# Dynamic Metalanguage Customisation with the Dictionary Application TshwaneLex

GILLES-MAURICE DE SCHRYVER

Department of African Languages and Cultures, Ghent University
Rozier 44, 9000 Gent, Belgium
gillesmaurice.deschryver@UGent.be
&
TshwaneDJe Human Language Technology
P.O. Box 299, Wapadrand 0050, Tshwane (Pretoria), South Africa
gillesmaurice.deschryver@tshwanedje.com

DAVID JOFFE

TshwaneDJe Human Language Technology
P.O. Box 299, Wapadrand 0050, Tshwane (Pretoria), South Africa
david.joffe@tshwanedje.com

In the present contribution, the point of departure is the dynamic metalanguage customisation that is realised in real time on the Web for reference works produced with the lexicography software TshwaneLex. This unique feature is absent from even the best electronic dictionaries currently on the market. It is shown that, to achieve this type of customisation, functionality beyond straightforward XML had to be implemented in TshwaneLex. This extra functionality has been made available to the dictionary compilers through a user-friendly editor dialog as part of the fully customisable and built-in DTD. Once set up, the language and format of all metalanguage can not only be easily changed at any point during compilation, but dictionaries can also be customised for particular target users or particular dictionaries (e.g. pocket versus unabridged editions) when outputting for print, while truly instantaneous tailoring is effectively made possible for electronic and online dictionaries.

## The dictionary writing system TshwaneLex

TshwaneLex is professional off-the-shelf lexicography software written in the C++ programming language and built using wxWidgets (an Open Source application development library). The stand-alone version requires a PC with Microsoft Windows 98 / Me / 2000 / XP. For full Unicode support, Windows 2000 or XP is recommended. Storage space and memory requirements are dependent on the size of the dictionary project.

TshwaneLex is currently being used as the lexicographic backbone for several projects at Oxford University Press, Macmillan and Van Dale Lexicografie, among others. Several dozen (smaller) dictionary projects around the world and for a multitude of different languages also make use of the software.

Over the past few years, TshwaneLex has been covered in a number of publications. A general overview and a first elaboration on some computational aspects of TshwaneLex may be found in Joffe *et al*. (2003a), respectively Joffe *et al*. (2003b). Secondly, a lexicographic perspective and an in-depth study of a real-world online lexicographic application are offered in Joffe & De Schryver (2004), respectively De Schryver & Joffe (2004). Thirdly, the fully customisable and built-in DTD editor of TshwaneLex, as well as more advanced DTD aspects are reported on in Joffe & De Schryver (2005), respectively De Schryver & Joffe (2005). Readers are invited to consult those publications for background information, if they so wish, before proceeding with the current discussion.

## Beyond straightforward XML in TshwaneLex & Problem statement

The advanced dictionary-compilation-specific functionality built into TshwaneLex, such as Linked View (whereby implicit links between the two sides of a bidirectional bilingual dictionary are automatically made visible for the lexicographer), Automatic Reversal (whereby single articles or even an entire (semi-)bilingual dictionary may be reversed by the software), or Cross-reference Tracking (whereby cross-reference integrity is ensured at all times by means of the automatic updating of target homonym and sense numbers whenever these change), strongly differentiate TshwaneLex from any ordinary generic XML editor.

Even so, the TshwaneLex DTD system is ‘modelled’ after the XML DTD system, meaning that most of the major components of XML DTDs such as elements, attributes, attribute types, child relations, etc. have been implemented. In some cases, however, ‘special extras’ beyond straightforward XML had to be put into place, precisely to add features that make TshwaneLex more powerful as a dictionary editing environment. The present contribution deals with one such extra, namely the need to be able to dynamically customise the metalanguage, and this throughout compilation, at output stage, as well as during electronic and online use.

Papers in Computational Lexicography, COMPLEX 2005
Eds. Ferenc Kiefer, Gábor Kiss & Júlia Pajzs
Budapest: Linguistics Institute, Hungarian Academy of Sciences

To begin with, and referring to the nature of the metalanguage in bilingual paper dictionaries, Honselaar states:

> Naturally, the meta-language is [in] the native language of the target group. So, in an English-Swedish dictionary for English speakers, comments will be in English. If a set of dictionaries X-Y and Y-X is meant for speakers of both X and Y, the meta-language may consist of words and abbreviations that are common to both languages. A neutral medium such as Latin may also be used. (Honselaar 2003: 324)

This is indeed how lexicographers *used* to go about it, and opting for one of the options is still the case when publishers intend to print only one set of dictionaries to cover all markets simultaneously. In an electronic environment, this need not be the case anymore, of course. In order to illustrate this, two screenshots are shown in Figure 1 reflecting typical instances of the customisation of the output-language in the *Linguistics Terminology Sesotho sa Leboa (Northern Sotho) – English* (Taljard & De Schryver 2003), an online dictionary produced with TshwaneLex.

When using the online dictionary with the interface in English, looking up a word like **karolopolelo** will – apart from the English translation equivalent 'word class, part of speech' – also return the POS tag (noun), label (linguistics) and the cross-reference marker text (SYNONYM) in English. For users who use the Sesotho sa Leboa interface, however, this same information will be *customised* for them, and POS tags, labels and cross-reference marker texts are all displayed in Sesotho sa Leboa (here respectively as **leina**, **popopolelo** and **LEHLALOŠETŠAGOTEE**).

According to De Schryver (2003: 12) the terminology list shown in Figure 1 contained a world's first for any Web dictionary as, at the time, no other Web

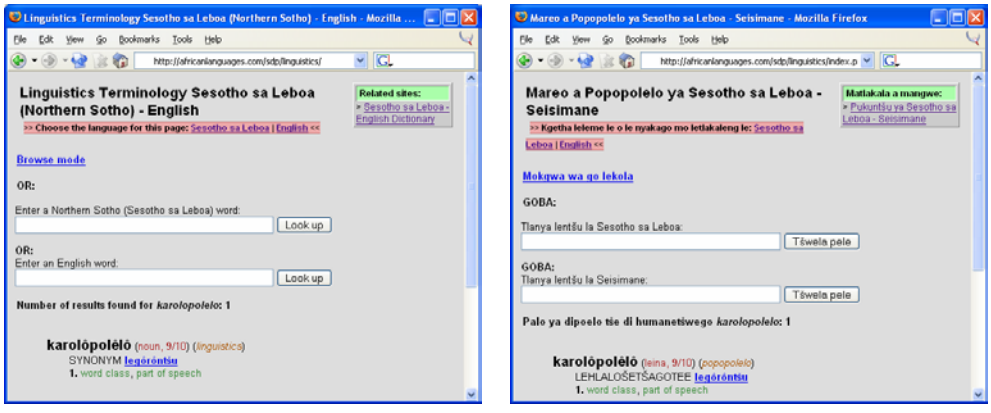


**Figure 1.** Looking up in an online dictionary produced with TshwaneLex, with the interface in English (left) versus Sesotho sa Leboa (right).


**Figure 2.** Looking up in the English – French side of *Le Grand Robert & Collins Électronique* (2003), with the interface in English (left) versus French (right).

dictionary dynamically customised the output-language of POS tags, usage labels and cross-reference marker texts depending on the interface-language chosen. Unfortunately, up to this day, even the better commercial bilingual electronic dictionaries do indeed not achieve this, as is illustrated in Figure 2 for *Le Grand Robert & Collins Électronique* (2003), a bidirectional bilingual French – English / English – French dictionary.

Although the entire interface text is either presented in English or in French depending on the option the user chose, the metalanguage itself is *not* customised. In the example from the screenshots in Figure 2, where the word 'metalanguage' is being looked up, even when consulting the electronic dictionary in a French environment, the POS is still indicated in English as 'noun', and the label is still indicated in English as 'Linguistics', instead of '**nom**' and '**Linguistique**' respectively.

Also note that, while the small 'f' and 'm' in superscript (following the translation equivalents) might stand for both 'feminine noun / **nom féminen**' and 'masculine noun / **nom masculin**' respectively, these abbreviations clearly have not been conceptualised as being part of the metalanguage. If one clicks on 'f' one obtains a new window with two options: 'F - nm' and 'F - abr'. The first leads to "**F, f** … nom masculin …", the second to "**F** … abréviation … franc … Fahrenheit … frère". The options for 'm' lead to: (1) the letter M, m, (2) me / m', (3) the abbreviation for metre, (4) m' (cross-referred to (2)), and (5) the abbreviation for mister. In other words, if one does not already know what these abbreviations stand for, one receives *no* guidance at all, with the first option for 'f'

as ‘nom masculin’ definitely confusing. One is thus forced to go into the help files attached to this electronic dictionary, to the sub-section ‘Symbols and abbreviations / Symboles et abréviations’. This links in with Atkins’ statement:

> All metalanguage should be in the user’s mother tongue (L1). This will obviously involve reduplication of effort at the compiling stage, but in an online dictionary should not result in redundant information at the point of use. (Atkins 1996: 525)

While it is true that presenting just one language to the user should not result in redundant information being offered, it is *not* true that preparing this kind of information involves a reduplication of effort. By and large, the metalanguage of a dictionary is predictable, with POS tags, labels, cross-reference marker texts, and the like, all belonging to closed sets. In a truly modern dictionary compilation program all these metalanguage elements should therefore be selectable from lists (and should thus never be typed in when compiling articles). If one now designs the software in such a way that each of those lists can have as many (customisable) ‘linked variant / alternative lists’ as one wants, then the metalanguage of an entire dictionary can be swapped from one language to another, or from a long form to an abbreviated form, etc., with just one instruction. This is precisely how TshwaneLex was designed.

**On attribute lists in TshwaneLex**

XML DTDs are too limited when it comes to the handling of ‘closed lists’ for practical lexicographic use. It was strongly felt that these were required in TshwaneLex, however, as should be clear from the problem statement above.

In TshwaneLex, lists are stored in a single, central place at the beginning of the XML file. Example [1] shows a section of the PyaSsaL file in this regard, PyaSsaL (Mojela *et al*. 2004) being the in-progress and PanSALB-sponsored monolingual dictionary for Sesotho sa Leboa compiled with TshwaneLex at one of South Africa’s eleven National Lexicography Units:

```
[1]   <dtdlist id="2" name="Part of speech">
        <dtdlistitem id="8" name="noun"/>
        <dtdlistitem id="9" name="pl noun"/>
        <dtdlistitem id="10" name="verb"/>
        <dtdlistitem id="11" name="adjective"/>
      …
        <labelset name="Sesotho sa Leboa">
          <label listitemid="8" name="leina ka botee"/>
          <label listitemid="9" name="leina ka bontši"/>
          <label listitemid="10" name="lediri"/>
```

```
      <label listitemid="11" name="lehlaodi"/>
…
    </labelset>
    <labelset name="Sesotho sa Leboa (abbreviated)">
      <label listitemid="8" name="l.bot."/>
      <label listitemid="9" name="l.bon."/>
      <label listitemid="10" name="ldr."/>
      <label listitemid="11" name="lhl."/>
…
    </labelset>
…
</dtdlist>
<dtdlist id="3" name="Noun class">
    <dtdlistitem id="18" name="0   0/6"/>
    <dtdlistitem id="19" name="1   1/-"/>
    <dtdlistitem id="20" name="1   1/2"/>
    <dtdlistitem id="21" name="1a  1a/2a"/>
    <dtdlistitem id="22" name="2   1/2 p"/>
    <dtdlistitem id="23" name="2a  1a/2a p"/>
…
    <labelset name="Sesotho sa Leboa">
      <label listitemid="18" name="%b0%b/6"/>
      <label listitemid="19" name="%b1%b/-"/>
      <label listitemid="20" name="%b1%b/2"/>
      <label listitemid="21" name="%b1a%b/2a"/>
      <label listitemid="22" name="1/%b2%b"/>
      <label listitemid="23" name="1a/%b2a%b"/>
…
    </labelset>
…
</dtdlist>
```

Each item in the list is given a unique ID. Internally, when one selects a list item on an attribute, it stores, for that attribute in the document, a list of the list item IDs that are selected, rather than the text of the selected items. So for the article **lengwalo**[1] in PyaSsaL, one might have:

[2]
```
<Lemma id="3028" LemmaSign="lengwalo"
HomonymNumber="1" Pronunciation="lengwalô"
PartOfSpeech="8" NounClass="28">
  <Sense id="3029">
    <References id="23888"/>
    <DEF id="3030" Definition="pampiri yeo go
    ngwadilwego atrese le ditaba goba melaetša go
    yona; gantši e phuthelwa ka gare ga omfolopo ya
    romelwa, gantši ka poso"/>
```

```
      <E.G. id="3031" Example="Maphutha o amogela ~
      la go tšwa go kgoro ya thuto"/>
    </Sense>
    <Sense id="3032">
      <DEF id="3033" Definition="pampiri ya bohlatse
      yeo e bontšhago gore motho o na le ditshwanelo
      goba dithuto tše di itšego"/>
      <E.G. id="3034" Example="Ke tla hwetša khuetšo
      le nna ka hwetša ~ la matriki"/>
    </Sense>
  </Lemma>
```

As may be seen from example [2], for 'PartOfSpeech' TshwaneLex stores, internally, '8' rather than '**leina ka botee**' 'singular noun'. This effectively makes it possible to change the text corresponding to '8' in just one place, or to define a substitute text label such as an abbreviated form '**l.bot.**', or even to create a translated version in another language. Likewise, in example [2] the 'NounClass' is stored as '28' rather than '**5**/6', which again means that such a notation may be changed throughout the entire dictionary to, say, '**le-**/ma-' in one go. Clearly, it is thanks to features such as these (singular and plural cross-reference types are handled in a similar way) that the entire metalanguage may easily be customised in dictionaries compiled with TshwaneLex. This further also allows labels to be Unicode text and to consist of any character(s), unlike the 'enumerated list' XML attribute type.

Of course, to do this, and to provide a self-explanatory interface for this – see in this regard Figure 3, which shows one of the tabs of the DTD editor dialog – TshwaneLex is taking care of a number of aspects 'behind the scenes' that an ordinary generic XML editor does not do. Further note that there are two different list types. In the first the lexicographer can only select one item from the list ('one of'), in the second zero or more items may be selected (the latter, again not possible with the XML DTD 'enumerated list' type). For the second type, a difference is also made between 'sorted' and 'unsorted'. For the sorted type, the order of the output of selected list items will always be the same as the order of the items defined centrally for the list. For the unsorted type, the order of the output of selected list items will be the same as the order in which they are selected by the lexicographer. Lastly, also note that any field can be converted from a free text field to a closed list (and vice versa) at any time in TshwaneLex.

Once an attribute list and its alternates have been set up in the DTD (cf. Figure 3), lexicographers may immediately use those to compile their articles, as may be seen from the F2 sub-window in Figure 4. Swapping to an alternate label set for attribute lists (or cross-reference texts) may easily be done under the F4 sub-window, as shown in Figure 5. Changes take immediate effect throughout the entire dictionary database, as seen in the preview, as well as when exporting data.

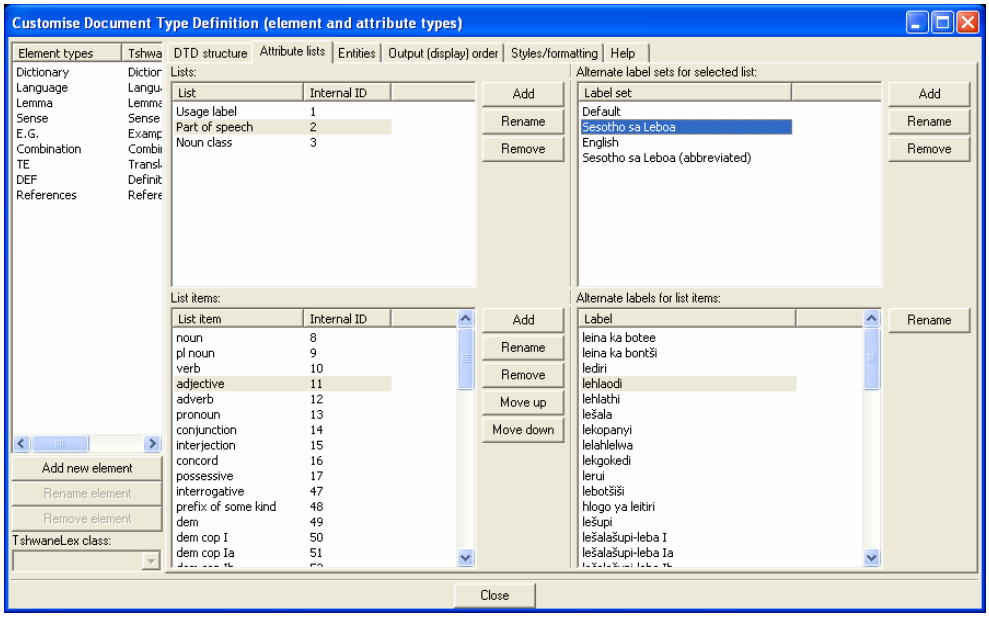


**Figure 3.** TshwaneLex 'attribute lists' editor dialog.

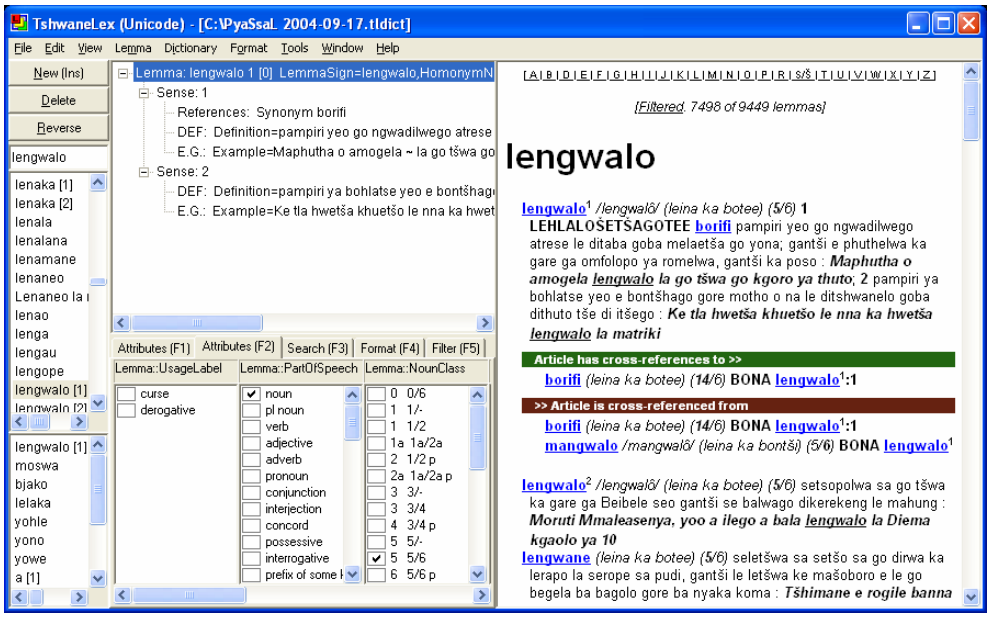


**Figure 4.** Selecting list items (under F2) while compiling in TshwaneLex.

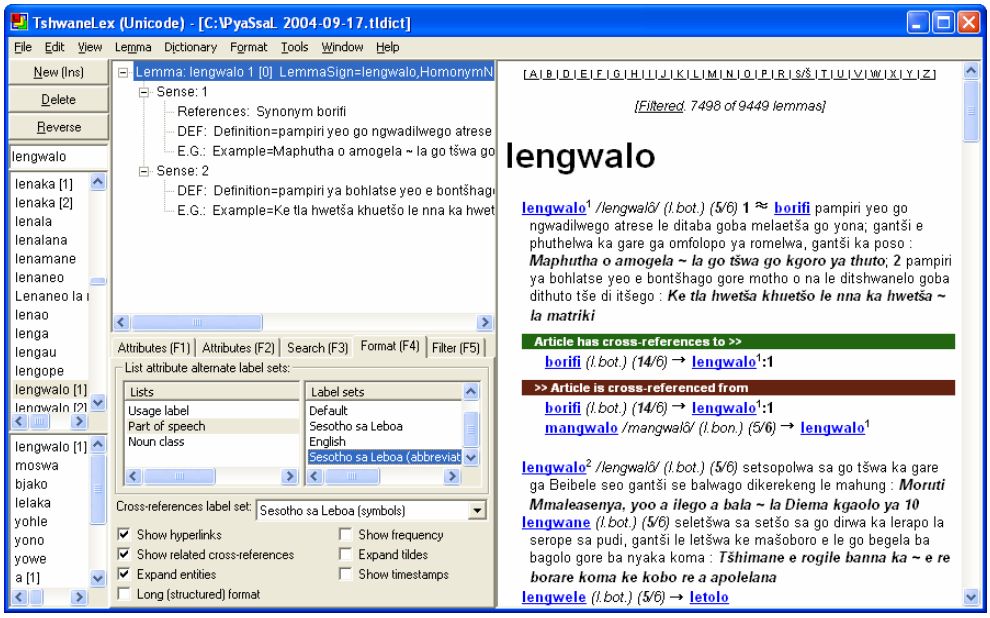


**Figure 5.** Varying the metalanguage (under F4) while compiling in TshwaneLex.

### Dynamic metalanguage customisation with TshwaneLex

The attribute-list system described above provides a powerful yet still easy to use method for customising the metalanguage. This functionality is furthermore carried through to other areas in the system, such as cross-reference type labels, as can be seen in Figure 5. Additionally, still other mechanisms for customisation are available for situations where labels may be embedded within other fields, as for example the French 'f' and 'm' gender labels. These may be defined as so-called XML entities (e.g. '&f;' and '&m;') which are replaced in the output with labels configured in a single, central place. This not only allows electronic or online dictionary software to easily customise the language of these labels, but also allows the software to be *aware* that these labels are part of the metalanguage, and to thus provide a more meaningful response should the user click on them. Clearly, with TshwaneLex, a truly powerful set of tools to fully customise the language of the metalanguage is put into the hands of the lexicographer for the very first time.

### References

**Atkins, B.T. Sue**. 1996. 'Bilingual Dictionaries: Past, Present and Future', in Martin Gellerstam *et al*. (eds.). 1996. *Euralex '96 Proceedings I-II*: 515–546.

Gothenburg: Department of Swedish, Göteborg University.

**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', in Geoffrey Williams and Sandra Vessier (eds.). 2004: 187–196.

**De Schryver, Gilles-Maurice and David Joffe**. 2005. 'One database, many dictionaries – varying co(n)text with the dictionary application TshwaneLex', in Vincent B.Y. Ooi *et al*. (eds.). 2005: 54–59.

**Honselaar, Wim**. 2003. 'Examples of design and production criteria for bilingual dictionaries', in Piet van Sterkenburg (ed.). 2003. *A Practical Guide to Lexicography*: 323–332. Amsterdam: John Benjamins Publishing Company.

**Joffe, David and Gilles-Maurice de Schryver**. 2004. 'TshwaneLex – A State-of-the-Art Dictionary Compilation Program', in Geoffrey Williams and Sandra Vessier (eds.). 2004: 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.). 2005: 108–114.

**Joffe, David, Gilles-Maurice de Schryver and D.J. Prinsloo**. 2003a. 'Introducing TshwaneLex – A New Computer Program for the Compilation of Dictionaries', in Gilles-Maurice de Schryver (ed.). 2003. *TAMA 2003 South Africa: CONFERENCE PROCEEDINGS*: 97–104. Pretoria: $(SF)^2$ Press.

**Joffe, David, Gilles-Maurice de Schryver and D.J. Prinsloo**. 2003b. 'Computational features of the dictionary application "TshwaneLex" '. *Southern African Linguistics and Applied Language Studies* 21/4: 239–250.

*Le Grand Robert & Collins Électronique*. 2003. Dictionnaires Le Robert / VUEF.

**Mojela, M.V.** (Editor-in-Chief), **M.P. Mogodi, M.C. Mphahlele and M.R. Selokela** (Compilers). 2004. *Pukuntšutlhaloši ya Sesotho sa Leboa ka Inthanete* [Explanatory Sesotho sa Leboa Dictionary on the Internet]. Available from: http://africanlanguages.com/psl/

**Ooi, Vincent B.Y.** *et al*. (eds.). 2005. *Words in Asian Cultural Contexts, Proceedings of the 4th Asialex Conference*. Singapore: Department of English Language and Literature & Asia Research Institute, National University of Singapore.

**Taljard, Elsabé and Gilles-Maurice de Schryver**. 2003. *Online Linguistics Terminology Sesotho sa Leboa (Northern Sotho) – English*. Available from: http://africanlanguages.com/sdp/linguistics/

*TshwaneLex*. 2002-2005. Available from: http://tshwanedje.com/tshwanelex/

**Williams, Geoffrey and Sandra Vessier** (eds.). 2004. *Proceedings of the Eleventh EURALEX International Congress*. Lorient: Faculté des Lettres et des Sciences Humaines, Université de Bretagne Sud.