The users and uses of TshwaneLex One

Gilles-Maurice de Schryver & David Joffe

TshwaneDJe Human Language Technology, Pretoria, South Africa
Department of African Languages and Cultures, Ghent University, Belgium
E-mail: {gillesmaurice.deschryver,david.joffe}@tshwanedje.com
Web: http://tshwanedje.com/

Abstract Ten months after the release of the dictionary compilation software TshwaneLex 1.0, and just days away from the launch of TshwaneLex 2.0, this paper presents a snapshot of the various users and uses of TshwaneLex to date.

1. Introduction

Development of the commercial, off-the-shelf dictionary compilation software TshwaneLex began in May 2002. This followed an in-depth study of the then-available packages for and approaches to dictionary compilation, as well as a survey of lexicographers' dreams with regard to 'the dictionary of the future' (De Schryver 2003). During the development of TshwaneLex, early adopters included numerous teams in especially Africa and Europe. Following the launch of TshwaneLex 1.0 in September 2005, the client base quickly grew to well over a hundred users. Since then, seven free upgrades within the version one range have been released. Now, in June 2006, just days away from the launch of TshwaneLex 2.0, it seems like an appropriate moment to take stock of the users and uses of 'TshwaneLex One'.

Two cautionary notes are in order. Firstly, the current field report will to some extent be self-censored, as commercial clients typically do not wish to divulge their plans until their products have reached the market. Secondly, as was the case with the early adopters, around 40% of the current users have already migrated to the next version, 'TshwaneLex Two', so some aspects of the latter will also be touched upon.

2. TshwaneLex in a nutshell

TshwaneLex is a dictionary writing system to compile *any* type of dictionary with, for *any* language(s). It is not a CQS (corpus-query system), DTP (desktop publishing) software, nor is it a 'generic XML editor'. Rather, the goal was to create a more specialised tool specifically to optimise and assist with dictionary compilation, and to be as 'user-friendly' as possible in that regard. Observe that this was the initial focus, and that CQS and DTP features as well as increased XML support have been steadily added as the client base has grown and clients have requested this or that feature. With the 'Web as Corpus' (Kilgarriff & Grefenstette 2003) in mind, direct links between TshwaneLex and Google text and image searches have for example been implemented, export options have multiplied with more possibilities for both online (e.g. one article per HTML page) and hardcopy (e.g. first/last lemma on each page in running header) options, while it is now also possible to import XML documents.

In TshwaneLex, a strict separation is made between the actual dictionary contents (the *data*), the structure of each article (the *dictionary grammar* or *DTD* (document type definition)), and the way those contents, given a certain structure, (may) look (the *formatting* or *style*). Each of those levels is fully customisable, with the data level further subdividable into unique and repetitive (*metalanguage*) material.

Among the many dictionary-compilation-specific features built into TshwaneLex are fully automated cross-reference integrity tracking and updating (Joffe *et al*. 2003), dynamic metalanguage customisation (De Schryver & Joffe 2005b), multidimensional lexicographic

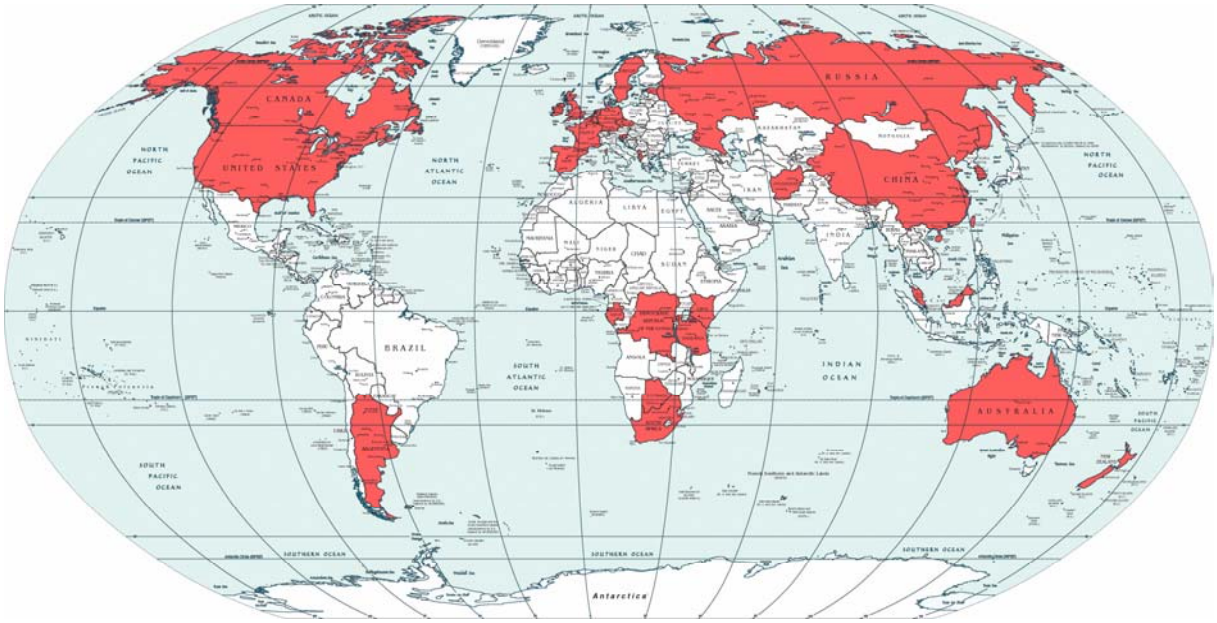
Rulers to help manage projects (De Schryver 2005), completely customisable sorting options (De Schryver & Joffe 2005a), etc., and specifically for bilingual and multilingual dictionary projects, powerful reversal and linked-view features (De Schryver & Joffe 2005a).

TshwaneLex One can be run from any stand-alone PC, and neither additional software nor knowledge of databases is required. (Note that porting to other platforms such as Mac and Linux is planned, while TshwaneLex Two contains network support.) Basically, TshwaneLex has its own 'internal format' for processing data in-memory which it always uses, and has a generic 'input/output layer' behind which backends/plugins exist (and more can be created) for loading/saving data from/to different underlying formats, including (1) the native TshwaneLex dictionary file (.tldict), (2) XML format, (3) a relational database, etc. (cf. the section 'extendible I/O architecture' in Joffe *et al*. (2003)). The 'internal format' can approximately be compared to a parsed XML document object. So internally TshwaneLex does not hold the data as XML, but rather, more like XML that has already been parsed into an in-memory structure. If saving to XML, the XML backend re-generates XML from the document object. If saving to a relational database, the relational database saves for instance rows to tables using SQL, and so on.

3. Basic user and usage statistics

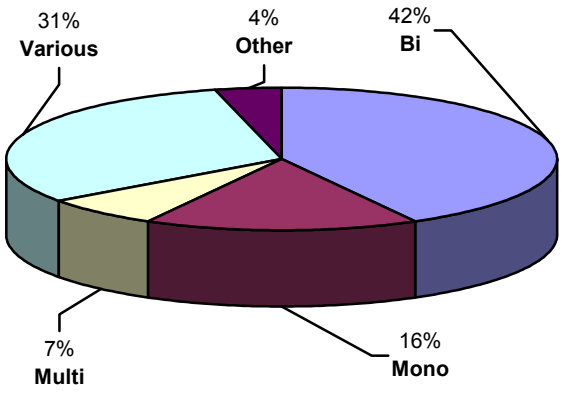
From the start, it has been the intention to cater for both commercial and academic projects, with in the latter case some level of philanthropy for languages listed in the *UNESCO Red Book of Endangered Languages*. With currently 195 users of TshwaneLex, the breakdown is as follows: 49% commercial, 47% academic, and 4% philanthropic. Around 29% of the users work on their own (in isolation), while 71% work on a project in group – in each case the software may be used to work on either a single or several projects simultaneously.

The family of TshwaneLex users presently spreads across the world, as is depicted on the following map:



TshwaneLex users are found in Afghanistan, Albania, Argentina, Australia, Belgium, Botswana, Canada, China, the Czech Republic, the Democratic Republic of the Congo, Estonia, France, Gabon, Germany, Ireland, Kenya, Luxembourg, Macao (China), Malaysia, the Netherlands, New Zealand, Russia, Rwanda, Slovenia, South Africa, South Korea, Spain, Sweden, Taiwan, Tanzania, the U.K., the U.S.A., and Wales (U.K.).

The number of different languages dealt with in TshwaneLex is even more diverse and currently approaches one hundred, among them: Afrikaans, Albanian, Alor Malay, Arabic, Acehnese, Bai, Balinese, Basque, Belarusian, Breton, Buginese, Bulgarian, Catalan, Chinese, Cilubà, Croatian, Czech, Danish, Dutch, East Javanese, English, Estonian, Finnish, French, Gaelic, German, Gimán, Haitian, Hmong, Iban, Icelandic, Indonesian, Inezeño Chumash, Irish, isiNdebele, isiXhosa, isiZulu, Italian, Jakarta Malay, Japanese, Javanese, Javindo, Kinyarwanda, Kiswahili, Korean, Kupang Malay, Ladino, Latin, Lingála, Low German, Macedonian, Madurese, Makassarian, Malay, Menadonesian, Minangkabau, Moluccan, Muna, Norwegian, Old English, Papiamento, Pashto, Petjoh, Picard, Polish, Polynesian, Portuguese, Romanian, Rotinese, Russian, Sahu, Sasak, Scots, Sesotho, Sesotho sa Leboa, Setswana, Singhalese, siSwati, Slovenian, Spanish, Sranantongo, Sundanese, Surinamese Javanese, Swedish, Terik, Ternate Malay, Tshivenda, Ukrainian, Virgin Islands Creole English, Walloon, Welsh, and Xitsonga.



Looking at the types of dictionaries that are being compiled with TshwaneLex, one notices that half the projects treat at least two languages (bilingual and semi-bilingual: 42%, trilingual and multilingual: 7%) versus only 16% that are truly monolingual. In every three out of ten projects (31%) a combination of types is being produced simultaneously, and in another 4% the focus is on the use of TshwaneLex to teach (meta)lexicography, to produce historical and dialect dictionaries or even pictionaries and encyclopaedias. Across the various types, roughly one fifth of the projects deal with LSP (language for specific purpose) dictionaries.

The extent/size of projects for which TshwaneLex is currently being used varies widely, from very small lexica to huge multi-volume reference works. To give an idea of a project between these extremes, the latest 1,552-page A4-size Afrikaans–English desktop dictionary by *Pharos* (Du Plessis *et al*. 2005) can easily be handled as a single TshwaneLex file on a single PC, with some statistics as follows:

- over 77,000 main entries;
- over 200,000 when including all sub-entries;
- over 2,400,000 elements (nodes) in the document tree;
- which corresponds to an 86MB TshwaneLex file;
- or exported as Unicode text, about 36MB;
- which translates to approximately 18 million characters.

4. Sorts of issues arising in different sorts of circumstances

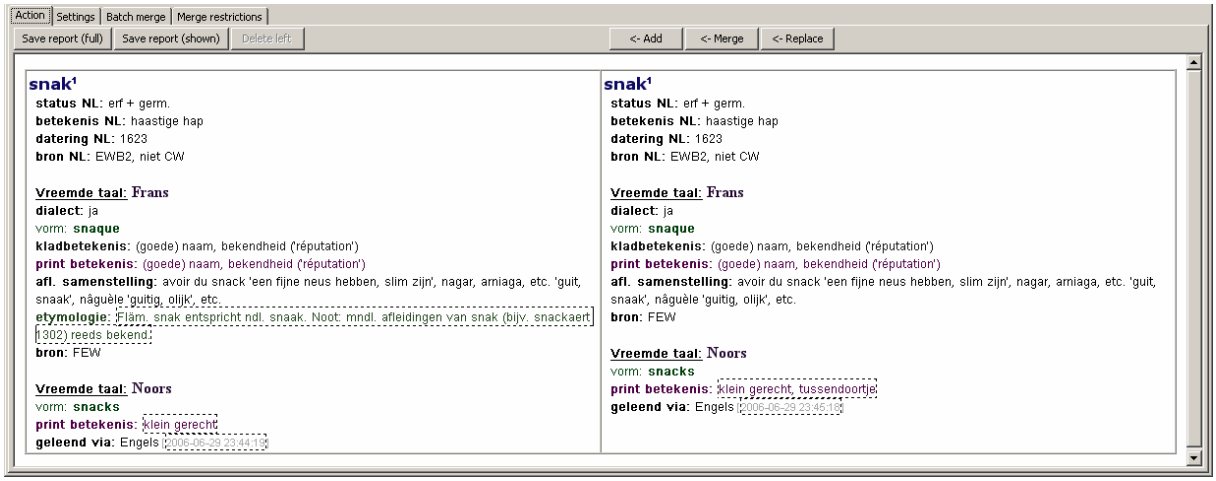
Clearly, with this wide geographical and typological coverage of users and uses, TshwaneLex simply had to support Unicode (as well as left-to-right and right-to-left scripts) on all levels. A flexible DTD with linked styles system that anyone *without programming* skills could set up also had to be, and *is*, part of the standard TshwaneLex package (Joffe & De Schryver 2005).

Perhaps surprisingly, the current needs did not include network support nor over-complex workflow modules. Conversely, all large teams (with on average around ten, but in one case up to thirty users) wished to have **advanced compare/merge tools** at their disposal. A special

effort was therefore put into the development of these. When teams work in a distributed approach on a single project, three typical cases present themselves:

- different ‘chunks’ (e.g. different alphabetic sections, words belonging to different word classes, or even different semantic fields) are being worked on, and are simply merged periodically into a main database, after which that main database is redistributed to all compilers;
- each compiler focuses on certain aspects of each article only (phonetics, definitions, examples, etc.), with the same TshwaneLex file being sent from one compiler to the next. This approach is sometimes combined with the previous one;
- in especially multilingual setups, where up to a dozen languages are being worked on in parallel, each compiler focuses on his/her respective language(s), with their data then being merged periodically, and a new version of the main database being redistributed to all project members.

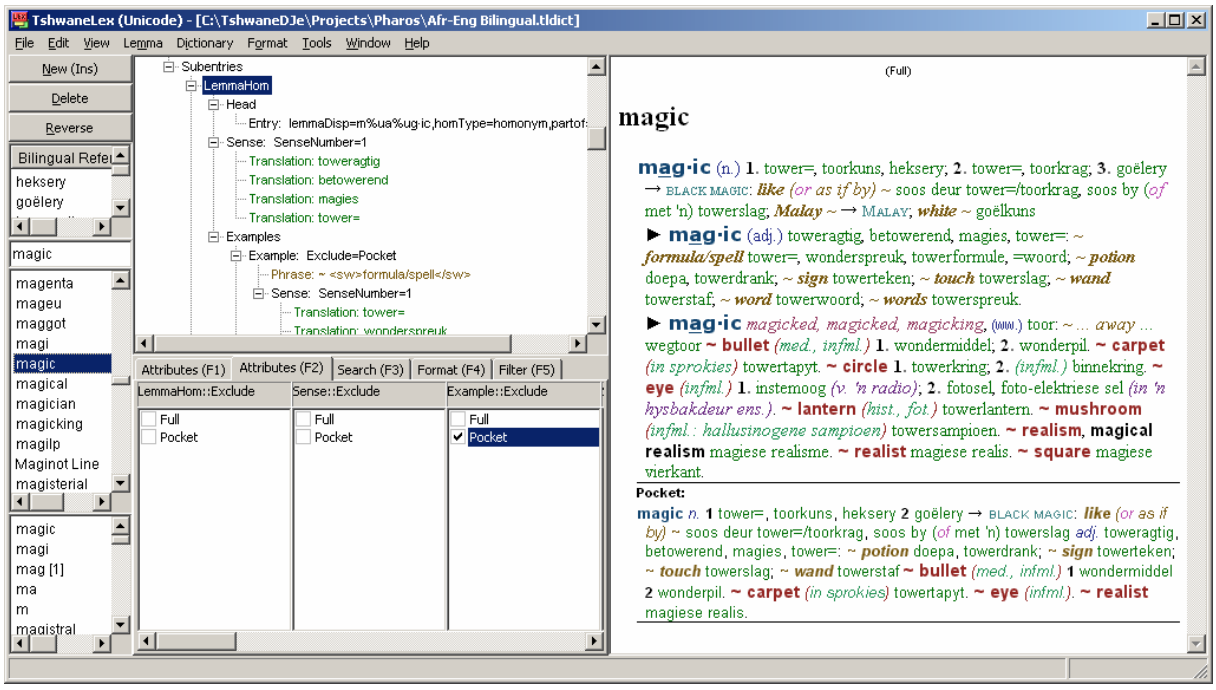
The latter approach is illustrated in the screenshot below, using data that is being prepared by a team of over a dozen compilers under the guidance of Nicoline van der Sijs, for her forthcoming book *Nederlands in de wereld*.



In this project, an inventory is made of all the Dutch words that have entered other languages over the centuries. In TshwaneLex, each Dutch word has a treatment of its own (in the screenshot *snak* [homonym 1] ‘snack’ and the block immediately underneath it), and then linked to that any number of *Vreemde taal* ‘foreign language’ blocks (here, and so far, for French and Norwegian). In the compare/merge illustrated here, the data from the compiler focusing on the Nordic languages, which includes Norwegian, is being merged into the main database. In this case, the synonym *tussendoortje* will be added to *klein gerecht* which was already in the database. Observe that, in an earlier compare/merge pass, the updated/new material from the compiler focusing on French had already been added, and those changes will of course not be overwritten. Needless to say, after having clicked all the necessary merge restrictions, combining databases is a fully automatic and seamless process.

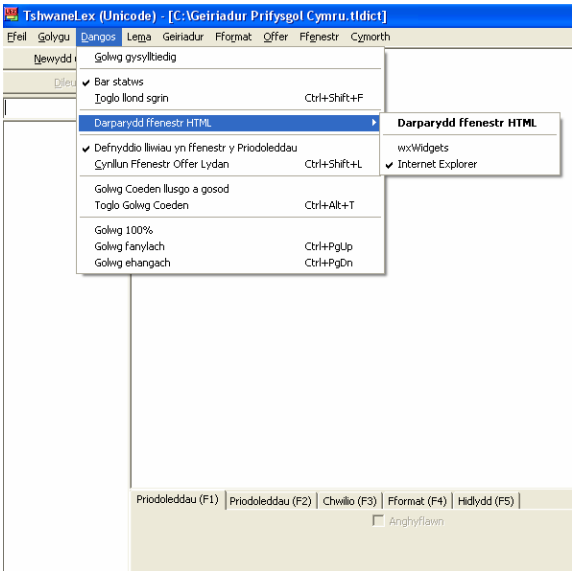
A second aspect that is becoming increasingly important is the notion to be able to ‘extract’ a multitude of dictionaries, each with their own characteristics, from a single, large database. Hence, with a single click, one typically wants to extract a pocket versus a desktop dictionary, or following another click a print edition versus an online version, or even a semi-bilingual versus a monolingual dictionary, all from the same TshwaneLex file, and in each case with the metalanguage in the appropriate format/language. In TshwaneLex this is achieved by means of allowing for **multiple sets of styles** to be set up, and in version two, a sophisticated

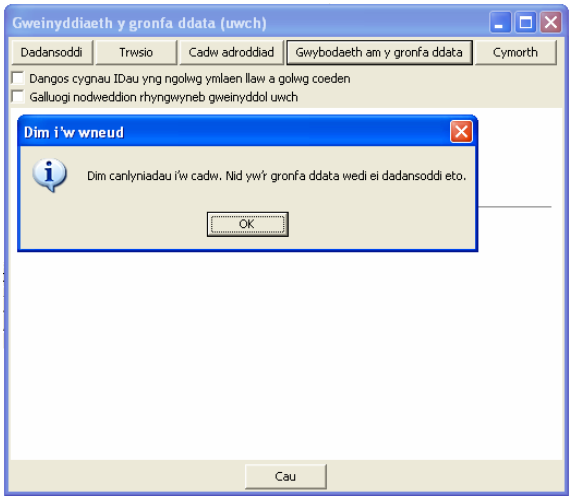
new **‘masks’ feature**. These aspects are exemplified below for the above-mentioned *Pharos* dictionary, where several different dictionary projects are currently being integrated into one unique TshwaneLex file. (Note that in order not to divulge the publisher’s plans, only two styles are shown here, ‘Full’ and ‘Pocket’.)



In this screenshot, the option was chosen to display the various styles simultaneously in the preview area (the right half of the screen), so the various views (here ‘Full’ and ‘Pocket’) of every article can be seen concurrently. Note for example the different styles for lemma signs and parts of speech in the different editions, but also the automatic (re)numbering when outputting selected levels of the data.

Thirdly, and hardly surprising given that TshwaneLex is being used in all corners of the world, the wish was quickly voiced to enable the easy localisability of the GUI (graphical user interface). See in this regard the screenshots below for the (in-progress) translation into Welsh.





The localisation of the TshwaneLex GUI is put at every user's fingertips with the self-explanatory **built-in localisation editor**. Particularly handy is the fact that the results of the localisation can be seen in real time within TshwaneLex itself. At present, several localised versions of TshwaneLex are already in use in Asia, Europe and Africa, in among others (and respectively) Chinese, German and Cilubà.

5. Conclusion and outlook

Looking back, and keeping in mind that several dictionary writing systems did not quite make it in the past, the creation and distribution of TshwaneLex has become a true success story. Any licenses acquired henceforth will automatically be version two licenses. This second version of course includes everything the first version has, but also contains some significant improvements and a battery of new features. The already-mentioned ability to import XML, better network/multi-user support, and a more versatile approach to the concept of 'one database, many dictionaries', are some of them. Interlinked search features and filters, and numerous user interface improvements that help speed up compilation work, such as click-in-preview-to-edit, highlight selected element, or an optional pop-up window for work on long definitions, are but some of the others. No doubt, the family of TshwaneLex users will continue to grow, and with new users come new uses, and thus exciting new features.

References

De Schryver, G-M. 2003. Lexicographers' Dreams in the Electronic-Dictionary Age. *International Journal of Lexicography* 16.2: 143–199.

De Schryver, G-M. 2005. Concurrent Over- and Under-Treatment in Dictionaries – The *Woordeboek van die Afrikaanse Taal* as a Case in Point. *International Journal of Lexicography* 18.1: 47–75.

De Schryver, G-M & D. Joffe. 2005a. One database, many dictionaries – varying co(n)text with the dictionary application TshwaneLex. In Ooi, V.B.Y., A. Pakir, I. Talib, L. Tan, P.K.W. Tan & Y.Y. Tan (eds.). 2005. *Words in Asian Cultural Contexts, Proceedings of the 4th Asialex Conference, 1-3 June 2005, M Hotel, Singapore*: 54–59. Singapore: Department of English Language and Literature & Asia Research Institute, National University of Singapore.

De Schryver, G-M & D. Joffe. 2005b. Dynamic Metalanguage Customisation with the Dictionary Application TshwaneLex. In Kiefer, F., G. Kiss & J. Pajzs (eds.). 2005. *Papers in Computational Lexicography, COMPLEX 2005*: 190–199. Budapest: Linguistics Institute, Hungarian Academy of Sciences.

Du Plessis, M. *et al.* 2005. *Pharos Afrikaans–Engels / English–Afrikaans Woordeboek / Dictionary*. Cape Town: Pharos.

Joffe, D. & G-M de Schryver. 2005. Representing and describing words flexibly with the dictionary application TshwaneLex. In Ooi, V.B.Y., A. Pakir, I. Talib, L. Tan, P.K.W. Tan & Y.Y. Tan (eds.). 2005. *Words in Asian Cultural Contexts, Proceedings of the 4th Asialex Conference, 1-3 June 2005, M Hotel, Singapore*: 108–114. Singapore: Department of English Language and Literature & Asia Research Institute, National University of Singapore.

Joffe, D., G-M de Schryver & D.J. Prinsloo. 2003. Computational features of the dictionary application "TshwaneLex". *Southern African Linguistics and Applied Language Studies* 21.4 (Special issue on 'Human Language Technology in South Africa: Resources and Applications'): 239–250.

Kilgarriff, A. & G. Grefenstette. 2003. Special Issue on the Web as Corpus. *Computational Linguistics* 29.3: 333–502.

TshwaneLex. 2002–2006. Available from `http://tshwanedje.com/tshwanelex/`