

Title Page

Gilles-Maurice de Schryver

(Ghent U, Belgium, &

5 U of the Western Cape, South Africa)

Residentie Wellington

F. Rooseveltlaan 381

B-9000 Ghent (Belgium)

Tel. +32 (0)9 225 70 39

10 E-mail [gillesmaurice.deschryver@UGent.be](mailto:gillesmaurice.deschryver@UGent.be)

**Running head:** short title = 'Tools to  
Support the Design of a Macrostructure'

**Subject index items:** underlined in red

15 **Abstract:** Two new tools to support the  
design of a dictionary's macrostructure  
are introduced. The first tool is a  
Part-of-Speech Ruler, which is shown to  
be variable for small dictionaries, and  
20 rather stable for (very) large ones. The  
second tool is a Multidimensional  
Lexicographic Ruler, or simply Ruler. A  
largely generic characterisation of what  
Rulers are (instruments to guide the  
25 relative alphabetical breakdown in  
semasiological dictionaries), of how  
they may be built (by drawing on either  
corpus or dictionary data, or both), and  
of how they may be integrated into  
30 dictionary compilation software (by  
being built into TshwaneLex), is then  
presented in detail.

**Three main keywords:** macrostructure,  
lemma-sign list, Rulers

## 121. Tools to Support the Design of a Macrostructure (Nomenclature)

- 40
1. Defining a Dictionary's Macrostructure
  2. On the Need for Rulers, Part 1
  3. Part-of-Speech Rulers ('POS Rulers')
  4. On the Need for Rulers, Part 2
  5. Multidimensional Lexicographic Rulers ('Rulers')
  6. Characterising POS Rulers and Multidimensional
- 45 Lexicographic Rulers
7. Integrating Rulers with Dictionary Compilation Software
  8. Select Bibliography

### 50 **1. Defining a Dictionary's Macrostructure**

A dictionary's macrostructure, nomenclature or lemma-sign list is, in simple terms, the inventory of all the

55 headwords in that dictionary. Each of those lemma signs (headwords) is a canonical form, representing an entire paradigm of morphologically-related forms. The use of dictionary citation

60 forms to group related forms is both a space-saving device that stems from the times of the paper dictionary (now, up to a point, arguably unnecessary in electronic dictionaries), and a device

65 that, implicitly, shows/teaches the morphology of a particular language (which remains a useful feature). In many dictionaries for Indo-European languages, for instance, the dictionary

70 citation form for the verbs is the infinitive form, and the dictionary user is expected to know that all (regular) verbal inflections need to be looked up

under that single canonical form.

75 Conversely, for some word classes the  
dictionary citation form is the only  
member of the paradigm, and in many  
languages this is for instance the case  
for some of the function words. The  
80 process to go from the various  
underlying forms to the canonical forms  
is often referred to as lemmatisation  
(cf. article 120).

A dictionary's macrostructure, then,  
85 is not merely the list of lemma signs,  
as suggested in the first sentence of  
this contribution, but also necessarily  
includes information on the word class  
and morphology of that lemma sign. (1)  
90 to (3) below, therefore, constitute  
three lemmas in the macrostructure of an  
English dictionary:

(1) **record** *noun* [<sup>pl.</sup> records] ...

(2) **record** *verb* [<sup>3p sg</sup> records, *pres. part.*  
95 recording, *past & past part.* recorded] ...

(3) **record** *adjective* [*no comparative nor  
superlative*] ...

Each lemma thus consists of three linked  
components: a lemma sign, a word class,  
100 and information on the morphology. The  
first component is found in all  
dictionaries, the second and third  
components are optional -- although most  
dictionaries do at least include the  
105 word class, while the morphological  
information is, when not mentioned,  
understood.

Observe that different dictionaries will present the macrostructure in varying ways. Most will present (1) to (3) above as homonyms, with numbers in superscript to differentiate between the three lemma signs, possibly with other types of information (such as pronunciation) intersecting the linear sequence shown. More radical options are also found, such as in the COBUILD series of learners' dictionaries, where a long single article combines (1) to (3), and the word classes are presented in an Extra Column, one word class for each and every sense. Given COBUILD senses are ordered according to frequency of use, the word classes are accordingly distributed, with in this case noun, verb, and adjective senses intermixed. Although the macrostructure is distributed 'all over' a dictionary article in COBUILD dictionaries, it should be kept in mind that this is merely a presentational issue. Or, if one prefers, an alternative approach to lemmatisation.

Lemmatisation is thus both language-dependent and dictionary-dependent. Within a certain language, each word class -- also known as grammatical class or part of speech (POS) -- is lemmatised in its own way. For languages with long lexicographic traditions, well-established lemmatisation approaches

have already been devised for the various POSs; for languages that have not yet been reduced to writing or for  
145 which dictionaries have not yet been compiled, various systems may be devised that draw on the morphology of those or cognate languages. Yet, as anyone comparing dictionaries will quickly  
150 notice, even within the same type of dictionaries for the same language, different lemmatisation approaches are in use. COBUILD-style lemmatisation versus lemmatisation in competing  
155 learners' dictionaries for English are a case in point. Generalising across languages, one may say that dictionaries vary in their degree of lumping versus splitting (on this macrostructural  
160 level). Given one deals with a continuum between these two extremes, the variations are endless. For many African languages, for instance, lumping is associated with stem-based dictionaries  
165 where the entries are huge and/or modular in design; whereas splitting is associated with word-based dictionaries with far shorter entries (cf. articles 64 to 70). Similarly, looking at  
170 lemmatisation from the dictionary angle in for instance reference works for Semitic languages, one sees that the distribution of the information is entirely different depending on whether  
175 all inflections of the roots are grouped

under the radicals (usually three  
consonants), or whether (as is the case  
for some modern, basic dictionaries)  
inflected forms are placed in  
180 alphabetical sequence in the  
macrostructure. This macrostructural  
lumping/splitting issue reappears for  
Chinese and Japanese dictionaries (cf.  
articles 61 and 62), and basically for  
185 any dictionaries for any other language.

Given the design of a dictionary's  
macrostructure thus depends on the  
language as well as type of dictionary  
dealt with; there are no generic  
190 computational tools to assist the  
lexicographer with this task.  
Computationally, however, this task can  
be brought back to a simple  
lemmatisation issue, for which see  
195 article 120, and Section 2 below.

## **2. On the Need for Rulers, Part 1**

Today's dictionaries are based on  
corpus data -- the better ones always  
200 have. Whereas such corpora used to  
consist of up to millions of citation  
slips stored in shoebox-like drawers,  
today's corpora are electronic  
collections of text (but also of  
205 multimedia, i.e., text, audio, computer  
graphics, and video -- all interlinked).  
When it comes to macrostructural  
decisions, corpora are especially useful  
in that they make it possible to

210 separate the frequent and average from  
the one-offs; to distinguish the typical  
from the oddities. Trivially, then, the  
frequent, average and typical are to be  
selected for the core of general-purpose  
215 dictionaries, while the one-offs and  
oddities ought to find their way into  
hard-word lists (only), large native-  
speaker dictionaries (as opposed to  
learners' dictionaries), etc. At face  
220 value, drawing up the macrostructure of  
a particular dictionary therefore seems  
straightforward, as one may use  
occurrence frequencies (summed on the  
level of the canonical forms) as main  
225 arbiter for decisions on inclusion  
versus omission. One may also require  
that the items that make up a paradigm  
are evenly spread across at least a  
certain number of sub-corpora, where  
230 those sub-corpora represent a number of  
different (transcribed) text types or a  
number of different genres. Depending on  
the kind of dictionary being compiled,  
one may also wish to focus on words that  
235 have recently entered the language, or,  
conversely, only include words with an  
even distribution across the decades  
(with a further option to either drop or  
mark archaic words and/or uses).

240 Furthermore, during actual dictionary  
compilation, several other factors also  
play a role, such as (1) the need/wish  
to complete certain paradigms of

interlinked, closed-class or  
245 grammatically similar words, (2)  
pedagogical, cultural or even political  
restraints/obligations, (3) purely  
commercial or academic pressures for the  
addition or removal of this or that  
250 word, etc. The log files and feedback  
forms attached to online dictionaries  
may furthermore indicate that the users  
truly want to see certain words treated  
(cf. article 24, Section 3). Another  
255 often-overlooked aspect is the one  
linked to the fact that dictionaries are  
at times released without them having  
been properly completed, in order to  
meet (unfortunate) deadlines.  
260 Consequently, regardless of the actual  
lemmatisation approach (or even the  
absence of any such approach), published  
dictionaries at least need to be  
representative of a lexicon's  
265 distribution at all times. Implementing  
this concept leads to a novel set of  
tools, known as 'Rulers'.

Indeed, during modern dictionary  
compilation one is working in all  
270 alphabetical categories simultaneously,  
constantly making changes throughout the  
(growing) dictionary database. In order  
to keep such a project in shape, a tool  
is required that allows for the  
275 continuous monitoring of various  
dictionary aspects. More in particular,  
and with specific reference to a

dictionary's macrostructure, one needs  
 to be able to easily check whether or  
 280 not items that belong to certain POSs  
 are not accidentally over- or  
 underrepresented, and one needs a way to  
 monitor the relative distribution of the  
 lemma signs across the alphabetical  
 285 categories.

Note that reference is made to the distribution of  
 'alphabetical categories' henceforth. In most dictionaries  
 for Semitic languages the corresponding distribution refers  
 to triplets of consonants, for Chinese and Japanese  
 290 dictionaries covering Kanji characters the corresponding  
 distribution may refer to stroke counts, and so on. Rather  
 than single letters, the 'alphabet' may also contain  
 digraphs, trigraphs, etc. All 'alphabetical' claims that  
 follow remain applicable however, mutatis mutandis.

295 At all times monitoring the POS as  
 well as the alphabetical-category  
 distribution must be potentially  
 combined with frequency thresholds. In  
 other words, given a certain frequency  
 300 range of the dictionary database, are  
 the POSs well distributed and are the  
 alphabetical categories well  
 distributed?

### 305 **3. Part-of-Speech Rulers ('POS Rulers')**

In the present section, the attention  
 goes to the so-called 'POS Ruler'.  
 Intuitively, one may look at two extreme  
 cases first. Imagine one would want to  
 310 compile a dictionary of the top 100  
 items only, culled from a lemmatised  
 frequency list. It is well known that  
 such a list will, for most of the  
 world's languages, contain mainly

315 function words such as (here for  
English): the, of, and, a, but, or, etc.  
At the other extreme, imagine one starts  
with a billion-word corpus and  
undertakes to lemmatise all forms.  
320 Obviously nearly all, if not all,  
function words -- being members of  
closed classes -- will be in the  
resulting dictionary. As for content  
words, while going down the frequency  
325 list, 'new nouns' will typically be  
personal names, names of products and  
companies, place names, etc. Increasing  
such a mega-corpus further might, in  
addition to new nouns, reveal 'new  
330 verbs' or 'new adjectives' every now and  
then, but the likelihood of say 'new  
conjunctives' being discovered is about  
nil. If one thus looks at the  
distribution of the POSs in a billion-  
335 word corpus, compared to the  
distribution in a corpus twice as large,  
then one will notice that the percentage  
allocation to the nouns continues to  
grow: the larger the corpus the more  
340 nouns relatively speaking. All real-  
world dictionaries fall somewhere in-  
between those extremes.

These intuitive thought experiments  
are confirmed when one studies the  
345 distribution of the POSs in corpora. See  
in this regard, for English, Figs. 121.1  
and 121.2. These graphs were constructed  
using the POS information for the

unlemmatised types in the 100-million-  
350 word British National Corpus (BNC),  
reworking data provided by Leech et al.  
(2001).

Fig. 121.1 shows the distribution of  
the POSs, at each rank anew, for the  
355 most frequent 1,000+ types in the BNC.  
One may clearly deduce from this graph  
that function words and verbs dominate  
the top-frequent ranks in an English  
corpus. The percentage of nouns grows  
360 steadily as one goes down the frequency  
list. At the 1,000+ mark the overall  
percentage of nouns already stands at  
40%, that of the verbs at 20%, while the  
function words shrank to 16% of the  
365 total (whereas these still represented  
roughly two thirds at the 100 mark).

Fig. 121.2 zooms out, and shows the  
distribution for each single rank down  
to all items that occur at least ten  
370 times per million tokens in the BNC, of  
which there are 7,726 in all. The  
allocation to the nouns at the 7,000+  
mark now stands at 52%, that to the  
verbs grew to 22%, while the function  
375 words shrank to a mere 4% of the total.  
These graphs can be extended down to any  
rank, while the same type of  
calculations can of course also be  
performed on lemmatised frequency lists,  
380 with similar results.

From Figs. 121.1 and 121.2 one can  
further deduce that there are as many

POS Rulers as there are ranks in a  
corpus-derived frequency list and/or as  
385 there are lemma signs in a dictionary. A  
dictionary of 10,000 items will have a  
certain POS distribution, one of 50,000  
items another one, and so on. With a POS  
Ruler at hand, one may thus give  
390 informed answers to claims (often found  
in reviews of dictionaries) such as:  
"Your dictionary contains too many  
nouns; you should have included far more  
verbs instead!" Perhaps surprisingly,  
395 but the reply to such a claim will  
depend on the size of the dictionary in  
question. Graphs such as those presented  
in Figs. 121.1 and 121.2 also indicate  
why the production of especially small  
400 or pocket dictionaries is non-trivial  
from a corpus perspective. If counts  
from a single corpus are used and  
nothing else, then the percentage of  
function words in a reference work with  
405 just a few thousand lemma signs will be  
relatively high. Many dictionary  
compilers find this unacceptable, and  
prefer to focus on content words like  
nouns and verbs rather, with the obvious  
410 skewing as a result. One example of how  
to deal with the creation of the  
macrostructure of small dictionaries has  
been proposed by De Schryver/Prinsloo  
(2003). They propose to work with two  
415 corpora, a large general-language one,  
and a smaller customised corpus

consisting of the same type of material  
the intended target user group of the  
dictionary will encounter. About four-  
420 fifths of the dictionary's lemmas are  
basically the top section of the  
lemmatised frequency list derived from  
the customised corpus, with the  
remaining fifth having low or even zero  
425 frequencies in the customized corpus,  
but chosen for their top frequencies in  
the general-language corpus. For large  
dictionaries on the other hand, and  
returning to Figs. 121.1 and 121.2,  
430 given POS Ruler changes become  
asymptotic with increasing dictionary  
sizes, one may conclude that rather  
stable POS distributions may indeed be  
derived for (very) large dictionaries.

435

#### **4. On the Need for Rulers, Part 2**

In a semasiological dictionary (i.e.,  
one in which the lemma signs are  
presented in alphabetical sequence), the  
440 different alphabetical categories are of  
course not equal in size. This is so  
obvious and trivial that one ought not  
to mention it. Even so, in one of the  
seventy texts included in the first  
445 anthology of lexicography (Hartmann  
2003), Serianni (II 198) writes with  
regard to a certain historical  
dictionary of modern Italian: "Whereas  
917 pages were sufficient to cover the  
450 letter A, 3058 were required for P and

3611 for S; adding up the total number of pages, A + P + S (= 7586 pages), the percentages are as follows: A 12.1%, P 40.3%, S 47.6%, with an evident  
455 imbalance to the disadvantage of letter A." This is a disturbing claim. The only way to know whether A is out of balance or not, is to compare the size of A with the allocation predicted by some kind of  
460 instrument that reflects the alphabetical distribution of the Italian (historical) lexicon. Claiming or implying that all alphabetical categories should be nicely balanced out  
465 compared to one another throughout a dictionary is simply outrageous. Imagine a general-language English dictionary where the letter X would be as large as the letter S -- that would be an  
470 extraordinary feat indeed. But then, the letters S and X may be of comparable sizes -- in a general Tsonga dictionary for example. So, yes, different languages simply have different word-  
475 initial letter distributions, but by no means are such distributions 'even' across the alphabet. One is therefore in need of an instrument that reflects the alphabetical distribution in  
480 dictionaries. That instrument -- as it turns out, a second type of ruler -- has been termed a 'Multidimensional Lexicographic Ruler', or more often just 'Ruler' (in contrast to 'POS Ruler').

485 The early beginnings of the design of  
a Ruler can be traced back to 1999-2000,  
when a team at the University of  
Pretoria was in search of an informed  
way to combat various types of  
490 inconsistencies encountered in the  
existing South African dictionaries (cf.  
e.g., De Schryver/Prinsloo 2000: 293-  
297, 2001: 376-380; Prinsloo/De Schryver  
2001: 192-195). On a macrostructural  
495 micro-level, for instance, the  
researchers noticed that the compilers  
of a number of bilingual dictionaries  
missed out on items likely to be looked  
for by their target users, presumably  
500 simply because these items did not cross  
the lexicographers' way during  
compilation. Cross-comparing the  
nomenclature of various dictionaries is  
of course a 'classic' in lexicography.  
505 In the mid-1980s, for example, Crystal  
contrasted sample pages from comparable  
English dictionaries. Focussing on lemma  
signs he observed that "the discrepancy  
factor (that is, the number of head  
510 words not shared divided by the number  
of head words shared) can be as much as  
30 per cent" (1986: 75). Likewise,  
Herbst (1990: 1380-1381) counted the  
number of 'corresponding entries'  
515 between fixed points in several English  
learners' dictionaries and arrived at a  
similar conclusion. For Afrikaans, Gouws  
(1985: 14-15) noted with concern that

the Woordeboek van die Afrikaanse Taal

520 (WAT) started to make use of more and more pages with each published volume as compared to the "number of pages for corresponding alphabetical stretches" in a standard Afrikaans desk dictionary.

525 All these tests point in the same direction, namely that there is a dire need for a sound measurement instrument with regard to alphabetical distribution to assist dictionary makers.

530

### **5. Multidimensional Lexicographic Rulers ('Rulers')**

From the start it was clear that the to-be-developed Ruler should: (1)

535 preferably be based on distributions derived from corpora in the current era of corpus lexicography, (2) preferably take advantage of distributions in existing dictionaries on the condition  
540 that these had been compiled according to systematic principles, and (3) be made to reflect the envisaged dictionary lemmatisation policies.

In a way, this aim thus addresses one  
545 of the fundamentals in lexicography, namely the overall distribution of the lemma signs in the central text of a semasiological dictionary. Surprisingly, however, just a handful of studies have  
550 been devoted to this issue. In the mid-twentieth century E.L. Thorndike (manually) devised a 'block system of

distribution of dictionary entries by  
initial letters', which has been  
555 discussed and reprinted by Landau in his  
textbook (1984: 241-243). Svensén, who  
read and reviewed the first edition of  
Landau's textbook (Svensén 1992), then  
wrote in his own textbook (1993: 242):  
560 "A decision must also be made as to what  
fraction of the whole dictionary each  
initial letter may occupy, so that the  
size of the finished dictionary can be  
kept under control during the course of  
565 the project. The percentages for each of  
the various initial letters in a given  
entry language are fairly constant, and,  
if such calculations have not already  
been done by others, it is wise to  
570 examine the distribution in a number of  
representative dictionaries."

The only other reference found in the  
literature in this regard is the  
following by Coutsogeorgopoulos et al.  
575 (2000: 128) who discuss the creation of  
a lemma-sign list for a Greek (to  
English) dictionary: "First, the number  
of lemmas for each letter of the  
alphabet of the Greek language is  
580 defined. The frequency of occurrence of  
each word extracted from the corpus as  
well as reliable bilingual or  
monolingual Greek dictionaries are  
consulted to balance the number of  
585 lemmas contained under separate letters  
of the Greek alphabet."

The existing literature, then, does not go beyond providing a block system for (American) English, the advice to  
590 examine existing dictionaries, and the suggestion to also consult corpora. In contrast, in De Schryver's PhD thesis (2004: 209-224) every step in the design of Multidimensional Lexicographic Rulers  
595 is accompanied by a solid argumentation. The overarching finding of the undertaken research is that there is a remarkable consistency for both dictionary and corpus distributions.

600 As far as dictionaries are concerned, alphabetical distributions across various existing dictionaries compare very well. Moreover, distributions based on space-allocation measurements  
605 correlate strongly with distributions in which exact lemma-sign counts are used instead. In addition, deriving such dictionary statistics from reference works several decades apart once more  
610 results in near-identical distributions.

On the corpus side, an alphabetical breakdown of all the types in a full corpus is equivalent with the alphabetical breakdown of the types in  
615 subsections of that corpus, as well as with the alphabetical breakdown of the top-frequent types only. In addition, for the Germanic languages English and Afrikaans, both unlemmatised and  
620 lemmatised corpus data once again result

in near-identical alphabetical distributions.

Implicit in the last finding is a surprising correlation, one which  
625 suggests that one has come full circle. Given that corpus lemmatisation produces canonical forms, and that a list of canonical forms corresponds with a lemma-sign list, which in turn  
630 correlates with space allocations in a dictionary, the experiments simply indicate that a direct comparison between corpus types and alphabetical categories in dictionaries is possible.  
635 If this is confirmed, then corpus data (full or partial, all or top-frequent only, etc.) and dictionary data (whether based on page-allocation measurements or lemma-sign counts, derived from one or  
640 more dictionaries, etc.) should also correlate well. At that point, dictionary data and corpus data may successfully be united (averaged) to produce a reliable and stable Ruler.  
645 This hypothesis will now be tested (and confirmed), taking English as an example.

In Tables 121.1 and 121.2 the distributions in four different sources  
650 are compared. These four sources are (1-2) the exact lemma-sign counts in two American dictionaries (Newbury 1999 and Heritage 2000), (3) the page counts in a British dictionary, namely the Macmillan

655 English Dictionary for Advanced Learners  
(MEDAL, Rundell 2002), and (4) the  
letter-initial distribution in a corpus  
of world Englishes (compiled at the  
University of Pretoria by R. Gauton). As  
660 may be deduced from the matrix in Table  
121.2, the four components correlate  
well, and this although exact lemma-sign  
counts, precise page measurements and  
plain corpus breakdowns are compared  
665 with one another, and this for  
respectively American, British and World  
Englishes. It is thus clear that Rulers  
may indeed be built by averaging  
dictionary data on the one hand and  
670 corpus data on the other -- which is  
also what is done in the penultimate  
column of Table 121.1.

The list of values in bold seen in  
Table 121.1 is referred to as a  
675 Multidimensional Lexicographic Ruler, in  
this case for general-language English  
dictionaries, as this tool operates on  
multiple levels: obviously the  
macrostructural one, which blends into  
680 the microstructure, but also on the  
level of the planning and management of  
dictionary projects. Stemming from the  
appearance of Rulers as true 'physical  
rulers', their first and natural  
685 function could be said to be to measure  
various dictionary aspects. As such any  
existing dictionary may be scrutinised,  
and measurements on either lemma-sign or

space-allocation levels, or both, may be  
690 performed in order to evaluate the  
'overall balance' in that particular  
dictionary. One thus sees that the Ruler  
as a 'measurement instrument' is also an  
'evaluation instrument'. If such an  
695 evaluation would reveal certain  
dictionary stretches to be over- or  
under-treated on either lemma-sign or  
space-allocation level, or both, and the  
observed deviations cannot be explained  
700 except as instances of inconsistent  
treatment on the side of the  
lexicographer(s), then the Ruler 'rules'  
that the observed imbalances ought to be  
rectified in revised editions. An  
705 example of a detailed analysis of a  
large multi-volume dictionary project  
along those lines can be found in De  
Schryver (2005).

For dictionaries that need to be  
710 compiled from scratch, or even for  
dictionaries that are still being  
compiled, a Ruler can also be used to  
predict various dictionary aspects. The  
Ruler as a 'measurement/evaluation  
715 instrument' is thus also a 'prediction  
instrument'. Continuing this line of  
thinking, if one is able to predict  
various aspects, then one may logically  
also use a Ruler to plan and thus to  
720 manage various aspects of a dictionary  
project. The Ruler as a  
'measurement/evaluation and prediction

instrument' is thus also a 'management  
instrument'. (The latter was also  
725 implicit in the quote from Svénson at  
the start of Section 5.)

## **6. Characterising POS Rulers and Multidimensional Lexicographic Rulers**

730 Thus far two rulers, viz., a POS Ruler  
and a Multidimensional Lexicographic  
Ruler have been introduced as valuable  
tools to support the design and  
monitoring of a dictionary's  
735 macrostructure. While the former is  
dynamic in that there are as many POS  
Rulers as there are ranks in a corpus-  
derived frequency list and/or as there  
are lemma signs in a dictionary, the  
740 latter is very stable indeed. It was  
nonetheless shown that POS Rulers do  
'stabilize', but for (very) large  
dictionaries only. In contrast, and in  
its most basic form, a Multidimensional  
745 Lexicographic Ruler, or simply a Ruler,  
is a unique abstract entity since a  
single series of percentages  
straightforwardly correlates with the  
relative alphabetical allocation in  
750 semasiological dictionaries. As such,  
each alphabetical category is assigned a  
certain percentage, reflecting the  
relative size of that category.  
Different languages, and even different  
755 types of dictionaries for a specific  
language (with possibly varying

lemmatisation approaches), have  
different Rulers. However, given a  
certain lemmatisation policy, a general-  
760 language Ruler for a particular language  
is valid for all general-language  
dictionaries for that particular  
language, and is also stable over time.  
Just as physical rulers with which one  
765 measures, Rulers can be made as fine-  
grained as one wishes by simply breaking  
down the alphabetical categories into  
smaller sections. And just as the human  
rulers who govern us, Rulers can be  
770 called in to manage dictionary projects.

Reformulated, a Ruler is a powerful  
instrument with which measurements/  
evaluations and predictions/monitoring  
can be performed on various macro- and  
775 microstructural dictionary levels. These  
levels are meta-levels since one is not  
taking any individual semantics (e.g.,  
the number of senses) or individual  
grammar (e.g., the POSs) of single lemma  
780 signs into account; one rather works  
with averages across a dictionary. It is  
precisely because averages are used that  
the composing components (viz., corpus  
counts and dictionary measurements)  
785 crystallised into a single Ruler.

To conclude this section, and  
summarizing the research presented in De  
Schryver (2004: 209-224), Rulers may be  
characterised and built as follows:

790 (1) A Ruler is an instrument to guide

the relative alphabetical distribution  
in semasiological dictionaries. Ideally,  
this instrument averages counts derived  
795 corpus, and the exact measurement of  
space allocations in existing  
dictionaries. Moreover, both the corpus  
and the reference dictionaries should  
cover the same domain as the intended  
800 dictionary for which the alphabetical  
distribution is required, as general-  
language and special-purpose reference  
works behave differently on the  
macrostructural axis.

805 (2) The research also showed that  
these conditions may successfully be  
weakened or strengthened along the  
following lines:

(2.1) On the corpus side:

810 \* In the absence of a large  
automatically lemmatised corpus, and for  
languages with a simple word-initial  
morphology as well as very regular  
inflection patterns, the data may also  
815 be derived from an unlemmatised corpus  
(and such a corpus might even be  
preferred for some types of  
dictionaries).

\* As dictionary-lemmatisation policies  
820 vary greatly, a corpus that is neither  
fully lemmatised nor fully unlemmatised  
might be required, but rather a corpus  
with an automated tailored  
lemmatisation.

825 \* If computational support for an  
automated lemmatisation is not  
available, a manually lemmatised corpus  
can be used (and, though laborious to  
construct, might even result in the best  
830 type of lemmatisation for the task at  
hand).

\* As a shortcut, drawing data from a  
(top-frequent) section only of a corpus  
will largely result in a comparable  
835 alphabetical breakdown.

(2.2) On the reference-dictionaries  
side:

\* If exact space-allocation measurements  
cannot be made or are not available,  
840 lemma-sign counts are a good  
approximation (and might even be  
favoured in some cases).

\* For general-language dictionaries data  
from reference works spanning several  
845 decades may be employed instead of  
solely time-restricted dictionary data.

(2.3) On the corpus and reference-  
dictionaries side:

\* In case either corpora or either  
850 comparable dictionaries are not  
available, the instrument to guide  
alphabetical distribution may be:

(i) based solely on one or more  
comparable dictionaries, or  
855 (ii) based solely on corpus data.

\* Guidance may also be provided for  
stretches smaller (or larger) than  
entire alphabetical categories.

(3) The core practical implementations  
860 are:

\* The devised measurement/evaluation  
instrument can point out imbalances  
across the alphabetical categories  
and/or across any dictionary stretches  
865 in existing dictionaries small and  
large, imbalances that can be addressed  
in revised editions.

\* The devised measurement/evaluation  
instrument is especially useful for very  
870 large (and thus long-term) multi-volume  
dictionary projects in progress, where  
imbalances cannot only be addressed in  
revisions, but where reliable strategies  
can be set up to effectively steer or  
875 (re)direct future compilation so as to  
prevent major (additional)  
inconsistencies. At this point the  
measurement/evaluation instrument thus  
becomes a prediction instrument.

880 \* The devised measurement/evaluation and  
prediction instrument enables one to put  
forward guidelines which are especially  
useful when used as a framework to  
assist in the compilation of new  
885 dictionaries. These guidelines revolve  
around:

(i) space allocation per alphabetical  
category or any dictionary stretch  
(e.g., expressed as a certain number of  
890 pages in hardcopy dictionaries), and

(ii) the number of lemma signs per  
alphabetical category or any dictionary

stretch.

\* The devised measurement/evaluation and  
 895 prediction instrument further also  
 suggests an average article length  
 (expressed in number of articles per  
 page, in number of column-lines per  
 article, or in number of words or  
 900 characters per article).

\* The devised measurement/evaluation and  
 prediction instrument is also a  
 management instrument in that it can be  
 used to allocate work and to keep track  
 905 of the lexicographers' progress.

## **7. Integrating Rulers with Dictionary Compilation Software**

In an ideal lexicography project,  
 910 compilers have access to both a POS  
 Ruler and a Multidimensional  
 Lexicographic Ruler at all times. In the  
 professional dictionary compilation  
 software TshwaneLex, for instance, both  
 915 have also seamlessly been integrated. On  
 the one hand TshwaneLex allows for the  
 extraction of all lemma signs with a  
 certain POS, within any frequency range,  
 and to have the allocation calculated.  
 920 These values may then be compared with  
 the POS Ruler values for the same  
 frequency range. On the other hand the  
 Ruler 'percentages' may be input, and at  
 any one point TshwaneLex may be asked to  
 925 visualize the dictionary database, or  
 any section of it, compared to the

Ruler. In Fig. 121.3, for instance, the distribution across the alphabet of the number of lemma signs that are marked as  
 930 'completed' in a monolingual dictionary for Northern Sotho, is compared to the Ruler breakdown. The graph in Fig. 121.4 compares the Ruler to both the space allocation (whereby characters are  
 935 counted) and the number of lemma signs per alphabetical category. From these visualisations one may immediately conclude that, at that point in time, E and I were under-treated, while M and S  
 940 were over-treated. TshwaneLex also allows any number of frequency bands to be marked, through the automatic and dynamic display of the relevant lemma signs in varying colours, or the  
 945 addition of star ratings, etc.

In conclusion one can therefore say that POS Rulers and Multidimensional Lexicographic Rulers are practical tools that enable lexicographers to keep their  
 950 dictionary databases balanced and representative -- on the macrostructural level and beyond.

## 8. Select Bibliography

- 955 Coutsogeorgopoulos, H./Kokkinakis, G./Dermatas, E. (2000):  
 KORAI: A Large Electronic Greek - English Dictionary with Spoken Pronunciation. In: Workshop Proceedings of COMLEX 2000. Kato Achaia, 127-130.
- 960 Crystal, D. (1986): The ideal dictionary, lexicographer and user. In: Ilson, R.F. (ed.), *Lexicography: An emerging international profession*. Manchester, 72-81.
- De Schryver, G.-M. (2004): *Concepts and Tools for Lexicography in the Electronic Age -- A case study of*

- 965 dictionary compilation in South Africa. PhD Thesis.  
Ghent University.
- De Schryver, G.-M. (2005): Concurrent Over- and Under-Treatment in Dictionaries -- The Woordeboek van die Afrikaanse Taal as a Case in Point. In: *International Journal of Lexicography* 18, 1, 47-75.
- 970 De Schryver, G.-M./Prinsloo, D.J. (2000): Electronic corpora as a basis for the compilation of African-language dictionaries, Part 1: The macrostructure. In: *South African Journal of African Languages* 20, 4, 291-309.
- 975 De Schryver, G.-M./Prinsloo, D.J. (2001): Corpus-based Activities versus Intuition-based Compilations by Lexicographers, the Sepedi Lemma-Sign List as a Case in Point. In: *Nordic Journal of African Studies* 10, 3, 374-398.
- 980 De Schryver, G.-M./Prinsloo, D.J. (2003): Compiling a lemma-sign list for a specific target user group: The Junior Dictionary as a case in point. In: *Dictionaries* 24, 28-58.
- Gouws, R.H. (1985): Die sewende deel van die Woordeboek van die Afrikaanse Taal. In: *Standpunte* 38, 1, 13-25.
- 985 Hartmann, R.R.K. (ed.) (2003): *Lexicography: Critical Concepts* (3 volumes). London.
- Herbst, T. (1990): *Dictionaries for Foreign Language Teaching: English*. In: Hausmann, F.J. et al. (eds.), *Wörterbücher, Ein internationales Handbuch zur Lexikographie*. Berlin, 1379-1385.
- 990 Landau, S.I. (1984): *Dictionaries: The Art and Craft of Lexicography*. New York.
- Leech, G.N./Rayson, P./Wilson, A. (2001): *Companion Website for: Word Frequencies in Written and Spoken English: based on the British National Corpus*. Available: <http://ucrel.lancs.ac.uk/bncfreq/> [Access 1 July 2007].
- 995 Prinsloo, D.J./De Schryver, G.-M. (2001): Taking Dictionaries for Bantu Languages into the New Millennium -- with special reference to Kiswahili, Sepedi and isiZulu. In: Mdee, J.S./Mwansoko, H.J.M. (eds.), *Makala ya kongamano la kimataifa Kiswahili 2000. Proceedings*. Dar es Salaam, 188-215.
- 1000 Serianni, L. (2003): A Survey of Contemporary Italian Lexicography. In: Hartmann, R.R.K. (ed.), II 195-210.
- 1005 Svensén, B. (1992): Book Review: Sidney I. Landau. 1984. *Dictionaries: The Art and Craft of Lexicography*. In: *International Journal of Lexicography* 5, 1, 79-83.
- Svensén, B. (1993): *Practical Lexicography: Principles and Methods of Dictionary-Making*. Oxford.
- 1010 TshwaneLex. Available: <http://tshwanedje.com/tshwanelex/> [Access: 1 July 2007].

Gilles-Maurice de Schryver  
(Ghent University, Belgium, & University of the Western  
Cape, South Africa)

Fig. 121.1

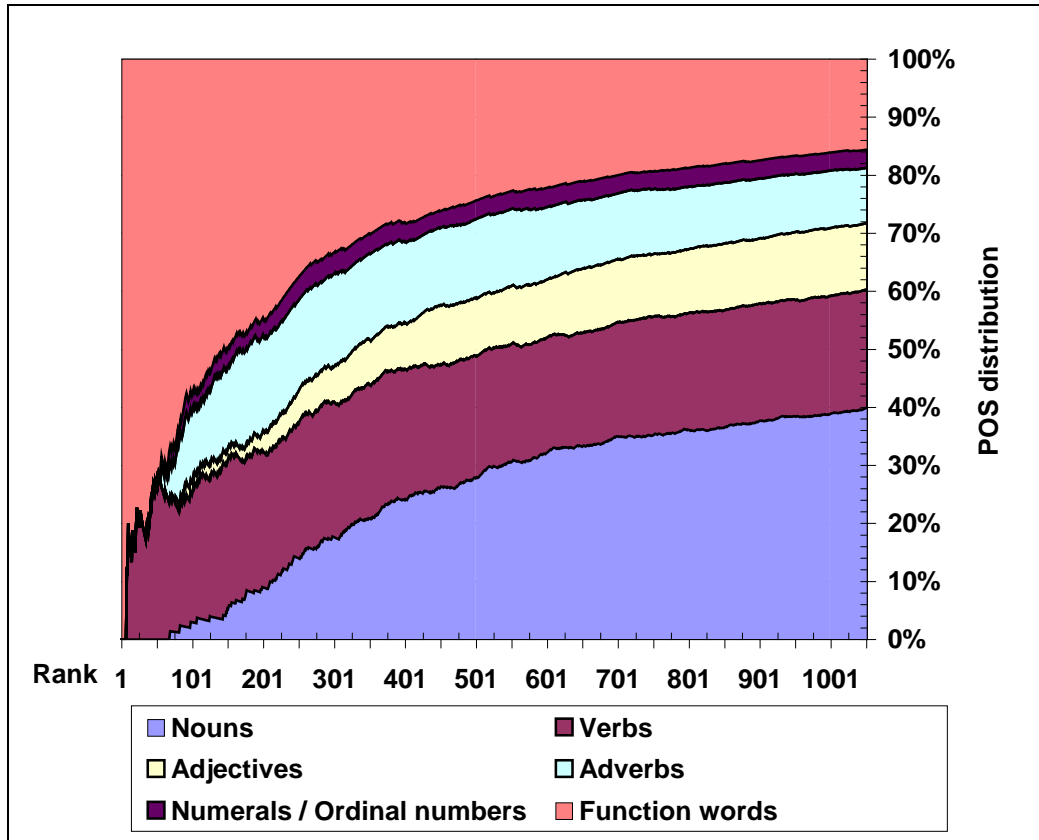


Fig. 121.1: POS (Part-of-Speech) distribution of the top 1,000+ types in the unlemmatised BNC (British National Corpus)

Fig. 121.2

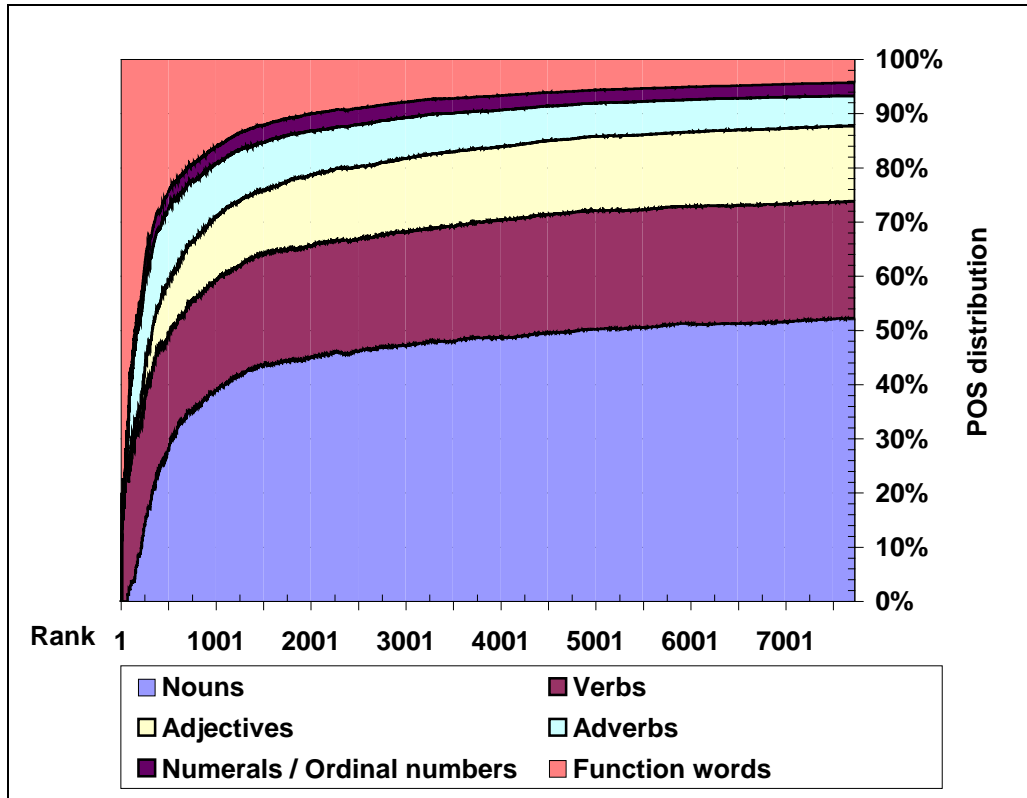


Fig. 121.2: POS (Part-of-Speech) distribution of the top 7,000+ types in the unlemmatised BNC (British National Corpus)

Table 121.1

Table 121.1: The design of an English general-language Ruler

|          | D1              |        | D2               |        | D3                    |        | C                                |        | D1-D3+C          |          |
|----------|-----------------|--------|------------------|--------|-----------------------|--------|----------------------------------|--------|------------------|----------|
|          | Newbury<br>1999 |        | Heritage<br>2000 |        | MEDAL<br>Rundell 2002 |        | English corpus<br>(12.5M tokens) |        | English<br>Ruler |          |
|          | lemma<br>signs  | %      | lemma<br>signs   | %      | pp.                   | %      | all<br>types                     | %      | %                |          |
| <b>A</b> | 1,302           | 5.48   | 5,800            | 6.31   | 82.145                | 4.94   | 7,697                            | 6.51   | <b>5.81</b>      | <b>A</b> |
| <b>B</b> | 1,256           | 5.28   | 6,031            | 6.57   | 105.752               | 6.37   | 7,210                            | 6.10   | <b>6.08</b>      | <b>B</b> |
| <b>C</b> | 2,059           | 8.66   | 8,752            | 9.53   | 155.822               | 9.38   | 10,759                           | 9.10   | <b>9.17</b>      | <b>C</b> |
| <b>D</b> | 1,606           | 6.75   | 4,564            | 4.97   | 89.369                | 5.38   | 6,590                            | 5.57   | <b>5.67</b>      | <b>D</b> |
| <b>E</b> | 935             | 3.93   | 3,166            | 3.45   | 56.107                | 3.38   | 4,595                            | 3.88   | <b>3.66</b>      | <b>E</b> |
| <b>F</b> | 1,269           | 5.34   | 3,794            | 4.13   | 86.346                | 5.20   | 4,646                            | 3.93   | <b>4.65</b>      | <b>F</b> |
| <b>G</b> | 820             | 3.45   | 3,305            | 3.60   | 57.846                | 3.48   | 4,193                            | 3.54   | <b>3.52</b>      | <b>G</b> |
| <b>H</b> | 993             | 4.18   | 3,915            | 4.26   | 68.668                | 4.13   | 4,681                            | 3.96   | <b>4.13</b>      | <b>H</b> |
| <b>I</b> | 854             | 3.59   | 2,927            | 3.19   | 57.294                | 3.45   | 4,244                            | 3.59   | <b>3.45</b>      | <b>I</b> |
| <b>J</b> | 241             | 1.01   | 946              | 1.03   | 14.911                | 0.90   | 1,722                            | 1.46   | <b>1.10</b>      | <b>J</b> |
| <b>K</b> | 162             | 0.68   | 1,285            | 1.40   | 12.561                | 0.76   | 1,887                            | 1.60   | <b>1.11</b>      | <b>K</b> |
| <b>L</b> | 814             | 3.42   | 3,293            | 3.59   | 64.299                | 3.87   | 4,452                            | 3.76   | <b>3.66</b>      | <b>L</b> |
| <b>M</b> | 1,171           | 4.92   | 5,537            | 6.03   | 80.131                | 4.82   | 7,634                            | 6.45   | <b>5.56</b>      | <b>M</b> |
| <b>N</b> | 550             | 2.31   | 2,222            | 2.42   | 33.182                | 2.00   | 2,829                            | 2.39   | <b>2.28</b>      | <b>N</b> |
| <b>O</b> | 701             | 2.95   | 2,292            | 2.50   | 44.537                | 2.68   | 2,753                            | 2.33   | <b>2.61</b>      | <b>O</b> |
| <b>P</b> | 1,911           | 8.04   | 7,398            | 8.05   | 133.874               | 8.06   | 8,636                            | 7.30   | <b>7.86</b>      | <b>P</b> |
| <b>Q</b> | 117             | 0.49   | 454              | 0.49   | 7.575                 | 0.46   | 603                              | 0.51   | <b>0.49</b>      | <b>Q</b> |
| <b>R</b> | 1,387           | 5.83   | 3,864            | 4.21   | 88.093                | 5.30   | 5,475                            | 4.63   | <b>4.99</b>      | <b>R</b> |
| <b>S</b> | 2,499           | 10.51  | 10,495           | 11.43  | 209.336               | 12.60  | 12,299                           | 10.40  | <b>11.23</b>     | <b>S</b> |
| <b>T</b> | 1,352           | 5.69   | 4,910            | 5.35   | 94.579                | 5.69   | 5,886                            | 4.98   | <b>5.43</b>      | <b>T</b> |
| <b>U</b> | 577             | 2.43   | 1,735            | 1.89   | 32.098                | 1.93   | 2,839                            | 2.40   | <b>2.16</b>      | <b>U</b> |
| <b>V</b> | 306             | 1.29   | 1,673            | 1.82   | 17.869                | 1.08   | 2,371                            | 2.00   | <b>1.55</b>      | <b>V</b> |
| <b>W</b> | 754             | 3.17   | 2,587            | 2.82   | 61.411                | 3.70   | 3,140                            | 2.65   | <b>3.08</b>      | <b>W</b> |
| <b>X</b> | 10              | 0.04   | 137              | 0.15   | 0.467                 | 0.03   | 127                              | 0.11   | <b>0.08</b>      | <b>X</b> |
| <b>Y</b> | 98              | 0.41   | 383              | 0.42   | 5.290                 | 0.32   | 549                              | 0.46   | <b>0.40</b>      | <b>Y</b> |
| <b>Z</b> | 34              | 0.14   | 387              | 0.42   | 1.790                 | 0.11   | 465                              | 0.39   | <b>0.27</b>      | <b>Z</b> |
|          | 23,778          | 100.00 | 91,852           | 100.00 | 1,661.352             | 100.00 | 118,282                          | 100.00 | <b>100.00</b>    |          |

Table 121.2

Table 121.2: Pearson correlation coefficients  $r$  between the different components of the English general-language Ruler

| $r$      | Newbury      | Heritage     | MEDAL        | Corpus       |
|----------|--------------|--------------|--------------|--------------|
| Newbury  | <b>1.000</b> | 0.964        | 0.981        | 0.966        |
| Heritage |              | <b>1.000</b> | 0.979        | 0.992        |
| MEDAL    |              |              | <b>1.000</b> | 0.966        |
| Corpus   |              |              |              | <b>1.000</b> |

Fig. 121.3

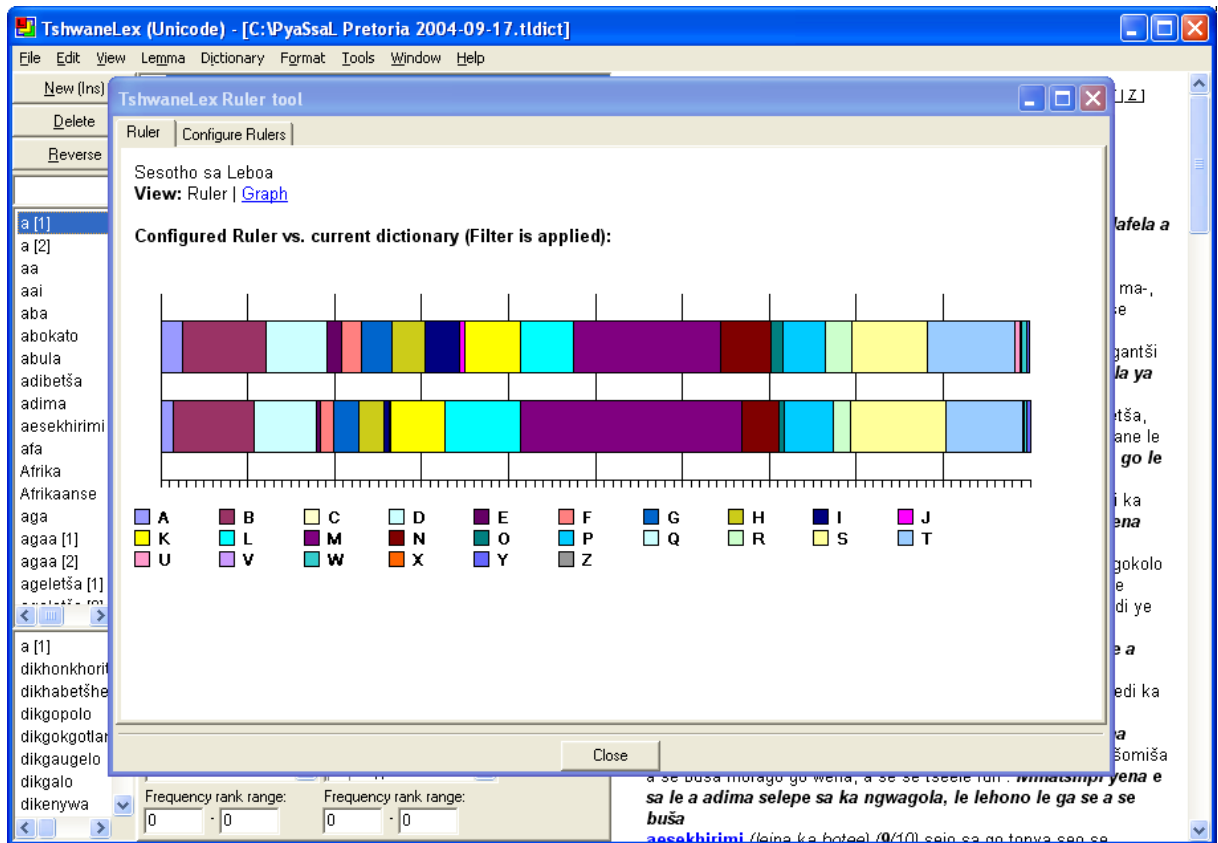


Fig. 121.3: TshwaneLex's Ruler tool: Physical Rulers  
(Data © 2004 Sesotho sa Leboa NLU)

Fig. 121.4

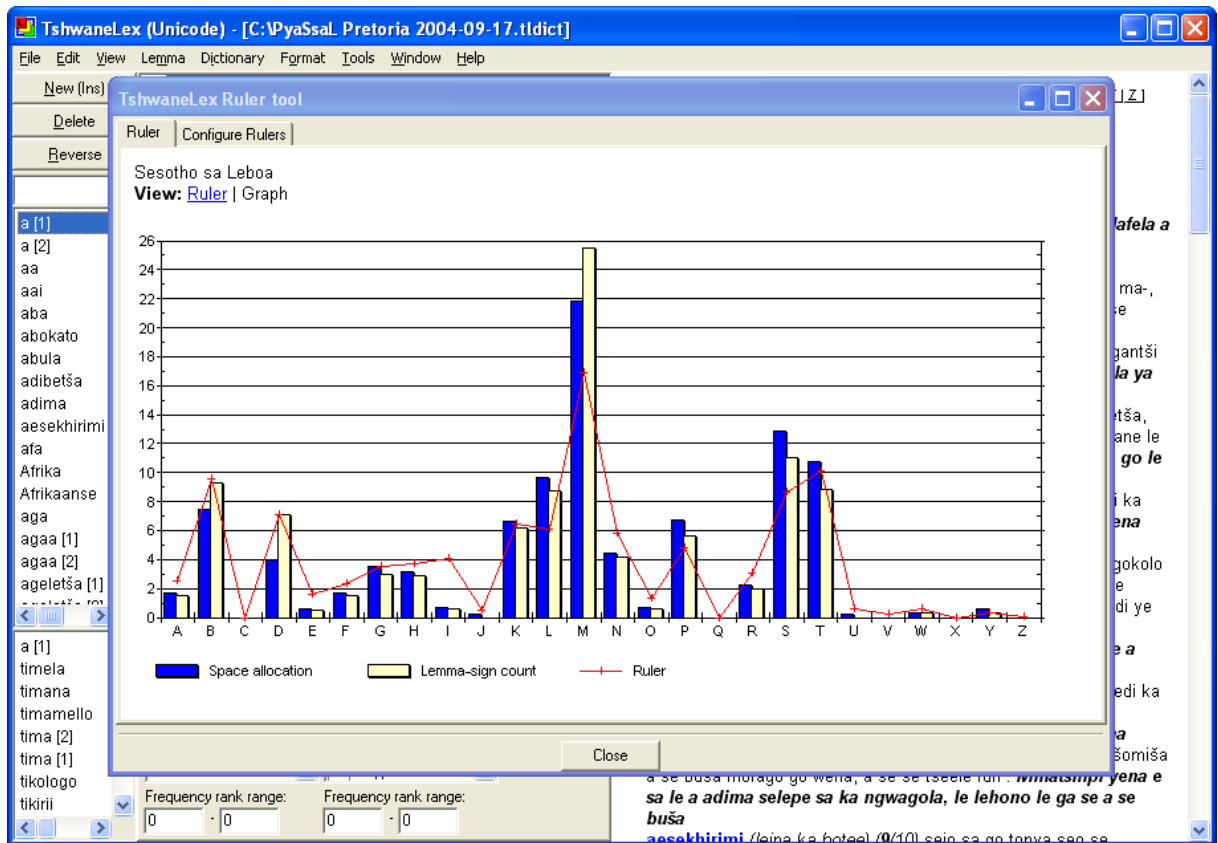


Fig. 121.4: TshwaneLex's Ruler tool: Over-Under graph (Data © 2004 Sesotho sa Leboa NLU)