*Nordic Journal of African Studies 11(2): 249-265 (2002)*

# Towards an 11 x 11 Array for the Degree of Conjunctivism/Disjunctivism of the South African Languages

D. J. PRINSLOO
*University of Pretoria, South Africa*
& GILLES-MAURICE DE SCHRYVER
*Ghent University, Belgium & University of Pretoria, South Africa*

## ABSTRACT

In this article a measurement instrument for the degree of conjunctivism / disjunctivism of the South African languages is presented. Following a discussion on conjunctivism versus disjunctivism, both absolute and relative approaches towards this measurement instrument are experimented with. Three potential absolute instruments are derived: one based on word length, one on sentence length, and one on the standardised type/token ratio. All of them pose problems. The search for a relative instrument is more successful. Although large sets of parallel texts would provide the ideal data, two-by-two parallel corpora offer a good substitute. The final 11 x 11 array is also compiled in this way. Applications of the 11 x 11 array in several fundamental and applied linguistic fields are reviewed. The fields include translation, academic writing, corpus linguistics, and theoretical reflections about spellcheckers and multi-dimension dictionary Rulers. A complete Bantu Array could be the ultimate goal.

*Keywords: conjunctivism, disjunctivism, measurement instrument, parallel corpora, 11 x 11 array, isiZulu, isiXhosa, siSwati, isiNdebele, Sepedi, Sesotho, Setswana, Tshivenda, Xitsonga, Afrikaans, English*

## INTRODUCTION

Since 1994, nine Bantu languages have – in addition to English and Afrikaans – become official languages of South Africa. The first attempts to reduce these African languages to writing, however, date from the early 19th century. Mainly as a result of distinct phonological processes, and thus for practical reasons (see Louwrens 1991: 1-12), different writing systems developed for these nine Bantu languages. The languages belonging to the Nguni group (isiZulu, isiXhosa, siSwati and isiNdebele) are said to be written conjunctively, while those from the Sotho group (Sepedi, Sesotho, Setswana), as well as Tshivenda and Xitsonga, adhere to a disjunctive writing system. Compare the following simple example in which the grammatical structure of a Sepedi and isiZulu phrase is exactly the same but is written as four words in Sepedi and as only one word in isiZulu:

| | | | | | |
|---|---|---|---|---|---|
| **Sepedi** | *ke a mo rata* | *ke* | *a* | *mo* | *rata* |
| | 'I love him/her' | I | [pres.] | him/her | love |
| **isiZulu** | *ngiyamthanda* | *ngi-* | *-ya-* | *-m-* | *-thanda* |
| | 'I love him/her' | I | [pres.] | him/her | love |

Despite the fact that especially disjunctivism poses problems for the notion *linguistic word*, the fact of the matter is that these opposite writing systems are now fairly well entrenched.

Nonetheless, over the past few years the need has arisen, and this in several fundamental and applied linguistic fields (as will be seen in §4), for the availability of an instrument with which the degree of conjunctivism / disjunctivism of the *orthographic word* could be measured. Formulated differently, departing from the current spelling for each of the nine Bantu languages, the requirement is to arrive at a scientifically sound tool which would reveal the 'density' of orthographic words in each particular language.

## 1. Conjunctivism versus Disjunctivism

It is important to keep in mind that a conjunctive versus a disjunctive writing system is in principle only an orthographical convention. Van Wyk (1995: 84) rightfully points out that one tradition is not to be regarded as superior to another. For the Bantu languages spoken in South Africa the unfortunate situation arose whereby one system is considered by its followers as more scientific than the other. In particular, those favouring conjunctivism tend to view it as superior to such an extent that even a traditionally disjunctively-written language such as Sepedi is sometimes treated in the same way as the conjunctively-written languages in lemmatisation for dictionaries (such as in Ziervogel & Mokgokong's (1975) *Comprehensive Northern Sotho Dictionary*). The linguistic issues for the Nguni languages are succinctly summarised by Louwrens as follows:

> The reason why Sotho and Nguni writers decided on different methods of word division in their works is not so much a scientific one as a practical one, since it mainly concerns the fundamental differences that exist between the *phonological systems* of the Sotho and Nguni language groups. Phonological processes such as vowel elision, vowel coalescence and consonantalisation which are very much less productive in the Sotho languages than is the case in the Nguni languages, render the disjunctive method of word division a highly impractical proposition in Nguni. (Louwrens 1991: 2)

For the Sotho languages he offers two reasons in favour of a disjunctive way of writing, and supports this with examples from Sepedi:

> In the Sotho languages, on the other hand, disjunctivism presents very few problems, since most formatives in these languages constitute

> syllables and can therefore easily be written disjunctively. ... A further reason why the conjunctive method of writing was not as acceptable to the Sotho languages as the disjunctive one, is because of their lack of semi-vowels between syllables which consist of a vowel only. If the conjunctive method of writing had been followed, it would have made the reading as well as the pronunciation of words such as the following a cumbersome task in the Sotho languages:
>
> | | |
> |---|---|
> | *gaaaapee* (ga-a-a-apee (mae)) | She doesn't boil them (the eggs) |
> | *oaoômiša* (o-a-o-ômiša (morôgô)) | She causes it (the morôgô) to become dry |
>
> (Louwrens 1991: 2-3)

Linguistic words such as *gaaaapee* and *oaoomiša* are of course typical examples of creations by grammarians who base their arguments on introspection. Although neither *ga a a apee* or *o a o omiša* occur in the 5.8-million-word *Pretoria Sepedi Corpus*, no one will dispute the sound arguments quoted above.

If conjunctivism versus disjunctivism occurs on orthographic-word level, it clearly also affects the orthographic-word density on sentence level. In order to give a series of real-life examples, a randomly selected excerpt from a text on South Africa's *National Qualifications Framework* (NQF)[1] is shown below, and this with translations in all other official South African languages. The eleven excerpts, nine for the Bantu languages, plus those for English and Afrikaans, are arranged according to increasing number of orthographic words used:

| | |
|---|---|
| **Siswati** *(26 orthographic words)* | Nga-1994 umphakatsi wemhlabawonkhe waba ngubofakazi bekutalwa kwentsandvo yelinyenti waphindze wemukela iNingizimu Afrika lensha njengelilunga lelisha lelive lamhlabawonkhe. Ngekwemukela loko kwetsenjwa, lelive latsatsa tincabhayi letihambisana naleso sikhundla. |
| **isiNdebele** *(28 orthographic words)* | Ngo-1994 umphakathi wephasi loke wabona ukuzalwa kombuso omutjha wentando yesitjhaba wamukela iSewula Afrika njengelungu elitjha lomphakathi wesphasi loke. Ekwamukeleni irhahla lelo, inarha lena yajamelana nokutjhijilelwa okutholakala ebujamweni obunjalo. |
| **isiXhosa** *(30 orthographic words)* | Ngonyaka ka-1994 uluntu lwehlabathi liphela lakubona ukuzalwa kwentando yesininzi kwaye lwamkela uMzantsi Afrika omtsha njengelona lungu litsha kubuhlanti behlabathi liphela. Ekulamkeleni elo wonga, eli lizwe layithabatha imingeni enxulumene naloo ndawo. |
| **isiZulu** *(33 orthographic words)* | Nonyaka ka-1994 umhlaba wonke wabona ukuzalwa kabusha kweNingizimu Afrika yentandoyeningi futhi wamukela iNingizimu Afrika entsha njengelungu lomphakathi womhlaba wonke jikelele. Ekwamukeleni lokho kuhlonishwa leli zwe laseNingizimu Afrika lona lathatha izinselelo ezihambisana naleso sikhundla. |
| **Afrikaans** *(38 orthographic words)* | In 1994 het die internasionale gemeenskap die geboorte van 'n nuwe demokrasie aanskou en die nuwe Suid-Afrika as die jongste lid in sy wêrelddorp verwelkom. Met die aanvaarding van hierdie eer het Suid-Afrika ook die gepaardgaande uitdagings aanvaar. |

[1] Source = http://www.saqa.org.za/nqf/overview.html (*spelling errors were corrected*).

| | |
|---|---|
| **English** *(41 orthographic words)* | In 1994 the international community witnessed the birth of a new democracy and welcomed the new South Africa as the most recent member of its global village. In accepting that honour, this country took on the associated challenges of that position. |
| **Xitsonga** *(42 orthographic words)* | Hi 1994, vaaki va misava hinkwayo va vonile ku tswariwa ka demokhirasi leyintshwa ni ku amukela Afrika-Dzonga leyintshwa tani hi xirho lexintshwa emutini wa matiko hinkwawo. Hi ku amukela ku fundzhiwa loku, tiko leri ri tekile mintlhontlho leyi fambelanaka ni xiyimo lexi. |
| **Setswana** *(42 orthographic words)* | Ka 1994 setšhaba sa boditšhabatšhaba se ne sa bogela botsalo ba demokerasi e ntshwa le go amogela Aforika Borwa yo montshwa jaaka tokololo e ntshwa ya motse wa lefatshe. Mo go amogeleng tlotlo eo, naga e e tsere kgwetlho ya maemo ao. |
| **Tshivenda** *(44 orthographic words)* | Nga 1994, tshakha dzothe dzo vhona u bebwa ha demokirasi ntswa khathihi na u tanganedza Afurika Tshipembe liswa sa murando muswa wa dzhango lothe. Ili shango lo tanganedza iyo thompho nga u vha na vhudifhinduleli kha khaedu dzothe dza zwi tshimbilelanaho na vhuimo honoho. |
| **Sepedi** *(50 orthographic words)* | Ka ngwaga wa 1994 setšhaba sa boditšhabatšhaba se bone pelego ya temokrasi ye mpsha gape sa amogela Afrika Borwa ye Mpsha bjalo ka leloko le lefsa kudu motseng wa sona wa lefase ka bophara. Ge go amogelwa tlotlo yeo, naga ye e tšere dihlotlo tšeo di amanago le maemo ao. |
| **Sesotho** *(61 orthographic words)* | Ka selemo sa sekete, makgolo a robong, mashome a robong a metso e mene setjhaba se mose ho mawatle se ile sa thabela ho bona ho ba le demokrasi mona Afrika Borwa mme se amohela setho se setjha e leng Afrika Borwa. Bakeng sa ho amohela tlotla eo, naha ena ya rona e tadimane le ditlhaselo bakeng sa ho ntshetsa pele. |

The rationale for including English and Afrikaans will become apparent further below, yet it should already be clear that, in order to convey (roughly) the same content, there is indeed a great divide between the number of orthographic words used in the Nguni languages (Siswati, isiNdebele, isiXhosa and isiZulu), as opposed to the number of orthographic words used in Xitsonga, Tshivenda and the Sotho languages (Setswana, Sepedi and Sesotho).[2] Number-of-word wise, English and Afrikaans happen to be in-between the two traditions. Conversely, it should also be clear that the length (expressed in number of letters used) of disjunctively-written orthographic words is obviously shorter than that of the conjunctively-written ones.

[2] Note, however, that in the Sesotho text the year '1994' is written in words which results in the use of eleven orthographic words.

# 2. Towards an Absolute Approach for Determining the Degree of Conjunctivism / Disjunctivism

## 2.1 Simulating a Conjunctive Way of Writing for Disjunctively-Written Languages and Vice Versa

A first way in which the observed conjunctivism versus disjunctivism could be quantified, would be to consider each language in isolation, and thus to pursue a language-internal or *absolute* measurement instrument. It should be apparent from the above that any of the Bantu languages can actually be written as conjunctively (or as disjunctively for that matter) as any other. Formulated differently, one can simulate a conjunctive way of writing for disjunctively-written languages and vice versa. For example, the above Sepedi NQF excerpt can be written with an isiZulu-like spelling as follows (where underscores connect the formatives which now form single orthographic words):

| | |
|---|---|
| **Sepedi** *(31 'isiZulu-like' orthographic words)* | Ka_ngwaga wa_1994 setšhaba sa_boditšhabatšhaba se_bone pelego ya_temokrasi ye_mpsha gape sa_amogela Afrika Borwa ye_Mpsha bjalo_ka_leloko le_lefsa kudu motseng wa_sona wa_lefase ka_bophara. Ge go_amogelwa tlotlo yeo, naga ye e_tšere dihlotlo tšeo_di_amanago le_ao maemo. |

When Sepedi is written as if it were isiZulu, Sepedi needs roughly the same number of orthographic words as isiZulu (31 versus 33 words respectively). Likewise, the isiZulu excerpt can be written with a Sepedi-like spelling as follows (where vertical bars separate the formatives which are now separate orthographic words):

| | |
|---|---|
| **isiZulu** *(54 'Sepedi-like' orthographic words)* | No\|\|nyaka ka-\|\|1994 umhlaba wonke wa\|\|bona uku\|\|zalwa ka\|\|busha kwe\|\|Ningizimu Afrika ye\|\|ntando\|\|ye\|\|ningi futhi wa\|\|(a)mukela iNingizimu Afrika e\|\|ntsha njenge\|\|lungu lo\|\|mphakathi wo\|\|mhlaba wonke jikelele. Ekwamukeleni lokho ku\|\|hlonishwa leli zwe la\|\|se\|\|Ningizimu Afrika lona la\|\|thatha izinselelo ezi\|\|hambisana na\|\|leso si\|\|khundla. |

Thus, when isiZulu is written as if it were Sepedi, roughly the same number of orthographic words as Sepedi are needed (54 versus 50 words respectively).

## 2.2 Three Possible Absolute Rulers

The previous simulations, although evidently rather crude (see §3.1 below), are important, as they also mean that an isiZulu-like orthography can be written even *more* conjunctively, and, conversely, that a Sepedi-like orthography can be written even *more* disjunctively. If one would engage in the latter, one would increasingly move away from linguistic words towards morphemes and finally

to single letters, while increasing conjunctivism would in the extreme case result in one word per phrase. One could thus decide on a certain point on this continuum, and then compare the actual density of the current orthography for each language with that chosen point. With this approach one would arrive at an absolute (as language-internally derived) instrument for the measurement of the degree of conjunctivism / disjunctivism. This method could then be repeated for each of the Bantu languages anew. The two extreme cases, viz. (i) comparing each orthographic word with the single composing letters (extreme disjunctivism), and (ii) comparing each orthographic word with the single phrase of which it forms a part (extreme conjunctivism), are rather easy to compute.

Extreme disjunctivism simply implies that one needs to calculate the overall average orthographic-word length. In the 5.8-million-word *Pretoria Sepedi Corpus*, the overall average orthographic-word length is 3.88; in the 5.0-million-word *Pretoria Zulu Corpus* it is 7.18. The actual distribution in % of the average length of the orthographic words in Sepedi and isiZulu is shown in Figures 1 and 2.

**Figure 1.** Distribution in % of the average length of orthographic words in Sepedi (overall average = 3.88).

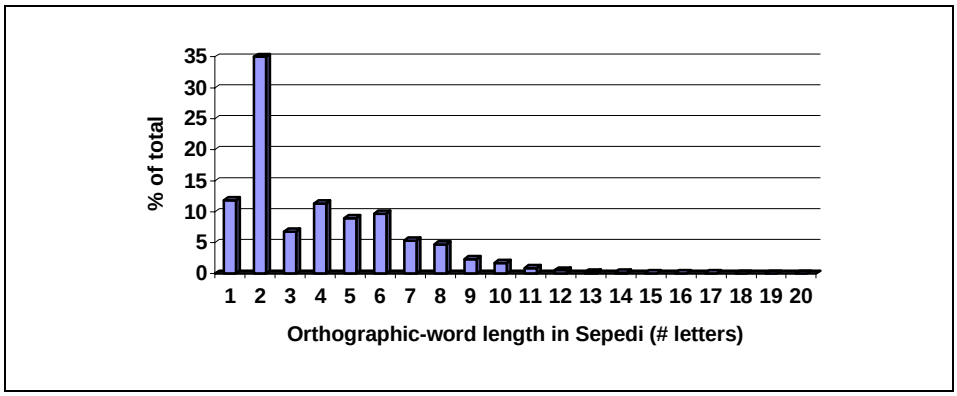
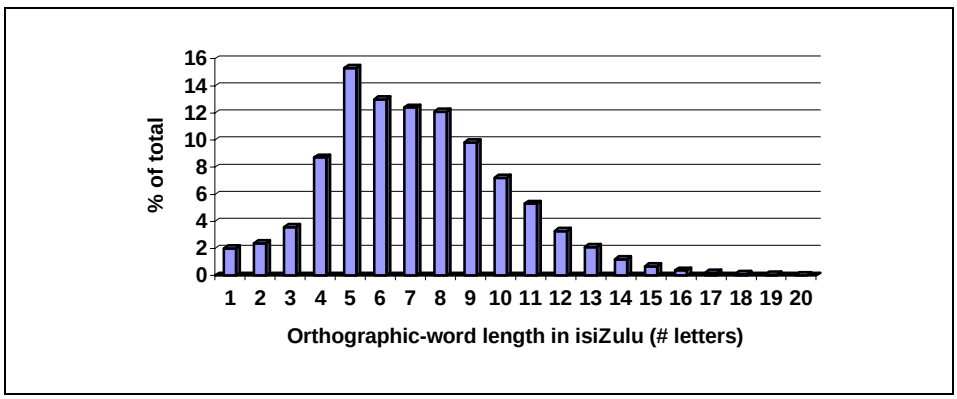



**Figure 2.** Distribution in % of the average length of orthographic words in isiZulu (overall average = 7.18).




From Figure 1 one can see that a staggering 35% of all running words in a Sepedi corpus are just two letters long. The effect of the CV structure of Bantu

languages is also apparent in a disjunctively-written language like Sepedi, as Figure 1 shows more 4-letter words than 3-letter words, and more 6-letter words than 5-letter words. The pattern seen in Figure 2 for isiZulu is more Gaussian, in that the spread of the average orthographic-word length has the characteristics of a normal statistical distribution (the so-called Bell Curve). As no languages have words of zero letters long, one can thus use a positive scale for which the following is roughly valid: the higher the overall average orthographic-word length, the higher the degree of conjunctivism. This is not entirely true, however, as the same sound might be spelled with a different number of letters in various languages, e.g. *š* versus *sh*, or *c* versus *tsh*, etc. An absolute measurement instrument for the degree of conjunctivism / disjunctivism can thus only be *approximated* with a ruler based on overall average orthographic-word lengths.

Extreme conjunctivism, on the other hand, actually implies that one needs to calculate the average number of orthographic words per phrase. Corpus-query software cannot normally compute this, but it can do so for the sentence level. In the above-mentioned corpora, the average number of orthographic words per sentence is 25.99 for Sepedi, and 12.15 for isiZulu. Again, a positive scale can be used for which the following is roughly valid: the lower the average number of orthographic words per sentence, the higher the degree of conjunctivism. Punctuation is, however, very differently used across languages, making an absolute ruler for the degree of conjunctivism / disjunctivism based on the average number of orthographic words per sentence *rather tenuous*.

A third absolute ruler that could be proposed is based on the *standardised type/token ratio*. 'Tokens' are all the *running* orthographic words in a corpus, 'types' all the *different* orthographic words. For example, the 5.8-million-word Sepedi corpus has 5,764,861 tokens, but only 148,714 types, while the 5.0-million-word isiZulu corpus has 5,001,456 tokens, and as many as 671,859 types. A standardised type/token ratio is computed every *n* tokens. We set *n* at 1,000. In the words of Scott this means that:

> … the ratio is calculated for the first 1,000 running words, then calculated afresh for the next 1,000, and so on to the end of your text or corpus. A running average is computed, which means that you get an average type/token ratio based on consecutive 1,000-word chunks of text. … The number shown is a percentage of new types for every n tokens. That way you can compare type/token ratios across texts of differing lengths. (Scott 1999)

The assumption in using this ratio as a measurement instrument is that the percentage of new types (or thus of new orthographic words) for every *n* (1,000 in our case) tokens will be higher for a conjunctively-written language than for a disjunctive one. This is confirmed, the standardised type/token ratio is 34.12 for Sepedi, and 69.76 for isiZulu. Just as the two previous absolute rulers, also the current approach is *not without question marks*, as a standardised type/token ratio is very much dependent on the type of data present in a corpus.

In Table 1, the data for the three possible rulers have been calculated for all eleven South African languages.[3]

**Table 1.** Three possible absolute rulers as a measurement instrument for the degree of conjunctivism / disjunctivism of the South African languages.

| | | | Absolute ruler 1 | Absolute ruler 2 | Absolute ruler 3 |
|---|---|---|---|---|---|
| ***Language*** | ***Tokens*** | ***Types*** | **Average word length (letters/word)** | **Average sentence length (words/sentence)** | **Standardised type/token ratio (*n* = 1,000)** |
| isiZulu | 5,001,456 | 671,859 | 7.18 | 12.15 | 69.76 |
| siSwati | 313,576 | 59,422 | 7.15 | 14.41 | 67.85 |
| isiNdebele | 1,033,965 | 193,996 | 7.08 | 9.67 | 69.06 |
| isiXhosa | 2,994,006 | 127,507 | 5.88 | 18.10 | 66.46 |
| Afrikaans | 4,817,239 | 284,411 | 4.56 | 19.73 | 43.97 |
| English | 12,545,938 | 118,193 | 4.35 | 16.72 | 41.83 |
| Xitsonga | 3,513,950 | 109,613 | 4.29 | 23.80 | 35.82 |
| Tshivenda | 2,462,243 | 102,386 | 4.07 | 28.06 | 38.15 |
| Setswana | 3,705,417 | 123,896 | 3.89 | 19.78 | 35.31 |
| Sepedi | 5,764,861 | 148,714 | 3.88 | 25.99 | 34.12 |
| Sesotho | 3,159,568 | 97,375 | 3.88 | 26.76 | 35.75 |

The rows in Table 1 have been sorted according to the first ruler. There is a rather good correspondence between the ruler based on the overall average length of orthographic words, and the ruler based on the standardised type/token ratio. Utilising the average number of orthographic words per sentence, however, seems less successful as a tool to differentiate clearly between conjunctivism and disjunctivism.

## 3. Towards a Relative Approach for Determining the Degree of Conjunctivism / Disjunctivism

As the three rulers introduced above do not seem all that satisfactory, a measurement instrument based on very different principles will be pursued in this section. Indeed, instead of a language-internal tool, an inter-language tool will be designed, so that the degree of conjunctivism / disjunctivism can be expressed in *relative* rather than in absolute terms. As the suggested approach is based on a large set of parallel texts, some considerations regarding the latter will precede the presentation of the measurement instrument itself.

[3] These data are derived from various corpora available in the department of African languages at the University of Pretoria. All of these corpora were compiled by D.J. Prinsloo and/or G-M de Schryver, except for the English one which was brought together by R. Gauton.

## 3.1 PARALLEL TEXTS: IDEAL TEXTBOOK EXAMPLES VERSUS REAL-WORLD OCCURRENCES

In their comparisons of conjunctively versus disjunctively written Bantu languages, linguists tend to cite ideal examples where there is complete similarity / correspondence between the examples from different languages. Compare the following sections presented in boldface for Sepedi and isiZulu.[4]

**English** Membership of the ANC is open to all South Africans above the age of 18 years, ...

**Sepedi** **Boleloko bja ANC bo buletšwe** batho ka moka ba Afrika Borwa, ba mengwaga ya go feta ye 18, ...

**isiZulu** **Ubulunga be-ANC buvuleleke** kuzo zonke izakhamizi zaseNingizimu Afrika ezineminyaka engu 18 nengaphezulu, ...

The Sepedi possessive construction *Boleloko bja ANC* and the isiZulu possessive construction *Ubulunga be-ANC* 'Membership of the ANC' have the same structure, i.e. noun + possessive concord + noun. The verbs following these possessive constructions, *bo buletšwe* and *buvuleleke* respectively, also have the same structure namely subject concord of class 14 followed by a verb stem. The number of orthographic words used in the two languages differs however. In these examples the word order is exactly the same for the two languages and consecutive and direct comparison of each structure in terms of conjunctivism versus disjunctivism is possible.

However, quite a number of factors impact negatively on this assumed similarity between the Bantu languages as will be illustrated in terms of a randomly selected section from the same set of parallel texts.

**English** The gold represents the mineral and other natural wealth of South Africa, which belongs to all its people, but which has been used to benefit only a small racial minority.

**Sepedi** Gauta e emela diminerale le mahumo a mangwe a naga, ao e lego a batho ka moka, gomme wona a ile a šomišetšwa fela ke sehlophana se se nnyane.

**isiZulu** Kanti igolide limele umcebo ombiwa phansi kanye nomnotho wemvelo waseNingizimu Afrika, umnotho ongowabantu bonke, kodwa usetshenziswa idlanzana labamhlophe ukuzizuzela bona bodwa.

**Observation 1**: A neutral or factual translation is given for *gold* in English = *the gold* = Sepedi *gauta*, but in isiZulu a conjunctive *kanti* 'just so, in fact, indeed' was added. The isiZulu version could just as well begin with *Igolidi limele* ... and likewise, the English and Sepedi with *In fact gold represents* ... and *Ka nnete, gauta e emela* ... respectively. The isiZulu translation is in our view not necessarily less acceptable. It would however be incorrect to

[4] Source = http://www.anc.org.za/about/anc.html.

conclude that isiZulu is more disjunctive than Sepedi in this case. The compiler simply added a word to focus the attention on the symbolism of *gold*.

**Observation 2**: The Sepedi equivalent of the English phrase *other natural wealth*, i.e. *mahumo a mangwe*, simply refers to 'other wealth'. As the concept *natural* was not translated, *other wealth* (two words) should be contrasted with *mahumo a mangwe* (three words) and not *other natural wealth* with *mahumo a mangwe.*

**Observation 3**: The English and isiZulu versions refer to *of South Africa* and *waseNingizimu Afrika* 'of South Africa' but this is simply translated in Sepedi as *a naga* 'of the country' instead of *a Afrika Borwa* (of Africa South) 'of South Africa'. Once again, *waseNingizimu Afrika* (two words) can be directly compared to *a Afrika Borwa* (three words) but *waseNingizimu Afrika* cannot be directly compared to *a naga* since the literal meanings differ.

**Observation 4**: The concept *all* is translated in isiZulu by means of a quantitative pronoun. For a class 2 noun, to which *abantu* 'people' belongs, it is *bonke*, thus *abantu bonke* (people all) 'all people'. Sepedi has a similar set of quantitative pronouns and the concept could therefore have been translated as *batho bohle* (people all) 'all people', expressed by two orthographic words in both languages. The compiler for Sepedi nevertheless opted for an alternative strategy, namely a particle group *ka moka* (with the whole) 'all', thus *batho ka moka* 'all the people', a total of three orthographic words. It would be incorrect to conclude that Sepedi is more disjunctive than isiZulu in this case.

**Observation 5**: The Sepedi text refers to *sehlophana se se nnyane* as equivalent for the isiZulu *idlanzana* 'small group'. Sepedi uses a noun with a diminutive suffix within an adjective construction, and isiZulu only a noun with a diminutive suffix. The researcher could incorrectly conclude that *four* orthographic words are used in Sepedi and only one orthographic word in isiZulu to express this concept. Firstly, the Sepedi phrase should have been written as *sehlophana se senyane*, thus *three* orthographic words (with a single *n-* in the adjective stem). Secondly, *sehlophana* already means 'a small group' and is thus structurally (noun + diminutive suffix), and semantically (small group) equal to *idlanzana*. The fact that the Sepedi compiler opted for an (additional) adjective with the stem *-nyane*, thus emphasizing the idea of a *very* small group / minority is quite acceptable but cannot be directly compared to *idlanzana* in terms of conjunctivism versus disjunctivism.

**Observation 6**: Finally, the three versions differ substantially in respect of reference to this 'small group'. The Sepedi version simply refers to 'a very small group'. The English version adds the concept of *race* whilst the isiZulu one refers to Whites (*labamhlophe* 'of Whites'). No conclusions in respect of conjunctivism versus disjunctivism can be drawn since there is no direct 1-1 equivalence between *a small racial minority*, *a very small group* and *a small*

*group of Whites* in the English, Sepedi and isiZulu versions respectively. These three phrases simply give different information.

It is clear that even in such very short, randomly selected parallel paragraphs, quite a number of differences between the translations are evident which should be considered in the study of conjunctivism versus disjunctivism.

## 3.2 ORTHOGRAPHIC-WORD RATIOS ACROSS PARALLEL TEXTS

Despite the seemingly grim picture sketched in the previous section, it *is* true that there is, from the point of view of average number of words, considerable consistency between certified translations. This can be verified as follows. In Figure 3 a set of 26 parallel texts in Sepedi and isiXhosa, totalling 170,298 orthographic words in Sepedi and 101,851 in isiXhosa, is compared.[5] For each text, the ratio of the number of orthographic words in Sepedi to the number of orthographic words in isiXhosa is shown.

**Figure 3.** Ratio of the number of orthographic words in Sepedi to the number of orthographic words in isiXhosa in 26 parallel texts ($r$ = 0.996).

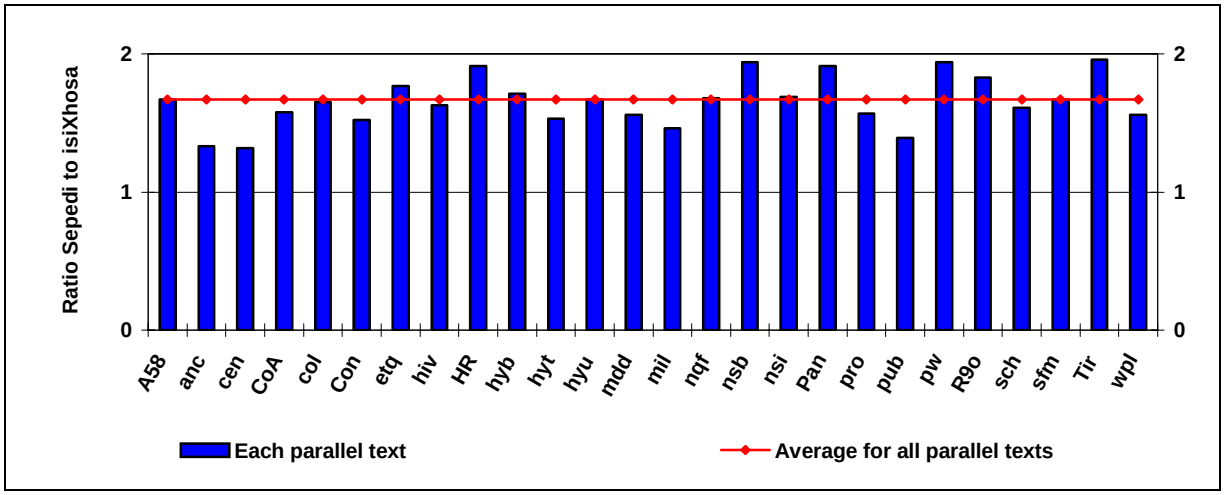



Figure 3 clearly indicates that each individual ratio is very close to the overall ratio of 1.67. This is confirmed by a calculation of the correlation coefficient $r$ which is as high as 0.996. In other words, no matter the topic of the parallel text nor the author of the translation, one orthographic word in isiXhosa will always roughly correspond to 1.67 orthographic words in Sepedi (or, conversely, one orthographic word in Sepedi will always roughly correspond to 0.60 orthographic words in isiXhosa).

This *stable pattern* is of extreme importance – and might even sound surprising in the light of §3.1 – yet forms the basis of the relative measurement instruments that will be introduced below. Two groups of parallel texts were

[5] For the sources of all these parallel texts, see De Schryver (2002: Appendix 2).

assembled for the purpose of creating these measurement instruments. Firstly, 10 sets of eleven-language parallel texts were brought together, totalling 348,467 orthographic words. Secondly, another 35 sets of parallel texts, with at least one text in a South African Bantu language and then one or more parallel texts in any of the other South African languages, were also compiled. Together, the 45 sets total more than 2 million orthographic words. A detailed description of this eleven-language parallel corpus can be found in De Schryver (2002: §1.3). The stable pattern illustrated in Figure 3 was found throughout all the language combinations of this eleven-language parallel corpus.

## 3.3 Towards an 11 x 11 Array Using a Set of 10 Eleven-Language Parallel Texts

The first group of parallel texts, being a set of 10 eleven-language parallel texts, can be used as follows. The number of orthographic words in each of the 10 texts can be summed for each language, upon which the resulting sum for each language can be compared to that of any other one. This is shown in Table 2.

**Table 2.** 11 x 11 array based on orthographic word counts across a set of 10 eleven-language parallel texts.

| | **Siswati** | **isiNdebele** | **isiXhosa** | **isiZulu** | **Afrikaans** | **English** | **Xitsonga** | **Setswana** | **Tshivenda** | **Sepedi** | **Sesotho** |
|---|---|---|---|---|---|---|---|---|---|---|---|
| ***sum →*** | *22,054* | *22,362* | *22,675* | *23,948* | *31,869* | *32,320* | *33,884* | *37,535* | *38,603* | *38,716* | *44,501* |
| **Siswati** | **1.00** | 1.01 | 1.03 | 1.09 | 1.45 | 1.47 | 1.54 | 1.70 | 1.75 | 1.76 | 2.02 |
| **isiNdebele** | 0.99 | **1.00** | 1.01 | 1.07 | 1.43 | 1.45 | 1.52 | 1.68 | 1.73 | 1.73 | 1.99 |
| **isiXhosa** | 0.97 | 0.99 | **1.00** | 1.06 | 1.41 | 1.43 | 1.49 | 1.66 | 1.70 | 1.71 | 1.96 |
| **isiZulu** | 0.92 | 0.93 | 0.95 | **1.00** | 1.33 | 1.35 | 1.41 | 1.57 | 1.61 | 1.62 | 1.86 |
| **Afrikaans** | 0.69 | 0.70 | 0.71 | 0.75 | **1.00** | 1.01 | 1.06 | 1.18 | 1.21 | 1.21 | 1.40 |
| **English** | 0.68 | 0.69 | 0.70 | 0.74 | 0.99 | **1.00** | 1.05 | 1.16 | 1.19 | 1.20 | 1.38 |
| **Xitsonga** | 0.65 | 0.66 | 0.67 | 0.71 | 0.94 | 0.95 | **1.00** | 1.11 | 1.14 | 1.14 | 1.31 |
| **Setswana** | 0.59 | 0.60 | 0.60 | 0.64 | 0.85 | 0.86 | 0.90 | **1.00** | 1.03 | 1.03 | 1.19 |
| **Tshivenda** | 0.57 | 0.58 | 0.59 | 0.62 | 0.83 | 0.84 | 0.88 | 0.97 | **1.00** | 1.00 | 1.15 |
| **Sepedi** | 0.57 | 0.58 | 0.59 | 0.62 | 0.82 | 0.83 | 0.88 | 0.97 | 1.00 | **1.00** | 1.15 |
| **Sesotho** | 0.50 | 0.50 | 0.51 | 0.54 | 0.72 | 0.73 | 0.76 | 0.84 | 0.87 | 0.87 | **1.00** |

Every value in Table 2 is simply the ratio of the corresponding total orthographic word counts. From this table one can for instance conclude that one orthographic word in Tshivenda corresponds to 0.62 orthographic words in isiZulu (23,948 / 38,603), or conversely that one orthographic word in isiZulu corresponds to 1.61 orthographic words in Tshivenda (38,603 / 23,948).

Unfortunately, the 10 sets of parallel texts only add up to corpora with total sizes ranging from 22,054 orthographic words for Siswati, to 44,501 for Sesotho. The values in Table 2 are thus still rather fragile.

## 3.4 Towards an 11 x 11 Array Using Sets of Two-by-Two Parallel Corpora

In order to arrive at even more reliable values, similar calculations as those presented under §3.3 can be done, but now on larger corpora. Instead of working simultaneously with eleven parallel corpora, of which there are unfortunately not enough available at this point in time, one can also calculate ratios between *pairs of parallel corpora*. In this way, and with the parallel corpora totalling 2-million-plus orthographic words mentioned in §3.2, up to 32 parallel sets can be compared to one another, constituting up to a quarter million words per language. The results of all these two-by-two comparisons can be seen in Table 3.

**Table 3.** 11 x 11 array based on orthographic word counts derived from 55 two-by-two parallel corpora.

| | **isiNdebele** | **Siswati** | **isiXhosa** | **isiZulu** | **English** | **Afrikaans** | **Xitsonga** | **Setswana** | **Tshivenda** | **Sepedi** | **Sesotho** |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **isiNdebele** | **1.00** | 1.01 | 1.01 | 1.04 | 1.41 | 1.41 | 1.61 | 1.63 | 1.67 | 1.73 | 1.77 |
| **Siswati** | 0.99 | **1.00** | 1.03 | 1.04 | 1.41 | 1.41 | 1.61 | 1.62 | 1.69 | 1.72 | 1.77 |
| **isiXhosa** | 0.99 | 0.97 | **1.00** | 1.01 | 1.36 | 1.37 | 1.58 | 1.58 | 1.75 | 1.67 | 1.71 |
| **isiZulu** | 0.96 | 0.97 | 0.99 | **1.00** | 1.32 | 1.34 | 1.54 | 1.55 | 1.58 | 1.60 | 1.66 |
| **English** | 0.71 | 0.71 | 0.74 | 0.76 | **1.00** | 1.00 | 1.15 | 1.16 | 1.19 | 1.24 | 1.25 |
| **Afrikaans** | 0.71 | 0.71 | 0.73 | 0.75 | 1.00 | **1.00** | 1.15 | 1.16 | 1.19 | 1.23 | 1.24 |
| **Xitsonga** | 0.62 | 0.62 | 0.63 | 0.65 | 0.87 | 0.87 | **1.00** | 1.01 | 1.05 | 1.06 | 1.08 |
| **Setswana** | 0.62 | 0.62 | 0.63 | 0.64 | 0.86 | 0.86 | 0.99 | **1.00** | 1.03 | 1.07 | 1.08 |
| **Tshivenda** | 0.60 | 0.59 | 0.57 | 0.63 | 0.84 | 0.84 | 0.96 | 0.97 | **1.00** | 1.03 | 1.08 |
| **Sepedi** | 0.58 | 0.58 | 0.60 | 0.62 | 0.81 | 0.81 | 0.94 | 0.94 | 0.97 | **1.00** | 1.02 |
| **Sesotho** | 0.57 | 0.57 | 0.58 | 0.60 | 0.80 | 0.80 | 0.92 | 0.92 | 0.93 | 0.98 | **1.00** |

It is important to realise that every (i, j) and (j, i) value in Table 3 is based on a different set of two-by-two corpora – hence 55 different pairs in all ((11x11-11)/2). Despite this, the *internal consistency* of this 11 x 11 array is truly remarkable.

As before, the degree of conjunctivism / disjunctivism of each language can be compared to that of any other, thus in relative terms. For instance, Table 3 indicates that one orthographic word in Siswati corresponds to 1.62 orthographic words in Setswana (or one orthographic word in Setswana to 0.62 orthographic words in Siswati). English and Afrikaans also happen to separate the conjunctive and disjunctive languages, creating four quadrants. In Quadrant 1 (top left), conjunctive orthographies are compared to conjunctive ones, and in Quadrant 4 (bottom right) disjunctive orthographies are compared to disjunctive ones. As a rule of thumb, the ratios can be thought of as being 1.0 in those quadrants. In Quadrant 2 (top right) conjunctive orthographies are compared to disjunctive ones. Here the ratio is on average 1.6, meaning that a disjunctively-written

South African language will on average have 60% more orthographic words than a conjunctively-written one. Finally, in Quadrant 3 (bottom left), disjunctive orthographies are compared to conjunctive ones. The average ratio is 0.6, which thus implies that a conjunctively-written South African language has an average of 40% fewer orthographic words than a disjunctively-written one.

One can also make comparisons with English and Afrikaans. The first thing one notices is of course the closeness of English and Afrikaans: it is indeed fair to assume that both languages use an equal number of orthographic words to convey the same content. Furthermore, conjunctively-written languages use on average 25 to 30% less orthographic words than English / Afrikaans, while the disjunctively written languages use an average of 15 to 25% more.

## 4. USES OF THE DEGREE OF CONJUNCTIVISM / DISJUNCTIVISM

In this last section we list some of the uses of the designed measurement instrument for the degree of conjunctivism / disjunctivism of the South African languages. Actually, these uses *prompted* the very design of this instrument.

**Use 1**: A recurrent issue for translators in today's South Africa is to be able to draw up reasonably consistent quotations for translations based on the number of words to be translated, especially if the translator prefers to take the expected number of words in the end product into account. In the light of the above, it goes without saying that one cannot have one single 'word fee' across all languages. Rather, Table 3 indicates, for instance, that the 'word fee' for isiNdebele should be roughly 1.67 times that of the one for Tshivenda. Related to this issue, translators are often asked to make a prediction of the expected number of pages, or are even required to fit their translation into a certain number of pages. Table 3 shows that a translation of a 10-page English flyer into Sepedi will result in an 11- or perhaps even a 12-page flyer (as more words mean more spaces, and thus more space). If the translated flyer also has to fit in the same number of pages, which is also often the case, then these values give a rough indication of how much should be cut / expanded.

**Use 2**: South African journals appreciate it if articles include a translation of the abstract in one of the official African languages. If the requirement is 'at least 100 words', and the English or Afrikaans version is only just above that, then chances are real that the translation for a conjunctively-written language will fall short of 100. In the Appendix one can find such a case. Even though the English abstract is 134 orthographic words long, the isiNdebele version is only 80 orthographic words long. The ratio is 0.60, thus worse than the 0.71 shown in Table 3.

**Use 3**: Another highly functional use is found in corpus linguistics. Until the late 1980s the unofficial standard size, or in terms of Leech (1991: 22), the

'going rate' for electronic corpora remained at roughly one million running words. Today, the world aims at copying the size of the *British National Corpus*, which stands at 100 million running words. Table 3 tells us, e.g., that a 5-million-word isiZulu corpus is in fact equivalent to a 6.6-million-word English corpus (cf. De Schryver & Gauton 2002). A 5.8-million-word Sepedi corpus, on the other hand, is only equivalent to a 4.7-million word English corpus.

**Use 4**: Currently, the data in Table 3 are used in the development of spellcheckers for the Bantu languages. As will be shown in Prinsloo & De Schryver (forthcoming), conjunctivism and disjunctivism result in very different needs.

**Use 5**: In lexicography conjunctivism and disjunctivism have direct implications for the so-called stem- versus word-tradition of lemmatisation. The derived relative conjunctivism / disjunctivism values are being linked to multi-dimensional Rulers for the Bantu languages.

**Use 6**: Any other language, and especially relevant, any other Bantu language, can be added to the 11 x 11 array using similar principles. A complete 'Bantu Array' would indeed prove to be a highly valuable tool.

## 5. CONCLUSION

In this article a measurement instrument for the degree of conjunctivism / disjunctivism of the South African languages has been proposed. It was first indicated that these two opposite writing traditions are primarily a matter of convention, and that one should not be considered more scientific than the other. Examples clearly illustrated that it is in fact possible to simulate a conjunctive way of writing for disjunctively-written languages and vice versa.

The possibilities for the design of an absolute measurement instrument were investigated, and three potential rulers were derived: one based on the overall average orthographic-word length, the other based on the average number of orthographic words per sentence, and a third one based on the standardised type/token ratio. All of these were found to have shortcomings.

The search for a relative measurement instrument was more successful. It was found that large sets of parallel texts would provide the ideal data. In the absence of this, two-by-two parallel corpora offer a good substitute, and the final 11 x 11 array, based on 55 different pairs of corpora, was actually compiled in this way.

Applications of the designed relative measurement instrument for the degree of conjunctivism / disjunctivism of the South African languages in several fundamental and applied linguistic fields were also reviewed. These include the fields of translation, academic writing, corpus linguistics, and theoretical reflections about spellcheckers and multi-dimension dictionary Rulers. A complete Bantu Array could be the ultimate goal.

## REFERENCES

De Schryver, Gilles-Maurice. 2002.
*Web for/as Corpus. A Perspective for the African Languages*. **Nordic Journal of African Studies** 11.

De Schryver, Gilles-Maurice and Rachélle Gauton. 2002.
*The Zulu locative prefix ku- revisited: A corpus-based approach*. **Southern African Linguistics and Applied Language Studies** 20 (3).

Leech, Geoffrey N. 1991.
The State of the Art in Corpus Linguistics. In *English Corpus Linguistics: Studies in Honour of Jan Svartvik*, K. Aijmer and B. Altenberg (eds.), pp. 8-29. London: Longman.

Louwrens, L.J. 1991.
*Aspects of Northern Sotho Grammar*. Pretoria: Via Afrika Limited.

Prinsloo, D.J. and Gilles-Maurice de Schryver. forthcoming.
*A perspective on spellcheckers for the African languages and Afrikaans*.

Scott, Mike. 1999.
*WordSmith Tools version 3*. Oxford: Oxford University Press. See for this software also http://www.lexically.net/wordsmith/index.html.

Van Wyk, Egidius B. 1995.
*Linguistic Assumptions and Lexicographical Traditions in the African Languages*. **Lexikos** 5 (AFRILEX-reeks/series 5B: 1995): 82-96.

Ziervogel, Dirk and Mokgokong, Pothinus C.M. 1975.
*Pukuntšu ye kgolo ya Sesotho sa Leboa, Sesotho sa Leboa–Seburu/Seisimane / Groot Noord-Sotho-woordeboek, Noord-Sotho–Afrikaans/Engels / Comprehensive Northern Sotho Dictionary, Northern Sotho–Afrikaans/English.* Pretoria: J.L. van Schaik.

**Appendix.** Abstracts of articles and 'number of words'.

| Abstract [1] | Okumumethweko Ngobufitjhani [2] |
|---|---|
| **Drawing up the macrostructure of a Nguni dictionary, with special reference to isiNdebele. [13]** | **Ukutlhama ingaphakathi lesiHlathululi-mezwi emalini wamaNguni, siqalise elimini lesiNdebele. [8]** |
| In this article a four-step methodology is proposed for the creation of the lemma-sign list of a Nguni-language reference work. The theoretical principles are illustrated throughout with a full-scale case study revolving around isiNdebele. For the suggested approach raw corpus data is utilised, and only standard, straightforward | I-atikili le iveza iindlela ezine zobusayensi ekungizo eziphakamiswako bona zingasetjenziswa ekwakheni irhelo lamagama elisiboniso emalini wamaNguni. Indlela esetjenzisiweko itjengiswa ngokupheleleko ngokobana kusetjenziswe isiNdebele. Ukuze isiphakamiswesi siphumelele, ikhophasi engakenziwa litho isetjenzisiwe begodu |

| | |
|---|---|
| and widely-available software tools are required to process the data. Apart from the inherent value of having an entire macrostructure at one's disposal right from the start of a dictionary project, it is shown how such a list can also be used for both predictions and measurements on lemma-sign, page and time levels. As such, drawing up the macrostructure of a dictionary, automatically leads to a "ruler" with which the entire lexicographic process can successfully be monitored. Specifically for isiNdebele, suggestions are made for the way ahead. [134] | kusetjenziswe ihlelo lekhomphuyutha elijayelekileko nelaziwa khulu. Ngaphandle nje kwegugu lokufuna ingaphakathi lesiHlathululi-mezwi ekuthomeni, kuyatjengiswa bonyana irhelweli lingaba lisizo elingangani ekuboneni ngaphambili nokulinganisa inani lamagama, amakhasi kanye namazinga weenkhathi. Alo-ke ngokuqala ingaphakathi lesiHlathululi-mezwi, lokho kusitjhingisa "eruleni", ekungiyo-ke engakghona ukutjheja woke umsebenzi wokuhlathulula. Ngokuqalisa ehlangothini lesiNdebele, iimphakamiso ziyavezwa ukuze kuragelwe phambili. [80] |