

Why Opting for a Dedicated, Professional, Off-the-shelf Dictionary Writing System Matters

Gilles-Maurice de Schryver

Department of African Languages and Cultures, Ghent University (Belgium)

Xhosa Department, University of the Western Cape (South Africa)

TshwaneDJe HLT

Abstract

In this workshop *TLex* (aka *TshwaneLex*) is analysed. *TLex* is a professional, feature-rich, fully internationalised, off-the-shelf software application suite for compiling dictionaries or terminology lists. It has been adopted by many major publishers, government organisations and individuals worldwide.

TLex contains numerous specialized features that allow one to dramatically reduce dictionary production time and costs, and to increase the quality and consistency of one's dictionaries (from single-user projects to large teams). These features include an integrated corpus query system, real-time preview, full customisability, an advanced styles system, smart cross-references with tracking and auto-updating, automated lemma reversal, automated numbering and sorting, export to MS Word and typesetting systems (such as InDesign, Quark and XPP), multi-user support for managing teams, etc. The data can be published to hardcopy, the Web, or CD-ROM / DVD / software.

TLex can be used for all languages, for all types of dictionaries, and supports industry standards such as XML and Unicode.

1. Software for Lexicography: A History of Pointless Reinventions

This workshop deals with dictionary writing systems (DWSs), also known as lexicographic workbenches (Ridings 2003), dictionary compilation programs (Joffe & De Schryver 2004), dictionary (content) management systems (Alegria et al. 2006, Langemets et al. 2010), or dictionary editing systems (Svensén 2009) – amongst others. In one of the first scientific descriptions of a DWS, the co-designer claimed:

“It is extremely unlikely that a project will find a commercial product that will meet its needs off the shelf.” (Ridings 2003: 214)

Indeed, Ridings's own *Onoma*, as well as earlier DWSs such as *Compulexis* or *GestorLEX*, were applications that had to be heavily redesigned for each dictionary project anew. To this day, this trend has persisted, with virtually every dictionary publisher reinventing the wheel, and either developing their own in-house system or acquiring a more generic platform (e.g. a generic XML editor) which is then turned into a DWS. Examples include *Pasadena*, developed for the *Oxford English Dictionary* (Thompson 2005), *DicSy* for Norstedts, Sweden's leading publisher of dictionaries (Svensén 2009: 423), or the *ANW article editor* for the *Instituut voor Nederlandse Lexicology* (INL) in the Netherlands (Niestadt 2009).

Every dictionary project at every academic institution also seems to be developing its own variant. Examples include the *Dictionary CMS* at the University of the Basque Country (Alegria et al. 2006), *Jibiki* at the Savoie University (Mangeot 2006), *DEB* at Masaryk University (Pala & Horák 2006), and so on. The main argument in this workshop is that this is an unfortunate state of affairs, as excellent off-the-shelf DWSs not only exist, but using those as a starting point would benefit both dictionary compilers and developers of DWSs – benefits which would speed up the creation of ever-better DWSs, with which ever-better dictionaries could then be compiled.

2. Generic Dictionary Writing Systems

The main characteristics and general requirements of DWSs have recently been summarized in Atkins & Rundell (2008: 113-117) and Svensén (2009: 422-425). More detailed accounts, including the presentation of actual DWSs, can be found in the proceedings of the last two DWS Workshops: DWS 2004 (cf. Smrž et al. 2004) and DWS 2006 (cf. De Schryver 2006a).

Four professional DWSs can currently be considered generic enough to function as a starting point for any new dictionary project: *TLex*, also known as *TshwaneLex* (Joffe et al. 2003), the *IDM DPS* (McNamara 2003), *iLex* (Erlandsen 2004), and *ABBYY Lingvo Content* (Kuzmina & Rylova 2010). In what follows we will analyse one of these, *TLex*, and compare this off-the-shelf DWS with a DWS developed in-house from scratch, namely the *ANW Article Editor* used at the INL.

3. A Brief Comparison of a Reinvention with a Generic Tool

According to the developer of the *ANW Article Editor*, there were four main reasons why a customised DWS simply (quote) “had” (unquote) to be developed at the INL:

- (1) the need to ‘inherit’ information from the ‘head’ of a dictionary article, and to see that information automatically ‘reflected’ at the various senses;
- (2) the need for a clear overview of the complex ANW article structure;
- (3) the need for a maximally user-friendly product; and
- (4) the need to build in ANW-specific functionality.
(Niestadt 2009: 216)

Compared to the off-the-shelf *TLex*, however, none of these so-called obstacles really looks like an obstacle:

- (1) can simply be achieved with the use of *TLex*’s built-in scripting language (Lua);
- (2) is fully in the hands of the person designing the DTD (document type definition DTD; i.e. Tree Views can be made as straightforward or as complex as one wishes);
- (3) *TLex* is known for its insistence on user-friendliness; and
- (4) *TLex* is maximally extendable.

Although the equivalent of seven person-months was spent on developing the *ANW Article Editor*, the developer admits that there are four areas where the newly built tool is lacking:

- (1’) there is no automated cross-reference tracking;
- (2’) there is no detailed search functionality;
- (3’) there are no options to overrule the formatting within a field; and
- (4’) most DTD changes are not yet automatically transferred to the interface.
(Niestadt 2009: 220-221)

Problems (1’) to (4’) are exactly examples of areas where a professional DWS such as *TLex* has already put its genius. Literally years of person-months were devoted in getting aspects such as (1’) to (4’) working correctly, so the gain of redesigning a tool from scratch is questionable.

4. *TLex* (aka *TshwaneLex*)

What, then, are the main advantages of a dedicated, professional, off-the-shelf DWS such as *TLex*? This is the main topic of the workshop. Given the workshop is conceived as a hands-on session, we will limit ourselves to an enumeration in bullet form here. (For the source of the text below, see “*TLex*” in the References.)

4.1. Who is *TLex* Intended For?

- Dictionary publishing houses
- Individual dictionary compilers
- Dictionary development teams
- Terminology managers and practitioners
- Government and other organisations compiling dictionaries or terminology lists
- Organisations with a need to produce, manage and distribute terminology internally (e.g. publish on an intranet)
- For the production of:
 - Monolingual, bilingual or multilingual dictionaries (paper, electronic, online/intranet)
 - Multilingual terminology lists (cf. Joffe & De Schryver 2005a)
 - Other explanatory dictionaries (e.g. economic terms, mining terms, etc.)
 - Large historical dictionaries
 - Any other kind of reference works (e.g. encyclopedia, thesaurus, etc.; cf. De Schryver & Joffe 2005b)

4.2. Benefits of *TLex*

Using dedicated dictionary compilation software rather than general-purpose tools such as word processors or generic XML tools provides significant benefits in terms of both dictionary development time and output quality for both individual lexicographers as well as lexicographic (or terminology compilation) teams:

- Reduced project completion time, thanks to (amongst others) various levels of automation, such as automatic numbering, lemma reversal, cross-reference tracking/updating, and error checking
- Increased consistency in the treatment of articles, thanks to features such as the article filter
- Leads to more consistent and balanced treatment of both languages in a bilingual dictionary
- Improved teamwork and team communication
- More easily scalable to larger team sizes

4.3. Primary Features of *TLex*

- Fast
- User-friendly: *TLex* does not require advanced computer literacy skills – if one can use a word processor, one will be able to learn *TLex*
- Automatic sense numbering
- Automatic homonym numbering
- Automatic cross-reference tracking and updating of homonym and sense numbers
- Immediate WYSIWYG (what you see is what you get) article preview
- Immediate preview of cross-referenced articles and cross-referencing articles
- Integrated corpus (cf. De Schryver & De Pauw 2007)

- Full Unicode support – supports virtually all of the world's languages
- Easily enter any phonetic symbol (IPA; phonetic extensions)
- Fully customisable and highly flexible (create any fields and structures relevant to one's dictionary)
- Network and multi-user (team) support – supports all major database servers (e.g. MS SQL Server, Oracle, PostgreSQL)
- Management tools: Assign tasks to users and monitor user or team progress
- Export to:
 - Microsoft Word format, RTF, HTML, XML, CSV
 - Corel WordPerfect and OpenOffice (via RTF format)
 - Adobe InDesign and QuarkXPress
- Import from:
 - Wordlists
 - CSV (may also import corpus frequency counts)
 - XML or word frequency counts from corpus query software
 - Custom, i.e. conversion of existing data
- Various features for generating 'multiple dictionaries from one database' (cf. De Schryver & Joffe 2005c)
- Customisable styles (font, colour, etc.) for every field in the dictionary
- Customisable language of the metalanguage (cf. De Schryver & Joffe 2005a)
- Bilingual dictionaries: Automated lemma reversal
- Bilingual dictionaries: Side-by-side bilingual editing and linked-view mode
- Bilingual dictionaries: translation-equivalent fanouts
- Multimedia: Allows sound (e.g. pronunciation) recordings to be linked to any field
- Multimedia: Allows images to be added to entries
- IME Windows soft-keyboard support
- Right-to-left language support (Hebrew, Arabic, etc.)
- Fast full-dictionary text search
- Filter: define criteria for viewing/exporting a subset of the dictionary based on specific characteristics
- Dictionary compare/merge feature: Integrate work done by different team members (cf. De Schryver & Joffe 2006)
- A unique Ruler Tool to ensure a balanced treatment on multiple levels (cf. De Schryver 2005)
- Automatic checking for various dictionary errors
- Electronic dictionary (CD-ROM / DVD) software module available (cf. Joffe, MacLeod & De Schryver 2008)
- Place dictionaries online, using:
 - Online dictionary module (cf. De Schryver & Joffe 2004)
 - Direct export to static HTML (cf. De Schryver & Joffe 2005b)
- Scripting language (Lua)
- Customisable DTD (cf. Joffe & De Schryver 2005b)
- Fully localisable (cf. De Schryver 2006b)

4.4. System Requirements for *TLex*

- Intel Mac
- OR
- Windows PC (Windows 2000 / XP / Vista / 7)

4.5. Screenshots of TLex

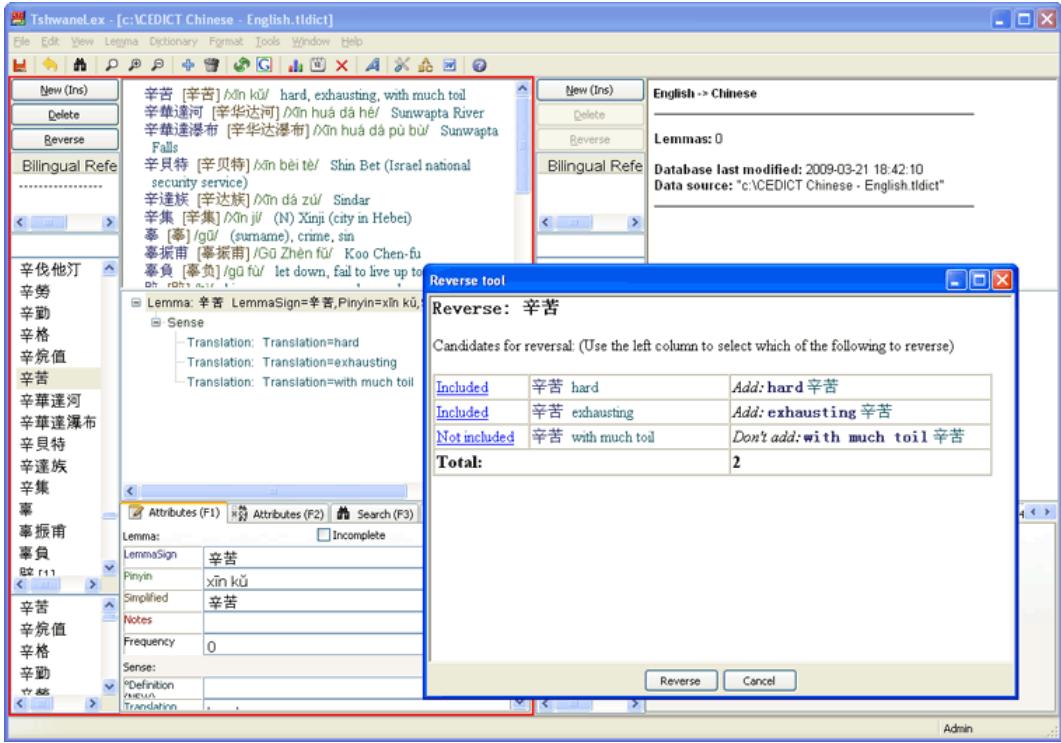


Figure 1: TLex screenshot showing the lemma reversal tool. When auto-reversing lemmas, individual word senses and combinations can be easily selected or deselected for reversal.

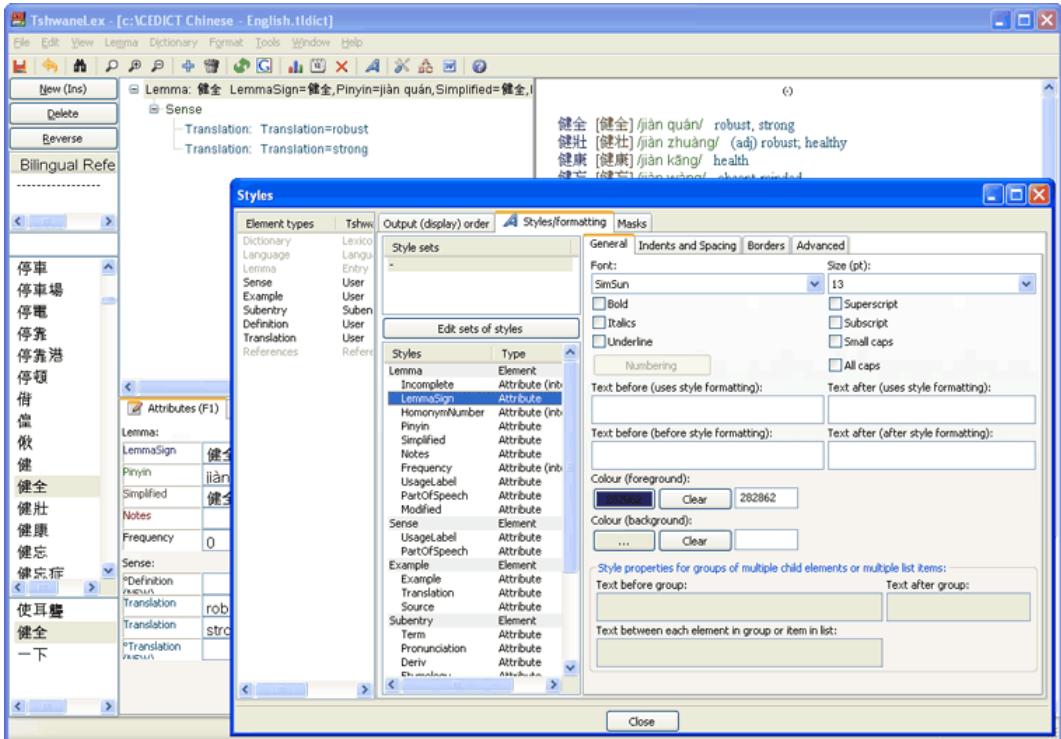


Figure 2: Customising the styles of different fields in TLex.

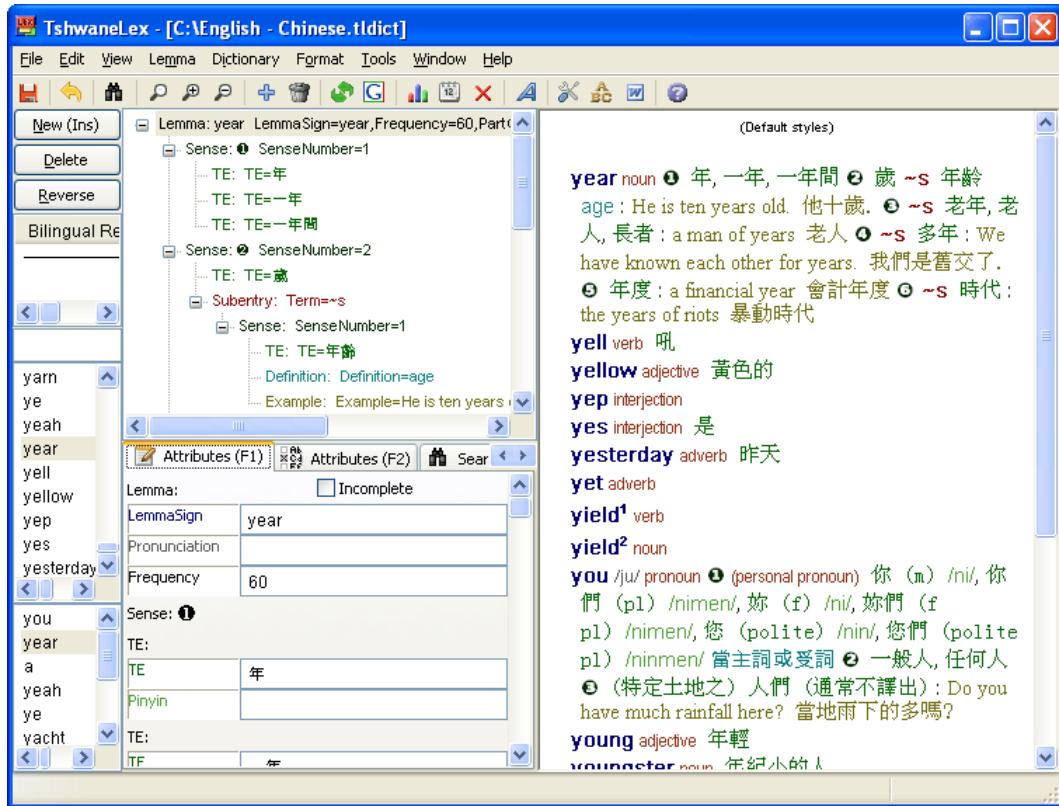


Figure 3: Automatic sense numbering in *TLex*. The numbering style can easily be changed at any time.

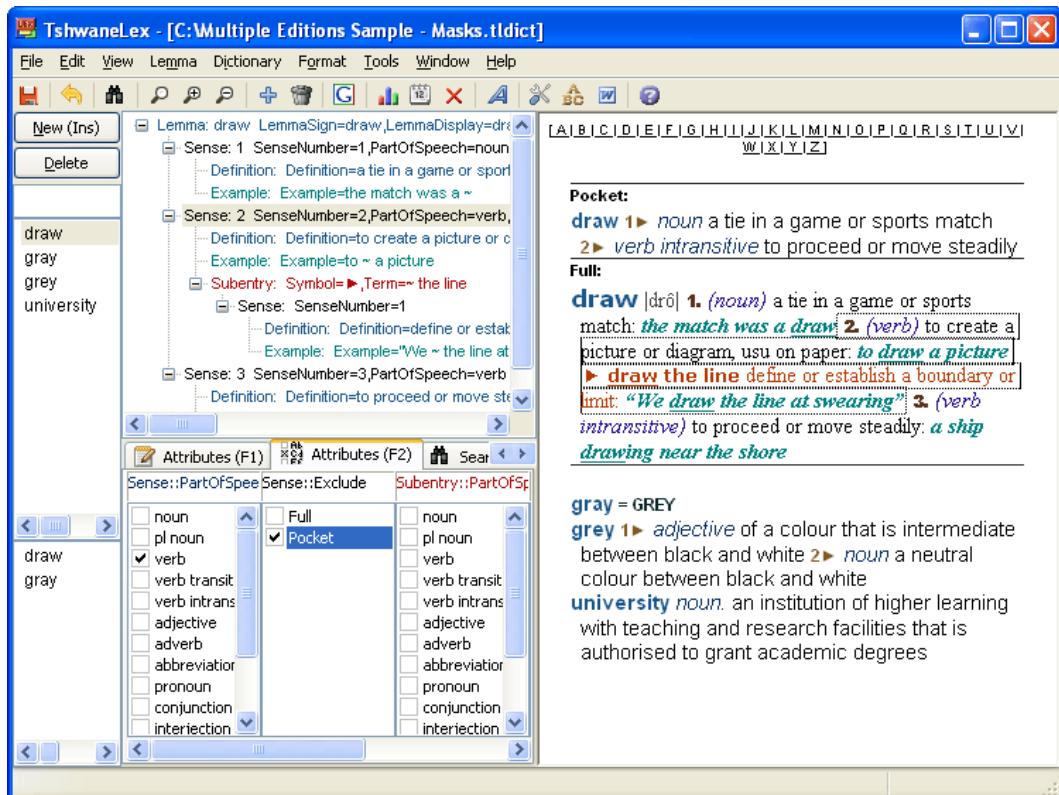


Figure 4: Screenshot demonstrating the ‘multiple dictionaries from one database’ features in *TLex*.

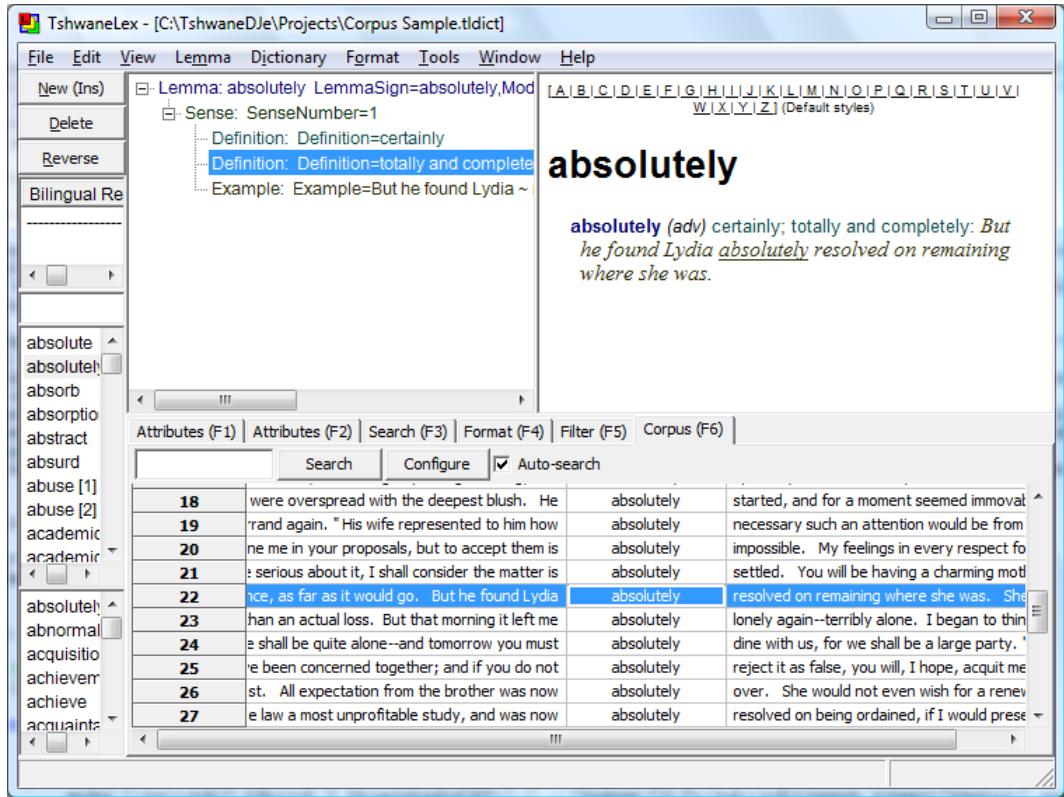


Figure 5: *TLex* integrated corpus query tool, with auto-extracted usage example.

	ID	Action	Date	User Agent			
1.	f18f20371985ffb4d715e574684c022c[1]	inverse	129.142.147.2	www.ltk.dk	jump	2003-11-13 16:08:10	Mozilla/4.0 (compatible; MSIE 6.0; Windows 98; Win 9x 4.90)
2.	f18f20371985ffb4d715e574684c022c[2]	inverse	129.142.147.2	www.ltk.dk	read	2003-11-13 16:10:24	Mozilla/4.0 (compatible; MSIE 6.0; Windows 98; Win 9x 4.90)
3.	f18f20371985ffb4d715e574684c022c[3]	inverse	129.142.147.2	www.ltk.dk	read past tense	2003-11-13 16:10:44	Mozilla/4.0 (compatible; MSIE 6.0; Windows 98; Win 9x 4.90)
4.	f18f20371985ffb4d715e574684c022c[4]	inverse	129.142.147.2	www.ltk.dk	have read	2003-11-13 16:11:02	Mozilla/4.0 (compatible; MSIE 6.0; Windows 98; Win 9x 4.90)
5.	f18f20371985ffb4d715e574684c022c[5]		129.142.147.2	www.ltk.dk	tobetsa	2003-11-13 16:12:03	Mozilla/4.0 (compatible; MSIE 6.0; Windows 98; Win 9x 4.90)

Figure 6: The *TLex* online dictionary software module provides the option to keep extensive, detailed (anonymous) usage logs. An intricate and multifaceted log file tracks every single action of every single user – date and time stamping every lookup, ordering founds and not-founds, monitoring long-term vocabulary retention, etc. – with a multitude of customisable summaries being presented to the lexicographers. In this screenshot the general logging of dictionary lookups is shown, with a date/time range set.

References

- ABBYY Lingvo Content [online] http://www.abbyy.com/lingvo_content/ [Access date: 1 June 2011].
- Alegria, I. et al. (2006). “A Dictionary Content Management System”. In Corino, E., Marello, C. and Onesti, C. (eds.). *Atti del XII Congresso Internazionale di Lessicografia, Torino, 6-9 settembre 2006 / Proceedings XII Euralex International Congress, Torino, Italia, September 6th-9th, 2006*. Alessandria: Edizioni dell’Orso. Vol. I. 105-109.
- Atkins, B. T. S. and Rundell, M. (2008). *The Oxford Guide to Practical Lexicography*. New York: Oxford University Press.
- 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. (ed.). (2006a). *DWS 2006: Proceedings of the Fourth International Workshop on Dictionary Writing Systems, Tuesday 5th September 2006, Turin, Italy (Pre-EURALEX 2006)*. Pretoria: (SF)² Press.
- De Schryver, G.-M. (2006b). “Internationalisation, Localisation and Customisation Aspects of the Dictionary Application *TshwaneLex*”. *Lexikos* 16. 222-238.
- De Schryver, G.-M. and De Pauw, G. (2007). “Dictionary Writing System (DWS) + Corpus Query Package (CQP): The Case of *TshwaneLex*”. *Lexikos* 17. 226-246.
- De Schryver, G.-M. and Joffe, D. (2004). “On How Electronic Dictionaries are Really Used”. In Williams, G. and Vessier, S. (eds.). *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. Vol. I. 187-196.
- De Schryver, G.-M. and Joffe, D. (2005a). “Dynamic Metalanguage Customisation with the Dictionary Application *TshwaneLex*”. In Kiefer, F., Kiss, G. and Pajzs, J. (eds.). *Papers in Computational Lexicography, COMPLEX 2005*. Budapest: Linguistics Institute, Hungarian Academy of Sciences. 190-199.
- De Schryver, G.-M. and Joffe, D. (2005b). “From *TshwaneLex* to *TshwanePedia*: Creating and Flexibly Maintaining Online Encyclopaedias”. *Lexikos* 15. 306-311.
- De Schryver, G.-M. and Joffe, D. (2005c). “One database, many dictionaries – varying co(n)text with the dictionary application *TshwaneLex*”. In Ooi, V. B. Y. 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 and Asia Research Institute, National University of Singapore. 54-59.
- De Schryver, G.-M. and Joffe, D. (2006). “The users and uses of *TshwaneLex One*”. In De Schryver, G.-M. (ed.). *DWS 2006: Proceedings of the Fourth International Workshop on Dictionary Writing Systems, Tuesday 5th September 2006, Turin, Italy (Pre-EURALEX 2006)*. Pretoria: (SF)² Press. 41-46.
- Erlanson, J. (2004). “iLex – new DWS”. In Smrž, P. et al. (eds.). *Presentations at the Third International Workshop on Dictionary Writing Systems (DWS 2004)* [online]. <http://nlp.fi.muni.cz/dws2004/pres/> [Access date: 1 June 2011].
- IDM DPS [online]. http://www.idm.fr/products/dictionary_writing_system_dps/27/ [Access date: 1 June 2011].
- iLex [online]. <http://www.emp.dk/> [Access date: 1 June 2011].
- Joffe, D. and De Schryver, G.-M. (2004). “*TshwaneLex* – A State-of-the-Art Dictionary Compilation Program”. In Williams, G. and Vessier, S. (eds.). *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. Vol. I. 99-104.
- Joffe, D. and De Schryver, G.-M. (2005a). "From *TshwaneLex* to *TshwaneTerm*: Tailoring Terminology Management for South Africa". *Lexikos* 15. 312-315.
- Joffe, D. and De Schryver, G.-M. (2005b). "Representing and describing words flexibly with the dictionary application *TshwaneLex*". In Ooi, V. B. Y. 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 and Asia Research Institute, National University of Singapore. 108-114.
- Joffe, D., De Schryver, G.-M. and Prinsloo, D. J. (2003). "Computational Features of the Dictionary Application *TshwaneLex*". *Southern African Linguistics and Applied Language Studies* 21 (4). 239-250.
- Joffe, D., MacLeod, M. and De Schryver, G.-M. (2008). "Software Demonstration: The *TshwaneLex* Electronic Dictionary System". In Bernal, E. and DeCesaris, J. (eds.). *Proceedings of the XIII EURALEX International Congress (Barcelona, 15-19 July 2008)* (Serie Activitats 20) [CD-ROM]. Barcelona: Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra. 421-424.
- Kuzmina, V. and Rylova, A. (2010). "ABBYY Lingvo Electronic Dictionary Platform and Lingvo Content Dictionary Writing System". In Granger, S. and Paquot, M. (eds.). *eLexicography in the 21st Century: New Challenges, New Applications. Proceedings of eLEX 2009, Louvain-la-Neuve, 22-24 October 2009* (Cahiers du Cental 7). Louvain-la-Neuve: Presses universitaires de Louvain. 419-423.
- Langemets, M., Loopmann, A. and Viks, Ü. (2010). "Dictionary Management System for Bilingual Dictionaries". In Granger, S. and Paquot, M. (eds.). *eLexicography in the 21st Century: New Challenges, New Applications. Proceedings of eLEX 2009, Louvain-la-Neuve, 22-24 October 2009* (Cahiers du Cental 7). Louvain-la-Neuve: Presses universitaires de Louvain. 425-429.
- Mangeot, M. (2006). "Dictionary Building with the Jibiki Platform". In Corino, E., Marello, C. and Onesti, C. (eds.). *Atti del XII Congresso Internazionale di Lessicografia, Torino, 6-9 settembre 2006 / Proceedings XII Euralex International Congress, Torino, Italia, September 6th-9th, 2006*. Alessandria: Edizioni dell'Orso. Vol. I. 185-188.
- McNamara, M. (2003). "Dictionaries for All: XML to Final Product". In *Proceedings of the XML Europe 2003 Conference & Exposition. Powering the Information Society*. London.
- Niestadt, J. (2009). "De ANW-artikeleditor: software als strategie" [The ANW article editor: software as a strategy]. In Beijk, E. et al. (eds.). *Fons Verborum. Feestbundel voor prof. dr. A. M. F. J. (Fons) Moerdijk, aangeboden door vrienden en collega's bij zijn afscheid van het Instituut voor Nederlandse Lexicologie*. Amsterdam: Gopher BV. 215-222.
- Pala, K. and Horák, A. (2006). "From Web Pages to Dictionary: A Language-Independent Dictionary Writing System". In Corino, E., Marello, C. and Onesti, C. (eds.). *Atti del XII Congresso Internazionale di Lessicografia, Torino, 6-9 settembre 2006 / Proceedings XII Euralex International Congress, Torino, Italia, September 6th-9th, 2006*. Alessandria: Edizioni dell'Orso. Vol. I. 199-204.
- Ridings, D. (2003). "Lexicographic Workbench: A Case History". In Van Sterkenburg, P. (ed.). *A Practical Guide to Lexicography* (Terminology and Lexicography Research and Practice 6). Amsterdam: John Benjamins. 204-214.
- Smrž, P. et al. (eds.). (2004). *Presentations at the Third International Workshop on Dictionary Writing Systems (DWS 2004)* [online]. <http://nlp.fi.muni.cz/dws2004/pres/> [Access date: 1 June 2011].
- Svensén, B. (2009). *A Handbook of Lexicography. The Theory and Practice of Dictionary-Making*. Cambridge: Cambridge University Press.

- Thompson, L. (2005). “Pasadena: A Brand New System for the *OED*”. *OED News* December 2005. 4.
- TLex (also known as *TshwaneLex*) [online]. <http://tshwanedje.com/tshwanelex/> [Access date: 1 June 2011].