Designing a Measurement Instrument for the Relative Length of Alphabetical Stretches in Dictionaries, with special reference to Afrikaans and English

D.J. Prinsloo & Gilles-Maurice de Schryver
Department of African Languages, University of Pretoria
Pretoria 0002, South Africa
prinsloo@postino.up.ac.za || gillesmaurice.deschryver@rug.ac.be

Abstract

In this paper it is argued that the lexicographer, in planning the macrostructure of his/her dictionary, should calculate the breakdown into alphabetical stretches in terms of overall length and the number of lemma signs to be treated. This should be done by using the suggested breakdown of types in a corpus as well as the average breakdown in a number of other dictionaries as an instrument of measurement and prediction. This will prevent situations where the lexicographer 'gets tired' towards the end or where a specific category is enthusiastically over-treated resulting in imbalances with regard to the other alphabetical categories. The development of the measurement and prediction instrument is illustrated for Afrikaans and English lexicography.

Introduction

Table 1 shows a typical inconsistency on the macrostructural level where the compilers of a number of dictionaries accidentally missed out on words most likely to be looked for by their target users, simply because these words did not cross the compilers' way during the compilation process.

D1	D2	D3	D4	D5	D6	D7	D8
English-Setswana *Snyman 1990*	English-Setswana *Matumo 1993*[4]	English-Sepedi *Kriel 1976*[4]	English-Sepedi *Kriel et al. 1997*[4]	English-Afrikaans *Juta 1983*[6]	English-Afrikaans *Kromhout 1997*[13]	LDOCE3 (English) *Summers 1995*[3]	COBUILD3 (English) *Sinclair 2001*[3]
funnel	—	funnel	funnel	funnel	funnel	0 / 0	◆◇◇◇◇
—	funny	funny	funny	funny	funny	S1 W3	◆◆◆◇◇
—	—	—	—	funny-bone	—	0	0
—	fur	fur	fur	fur	fur	0 / 0	◆◆◇◇◇
—	—	—	—	furbelow	—	—	—
—	—	—	—	—	~ cloak	—	—
—	—	—	—	—	~ coat	—	—
—	fur-cap	fur-cap	—	—	—	—	—
—	—	fur-lined	—	—	—	—	—
—	furbish	—	—	—	—	0	—
—	—	furcate	—	furcate	—	—	—
furious	furious	furious	furious	furious	furious	0	◆◆◇◇◇
—	—	—	—	furl	—	0	0
—	—	furlong	furlong	—	—	0	◆◇◇◇◇
—	—	furlough	—	furlough	—	0	0
—	furnace	furnace	furnace	furnace	furnace	0	◆◇◇◇◇
—	—	furnish	furnish	furnish	furnish	0	◆◇◇◇◇

—	—	—	—	—	~er	—	—
—	—	—	—	furnishings	—	0	◆◇◇◇◇
—	—	furniture	furniture	furniture	furniture	S2 W3	◆◆◆◇◇
—	—	—	—	—	piece of ~	—	—
—	—	furore	furore	furore	—	0	◆◇◇◇◇
—	—	—	—	furrier	—	0	0
—	furrow	furrow	furrow	furrow	furrow	0 / 0	◆◇◇◇◇
—	—	—	—	furry	—	0	◆◇◇◇◇
—	further	further	further	further	further	S1 W1 / 0 / 0	◆◆◆◆◆
furthermore	—	—	furthermore	furthermore	~more	W3	◆◆◇◇◇

Table 1: Comparisons of the macrostructure between the fixed points *funnel* and *furthermore* in various bilingual dictionaries, including the respective frequencies as found in LDOCE3 and COBUILD3

In Table 1 the first three dictionaries (D1-3) are small desktop ones, while the next three (D4-6) are pocket editions. These six dictionaries viewed together cover 27 items in the stretch *funnel* to *furthermore*. D5 offers treatment for 19, D3 for 15, D6 for 14, D4 for 12, D2 for 8, while D1 treats only 3. The latter misses out on commonly used words such as *funny*, *furniture*, *further*, etc., while D2 and D3, e.g., do not treat the frequently used item *furthermore*. One cannot but deplore the fact that precious space has been allocated to words which are unlikely to be looked up by the target users (such as *fur-cap*, *fur-lined*, etc.), whilst highly used words have been omitted.

Already in the mid-1980s Crystal compared sample pages from comparable English dictionaries. Focussing on lemma signs he observed that "the discrepancy factor (that is, the number of head words not shared divided by the number of head words shared) can be as much as 30 per cent" [1986: 75]. The data in Table 1 wholly confirm this.

The aim of this paper is to analyse inconsistencies in Afrikaans and English dictionaries when it comes to space allocation or the relative length of alphabetical stretches, by treating certain sections of the lemma-sign list more exhaustively than others. 'Space allocation' and the 'relative length of alphabetical stretches' will be used as broad cover terms for physical dictionary space, i.e. the number of pages used for treating a specific alphabetical category / stretch in the dictionary. Over- and under-treatment will be viewed in terms of (a) the number of lemma signs treated within the specific alphabetical category / stretch, and (b) the average length of the dictionary articles.

Lexicographers are quick to point out that a dictionary cannot and should not try to cover or treat 'all' the lexical items from a particular language, let alone try to find 'all' the senses of the selected lemma signs, and that it is impossible to quote the totality of a word's occurrences in written and spoken communication. If space restriction is applicable to the dictionary as a whole, it must also be relevant to *each individual alphabetical category* in the dictionary. The compiler of a dictionary can over-treat the alphabetical category **A** to such an extent that the remaining categories are under-treated in order to fit into the physical number of pages allocated to the dictionary as a whole, or the compiler can simply grow tired towards the end of the alphabet. The question is thus: what urges the compiler to move on to

B? Is it merely the fact that one is limited to x pages by the publisher and that **A** should therefore not fill more than x divided by the number of letters in the alphabet? Or does one intuitively reach a stage where, taking into consideration issues such as the proficiency level of the target user group and the physical limitations on the dictionary, it is 'time to move on'? What should the breakdown for each alphabetical category in terms of physical length be? Are there reliable ways to determine whether the space allocation to each alphabetical category can be justified, especially in respect of each category weighted up against each other category?

Thorndike's Block System

Already half a century ago, Edward L. Thorndike worked out a system in an effort to answer the latter question. He studied the American lexicon and devised a 'block system of distribution of dictionary entries by initial letters' [Landau 2001: 360–362]. Thorndike divided the alphabet into 105 blocks: 6 for **A** (A1: a-adk, A2: adl-alh, A3: ali-angk, …), … 1 for **J** (J50: j-jz), … 3 for **W** (… , W104: wit-wz) and 1 for **XYZ** (XYZ105: x-zz). With approximately the same weight assigned to each of those blocks, this series supposedly reflects the 'distribution of lexical units throughout the alphabet'. Focusing on the alphabetical categories as a whole, 6 blocks for **A** corresponds with 5.71% of the entire alphabet, 1 block for **J** with 0.95%, 3 blocks for **W** with 2.86%, etc. Our transformation of 'Thorndike blocks' into percentages is shown in Figure 1, together with the distribution of our actual lemma-sign count for two recent American dictionaries [Newbury 1999; Heritage 2000[4]] and one early American dictionary [Webster 1913].

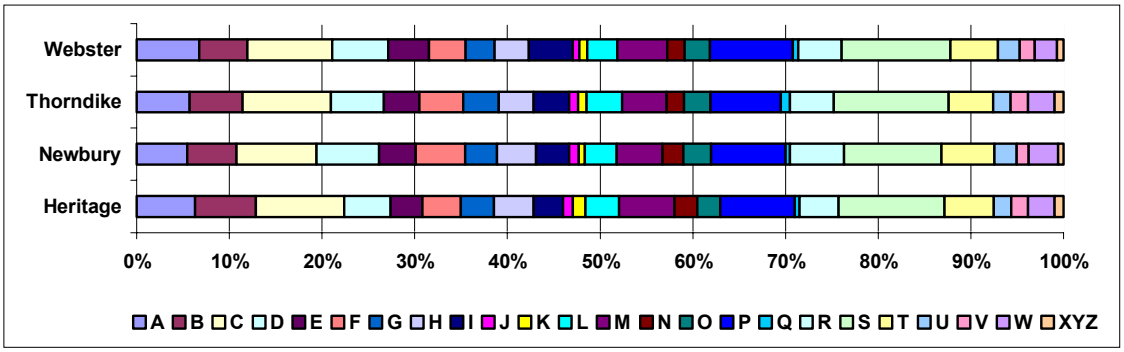

Figure 1: Thorndike's block system compared to various American dictionaries

Whether compiled half a century before or half a century after Thorndike devised his block system, Figure 1 shows that the lemma-sign distribution in the selected American dictionaries indeed seems to *roughly* follow the breakdown suggested by Thorndike.

Space Allocation to Alphabetical Stretches in Afrikaans

As a point of departure for Afrikaans, five different dictionaries were chosen and the space per alphabetical category was calculated both in terms of the number of pages and the percentage of dictionary space each of these categories fills in the dictionary. The data are shown in Table 2.

	D1		D2		D3		D4		D5		D1-5
	Afrikaans-Sepedi *Kriel 1983*[3]		Afrikaans *Kritzinger & Eksteen 1984*[4]		Afrikaans-English *Grobbelaar 1987*		Afrikaans-Sepedi *Ziervogel & Mokgokong 1988*[4]		Afrikaans *Odendal & Gouws 2000*[4]		
	pp.	%	pp.	%	pp.	%	pp.	%	pp.	%	%
A	24.2	**9.00**	29.1	5.98	35.8	5.71	4.0	6.32	56.8	**4.14**	6.23
B	33.3	**12.38**	39.0	8.02	46.3	7.38	5.5	8.69	68.0	**4.96**	8.29
C	1.0	0.37	1.4	0.29	2.6	0.41	0.1	0.16	4.4	0.32	0.31
D	15.1	5.61	21.5	4.42	28.0	4.47	3.1	4.90	58.0	4.23	4.73
E	6.6	2.45	9.8	2.01	13.6	2.17	1.3	2.05	27.4	2.00	2.14
F	3.9	1.45	6.7	1.38	7.8	1.24	0.5	**0.79**	23.4	1.71	1.31
G	16.4	6.10	23.0	4.73	39.4	6.28	3.4	5.37	97.9	7.14	5.92
H	12.5	4.65	16.2	3.33	28.3	4.51	2.3	3.63	84.4	6.16	4.46
I	6.8	2.53	8.7	1.79	13.2	2.11	1.0	1.58	57.4	**4.19**	2.44
J	2.0	0.74	2.4	0.49	4.4	0.70	0.6	0.95	11.9	0.87	0.75
K	20.8	7.73	45.2	9.29	48.8	7.78	4.2	6.64	131.3	9.58	8.20
L	10.0	3.72	22.8	4.69	21.0	3.35	2.8	4.42	40.9	2.98	3.83
M	10.2	3.79	24.5	5.04	26.6	4.24	2.6	4.11	51.6	3.76	4.19
N	5.2	1.93	11.2	2.30	14.7	2.34	1.7	2.69	25.8	1.88	2.23
O	10.9	4.05	25.2	5.18	51.7	**8.25**	3.8	6.00	72.7	5.30	5.76
P	5.2	**1.93**	22.5	4.62	27.4	4.37	2.4	3.79	59.8	4.36	3.82
Q	0.0	0.00	0.1	0.02	0.2	0.03	0.0	0.00	0.3	0.02	0.01
R	7.0	2.60	19.8	4.07	19.8	3.16	2.1	3.32	55.3	4.03	3.44
S	26.2	9.74	66.6	13.69	74.2	11.83	7.6	12.01	192.7	14.05	12.26
T	9.9	3.68	21.5	4.42	26.3	4.19	2.2	3.48	55.9	4.08	3.97
U	5.0	1.86	7.7	1.58	11.4	1.82	1.1	1.74	28.2	2.06	1.81
V	25.5	9.48	40.5	8.32	56.8	9.06	7.3	11.53	123.4	9.00	9.48
W	10.7	3.98	19.8	4.07	26.9	4.29	3.3	5.21	40.7	2.97	4.10
X	0.1	0.04	0.2	0.04	0.4	0.06	0.1	0.16	0.7	0.05	0.07
Y	0.4	0.15	0.9	0.18	1.1	0.18	0.2	0.32	1.8	0.13	0.19
Z	0.1	0.04	0.2	0.04	0.3	0.05	0.1	0.16	0.5	0.04	0.06
	269.0	100.00	486.5	99.99	627.0	99.98	63.3	100.02	1371.2	100.01	100.00

Table 2: Breakdown of alphabetical categories in five randomly selected Afrikaans dictionaries

From Table 2 it is clear that the space (expressed in %) allocated to the different alphabetical categories is reasonably consistent across the five dictionaries, with the exception of individual discrepancies such as those indicated in bold.

When these dictionaries, D1 to D5, are compared with the huge multi-volume overall-descriptive *Woordeboek van die Afrikaanse Taal* (WAT), in compilation for the past 75 years and published up to the letter **O** (volume XI), huge discrepancies are manifest. In Table 3 space allocation in WAT is compared to that of D1-5. Space allocation in the latter dictionaries has been recalculated to only cover the categories **A** to **O** in order to enable comparison with WAT.

	D1	D2	D3	D4	D5	D1-5	DIFFERENCE		WAT		
	pp.	pp.	pp.	pp.	pp.	%	abs. %	rel. %	%	pp.	
A	24.2	29.1	35.8	4.0	56.8	10.18	-5.82	**-57.18**	4.36	316.0	**A**
B	33.3	39.0	46.3	5.5	68.0	13.52	-9.68	**-71.62**	3.84	278.3	**B**
C	1.0	1.4	2.6	0.1	4.4	0.51	+0.44	+86.45	0.95	68.7	**C**
D	15.1	21.5	28.0	3.1	58.0	7.76	-2.08	-26.86	5.68	411.7	**D**
E	6.6	9.8	13.6	1.3	27.4	3.51	-0.40	-11.40	3.11	225.7	**E**
F	3.9	6.7	7.8	0.5	23.4	2.16	-0.03	-1.25	2.13	154.6	**F**
G	16.4	23.0	39.4	3.4	97.9	9.75	-2.97	-30.40	6.79	492.3	**G**
H	12.5	16.2	28.3	2.3	84.4	7.33	-0.62	-8.49	6.71	486.7	**H**
I	6.8	8.7	13.2	1.0	57.4	4.01	-0.66	-16.45	3.35	243.2	**I**
J	2.0	2.4	4.4	0.6	11.9	1.24	+0.28	+22.70	1.52	110.3	**J**
K	20.8	45.2	48.8	4.2	131.3	13.54	+20.64	**+152.42**	34.19	2479.0	**K**
L	10.0	22.8	21.0	2.8	40.9	6.33	+0.23	+3.66	6.56	476.0	**L**
M	10.2	24.5	26.6	2.6	51.6	6.92	+1.62	+23.40	8.54	619.4	**M**
N	5.2	11.2	14.7	1.7	25.8	3.69	-0.40	-10.73	3.29	238.8	**N**
O	10.9	25.2	51.7	3.8	72.7	9.53	-0.56	-5.84	8.98	650.9	**O**
	178.9	286.7	382.2	36.9	811.9	99.98	***r* = 0.608**		100.00	7251.6	

Table 3: Breakdown of alphabetical categories in D1-5 compared to WAT (up to **O**)

The discrepancies are big for the categories **A** and **B** and *extremely* big for **K**. This comparison thus suggests a serious over-treatment of **K** and an under-treatment of **A** and **B** in WAT. From a WAT angle it could of course be argued that D1-5 over- or respectively under-treated these categories. However viewed, the correlation coefficient *r* between the D1-5 breakdown and the WAT breakdown is as low as 0.608.

What one really needs at this point is a *dictionary-external arbiter* that could be used as a reliable basis on which to base relative lengths of alphabetical stretches. As we will seek that arbiter from corpus data, we must first check which kind of corpus data one actually needs.

Corpora: Lemmatised versus Unlemmatised Corpora

For Indo-European languages like Afrikaans or English, there is no one-to-one relation between the 'corpus types' (derived from a 'raw corpus') and the 'canonical (or dictionary citation) forms'. Taken at face value, one would therefore need to do counts on a 'lemmatised corpus'. The (automated) lemmatisation of a corpus is dependent on the (automated) POS-tagging of that same corpus, and the latter is never error-free. Consequently, a lemmatised 'item-POS pair' corpus is never error-free either. Moreover, different dictionaries have different 'dictionary lemmatisation policies', with some favouring a strict canonical approach, others deviating from it to assign lemma-sign status to problematic items. Depending on the needs, one will therefore require counts derived from either a lemmatised or an unlemmatised corpus. It is thus crucial to compare the alphabetical breakdown for the extremes: lemmatised *versus* unlemmatised, and this for both item-POS types and for true raw types. One could assume that the distribution of (a) the 'item-POS types *versus* canonical forms' and of (b) the 'raw types *versus* canonical forms', is rather even across the different alphabetical categories, so that it doesn't matter which kind of corpus one queries. As an illustration, we will verify the first assumption for English and the second for Afrikaans.

For English, the 100-million-word *British National Corpus* (BNC) was queried. The complete BNC contains 100,106,029 tokens and 939,028 item-POS types. We used a subset by focussing on those items that occur over 5 times only, which left us with 96,516,432 tokens and 196,575 item-POS types. As far as the unlemmatised breakdown is concerned, all the item-POS types per alphabetical category in this BNC subset were counted. For the lemmatised breakdown, however, the daunting task was undertaken to manually lemmatise the top-frequency section of this BNC subset. Lemmatisation was done with a hypothetical LDOCE-type of dictionary in mind, thus separating, for instance, nominal and verbal forms of the same item. Once the top-frequency section of the BNC subset had been manually lemmatised, the first 10,000 lemmatised canonical forms were cut off, and the breakdown was calculated for them. The BNC results are shown in Table 4a.

British National Corpus (BNC)							
	Unlemm. BNC (freq. > 5)		DIFF.		Lemmatised BNC (top 10,000)		
	item-POSs	%	abs. %	rel. %	%	canon. forms	
A	11486	5.84	+0.75	+12.78	6.59	659	**A**
B	12393	6.30	-1.26	-20.06	5.04	504	**B**
C	18040	9.18	+1.36	+14.85	10.54	1054	**C**
D	11174	5.68	+0.58	+10.13	6.26	626	**D**
E	7476	3.80	+0.72	+18.85	4.52	452	**E**
F	8575	4.36	+0.63	+14.39	4.99	499	**F**
G	6905	3.51	-0.78	-22.28	2.73	273	**G**
H	7979	4.06	-0.75	-18.45	3.31	331	**H**
I	6498	3.31	+1.00	+30.38	4.31	431	**I**
J	2137	1.09	-0.36	-32.85	0.73	73	**J**
K	2918	1.48	-0.87	**-58.91**	0.61	61	**K**
L	7332	3.73	-0.30	-8.04	3.43	343	**L**
M	11681	5.94	-1.20	-20.23	4.74	474	**M**
N	5017	2.55	-0.57	-22.42	1.98	198	**N**
O	4573	2.33	+0.20	+8.75	2.53	253	**O**
P	14355	7.30	+0.77	+10.51	8.07	807	**P**
Q	915	0.47	-0.03	-5.47	0.44	44	**Q**
R	10843	5.52	+0.68	+12.40	6.20	620	**R**
S	22071	11.23	-0.02	-0.16	11.21	1121	**S**
T	10101	5.14	-0.08	-1.53	5.06	506	**T**
U	3654	1.86	-0.16	-8.54	1.70	170	**U**
V	2813	1.43	+0.24	+16.70	1.67	167	**V**
W	5944	3.02	-0.09	-3.10	2.93	293	**W**
X	263	0.13	-0.08	-62.63	0.05	5	**X**
Y	797	0.41	-0.12	-28.47	0.29	29	**Y**
Z	635	0.32	-0.25	-78.33	0.07	7	**Z**
	196575	99.99	***r* = 0.977**		100.00	10000	

Pretoria Afrikaans Corpus (PAfC)							
	Unlemm. PAfC (top freq.)		DIFF.		Lemmatised PAfC (top freq.)		
	raw types	%	abs. %	rel. %	%	canon. forms	
A	574	5.74	+0.26	+4.50	6.00	419	**A**
B	838	8.38	+0.22	+2.67	8.60	601	**B**
C	64	0.64	-0.50	-77.63	0.14	10	**C**
D	519	5.19	-0.24	-4.56	4.95	346	**D**
E	279	2.79	-0.10	-3.53	2.69	188	**E**
F	145	1.45	-0.16	-11.14	1.29	90	**F**
G	850	8.50	+1.21	+14.19	9.71	678	**G**
H	434	4.34	-0.33	-7.64	4.01	280	**H**
I	231	2.31	-0.05	-2.08	2.26	158	**I**
J	124	1.24	-0.47	**-37.65**	0.77	54	**J**
K	552	5.52	+0.11	+1.93	5.63	393	**K**
L	342	3.42	-0.41	-12.09	3.01	210	**L**
M	447	4.47	-0.86	-19.29	3.61	252	**M**
N	251	2.51	-0.16	-6.46	2.35	164	**N**
O	629	6.29	+1.25	+19.95	7.54	527	**O**
P	408	4.08	-0.40	-9.82	3.68	257	**P**
Q	—	—	—	—	—	—	**Q**
R	333	3.33	-0.44	-13.16	2.89	202	**R**
S	1022	10.22	+0.06	+0.58	10.28	718	**S**
T	434	4.34	-0.16	-3.68	4.18	292	**T**
U	179	1.79	+0.24	+13.57	2.03	142	**U**
V	901	9.01	+1.21	+13.45	10.22	714	**V**
W	399	3.99	-0.02	-0.61	3.97	277	**W**
X	3	0.03	-0.02	-52.28	0.01	1	**X**
Y	26	0.26	-0.13	-50.44	0.13	9	**Y**
Z	16	0.16	-0.12	-73.16	0.04	3	**Z**
	10000	100.00	***r* = 0.991**		99.99	6985	

Tables 4a & 4b: Breakdown of alphabetical categories in the BNC & PAfC

The data in Table 4a support the assumption that the distribution of the 'item-pos types *versus* canonical forms' is rather even across the different alphabetical categories.[1]

An analogous test was done for Afrikaans. Firstly, the top 10,000 raw types from the 4.5-million-word *Pretoria Afrikaans Corpus* (PAfC) were selected, and the breakdown was calculated for them. Secondly, this list of 10,000 raw types was manually lemmatised with a COBUILD-type of dictionary in mind, after which the new breakdown was calculated. The outcome of this test is summarised in Table 4b. The data clearly support the assumption that the distribution of the 'raw types *versus* canonical forms' is rather even across the different alphabetical categories.

The experiments (and corresponding statistics) discussed in this section indicate a neat correlation between lemmatised and unlemmatised corpora, thus enabling a direct comparison between corpus types and alphabetical stretches.

Corpora: Averaging Corpus Data and Dictionary Data

At this point we are in a position to compare the observed stable corpus pattern of suggested 'percentage allocation to alphabetical stretches' with the actual 'space allocation' in both D1-5 and WAT. A comparison between PAfC and D1-5 immediately reveals that this average represents a very sound breakdown indeed. This is shown in Table 5.

	PAfC	DIFF.		D1-5	
		abs.	rel.		
	%	%	%	%	
A	5.74	+0.49	+8.55	6.23	**A**
B	7.17	+1.11	+15.50	8.29	**B**
C	1.21	-0.90	**-74.29**	0.31	**C**
D	5.03	-0.30	-6.04	4.73	**D**
E	2.88	-0.74	-25.69	2.14	**E**
F	1.69	-0.38	-22.23	1.31	**F**
G	6.60	-0.67	-10.18	5.92	**G**
H	4.44	+0.02	+0.34	4.46	**H**
I	2.73	-0.29	-10.56	2.44	**I**
J	1.10	-0.35	-31.56	0.75	**J**
K	6.11	+2.09	+34.25	8.20	**K**
L	3.96	-0.13	-3.28	3.83	**L**
M	4.66	-0.47	-10.15	4.19	**M**
N	3.05	-0.82	-26.97	2.23	**N**
O	6.22	-0.46	-7.47	5.76	**O**
P	3.98	-0.17	-4.20	3.82	**P**
Q	0.00	+0.01	>	0.01	**Q**
R	3.90	-0.46	-11.83	3.44	**R**
S	10.85	+1.41	+13.03	12.26	**S**
T	4.80	-0.84	-17.39	3.97	**T**
U	1.87	-0.06	-3.11	1.81	**U**
V	7.24	+2.24	+31.00	9.48	**V**
W	3.94	+0.17	+4.27	4.10	**W**
X	0.12	-0.05	-40.96	0.07	**X**
Y	0.47	-0.28	-59.21	0.19	**Y**
Z	0.25	-0.19	-74.80	0.06	**Z**
	100.01	***r* = 0.976**		100.00	

Table 5: D1-5 *versus* PAfC

	PAfC & D1-5	CURR.	SUGG.	SUGG.	
	%	pp.	pp.	lemmas	
A	5.98	316.0	731.0	9982	**A**
B	7.73	278.3	944.2	12893	**B**
C	0.76	68.7	92.8	1267	**C**
D	4.88	411.7	595.7	8135	**D**
E	2.51	225.7	306.3	4182	**E**
F	1.50	154.6	183.4	2504	**F**
G	6.26	492.3	764.6	10441	**G**
H	4.45	486.7	543.3	7419	**H**
I	2.58	243.2	315.3	4306	**I**
J	0.92	110.3	112.9	1541	**J**
K	7.16	2479.0	874.2	11938	**K**
L	3.90	476.0	**476.0**	**6500**	**L**
M	4.42	619.4	540.5	7380	**M**
N	2.64	238.8	322.6	4405	**N**
O	5.99	650.9	731.5	9989	**O**
P	3.90		476.4	6505	**P**
Q	0.01		0.9	12	**Q**
R	3.67		447.9	6116	**R**
S	11.56		1411.7	19278	**S**
T	4.39		535.9	7317	**T**
U	1.84		224.7	3069	**U**
V	8.36		1020.8	13940	**V**
W	4.02		491.0	6705	**W**
X	0.09		11.6	158	**X**
Y	0.33		40.3	551	**Y**
Z	0.16		19.5	266	**Z**
	100.01	7251.6	12215.0	166799	

Table 7: Pages and lemma-signs in WAT

From Table 5 one can see that there is a rather good correlation between the average of the five dictionaries and the corpus suggestion. The correlation coefficient *r* is as high as 0.976.

Space allocation in WAT will now be compared to that suggested by D1-5 and PAfC respectively, and also against the average of PAfC & D1-5. Block I in Table 6 compares WAT with D1-5, Block II compares WAT with PAfC, while Block III compares WAT with the average of the dictionary data (Block I) and the corpus data (Block II). Once again, in order to enable comparison with WAT, the data have been recalculated to cover only the categories **A** to **O**.

Block I					Block II						Block III				
D1-5	DIFFER.		WAT		P Af C		DIFFER.		WAT		PAfC & D1-5	DIFFER.		WAT	
	abs.	rel.			raw		abs.	rel.				abs.	rel.		
%	%	%	%		types	%	%	%	%		%	%	%	%	
10.18	-5.82	**-57.18**	4.36	**A**	10085	9.17	-4.81	**-52.48**	4.36	**A**	9.67	-5.32	**-54.95**	4.36	**A**
13.52	-9.68	**-71.62**	3.84	**B**	12606	11.46	-7.63	**-66.52**	3.84	**B**	12.49	-8.65	**-69.28**	3.84	**B**
0.51	+0.44	+86.45	0.95	**C**	2123	1.93	-0.98	-50.93	0.95	**C**	1.22	-0.27	-22.30	0.95	**C**
7.76	-2.08	-26.86	5.68	**D**	8837	8.04	-2.36	-29.35	5.68	**D**	7.90	-2.22	-28.12	5.68	**D**
3.51	-0.40	-11.40	3.11	**E**	5055	4.60	-1.48	-32.29	3.11	**E**	4.05	-0.94	-23.24	3.11	**E**
2.16	-0.03	-1.25	2.13	**F**	2968	2.70	-0.57	-21.01	2.13	**F**	2.43	-0.30	-12.23	2.13	**F**
9.75	-2.97	-30.40	6.79	**G**	11590	10.54	-3.75	-35.58	6.79	**G**	10.15	-3.36	-33.09	6.79	**G**
7.33	-0.62	-8.49	6.71	**H**	7803	7.10	-0.38	-5.41	6.71	**H**	7.21	-0.50	-6.97	6.71	**H**
4.01	-0.66	-16.45	3.35	**I**	4789	4.35	-1.00	-22.99	3.35	**I**	4.18	-0.83	-19.85	3.35	**I**
1.24	+0.28	+22.70	1.52	**J**	1928	1.75	-0.23	-13.24	1.52	**J**	1.50	+0.02	+1.65	1.52	**J**
13.54	+20.64	**+152.42**	34.19	**K**	10738	9.76	+24.42	**+250.10**	34.19	**K**	11.65	+22.53	**+193.35**	34.19	**K**
6.33	+0.23	+3.66	6.56	**L**	6962	6.33	+0.23	+3.69	6.56	**L**	6.33	+0.23	+3.67	6.56	**L**
6.92	+1.62	+23.40	8.54	**M**	8191	7.45	+1.09	+14.68	8.54	**M**	7.18	+1.36	+18.88	8.54	**M**
3.69	-0.40	-10.73	3.29	**N**	5364	4.88	-1.58	-32.49	3.29	**N**	4.28	-0.99	-23.12	3.29	**N**
9.53	-0.56	-5.84	8.98	**O**	10932	9.94	-0.96	-9.71	8.98	**O**	9.74	-0.76	-7.81	8.98	**O**
99.98	***r* = 0.608**		100.00		109971	100.00	***r* = 0.463**		100.00		99.98	***r* = 0.550**		100.00	

Table 6: WAT *versus* D1-5 and PAfC (up to **O**)

The data in Block II of Table 6 amply confirm that, when evaluated against PAfC, the WAT breakdown is way out compared to the D1-5 breakdown (shown in Table 5). Indeed, the correlation coefficient *r* between PAfC and WAT is as low as 0.463, compared to a correlation of 0.976 between PAfC and D1-5. The latter correlation means that the *average* of the space allocation to alphabetical categories suggested by PAfC & D1-5 may successfully be used as a comparison instrument (Block III). Based on the breakdown suggested by the average of PAfC & D1-5, a calculation can now be proposed of what the breakdown for **A** to **O** in WAT *should have been* and what the breakdown in future volumes *will have to be*. The latter can be of use to the compilers since the final few categories in the alphabet (**P** to **Z**, roughly 2/5 of the dictionary) still have to be compiled, whilst the former can be kept in mind for future revisions of the dictionary.

Since **L** represents, in terms of Botha, the start of the 'new' WAT following "a drastic revision of the editorial process" [1994: 423], the calculations could be done to reflect the ideal situation taking **L** as the basis. As **L**, with 476.0 'new' WAT pages, should be allocated 3.90% of the dictionary space, every 1% corresponds with 122.1 pages. From this, one derives the data shown in Column 4 of Table 7. Furthermore, since Feinauer [1996: 234] states that the number of lemma signs treated in **L** is roughly 6500, the calculations can be extended to include the proposed number of lemma signs for the *entire* dictionary, as shown in Column 5 of Table 7. Dividing Column 5 by Column 4 further indicates that every page in the dictionary should contain an average of 13.66 lemma signs. Finally, as the WAT layout of the letter **L** has three columns of each 80 lines per page, every lemma sign in the

dictionary should consist of an average of 17.58 column-lines. The column-line being the 'basic unit of length' in a dictionary [Landau 2001: 375], we see that we have arrived at an extremely powerful guidance tool indeed. For instance, the following prediction can be used as a guidance by the WAT compilers of the next alphabetical stretch: c. 6505 lemma signs should be singled out for the category **P**, c. 476.4 pages will be required to treat them, and during the compilation c. 17.58 column-lines should be devoted to each lemma sign.

The value of such a guidance tool cannot be emphasized enough. The current Editor-in-Chief of the WAT was recently quoted in *Beeld* (21 December 2001: 19) stating that roughly a quarter of all Afrikaans words start with the letter **S**. Tables 5 and 7, however, suggest that the allocation should only be slightly more than *one tenth* of the total. Such information is *crucial* in the planning and management of especially a major dictionary project (compare Moerdijk [1998: 330–334]).

Under-treatment and/or Over-treatment

The sketched strategy will successfully highlight deviations in the following cases:

- *under-treatment* in terms of the number of lemma signs and/or article length;
- *over-treatment* in terms of the number of lemma signs and/or article length.

However, comparison with other dictionaries and/or corpora as measurement instruments will not be able to detect the following inconsistencies:

- *under-treatment* in terms of the number of lemma signs occurring simultaneously with *over-treatment* in terms of article length within the same alphabetical category;
- *over-treatment* in terms of the number of lemma signs occurring simultaneously with *under-treatment* in terms of article length within the same alphabetical category.

In the latter two cases the *exact* number of *lemma signs* in the dictionary and their distribution across the different alphabetical categories must be compared directly with the statistics obtained from the corpus (and other dictionaries).

As an illustration, an analysis of **K** in WAT reveals over-treatment in terms of article length *and* the number of lemma signs entered, so it was obviously detected with the proposed measurement instrument. This 'over-treatment' of **K** compared to the previous letters in WAT was also noted by leading linguists such as Combrink [1979], Gouws [1985] and Swanepoel [1989]. Major points of criticism generally expressed against WAT revolve around *excessive* encyclopaedic treatment of lemma signs, *overemphasis* on and *highly* complicated defining style of technical terms, *overuse* of colour-plates, *unnecessary* treatment of transparent compounds, etc.

Conclusion

A tool to guide alphabetical breakdown in dictionaries may successfully be based on frequency data derived from an unlemmatised corpus, supplemented by counts from existing dictionaries. Further:

- this devised measurement instrument can point out serious alphabetical-stretch imbalances in existing dictionaries, imbalances which can be addressed in revised editions;
- the devised measurement instrument is especially useful for large (and thus long-term) multi-volume dictionary projects in progress, where imbalances cannot only be addressed in revisions, but more importantly, reliable strategies can be set up to steer the future compilation – at this point the 'measurement' instrument thus becomes a 'prediction' instrument;
- the devised measurement and prediction instrument enables to put forward guidelines for space allocation (i.e. number of pages) and for the number of lemma signs per alphabetical stretch;
- the devised measurement and prediction instrument makes it possible to suggest an average article length (expressed in number of articles per page, or even in number of column-lines per article).

Endnotes

1. Instead of a manual lemmatisation, software can be called in to derive lemmatised lists from unlemmatised ones. Two attempts for the BNC can be found on the Internet, both with rather high cut-off frequencies however. Kilgarriff [1996] presents a "lemmatised frequency list for the 6,318 [lemmas] with more than 800 occurrences in the whole 100M-word BNC"; and Leech, Rayson & Wilson [2001] uploaded approximate figures for a lemmatised frequency list of the whole BNC "down to a minimum of 10 occurrences of a lemma per million". If one extracts the length of the alphabetical stretches from those lemmatised lists (respectively 6,318 and 6,531 items long), and compares them to the stretches in the unlemmatised BNC (freq. > 5), just as was done in Table 4a, then one obtains correlation coefficients *r* of 0.974 and 0.972 respectively. These compare rather well with the value of 0.977 found after a manual lemmatisation – even though the particular approach to the lemmatisation itself was slightly different in each case.

References

Beeld, 21 December 2001, p. 19. 'WAT' so lank?

[Botha 1994] Botha, W., 1994. An About-Turn Halfway Through the Completion of a Multi Volume Overall-Descriptive Dictionary – Gallantry of Folly?, in: W. Martin et al. (eds.) *Euralex 1994 Proceedings, Papers submitted to the 6th EURALEX International Congress on Lexicography in Amsterdam, The Netherlands*, pp. 419–425, Vrije Universiteit, Amsterdam.

[Combrink 1979] Combrink, J.G.H., 1979. Die sesde deel van die W.A.T., in: *Standpunte, 32 (2)*, pp. 49–64.

[Crystal 1986] Crystal, D., 1986. The ideal dictionary, lexicographer and user, in: R.F. Ilson (ed.) *Lexicography: An emerging international profession* (Fulbright Papers 1), pp. 72–81, Manchester University Press, Manchester.

[Feinauer 1996] Feinauer, I., 1996. Die negende deel van die Woordeboek van die Afrikaanse Taal, in: *Lexikos 6* (AFRILEX-reeks/series 6: 1996), pp. 233–271.

[Gouws 1985] Gouws, R.H., 1985. Die sewende deel van die *Woordeboek van die Afrikaanse Taal*, in: *Standpunte, 38 (1)*, pp. 13–25.

[Grobbelaar 1987] Grobbelaar, P. (ed.), 1987. *Reader's Digest English – Afrikaans Dictionary*, Reader's Digest Association South Africa, Cape Town.

[Heritage 2000[4]] *The American Heritage® Dictionary of the English Language, Fourth Edition*, 2000[4]. Houghton Mifflin, Boston, <http://www.bartleby.com/61/>.

[Juta 1983[6]] Juta, 1983[6]. *Juta se sakwoordeboek / Juta's Pocket Dictionary, Afrikaans – Engels, English – Afrikaans*, Juta en Kie, Kenwyn.

[Kilgarriff 1996] Kilgarriff, A., 1996. *BNC database and word frequency lists*, <ftp://ftp.itri.bton.ac.uk/bnc/>.

[Kriel 1976[4]] Kriel, T.J., 1976[4]. *The New English – Northern Sotho Dictionary, English – Northern Sotho, Northern Sotho – English*, Educum Publishers, Johannesburg.

[Kriel 1983[3]] Kriel, T.J., 1983[3]. *Pukuntšu woordeboek, Noord-Sotho – Afrikaans, Afrikaans – Noord-Sotho*, J.L. van Schaik, Pretoria.

[Kriel et al. 1997[4]] Kriel, T.J., D.J. Prinsloo and B.P. Sathekge, 1997[4]. *Popular Northern Sotho Dictionary, Northern Sotho – English, English – Northern Sotho*, Pharos, Cape Town.

[Kritzinger & Eksteen 1984[4]] Kritzinger, M.S.B. and L.C. Eksteen, 1984[4]. *Kompakte Afrikaanse Woordeboek*, J.L. van Schaik, Pretoria.

[Kromhout 1997[13]] Kromhout, J., 1997[13]. *Klein woordeboek / Little Dictionary, Afrikaans – Engels, English – Afrikaans*, Pharos, Cape Town.

[Landau 2001] Landau, S.I., 2001. *Dictionaries: The Art and Craft of Lexicography (2nd edition)*, Cambridge University Press, Cambridge.

[Leech, Rayson & Wilson 2001] Leech, G., P. Rayson and A. Wilson, 2001. Companion Website for: *Word Frequencies in Written and Spoken English: based on the British National Corpus*, <http://www.comp.lancs.ac.uk/ucrel/bncfreq/>.

[Matumo 1993[4]] Matumo, Z.I., 1993[4]. *Setswana – English – Setswana Dictionary*, Macmillan Botswana Publishing Company, Gaborone.

[Moerdijk 1998] Moerdijk, A., 1998. Het WNT in cijfers, in: F. Heyvaert et al. (eds.) *Het grootste woordenboek ter wereld. Een kijkje achter de kolommen van het Woordenboek der Nederlandsche Taal (WNT)*, pp. 319–338, Sdu Uitgevers, Den Haag.

[Newbury 1999] *The Newbury House Online Dictionary of American English*, 1999. Monroe Allen Publishers, Inc., <http://nhd.heinle.com/>.

[Odendal & Gouws 2000[4]] Odendal, F.F. and R.H. Gouws, 2000[4]. *HAT. Verklarende Handwoordeboek van die Afrikaanse Taal*, Perskor, Midrand.

[Sinclair 2001[3]] Sinclair, J.M. (ed.), 2001[3]. *Collins COBUILD English Dictionary for Advanced Learners*, HarperCollins Publishers, London.

[Snyman 1990] Snyman, J.W. (ed.), 1990. *Dikišinare ya Setswana – English – Afrikaans Dictionary/Woordeboek*, Via Afrika Limited, Pretoria.

[Summers 1995[3]] Summers, D. (director), 1995[3]. *Longman Dictionary of Contemporary English, Third Edition*. Longman Dictionaries, Harlow.

[Swanepoel 1989] Swanepoel, P.H., 1989. Botsing tussen leksikografiese beginsel en leksikografiese praktyk; enkele gedagtes oor die terminologiebeleid van die *WAT*, in: *Suid-Afr. Tydskrif vir Taalkunde, 7 (1)*, pp. 6–19.

[WAT 1926 –] *Woordeboek van die Afrikaanse Taal*, 1926 – . Bureau of the WAT, Stellenbosch, <http://www.sun.ac.za/wat/index.htm>.

[Webster 1913] *Webster's Revised Unabridged Dictionary* (edited by Noah Porter), 1913. G. & C. Merriam Co., <http://humanities.uchicago.edu/forms_unrest/webster.form.html>.

[Ziervogel & Mokgokong 1988[4]] Ziervogel, D. and P.C.M. Mokgokong, 1988[4]. *Klein Noord-Sotho woordeboek, N.-Sotho – Afrikaans – English, Afrikaans – N.-Sotho, English – N.-Sotho*, J.L. van Schaik, Pretoria.