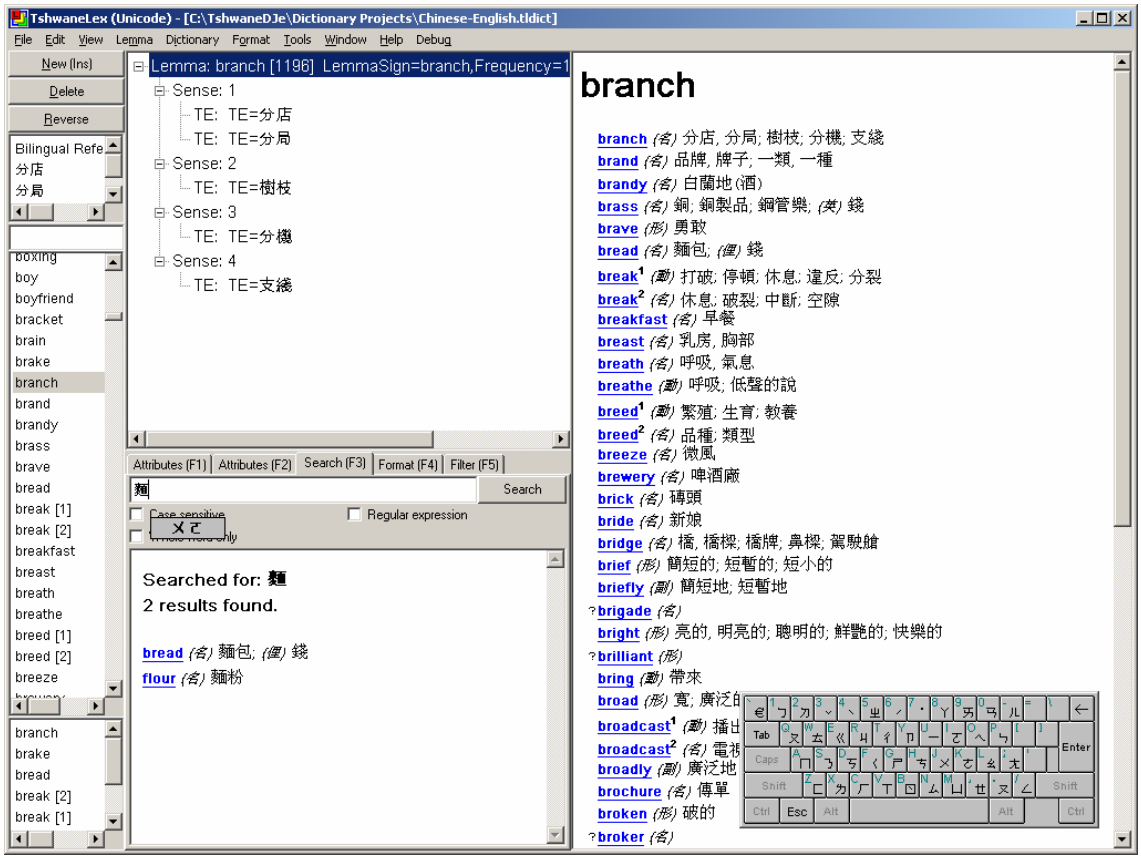

Figure 1. Unicode support for a bilingual English-Chinese dictionary in TshwaneLex

Generating multiple dictionaries from a single database

Elsewhere in this volume (cf. Joffe & De Schryver 2005) the main aspects of the customisable and multilayered DTD editor dialog are presented. Not only can the dictionary grammar for any project be flexibly configured and then kept under control with the built-in DTD, given that all elements and attributes are also linked to a comprehensive style system for generating the output (and preview), one single database can efficiently hold several dictionaries. Broadly speaking, this is achieved by doing two things: Firstly, by making use of multiple element 'categories' to which the various data attributes are assigned by the lexicographer depending on which dictionary or dictionaries they should appear in, and secondly by defining a different set of styles for each 'view' of the database, i.e. for each dictionary. Certain element categories are made visible or invisible in each style, which thus effectively functions as a kind of "mask" that filters and reveals only the portions of data to be shown for the current dictionary. Additionally, this also allows a different 'look' to be defined for each dictionary. These features are illustrated in Figures 3 and 4, which respectively show the desktop and pocket editions of a bidirectional French-Dutch dictionary (© 2005 Van Dale Lexicografie). One hotkey allows the lexicographer to switch between the two views, and thus also the two dictionaries. The extent of co-text and context for the production of any particular dictionary may thus easily be decided on at the output stage.

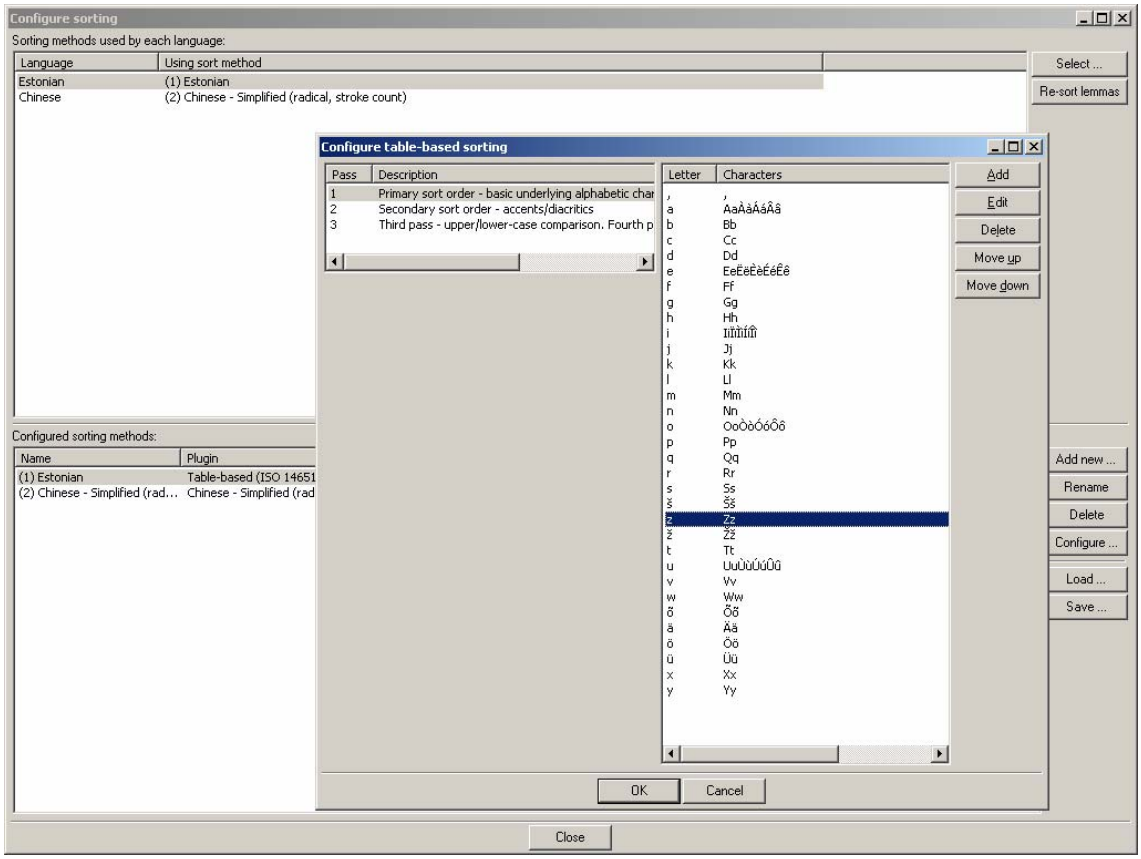

Figure 2. Configuring table-based sorting for Estonian in TshwaneLex

This feature may also be tied in with customising the language of the meta-language, as described earlier, potentially being used to customise aspects of the dictionary output further according to the language of the target user of the dictionary. For example, in a bidirectional Japanese-English dictionary, the information in some fields may inherently be primarily only useful to either a Japanese or English mother-tongue speaker. Lexicographers sometimes have to make editorial decisions and compromises based on assumptions about the language of the target market; by customising the output from a single database this need not be the case. In an electronic dictionary, one could take still other factors into account, such as the level of the user, presenting different views of the dictionary to beginner or advanced language learners.

Linked View mode for bilingual dictionary editing

Several innovative functions assist in bilingual dictionary compilation, such as side-by-side editing, automated reversal and Linked View mode. When in side-by-side editing mode, the screen is split in two down the middle, and the lexicographer can work on either side of a bilingual dictionary by simply moving between the windows. When in Linked View mode, as in the case of Figures 3 and 4, related articles in the reverse side of a bilingual dictionary are automatically displayed. For instance, from the left-hand side of Figure 3 one sees that the Dutch words ‘bagagedepot’, ‘statiegeld’, ‘lege fles’, ‘instructie’, ‘kwartierarrest’ and ‘(het) nablijven’ have been used as translation equivalents for the French word ‘consigne’.

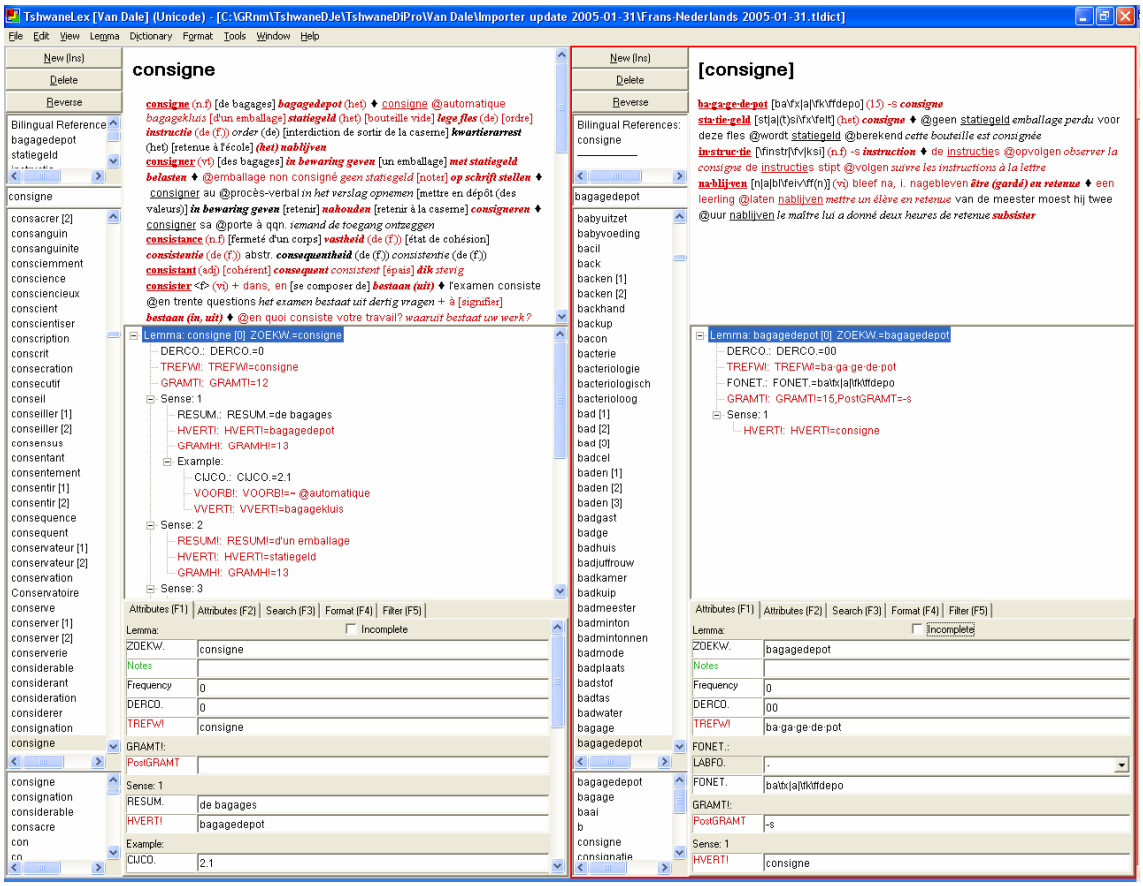

Figure 3. Desktop-edition view of a bidirectional French-Dutch dictionary in TshwaneLex, from the same database as the pocket edition

When in Linked View mode, TshwaneLex automatically shows all and only those articles that have these translation equivalents as lemma signs, in this case ‘bagagedepot’, ‘statiegeld’, ‘instructie’ and ‘nablijven’ as may be seen from the right-hand side of Figure 3.

The Linked View mode feature thus allows the lexicographer to attempt to honour the reversibility principle, that is, the condition whereby all lexical items presented as lemma signs or translation equivalents in the X-Y section of a dictionary are respectively translation equivalents and lemma signs in the Y-X section of the dictionary (cf. e.g. Tomaszczyk 1988: 290; Gouws 1989: 162; Gouws 1996: 80). The reversibility principle has always been a crucial but hitherto little-looked into requirement in lexicography, now easily made accessible in TshwaneLex.

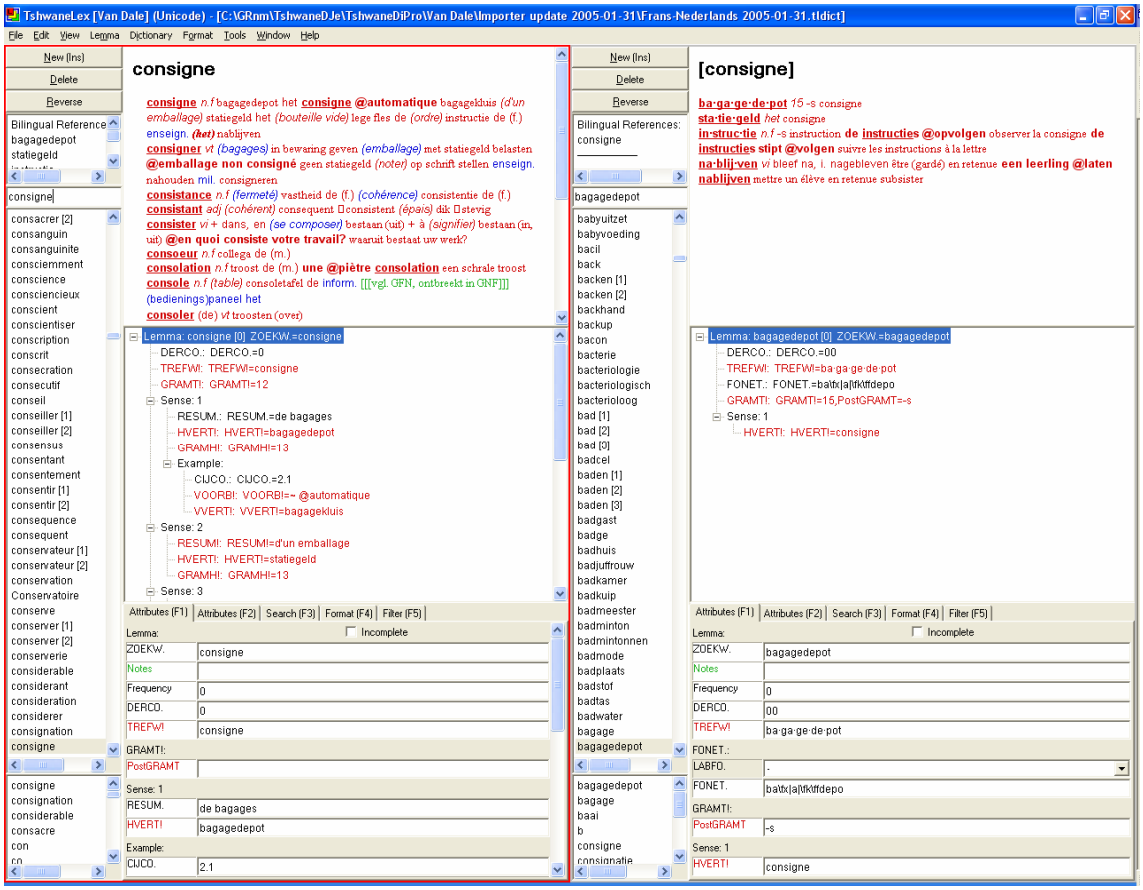

Figure 4. Pocket-edition view of a bidirectional French-Dutch dictionary in TshwaneLex, from the same database as the desktop edition

References

De Schryver, Gilles-Maurice and David Joffe (2004), 'On how electronic dictionaries are really used', in Geoffrey Williams and Sandra Vessier (eds), pp. 187–196.

Gouws, Rufus H. (1989), *Leksikografie* (Pretoria: Academica).

Gouws, Rufus H. (1996), 'Idioms and collocations in bilingual dictionaries and their Afrikaans translation equivalents', *Lexicographica: International Annual for Lexicography* 12: 54–88.

Joffe, David and Gilles-Maurice de Schryver (2004), 'TshwaneLex – A state-of-the-art dictionary compilation program', in Geoffrey Williams and Sandra Vessier (eds), pp. 99–104.

Joffe, David and Gilles-Maurice de Schryver (2005), 'Representing and describing words flexibly with the dictionary application TshwaneLex', in Vincent B.Y. Ooi *et al*. (eds), *Words in Asian Cultural Contexts, Proceedings of the 4th Asialex Conference, 1-3 June 2005, M Hotel, Singapore* (Singapore: Department of English Language and Literature & Asia Research Institute, National University of Singapore), pp. 108–114.

Tomaszczyk, Jerzy (1988), 'The bilingual dictionary under review', in Mary Snell-Hornby (ed.), *ZüriLEX'86 proceedings, papers read at the EURALEX international congress, University of Zürich, 9-14 September 1986* (Tübingen: A. Francke Verlag), pp. 289–297.

TshwaneDJe HLT (2002-2005), Online info, http://tshwanedje.com/ (accessed: 31 March 2005).

TshwaneLex (2002-2005). Online info, http://tshwanedje.com/tshwanelex/ (accessed: 31 March 2005).

Williams, Geoffrey and Sandra Vessier, editors (2004), *Proceedings of the eleventh EURALEX international congress, EURALEX 2004, Lorient, France, July 6-10, 2004* (Lorient: Faculté des Lettres et des Sciences Humaines, Université de Bretagne Sud).