From *TshwaneLex* to *TshwanePedia*: Creating and Flexibly Maintaining Online Encyclopaedias*

Gilles-Maurice de Schryver, *Department of African Languages and Cultures, Ghent University, Ghent, Belgium and TshwaneDJe HLT, Pretoria, Republic of South Africa (gillesmaurice.deschryver@UGent.be)*, and David Joffe, *TshwaneDJe HLT, Pretoria, Republic of South Africa (david.joffe@tshwanedje.com)*

Abstract: The addition of a restricted number of features to the dictionary (compilation) software *TshwaneLex* suffices to turn this application into a tool for the creation and maintenance of encyclopaedias. This article gives a brief overview of those extra features, using the online encyclopaedia of the *James Randi Educational Foundation* (JREF) as case study.

Keywords: LEXICOGRAPHY, DICTIONARY, ENCYCLOPAEDIA, SOFTWARE, ONLINE, TSHWANELEX, TSHWANEPEDIA, JAMES RANDI EDUCATIONAL FOUNDATION

Samenvatting: Van *TshwaneLex* naar *TshwanePedia*: het samenstellen en flexibel herzien van on line encyclopedieën. De toevoeging van een beperkt aantal uitbreidingen aan de woordenboek(aanmaak)software *TshwaneLex* is voldoende om dit programma met succes voor de samenstelling en herziening van encyclopedieën in te zetten. Dit artikel geeft een kort overzicht van die extra uitbreidingen, en gebruikt de on line encyclopedie van de *James Randi Educational Foundation* (JREF) als illustratie.

Sleutelwoorden: LEXICOGRAFIE, WOORDENBOEK, ENCYCLOPEDIE, SOFTWARE, ON LINE, TSHWANELEX, TSHWANEPEDIA, JAMES RANDI EDUCATIONAL FOUNDATION

1. The off-the-shelf dictionary (compilation) software TshwaneLex

In South Africa, the dictionary (compilation) software *TshwaneLex* is well-known. Development of the application started in Pretoria in mid-2002, and already one year later a first release was in use at the Sesotho sa Leboa *National Lexicography Unit* (NLU). Since then, all members of the eleven NLUs have come into contact with TshwaneLex, either through training sessions organised by the *Pan South African Language Board* (PanSALB) and/or simply as a result of

* An earlier version of this article was presented at the Tenth International Conference of the African Association for Lexicography, organised by the Sesiu sa Sesotho Lexicography Unit, University of the Free State, Bloemfontein, Republic of South Africa, 13–15 July 2005.

the fact that they use TshwaneLex on a daily basis in their respective units. Several commercial dictionary publishers in South Africa, including *Oxford University Press* and *Pharos Dictionaries*, also use or are in the process of acquiring TshwaneLex. Reports of the first South African products compiled and placed online with TshwaneLex may be found in *Lexikos* 13 (De Schryver 2003: 10-12) and *Lexikos* 14 (De Schryver et al. 2004: 56-57, 66).

Over the years, TshwaneLex has also been well-received at all major international lexicography conferences, including TAMA 2003 (Johannesburg), AFRILEX 2003 (Windhoek), DWS 2003 (Brighton), EURALEX 2004 (Lorient), AFRILEX 2004 (Libreville), DWS 2004 (Brno), ASIALEX 2005 (Singapore), PALMA 2005 (Kuala Lumpur), LEXICOM 2005 (Brno), COMPLEX 2005 (Budapest), and AFRILEX 2005 (Bloemfontein). At each of those meetings, the then-latest features of TshwaneLex were introduced, features about which one can read more in the proceedings of each of those conferences.

Today, there are TshwaneLex users in the four corners of the world: from Papua New-Guinea and China in the East, to the United States in the West, from Estonia and Ireland in the North, to South Africa in the South. The dictionary projects are either government-sponsored (e.g. at the *Royal National Academy of Medicine* in Spain, or at the *Research Centre of African Languages and Literatures* in Congo), commercial (e.g. at *Van Dale Lexicografie* in the Netherlands, or at *Macmillan* in Botswana), or private (with users in Japan, Macao, Afghanistan, Albania, Slovenia, the Czech Republic, Germany, Luxembourg, France, the United Kingdom, Kenya, etc.).

Clearly, in order to cover such a wide variety of projects and languages, each with its own unique dictionary structure and needing its own script(s), TshwaneLex had to be a truly off-the-shelf application. To attain this, the software was built around three core concepts: user-friendliness, language-independency, and full customisability. User-friendliness is achieved by means of close cooperation between the developers of the software and numerous beta testers around the world. The language-independent nature of the application is realised thanks to full Unicode support on all levels, which also allows for the simultaneous use of various left-to-right and right-to-left scripts. Customisability is brought about by, among others, a powerful Document Type Definition (DTD) editor and linked styles system. This third aspect, customisability, turned out to be so powerful that it led to two adaptations of the basic TshwaneLex code: *TshwanePedia* for the production of encyclopaedias, and *TshwaneTerm* for the management of terminology. In this article we will be concerned with the former, and in a subsequent one (cf. Joffe and De Schryver 2005a) we will look into the latter.

2. From TshwaneLex to TshwanePedia

One of the most important aspects we felt had to be in TshwaneLex was a high degree of built-in customisability, as each dictionary project has its own struc-

ture and styles, or "style guide". To this end, we built functionality into TshwaneLex to allow end-users to customise the DTD. The DTD defines the structure of articles in the dictionary, and the fields that appear in a specific dictionary. Tied in with the DTD is the styles system, which allows one to customise the entire formatting for all fields (e.g. bold/italics, Times New Roman/Arial, as well as common punctuation to appear before, after or between fields). An in-depth (technical) discussion of the multilayered TshwaneLex DTD editor may be found in Joffe and De Schryver (2005).

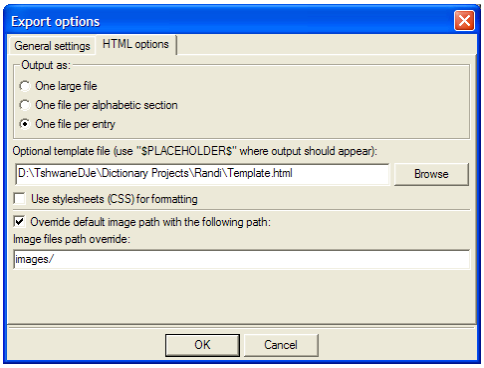
What is important here is that this customisability allows for the creation of other types of reference works with TshwaneLex, not just 'dictionaries'. For example, several TshwaneLex users have (ab)used the software for the creation of bibliographies, address databases, and even diaries.

When the *James Randi Educational Foundation* (JREF) approached us to place their 'Encyclopedia of Claims, Frauds, and Hoaxes of the Occult and Supernatural' (Randi 1995) online, we realised that TshwaneLex was indeed flexible enough to handle such a project. At the same time, however, we seized the opportunity to add a string of additional features to turn Tshwane*Lex* into Tshwane*Pedia*. The extra features, although predominantly useful for the compilation of encyclopaedias, have been 'fed back' into TshwaneLex, thus becoming available for dictionary compilation as well. Three issues will be focused upon, viz. 'window layout', 'multimedia' and 'export' features.

The first difference one notices when comparing a typical dictionary with a typical encyclopaedia, is that encyclopaedia entries are generally much longer than dictionary articles. Additionally, whereas the data of a single dictionary article is normally broken up into many different chunks, with each chunk being placed in a separate and carefully thought-out field in the DTD, encyclopaedia entries are more straightforward. Although it remains important for the compilers of an encyclopaedia to be able to see the 'structure' of the entries they are compiling (in the Tree View), more (horizontal) space is thus often needed for the various input boxes. For that reason, a so-called optional 'Wide Tools window layout' was implemented, which is accessible with a single 'hotkey'. Addendum 1 shows a screenshot of the TshwaneLex interface with the wide view enabled. (Note that when the wide view is not enabled, the entire right side is taken up by a preview of the encyclopaedia entries.)

Secondly, encyclopaedias also typically contain far more illustrations throughout. A new (multimedia) data type 'Image file' was added to that intent. In the DTD for the encyclopaedia shown in Addendum 1, images (and their captions) may be added following any paragraph. All images are stored in a central place, and whenever compilers want to add a new image, they can simply use the 'Browse …' button to select a stored image. This has been taken one step further. Given that the corpus of the future is the Web, TshwaneLex already had a hotkey to launch a *Google Web* search for the lemma sign one is working on, the idea being that one can simply select/adapt corpus lines from the Web. This functionality has been extended to the images, with another hotkey now also launching a *Google Images* search.

The third extra feature concerns flexibility of the export, especially with online encyclopaedias in mind. TshwaneLex already provided several methods for placing reference works online. The online dictionaries described in *Lexikos* 13 (De Schryver 2003: 10-12) and *Lexikos* 14 (De Schryver et al. 2004: 56-57, 66), for example, were placed on the Web with the 'TshwaneLex online software module'. This is a customisable set of PHP scripts that provide functions for creating a search interface where the user can enter words, to perform searches on a TshwaneLex file stored in a MySQL database, and to generate HTML output. In order to decrease the load on the web server, one may rather wish to generate 'static' output, where the reference work is placed online as a pre-generated file or set of files. In this regard the 'Export HTML' features were extended, with options to create one file per alphabetical category or even one file per encyclopaedia entry, in addition to one single large file. In the screenshot shown below, for instance, the output will be generated as one file per entry, with the data for each of those entries being 'dropped' into a template file.



3. Placing the JREF encyclopaedia online

The first edition of James Randi's encyclopaedia was published as a hardcopy in 1995, by St. Martin's Press in New York. In 1996 the not-for-profit *James Randi Educational Foundation* (JREF) was founded "to promote critical thinking by reaching out to the public and media with reliable information about paranormal and supernatural ideas so widespread in our society today" (JREF 1999–2005). When the JREF website was launched in 1999, the idea rose to link the *entire* contents of James Randi's encyclopaedia to the site. Various attempts produced mixed results over the years, and in mid-2005 *TshwaneDJe HLT* (the company which created TshwaneLex) agreed to undertake the task.

The encyclopaedia was received as a set of WordPerfect files, and these were parsed and then imported into TshwaneLex. The material was proofread, corrected and extended, and a TshwaneLex plug-in was created to transform the (implicit) cross-references into hyperlinks. In the process, the three adaptations mentioned above were made to the software. The encyclopaedia pages,

one per alphabetical category, were uploaded to the JREF site on July 28, 2005, and James Randi 'announced' this one day later in his weekly column.

In just four days' time, from August 1 to 4, no less than *sixty* 'bloggers' referred to and commented on the encyclopaedia — an overwhelming response. Today's bloggers clearly complement the feedback strands used so far in our research on (online) dictionary use: "a well thought out log file has been unobtrusively keeping track of all aspects of dictionary use, while an online feedback form has allowed for a more traditional and open way of receiving feedback" (De Schryver and Joffe 2004: 188). One blogger pointed out that large HTML files are cumbersome (= 'implicit feedback'); in the update that went live on August 9, 2005, seven hundred pages were automatically exported and uploaded, one per entry, instead of one per alphabetical category (= 'reaction'). See Addendum 2 for an example of the online encyclopaedia in this regard.

As one can see, *creating* and subsequently flexibly *maintaining* an online encyclopaedia has now become available at every compiler's fingertips thanks to TshwanePedia (just as was already the case for online dictionaries thanks to TshwaneLex).

References

De Schryver, Gilles-Maurice. 2003. Online Dictionaries on the Internet: An Overview for the African Languages. *Lexikos* 13: 1-20.

De Schryver, Gilles-Maurice and David Joffe. 2004. On How Electronic Dictionaries are Really Used. Williams, G. and S. Vessier (Eds.). 2004. *Proceedings of the Eleventh EURALEX International Congress, EURALEX 2004, Lorient, France, July 6–10, 2004*: 187-196. Lorient: Faculté des Lettres et des Sciences Humaines, Université de Bretagne Sud.

De Schryver, Gilles-Maurice, Elsabé Taljard, M.P. Mogodi and Salmina Maepa. 2004. The Lexicographic Treatment of the Demonstrative Copulative in Sesotho sa Leboa — An Exercise in Multiple Cross-referencing. *Lexikos* 14: 35-66.

Google. 1998–2005. Google Search Engine [online]. Available: <http://www.google.com/>.

Joffe, David and Gilles-Maurice de Schryver. 2005. Representing and describing words flexibly with the dictionary application TshwaneLex. Ooi, V.B.Y., A. Pakir, I. Talib, L. Tan, P.K.W. Tan and 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.

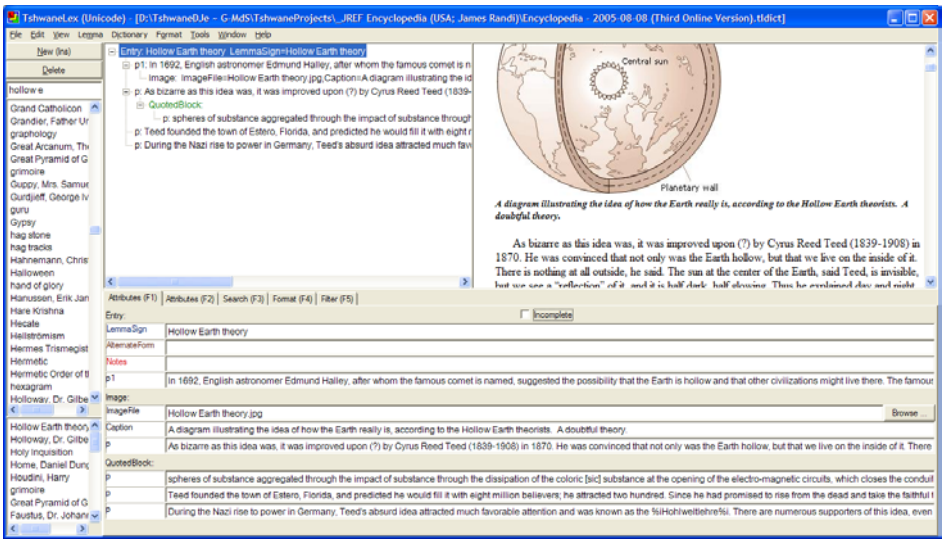
Joffe, David and Gilles-Maurice de Schryver. 2005a. From *TshwaneLex* to *TshwaneTerm*: Tailoring Terminology Management for South Africa. *Lexikos* 15: 312-315.

JREF. 1999–2005. James Randi Educational Foundation [online]. Available: <http://randi.org/>.

Randi, James. 1995. *An Encyclopedia of Claims, Frauds, and Hoaxes of the Occult and Supernatural: James Randi's Decidedly Skeptical Definitions of Alternate Realities*. New York: St. Martin's Press.

TshwaneDJe HLT. 2003–2005. TshwaneDJe Human Language Technology [online]. Available: <http://tshwanedje.com/>.

Addendum 1: Screenshot of TshwaneLex, in 'Wide Tools window layout'



Addendum 2: Screenshot of one entry of the JREF's online encyclopaedia