Optimising Data Utilisation in Lexicography: The Case of the *Khoekhoegowab Dictionary*

Wilfrid Heinrich Gerhard Haacke, *Department of Language and Literature Studies, University of Namibia, Windhoek, Namibia (whaacke@gmail.com)*

Abstract: Despite of rapid progress in Southern Africa in the direction of multifunctionality of lexical databases through the advent of generic lexicographic software, a considerable number of lexicographic projects — especially in Khoe and Saan languages — still use or have recently used a word processor with the sole objective of compiling a printed dictionary. Hence the present paper expounds on the case of the *Khoekhoegowab Dictionary Project,* how in the early 1990s some off-the-shelf DOS-based database software was configured as part of a "home-grown" custom-made dictionary writing system. It is demonstrated in a non-technical way that the use of a structured database with fully-fledged retrieval facilities allows for the far-reaching elimination of human error in a dictionary, for the automatisation of processes like language reversal and sorting, and, finally, for the significantly enhanced usability of the data for purposes other than fixed media dictionary compilation. Compiling a dictionary without extensive query facilities as offered by tabular databases, is argued to be a lost opportunity, as it should be possible to utilise lexicographic data for more than just lexicography. By 2010 the data was accommodated in open source software to ensure its optimal survival in digital form for future use.

Keywords: AUTOMATISATION, COMPILATION SOFTWARE, DATA RETRIEVAL, DATABASE CONFIGURATION, DATABASE REPORT, FLAT-FILE DATABASE, FORM, INFORMATION GENERATION, KHOEKHOE, KHOESAAN DICTIONARIES, LEXICOGRAPHY, LOOKUP FACILITIES, MULTIFUNCTIONALITY, QUERY FACILITIES, RETRIEVAL FACILITIES, SOFTWARE, TONES

Opsomming: Die optimalisering van datagebruik in die leksikografie: Die geval van die *Khoekhoegowab Woordeboek*. Verskeie leksikografiese projekte — veral in die Khoe- en Saantale — gebruik 'n woordverwerker met die uitsluitlike doel om 'n gedrukte woordeboek saam te stel, of het dit tot onlangs nog so gebruik. Dit gebeur ondanks die opkoms van generiese leksikografiese sagteware, wat gelei het tot die vinnige vooruitgang van die multifunksionaliteit van leksikale databasisse in Suider-Afrika. Daarom gaan hierdie artikel in op die geval van die *Khoekhoegowab Woordeboekprojek,* hoe gekoopte DOS-gebaseerde databasissagteware vroeg in die 1990's as deel van 'n "tuisgemaakte" doelgemaakte woordeboekskryfstelsel opgestel is. Daar word op 'n nietegniese manier getoon dat die gebruik van 'n gestruktureerde databasis met volwaardige herwinningshulpmiddels verreikende uitskakeling van menslike foute in 'n woordeboek bewerkstellig, die outomatisering van prosesse soos taalomkering en sortering toelaat, en uiteindelik die bruikbaarheid van die data aansienlik verhoog vir doeleindes buiten woordeboeksaamstelling in vaste mediums. Daar word geargumenteer dat die samestelling van 'n woordeboek

sonder uitgebreide navraaghulpmiddels soos voorsien deur tabellariese databasisse, 'n verspeelde geleentheid is, aangesien dit moontlik behoort te wees om leksikografiese data te benut vir meer as net leksikografie. Teen 2010 is die data in oopbronsagteware geberg om te verseker dat dit optimaal voortleef in digitale vorm vir toekomstige gebruik.

Sleutelwoorde: OUTOMATISERING, SAGTEWARE VIR WOORDEBOEKSAMESTELLING, DATAHERWINNING, DATABASISKONFIGURASIE, DATABASISVERSLAG, PLATLÊERDATABASIS, VORM, INLIGTINGSGENERERING, KHOE, KHOESAANWOORDEBOEKE, LEKSIKOGRAFIE, NASLAANHULPMIDDELS, MULTIFUNKSIONALITEIT, NAVRAAGHULPMIDDELS, HERWINNINGSHULPMIDDELS, SAGTEWARE, TONE

1. Introduction

In a time of rapidly developing computational practices in lexicography it should amount to carrying coal to Newcastle to argue for the need of multifunctionality of lexical databases. However, despite of the very recent advances in software design there are still a number of recent if not current lexicographic projects in Southern Africa that operate with the sole objective of producing a printed dictionary, and that have pursued or still pursue this aim by means of a common word processor without resorting to an underlying database format or any dedicated lexicographic software. Most of these dictionaries deal with either Khoe or Saan languages, or Bantu languages outside South Africa, as the Bantu languages of South Africa are now catered for by the National Lexicography Units of the *Pan South African Language Board*, which by now all use *Tshwanelex*. The following titles are instances of Khoe or Saan dictionaries published in the last twenty years by means of common word processors (or, at best, *Toolbox*), or which are still in preparation:

— 1994. Dickens, P. *English–Ju|'hoan Ju|'hoan–English Dictionary*. (371 p.)

— 1994/2009. Traill, A. *A !Xóõ Dictionary*. (292 p.)

— 2001 (fourth prelim. edition). Visser, H. *Naro Dictionary* (240 p.)

— 2003. Kilian-Hatz, C. *Khwe Dictionary* (395 p.) based *i.a.* on the fieldwork data of Oswin Köhler collected since the 1960s)

— 2004. Weich, F. *San Dictionary/San-woordeboek. San–Afrikaans–English/ English–San–Afrikaans/Afrikaans–San–English*. (377 p.)

— 2008. König, C. and Heine, B. *A Concise Dictionary of Northwestern !Xun* (186 p.)

— Traill, A., A. Chebanne and H. Nakagawa (in prep.) *A Trilingual Dictionary in !Xóõ, English and Setswana*.

— Nakagawa, H., K. Sugawara and J. Tanaka (in prep.) *Ghui Dictionary*.

— Andy Chebanne uses Toolbox to collect lexical data from Shua and Tciretcire of the Eastern Kalahari;

— so do Christfried Naumann and Tom Gueldemann in their ongoing fieldwork on Western !Xoon.

The main reason for not using dedicated software is that projects like these usually are hamstrung by a historic legacy in that the compilation of the (lexical) corpus was started single-handedly in index card mode when — in several cases — personal computers did not exist yet, let alone generic software for dictionary compilation.[1] This use of obsolete methods may warrant a closer look at the case of the *Khoekhoegowab Dictionary Project* (originally registered as *Nama Dictionary Project* in 1981). This paper will expound on the way how in the early 1990s some off-the-shelf DOS-based database software was configured as part of a "home-grown" custom-made dictionary writing system, and how by 2010 the data was accommodated in open source software to ensure its optimal survival in digital form. The paper will then provide instance how such lexical data can be utilised significantly beyond its primary purpose of serving as lemmas in a fixed media dictionary or glossary. It is suggested that the time, cost and effort spent on converting lexical data that have been compiled in a word processor document into a structured database with fully-fledged retrieval facilities, is — in the long run — amply compensated for

— by the far-reaching elimination of human error in the dictionary;

— by the automatisation of processes like language reversal and sorting; and, finally,

— by the significantly enhanced usability of the data for purposes other than dictionary compilation.

To today start compiling a dictionary without extensive retrieval facilities as offered by databases, is a waste of opportunities that can no longer be justified. It should be possible to utilise lexicographic data for more than just lexicography.

2. The Beginnings of the *Nama Dictionary Project*

The "*Nama Dictionary Project*" (henceforth ***NDP***) commenced with data collection in 1981; that is, at least in Namibia, in the era of typewriters and handwritten index cards.

Pastor Eliphas Eiseb as Khoekhoe speaker and co-author, worked for the project full-time for eleven years (until 1992), collecting data by perusing the existing literature in Khoekhoegowab, when the team did not work in plenum (in the afternoons). Electronic text corpora for Khoekhoegowab did not exist in

those days of off-set printing, nor were corpus-query tools available for the analysis of text corpora and the compilation of concordances (cf. Van Sterkenburg 2003: 195 ff. and De Schryver and De Pauw 2007 on the development of digital resources). Even if electronic text corpora had been produced by scanning, they would have been of only limited value as the revised and standardised orthography of Khoekhoegowab had not yet established itself in the literature and lemmas would hence not have been recognised automatically. So the selection of lemmas depended almost solely on the discretion and memory of Eiseb while perusing the literature and recordings. The electronic database comprises over 24 500 Khoekhoe records, of which some 2 700 are illustrative examples. In this paper the term database will be used not in its widest sense of any accumulation of data but in the more technical sense of a tabular, structured database.

2.1 The Lexicographic Concept of the *Khoekhoegowab Dictionary*

The aim of the *NDP* was (and is) of a dual nature:

— firstly *practical*, to provide a comprehensive bilingual dictionary for general use by Khoekhoe as well as non-Khoekhoe speakers;

— secondly *academic*, to document the atrophying lexicon of the last surviving language of the KHOEKHOE branch of the KHOE (Central Khoesaan) family for comparative and other linguistic purposes.[2]

These two aims require rather diverging lexicographic approaches and procedures, and a combination of them unavoidably requires compromises. To do them full justice such aims can hardly be pursued in one and the same dictionary. Circumstantial constraints and urgency, however, demanded that several needs be addressed in the same project, as is so often the case in the early stages of lexicographic documentation of a language. The fact that the target users were to be of a maximally diversified nature required even more systematic planning of the lexicographic procedures, in an attempt to strike an acceptable balance. For an elaboration on these considerations see Haacke (1998).

The *Khoekhoegowab Dictionary* (***KhD***) was planned to cover the lexicon as widely as possible, including even archaic and obsolete words, albeit labelled accordingly. One reason is that it is hoped that valuable data for comparative reconstruction of proto-Khoe might be preserved. The other reason is that it is hoped that some of the culture-specific words are not replaced by modernisms but may be revived and thus saved from oblivion — like the glossonym *Khoekhoegowab* itself.

From its inception in 1981 to 1989 the *KhD* was compiled on handwritten index cards (with a back-up set), to be eventually typed after completion of all index cards. In 1988, when I was engaged in doctoral studies in London, I met

the phonologist and programmer Professor Jonathan Kaye at the School for Oriental and African Studies. When I approached him about my needs for customised lexicographic software, he immediately consented to develop such software.

The main reasons why I needed *customised* software was that

— lemmas were to be sorted by phonemes, rather than by letters; i.e. phonemes represented by polygraphs, such as clicks and their releases, should have their stipulated position in the alphabet;

— the Lepsius click symbols (/, ǁ, !, ǂ) were to be integrated into the officially stipulated alphabetising sequence;

— lemmas were to be sorted according to four tones, after they had been sorted by segmental phonemes.

Kaye's concept provided for two main stages, strictly distinct from each other: firstly, a **pre-dictionary stage** in which the data would be entered into a database by means of an off-the-shelf generic database management program. Secondly, a **dictionary compilation stage** in which these data are converted into a print-ready presentation layout format by means of compilation software using a conventional word processor and software that he would write. Separate compilation programmes serve the compilation of the *Khoekhoegowab–English Dictionary*, the conversion into a *Khoekhoegowab–English Glossary*, and the reversal to an *English–Khoekhoegowab Index* or *Glossary*. The strength of this concept rests in the provision that the data is to be stored in one common database file, rather than entering it into one or more text files from the start. The fact that this strategy was adopted in 1989 is evidence of farsighted programming. It made a pivotal difference to the project.

The database management program to be used was *File Express*; the word processor *Word Perfect 6.0a* (mainly because of its superior macro facilities in those early years of Windows platforms), and the programming language was *Icon*.

3. The Configuration of the *FE* Database Format for *NDP5*

File Express (***FE***) is described by its producers *Expressware* as flat-file database management software with basic relational lookup facilities, first published in 1984; it uses a DOS platform and was written for the IBM Personal Computer. In a flat-file database the data are organised in a single two-dimensional matrix consisting of rows and columns, as opposed to a relational database, where further tables may be incorporated in the main table in a hierarchical structure. A typical flat-file application would be an address-book in tabular format, with the data for each person filling one row, while

information *categories* like street name, town, postal code, etc. would be assigned to respective columns.

FE was, at its time, considered to be the "friendliest, most intuitive database program available anywhere, at any price", according to the prolific freelance author on computer matters, Alfred Glossbrenner. This user friendliness manifests itself right away in the arrangement that the primary user interface is not the tabular database layout with rows and columns. Rather, an input screen or Form is displayed which transparently presents one record at a time, for easy entering or editing of data. The layout of the form is designed by the user. It is one of the credits of *FE* that even novices without programming skills can configure the structure of the database by simply designing a Form. It was not common in those days that software would allow novices to create a database without the assistance of a programmer.

The configuration of this Form is of utmost importance as it is this design of the database format that determines what categories of information can be retrieved and hence, what determines the extent of multifunctionality of the database. The header file of the *Nama Dictionary Project* uses the acronym "NDP**5**". This indicates that it was only the fifth version of the Form layout that was found to be satisfactory. It was progressively improved by trial and error.

In the record format a record is made up of "Fields" that contain specific types of information (in the case of the *NDP* displayed in horizontal lines; cf. also Figure 6 below). These fields of the Form display the data of the respective *column* of the table format.

Figure 1 displays the configuration of a record as defined on the database definition screen for the *NDP*. This set-up amounts to a Form in current database software like *OO BASE*. 18 fields were specified, each identified by a number and an arbitrary "field name" that serves as descriptive mnemonic for the user.

Only the fields displayed in bold print (F2, F4, F5 and F7-F13) eventually appear in the compiled dictionary. The other fields have housekeeping functions. The sequence of data types within a record was already determined by the ordering of the fields in the present Form. Of interest here is only the column for –length–. In "length" the field length has to be specified (in bytes), as *FE* uses fixed-length fields. Hence it is important in the interest of saving data volume, to configure a field as short as possible. One record in the above layout has a length of 563 bytes, irrespective of whether the space has been maximally populated or not. The complete data file of the *NDP* as used for the *Dictionary* had a size of just under 13.5MB. More will be said later about the saving of field length through the use of codes. All fields of a record need manual completion, but need not be obligatorily populated, as long as at least one field has data.

-field name-	-type-	-length-	-format-
1 KEY ENTRY	C	15	Character
2 ENTRY	**C**	**80**	**Character**
3 ENTRY ALLOC	C	30	Character
4 EXAMPLE	**C**	**50**	**Character**
5 PRONUNC	**C**	**50**	**Character**
6 CATCHWORD	C	22	Character
7 PT SPEECH	**C**	**2**	**Character**
8 LABEL	**C**	**2**	**Character**
9 AFR RENDERING	**C**	**130**	**Character**
10 SCIENTIFIC	**C**	**40**	**Character**
11 ADD COMMENT	**C**	**3**	**Character**
12 X-REFERENCE	**C**	**33**	**Character**
13 LOAN	**C**	**32**	**Character**
14 SOURCE	C	4	Character
15 TONE CONSTIT	C	16	Character
16 ENGL ENTRY	C	45	Character
17 NOTES?	C	2	Character
18 STAGE/DATE	C	6	Character

Record Length: 563 bytes

Figure 1: Database Specifications for a Record in *NDP5*

It should be kept in mind that this database structure was conceived in 1989 primarily, if not solely, with the intention to compile a bilingual Khoekhoe–English dictionary with tone marking. Initially the ideal of multifunctionality did not feature prominently in the design, other than using the lexicographic data for a tonological analysis. After all, the concept of multifunctionality was, at that stage, not very prevalent in lexicography in general. Once the compilation software for the *NDP*, i.e. *NDP5* had been finalised by 1992 it was no longer possible to add further fields to the *FE* database, as a change of field number would have lead to mismatches in the compilation software (unless the field was non-printable and was added after the last field that was addressed by the compilation software). The only possibility to extend options of information generation that does not involve major surgery is to add further metalinguistic codes that would be accommodated in existing fields. Examples will be presented in the next section.

The fields are now briefly introduced.

F1, *KEY ENTRY*, merely serves for sorting to subsume all records into one article under the main lemma as specified in F1. The contents of F1 do not appear in print. Any record that has an empty F1 serves as main lemma. The Khoekhoe–English dictionary contains 4 273 main lemmas, i.e. 4 273 articles.

F2, *ENTRY*, contains the Khoekhoe lemma to be printed, unless the record contains an example in F4, in which case F2 is empty. F2 and F4 thus are mutually exclusive. F2 is specified for citation tone. The mode of tone codification will be explained further on. Dialect variants are marked in the running text with one of the following labels:

[Bz]	Bondelzwarts	**[V]**	Vaalgras Nama
[D]	Damara	**[T]**	Topnaar, ǂAonîn
[Hm]	Hai‖om	**[ǂA]**	ǂĀkhoe
[S]	Sesfontein Damara	**[ǂD]**	ǂAodama

Figure 2: Labels for Dialect Variants

Data for a specific dialect can thus be extracted by searching for these labels.

F3, *ENTRY ALLOCATION*, is perhaps a misnomer, as more aptly it should have been called Example Allocation. It specifies the lemma (F2) that an example as provided in F4 has to illustrate, i.e. that F4 has to follow in the sort during compilation. F3 thus is not printed. As the lemma of F2 in turn is assigned to a head lemma by means of F1, all records from main lemma to sub-lemmas with their examples will be assigned to the same article in ordered fashion.

F4, *EXAMPLE*, contains example sentences to illustrate mainly usage (esp. verb valency) or collocations. 2 703 records (11%) of the total of 24 520 Khoekhoe records are examples.

F5, *PRONUNCiation*, provides pronunciation of a lemma if the standardised spelling (in F2) does not readily reflect it; cf. Figure 3.

Pre-dictionary database input:

F2 ENTRY	aama
F5 PRONUNC	**(pronunc.: a2a2mâ3a2)**
F7 PT SPEECH	t
F9 RENDERING	(go on) pub-crawl, ...

Dictionary output:

aama {āma} (pronunc.: ààmáˆà) *v.t* (go on) pub-crawl,

Figure 3: Information on Pronunciation

F6, *CATCHWORD*, is a very important retrieval tool. It represents lemmas of F2 without tone marking, so that the record can be searched for in a simple query without the tone pattern having to be known — which is the normal situation when one wants to consult a Khoekhoe entry. F6 is not printed.

F7, *PT SPEECH*, specifies the word category of lemmas. As certain abbreviations contain up to eight characters, e.g. *v.i.stat* (stative intransitive verb), all abbreviations are assigned a pre-ordained code consisting of maximally two characters, so as to save space in the fixed-length records. These codes of the database are automatically replaced with the full abbreviation by the dictionary compilation software. While codes were used to save space, they — as a significant spin-off — bring the advantage that the abbreviations that the codes are automatically replaced with will be absolutely consistent with no scope for human error.

F8, *LABEL*, specifies usage labels. Again, preordained codes of maximally two characters are used, so as to save space (while ensuring consistency). The following labels (in italics) appear in print after the conversion:

arch.	archaic	*lit.*	literal
bot.	botanical	*math.*	mathematics
colloq.	colloquial	*med.*	medical
derog.	derogatory	*meteor.*	meteorology
did.	didactic	*mil.*	military
dign.	dignified	*mod.*	modern
euphem.	euphemism	*mus.*	music(al)
exagg.	exaggeration	*myth.*	mythol./myth.
fig.	figurative	*obs.*	obsolete
gen.	general	*obsc.*	obscene
geom.	geometry	*ornith.*	ornithology
hort.	horticulture	*phys.*	physiology
id.	idiomatic	*poet.*	poetry/poetical
inf.	informal	*pr.n.*	praise name
joc.	jocular	*vulg.*	vulgar
jur.	juridicial	*zool.*	zoology
ling.	linguistics		

Figure 4: Usage labels

Obviously any of these labels can be summoned to create a report for some kind of semantic or stylistic investigation, e.g. to create a list of zoological

terms.

F9, *RENDERING*, contains the English ready equivalents. Different senses of a lemma are not structured into different fields of sub-fields of a record. Rather, each sense is numbered and accommodated in a separate record. The sorting procedure will, during compilation, sequence these senses in succession in the article. The following extract from an article is, for instance, compiled out of three records:

> **!gãũ**^{1} {**!gâu**} *v.t/i* 1 cross (*e.g.* river, street); go across (*e.g.* field); s.a. ǂHǏǏ; 2 *fig.* bec. drunk/intoxicated/inebriated, s.a. |HÒRŐ5; 3 ferment.

Should one, for whatever purposes, wish to extract a report on Khoekhoe words with multiple senses, one can simply submit a query searching for the figure "1" in F9.

F10, *SCIENTIFIC*, contains the scientific names for zoological, ornithological or botanical words.

F11, *ADDitional COMMENT*, provides a slot for encyclopedic information, for instance about cultural practices.

F12, *X-REFERENCE*, accommodates cross-references. More will be said about this later.

F13, *LOAN*, provides sources of loan words, occasionally also recipients of Khoekhoe words, or it draws attention to calques; e.g.

> **pűrúkhòè**b … (< Dutch pl. broeken),[3] or
> **ànídànà** … (cf. Afr. hoenderkop).

Needless to say, as a spin-off this field provides a rich source of material for a discussion of language contact or loaning. F13 is the last field that provides text for the printing template.

F14, *SOURCE*, merely serves for "household" purposes, as it refers to the source of a lemma for purposes of verification. This field, like all subsequent fields, is not printed.

F15, *TONE CONSTITuents*, provides information in numerical form on the underlying tone of lemmas, as well as on tone rules that apply to a compound lemma. This field is of utmost importance to the data base and has provided the tool for analysing the tonological system of Khoekhoegowab, as will be further discussed below.

F16, *ENGLish ENTRY*, provides a hand-picked selection of key words from the English renderings provided in F9. Each of these entries in F16 will serve for the reversal of Khoekhoe–English to English–Khoekhoe in the compilation program, that is, each respective English word in F16 will be paired with the Khoekhoe contents of F2, *ENTRY* to form a lemma in the English–Khoekhoe Index. The contents of F16 will thus not appear in print in the Khoekhoe–English *Dictionary*, only in the English–Khoekhoe *Index* or *Glossary* as separate entries in alphabetically ordered sequence.

F17, *NOTES?*, is a general purpose household field in which the editor can *ad libitum* provide for the identification of various types of data, but all represented by a preordained code of maximally two characters. The metalinguistic notes that were made mostly convey linguistic or cultural information in order to allow the extraction of relevant data. This field also contains notes concerning the stage of editing, e.g., if a term needed further investigation or discussion with certain key consultants.

The following list of codes (Figure 5) gives an idea what kind of information can be retrieved at the present stage (the choice of mnemonic label is trivial). As said, further codes can be added *ad libitum* in this field in future, as they do not affect the basic configuration of the database:

Phonotactic information:

s	segmental peculiarities/variants
v	variants (regional or dialectal), e.g. *!a**a**b/!a**e**b* (river)
r	historical reduction (by elision of segments, esp. CV**C**V > CVV)
i	insertion (epenthesis, vowel anticipation, inversion), e.g. *horaga/ ho**a**raga* (whole)
3	trisyllabic root

Tonological information:

f	derivation of deverbal noun by means of flip-flop

Lexical information:

p	incorporation (of object noun into verb)
o	incorporation (of postposition/postpositional phrase into verb)
d	ideophone
m	idiomatic usage
g	"ergative" pair

Editing tasks:

?	problematic, needs further attention
z	scientific identification required
=	re-checking completed
x, e	manual editing required in Word Perfect file
T	enquire with Topnaars

Figure 5: Labels for household purposes

F18, *STAGE/DATE* provides the month and year when a record was discussed by the team. The date is inserted manually, as this field also provides for some codes indicating the stage of completion. These codes, as also the relevant codes of F17, allowed final rounds of editing to deal with remaining problems.

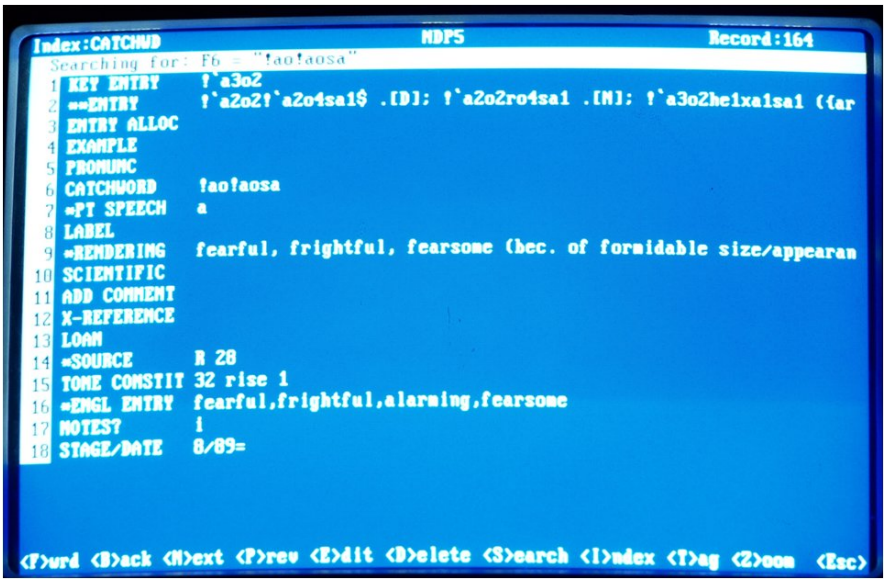

Figure 6: Photo of a Record as displayed on the *File Express* Database Screen

Figure 6 presents an example of a record as displayed in the *NDP5* configuration of *File Express* followed by Figure 7 displaying an extract of the finally compiled dictionary with the same record (underlined) in context:

intimidation; **!àò!àò-àò**b/s, **~rò-àò**b/s {**!ao!ao-ao**b/s, **~ro-ao**b/s} *n.* intimidator; s.o. who frightens/*etc.*; **!àò!àò(hè)sà**, **~rò(hè)sà** {**!ao!ao(he)-sa**, **~ro(he)sa**} *a.* (wh. h.b.) frightened/scared; **!àò!áósà** [D]; **!àòrósà** [N]; **!áòhèxàsà** (arch.) {**!ao!aosa**; **!aorosa**; **!aohexasa**} *a.* fearful, frightful, fearsome (bec. of formidable size/appearance); alarming; **!àò!áó-sàsì**b [D]; **!àòrósàsì**b [N]; **!áò(hè)sàsì**b {**!ao!aosasi**b; **!aorosasi**b; **!ao(he)sasi**b} *n.* fearfulness, frightfulness, fearsomeness; ...

Figure 7: The same Record (Figure 6) converted by the Compilation Software into a Lemma within an Article

4. The Benefits of a Database

In this section some of the major benefits of the Khoekhoe database are introduced.

4.1 The Query System

FE uses an extremely user-friendly and intuitive syntax for queries. There is no need to use standard SQL. It is a conversational program as natural language can be used to a large extent, e.g. "F1 is not empty" (which would yield all records that are not sub-lemmas, thus main lemmas). The beginning of the query is provided on screen: "*Find all records where:*" (for finding individual records) or "*Print all records where:*" (for producing reports) and the user must complete it with a minimal adherence to formulas, e.g. "F8 = z" or "F8 is z" would yield all records that are labelled as zoological terms.[4]

The power of database queries, as opposed to linear word processor searches, lies in the possibility to combine searches in *multiple* fields; e.g. *Print all records where:* "F6 = "..si" and F8 = "t" and F8 is not "o".[5] This query would generate a report listing all Khoekhoe transitive verbs (in the database) ending with the suffix "*-si*" that are not marked as "obsolete". The example should convey an impression what a powerful resource for especially but not only morphological investigation the database is. It is possible to combine more than two fields in a complex query.

4.2 Instances of further data utilisation of the *NDP*

Data can be selectively utilised for all sorts of publications on specific domains, ranging from smaller articles to books and different versions of the *KhD*. Below (4.2.2) it will be expounded on how the present Khoekhoegowab-oriented database has already been used for three dictionaries/glossaries, each with reversal of source and target languages. Currently the database is being extended to some endangered dialects for a second, enlarged edition (4.2.3).

An instance of a smaller publication was a list of some 350 botanical names: *A preliminary list of Khoekhoe (Nama/Damara) plant names* (Eiseb, Giess and Haacke 1991). Such intermediary publications may furthermore be useful for satisfying funding agencies that a project produces tangible results, even if the final dictionary is still years away.

Lexicographic databases are attractive to builders of computer applications of various kinds. The *NDP5* database is currently used by a postgraduate student in South Africa to extract morphological data for a Master's thesis on the development of a morphological parser for Khoekhoegowab.[6]

A Khoekhoe spell checker is an obvious utility that should emanate from this electronic database. At this stage, however, there are no substantive agreements with experts yet.

Plans are under way to use the *NDP5* database extensively for writing a reference grammar of Khoekhoegowab. The *NDP5* database provides the most comprehensive Khoekhoe corpus on morphological information that exists. Data have been extracted for university study guides on Khoekhoegowab grammar, and have been supplied to other researchers for their purposes.

4.2.1 Tonological Analysis of Khoekhoegowab

The *KhD* was to be marked for tone, as Khoekhoegowab is a fully-fledged tone language. Impressionistic annotation of words with diacritical tone marks, was out of the question, however, as a tonological system is as systematic and rule governed as the segmental phonology of a language. Hence a systematic investigation of the tonology of Khoekhoe was a prerequisite for marking tone in the dictionary. On the other hand, the lexicon of the database was to provide the data for such a study. It is for this reason that the field F15, *TONE CONSTITuents*, was introduced. This field, which is not printed, lists the underlying lexical tone melodies that serve as input to compound lemmas in F2, ENTRY; e.g. Figure 8:

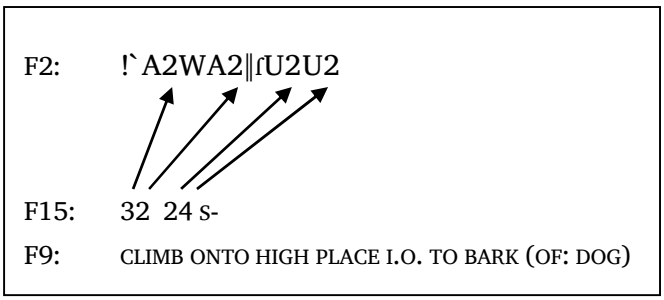

Figure 8: Tonal Input (F15) and Output (F2) for *!àwà ǁhùù*

The four tonemes that were postulated for Khoekhoegowab are marked in the *FE* database by Arabic numbers from 1 to 4 for the lowest to the highest toneme respectively, and the tone number follows on the t.b.u. (as ASCII codes do not provide for diacritics). F15 in the above example (Figure 8) indicates that the two disyllabic roots *!áwà* (climb) and *ǁhùúí* (bark) underlyingly have the bimoraic tone melodies 32 and 24 respectively. The codes "s" and "-" in F15 respectively indicate that in the compound as provided in F2 the first root undergoes a "**s**witching" rule (better known as "flip-flop", by analogy to Chinese), and that the second root changes from the citation melody (+) to a sandhi melody (-).[7] The correlation of the tonal input and output information in F15 and F2 was instrumental for the analysis of the lexical tonology by way of a doctoral thesis (Haacke 1999). No entry in F2 was marked for tone unless the relevant rule application was understood. In this way the reliability of the tonal marking was significantly enhanced. The metalinguistic codes for tone rules in F15 furthermore permit queries that provide all instances of a particular rule application (or combinations thereof) contained in the database. Such a retrieval is not possible with a word processor.

The device to enter tones as numeric characters in the database moreover

allows for sorting of homographs in consistent tonal sequences. In the compilation program a specific routine moves all numeric characters to the end of the word, before character-by-character sorting is done, e.g. *o2ra2* > *ora**22*** and *o1ra3* > *ora**13***. This causes *o1ra3* (***òrá*** "eat raw") to be placed before *o2ra2* (***òrà*** "raw"). After sorting, the conversion to diacritics is done on the original entry.

Again, the possibility to combine different fields in one query was of utmost importance in the investigation of tonology, so as to establish in optimal detail the interrelation between the segmental and suprasegmental phonology. A query like

"F2 = "m.." and F15 = "4..""

would, for instance, generate a report listing all lemmas that have "m" as initial consonant/segment and a "double high" toneme "4" as first tone, i.e. on the t.b.u. following the specified first consonant; cf. Figure 9:

Find all records where:
F2 = "m.." and F15 = "4.."
i.e.
Find all records where:
the first segment is "m" and the first tone is "4".

Figure 9: Correlating Segmental with Tonal Data to establish Depressor Consonants

The following Table from Haacke (1999: 56), Figure 10, presents a synoptic overview of the quantitative distribution of C_1 of roots with the six major tonal melodies (listed in the top line) of Khoekhoegowab in the 2 150 roots that were in the database at that stage. This list conveys crucial data of the phenomenon that Khoekhoegowab has depressor consonants that lower existing tonal melodies: The depressor consonant ***h***, ***m***, ***kh***, ***ts*** and the click (X) releases X***h***, and X***n*** depress the tonal melodies /22/ and /32/ to /13/ and /12/ respectively (instances underlined). This phenomenon of pitch-evolution through phonetically conditioned tone splits is known as tonogenesis — a phenomenon prevalent in *i.a.* Sino-Tibetan languages. The present database was instrumental in investigating to what extent tonogenesis occurs in Khoekhoegowab (and, hence, is likely to be found in related Khoe languages); it allows retrieval of every supporting example but also every counter example, and the record counting facility allows for statistical assessment of the frequency of instances.

	/12/	/13/	/22/	/32/	/24/	/43/	Total
b/p	0	7	1	0	2	1	*11*
d/t	12	16	2	8	12	5	*55*
g/k	12	15	6	9	12	10	*64*
ʔ(V)	3	13	11	15	25	13	*80*
h	**5**	**9**	**0**	**0**	6	7	*27*
m	**3**	**3**	**0**	**0**	4	3	*13*
n	5	5	5	2	12	4	*33*
s	**1**	**12**	**12**	**14**	21	13	*73*
x	**0**	**5**	**3**	**14**	19	14	*55*
kh	**14**	**17**	**0**	**0**	3	7	*41*
ts	**19**	**10**	**0**	**0**	14	8	*51*
Total:	*74*	*112*	*40*	*62*	*130*	*85*	***503***
ǀ	**0**	**6**	**20**	**21**	36	17	*100*
ǀg	22	16	3	12	15	12	*80*
ǀh	**31**	**24**	**1**	**2**	26	16	*100*
ǀkh	**0**	**4**	**8**	**15**	26	23	*76*
ǀn	**8**	**14**	**0**	**1**	15	8	*46*
Total:	*61*	*64*	*32*	*51*	*118*	*76*	***402***
ǁ	**0**	**6**	**16**	**21**	29	15	*87*
ǁg	14	18	10	15	24	12	*93*
ǁh	**19**	**23**	**0**	**0**	17	7	*66*
ǁkh	**0**	**1**	**16**	**17**	24	15	*73*
ǁn	***14***	***14***	**1**	**5**	21	13	*68*
Total:	*47*	*62*	*43*	*58*	*115*	*62*	***387***
!	**1**	**10**	**21**	**25**	34	9	*100*
!g	25	28	9	18	23	14	*117*
!h	**28**	**26**	**1**	**3**	16	19	*93*
!kh	**0**	**4**	**12**	**22**	30	15	*83*
!n	**25**	**24**	**1**	**11**	30	17	*108*
Total:	*79*	*92*	*44*	*79*	*133*	*74*	***501***
ǂ	**0**	**8**	**13**	**16**	28	11	*76*
ǂg	13	18	12	9	21	13	*86*
ǂh	**25**	**18**	**0**	**2**	14	18	*77*
ǂkh	**0**	**0**	**21**	**14**	15	6	*56*
ǂn	**17**	**15**	**1**	**3**	15	11	*62*
Total:	*55*	*59*	*47*	*44*	*93*	*59*	***357***
Grand Total:	***317***	***389***	***206***	***294***	***588***	***356***	***2 150***
Percentage:	**14,7**	**18,1**	**9,6**	**13,7**	**27,3**	**16,5**	**100%**
		27,7					
		28,4					

Figure 10: Distribution of C_1 with the Six Major Tonal Melodies in Disyllabic Skeleta

4.2.2 Tagging of Records for producing different versions of the dictionary

Databases are re-usable for other kinds of dictionaries or reference works. This possibility is enhanced by *FE*'s facility to tag records, i.e. to earmark them in an *ad hoc* way for a particular purpose. This facility was used for extracting the 10 150 records that have served as input to the Khoekhoegowab–English part of the *Khoekhoegowab–English/English–Khoekhoegowab Glossary/Mîdi Saogub* (Haacke and Eiseb 1999), and subsequently to the *Khoekhoegowab–Afrikaans Afrikaans–Khoekhoegowab Glossarium/Mîdi Saogub* (Haacke, Eiseb and Gericke 2010).

These glossaries are important by-products that had not been planned for initially, but could be created thanks to the software design. When the first trial page of the *KhD* in book format was produced by means of the new software, co-author Eliphas Eiseb was of the opinion that the Khoekhoe speakers would not use the dictionary, as it was too complicated. The tone marking with diacritics, and also the arrangement of lemmas in articles would be disconcerting for users with limited referencing skills. The project leader took heed of his opinion and requested the programmer, Jonathan Kaye to write another compilation program that would produce a Khoekhoe–English glossary without tone marking, using just the standard orthography. This book is now used in schools by MT-speakers. This solution of providing for two books derived from one database satisfactorily addresses the rather ambitious if not unrealistic aim to cater for a wide-fanned spectrum of users ranging from moderately literate MT-speakers to academics.

A further by-product that resulted from the tagged records via the extracted database of the *Glossary* is the *Khoekhoegowab–Afrikaans Afrikaans–Khoekhoegowab Glossarium/Mîdi Saogub*, which was commissioned by PANSALB (*Pan South African Language Board*) for use by the Khoekhoe descendants in South Africa. The commission was to replace English with Afrikaans. As by then contact with Jonathan Kaye had been lost, who had moved on from London to China, the compilers were obliged to use the existing database configuration, since the compilation software is linked to specific fields of it. Hence improvisation was called for:

The database with its 18 fields was cloned and

— the English data of F9 RENDERING transferred to the non-printable F11, which in the original database has the ADDitional COMMENT;

— the data of the original F16 ENGL ENTRY transferred to the non-printable F15, which originally has the TONE CONSTITuents.

This freed F9 for entering the AFR RENDERING and F16 for the AFR ENTRY to accommodate the key words for the Afrikaans–Khoekhoe reversal. The contents of the original F11 and F16 could be dispensed with in the compilation of the *Glossarium*, as this information was not needed. As the *Glossarium* was intended for a target group that wants to re-acquire the Khoekhoe language, it

was imperative, however, that tonal information was supplied somewhere. This required that the tonal decodification was re-activated in the macro that produces the Khoekhoe–Afrikaans part of the *Glossarium*, while the Afrikaans–Khoekhoegowab part was kept in the standard orthography without tone marking, as in the original *Khoekhoegowab–English Glossary.*

-field name-	-type-	-length-	-format-
1 KEY ENTRY	C	15	Character
2 ENTRY	**C**	**80**	**Character**
3 ENTRY ALLOC	C	30	Character
4 EXAMPLE	**C**	**50**	**Character**
5 PRONUNC	**C**	**50**	**Character**
6 CATCHWORD	C	22	Character
7 PT SPEECH	**C**	**2**	**Character**
8 LABEL	**C**	**2**	**Character**
9 AFR RENDERING	**C**	**130**	**Character**
10 SCIENTIFIC	**C**	**40**	**Character**
11 RENDERING	C	130	Character
12 X-REFERENCE	**C**	**33**	**Character**
13 LOAN	**C**	**32**	**Character**
14 SOURCE	C	4	Character
15 ENGL ENTRY	C	16	Character
16 AFR ENTRY	C	45	Character
17 NOTES?	C	2	Character
18 STAGE/DATE	C	6	Character

Figure 11: Database Modification for the *Glossarium*

4.2.3 New Uses: Updating and Expansion of Dialect Data

Haacke has for several years now been working on eliciting dialect equivalents for over 6 785 English lemmas of the *Glossary.* The dialects concerned are endangered dialects on the northern and northwestern periphery of Khoekhoegowab, viz. ǂĀkhoe and Haiǁom, as well as Sesfontein Damara in Kaokoland. These dialect variants, which are to be integrated into the main database, are to be added to a second, enlarged edition of the *KhD*. Obviously such data might also find their way into further publications like a comparative dictionary.

4.3 Editorial Advantages of the Database

The editorial advantages of a database are immeasurable. It is hard to say whether the main *KhD* with over 24 000 entries would ever have been published if no dedicated software had become available. The reversal of the Khoekhoe–English dictionary to English–Khoekhoe — albeit in *Index* form — would, because of time constraints, not have taken place. For certain, the late co-author Eliphas Eiseb would never have seen the fruit of his life's vocation, and certainly neither of the two glossaries would ever have appeared.

The editorial advantages of a database can only be identified very briefly here:

4.3.1 Reversal of Source and Target Language

The automatic reversal of source and target language, i.e. in our case Khoekhoegowab–English to English–Khoekhoegowab, obviously is one of the most compelling reasons for using a Dictionary Writing System. The reversal lists the English (or Afrikaans) lemmas that were identified in F16, in alphabetical order with the entire contents of F2 ENTRY or F4 EXAMPLE attached, whichever the case may be. As there will be numerous repetitions in F16 of specific English lemmas, but with different associations of F2 or F4, careful editing is required to conflate such multiple entries wherever the meaning warrants it. As this automatic reversal is bound to be riddled with lacunae of English concepts that do not occur in the English rendering (F9, F16) of Khoekhoe concepts — for no Khoekhoe terms for English concepts were coined by the authors, it is called an *Index*. Ideally users should follow an onomasiological approach by also consulting the Khoekhoe–English *Dictionary* for each English lemma, to get a more contextualised understanding of the Khoekhoe word.

4.3.2 Automatic Conversion from Tonal to Standard Orthography

As it is an essential feature of the database that all lemmas are marked for tone, the standard orthography had to be systematically adapted so as to provide the required tone-bearing units by replacing single vowels marked with a macron for length, with two identical vowels (as actually is historically correct, moreover). Each vowel, being a syllable peak, then bears a tone mark. The compilation software for the *KhD* automatically provides the version in standard spelling without tone marks after the entry with tone marks, indicated by curly brackets; e.g.

> **áà**b/-i {**ā**b/-i} *n.* drink, beverage.

This was considered an essential aid for the less versatile user. In the *Glossary* the compilation software replaces the tone-marked entry and provides solely

the versions in standard orthography, both, in the Khoekhoegowab–English part and the English–Khoekhoegowab part, e.g.

āb/-i *n.* drink, beverage.

These automatisation processes are a very significant editorial aid in that they not only save immense amounts of time but also preclude human error in transcribing.

4.3.3 Speeding up editing and proofing

Using a database is a pivotal device of drastically reducing human error and inconsistency in editing.

A typographical error in spelling or tone assignment in the raw corpus data will more often than not be revealed by a resultant missort of the record in the database, which by itself usually is conspicuous. Proofing of spelling thus happens *en passant* over the years of data compilation and should have been largely concluded by the time the dictionary compilation stage is initiated.

Checking of editorial matters like bolding, italicisation, metalinguistic labels and abbreviations, numbering of sub-senses or of homographs, can be done systematically by filtering out the form fields (or spreadsheet columns) concerned. The fact that the preordained codes in the database are automatically replaced by the metalinguistic labels and abbreviations in the dictionary, already ensures consistency, unless a faulty code was typed for a start and hence will not convert. As *FE* does not allow selective use of font styles or diacritics and is limited to ASCII type characters, such editorial choices have to be configured by preordained coding in the database, so that they can be automatically converted by the writing program during the dictionary compilation phase. Italics, for instance, were indicated in the database by curly brackets.

Pre-dictionary database input:	
F10 SCIENTIFIC:	{Olea europea} subsp. {africana}
Dictionary output:	
	Olea europea subsp. *africana*

Figure 12: Italics in Scientific Names

The use of a database form will ensure adherence to the style manual and guide the analysis of a record; in short, it will significantly enhance data integrity. Detecting all such inconsistencies and omissions in dictionary articles that

were drafted with a word processor is a Sisyphean task indeed. When I once replied to the question how often I had proofread the print-version of the *KhD*, that I had proofread it only once because of time constraints, I was looked at silently in disbelief. I had to explain that the use of a database that had been fine-tuned over years had largely obviated proofreading of the compiled dictionary file, other than for checking that the compilation software had not slipped up on editorial matters.

4.4 Limitations of the *NDP5* Dictionary Writing System

It should be emphasised at the outset that the DWS as custom-designed by Jonathan Kaye in the early 90s was never intended as a commercial application. It was designed exclusively to meet my requirements as specified earlier, at a time when no commercially available software would have met my particular requirements. So there was never an intention to make it commercially viable. However, for the sake of objectivity a superficial comparison to current state-of-the-art expectations of DWSs as set out by, for instance, Joffe and De Schryver (2004) and by De Schryver and De Pauw (2007) may be opportune.

It was stated before that the software does not provide for a Corpus Query Package as at the time of its creation electronic text corpora hardly existed in Khoekhoegowab.

Most outstanding is the absolute separation of the pre-dictionary database stage and the dictionary compilation phase. While working in the database there is no possibility to obtain a WYSIWIG preview of a lemma as it will appear in print.

The software does not provide for automatic tracking and updating of cross-references. Checking that all cross-reference links are correct — let alone exist, consumes a considerable amount of time and invites human error. The *KhD* contains 3 706 cross-references. They needed manual checking of the paired records — a task best done on two PCs standing next to each other.

The software thus provides no automated controls for editing and consistency checks. Yet — as pointed out above — the actual editing is significantly facilitated by taking place in the database instead of in the final dictionary text, as would be the case when using a word processor.

Having pointed out the above limitations I wish to come to the cardinal advantage of the ***NDP*** software: that the ageing legacy software of the early 1990s has been replaced by open source software that should ensure the long-term survival of the primary database, while essentially still following the same basic concepts of the original database configuration and compilation software.

5. Conclusion: Transition to Open Source Software

In 2001 a contractual agreement was set up between PANSALB, the Publisher

and Haacke that the *Khoekhoegowab–English/English–Khoekhoegowab Glossary/Mîdi Saogub* should be converted to a *Khoekhoegowab–Afrikaans Afrikaans–Khoekhoegowab Glossarium/Mîdi Saogub* (see above). When in 2007 the project was ready to embark on the Dictionary Compilation Stage so as to convert the modified database into print-ready format, disaster struck: The custom-made software failed to perform because of its dependence on outmoded operating systems. Unless we could get it to work, the project was doomed to failure and the converted database of no use. I was fortunate to re-establish contact with Jonathan Kaye, who by then had abandoned Windows for Linux (Debian). In an almost daily shuttle of files per email over three months he admirably converted the DOS database into the camera-ready Word Perfect dictionary format in a multitude of ad hoc measures employing LINUX, thanks to his advanced programming skills.

The all too obvious lesson of the crisis was that the data had to be rescued into non-proprietary software before it would become inaccessible through the advance of Microsoft operating systems. Again Jonathan Kaye came to the rescue. He transferred the *File Express* database to an *OpenOffice CALC* spreadsheet and replaced the dictionary compilation procedures that were written with the aid of Icon and Word Perfect facilities originally, by formulas and macros that operate within CALC. ASCII characters were replaced by Unicode characters, which required that the Khoekhoe data are encoded differently now. In order to add and/or edit data for a next publication, as well as to retrieve data (the processes formerly done in *File Express*) the data have to be transferred from the *OpenOffice CALC* spreadsheet into an *OpenOffice BASE* database. *BASE* is a graphical front-end for accessing databases. It allows the execution of queries and reports, and the editing and adding of data by using a customised Form as user-interface. Forms are data input and output masks and are the actual editing tool. Once the pre-dictionary data compilation phase has eventually been completed, the data has to be transferred back into the *CALC* spreadsheet in order to compile the *Khoekhoegowab–English Dictionary* and the *English–Khoekhoegowab Index* in print-ready format.

The conversion of the database and dictionary compilation software from legacy formats to robust non-proprietary software subject to a GNU General Public License, thus has made possible if not ensured the survival of the only lexical corpus in database format of a Khoesaan language after the demise of the compilers. Apart from ensuring that more publications can be generated by the present proprietor of the database while it is not public domain as yet, the way is open to eventually subscribe to "best practice" aspirations. "Best practices", according to the E-MELD School of Best Practises, are "practices which are intended to make digital language documentation optimally long-lasting, accessible, and re-usable by other linguists and speakers".[8] This, to the best of my knowledge, is a challenge that state-of-the-art lexicographic software currently in vogue in South Africa and lexicographic projects in South Africa still have to face.

To conclude: This case history was presented in the hope to show that, firstly, lexicographers in this millennium need to avail themselves to dedicated lexicographic software with complex retrieval facilities, software that allows them to get an optimal return for their investment by ensuring that these usually formidable repositories of knowledge can be tapped not only for producing a dictionary but are reusable for other purposes.

Secondly, lexicographers should take heed of the current trend of following "best practices" in Language Documentation by resorting to open source software so as to ensure optimal survival of the multipurpose database, for coming generations to build on.[9]

Endnotes

1. As recently as 2004 Joffe and De Schryver (2004: 99) stated that "dedicated tools to assist lexicographers are not readily available".
2. *Khoekhoegowab* (literally: "*Khoekhoe*-language") is the revived original name of the language that in the late 19th century became better known as "Nama", and subsequently as "Nama/Damara" (cf. Haacke 2011). This glossonym was officially reinstated through the initiative of Eliphas Eiseb. For the sake of brevity the language is mostly referred to here as *Khoekhoe*.
3. The bolded vowels *oe* of the plural form of *purukh**oe**n*, (pairs of trousers), just as in the loan *rokh**oe**n* (skirts, < Du. *rokk**en***), indicate that these words are not loans from the more recent Afrikaans, but date back to first contacts of Khoekhoe with Europeans. These words in turn were borrowed from Khoekhoegowab by several Bantu languages in Southern Africa.
4. "z" is the metalinguistic code used in the database for the printed abbreviation "*zool.*".
5. As F6 lists lemmas without tone specifications, all relevant lemmas are listed irrespective of their tonal profile; two dots ".." represent a wildcard meaning "anything"; "t" is the metalinguistic code used for "*v.t.*", and "o" is the metalinguistic code for "*obs.*".
6. Hendrina du Plessis *Natural Language Processing of Khoekhoegowab (Previously Nama/Damara).* UNISA.
7. Sandhi: tonal modification by syntactic/morphological context.
8. **E_MELD:** ***E****lectronic* ***M****etastructure for* ***E****ndangered* ***L****anguages* ***D****ata*; http://emeld.org/school/what.html.
9. For an introduction to the relatively new discipline of Language Documentation cf. *i.a.* Himmelmann (2006).

References

De Schryver, G.-M. and G. de Pauw. 2007. Dictionary Writing System (DWS) + Corpus Query Package (CQP): The Case of *TshwaneLex*. *Lexikos* 17: 226-246.

Dickens, P. 1994. *English–Ju|'hoan Ju|'hoan–English Dictionary.* Quellen zur Khoisan-Forschung 8. Cologne: Köppe Verlag.

Eiseb, E., W. Giess and W. Haacke. 1991. A Preliminary List of Khoekhoe (Nama/Damara) Plant Names. *Dinteria* 21: 17-30.

Haacke, W.H.G. 1998. A Khoekhoe Dictionary in the Making: Some Lexicographic Considerations. Schladt, M. (Ed.). 1998. *Language, Identity, and Conceptualization among the Khoisan*: 35-64. Quellen zur Khoisan-Forschung 15. Cologne: Köppe Verlag.

Haacke, W.H.G. 1999. *The Tonology of Khoekhoe (Nama/Damara)*. Quellen zur Khoisan-Forschung 16. Cologne: Köppe Verlag.

Haacke, W.H.G. 2011. *Nama* als Sprachbenennung in der Koloniallinguistik Deutsch-Südwestafrikas: zwischen Endonym und Exonym. Stolz, Th., Chr. Vossmann and B. Dewein (Eds.). 2011. *Kolonialzeitliche Sprachforschung. Die Beschreibung afrikanischer und ozeanischer Sprachen zur Zeit der deutschen Kolonialherrschaft*: 139-160. Berlin: Akademie Verlag.

Haacke, W. and E. Eiseb. 1999. *Khoekhoegowab–English/English–Khoekhoegowab Glossary/Mîdi Saogub.* Windhoek: Gamsberg-Macmillan.

Haacke, W. and E. Eiseb. 2002. *A KHOEKHOEGOWAB DICTIONARY with an English–Khoekhoegowab Index.* Windhoek: Gamsberg-Macmillan.

Haacke, W., E. Eiseb and C. Gericke. 2010. *Khoekhoegowab–Afrikaans Afrikaans–Khoekhoegowab Glossarium/Mîdi Saogub.* Windhoek: Macmillan.

Himmelmann, N.P. 2006. Language documentation: What is it and what is it good for? Gippert, J., N.P. Himmelmann and U. Mosel. 2006. *Essentials of Language Documentation*: 1-30. Berlin: Mouton de Gruyter.

Joffe, D. and G.-M. de Schryver. 2004. Tshwanelex — A State-of-the-Art Dictionary Compilation Program. Williams, Geoffrey and Sandra Vessier (Eds.). 2004. *Proceedings of the Eleventh EURALEX International Congress, EURALEX 2004, Lorient, France, July 6–10, 2004*: 99-104. Lorient: Faculté des Lettres et des Sciences Humaines, Université de Bretagne Sud.

Kilian-Hatz, C. 2003. *Khwe Dictionary. With a Supplement on Khwe Place Names of West Caprivi by Matthias Brenzinger.* Namibian African Studies 7. Cologne: Köppe Verlag.

König, C. and B. Heine. 2008. *A Concise Dictionary of Northwestern !Xun.* Quellen zur Khoisan-Forschung 21. Cologne: Köppe Verlag.

Nakagawa, H., K. Sugawara and J. Tanaka. In preparation. *Ghui Dictionary.*

Traill, A. 1994, 2009. *A !Xóõ Dictionary*. Quellen zur Khoisan-Forschung 9. Cologne: Köppe Verlag.

Van Sterkenburg, P. (Ed.). 2003. *A Practical Guide to Lexicography*. Terminology and Lexicography Research and Practice 6. Amsterdam/Philadelpia: John Benjamins.

Visser, H. 2001. *Naro Dictionary: Naro–English, English–Naro*. Fourth edition. D'Kar: Kuru Development Trust.

Weich, F. 2004. *San Dictionary/San-woordeboek. San–Afrikaans–English/English–San–Afrikaans/Afrikaans–San–English*. Johannesburg: Macmillan South Africa.