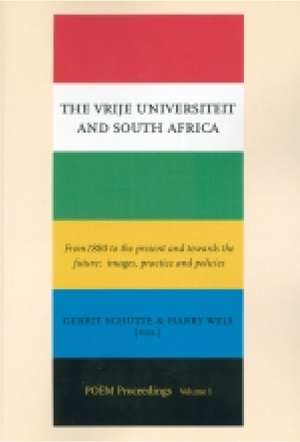
The Vrije Universiteit And South Africa ~ ANNA And A ‘New’ Lexicography For South Africa


THE VRIJE UNIVERSITEIT
AND SOUTH AFRICA
From 1880 to the present and towards the future: images, practice and policies
GERRIT SCHUTTE & HARRY WELS
[EDS.]
POEM Proceedings Volume 1

Introduction

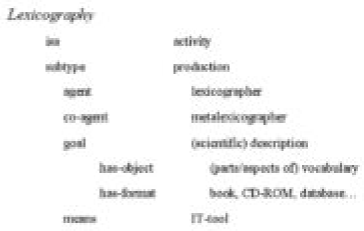
In this paper we will try to make clear that the ANNA-project, in its own way, is one of the (possible) examples to show that different/changing situations, needs and target groups (may) require different/new approaches, models and products.

We will do so by taking the following steps:

– First, some basic terminology will be given.
– Secondly, the lexicographical situation in South Africa will be outlined.
– Thirdly, the ANNA-project itself will be presented.
– Next the ‘new’ features of the ANNA-project will be highlighted.
– To end with, the pros and cons of ANNA in a ‘new’ South Africa will be discussed.

Lexicography

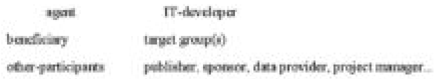
We can define lexicography in at least two ways. The first ‘classical’ definition could read as follows: ‘Lexicography is the description of one or more aspects of one or more vocabularies in function of one or more target groups or users’. For the second, more ‘formal’, definition the following ‘frame’**[i]** could be used:



Lexicography	
isa	activity
subtype	production
agent	lexicographer
co-agent	metalexicographer
goal	(scientific) description
has-object	(parts/aspects of) vocabulary
has-format	book, CD-ROM, database…
means	IT-tool

Although the latter definition differs *formally* from the former, it does not really do so from the point-of-view of *content*. One can easily paraphrase the above frame-based definition as: ‘Lexicology is an activity which leads/should lead to a product made by lexicographers/metalexicographers with the aim to come to a (scientific) description, etc'.

Next to the possibilities that a frame-based definition offers on the levels of explicitness and consistency, it serves our purpose better than the more traditional one in two ways:



agent	IT-developer
beneficiary	target group(s)
other-participants	publisher, sponsor, data provider, project manager…

- First, it makes clear the fact that lexicography as an activity is no longer to be considered as a solitary act of a lexicographer, but rather as a scene on which next to lexicographers, different players, such as metalexicographers (theoreticians, designers of models/theories upon which to base lexicographical practice), tool developers, project managers, data providers, users, and publishers play a role.
- Secondly, as one can observe, a frame has a stable side (the left hand side, the slots) and a variable side (the right hand side, the fillers). It is to be expected that changes in the fillers over time will entail changes in the character of lexicography itself and so lead to a 'new' lexicography. In particular this is what has happened to the fillers for the 'format', 'means', 'beneficiary', and 'other-participants'-slots during the last decades.

The lexicograpical situation in South Africa**[ii]**

The situation before 1994

This part of the article is mainly based on Van Schalkwyk 2003. In what follows, we will give a brief outline bullet-wise of the most important lexicographical 'facts' of the pre-1994 period:

- Two official languages existed (English and Afrikaans).
- their wake two big government-subsidized lexicographical projects were carried out:
- the Dictionary of SA English on Historical Principles (1968-1996);
- the WAT (= Woordeboek van die Afrikaanse Taal) (1926 – …) (up until 2003, 11 volumes (A-O) have been published).
- For African (black) languages three university projects have been started up:

Xhosa (1968 – …) (Greater isiXhosa Dictionary), Zulu (1977 – …) (isiKazamasu Dictionary) and Sesotho sa Leboa (1988 – …) (English – Pedi v.v. Dictionary).
– Furthermore there existed bilingual wordlists between English/Afrikaans and one or more of the African languages mainly meant to give (rapid) access to the 'white' languages for the 'black' language speaker.

The situation after 1994
A new lexicographical policy/regulation takes place from the side of government leading to:

– The establishment of 11 official languages in 1996: the existing ones, English and Afrikaans, were augmented by Xhosa, Zulu, Ndebele, Swati (= members of the Nguni family), Sesotho sa Leboa, Sesotho, Setswana (= members of the Sotho family), Xitsonga, and Tshivenda.
– PANSALB (= Pan South African Language Board) was established in 1996 for the development and stimulation of all official South African languages and, in particular, those which have been marginalized up until now; PANSALB should also promote multilinguality.
– In 1998/99 lexicographical units were established for all South African official languages with the aim to come to monolingual data/databases, serving as a basis for all kinds of lexicographical products.

Van Schalkwyk (2003) reports with regard to the state-of-affairs of the lexicographical units: 'In die tussetyd is reeds personeel aangestel. Op die oomblik is die soektog na hoofdredakteurs aan die gang'. ['In the meantime staff has already been appointed. At the moment the search for editors-in-chief is going on'.] This quotation makes clear that, apart from the specific descriptive problems for African languages, the most outstanding problem of South African lexicography at the moment, is the fact that lexicographers/ metalexicographers with a formal training are scarce, making the process of starting up lexicographical projects for Africa languages an extremely slow one.

The ANNA-project: some facts and figures
ANNA is an acronym for *Afrikaans-Nederlands, Nederlands-Afrikaans*. As a dictionary project it came to the fore in 1999 when a *Feasibility and Definition Study* was undertaken (see Martin, Gouws and Renders 1999) in order to find out whether such as dictionary was needed and if so, which features it should have. In 2000 it was decided to definitely start up the project, expecting it to be finalized

in 2006/7. ANNA is mainly subsidized by private money, namely by two Dutch foundations: ZASM (= *Zuid Afrikaanse Spoorweg Maatschappij*, 'South African Railway Company', Amsterdam) as the main sponsor, and the Van den Berch van Heemstede Foundation, The Hague. The Universities of Potchefstroom and Stellenbosch have also financially contributed. The work is carried out in close co-operation between the universities of Stellenbosch, Port Elisabeth, Vrije Universiteit Amsterdam, and the Limburg University Centre (Belgium); the Vrije Universiteit acting as a co-ordinator.

The production is also expected to be a co-operative effort between South Africa and the Low Countries, as Pharos (for the electronic version) and Van Dale (for the paper version) are the intended publishers. The main features of the final ANNA-product can be summarized as follows:

ANNA
– aims to be a bilingual dictionary/database;
– not in the traditional sense of the word;
– making use of modern, new technology (tools) and new metalexicographical insights (models);
– which should lead not only to an innovative, contrastive, comparative dictionary/database between Afrikaans and Dutch;
– but also to new models and tools which can be used by other language pairs such as the African languages spoken in South Africa.

This last feature was one of the basic requirements of the partners to participate in the project and one of the main issues that has been investigated in the pilot study. In what follows this 'innovative' feature of ANNA will be dealt with in more detail.

ANNA and the 'new' lexicography
In order to make clear what is new in ANNA, we will start from a very simple but concrete example: the Dutch word *zalm*, with its Afrikaans equivalent *salm* (English: 'salmon') and show what ANNA does and what it does not (does no longer).

What ANNA DOES NOT
A traditional bilingual Dutch-Afrikaans v.v. dictionary (NA/AN) would contain two entries for 'salmon', one in the NA-part zalm salm, and one in the AN-part salm ®

zalm. One could imagine these entries to look as follows:

NA-part:
ZALM**[iii]**

1 [kind of fish] *salm*
een plakje zalm ’n *repie salm*; verse zalm *vars salm*; een blikje zalm ’n *blikkie salm*; gerookte zalm *gerookte salm*; <fig.> *het neusje van de zalm die allerbeste.*

AN-part:
SALM
1 [kind of fish]
’n repie salm een *plakje zalm*; vars salm *verse zalm*; ’n blikkie salm een *blikje zalm*; gerookte salm *gerookte zalm*.

As one can observe, closely related languages, such as Dutch and Afrikaans, resemble each other strongly in their ‘complementary’ parts. They are not each other’s mirror image, yet they come very close. The way they are treated above shows a very high degree of redundancy, that is why we will prefer another descriptive model, the amalgamation model, illustrated in what follows.

ANNA does not treat two closely related languages apart, but unifies them, malagamates them into one, single macrostructure. In other words, it treats zalm/salm as if the two entries were the same (or variants of a common entry). This has the advantage that one can both reduce redundancy (see the preceding paragraph) and enhance/optimalise comparability/contrastivity. For instance, in the case of zalm/salm it will become clear at a single glance what is similar/different between them as the entry below (which is an ANNA-entry) shows.

ZALM/SALM
1 [a kind of fish]
<< >> verse zalm *vars salm*; een blikje zalm ’n *blikkie salm*; gerookte zalm *gerookte salm*;
>> << een plakje zalm ’*n repie salm*;
<fig.> het neusje van de zalm *die allerbeste.*

The symbols used are to be interpreted as follows:

<< >>: similar, non-contrastive examples/combinations;
>> <<: different, contrastive examples/combinations;
<fig.>: idioms and figurative usage.

If one is only interested in differences between the two languages (from a combinatorial point-of-view) one could for instance skip the << >>-section. This way the user can define his own 'paths' through the data, depending on his/her interests and needs.

What ANNA CAN DO (for a 'new' lexicography in South Africa)
It will have been clear from the preceeding that it is the ambition of the ANNA-project that some of its material and immaterial infrastructure can be used *beyond* the project itself. This is the case for the amalgamation model (immaterial infrastructure) and for the tool to deal with it (material infrastructure). In what follows we briefly present both possibilities.

The amalgamation model
The amalgamation model has proven its value for the ANNA-project in that it successfully unifies the macrostructures of two closely related languages and optimises their comparability. Its main features are the following:

– The two macrostructures are unified into one, amalgamated structure.
– This is done on the basis of formal and semantic relatedness.**[iv]**
– This amalgamation leads to different lemma-types:

* Combined lemmata (cognates) (A + D) such as:
– absolute cognates (e.g. A *tafel*/D tafel (E *table*));
– absolute cognates with systematic morphological/orthographical differences (e.g. A *ontsnap*/D *ontsnappen* (E escape));
– partial cognates (contrary to absolute cognates, partial cognates do not share all meanings, e.g. *robot* A/D = E automaton, A only = traffic light);

* Non-combined lemmata (non-cognates) which formally differ, such as D *verkeersdrempel*/A *spoedwalletjie* (E speedwall) and false friends which semantically differ, such as:
– D *mus* = A *mossie* (= E sparrow);
– A *mus* = D *muts* (= E cap).
* The model can reduce or leave out non contrastive, redundant examples/combinations.

* The model guarantees optimal and direct comparability between two closely related languages.
* The model has proven to be exportable to other closely related languages such as languages from the Nguni and the Sotho family (see Mashamaite 1995).

ANNA, OMBI and the Hub-and-Spoke Model

In order to come to an efficient language infrastructure, resources which have to function within a multilingual environment should be developed with a view on their *usability* and *linkability* with other resources of that environment. Therefore they should use adequate instruments such as editors with linking and reversing capacity. ANNA makes use of such an editor, viz. OMBI (acronym for ‘OMkeertool voor BIlinguale bestanden (‘reversing tool for bilingual databases’)).**[v]** It would lead us too far to enter into details here; let it suffice to state that OMBI links two words of two different languages at meaning level: Dutch ‘paard’, for instance, is linked in its ANIMAL meaning to ‘horse’ (in its ANIMAL meaning) and, in its CHESS meaning, to ‘knight’ (in its CHESS meaning). Furthermore, OMBI specifies the relationship between these semantic units so that one can use them in a calculus in order to reverse them, block them, derive other values for them etc. In other words, OMBI does as what is illustrated in figure 1:

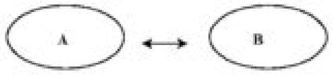


Fig. 1: Linking items from language A with those of language B and vice versa

Figure 1 – Linking items from language A with those of language B and vice versa

The fact that one has explicitly specified the relationships between the items makes that, when adding one language (C) to the data collection (A « B), one can infer the links between B and C by means of derivation rules. See fig. 2:

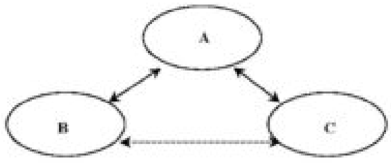


Fig. 2: Linking two languages (B and C) to one common language (A) and deriving the links between B and C

Figure 2 – Linking two languages (B

and C) to one common language (A) and deriving the links between B and C

Language A now functions as a common link. It is called the hub-language in analogy with air-traffic organization where often one does not fly directly from a (spoke) airport to another (spoke) airport, but via a hub, a central airport. B and C are the spoke-languages.

The Hub-and-Spoke Model is a model with a large potential. Its strength lies in the fact that it exploits the intra- and interlingual relations in and between languages and does so via a hub (instead of bidirectionally). One can imagine that in a multilingual context such as in South Africa, where there are 11 official languages, the model seems to be very promising. In such a situation (11 languages), 55 different pairs (= 110 bilingual dictionaries; taking both directions into account) could be derived. In a hub-and-spoke configuration one could suffice with ten direct links (10 spokes to 1 hub), the remaining 45 being generated as indirect spin-offs. Actually there is a gain in data-production from the third language onwards (see fig. 3). For more information on the Hub-and-Spoke Model see Martin and Heid (1998).

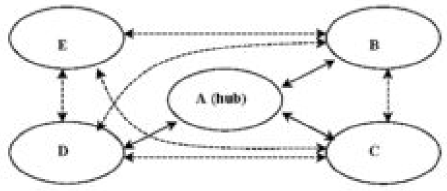


Figure 3 – Situation with five different languages in a HaS-configuration: ten language pairs; four directly linked (indicated by full lines); six indirectly linked (indicated by broken lines).

Conclusion: *PROS and CONS of the ANNA project*

As a conclusion one may state that there is (at least) one good argument NOT to carry out the ANNA-project and three good reasons to DO SO. We list them one after the other in what follows.

The communicative argument
There is no big communicative gap to bridge between Afrikaans and Dutch considering the fact that speakers from Afrikaans and Dutch can communicate with each other each speaking his/her own language. Of course, this communication will now and then lead to miscommunication, to misunderstanding, and to linguistic problems, but solving these are not in themselves sufficient conditions to justify the ANNA project.

The functional argument
If one accepts that language is a vehicle not only for basic communication, but also to properly express oneself in, be it in literature, in science or in everyday situations, then a bilingual dictionary is an important instrument to understand the subtleties and nuances of the other language, the other culture. In Afrikaans and Dutch this is the more so as there do not exist fully-fledged dictionaries Afrikaans-Dutch and vice versa at the moment. In this respect ANNA fills a gap both for Dutch and for Afrikaans language users.

The descriptive argument
Afrikaans, until very recently, has been described as a rather homogeneous and 'pure' language. ANNA starts from a Dutch database, the RBN (= *Referentie Bestand Nederlands, 'Reference Database of Dutch'*)**[vi]** which shows a rather varied picture of Dutch. It is the aim to link Dutch to Afrikaans as it is used now, being the mother tongue of more coloured than white people (5,5 million speakers in all, in a proportion 60 per cent coloured, 40 per cent white). This way ANNA can give an impetus to a 'new' view on the lexicographical description of Afrikaans.

The scientific argument
Last but not least, ANNA has developed a model (the amalgamation model together with the appropriate tools to make the model operative, see preceding section) which has a general linguistic value. The model enables closely related languages, no matter whether they are spoken by 'white', 'black', or 'coloured' linguistic communities, to be described in an own contrastive way. In doing so, attention is paid both to similarities and differences. More in particular contrasts which lie on the supralexical level (co-text (combinations) and context (wider pragmatic situation)) and which often pass unnoticed, can now be captured. We

hope that this last feature in particular will pave the way to a ‘new’, useful and co-operative lexicography in South Africa.

NOTES

[i] A frame is a schema to represent knowledge with. It consists of slots (general classes) and fillers (specifications of the slots). Frames and their variants (as graphs, networks etc.) are well-known in Artificial Intelligence literature.

[ii] This part of the article is mainly based on Van Schalkwyk 2003.

[iii] For clarity’s sake the metalanguage in the meaning resumés is represented in English here (see: kind of fish).

[iv] Words (from Dutch and Afrikaans) that are combined are declared to be cognates. In order to be considered a cognate, the word pair must share the ‘same’ form (deviations permitted), and at least one meaning.

[v] For more information on OMBI see Martin and Tamm 1996.

[vi] For more information on the RBN see Van der Vliet 2005 (forthcoming).

References

Martin, W. and Tamm, A. (1996) ‘OMBI, an editor for constructing reversible lexical databases’, in M. Gellerstamm et al. (Eds), *Euralex ’96 Proceedings, Göteborg: Göteborg University*, pp. 675-85.

Martin, W. and Heid, U. (1998) *On the construction of bilingual dictionaries*, Den Haag: CLVV.

Martin, W., Gouws, R. and Renders, L. (1999) *Haalbaarheids- en definitiestudie voor een woordenboek Afrikaans-Nederlands/Nederlands-Afrikaans.* Amsterdam: VU University Press.

Martin, W. and Gouws, R. (2000) ‘A new dictionary model for closely related languages: The Dutch-Afrikaans dictionary project as a case-in-point’, in U. Heid et al. (Eds), *The ninth Euralex International Congress Proceedings,* Stuttgart: University of Stuttgart, pp. 783-92.

Martin, W. (2003) ‘Lexicography, lexicology, linking and the hub-and-spoke model’, in W. Botha (Ed.), *’n Man wat beur.* Huldigingsbundel vir Dirk van Schalkwyk, Stellenbosch: Buro van die WAT, pp. 268-85.

Mashamaite, K.J. (1995) *The hub-and-spoke model: A recipe for making bilingual dictionaries between African languages in South Africa*, MA-thesis, Amsterdam: Vrije Universiteit.

Van der Vliet, H. (2005, forthcoming). *Digitale bronnen voor het Nederlands: Het referentiebestand Nederlands* (RBN).

Van Schalkwyk, D. (2003) *Lexicografie in Afrika met die klem op Suid-Afrika* (unpublished manuscript).