

# TOWARDS THE DEVELOPMENT AND APPLICATION OF REPRESENTATIVE LEXICOGRAPHIC CORPORA FOR THE GABONESE LANGUAGES

Léandre Serge SOAMI

Dissertation presented for the Degree of Doctor of Literature (in Lexicography) at the  
University of Stellenbosch

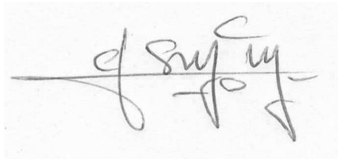


Promoter: Prof. R.H. GOUWS

March 2010

## Declaration

By submitting this dissertation electronically, I declare that the entirety of the work contained therein is my own, original work, that I am the owner of the copyright thereof (unless to the extent explicitly otherwise stated) and that I have not previously in its entirety or in part submitted it for obtaining any qualification.

A handwritten signature in black ink, appearing to read 'J. S. van der Merwe', written over a horizontal line.

Signature

22 February 2010

Date

Copyright © 2010 Stellenbosch University

All rights reserved

## **Abstract**

The compilation of dictionaries is a laborious activity and it takes time, money and staff to achieve the objectives of any dictionary project. Many dictionaries have been compiled using the lexicographers' personal intuition and guessing rather than being corpus based. That resulted in some dictionaries often being criticised by users because of the lack of representation of some important lexical items. This can probably be explained by the fact that most of these dictionaries were compiled in an era when theoretical lexicography was lacking or not well established. The last decades have witnessed the emergence of metalexicography as a theory directed also at dictionary planning in order to enhance the quality of lexicographic practice and the way in which the management and the compilation of dictionaries are dealt with. The planning of dictionaries takes into account not only the gathering of language material to be used but also the way in which this material will be treated and presented on both the macrostructural and the microstructural level as well as in the front matter texts and the back matter texts.

In order to enhance the quality of the presentation in dictionaries, this dissertation pleads in favour of the formulation of a data collection policy that takes into consideration all the different sources of material, written and spoken, used in the different phases of the compilation of a dictionary. The three phases that form the main focus of this study are the material acquisition phase, the material preparation phase and the material processing phase. The involvement of the speech community in the compilation of a lexicographic corpus ensures the collection of representative and balanced data, and the different needs of that community are central to the dictionary project. The different language materials can be organised into different corpus types.

The efficiency of a corpus resides in its capacity to provide different data types that can be included in the comment on semantics and the comment on form of each article in the central list of each dictionary. Some dictionaries lack a good representation of data in both these comments in the different articles. However, languages such as the Gabonese languages are in a privileged situation because they can still avoid the mistakes of other dictionary compilers by investing in corpus-based dictionaries at this early stage. Therefore, the establishment of lexicographic units

with multifunctional tasks can play an important role. In a multilingual environment such as Gabon the issue of language status needs to be dealt with carefully because it is realistic to choose a certain number of languages to function as official languages. Different alphabets are presented in this study and realistic choices are made.

The way in which the language material is organised will impact on the quality of the macrostructure and microstructure; this is essential because dictionaries are consulted most of the time for the spelling of a given lexical item, for a translation equivalent or for the explanation of the meaning of a lemma sign. The computerisation of a corpus is a focal point and needs to be done in a satisfactory manner that presents a clean and helpful corpus in order to provide the lexicographer with useful statistics, frequency word lists and the different concordance lines that are very important for the wording of definitions and the extraction of example sentences. This is why a corpus is seen as an indispensable tool in the improvement of the macro- and the microstructure of any type of dictionary.

## Opsomming

Die saamstel van woordeboeke is 'n moeisame aktiwiteit, en dit verg tyd, geld en personeel om die doelstellings van 'n woordeboekprojek te bereik. Talle woordeboeke is op grond van die navorsers se persoonlike intuïsie en raaiwerk saamgestel, in stede daarvan dat dit korpusgebaseerd is. Die gevolg is dat baie woordeboeke dikwels deur gebruikers gekritiseer word weens die gebrek aan verteenwoordiging van enkele belangrike leksikale items. Dít kan moontlik verklaar word deur die feit dat die meeste van hierdie woordeboeke saamgestel is in 'n era waartydens teoretiese leksikografie gebrekkig en nie goed gevestig was nie. In die afgelope dekades het metaleksikografie na vore getree as a teorie wat op woordeboekbeplanning gerig is ten einde die gehalte van die leksikografie-praktyk en die manier waarop die bestuur en samestelling van woordeboeke hanteer word, te verbeter. By die beplanning van woordeboeke word nie net die versameling taalmateriaal wat gebruik kan word in berekening gebring nie, maar ook die manier waarop hierdie materiaal op sowel makro- as mikrostrukturele vlakke, asook in die voorwerk en die agterwerk, hanteer en aangebied gaan word.

Ten einde die gehalte van die aanbieding in woordeboeke te verbeter, lewer hierdie proefskrif 'n pleidooi vir die formulering van 'n dataversamelingsbeleid wat al die verskillende materiaalbronne, hetsy skriftelik of mondelings, wat in die verskillende stadia van die samestelling van 'n woordeboek gebruik word, in ag neem. Die drie stadia wat die hooffokus van hierdie studie is, is die stadia waarin die materiaal aangeskaf, voorberei en verwerk word. Die spraakgemeenskap se betrokkenheid by die saamstel van 'n leksikografiese korpus verseker die versameling van verteenwoordigende en gebalanseerde data, en die verskillende behoeftes van sodanige gemeenskap is die kern van die woordeboekprojek. Die verskillende taalmateriale kan in verskillende korpussoorte georden word.

Die doeltreffendheid van 'n korpus berus op die vermoë daarvan om verskillende datasoorte te verskaf wat in die kommentaar op semantiek en die kommentaar op vorm van elke item in die sentrale lys van elke woordeboek ingesluit kan word. Sommige woordeboeke toon 'n gebrek aan goeie verteenwoordiging van data in albei hierdie soorte kommentaar in die verskillende items. Tale soos die Gaboenese tale is egter in 'n bevoorregte posisie, aangesien hulle nog die foute van ander

woordeboeksamestellers kan vermy deur op hierdie vroeë stadium in korpusgebaseerde woordeboeke te belê. Die stigting van leksikografiese eenhede met multifunksionele take kan dus 'n belangrike rol speel. In 'n veeltalige omgewing soos Gaboen moet die kwessie van taalstatus versigtig hanteer word, aangesien dit realisties is om 'n sekere hoeveelheid tale as amptelike tale te kies. Verskillende alfabette word in hierdie studie aangebied en realistiese keuses word gemaak.

Die manier waarop die taalmateriaal georden is, sal 'n uitwerking op die makro- en mikrostruktuur hê; dit is van belang omdat woordeboeke meestal vir die spelling van 'n gegewe leksikale item, vir 'n vertaalekwivalent of vir die verklaring van die betekenis van 'n lemmateken geraadpleeg word. Die rekenarisering van 'n korpus is 'n belangrike aspek en moet op 'n bevredigende wyse uitgevoer word wat 'n skoon en nuttige korpus lewer ten einde die leksikograaf van goeie statistieke, frekwensiewoordlyste en die verskillende konkordansielyne te voorsien, wat baie belangrik is vir die skryf van definisies en die onttrekking van voorbeeldsinne. Om hierdie rede word 'n korpus as 'n onmisbare instrument in die verbetering van die makro- en mikrostruktuur van enige soort woordeboek beskou.

## Table of Contents

Declaration .....	i
Abstract .....	ii
Opsomming .....	iv
Table of Contents .....	vi
Liste of Tables.....	xi
Liste of Figures .....	xii
Acknowledgements.....	xiii
CHAPTER I: GENERAL INTRODUCTION .....	1
1.1 Introduction.....	1
1.2 Aims.....	1
1.3 Background.....	2
1.4 Socio-geographical situation of Yipunu .....	4
1.5 Linguistic context of Yipunu .....	5
1.6 Preliminary study .....	6
1.7 Rationale and research problem.....	7
1.8 Theoretical approach.....	8
1.9 Impact .....	10
1.10 Methodological approach.....	11
1.11 Historical view of corpus usage and development .....	14
1.12 Motivations .....	18
1.13 Overview and dissertation structure.....	20
1.14 Concluding remarks .....	21
CHAPTER 2: THE FORMULATION OF A DATA COLLECTION POLICY.....	22
2.1 Introduction.....	22

2.2 Data collection theories.....	22
2.3 Principles of data collection.....	31
2.4 Corpus usage.....	33
2.5 Corpus types.....	34
2.5.1 Monolingual and multilingual corpora .....	35
2.5.2 Synchronic and diachronic corpora .....	36
2.5.3 Spoken and written corpora .....	36
2.5.4 Samples .....	36
2.5.5 Reference corpora .....	37
2.5.6 Monitor corpora .....	37
2.5.7 Parallel corpora .....	38
2.5.8 Comparable corpora.....	39
2.5.9 Multilingual corpora .....	39
2.5.10 Towards hybrid corpora.....	41
2.6 Dictionary basis .....	42
2.6.1 Sources of materials .....	44
2.6.1.1 The material acquisition phase.....	44
2.6.1.2 The material preparation phase.....	45
2.6.1.3 The material processing phase .....	45
2.6.2 Written materials.....	46
2.6.3 Spoken materials .....	50
2.6.4 Methods of language material collection.....	53
2.6.4.1 Methods of obtaining spoken sources.....	53
2.6.4.2 Methods of obtaining written sources.....	56
2.7 Concluding remarks .....	58



CHAPTER 3: CORPUS STRUCTURE .....	60
3.1 Introduction.....	60
3.2 Types of data.....	60
3.2.1 The comment on form.....	61
3.2.2 The comment on semantics.....	62
3.2.3 Corpus design as an aid to the representation of the comment on form .....	65
3.3 The main corpus.....	71
3.4 Corpus representativeness and balance.....	72
3.5 Use of tags.....	75
3.6 The role of a lexicographic unit in the process .....	77
3.6.1 Guthrie’s classification .....	79
3.6.2 Kwenzi’s classification .....	81
3.7 Language status and corpus building.....	86
3.8 On alphabets.....	88
3.8.1 Graphemes of Yipunu .....	92
3.8.1.1. Vowels .....	92
3.8.1.2. Consonants .....	93
3.8.2 The chosen orthography of Yipunu .....	94
3.8.3 Conjunctive vs. disjunctive.....	94
3.9 Standardising an orthography .....	96
3.10 Fonts.....	98
3.11 Corpus structure .....	99
3.12 Concluding remarks .....	100
CHAPTER 4: COMPUTERISING A CORPUS .....	102
4.1. Introduction.....	102

4.2. Data transfer methodology.....	102
4.2.1. Keyboarding.....	102
4.2.2. Scanning.....	105
4.2.3. Downloading.....	106
4.2.4. Cleaning the corpus.....	107
4.3. Corpus analysis .....	108
4.3.1. General statistics .....	110
4.3.2. Word frequencies in corpora.....	112
4.3.3. Word lists .....	118
4.3.4. Concordance lines in the SYC .....	120
4.3.5. Collocations .....	125
4.3.5.1. Word clusters .....	127
4.3.5.2. Sorting definitions and examples.....	128
4.3.6. The ruler as instrument to measure the macrostructure stretch .....	129
4.3.7. Projecting Yipunu databases.....	131
4.4. Corpora as means of enhancing the quality of the lexicographic representation .....	133
4.4.1. Corpora and semantic data.....	133
4.4.2. Corpora and pragmatic data.....	136
4.4.3. Corpora and grammatical data .....	140
4.5. Concluding remarks .....	147
CHAPTER 5: CORPORA AND THE COMPILATION OF DICTIONARIES .....	148
5.1. Introduction.....	148
5.2. About different typologies .....	148
5.2.1. Different classifications of dictionaries .....	149

5.2.2. Different types of macrostructure .....	153
5.2.3. Different types of microstructure .....	156
5.3. Lexicographic functions and corpora.....	159
5.4. Corpus for monolingual dictionaries .....	162
5.4.1. Encyclopaedic dictionaries .....	162
5.4.1.1. The macrostructure of encyclopaedic dictionaries .....	162
5.4.1.2. The microstructure of encyclopaedic dictionaries .....	166
5.4.2. Diachronic dictionaries .....	170
5.4.2.1. The macrostructure of diachronic dictionaries .....	170
5.4.2.2. The microstructure of diachronic dictionaries .....	173
5.4.3. Synchronic dictionaries.....	176
5.4.3.1. The macrostructure of synchronic dictionaries.....	177
5.4.3.2. The microstructure of synchronic dictionaries .....	178
5.5. Corpus for translation dictionaries.....	185
5.5.1. The macrostructure of translation dictionaries .....	188
5.5.2. The microstructure of translation dictionaries .....	191
5.6. Corpus for restricted dictionaries.....	196
5.6.1. The macrostructure of restricted dictionaries .....	196
5.6.2. The microstructure of restricted dictionaries .....	198
5.7. Data distribution: A corpus view .....	200
5.8. Concluding remarks .....	204
CHAPTER 6: GENERAL CONCLUSION.....	206
BIBLIOGRAPHY .....	215
APPENDIX: The top 3027 words extracted from the SYC .....	232

## Liste of Tables

<b>Table 3.1:</b> Alphabet of the Gabonese languages by André Raponda-Walker.....	88
<b>Table 3.2:</b> Scientific Alphabet for the Gabonese Languages.....	89
<b>Table 3.3:</b> Alphabet of the Gabonese languages.....	90
<b>Table 1.4:</b> Yipunu vowels.....	92
<b>Table 3.5:</b> Yipunu consonants.....	93
<b>Table 2.1:</b> Overall statistics of sources used in the SYC.....	111
<b>Table 4.2:</b> Frequency word list of the 200 top frequencies in the SYC.....	113
<b>Table 4.3:</b> 15 Hapex legomena extracted from the SYC.....	118
<b>Table 4.4:</b> Section of a simple alphabetical word list from the Yipunu corpus.....	119
<b>Table 4.5:</b> An alphabetic frequency word list from the Yipunu corpus.....	120
<b>Table 4.6:</b> Concordance lines for the associative <i>na</i> (with).....	121
<b>Table 4.7:</b> Concordance lines for the comparative <i>nana</i> (like) .....	122
<b>Table 4.8:</b> Collocation range for <i>na</i> within the horizon L5-R5 in SYC.....	126
<b>Table 4.9:</b> Example for Yipunu locative cluster <i>vha</i> .....	127
<b>Table 4.10:</b> List of different nouns and their frequencies from the SYC.....	143
<b>Table 4.11:</b> List of some verb forms and their frequencies from the SYC.....	144
<b>Table 4.12:</b> List of some locatives and possessives with their frequencies from the SYC.....	145

## Liste of Figures

<b>Figure 1.1:</b> Location of the Bapunu people.....	4
<b>Figure 3.1:</b> Phonetic word segmentation in <i>bubitu</i> (gums).....	68
<b>Figure 3.2:</b> Segment identification in <i>aghatsi</i> (he/she is not in).....	69
<b>Figure 3.3:</b> Spectrographic segment identification in <i>mitsogha</i> (suffering).....	70
<b>Figure 4.1:</b> Example of the phonological transcription extract from Kwenzi-Mikala (1997) .....	104
<b>Figure 4.2:</b> Alphabetical stretch measurement of the SYC.....	131
<b>Figure 5.1:</b> Representation of bilingual/multilingual dictionaries for the Gabonese languages.....	186

## Acknowledgements

I wish to express my sincerest appreciation to:

The Gabonese government, for the financial support throughout my studies at Stellenbosch University.

Prof R.H. Gouws for his academic guidance and patience throughout the years.

The completion of this dissertation was achieved through so much support (financial and emotional) and encouragement that I am forever indebted to so many colleagues, friends and family members.

I would like to thank especially the following people:

My grandmother Adèle Dianga, my mother Elisabeth Nyangui and my father Jacques Soami for everything you have done so far.

My other family members and friends:

Fernand Léandre Bakenda  
Mireille Bibalou  
Marcelle Anouchka Ella Ebane  
Serge Stéphane Ibinga  
Annette Madjinza  
Angèle Matsanga  
Nestor Mapangou  
Victor Misulu Mutamba  
Gaëlle M've  
Nina Ndembi Bouanga Mamboundou  
Hugues Steve Ndinga-Koumba-Binza  
Richard Orendo-Smith  
Gilles Saphou-Bivigat

## **CHAPTER I: GENERAL INTRODUCTION**

### **1.1 Introduction**

This section deals with the overall presentation of the research. This dissertation focuses on the development of representative databases for the Gabonese languages, taking Yipunu as a reference. In fact, what is done in Yipunu will serve as an example for the rest of the Gabonese languages in particular and other African languages in general. A comprehensive presentation of the Gabonese languages can be observed in paragraphs 3.6.1 and 3.6.2 with regard to their internal relations and their individual classification.

This chapter analyses the objectives related to the research and deals with the research problems and hypotheses of the study. Moreover, it presents the motivations for the study. It also gives a brief analysis of the impact of the study and provides a discussion on the methodological approach used in the study. Furthermore, the chapter deals with previous lexicographic studies on Yipunu and the social and geographical environment of Yipunu. Then in a final paragraph, the structure of the dissertation is presented.

### **1.2 Aims**

This work presents different aspects related to the gathering of language materials that are going to be used for the compilation of dictionaries in the Gabonese languages in particular and in the rest of the African languages in general. The study investigates different source materials needed for the corpus and discusses the different possible usages of these materials after being collected.

In addition, the aim of this work is to provide Gabonese lexicography with a model for the development of databases to be utilised for the compilation of dictionaries. This model needs to be adequately formulated in order to bring about the expected results and satisfy the needs of researchers in the Gabonese languages. The model is initially applied to Yipunu and will later be enlarged to make provision for the rest of the Gabonese languages. The way in which the material in Yipunu is gathered and applied will serve as an example for all the Gabonese languages. Throughout this dissertation it will be shown how the model is applied to the Yipunu context since this

language provides the material that is used as point of departure in this dissertation. The development of databases implies a sound material collection policy. This research also focuses on the formulation of such a policy, which can be found in Chapter 2 of this dissertation. In this work all the instructions regarding the design of a good corpus have been referred to carefully in order to avoid mistakes regarding the types and sources of language material collected and the methods of collection. Databases do not refer to the product, that is the corpus, but to the process, in other words the method of storage of the material collected. The two terms *corpus* and *database* must not be confused. As already mentioned, *database* refers to the programme used for storage and *corpus* has to be understood as explained and defined in paragraph 1.5. The implication is that after compiling a corpus, the material needs to be stored in a database in order to be utilised efficiently.

The main aim of this study is to provide a theoretical model for the way in which data will be utilised in a database for the compilation and development of dictionaries for Yipunu.

The secondary aims are

- to serve as a model for the rest of the Gabonese languages; and
- to emphasise the importance of language material in the compilation process of dictionaries.

The use of corpora in lexicographic research worldwide is a fundamental reason to take a real step forward in the development of the Gabonese languages. This is possible through the collection and digitalisation of language material from all these languages.

### **1.3 Background**

Surrounded by Cameroon and Equatorial Guinea in the north, by the Republic of Congo in the east and south and by the Atlantic Ocean in the west, Gabon is, like most African countries, a multilingual country. According to Kwenzi-Mikala (1988), 62 languages including dialects, are spoken in Gabon. These languages are unequally spread through the country. All research that has been done on these languages has primarily been based on linguistic studies that have used the Greenberg, Tervuren and



Welmers questionnaires of Doneux (1967), which are composed of lists of words in isolation and phrases and sentences in context. Gabon does not have a tradition of dictionaries, even if some attempts exist in Yipunu, Fang and the Omyene languages, to name a few (see page 26 for detail).

In this regard, a corpus directed at lexicographic needs seems to be the most suitable for the compilation of any dictionary according to the needs of the target users. Such a corpus can help to solve the problem of the compilation of a good dictionary by providing the right lexical items, example sentences and the context in which different data types are used. It will also help in the definition of some lexical items as part of the lexicographic process. It can also improve the quality of the first dictionaries in the Gabonese languages in terms of their macro- and microstructural treatment and presentation. Revision of existing dictionaries may also be possible if one uses the corpus. The macrostructure and the microstructure of existing dictionaries can be compared with the frequency counts to identify the lemmas that were mistakenly left out by the dictionary compiler and the ones that need to be included in the new version. It is still possible to compile a dictionary without data collection if one does not want to compile a good one, but nowadays it is generally accepted that language material collection is necessary in order to compile a good dictionary. Language material collection is the first step towards the compilation of a language material database. Such a database can be helpful to solve the problem of the standardisation of the Gabonese languages in general and Yipunu in particular and also for the choice of data to be included in or excluded from a dictionary.

For Prinsloo and De Schryver (2001), if African language linguistics is to take its rightful place in the millennium, the active compilation, querying and application of corpora should become an absolute priority.<sup>1</sup> A corpus, as underlined above, is a work that precedes the compilation of any dictionary. The model that will be applied in the Gabonese languages must focus on both the theoretical and practical aspects of corpus design. If dictionaries in these languages are to be compiled, the focus should be first on the development of corpora. Without corpora it is difficult to compile a good dictionary.

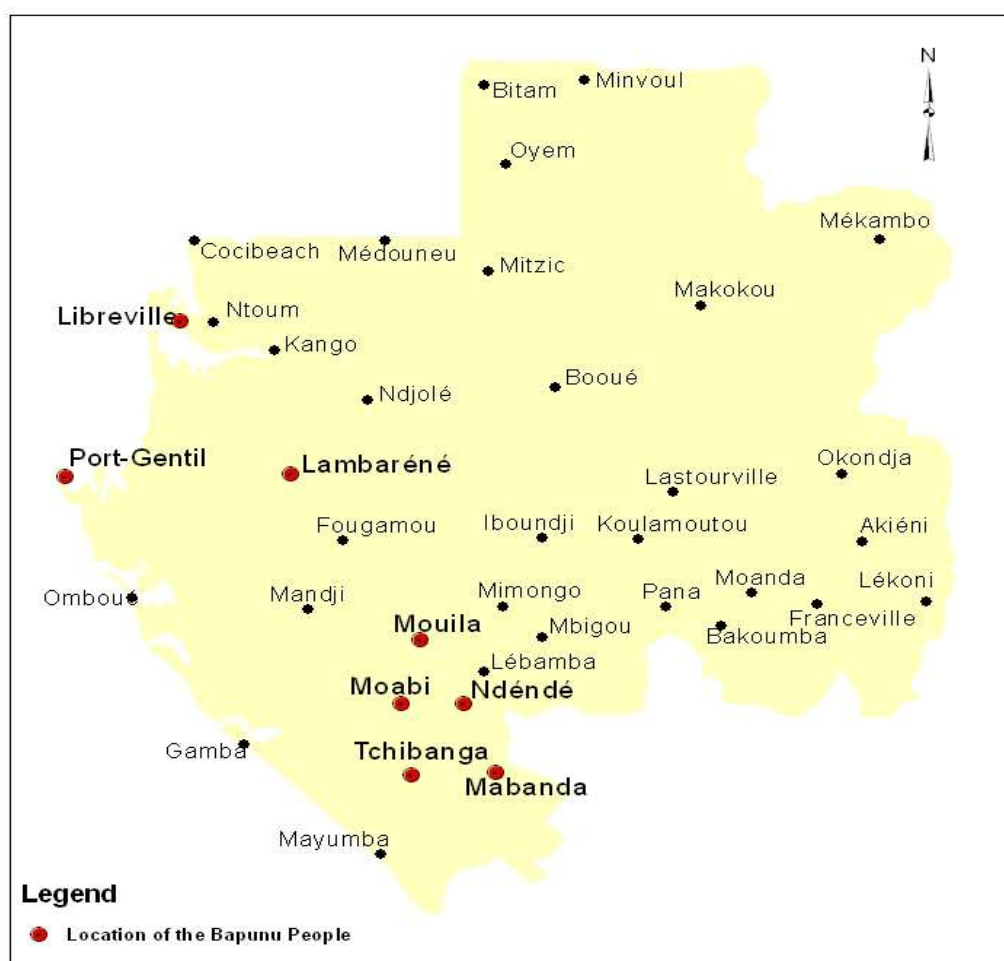
---

<sup>1</sup> Lecture in corpus building.

### 1.4 Socio-geographical situation of Yipunu

The term *Yipunu* is used to denote the language used by the Mupunu or Bapunu people. Yipunu is normally used in the southern region of Gabon in the Nyanga and Ngounié provinces. In these two areas the Bapunu are in a daily contact with the Nzebi, Massango, Gisira, Tsogo, Varama, Balumbu and Vili people, to name but a few.

The lack of employment has obliged the Bapunu to immigrate to what they consider as better locations. Consequently, the number of Bapunu people has tremendously grown in some areas such as Port-Gentil in the Ogooué Maritime Province, Libreville in the Estuaire Province and Lambaréné in the Moyen Ogooué Province. In these new regions contact with Fang, Omyene and other Gabonese languages has been established. The following map shows the dispersion of the Bapunu in the country:



**Figure 2.1:** Location of the Bapunu people

Source: designed by S. Idima and L.S. Soami

### 1.5 Linguistic context of Yipunu

According to some linguists of African languages, Yipunu is a Bantu language that belongs to Group B43, as shown by Guthrie (1948; 1953). It belongs to both the Niger-Congo and the subgroup Benue-Congo family, according to Greenberg (1955). The Bantu family is considered as a subgroup of a subgroup of Benue-Congo and not a subfamily coordinate with the others in Westermann's (1911) terminology; Greenberg also supports this conception. This is mentioned to portray the linguistic environment of the Bantu languages rather than getting involved in a discussion about Yipunu's relation with whatever group.

According to Mutaki and Tamanji (2000:1), Guthrie gives a typological classification and the Greenberg one is based on genealogical aspects. Guthrie's classification has also been used in Williamson (1971) and Sebeok (1971).

Mutaki and Tamanji (2000) add that typologically, Bantu languages are agglutinative compared to inflectional languages such as Greek and isolating languages such as Chinese. Agglutinative languages are languages in which boundaries between stems and affixes are clear-cut, as in Swahili, Lingala and Yipunu.

Yipunu and other Bantu languages form verbs by adding prefixes and suffixes to a verb stem. Prefixes, which come before the verb stem, indicate who (the subject), when (the time period) and what (the object); suffixes, which come after the verb stem, express prepositional phrases, causal relationships and passive voice.

As far as nouns are concerned, prefixes and are attached to the noun stem to mark a noun class. In Bantu languages, nouns normally consist of a prefix followed by a noun stem. The prefix can indicate number, for example *mutu* (one person) and *Batu* (people) in Yipunu. The same can be applied to Shona: *Mu/va* and *munhul/vanhu* respectively express whether one person or more than one is meant. Nouns and other parts of speech such as demonstratives, verbs and adjectives undergo changes for agreement in all Bantu languages.<sup>2</sup> This can be illustrated in the Yipunu examples

---

<sup>2</sup> Most of this information was taken from

[http://encarta.msn.com/encyclopedia\\_761565449/African\\_Languages.html](http://encarta.msn.com/encyclopedia_761565449/African_Languages.html)

*ivika igona imabulig* and *bivika biboti bimabuliga*, which mean respectively ‘the good chair is broken’ and ‘the good chairs are broken’.

This information about verb and noun classes together with other parts of speech, word formation and word structures can show why most Bantu languages are written conjunctively. It is also important to mention that the choice of writing a language conjunctively or disjunctively is arbitrary most of the time. The presentation of linguistic context of the Bantu languages can appear superficial but the reality is that we think it important to go back to the history of languages in order to understand the existing relations between languages. This information is provided here just to contextualise Yipunu by briefly presenting its linguistic environment for clarity and for the sake of information.

### **1.6 Preliminary study**

No studies have yet been done on lexicographic corpora in Gabon. All the data available in the Gabonese languages are mainly oriented toward linguistic studies. In addition, one can note the existence of some books by the Fondation Raponda Walker in the Gabonese languages regarding the teaching of some Gabonese languages. Some dictionaries compiled by missionaries as pioneers also exist. Nowadays researchers at the Université Lumière Lyon II are dealing with an important project for the compilation of a database of the names of fish in the Gabonese languages. This data, when available, can be added to the material already collected in the Gabonese languages. All these materials can form the point of departure for the database to which additional data will be added along with existing recordings on Gabonese radio and television.

Even if, a lack of metalexical work in databases in the Gabonese languages can be noted, some works in the field of practical lexicography exist, such as the Brown University Standard Corpus of Present-day American English (known as the Brown Corpus); the major lexicographical mega-corpus project. We can also mention the existence of the Collins Birmingham University International Language Database (COBUILD); the British National Corpus (BNC); and, finally, the Bank of English initiated at the University of Birmingham. This research will try to integrate all the practical benefits of the above projects.

Several doctoral dissertations of a metalexicographical nature have already been presented in some of the Gabonese languages. So far, Fang, Yipunu and Yilumbu are the only languages that have constituted the central focus of researchers in Gabonese lexicography. This critical situation should urge political decision makers and researchers to expand studies to the rest of the languages in Gabon. Realistic ideas around the debate on the number of languages are discussed in Chapter 3 and proposals are made so that too much time is not wasted.

The arrival of Gabonese students in South Africa to acquire practical lexicographic skills and theoretical knowledge has made it possible for the scientific environment in Gabon to be enriched by high-level researchers in metalexicography. This will make the compilation of dictionaries in the Gabonese languages easier in future. At the same time it will also help to improve the quality of the product to present to the general public by taking into consideration the users' needs. This is made possible by involving the research and training courses at the Bureau of the WAT and at the University of Stellenbosch in the Department of Afrikaans and Dutch and exposing them to international research of high calibre via seminars, workshops and conferences.

### **1.7 Rationale and research problem**

All the languages in Gabon are still in the process of development. They need real attention and language material in order to achieve such development and they should be managed carefully. All the languages lack adequate academic material in terms of handbooks and other didactic tools for their teaching and enough language material for the research to be endeavoured in these languages. The main emphasis of this study is on formulating and commenting on various theoretical and practical aspects of corpus development. It will in many ways help to fill the gap by providing Yipunu with some material and guidelines that can boost research in all scientific areas but more specifically in metalexicography. The researcher believes that the development of corpora is the best starting point. This research investigates the different ways in which corpora and databases can be utilised to enhance the compilation of dictionaries in the Gabonese languages.

In this regard, it is important to note that although many linguists may use the term *corpus* to refer to any collection of texts it is used here to refer to a body of text that is carefully sampled to be maximally representative of the language or language varieties and to be mainly used for the compilation of dictionaries. Here the term *corpus lexicography* is more appropriate than *corpus linguistics*, mistakenly used by some researchers as a generic concept. Corpus lexicography should be used in the metalexicographic environment and for practical lexicography purposes while corpus linguistics, as clearly indicated by the term, should be used for linguistic matters. Both terms should be seen as subsets of the activity within an empirical approach. Although corpus linguistics entails an empirical approach, empirical linguistics does not always entail the use of a corpus.

The last decades have shown tremendous development in corpus building and application, both in linguistics and in lexicography. In Europe and Africa, researchers are gathering material on the theoretical and practical levels. These activities have already shown results in such a way that many dictionaries are being published more rapidly and effectively.

The collection of materials in the Gabonese languages has been going on for quite some time in linguistics, anthropology and sociological studies. Many of these materials will be included in the whole corpus as the project is at the starting point and still growing. For the Stellenbosch Yipunu Corpus (SYC), most of the published and unpublished works have been taken into consideration.

### **1.8 Theoretical approach**

This thesis for the most part uses the theory of Wiegand, based on general principles in lexicography. This theory will be applied in the compilation of the model. Wiegand's theory gives an exposition of the four most important components of metalexicographic theory: constituent theories A, B, C and D, which are briefly discussed in Chapter 2 (see page 23 for detail).

Wiegand's general theory of lexicography makes provision for the distinction of many terms and concepts and avoids confusion by clearly defining them. He focuses on many areas, amongst which are systematic dictionary research (this is the general theory of lexicography) and research on the history of dictionaries, dictionary use and

dictionary criticism. Systematic dictionary research includes topics such as the purpose of dictionaries, the relation between lexicography and other disciplines, the organisation of lexicographic activities, lexicographic language research (e.g. research on the collection and processing of data as well as computer lexicography) and the theory of lexicographical description of language (e.g. dictionary typology and the structure of lexicographic texts).<sup>3</sup>

In many articles related to his theory of lexicography, Wiegand gives his opinion on many objections formulated by some researchers to lexicography with regard to its scientific features and its relation to other disciplines. The intension of this research is not to argue about the criticism levelled at lexicographers but rather to explain Wiegand's theoretical logic by expressing his ideas. Wiegand has expressed his thoughts on many aspects of lexicography. This research will not address all the aspects. The focus will be on some of them and they are given here for the sake of the presentation of the theory, as has already been pointed out.

To the question of whether lexicography is a science or not, Wiegand clearly states that lexicography is not a science. For him, to answer this question properly, it is necessary to understand the features of a science. According to Wiegand (1984:13) "scientific activities as a whole are aimed at producing theories, and this is not true of lexicographic activities". Scientific procedures are formulated and applied on theoretical principles and on a clearly identified object. In other words, a science as a theory and an object. Wiegand adds that even if lexicography is not a science, it is a scientific practice because both the process of dictionary compilation and the result of lexicographic activities are not theories. Lexicography is mainly orientated toward the production of reference works such as dictionaries. Furthermore, based on the above ideas, Wiegand (1989a) also states that this should not be understood that one does not use any theories, part of theories or scientific methods coming from these theories in the planning and compilation of dictionaries. Rather, what should be taken into consideration is the fact that the aim of lexicography is not to develop and test theories about specific phenomena.

Wiegand (1984b) also believes that lexicography is not a subdivision of applied linguistics because the target of lexicography is not to test linguistic theories and

---

<sup>3</sup> This information comes from Prof. R.H. Gouws (lectures).

methods. In the compilation of some special dictionaries, certain other disciplines play a prominent role. For a dictionary of dialects, for example, one would need geographic or folklore data. The researcher believes that lexicography is a scientific activity that uses the strong points of other theories as well as its own theoretical contributions in order to achieve its goals.

This research project also takes into consideration many aspects of other theories related to research on corpora and language analysis to fill the gap in Wiegand's theory since he did not elaborate extensively in this regard. For instance, this research will look at Prinsloo's argument when he says that a main dictionary is based on a main corpus. Prinsloo deals with South African Bantu languages as the main focus. In this context, all the African languages in general and the Gabonese languages in particular will benefit from the experience of their South African counterparts. According to Prinsloo (2000), since any modern dictionary is to derive its data from a corpus, the compilers have to build and query an electronic corpus for the specific language(s) first. The compilers cannot start the compilation of the main dictionary before they have access to such a satisfactory corpus.

The hub-and-spoke theory of Martin (1996) is implemented in the multilingual environment of Gabon in relation to the compilation of bilingual dictionaries for such a society. A more comprehensive analysis of the theory is provided in Chapter 5 of this dissertation.

This research also utilises a number of other relevant theories that are not mentioned here. The reader will come across these theories in all the chapters of the dissertation.

## **1.9 Impact**

The title of this study, 'Towards the development and application of representative lexicographic corpora for the Gabonese languages', speaks for itself in the sense that the research is primarily aimed at lexicographic activities. The application of this research could lead to the compilation of one or more major corpora for the Gabonese languages, which will allow different research projects to be performed. Apart from lexicographic purposes, the research can be appropriate for any linguistic studies in these languages, from morphology to syntactic analysis. It will also benefit research in sociology and anthropology.



This study will contribute to the reconsideration of the orthography of the Gabonese languages and will constitute a sound base for the compilation of academic manuals in these languages to meet the expectations of the government educational policy based on the teaching of the Gabonese languages on preprimary, primary, secondary and tertiary education levels.

Furthermore, the study will have an impact on the automatic development of the Gabonese languages by leading to the compilation of electronic dictionaries that can be downloaded on CD-ROM or in other forms. It will also enable the publication of the corpus on the Internet for international use. The study will also help to adjust the contextualisation of some theoretical aspects by giving them a particular orientation in terms of their practicability of usage in corpus lexicography.

### **1.10 Methodological approach**

This research applies different methodological approaches to corpus lexicography used in other studies. The type of research initiated by Prinsloo, De Schryver and other researchers for the development of corpora in Sepedi, Cilubà and Kiswahili and other African languages is followed and applied in the Gabonese context. It plays an important role in this work. This research also looks at some relevant work in both European and American situations where corpora have played a prominent role in the compilation of dictionaries. All these works are referred to in this dissertation.

In language studies the past decades have seen the development of many analytic methods. This study has implemented both the rational-theory-based and the empiricist-descriptive methods. These two approaches are often seen as separate and in competition with each other. However, some researchers have used corpora in order to test essentially rationalist grammatical theory rather than use them for pure description or the inductive generation of theory, as indicated by Schmied (1993).

However, the methods used in this study will be empirical and descriptive, using quantitative and qualitative analysis. It is internationally recognised that there are two main language material sources: written sources and oral or spoken sources. In this work, the focus is directed at both spoken and written sources and all the material available in Yipunu will be utilised. Even if there is a lack of a written tradition in the

Gabonese languages, some written material exists in some of the languages. To this is added, for Yipunu more specifically, oral recordings that were transcribed and transformed in electronic format. This was done with the help of mother-tongue speakers. Written sources such as linguistic and religious texts will also be used.

It is important to note that in order to meet expectations, three fieldwork sessions were conducted in Gabon, mainly in Tchibanga and Libreville. The first took place between December 2001 and February 2002, the second between December 2002 and March 2003 and the last one between December 2006 and March 2007. In total, 22 participants of different age, sex and class were recorded in order to obtain a general view of Yipunu. Then the recordings had to be transcribed before being used for the data analysis, as is shown in Chapter 4. This dissertation clearly explains the different methods used to transfer all the language material to electronic format. The oral material together with the identified written sources had to be saved in a computer file and will have to be stored in a computer database. Chapters 2 and 3 give an extensive description of this process.

From this brief discussion, it can be appreciated that both qualitative and quantitative analyses have something to contribute to corpus study. There has been a recent move in social science towards ‘multi-method’ approaches, which tend to reject the narrow analytical paradigms, in favour of the breadth of information that the use of more than one method may provide. As Schmieid (1993) explains, a stage of qualitative research is often a precursor of quantitative analysis, since before linguistic phenomena can be classified and counted the categories for classification must first be identified. In addition, Schmieid demonstrates that both corpus linguistics and lexicography could benefit as much as any other field from multi-method research. This is also a manner of opening ways for lexicographic research by looking at the methodological approaches used in other disciplines in order to select the necessary features that can benefit any lexicographic project.

The use of qualitative analysis ensures the completeness and detailed description of data in the corpus. Its utilisation will not allow one to assign frequencies to the linguistic features that are identified in the data, and uncommon phenomena receive (or should receive) the same amount of attention as more frequent phenomena. Another important feature of qualitative analysis is that it allows for normal

distinctions to be drawn because it is not necessary to adjust the data into a determined number of classifications.

Even the most unclear situation or ambiguous phenomenon intrinsic in human languages can be recognised in the analysis. This is a very important matter when it comes to the identification of particular data that need to be considered in the macrostructure and the treatment in the dictionary article and that can be applied in a situation where a given lemma sign can have a polysemic sense and impact in its contextual usage. All the senses of that lexical item are taken into consideration in a qualitative analysis.

One of the major problems with qualitative approaches to corpus analysis could be that their conclusions are usually based on a limited population and cannot be extended to wider populations with the same degree of certainty as could be done by using quantitative analyses. This is because once a lexical item has already been qualitatively selected for inclusion in the macrostructure, it does not display any relevancy in terms of the number of times it appeared in the corpus and the deductions of the research are not tested to discover whether they are statistically significant or due to chance.

The qualitative analysis of the SYC will allow the extraction from the corpus of the necessary data that could be used for the compilation of dictionaries in Yipunu. A total of more than a million words have been broken down to a mere 60 000 words constituting the number of types in the corpus. Further details on the matter are provided in Chapter 4. All these elements are essential to the analysis and conclusions that the lexicographer may use qualitatively to help in his or her decision making based on the availability of the data.

Quantitative analysis will enable the classification of some features, count them and even make provision for statistical models in an attempt to explain what is observed. This will certainly enable the observation of particular phenomena to be generalised to a larger population, in other words language data. The observations could also be extended to different corpora if the gathering of language material has been subjected to the use of good sampling methods. Thus, quantitative analysis will allow ascertaining in the whole corpus which phenomena are likely to be genuine reflections of the behaviour of a language or variety and which are merely chance occurrences.

The more basic task of just looking at a single language variety allows one to form a precise picture of the frequency and rarity of particular phenomena and thus their relative normality or abnormality.

It is important to emphasise that the representation of the data that emerge from quantitative analysis is not as relevant as the representation of that obtained from qualitative analysis. For statistical purposes, classifications have to be made in a certain way so that an item is labelled as pertaining to a given class and not to any other. To ensure that certain statistical tests provide reliable results, it is essential that minimum frequencies be obtained; that is to say, categories may have to be collapsed into one another, resulting in a loss of data richness, as indicated by Schmied (1993).

The use of a statistical programme such as WordSmith Tools will help to indicate the difference between quantitative data and qualitative data in the corpus. Quantitative investigation will take into consideration all data (relevant or irrelevant) in the corpus without any distinctions, and the usability of the material will be determined by the qualitative method as it will meticulously dissect these materials and extract the more beneficial ones.

### **1.11 Historical view of corpus usage and development**

The literature reveals that empirical data have been used in lexicography long before the discipline of corpus linguistics was invented. Samuel Johnson, for example, illustrated his dictionary with examples from literature, and in the 19th century the *Oxford Dictionary* used citation slips to study and illustrate word usage. Corpora, however, have changed the way in which lexicographers can look at language on the one hand, and on the other hand corpora have helped lexicographers to enhance the quality of dictionaries via the frequent use of text evidence.

Many researchers such as Chomsky, as mentioned by Landau (2001), and Bergenholtz, as quoted by Smith (1996), have expressed their reluctance to use corpora in quantitative analysis because they believe that it is an expensive endeavour and a time-consuming enterprise. Another supportive reason is that of Chomsky and his syntactic structures theory. Based on the infinite potential uses of language (performance), the ability (competence) of the native speaker to generate such an infinite variety could only be explained by an intuitive grasp of the grammatical rules

underlying such a generation of utterances. Moreover, according to Chomsky, syntactic structures set forth some of those rules or ‘transformations’ that, if properly understood, could account for the particular form of every possible utterance.

Chomsky and his followers were hostile to a quantitative approach to language study and sometimes ridiculed it because it could not account for the infinite variety of potential uses of language. Only the native speaker, they said, could do that. Landau (2001: 276) quotes Graeme Kennedy who remarks that Chomsky’s influence suppressed statistical approaches to language study and as a result corpus-based study of vocabulary “declined in influence from the 1950’s until its revival ... around the 1980’s”. In the same way Landau (2001: 276) mentions that Chomsky argues that linguistics is a branch of cognitive psychology, “that it can be based on intuitive data and isolated sentences, that corpus data is unrevealing, [and] that the study of language in use is essentially uninteresting”.

Despite this negative opinion of the use of a quantitative approach to language studies, many projects on corpus building have taken place worldwide. This situation has witnessed the development of those projects in different periods. Leech, as quoted by Landau (2001), observes that the term *corpus linguistics* did not exist in the 1950s but was first used in the 1980s. During this period, a number of corpus-based studies were undertaken in the United States and Britain.

The real development of corpora began with what is known as the first major computer-based study designed to be representative: the Brown Corpus, assembled in 1963–4. This paved the way for many other projects and stimulated the compilation of the Lancaster-Oslo/Bergen (LOB) Corpus in 1970 (Landau 2001).

To these corpus projects can be added the following:

- Randolph Quirk’s project of English Usage started in 1959 and was accomplished in 1989.
- The computerised spoken version of Quirk’s survey was undertaken by Jan Svartvik in 1975 and completed in 1980. It is known as the London-Lund Corpus.

- The Birmingham Corpus, which was begun in 1980 and later became the COBUILD (which stands for Collins Birmingham University International Language Database) Project, is a growing project that changed its name in 1997 to become the Bank of English.
- After the 1980s the Longman Lancaster English Language Corpus was created.
- The British National Corpus was begun in January 1992 and is still growing.
- Around 1995 the Cambridge International Corpus was created and is still developing.

To the above projects can be added the ones taking place in African languages, such as those for

- Sesotho
- IsiZulu
- Afrikaans
- Sepedi
- Swahili

These projects have been going on for some time and are still growing. They are all referred to in this research when necessary. The application of corpora in the study of linguistic phenomena over the years could benefit the development of corpora in lexicography. For instance, the LOB Corpus, undertaken in 1970 at the University of Lancaster and completed in 1978 by the University of Oslo in collaboration with the Norwegian Computing Centre in Bergen and the Brown Corpus, assembled in 1963–1964 at Brown University. These two corpora have been used for different research purposes around the world and were seen as groundbreaking events. Both the LOB Corpus and the Brown Corpus, as already mentioned above, contain roughly the same genres, sample sizes, and are assembled from the sources published in the same period, 1961. Another corpus known as the Kolhapur Indian Corpus was gathered with the same principles and is broadly parallel to Brown and LOB, although it was sampled 17 years later, in 1978.

Johansson and Norheim (1988), as cited by McEnery and Wilson (1993), mention that the Brown Corpus and the LOB Corpus have also been used as the basis of more complex aspects of language such as the use of the subjunctive. More specifically, the LOB and Brown corpora were initially used for the production of a word frequency comparison of written American and British English. Moreover, both these corpora contain valuable material that can allow learning the frequency of particular words within particular genres so that, for instance, British and American English can be compared with respect to the same genre. Clearly, such information is valuable in particular comparisons between varieties of the same language.

Once again, corpora in national variation studies have been used as an experiment for two theories of language variation. That of Quirk, Greenbaum, Leech, and Svartvik (1985), in relation with what he calls ‘common core’, and Braj Kachru’s understanding of national varieties as forming many unique ‘Englishes’ that are apparently different from one another in important ways (cf. McEnery & Wilson 1993).

Leitner’s (1991) point of view, as cited in McEnery and Wilson (1993), is that work on lexis and grammar comparing the Kolhapur Indian Corpus with Brown and LOB has supported the ‘common core’ hypothesis. Even if there is still scope for the extension of such work, Brown and LOB are the most referenced corpus-based works of this kind that one can consult when dealing with the same type of research.

There are few examples of dialect corpora at present around the world, two of which are the Helsinki Corpus of English Dialects and Kirk’s Northern Ireland Transcribed Corpus of Speech (NITCS). These two corpora are based on conversations with fieldworkers: in Kirk’s corpus from Northern Ireland and in the Helsinki Corpus from different English regions.

To be able to study prosody, discourse and register, to name a few, the researcher can take advantage of a corpus that in principle contains a number of text genres. By using a corpus, researchers will improve the results of their analyses. In line with what has been mentioned previously, a corpus needs to be sampled in a way that is representative enough to allow qualitative and quantitative studies. This can result in drawing conclusions about the target population in general.

For discourse analysis, for example, Macaulay (1991) suggests that dialectologists should pay attention to how the characteristic ‘flavour’ of a dialect may also reside in special norms for interaction, special types of speech event that may be embedded within a conversation and the use of elements whose functions are to smoothen interaction and conversation. Macaulay makes an effort to characterise the dialect of English in Ayr, Scotland, by quantifying the use of ‘discourse particles’ such as *I mean, know, you ken, oh*, and so on. These particles serve to keep the conversation flowing and at the same time give it a local and personal (‘you and me’) flavour. Moreover, it seems that, in a more significant way, dialect and discourse are other norms for organising conversation and interaction. Such aspects of speech culture involve genres such as narratives, children’s language games and the use of riddles and proverbs in ordinary speech.

### **1.12 Motivations**

There are many aspects that are not discussed in this dissertation because the subject of this research is not the compilation of a dictionary model as such but a process that guides the researcher towards the compilation of dictionaries, using language material. In other words, this dissertation is paving the road to any dictionary project that takes a corpus as a point of departure by showing different steps to be followed. The lexicographic theory makes provision for concepts such as macrostructure and microstructure as main components of the central list of the dictionary and other important aspects related to the lexicographic process. Corpora have to play an important role in the whole process in order to enhance the quality of dictionaries. The question will be how corpora can aid the compilation process in general.

The lexicographer takes satisfaction from delivering a product that is acknowledged by the users as being useful. This judgement would be based on the capability of the dictionary to help the users in their daily language activities. According to Wiegand (1996a), dictionaries as text type carriers are arranged into a macrotext that has a mediostructure and a metatext. For him, in lexicographic texts, the following structures are observed: the macrostructure, the outer and inner access structure, the textual frame structure, the textual wordlist structure, the microstructure, the article structure, the structure of the positions, the item structure, the addressing structure, the



mediostructure and the microarchitecture. Not all these elements are discussed in this study but they are essential in the determination of which items should be included in a dictionary article and in which order. The researcher believes that a corpus can help the lexicographer in decision making as it has many benefits as it is shown in chapter 2, 3, 4 and 5 of this dissertation.

A researcher who has access to a corpus or other material collection of machine-readable text can call up all the examples of a word or phrase from many millions of words of text in a few seconds. Dictionaries can be produced and revised much quicker than before, thus providing up-to-date information about language. Definitions can also be more complete and precise since a larger number of examples are examined.

Examples extracted from corpora can be easily organised into meaningful groups for analysis. For example, by sorting the right-hand context of the word alphabetically so that it is possible to see all instances of a particular collocate together. Furthermore, the way in which the corpus will be tagged will give information about the regional variety of a word, the author of the lexical item and the date of appearance of the word. It will also display the genre in which it belongs and the part of speech in which it can be classified. This will then facilitate the usage of the word when necessary.

Furthermore, the development of monitor corpora has its greatest role in dictionary building as it enables lexicographers to be aware of, for example, all new words entering the language, existing words changing their meanings or the balance of their use according to genre. However, finite corpora also have an important role in lexical studies, in the area of quantification. It is possible to produce rapidly reliable frequency counts and to subdivide the areas across various dimensions according to the varieties of language in which a word is used.

Additionally, the ability to call up word combinations rather than individual words and the existence of mutual information tools that establish relationships between co-occurring words mean that one can treat phrases and collocations more systematically than was previously possible. A phraseological unit may constitute a piece of technical terminology or an idiom, and collocations are important clues to specific word senses.

Wiegand's lexicographic theory explained before makes provision for elements such as macrostructure and microstructure, which are very important concepts in different lexicographic activities and dictionary compilation.

Many research questions need to be formulated if the use of corpora is to reach its goals:

- How can a corpus help to improve the macrostructure in the dictionary?
- How can a corpus help to improve the microstructural treatment of the lexical items in the dictionary article and the formulation of a microstructural programme?
- How can a corpus help to enhance the process of data distribution strategies as a dictionary is a carrier of text types?
- How can a corpus assist the fulfilment of the lexicographic functions in a dictionary?
- How can a corpus assist the lexicographer in the compilation of different types of dictionaries?
- What kind of material will be needed for what specific purposes?

All these questions are answered throughout this study in a consistent way.

### **1.13 Overview and dissertation structure**

This dissertation is divided into six chapters. Chapter 1, the introduction, highlights the building up of the research by indicating the aims and the problems related to the study, the methodological and theoretical approaches and the sociolinguistic and geographical situation of the language in question.

Chapter 2 deals with the comprehensive process of data collection by explaining how the planning of the collection envisaged in this project will be done. It also gives in detail some theoretical notes about the collection, the way in which the data will be collected and the implications pertaining to the collection.

Chapter 3 presents the way in which the corpus should be organised and how the data entries are linked to each other. After the collection of data, the lexicographer needs to

sort out all the material in such a way that it is useful and helpful in the compilation process of the dictionary. The way in which the data will be stored and arranged will facilitate the retrieval of the information.

Chapter 4 tackles the way in which different sources have been processed onto the computer. It explains the process from the keyboarding and the scanning methods to the way in which sources were downloaded from the Internet to the transfer of data in digital format. A focus on a general analysis of the corpus via a quantitative and qualitative method takes the reader through statistics, frequencies, concordances, collocations and lemmatisation.

Chapter 5 helps with the identification of the dictionary functions since they play an important role in the compilation of dictionaries. In this regard, there is no doubt about the importance of taking into consideration these theoretical issues when dealing with a particular dictionary. In this chapter, the researcher tries to stretch the statement that gives credit to the compilation of corpora according to dictionary types. The researcher also endeavours to demonstrate the relation between the corpus and the knowledge-oriented function and the communication-oriented function. In order to achieve this aim, it is necessary to give a background of typologies on dictionaries, macrostructure and microstructure. This will help with the lemmatisation and the treatment of the lexical items on different levels.

Chapter 6 provides a summary of the study, presents recommendations and draws the future perspectives of the research.

#### **1.14 Concluding remarks**

This section has helped to present the main theoretical aspects of the development of lexicographic corpora such as the model expressed in this research. The framework of this type of study necessitates further precision in order to fit the hypotheses related to the research by providing them with practical answers and explanations. All these notions need to be clarified and understood to ensure that the objectives targeted by the study are reached.

## **CHAPTER 2: THE FORMULATION OF A DATA COLLECTION POLICY**

### **2.1 Introduction**

The planning of data collection is part of the general process of any dictionary project. Therefore, the planning of any collection of material must be adapted according to the dictionary that the source is going to serve.

The data collection policy refers to the planning and the gathering of written and/or spoken material in a language. This implies a well-structured method of collecting data. When planning such a task one has to bear in mind the fact that, the sources have to be found according to the need and the genuine purpose of the research.

Data collection is an important activity that precedes the compilation of any dictionary. Data from which lexicographers draw their information and compile their dictionaries have to be chosen to suit the type of dictionary being planned and the needs of the users (Čermák 2003). The corpus serves the purpose, among others, of implementing a data collection policy.

The aim of this section is to deal with the comprehensive process of data collection by explaining how the planning of the collection in this project will be done. It will also give in detail some theoretical notes about the collection, the way the data will be collected and all the implications referring to the collection.

### **2.2 Data collection theories**

When formulating his general theory of lexicography, Wiegand (1984a:16) divided it into four subconstituents:

- A general section, constituent theory A, that entails the unification of the purpose of dictionaries, the relationship to other theories and the principles from the history of lexicography.
- An organisational section, constituent theory B, that serves the role of determining the basic rules for organising and coordinating all three areas of lexicographical activity.

- A theory of lexicographical language research, constituent theory C that consists of the set of all scientific methods that can be applied in lexicography composed by the theory of data collection and the theory of data processing.
- A theory of lexicographical language description, constituent theory D that is the class of all the presentations of the results of linguistic lexicography as texts about language (Wiegand 1984a:16).

All these different components of the theory will play a very important role throughout this work. In this section, the focus is on constituent theory C.

Wiegand calls data collection all the possible material that one needs for the compilation of the corpus that allows one to build a dictionary basis (this will be discussed in paragraph 2.5.) According to him, three different sources of material have to be identified, namely primary, secondary and tertiary sources:

- Primary sources are all the texts, spoken materials and conversations that are found in a specific language.
- Secondary sources are all the material coming from the existing dictionaries in that language.
- Tertiary sources are composed of the work done in the field of linguistics (it is of major importance to indicate the origins of sources and save all the data in a computer file).

These sources will be discussed in detail later when dealing with the relevant subsections.

In the same sense, Svensén (1993) proposes the existence of two main sources: primary and secondary sources. According to him, primary sources may be oral or written. He focuses on written sources since the dictionary he is dealing with does not take into consideration oral sources. Instead, he refers to Renouf (1987) for more information on oral sources. Primary sources include unprocessed material consisting of linguistic utterances, both oral and written. The latter group contains primarily dictionaries.

Svensén (1993) adds that oral sources are of importance as regards verifying collected material with the help of informants. The challenge of primary sources is to choose sources that will give the widest possible coverage of the language variants the

dictionary aims to deal with and from those chosen sources to make a selection of linguistic utterances that is as representative as possible.

Čermák (2003) states that nowadays, lexicographic resources may be viewed as primary (archive and corpus) or as secondary (fieldwork, other dictionaries and encyclopaedias and the Internet).

Landau (2001) makes a distinction between semiscripted speech and unscripted speech. The former includes transcriptions of radio or television talk shows and political interview programs and is produced in a controlled format that respects certain spoken conventions. In this context, the speaker does not have the freedom to say whatever he or she would like to. Unscripted speech is formed by elements such as conversations between friends or acquaintances or the informal conversation between workers. These are the two different types of spoken source according to Landau. Of course, the unlimited character of spoken material gives the researcher reason to pay more attention to it in order to gather a huge amount of data.

It is obvious that all the abovementioned researchers are talking about the same thing, written and spoken material, no matter what term is used. In the course of this study, particular attention will be given to both types of source in detail. The result of the collection of material is the constitution of a corpus base for any good dictionary nowadays. As for the Gabonese situation, an emphasis should be put on the gathering of a huge amount of materials for the compilation of new dictionaries, which will be based on a representative and balanced corpus to fully cover the Gabonese languages. The researcher believes that the success of the research does not reside only in the identification of different sources but also in the ability of these sources to help the lexicographer achieve his or her mission and objectives. The distinction between primary, secondary and tertiary sources seems to be applicable to languages with a long written tradition, and when it comes to most oral African languages, *oral* and *written* are sufficient terms to identify sources.

It must be mentioned here that missionaries compiled most of the existing dictionaries in Gabon during the colonial era. These dictionaries were mainly based on linguistic material and were compiled to allow the colonialists to understand the local languages. The information given in local languages is too restricted compared to that given in the colonial language.

The examples of one of the Fang dictionaries compiled by Galley in 1964, the Yipunu one compiled by Eglise Evangelique du Sud Gabon in 1966 and even the Pongoué dictionary compiled by Mgr Bessieux in 1847 show how urgently new dictionaries need to be compiled because these pioneerind works probably are full of old-fashioned words that need to be revised. What happened in Gabon is a picture of all colonised countries in Africa, whether they were English speaking or French speaking. The difference between these countries is not huge in terms of the situation experienced.

These works today are referenced and used as point of departure for scientific works. They need to be improved to fit the new challenges of today. Although in some African countries, such as South Africa, such works have already been revised and improved, matters in Gabon are still at the initial stage since new editions of these dictionaries still have to be compiled and published.

To avoid the mistake of the past of all the dictionary compilers in that the compilation was not corpus based, it will be beneficial for Gabonese lexicographers to compile large corpora in order to take into consideration the synchronic situations of these languages and to reflect the way they are spoken by the natives. By doing so, precautions will be taken and quality ensured. In other words, it is crucial to involve the speech community of the language the lexicographer is dealing with by recording native speakers from different ages and areas where the language is spoken. That will include all the different dialects in the corpus so that language varieties will be given space in the dictionary at a later stage.

One of the recurrent criticisms levelled at the content of dictionary articles and their compilers is that they rarely provide as much data as possible about the source from which a lexical item or citation sentences originate. All these details can be observed in a well-designed corpus, which is the reason why the material has to be gathered and ordered carefully. This urges lexicographers to pay more attention to sociolinguistic parameters in dictionaries; in other words, they must focus on greater precision and detail by presenting informative data on the community, the specific dialect and the age, the sex and even the level of education of the informants chosen among the community. When gathering and organising lexicographic data, the recording of the following parameters, among others, are of vital importance:

- the place where the research took place (the speech community)
- the community with which the lexicographer is in touch
- the gender of the informants
- the age of the informants
- the educational levels or occupations of the informants
- ethnicity
- the type of dialect, when applicable
- the year in which the research took place
- the author of the book or any other published document from which texts and citations were taken.

All these characteristics are valuable material that helps the lexicographer in the analysis of data. It is important for lexicographers to keep a record of the language type, the gender of the people and the different social stratifications of the people involved in the project. All the above listed variables are important in indicating social interactions among groups of people in any community. The mission of the lexicographer is to provide the community speaking the language he or she is busy compiling the dictionary for with a product that will serve its needs.

This can be successfully done only if the study includes informants from different age groups. This makes it possible to compare the data gathered from young people with that gathered from old people. By doing so, the lexicographer could immediately see the difference between the vocabularies used by youngsters and those pertaining to older people. These different vocabularies are of great importance in the compilation of different types of dictionary, as shown in Chapter 5.

More specifically, the corpus can allow the analysis of different discourses by different people from different regions speaking the same language in order to identify the pronunciation diversity among them. In a given region, people will pronounce the same words in different manners. Research in the field would benefit from the collection of large numbers of material to allow quantitative studies, if this material is meticulously gathered.



By doing so, the lexicographer should be able to carefully manage data and gain a more precise overview of the material at his or her disposal. That, of course, will facilitate the tagging of the material in order to fit the requirements of the lexicographic process ahead of the compilation of the dictionary/ies. That will ensure the success and fulfil the objectives of the product to be presented to the general public. It will also allow capturing the target users' attention and deliver a dictionary that better takes into consideration the needs of users. This process has been implemented in many projects whereby materials were assembled to be utilised for special purposes.

With regard to the situation presented above, the work done by Labov (1966) can be beneficial to the lexicographer in terms of the way in which materials are arranged and marked up to fit the research objectives. In his research, Labov (1966) with his classic *The Social Stratification of English in New York City* has influenced linguistic studies of change and variation. He used audio recordings of the speech of residents of New York's Lower East Side. The social characteristics of speakers allowed sophisticated comparisons between linguistic and social data.

The basic elicitation tool was to conduct a sociolinguistic interview in which the investigator asked the respondent many series of questions that had been designed to inspire connected speech and to encourage the use of particular linguistic forms. The alternatives that belong to a sociolinguistic variety are called variables or variants. The analysis ended in a pattern called class stratification.

According to Labov (1966) the study showed whether variety denotes class stratification. Certain variants are used most frequently by the highest status class, least frequently by the lowest status class and at intermediate frequencies by the classes in between, with the frequency matching their relative status. For Labov, the patterns of linguistic variation are critical for a full understanding of language change over time. He contends that sound change in progress can be observed but only by studying language in its social context.

Strassel *et al.* (2003) similarly report on the SLX Corpus of Classic Sociolinguistic Interviews, comprised of eight sociolinguistic interviews with a total of nine speakers that were conducted in the 1960s and 70s. All the interviews were conducted by William Labov or by one of his students. Labov mentions that these interviews are not

classical in the sense that they form part of a systematic sociolinguistic study of the speech community. Normally, what makes these interviews classic is that they represent classic solutions to the problems of achieving cross-cultural contact, reducing the effect of the ‘observer’s paradox’ and approximating the vernacular of everyday life. Most importantly, these interviews were conducted with extraordinarily gifted, memorable and fluent speakers.

Strassel *et al.* (2003) add that the corpus includes the complete interview recordings plus time-aligned verbatim transcripts for each speaker. Also included in the publication is a sociolinguistic variable survey that represents an overview of the intra- and inter-speaker variation attested in the corpus, highlighting a different range of phonological, phonetic, grammatical, lexical and stylistic variables. Finally, the publication includes a number of annotation tools that allow users to listen to each interview while browsing the corresponding transcripts and to display and hear each token identified in the variable survey. These tools can be extended to create new time-aligned transcripts or tag additional variables within the existing corpus.

Moreover, according to Strassel *et al.* (2003), the SLX Corpus was developed as part of the Data and Annotations for Sociolinguistics (DASL) Project. It is an examination of best practices in the use of digital speech corpora for the study of language variation. Containing classic interview material in what Strassel *et al.* call the ‘Labovian’ tradition, it is a valuable teaching tool for linguists. The recordings demonstrate successful interviewing techniques, the sound quality is high and the digitisation, segmentation and transcription of the data represent best practice in these areas. The variable survey highlights over 150 sociolinguistic variables attested in the corpus and suggests avenues for further research. Most importantly, the SLX Corpus provides both an example of a digital speech corpus developed specifically to support sociolinguistic research and a stable benchmark for training in sociolinguistic data collection, digitisation, segmentation, transcription, analysis and publication.

In addition, Kjellmer (1986) in his research used the Brown and LOB corpora to examine the masculine bias in American and British English. He analysed the occurrence of masculine and feminine pronouns and the occurrence of the terms *man/men* and *woman/women*. The research showed clearly that the frequencies of the female terms were much lower than the frequencies of the male items in both corpora.

The remark was made that the female items were more common in British English than in American English. A further supposition of Kjellmer, which was not a result of the analysis of the corpora, was that woman would be less ‘active’ and would more frequently be the *objects* rather than the *subjects* of verbs. Normally, according to statistics, men and women have analogous subject/object ratios.

Another idea around the sociolinguistics of language is introduced by Holmes (1994) who makes two important points about the methodology of these kinds of study: On the one hand, when classifying and counting occurrences, the context of the lexical item should be taken into consideration. In this regard, observation shows that while there is a nongender alternative for *policeman/policewoman*, namely *police officer*, there is no such alternative for the *-ess* form in *Duchess of York*. When analysing the gender bias in writing, from counts of ‘sexist’ suffixes, one sees that the second form should therefore be excluded.

On the other hand, Holmes (1994) points out the degree of difficulty when trying to classify a form when it is actively undergoing semantic change. Her observation is that the word *man* can either refer to a single male, as in the phrase *A 35-year-old man was killed*, or can have a standard meaning that refers to humanity as a whole, as in *Man has engaged in warfare for centuries*. In fact, in expressions such as *We need the right man for the job* it is confusing to determine accurately whether *man* in this context is gender specific or could be used alternatively with *person*. Holmes came up with the assumption that these specific indicators should motivate a more critical approach to data classification in further sociolinguistic work using corpora, both within and without the area of gender studies.

In other words, while the initial focus of research was on generalised gender differences, more recent studies have acknowledged and begun to explore diversity among women and men; this parallels more general developments in gender theory (cf. Mesthrie, Swann and Deumert. 2000). Many studies in the domain of gender have clearly shown that language functions as a kind of social mirror, reflecting social distinctions. Already in 1944 Furfey argued that the existence of different female and male forms of language meant that speakers were conscious of women and men as different categories of human beings.

Another opinion that developed in that field of research is the one of Mesthrie *et al.* (2000), also expressed by Fasold (1990). They are all in favour of the idea that during human communication, some aspects within the definition of language imply attention to the way language is played out in societies in its full range of functions. Language for these authors is not just denotational; it is also a term that refers to the process of conveying meaning, referring to ideas, events or entities that exist outside language.

While using language primarily for this function, a speaker will inevitably give off signals concerning his or her social and personal background. Consequently, language is said to be indexical of one's social class, status, region of origin, gender, age group, and so on. In a sociolinguistic sense, this indexical aspect of language refers to certain features of speech (including accent) that indicate an individual's social group; the use of these features is not exactly arbitrary since it signals that the individual has access to the lifestyles that are associated with that type of speech.

All these ideas revolve around the fact that the lexicographer has to plan the investigation of the data carefully to avoid discontinuities in the source of materials. In the above paragraphs, one has seen the benefit of social characteristics in corpus lexicography as they are helpful in both the organisation and analysis of data in order to enhance the quality of lexicographic representation. A fair balance between men and women should be observed in such a way that both sides are equally involved in the project in order to make a positive contribution to avoid biased results. The corpus lexicography will have everything to benefit by considering these two groups because the language is spoken mutually by them and they fully represent the community of speakers. If a corpus contains only materials coming from one of the genders, it will not be balanced enough to represent the way in which the language is spoken by the community.

The involvement of the community in the research project will enable the identification and inclusion of the different dialects of the language in question. Normally, this applies to languages that have clear dialect components. In the case of the Gabonese languages, Fang (Mazona) and Omyene (Myene) are the only two languages in which various dialects can be identified, and an exhaustive presentation of those dialects is given in the next chapter. For the representativity of a language, dialects should also be included even in the gabonese context. Thus, a satisfactory dictionary is one that, among other things, takes into consideration all the varieties of

a language. That will guarantee the coverage of that language in all the dimensions and the quasi-totality of the lexical items or vocabulary of the language. This can be achieved only if the corpus incorporates data from regions in which the language is spoken. The implications of such statements will be observed in the next chapters when dealing with dialects in corpus lexicography.

For African countries such as Gabon where written sources are rare, it is important to concentrate on oral ones because of the strong oral tradition. The success of corpora based on both oral and written sources in many dictionary projects, such as COBUILD and the British National Corpus (BNC), is apparent. It is therefore very important for the Gabonese languages that the necessary steps be followed because no good dictionary can be compiled nowadays without a good corpus.

### **2.3 Principles of data collection**

Smit (1996) points out that according to Schaefer (1979:356), lexicographical corpora have an important purpose. Schaefer adds that a text corpus is a finite set of texts in natural language collected for the purpose of linguistic or literary research. For him the texts in such collections may be

- systematically collected and ordered;
- in one or more languages;
- selected according to diachronic or synchronic points of view;
- containing examples of standard language or other varieties; and
- written or spoken

The material collected should be comprehensive or representative enough to cover the language to a large extent and also to be useful for the compilation of different dictionaries. This will allow and ensure the flexibility of the corpus and enlarge its aim. Therefore, it is quite difficult for any corpus to serve a general purpose if it is based on a restricted compilation.

A restricted corpus will limit the chances of using it in a comprehensive way but will promote its utilisation in a special situation. For instance, it will be convenient for the

compilation of dictionaries for special purposes. In this regard, mathematics, biology and linguistics, for example, as special target fields can be covered by such a corpus.

The different corpora should target not only language for general purposes but also language for special purposes. A corpus compiled in such a way, will be able to take into consideration all aspects and situations of the language in which the lexicographer is working in order to avoid leaving aside some important words. Among these words can be included words that are commonly used by the community of speakers but the corpus could not bring out because of the restriction of the material.

In other words, the lexicographer could miss words that may occur very frequently in the language but do not exist in that corpus. In order to fulfil efficiently the dictionary functions, the lexicographer must make sure of the usage of all the varieties of the language. It would be better to have standard forms as well as nonstandard ones from both oral and written sources.

Furthermore, the collection of material is done according to the quality of the planning of the language material policy, which is dynamic in the sense that it is drawn up with the following considerations in mind:

i) The type of dictionary to be compiled (e.g. monolingual, bilingual, technical or comprehensive).

If a comprehensive dictionary is to be compiled, the language material collection policy would seek to be representative of a wide range of media and sources.

(ii) The target reader or market (e.g. primary school pupils, high school pupils, specialists or language learners).

If the target reader is a primary school pupil, a language material policy that includes highly technical material in the corpus would be inappropriate.

(iii) The state and composition of the existing language data (e.g. supplementing existing data).

If the existing language data reflect imbalances in its composition, for example too many items from a certain period or by a certain author, the policy should be adjusted accordingly.

(iv) The flexibility of the database (e.g. new or existing projects utilising the same database or adding new types of data to the existing database).

If the policy is too restrictive regarding the information types of the material collection, the possibility of compiling other types of dictionary from the same corpus is also restricted. On the other hand, building up language data containing many information types may take more time than is available to complete a project. Consequently, a balance should be maintained between complexity (and the potential use of the corpus in compiling other dictionaries) and the time available to complete the initial project.

## **2.4 Corpus usage**

In order to be useful in a multilingual environment such as Gabon the corpus should be compiled on the basis that it has a multifunctional purpose. It is primarily aimed at the compilation of all types of dictionary: bilingual, monolingual, specialised, etymological, and so on. In addition, the corpus has to be designed in a comprehensive manner in order to meet the goals targeted by the dictionary compilers.

It is not beneficial for a corpus to be used only for one dictionary because, as said earlier, it must avoid restrictiveness. Rather, it is important that more than one dictionary can be compiled from the same corpus and therefore the corpus has to be designed accordingly. By doing so, lexicographers are targeting other purposes as well. Then the corpus instead of being used for the compilation of dictionaries only can also be useful to cover research in other domains. In the environment of the object language, the corpus can enable both stylistic analysis and the compilation of grammar books, which are very important tools for the development of languages because they can facilitate the identification of different lexical categories. The same corpus can be conjointly used for linguistic research more directly in sociolinguistics and dialectology studies. More generally, it will serve as a general source of language material.

Therefore, the aim is to encourage the compilation of a well-designed multifunctional corpus that can be used and adapted for a specific purpose. Such a corpus, if large enough, can be used to illustrate some specific points in lexicography. For instance, it

will be easier to deal with recurrent problems in lexicography such as homonyms if the context will help one to identify different meanings or senses (polysemy) to illustrate the relation between different senses of words. More precisely a well-designed corpus will be helpful for the improvement of the macrostructure and the microstructure of the dictionary. Furthermore, it will also help at a theoretical level to distinguish between lexicographic data and data for other disciplines.

Bo Svensén (1993) states that a corpus based on secondary material can be used in three different ways:

- as a basis for the dictionary as a whole, in cases where no special demands require that it has to be based on primary sources
- to supplement the primary material that forms the core of the dictionary
- to supplement a dictionary that is to be revised and enlarged

In addition, a corpus is helpful for the compilation of a frequency list, a lemma list, concordances, spelling checkers and rulers. All these terms will be discussed in detail in the coming chapters.

## **2.5 Corpus types**

In lexicography or language research there are different types of corpus, but Sinclair (1996) points out that so far little consensus as to what counts as a corpus and how corpora should be classified has emerged. Up to now, this position has not changed much. For him, the most important areas are

- the minimum conditions for any collection of language to be considered as a corpus; and
- the separation of corpora that record a language in ordinary use from corpora that record more specialised kinds of language behaviour.

Furthermore, there are many collections of language called corpora that do not meet these conditions, and there are some corpora available that record special and artificial behaviour. It is clear that, there are many types of dictionary that serve different functions and objectives according to the goals of the compilation thereof. The need for different types of corpus is already shown in the typology of dictionaries to be



compiled by the lexicographer. Normally, a specific type of dictionary should correspond to a specific type of corpus because of the data and the information required in a dictionary.

In linguistics or lexicography, many corpus types exist but here the focus is on the main types, as discussed in the following paragraphs.

### **2.5.1 Monolingual and multilingual corpora**

The monolingual corpus can be divided into two different types: The language for general purposes corpus or LGP corpus is designed for the compilation of dictionaries that provide a description of the language in general use: They are usually called general dictionaries or LGP dictionaries. It is also termed a reference corpus (Fail 2004). This kind of corpus generally contains millions of words and is usually representative of many aspects of a language.

The language for special purposes corpus or LSP corpus, designed for specialised dictionaries, is restricted to a set of linguistic elements from one or more specialised subject fields. In this category, one can also identify monolingual and bilingual corpora, which can help in the compilation of a specialised dictionary.

The LSP corpus is smaller than the LGP one and is quite often a private initiative. In this category what is known as a multilingual corpus can also be included. The multilingual corpora may indeed be *parallel*, meaning that they gather different translations of the same texts (or sometimes of texts written in different languages about the same subject).

A typical example, (and fairly well known in the corpus research community) is the bilingual English-French corpus of the debates in the Canadian Parliament known as a Canadian Hansard Corpus (cf. <http://www.isi.edu/natural-language/download/hansard/>.) Furthermore such corpora may, be ‘aligned’, which means that the corresponding position labels are inserted in the texts of each language, which allows relating each segment (paragraphs, even sentences) of text across the languages of the corpus. According to Gautier (1998), such corpora are suitable for language translations and the development of bilingual and multilingual dictionaries.

### **2.5.2 Synchronic and diachronic corpora**

This section deals with the problem of the choice of texts. The synchronic corpus gathers texts from the same period in time (practically, one 10-year window or several times this period at most) and bears a more or less precise testimony of the state of a language. A suitable reference of such a type is the *Cobuild* dictionary.

The diachronic corpus follows the evolution of the language during a chosen period, which may well be several centuries long. An example of this type of corpus is the ARTFL corpus of the French language, built through collaboration between the French national research body CNRS and Chicago University. It stretches from the 16<sup>th</sup> century to the present (URL: BALL, 1997), cited by Gautier (1998).

### **2.5.3 Spoken and written corpora**

The literature around these expressions shows considerable confusion over their use.

First, the term *spoken corpus* is to be distinguished from a speech corpus, which is a collection of recordings. Then there is a choice. The term *spoken* sometimes means a corpus of informal, impromptu conversation with no media involvement. On the other hand, it is used by some to mean any language whose original presentation is in oral form; that is, the speakers involved behave in oral mode.

If such a text is later presented in written form without change except for the transcription, it should be classified as spoken, a BBC Reith Lecture, for example. If, in time, a spoken corpus can be stored as sound waves as well as transcript, such a text may exist in two versions and a special kind of parallel corpus can be introduced (Sinclair, 1996). Similarly, Sinclair is in favour of the idea that any text composed to be presented in written form can be read out loud but its expression needs only change in ways required by the change of medium. It is, therefore, primarily a written text.

### **2.5.4 Samples**

Sampling has become a popular technique used in the assembly of corpora following the typical example of the Brown model. Samples are small in relation to texts such as newspapers, books and radio programmes and of a constant size (Sinclair 1996).

There is a clear difference between a text corpus or a whole text corpus and a samples corpus. According to Sinclair, this feature is just a remnant of the early restraints on corpus building and it confers no benefit on the corpus. The use of samples of a constant size gains only a spurious air of scientific method.

### **2.5.5 Reference corpora**

A reference corpus is one that is designed to provide comprehensive information about a language. It aims to be large enough to represent all the relevant varieties of the language and the characteristic vocabulary so that it can be used as a basis for reliable grammars, dictionaries, thesauri and other language reference materials.

The model for selection usually defines a number of parameters that provide for the inclusion of as many sociolinguistic variables as possible and prescribes the proportions of each text type that is selected. A large reference corpus may have a hierarchically ordered structure of components and subcorpora (Sinclair 1996). The reference corpus is a central pool or a dam in which all the types of data are kept in different files from various sources and genres.

A reference corpus can also be used in order to measure the deviation of special corpora. Reference corpora are at the heart of the future development of corpus-based work in Europe and elsewhere. One of the well-known projects of this kind is the Bank of English that consists of about 415 million words.

### **2.5.6 Monitor corpora**

It is becoming clearer and clearer that a limit on a corpus for any length of time is an unnecessary restriction. For some uses, it is essential to achieve a steady corpus size and constitution, but this is easy to devise within a large and constantly moving collection. One can ask the question about how to manage the large quantities of data that are foreseen for what is known as a monitor corpus.

According to Sinclair (1996), one of the first things to bear in mind is a corpus of a constant size so that the software of the day could cope with it. Such a corpus would be constantly refreshed with new material while equivalent quantities of old material

would be removed to archival storage. The constitution of the corpus would also remain parallel to its previous states.

That puts forward the idea of rate of flow as the best way of managing the corpus. Instead of setting, say, 10 million words as the proper proportion of a genre, the setting could just as easily be 10 million words a year or a month or a week. The language would then flow through the machine so that at any one time there would be a good sample available, comparable to its previous and future states.

Such a model opened up new prospects for those interested in natural language processing, and it added another dimension to contemporary corpora: the diachronic. New words could be identified and movements in usage could be tracked, perhaps leading to changes in meaning. Long-term norms of frequency distribution could be established, and a wide range of other types of information could be derived from such a corpus.

Some scholars were less than happy about disposing of the older texts as new ones came in. That problem, however, was solved by the fast-expanding power and memories of machines. There is no need to remove any text from modern systems. However, to manage a monitor corpus to the best advantage, it is convenient to divide it into batches of similar size and constitution.

Moreover, Sinclair (1996) believes that after a while, the balance of components of a monitor corpus will change. New sources of data will become available and new procedures will enable scarce material to become abundant. The rate of flow will be adjusted from time to time.

### **2.5.7 Parallel corpora**

A parallel corpus is a collection of texts, each of which is translated into one or more languages other than the original. The simplest case is when only two languages are involved: One of the corpora is an exact translation of the other. Some parallel corpora, however, exist in several languages. The direction of the translation also need not be constant, so that some texts in a parallel corpus may have been translated from language A to language B and others the other way around. The direction of the translation may not even be known.

Parallel corpora are objects of interest at present because of the opportunity offered to align original and translation and gain insights into the nature of translation. It is hoped that tools to aid translation would be devised. Probabilistic machine translation systems can, moreover, be trained on such corpora. Parallel corpora are created in the business of communication in multilingual societies, such as the United Nations, NATO, the EU and officially bilingual countries such as Canada (Sinclair, 1996).

### **2.5.8 Comparable corpora**

A comparable corpus is one that selects similar texts in more than one language or variety. There is yet no agreement on the nature of the similarity because there are very few examples of comparable corpora. One of the clearest is the International Corpus of English (ICE) (Greenbaum 1991) Corpora of around one million words in each of the many varieties of English around the world are being assembled following the same model, which prescribes genres and the target quantity of words to be gathered in each. Originally, the corpora were all to be gathered in the same year.

The purpose of a comparable corpus is to compare different languages or varieties in similar circumstances of communication while avoiding the inevitable distortion introduced by the translations of a parallel corpus (Sinclair, 1996).

### **2.5.9 Multilingual corpora**

At present there are not many multilingual corpora apart from parallel and comparable corpora. There are plenty of centres that have collected text material in several languages and some of these collections are corpora in their own right, but unless the collections share common features of selection, at least at the level of the comparable corpus, they are just text resources in different languages. It therefore seems unhelpful to use the term *multilingual corpus* (Sinclair 1996).

To some extent Atkins, Clear and Ostler (1992) have drawn up a similar corpus typology. Nine corpus types have been identified, sometimes using a different term to indicate them. In principle, they give some additional information that can be used on the types given above. The aim here is to mention their existence rather than to give

another classification. In the meanwhile, they match with the text typology or sources of material, written or spoken, which are dealt with in the next sections. The following types can be noted:

Types: FULL TEXT SAMPLE MONITOR

For *full text*, each text in the corpus is unabridged; for *sample*, size is to be defined, as well as the location of the sample within the full text and the method of selection of samples; for *monitor*, texts are scanned on a continuing basis and ‘filtered’ to extract data for the database but not permanently archived.

Types: CLOSED OPEN-ENDED

Types: SYNCHRONIC DIACHRONIC

A specific period must be designated for a synchronic corpus; this requires research into how short that period may be if the corpus is to be considered synchronic.

Types: GENERAL TERMINOLOGICAL

Terminologists must define conditions that must be created if a corpus is to be valid for terminological use; this forms part of text typology.

Types: MONOLINGUAL BILINGUAL PLURILINGUAL

Types: LANGUAGE(S) OF CORPUS

English, Russian, Japanese, Yipunu, Yilumbu and so on can constitute the language(s) of a corpus.

Types: SINGLE PARALLEL-2 PARALLEL-3...

Are all the texts in the corpus stand-alone or part of a parallel pair/trio, and so on of translated texts? (This applies only to bi- or plurilingual corpora.)

Types: CENTRAL SHELL

The central corpus is a selected body of texts, of manageable size, big enough for normal purposes. The shell, which may be the remainder of the electronic text library (ETL), is available for access when necessary.

Types: CORE PERIPHERY

These concepts are discussed by Leitner (1990) in relation to the ICE. The ‘core’ contains text types that are common to all varieties of English and therefore are

present in all the subcorpora while the ‘periphery’ contains text types that are specific to some subcorpora only.

All these different types of corpus have to be taken into consideration when planning the compilation of a dictionary in order to avoid confusion. It then becomes clear that each dictionary has to correspond to a specific type of corpus that one needs to adapt for a particular purpose. By doing so, the lexicographer would be able to achieve the genuine purpose the dictionary aims to serve.

When planning a dictionary for the Gabonese languages, the emphasis should be on the corpus on which the dictionary will be based. It is important that the planning phase of the dictionary also take the type of corpus into account. It is true that a multipurpose corpus is the ideal solution to avoid dispersion but such a corpus needs to be adapted accordingly to be workable. Languages are different from one another in such a way that one must always effect changes and adaptation here and there because what is good for one can cause problems for another. That will appeal to the lexicographer’s accuracy, knowledge and professionalism.

Hess *et al.* (1983:9–10), quoted by Smit (1996:126–127), stipulate that one can detect in every corpus an attempt to register a specific temporal or more or less discernible condition in a language. A dictionary based on such a corpus describes the vocabulary within this specific condition. It is possible to define ‘language’ within this context from a geographical, dialectal, historical or a specific person’s point of view. Thus the corpora would be compiled to contain texts from certain areas where a language is spoken, a certain era in which a language is or was spoken, a certain text type or a specific author. For all types of corpus, except for author corpora and corpora dealing with dead languages, the principle holds good that the corpora will always consist of a selection from the entire set of available texts.

### **2.5.10 Towards hybrid corpora**

The term *hybrid corpora* here refers to corpora that combine the features of the above types. Hybrid corpora should be more general by containing texts from every sector and domain, and the multifunctional feature should be their main characteristic. In other words, hybrid corpora must be used for the compilation of the different

dictionary types that the lexicographer intends to compile. That means that all the data needed for each dictionary should be extracted from the hybrid corpus. In fact, such corpora will contain historical data and synchronic data that will help with the compilation of historical and synchronic dictionaries.

More generally, hybrid corpora will accommodate data of a general nature that are necessary for the compilation of general monolingual dictionaries, that is general monolingual dictionaries with a large macrostructure and extensive microstructural treatment and the ones with a restricted macrostructure and limited microstructural treatment. In these two categories all the dictionaries of monolingual character can rightfully be included, from desk dictionaries to standard dictionaries. Bilingual and multilingual dictionaries can also be compiled using the same corpora, and after determining the number of languages, one can initiate the translation of the original corpora into that language or those languages.

At least, the hybrid corpus can help to solve the problem of having several corpus types that one cannot use efficiently for the compilation of dictionaries and it will ease the lexicographer's task. It will save energy and bring more focus on the work that needs time and patience. It seems easier to manage one large corpus than many smaller corpora. It is advisable for lexicographers to start compiling large single corpora for the compilation of dictionaries. Compiling a corpus is an ongoing activity that takes years to be completed. As long as languages are dynamic, one will not have a corpus that captures the whole vocabulary of a language.

## **2.6 Dictionary basis**

Compiling a dictionary basis is part of the general preparation phase of the dictionary conceptualisation plan, which lays the foundation for the structure, contents and presentation of the final product (Gouws 2001:67). It is the collection of sources from which the corpus for a given dictionary is to be compiled. Another issue to be dealt with in the general preparation phase is the identification, establishment, nature, extent as well as the description of a dictionary basis that suits the relevant dictionary project in the best possible way.

A dictionary basis, according to Gouws (2001:68) and Wiegand (1998:139), can be described as the total of the source language material for the specific lexicographic



process. This includes all the possible sources that accommodate such material as well as informants and mother-tongue speakers of the language who can assist the editorial staff in the building up of a materials collection. The dictionary basis will differ from dictionary project to dictionary project according to the typological nature of a dictionary.

According to Gouws (2001) and Wiegand (1998), the dictionary basis of a general monolingual or bilingual dictionary can be compiled from three types of source:

- Primary sources will usually be texts. The dictionary basis of a dictionary compiled for a language with a strong oral tradition can also use recordings of the orature as primary sources. Most of the African languages in general and the Gabonese languages in particular in this case can take advantage of their oral tradition to compile a well-designed dictionary basis, using such material effectively.
- Secondary sources are all the available dictionaries in a given language. It is important for lexicographers to be very careful not to perpetuate lexicographic failures of the past in the new dictionary projects.
- Tertiary sources comprise all other linguistic material that can be used, for example linguistic monographs, papers and grammars.

According to Gouws (2001:69), an early identification of the dictionary basis enables the lexicographer to apply a well-directed materials collection policy, which in its turn allows a more rapid macrostructural selection. All the established projects for the African languages and those at the starting stage need to invest energy and time in the building of a dictionary basis. The fact that many of these African languages have an oral tradition could be one of the difficulties to face when compiling a corpus for those languages. It is therefore easier for the languages with a written tradition because sources are already available.

This restriction in written material is in contrast with the availability of numerous oral sources. In many cases, a corpus compiled only from written material will not be fully representative of the lexical stock of the language. Consequently, after having determined the target user and the typological category of a dictionary to be compiled, lexicographers can ascertain whether the written sources will be sufficient to render the dictionary basis needed for the specific project or whether other sources. In this

regard, mother-tongue speakers who can convey orature, should be consulted to complement the written sources as a component of the dictionary basis.

The dictionary basis is part of the general preparation phase, which is itself one of the five subdivisions identified by Wiegand (1998:151) in the dictionary conceptualisation plan. The four other subdivisions are identified as follows:

- The material acquisition phase
- The material preparation phase
- The material processing phase
- The publishing preparation phase

The emphasis of this study is on the three phases that will be discussed in the coming sections: the material acquisition, preparation and processing phases.

### **2.6.1 Sources of materials**

Materials needed for the compilation of any dictionary come from various sources, which are grouped into written and spoken sources. The acquisition of materials is a very important phase that precedes the compilation of any dictionary. It is of a great importance to identify the material the lexicographer will utilise in the compilation process of the dictionary and plan the gathering thereof very well. The material can come from diverse sources such as mother-tongue speakers of a given language, works of reference and any single entity or domain.

#### ***2.6.1.1 The material acquisition phase***

Gouws (2001:69) states that the material acquisition phase precedes the compilation process and focuses on the gathering of speech material from the sources earmarked for the dictionary basis. A result of the material acquisition phase is the compilation of the database or lexicographic corpus, namely the collection of items gathered from primary, secondary and tertiary sources.

Gouws (2001) adds that in modern-day lexicography the material acquisition phase will inevitably lead to a corpus. No modern dictionary can be representative if it is not

based on a reliable corpus. The compilation of corpora has to be regarded as a highly skilled activity and ample provision needs to be made for this during the planning phase of any lexicographic endeavour. During the material acquisition phase the infrastructure for corpus building and therefore also the infrastructures regarding computational aspects have to be in place. A model for the compilation of dictionaries determines the need for a dictionary plan focusing on much more than the mere compilation process.

#### ***2.6.1.2 The material preparation phase***

Gouws (2001:69) suggests that with regard to the material preparation phase, lexicographers have to prepare the collection of material for the next steps of the lexicographic process. The materials recorded orally, for instance, have to be transcribed and scanned into a computer for eventual inclusion in the corpus. Throughout this phase, the opportunity is given to the staff of the dictionary project to sort the material in order to remove material that cannot be used. At some stage a corpus should be in sufficient order to allow the selection of the necessary data to be included as verbal illustrations in the dictionary article, such as citations and examples.

After sorting out the corpus lexicographers can proceed with the macrostructural selection to present the lexical items to be included as lemmata in the dictionary. Such work is done on the basis of the different types of dictionary available. The lemmata function as the most typical treatment units of a dictionary and once these treatment units have been selected, ordered and presented as guiding elements of their respective articles, lexicographers are in a position to apply the lexicographic treatment by activating the microstructural programme. The way in which this is done will be fully explained when dealing with the compilation of the dictionary based on the corpus throughout the work.

#### ***2.6.1.3 The material processing phase***

The material processing phase has a direct link with the application of the data distribution structure and the writing of the dictionary texts. According to

Bergenholtz, Tarp and Wiegand (1999), quoted in Gouws (2001), the data distribution structure of a dictionary determines the specific position of each data type in the dictionary as a so-called carrier of text types. Some data will be included in the texts accommodated in the front and back matter while other data will be part of dictionary articles, the texts constituting the central list of a dictionary. Most of the data used in the data distribution are taken from the corpus since it is the main constituent of the dictionary basis. This data are used in both the macrostructure and the microstructure.

In order to continue with the construction of the dictionary articles as texts, it is important as a first step to plan the microstructural programme and to finalise the choosing of the macrostructure. Gouws (2001) points out that the lemma functions as guiding element of each dictionary article and that the microstructural programme orders the entries included as part of the treatment of the lemma in such a way that they can be divided into two major components: the comment on semantics and the comment on form. Every data category included in the microstructure of a dictionary belongs to one of these components. The distinction between the comment on form and the comment on semantics applies to all general bilingual and monolingual dictionaries.

When dealing with a specific type of dictionary and type of data a comprehensive treatment will be applied to both the comment on form and the comment on semantics, when possible and necessary. Once again, the different sources from which the corpus is built are divided into written material and spoken material used as dictionary basis for the compilation process of the dictionary.

### **2.6.2 Written materials**

It is commonly accepted that written sources are classified as primary sources, as seen earlier. All published and unpublished books or documents in a specific language are considered as written materials. This statement can raise the problem of what is going to be considered as a book or a document when one knows that texts are written by individuals and others by several authors. Atkins *et al.* (1992) have already answered all these questions. In this study, every written document, published or not, in a given language is considered as written material. Therefore, the following nonexhaustive list of elements can constitute written sources (Atkins *et al.* 1992):

- Small advertisements in newspapers. The corpus builder might prefer to make a collection of these small advertisements and treat them as one text.
- An article in a newspaper or magazine. It may be convenient to treat one issue of a newspaper (and single issues of other periodicals) as one text.
- An academic Festschrift, scientific journal, and so on where the bibliographic data apply to the whole book but the papers differ linguistically to such an extent that they might be best treated as discrete texts.
- A poem. It is often more convenient to gather many short poems by the same author into collections and to treat each collection as a text.
- Published correspondence in which the letters are the products of two authors but constitute a single discourse. The critical apparatus, introduction and editorial material will be yet another author's intervention. The way in which this problem is handled will depend on the requirements of the corpus.

According to Atkins *et al.* (1992), the potential written sources to be included in a corpus are of an infinite number and can cover many areas and subjects. Among them are the following:

- Any written book in the language
- Periodicals
- Poems of any kind
- Novels
- Short stories
- Plays
- Essays
- Letters
- Advertisements
- Regulations and laws
- Articles
- Advice columns

- Horoscopes
- Announcements
- Reports
- Commentaries
- Features
- Advice programmes

They add that all these genres and types cover a large range of domains such as the following:

- Education
- Work
- Leisure
- Public affairs

The following functions can also be performed:

- Narrative
- Informative
- Expository
- Hortatory/persuasive
- Regulatory/instructional
- Reflective (giving one's opinion)
- Entertaining

The subject of those texts can be related to the following:

- Science
- Biology
- Chemistry
- Music
- Orchestra

- Opera
- Animals

In the same way, Landau (2001) says, text categories refer to the sources from which the material has been obtained, such as newspapers, magazines and journals, documents and personal communications such as private letters and diaries. By documents he means any type of miscellaneous printed material that is not a book, magazine or newspaper. Documents include, inter alia, junk mail (even though in the form of a letter), instructions for assembling a toy or piece of furniture, companies' annual reports, warnings on drug labels, signs and public notices, government pamphlets, legal notices such as jury summonses and the advertisements on cereal boxes.

Landau (2001) adds that genres usually apply to subject categories such as natural or applied science, politics and world affairs, business and economics, the fine arts, entertainment, *belles lettres* and popular fiction. He does not find many of these text types very practical. Landau thinks that many nonfiction text sources are hard to qualify as belonging to any one subject. Not only are they difficult (in many cases impossible) to label accurately, but one is also sceptical as to how useful they would be even if they were labelled. In examining the results of a corpus search, one can easily see from the header information and the text itself what the subject deals with. If all the hits one sees are in a business context, one knows that the subject is business without having to rely on a code for that genre.

In addition Landau (2001:279) reports on the genres used by the Brown Corpus alongside with the number of 2 000-word samples in each, as shown below:

- Press: Reportage
- Press: Editorial
- Press: Reviews
- Religion
- Skills and Hobbies
- Popular Lore
- *Belles lettres*, Bibliography, etc.

- Miscellaneous
- Learned and scientific writings
- Fiction: General
- Fiction: Mystery and Detective
- Fiction: Science
- Fiction: Adventure and Western
- Fiction: Romance and Love story

All these domains are of great importance when gathering written sources for a corpus. If all these elements could be found in all the languages in the world it would be very helpful, but it is easy to find them in a society or language with a well-established written tradition, such as European languages or some languages in Africa, for example Afrikaans or Kinyarwanda. For the rest of the languages, especially African languages, the emphasis should be on oral sources.

In the corpus of this study, the focus has mainly been on the digitalisation of the Bible, the most valuable written source found in the Yipunu language for now. If available, other written sources in the language will be added in future. For completeness of the sources of material to include in a dictionary, it is important to consult and use the existing dictionaries and linguistic books available in the language for their inclusion in the corpus if the material is needed. Therefore, in Yipunu the *Dictionnaire Français-Yipounou Yipounou-Français* should be taken into consideration as well as the grammar books published in the same language. This strategy will also be applied when dealing with the other Gabonese languages and adjustments should be made accordingly because these languages are different from each other and materials are specific to each language.

### **2.6.3 Spoken materials**

Like written sources, spoken ones are also primary source constituents. They are transcripts of recorded language of various origins. Since language use is primarily reflected in utterances, spoken language recordings are vital sources of language use.



According to Atkins *et al.* (1992), potentially the following elements can be registered as part of spoken sources:

- An informal face-to-face conversation
- A telephone conversation
- A lecture
- A meeting
- An interview
- A debate
- TV programmes
- Talks or speeches
- Radio programmes
- Entertainment (theatre, etc.)
- Riddles
- Tales and stories
- Parliamentary transcriptions

The above-mentioned genres of spoken material are also included in the ones given in detail by Kennedy (1998), quoted by Zang-Bié (2002). Kennedy gives a comprehensive categorisation of spoken texts, which are divided into monologue and dialogue:

### **Monologue**

- Formal texts
- Texts written to be read
- Prepared but unscripted public speeches
- Less formal texts
- Academic lectures
- Commentaries on public occasions

- Sport commentaries
- Demonstrations

### **Dialogue**

- Face-to-face dialogue
- Public discussion (e.g. on radio)
- Business or professional transaction
  - Client and professional interaction
  - Workplace interaction among colleagues
  - Commercial (sales) transaction
- Informal interaction
  - Within family
  - Among friends
- Telephone dialogue
  - Between interlocutors known to each other
  - Between interlocutors unknown to each other
- Structured interaction
  - Interview
  - Formal interaction, for example arts programme
- Less formal interaction, for example interviews with survivors of and witnesses to events such as accidents
- Debate
- Committee meeting

It is clear that this list of spoken sources is not exhaustive and it will be completed in the course of time when necessary.

In the Yipunu corpus being compiled credit will be given to all the possible oral sources. In countries such as Gabon where languages are generally based on oral activities, many events, ceremonies and so on are given orally. It is therefore advisable to record as much material as possible in order to fulfil the needs of the users. In this corpus, for the moment transcripts of some dialogues from ‘Rapidolangué’ stories and tales from Kwenzi-Mikala have been included, which give more contextual words and their usage than just collecting single words and sentences. This is not to assign less value to single words but just to explain what has been done. Kwenzi-Mikala’s books such as *Mumbwanga* and *Parémies* have been used in the corpus.

#### **2.6.4 Methods of language material collection**

There are many ways of collecting material, depending on the types of source and the purpose and use of the material. Each field of research has a typical method of collecting its sources to suit the needs of the collector. In language research, spoken and written material is dealt with, using various collection methods. Methods used by other disciplines can be beneficial to language research in certain ways. Of course, it can be strongly recommended to the fieldworker not to be bound to the existing methods. It is up to the researcher to adapt the techniques of material collection according to the situation of the field.

##### ***2.6.4.1 Methods of obtaining spoken sources***

For the collection of spoken material, it is necessary to organise fieldwork in order to interview language speakers and involve a large number of people. Fieldworkers have to be equipped with recording material and sets of questionnaires, such as the one of Greenburg-Tervuren-Welmers and others designed for specific circumstances, for example the one used by Nong, De Schryver and Prinsloo (2002) for dealing with loan words versus indigenous words in the Northern Sotho language. Fieldworkers may be members of the community carrying recording apparatus during their daily activities. They can be equipped with cassettes and digital recorders to make their work easy.

The bottom line here, says Menge (1982:550), is that researchers will have to choose the methods they want to use in accordance with their research purposes. For example, one can deal with investigations on vocabulary by effectively using the indirect method. This method is successfully used by the Bureau of the WAT as an aid to the compilation of its dictionaries in Afrikaans. Many restricted questionnaires on cards are being sent to the community and the Bureau receives them back as well.

All these methods will be experimented within the Gabonese environment and adjustments will be made when necessary in order to achieve the goal targeted by the research.

It is extremely important to obtain recordings from broadcasters, libraries, universities, parliament and everywhere where data can be found. Informants of different ages, gender, social class, educational levels, region and field of work must be selected. The recordings need to be transcribed and should be of both a spontaneous nature and prepared.

It is necessary for the recorder to participate in social activities when recording languages in oral societies. Ceremonies such as weddings, funerals, public talks, circumcisions, cultural dances and so on should be attended because huge amounts of material can be collected. When collecting data from artificial speech situations it will be better, according to Menge (1982:545), like researchers in dialectology to use the ‘classical’ method of interviewing informants to obtain information on their pronunciation, vocabulary, and so on. Informants have to speak into microphones and are recorded on tape or directly into the computer using software such as PRAAT.

Being involved in these ceremonies, the lexicographer can be directly in contact with the cultural data that need to be included in the dictionary later but in the corpus first because the corpus supplies the dictionary with all the necessary data. That is one of the reasons why the corpus once again needs to contain as much data as possible and from as many different genres and domains as possible.

A second classical method is to ask informants to tell stories while being recorded. Even if these types of conversation may be recorded in an environment known to the informant, she or he may feel obliged to use standard language instead of dialect because the researcher is an unknown person. In addition, Menge (1982:547) advises that the researcher should plan the selection of informants very carefully and that he

or she should speak the dialect or the language of the informants during the interview. If this is not possible one should hire a translator to lead the conversation. In case of secret recordings, one should be aware of the ethical problems.

Landau (2001:324) stipulates that speech is called semiscripted when it takes place in a controlled format in which certain spoken conversations are observed and in which the speakers are not free to say anything they please. He adds that of even greater value is unscripted speech, which is harder to acquire, requires a team of highly skilled people to prepare and takes much longer to convert into a usable form than written text.

Béjoint (1989:25), quoted by Smit (1996), defines informants as nonspecialists who are personally consulted about lexical items to be included in the dictionary. He contrasts them with the experts who are traditionally consulted on the meaning of specialised technical work. In addition, he says that in the compilation of small dictionaries in rare languages of which the compiler is not a native speaker, informants often have to be used.

There are many situations favourable to the collection of data the researcher can use when gathering oral material. Data can be collected from natural speech situations, as shown by Wodak (1982:541), quoted by Smit (1996). Wodak suggests that when using informants, one has to plan the situation very carefully with regard to the place and time of recording. He advises researchers not to repeat natural speech situations and adds that it is more difficult to guarantee the results of open speech situations but estimates situations found in, for example, schools, in courts of law and in business transactions to be easily interpretable.

Furthermore, Wodak (1982:541) declares that the researcher should be somewhat familiar with the people and the research environment; otherwise, the situation would not be natural. Informants may feel intimidated by the presence of the researcher and therefore the natural structure of the conversation may change. The researcher should also not invade the privacy of the informants. If it is not possible to record a natural speech situation one could resort to simulated natural speech situations, Wodak suggests.

#### ***2.6.4.2 Methods of obtaining written sources***

Written material, if available, may be obtained from libraries, publishers, universities, schools, booksellers, and so on by means of purchases, loans and donations. When possible, depending on the existence of sources, the material can be acquired, like during the COBUILD Project.

According to Renouf (1987:3), the task for the COBUILD Project was to investigate what they, themselves, their colleagues and visiting scholars regarded as the texts most typically being read in Britain and abroad. British Council Libraries around the world were asked to identify titles that were continuously popular with indigenous library users. Renouf adds that the team further consulted the best seller lists published weekly in major newspapers and the various books that appear periodically.

The acquisition of written material brings out the copyright problem since some sources are protected by law and cannot be used without permission. In this regard laws are different from one country to another and such matters must be dealt with differently according to each situation.

Landau (2001:327) thinks that the actual process of compiling a corpus involves a great deal of contact by phone or fax between the corpus acquisition editor and the various owners of the texts, most of which are under copyright. In addition, he suggests that one needs someone who is both persistent and patient in dealing with the owners, who are often busy and have more important things to do than listen to someone who wants something from them but from whom they can expect little in return.

One has to contact the owners of texts first by letter, explaining as simply and briefly as possible why the texts are needed and to what uses they will be put. Landau (2001) adds that it is wise to mention that one would be glad to cover their costs for preparing duplicate disks of the material requested. It is necessary for the corpus acquisition editor to establish personal contact with the owner or the owner's representative and convince that person why the text is so important and also to allay any fears the owner may have that the text might be compromised and copied illicitly because of its inclusion in a corpus.

Furthermore, Landau (2001:328) points out that one must obtain written permission from every owner to use the copy of the text in specified ways and for specified

purposes. The dictionary editor and corpus acquisition editor can create such a permission request, and it would not be a bad idea to show it to legal counsel before putting it to use.

Landau (2001:330) estimates that a national corpus with the prestige of government backing it can obtain almost any text it seeks but for other, independent organisations, success is not guaranteed. However, most people will be sympathetic to the request if it is explained to them as an educational tool used for research in compiling dictionaries.

The collection of material is not an easy task. One needs to plan it carefully in order to avoid mistakes that can lead the researcher to a work that does not yield good results. A team of fieldworkers has to be formed and the lexicographer has to make sure that the method of collection is synchronised and understood by everyone. The fieldworkers will then approach the communities and ask questions on several aspects, fields and areas to individuals and groups of people. In this regard, Eichhoff (1982:549) states clearly that questioning informants by means of written questionnaires is called the indirect method in contrast to interviews, which are the direct method. Some of the advantages of the indirect method are as follows:

- Informants can answer the questionnaires in their own time when they are alone and cannot be influenced by the presence of the researcher.
- The indirect method takes less time, needs fewer personnel and is less costly than the direct method. This means that one can cover greater geographical areas than with the direct method.

Some of the weak points in using the indirect method, mentioned by Eichhoff (1982), would be the following:

- One only has the normal alphabetical system of writing down linguistic data available. Nonspecialists do not have the means at their disposal to write down aspects of pronunciation, for example. Several linguistic phenomena cannot adequately be written down, such as prosody.
- When the researcher is not present, the informant cannot pose questions in case of difficulties. In addition, the researcher cannot immediately react on

insights gained by the answers of the informants or change the questions accordingly.

- One can pose considerably fewer questions in a questionnaire than in an interview.
- It is not possible for the researcher to select the right informants and to supervise the accuracy with which the informant answers the questions.
- Reactions such as hesitations, uncertainties, cheerfulness and objective comments that could provide useful insights cannot be recorded on paper.

It is also of vital importance to make clear to the owner of a text that the text will be part of a huge collection to which many others have also contributed. One must also mention that the primary use will be to analyse language use and that if any parts of the text are quoted, the quotations will be used to illustrate dictionary meanings and will be short. It also has to be emphasised that the corpus will be secure from unauthorised use and that under no circumstances will the texts be downloaded or printed out.

It is also of great importance to mention that one should not specify a word limit to quotations, promise the owners of texts a report or feedback of any kind or complementary copies of one's dictionary or promise that the entire corpus, of which their text will be a part, will not be copied. Because one cannot foresee the precise uses to which the corpus may be put, it is wise not to be more specific about such uses than one has to be.

## **2.7 Concluding remarks**

This section has dealt principally with the establishment of a data collection policy that precedes the compilation phase of any dictionary. It has also presented the different corpus categories that give meaning to the dictionary typology.

It will be difficult for African languages to base their corpus primarily on written material since the number of sources is too limited. Despite the fact that the collection of oral sources is costly and time consuming, African lexicographers face the challenge of investing themselves in the gathering of huge oral material collections in order to achieve their objectives.



One has to record a large variety of spoken material from as many different genres or topic areas as possible. When looking for material to be included in the corpus, it is very important to pay attention to the regulation in terms of the law that is in use in the specific area. The lexicographer should make the right decisions in the right environment at the right time.

It must be said again without hesitation that if Gabon is to overcome the challenges of dictionary compilation the first step would be to organise a national campaign on data collection in all the languages spoken in the nine provinces. This can be possible with the involvement of all the speakers in the community. Most important, the government should finance more projects related to the collection of material so that students and promoters at universities can plan their research activities more carefully.

Private and public radio stations and television should organise and adjust their programmes in such a way that they assign more time slots to the promotion of languages by allowing the general public to participate in debates. All the authors of novels and books written in any Gabonese language should be asked to make their works available for the research effort after making sure that it has been authorised. Researchers must attend traditional ceremonies such as funeral talks, weddings and folk dances to collect cultural data that cannot be found anywhere else. At the end of the collection the material should be transcribed and stored by trained people in order to be utilised effectively.

## **CHAPTER 3: CORPUS STRUCTURE**

### **3.1 Introduction**

This section deals with the way in which the corpus should be organised and how the data entries are linked to each other. After the collection of data, the lexicographer needs to sort out all the material in such a way that it will be useful and helpful in the compilation of the dictionary. The way in which the data will be stored and arranged will facilitate the retrieval of the information. It is the starting point of the data distribution phase. In this section, the categories of data to be included in the microstructure as part of the treatment of a lemma are discussed. It also deals with the way in which all the files should be organised, the writing system used in the corpus and how the words are combined. Some comments are also made regarding the tagging system that will be used, if there is such a need.

### **3.2 Types of data**

Corpora used nowadays in the compilation process of a dictionary make a substantial amount of data available that can be put into categories to be included in the microstructure as part of the treatment of a lemma. The lexicographer can then have at his or her disposal data for the specific dictionary and can start the dictionary conceptualisation plan in order to formulate a microstructural programme. The data categories will not be treated in detail in this section. A more comprehensive analysis will be given when dealing with the microstructure and the macrostructure in Chapter 5 in relation to dictionary typology.

According to Gouws and Prinsloo (2005:118, 119), the microstructural programme will determine the nature and extent of the microstructure, the article structure and the way in which the different slots in the article will be filled with data types. After the formulation of the microstructural programme and the complete selection of the macrostructure, lexicographers are in a position to construct the dictionary articles as texts in the central list of the dictionary. The lemma functions as guiding element of each dictionary article and the microstructural programme orders the entries included as part of the treatment of the lemma in such a way that the article displays a definite structure. The article structure can be divided into two major components: the

comment on form and the comment on semantics. Every data category included in the microstructure belongs to one of these components.

### **3.2.1 The comment on form**

Gouws and Prinsloo (2005:119) define the comment on form as the search field accommodating those data types that reflect on the form of the lemma sign that is the morphological, phonetic and orthographical form. The lemma sign is part of the comment on form because it conveys data regarding the spelling of the treatment unit. Once again, all of these will only be possible if supported by the corpus because it is very important for lexicographers to stay away from self-made constructions in languages.

It is better to use conventional principles and stick to the different ways in which theories, techniques and concepts are used to avoid criticism. Gouws and Prinsloo (2005:119) pursue the statement by saying that people often need orthographic guidance and their dictionary consultation procedure only goes as far as finding the lemma and retrieving the necessary spelling information from the lemma sign. The comment on form can also accommodate additional spelling guidance if the lexical item included as lemma has spelling variants.

The comment on form also provides users with information regarding the pronunciation of words. Gouws and Prinsloo (2005:119) stipulate that this falls under the comment on form because pronunciation has to do with the sound form of lexical items. Pronunciation can be presented in various ways and dictionaries differ in terms of the amount of pronunciation guidance on offer in a dictionary article. A typical treatment of the sound form of a word focuses on its phonetic representation and its stress pattern. Some dictionaries would give a comprehensive phonetic transcription, using the symbols from the International Phonetic Alphabet (IPA). For the Gabonese languages tone plays a very important role in this slot of the comment on form.

The comment on form can also include morphological data, discussed in Gouws and Prinsloo (2005:121, 122), as data regarding the morphology of the lemma as well as certain grammatical features. In the treatment of a lemma representing a noun, the

comment on form may include, where applicable, entries indicating morphological data such as the plural and the diminutive forms.

As far as the part of speech is concerned, according to Gouws and Prinsloo (2005:124), dictionaries are often consulted for verification of the part of speech of the word represented by the lemma sign. In the planning of their dictionaries, lexicographers need to give a clear indication of the extent of their presentation of part of speech data.

It is important that the phonetic transcriptions in dictionaries follow a certain norm. Using the IPA for the transcription process in dictionaries will help users to become familiar with the phonetic symbols compiled for almost all the languages in the world. Giving an orthographic transcription could be a solution but that specific transcription does not provide the user with the right way in which vowels and consonants are pronounced.

It is true that orthographic transcriptions could help to give an indication of word division and the point where the stress should be observed for a given word. Lexicographers are urged not always to avoid the use of scientific methods such as phonetic transcriptions just for the sake of pleasing users. The use of scientific methods ensures the quality of the product in the end. In this regard, if the users are not familiar with the special diacritics, they can be adequately trained in order to adapt to the situation.

### **3.2.2 The comment on semantics**

The concept comment on semantics is defined, by Gouws and Prinsloo (2005:125), as the search area accommodating those data types that reflect on the semantic and pragmatic features of the lexical item represented by the lemma sign. This component typically presents a range of data types. The nature and the extent of the comment on semantics are also determined by the type of dictionary, the dictionary user and the situations of usage. This includes mainly subcomments on semantics, definitions and translation equivalents.

In the comment on semantics data such as cotext and context entries can be integrated. For Gouws and Prinsloo (2005:127), if text production is a function of the dictionary

one would like to compile, the lexicographer has to assist the user to use the words presented by the lemma sign and the translation equivalents in active communication. This could be done, among others, by means of entries giving the relevant context or cotext for the lemma and the translation equivalents. The context of a given word can be regarded as the pragmatic environment in which it is typically used whereas the cotext refers to the syntactic environment in which it is typically used. This is usually indicated by means of illustrative example material such as collocations and example phrases and sentences. Context and cotext entries play an important role in both monolingual and bilingual dictionaries.

Lexicographic labels are also part of the comment on semantics. Gouws and Prinsloo (2005:129) assert that lexicographic labels are frequently employed in the comment on semantics to give explicit contextual guidance. As pragmatic markers, labels are used to relate an item in a dictionary to the world outside the dictionary and to mark either a macro- or a microstructural item in a dictionary. When a label is used to mark a lemma sign it implies that all the senses of the word represented by the lemma sign fall within the scope of the label.

Another category of data is etymological data, which Gouws and Prinsloo (2005:132) qualify to be part of the comment on semantics since it does not include only semantic data but all the data in a dictionary article not related to the form of the lemma. Many dictionaries contain entries giving some guidance regarding the origin of the word concerned. These etymological entries also form part of the comment on semantics.

In relation to what has been said above, a corpus that is not able to supplement such material is of less use since according to Schaefer (1979:359), quoted by Smit (1996), a text corpus has the following important functions or purposes:

- It forms material that one can use to formulate hypotheses on the object domain.
- It serves as a database for grammatical and lexical description as a complex empirical-descriptive act. This complex act consists of observation, comparison of phenomena within a language, segmentation and classification.
- It provides material for grammatical argumentations.

- It provides material on the basis of which one can confirm or refute hypotheses and test the adequacy of conclusions that were drawn from certain suppositions.
- It serves as a collection of citations that illustrate certain usages of words and syntagms.

All these criteria are basic for general research in language but do not bring out all the important features of the envisaged corpora one is looking for. As the main focus of the model proposed here is the compilation of dictionaries, the researcher agrees with these functions to some extent but the following features need to be taken into consideration in order to better fulfil the purpose. Apart from the above features, the corpus also should be able to assist the lexicographer in the following:

- Compilation of different types of dictionaries (monolingual and multilingual)
- Microstructural treatment of lexical items
- Disambiguation of senses of different lexical items
- Identification of the different cotext and context of the lexical items
- Retrieval of different collocations
- Definition of the lexical items
- Identification of the different equivalent relations

There are some other features that can be regarded by lexicographers as important for their different corpora that one has to identify when necessary and according to one's defined purpose.

This is confirmed by Renouf (1987) to whom, when constructing a text corpus, one seeks to make a selection of data that is in some sense representative, providing an authoritative body of linguistic evidence that can support generalisations and against which hypotheses can be tested.

It must be mentioned that this treatment is an indication of the different types or, better said, some types of data category one needs for the treatment in the microstructure. This data are of importance in the organisation of entries and dictionary articles. A comprehensive treatment will also be given when necessary.

### **3.2.3 Corpus design as an aid to the representation of the comment on form**

A formulation of a microstructural programme is one of the major tasks of the lexicographer. This is done as part of the dictionary conceptualisation plan. The lexicographer has to deal with the programme in a way that determines the nature and the extent of the microstructure, the article structure and the manner in which data types will be accommodated in different articles. The corpus will help the lexicographer to achieve his or her goal by facilitating the retrieval of the information that will allow the construction of the dictionary articles as texts in the central list of the dictionary.

As already mentioned, this process includes the determination of the article structure. The comment on form and the comment on semantics play a prominent role in the way data will be presented in the dictionary articles. The observation of some dictionaries today has shown that the comment on form and the comment on semantics are very often presented in dictionaries in such a neglected manner that it confuses the users and does not ensure the user-friendliness of the dictionary.

To improve the presentation of different data types in a dictionary, the lexicographer should make use of effective techniques that exist in this regard. Pronunciation, word divisions, morphological data, labels, syntactic data, and so on are very important data types that a dictionary user is looking for in a dictionary on a regular basis. To better serve the users, the lexicographer has to deal with this data carefully. Using human language technology (HLT) in this regard will help to solve problems in the dictionary articles (see page 70, 71 and 72 for detail).

The way in which corpora are compiled can influence the retrieval of the information needed for a specific task in a dictionary. If a corpus is compiled using archaic techniques, in other words not using adequate and reliable devices, manual counting of data, and so on, the result of the whole process will be biased. HLT is a solution to help achieve good results. HLT consists of a wide field of research. This study focuses on a short list of domains of this field by showing how they can assist the lexicographer in the selection of data to be included in dictionary articles. Here corpora are the basis of the whole analysis.

Roux and Bosch (2002:26) mention that “HLT’s are enabling technologies which are implemented in systems which allow humans to interact with computer systems in

different modes (through text or speech) by using natural everyday language”. The Sydney University Language Technology Research Laboratory comments as follows<sup>4</sup>:

... most of what is stored on computers, just like most of what lives on the web, is information in the form of various human languages, regardless of whether it is stored as text, images, sound files, hand movements, or multimedia presentations.

HLT holds much promise for the developing world, especially for user communities that have a low literacy rate, speak a minority language or reside in areas where access to conventional information infrastructure is limited. HLT mostly uses oral language in the development of specific types of interactive system. African languages that are more oriented towards oral communication are in a privileged position that could allow the compilation of electronic dictionaries based on such source material.

The relationship between HLT and corpora resides in various aspects. One of them is the fact that corpora enable the construction of databases, as already mentioned, and as such they are part of language technologies as well.

Moreover, corpora through the material they provide could represent the basic form of language resources and are therefore an indispensable foundation for language technology research for each natural language. For instance, the development of corpora is by itself a work of HLT due to the extent of technology involvement in the process of corpus building. In fact, corpus building requires not only appropriate software but also storage capacity and functionality as well as product format (electronic or paper format, text or speech, etc.) and usage proprieties (interactive, visual, etc.).

In addition, Roux and Bosch (2002) mention that dictionaries, especially those in electronic format, play an essential role in the development of HLT. According to the understanding of Roux and Bosch (2002), the development of HLT implies the construction of important corpora in view of creating interactive systems such as the following:

- multilingual telephone-based information systems

---

<sup>4</sup> [www.sultry.arts.usyd.edu.au](http://www.sultry.arts.usyd.edu.au)



- multilingual multimedia information systems
- multilingual automatic/machine-aided translation systems

Among other domains of HLT, corpora are useful in various aspects of speech technology. Speech technology (also known as speech processing) includes several subfields (cf. Cole, Mariani, Uszkoreit & Battista Varile 1998; Laver 2006):

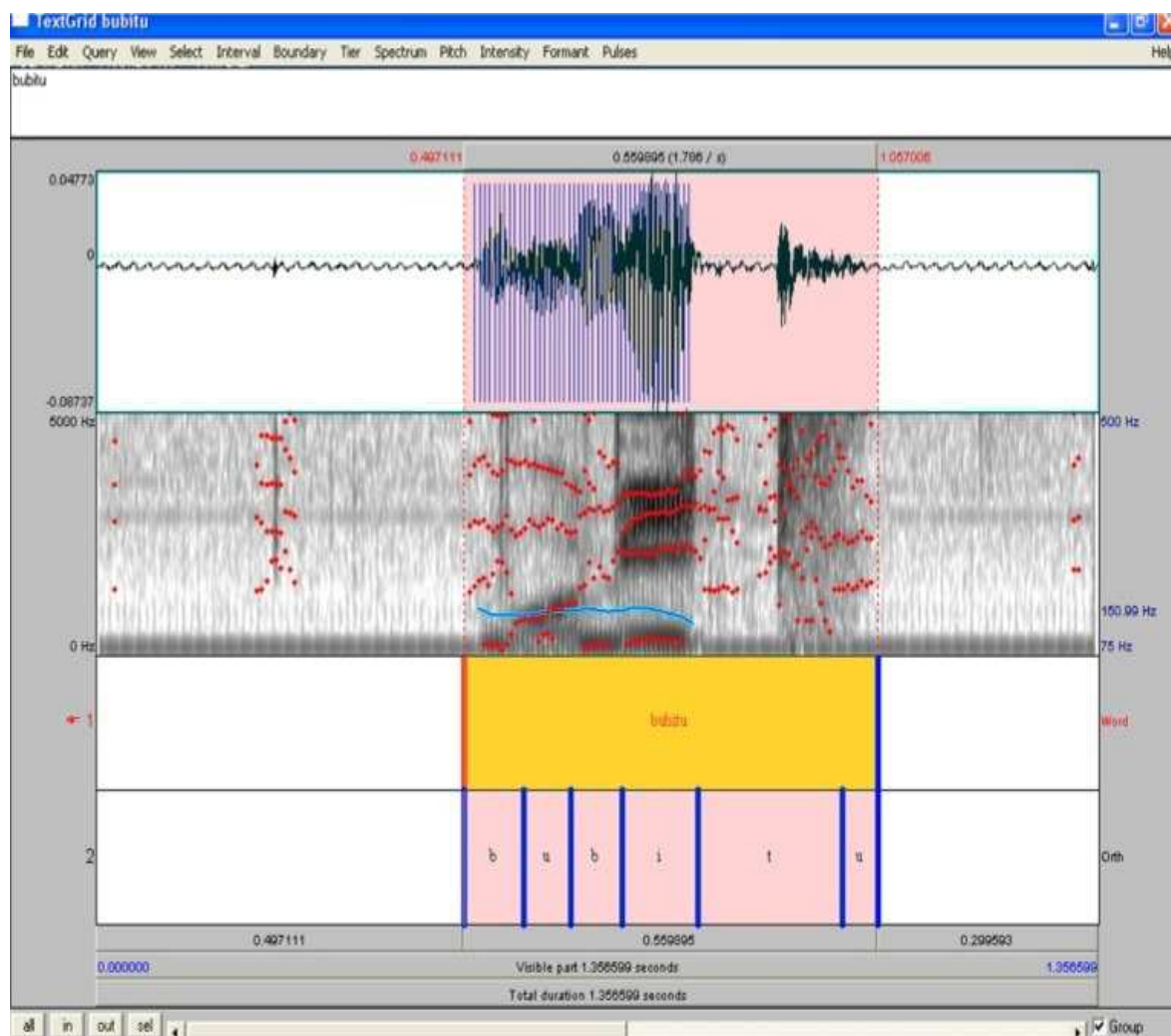
- Speech synthesis or the artificial synthesis of speech, which usually means computer-generated speech.
- Speech recognition, which deals with analysis of the linguistic content of a speech signal.
- Speaker recognition (or voice recognition), where the aim is to recognise the identity of the speaker.
- Speech compression (speech encoding and time-compressed speech).
- Multimodal interaction, which provides the user with multiple modes of interfacing with a system beyond the traditional keyboard and mouse input/output.

It is believed that the data and analyses resulting from this dissertation would benefit any research in speech technology should it be undertaken in the Gabonese languages, particularly in Yipunu.

In the specific domain of speech synthesis, a number of manipulations are carried out on speech data within the field of phonetic analysis. In this regard, data from speech corpora are useful for speech segmentation, as presented in Figure 3.1 below, which presents the phonetic segmentation of the Yipunu word *bubitu* (gums). The word was recorded using PRAAT (version 4.3.12)<sup>5</sup> as sound file for corpus-building purposes, an oral corpus. Such a corpus can be used for various linguistic and phonetic studies. Using these techniques, the lexicographer is in a position of providing the users with much of the accurate information they are looking for in a dictionary. The segmentation of words, the phonetic transcriptions, will be presented in such a way that a distinction is clearly made between syllables and sounds of different vowels or consonants. This is carefully done in the word presented below.

---

<sup>5</sup> PRAAT is a software for doing phonetics by computer ([www.praat.org](http://www.praat.org)).

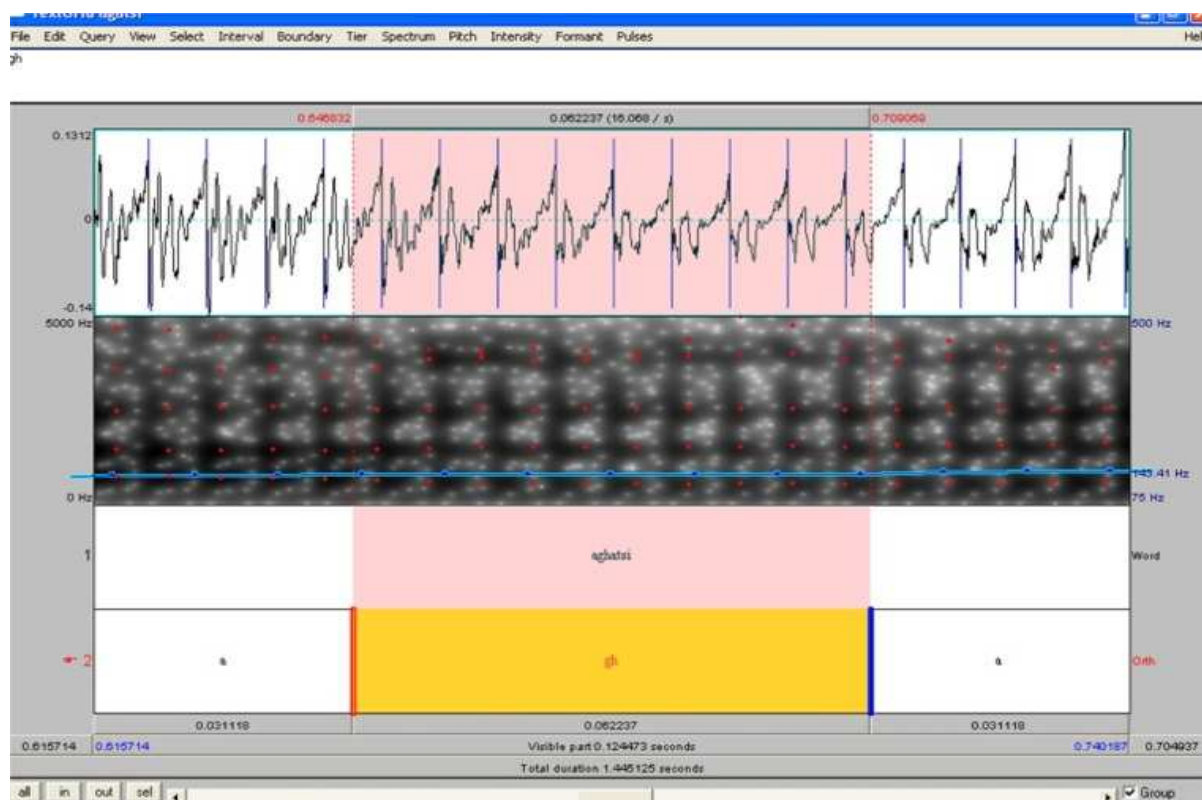


**Figure 3.1:** Phonetic word segmentation in *bubitu* (gums)

Figure 3.1 is a screenshot from PRAAT within a phonetic study. The screenshot shows the following levels of a phonetic study of the word *bubitu*:

- a waveform for a spectral analysis
- a spectrogram for formants and acoustic analyses
- a word tier (**Word**)
- an orthographic transcription tier (**Orth**) for word segmentation

Figure 3.1 also shows the word-time duration in milliseconds. A specific selection of the screenshot can also show the duration of each selected segment of the word. This is shown in the spectrogram presented in Figure 3.2 below.

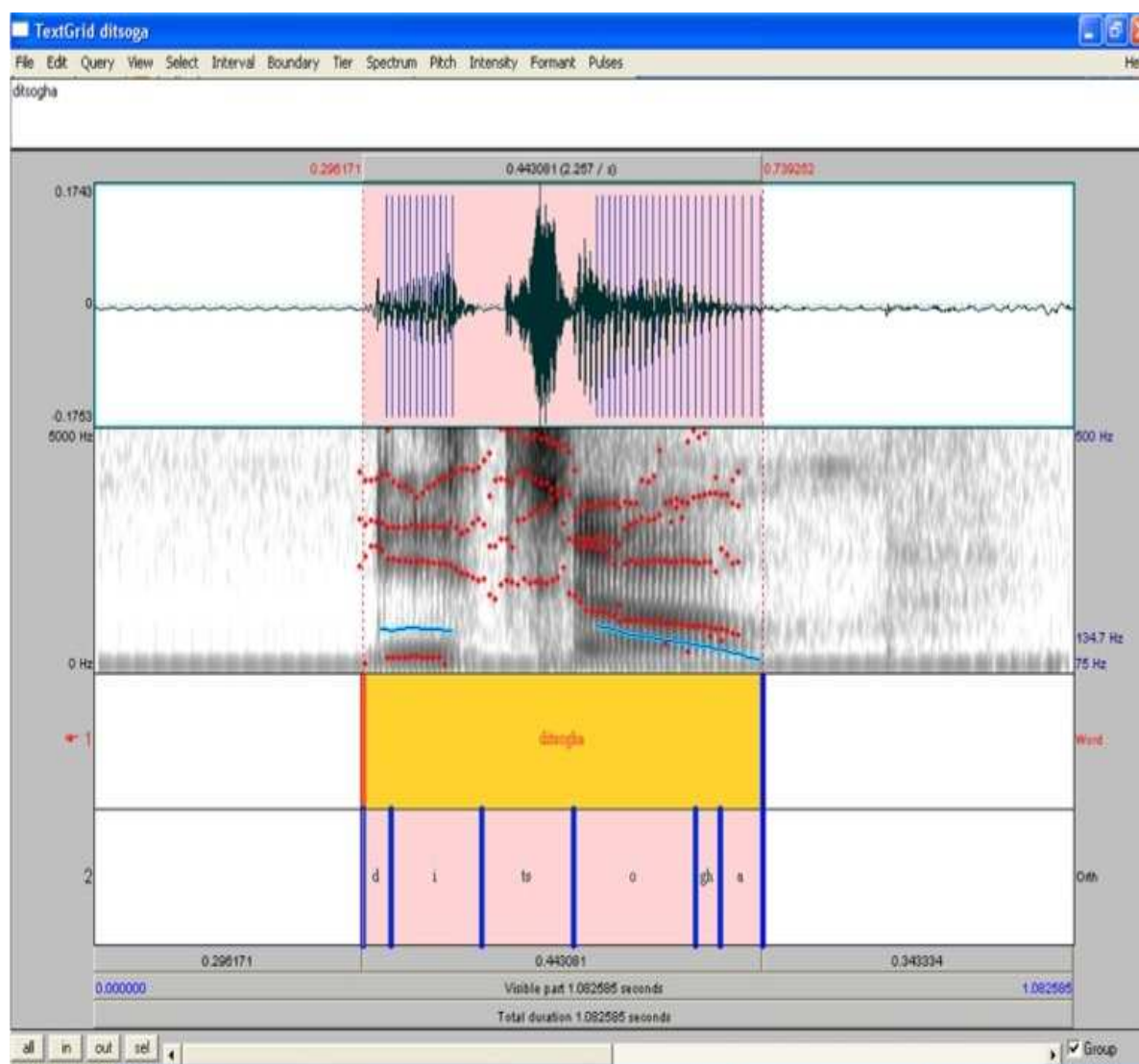


**Figure 3.2:** Segment identification in *aghatsi* (he/she is not in)

The waveform and formants in the spectrogram of the part ‘*agha*’ from *aghatsi* show a clear distinction between vowels before and after ‘*gh*’. Both vowel sounds are recognised as the segment [a] and consonantal characteristics are shown for the segment ‘*gh*’, which is phonetically transcribed [ɣ].

Furthermore, the oral corpus can help the researcher in the phonetic transcription of monolingual and bilingual dictionaries as well as electronic dictionaries by creating text files in digital formats.

In the development of languages, both written and spoken materials play a prominent role that one should really consider. With recorded material, for instance, one can create spectrograms to identify vowels and consonants, which will help the researcher in the pronunciation of the word as a whole. This is shown in the spectrogram presented in Figure 3.3 below.



**Figure 3.3:** Spectrographic segment identification in *ditsogha* (suffering)

A detailed phonetic study of the Yipunu sound system from speech data gathered in the Yipunu database suggested in this dissertation could include a number of other phenomena such as the following:

- vowel reduction at a word-final position
- vowel duration (vowel length and lengthening)
- vowel quality
- segment comparison

Spectrographic segment identification, segment identification and phonetic word segmentation are different techniques that can help the lexicographer to enhance the

quality of the comment on form presented in dictionaries. Thus, phonetic and orthographic transcriptions, word segmentation and even stress indication are different data types that can better be represented in dictionaries using oral corpora.

### **3.3 The main corpus**

Much has been said in Chapter 2 about corpus typology and the intention here is not to return to that. It is important to note that primary and secondary sources form part of the main source. For Renouf (1987), the first step towards achieving this aim is to define the whole of which the corpus is to be a sample. The simplest case will be when the whole is known and finite, such as Shakespeare's works or the Bible. In many cases, the whole is not so easily identifiable and some means of identification must be devised. One possible strategy is to take a library as a microcosm of the written language, or one can rely on the comprehensiveness of established bibliographical sources, as did the creators of the LOB Corpus (Hofland & Johansson 1982).

In this study, the Bible written in Yipunu has been effectively used to serve as the main written source since the language does not have comprehensive written sources. One has to start somewhere.

Svensén (1993:54) indicates that primary sources can be used as main sources if a general language dictionary is based solely on such material. This material would, in order to give the user a reasonably comprehensive result, need to be very extensive and would thus be difficult to manage. This is, in fact, the principal reason why such dictionaries have so far been fairly rare. However, modern computer technology has opened up possibilities in this area. Venturing into computerised material collection is costly and cannot be recommended to publishers of a single dictionary. It is possible only for an institution with considerable resources and long-term plans for compiling a large number of dictionaries of different types. This is the reason why the establishment of a lexicographic unit that will take into consideration the Gabonese government's projects is of vital importance. It will involve a team of lexicographers and other human resources for the gathering of material and the compilation of dictionaries.

In the case of secondary sources being used as main sources, the most common method is to choose a general dictionary by making the selection used for that dictionary the basis of the selection for the new one, Svensén (1993:57) suggests. It is then necessary to make sure that the dictionary has been compiled by competent and conscientious lexicographers.

Svensén (1993) adds that when a bilingual dictionary has to be compiled the use of secondary sources as main sources is risky. If one supposes that a French-English dictionary is to be used as main source for a French-Swedish dictionary, it is necessary to first make sure that the original dictionary really is intended as a passive dictionary for English speakers and not as an active dictionary for French speakers. In the reverse procedure of making an English-French dictionary the main source for a Swedish-French dictionary by transferring the English material into Swedish, it is necessary to first make sure that the original really is an active dictionary for English speakers and not a passive dictionary for French speakers.

For the Gabonese languages in general and Yipunu in particular, the main corpus will be based on primary sources, mostly huge amounts of oral sources and the available written sources from which potential lemma candidates will be included in the macrostructure. This will be done in comparison with the macrostructure of some existing dictionaries in the specific Gabonese language.

### **3.4 Corpus representativeness and balance**

Representativeness and balance are recurrent concepts in corpus lexicography and are very important principles in the compilation of corpora. They are extremely difficult to define and yet easy to work with. These two concepts overlap and the border between them is difficult to draw.

Svensén (1993:41) thinks that it is not enough to be certain that every word or expression in the dictionary actually occurs in the language or the form of language that the dictionary aims to cover. It is also necessary to make sure that every word or expression in a dictionary occurs often enough in the language, in other words that it is sufficiently representative of the language. Representativeness thus means that the word or expression occurs with a certain frequency in the general use of the language.

Svensén (1993) adds that this condition of representativeness primarily concerns active dictionaries. For passive dictionaries, other rules apply: The requirement of representativeness here means that a word or expression must occur frequently enough in the texts that the language users may come into contact with, regardless of whether or not they might actually consider using it. In addition, what is meant by representativeness must naturally vary from one dictionary to another. The limits are imposed ultimately by considerations of quantity, in other words the extent of the planned dictionary. One has to be certain that all the selected words and expressions are authentic and representative. It is also necessary to make sure that the dictionary covers the largest possible range within the area of language to be described.

This exposition is in line with what Landau (2001:331) says: that one should not compromise on the essential standard of representativeness of the corpus as a whole. A corpus is only good if it can be reasonably trusted to represent the way language is used by those people whose usage is interested in describing.

Regarding answering the questions, what are the ways in which representativeness can be achieved? and What or who is the corpus supposed to be representative of? Landau (2001:331) says that one way is by paying attention to text categories and genres and another is size and number of samples. Still other considerations relate to the time period covered and to geographic distribution. He adds that to be truly representative of all language produced, a corpus would have to include a very high proportion of spoken language, a reasonable amount of newspaper text and a tiny percentage of everything else; to be truly representative of all language consumed, the corpus would be dominated by popular newspapers and lowbrow fiction. Landau (2001:333) addresses this point thoughtfully:

Corpus compilers aim to produce ‘well-balanced’ collections of texts which collectively represent the full repertoire of spoken and written performance across the broadest possible range of contexts and genres. The reason for this somewhat odd definition of representativeness is tied to the aim of the dictionary maker to find examples of as many different usages as possible.

Ensuring the representativeness and the balance of the corpus in the Gabonese languages is a very demanding enterprise taking into consideration the subject field in which some of the existing material is published. To achieve the requirement of

representativeness and balance, an additional and well-targeted collection of material containing sources from diverse sectors is primordial. There is no accurate measurement as to the specific value of balance and representativeness in a corpus. In this regard, Gouws and Prinsloo (2005:24) affirm that the reality for most African languages is such that a neatly designed collection strategy is not possible and that the whole selection process eventually boils down to the collection of all *available* texts for the specific language. In many instances available texts have to be heavily supplemented by subcorpora compiled from oral data collections in order to reach a corpus size of a few million running words.

The gist of what has been said so far is that the lexicographer cannot build a perfect corpus. The representativeness and the balance of the corpus can only be reached while the corpus is growing, by enlarging it, taking into account the new areas of the language. Even in the European languages, which have a long written history, some corpus compilers took time to reach this point and some are still in the process.

Thus, there is nothing to lose by trying to compile the corpus in the way everyone is doing; otherwise, if things are not moving as planned, it will be advisable to follow Gouws and Prinsloo's suggestion which is in favour of the concept of *organic corpora*, which fits the African languages like a glove, quoting Atkins, Rundell and Weiner (1997):

A corpus compiler should first attempt to create a representative corpus. Then this corpus should be used and analysed and its strengths and weaknesses identified and reported. In the light of experience and feedback the corpus is enhanced by the addition or deletion of material and the circle repeated continually. This is the way to approach a balanced corpus.

Furthermore, Atkins *et al.* (1997) argue that

one should not try to make a comprehensive and watertight listing [...] rather, a corpus may be thought of as organic, and must be allowed to grow and live if it is to reflect a growing living language [...] In our ten years' experience of analysing corpus material for lexicographic purposes, we have found any corpus – however unbalanced – to be a source of information and indeed inspiration. Knowing that your corpus is unbalanced is what counts.



### 3.5 Use of tags

It is important to distinguish between tagged corpora and non-tagged corpora. Contrary to non-tagged corpora, tagged corpora are corpora in which different parts of speech can immediately be identified by an indicator. This is done manually, by tagging word by word in a corpus, or automatically, by using a special computer program, with the help of the lexicographer. Inserting labels into text is called text tagging and this should be done in accordance with the functions and the purpose of the corpus. Non-tagged corpora still exist, built just by gathering texts and saving them in their original formats. They are not of less use at all. They can allow very important manipulations. Having a great amount of raw text available allows one, for instance, to study word and phrase frequencies and collocations and even to build concordance lines in the language, and dictionaries can still be compiled from this kind of corpus.

Concordances may be obtained through an automatic search across the whole corpus for all the occurrences of a given word. If it is a synchronic corpus, concordances are of obvious interest to the lexicographer. However, even in a diachronic corpus, which potentially presents a history of the studied language, (for instance from the 17th century up to the present), a set of *dated concordances* may well allow one to study the history of ideas and trace the appearing and rise of a concept or of a new meaning. A common example of this is a study using the ARTFL corpus, which unveiled the interesting fact that the use of the word *revolution* in its meaning of ‘political upheaval’ increased significantly in French texts during the years preceding the French Revolution (Gautier 1998).

Right now the corpus at issue in this study is neither tagged nor annotated and the querying of that corpus is yielding good results. The researcher has managed to generate word and phrase frequencies and collocations in the SYC as well as concordances and the lemmatisation of inflexions and lexical items. This will be further discussed in chapters 4 and 5.

In the case of a multilingual corpus, it has already been mentioned that a common practice was to use tags to align different versions of the same text in different languages. Aligning may be accomplished through the use of identical numbers for the corresponding sections (sentences and paragraphs). Even for a monolingual

corpus, this structural tagging is also of use to mark the divisions of a text (pages, chapters, stanzas, etc.).

However, the first expected information in a text corpus is probably bibliographical: author, date of publication, publisher, type of text, commentary, and so on. There is a trend to some standardisation in this field (*Text Encoding Initiative*, 1996) (*URL: Corpus Encoding Standard*, 1997).

Linguistic tagging may also be applied to the words of the text by part of speech, in other words type of word or lexical item. It is then referred to as *part-of-speech* or *grammatical tagging* (French *étiquetage grammatical* or *assignation grammaticale* [DE LOUPY, 1995]). A *semantic tagging* of the type long used in margins by anthropologists, which means tagging by subject parts of the text, may also be used. In the same way Kennedy (1998:21) defines part-of-speech tagging as assigning a word class to all the words in a corpus. Among others, part-of-speech tagging provides crucial data for lemmatisation, parsing or advanced concordancing.

As far as lemmatisation is concerned, Hartmann and James (1998:83) think that lemmatisation should be understood as the reduction of a paradigm of variant word forms to a canonical form; hence a lemmatiser (also called morphological analyser) merges a certain paradigm of variants into a single canonical form. This is very important for lexicographic purposes.

Structural analysis of sentences is known as syntactic parsing. According to Kennedy (1998:21) corpora can also be parsed to show the sentence structure and the function in the sentences of the different word classes.

African languages that have a conjunctive orthography and few word delimiters can be helped by word tokenisation, which is segmenting a text containing conjunctively written words into free-standing words.

Landau (2001:334) believes in corpus tagging because all modern lexicographical corpora are annotated with tags (specific sequences of codes, often within angle brackets) to mark certain features of the text. The process of text tagging (called markup) has been standardised so that everybody will designate the same part of a text, such as the end of a paragraph, with the same code. This system of markup, the Standardised Generalised Markup Language (SGML), has been further elaborated for the description of whole texts in the Text Encoding Initiative (TEI).

### 3.6 The role of a lexicographic unit in the process

The establishment of a lexicographic unit in Gabon will play an important role in the process of dictionary compilation as it has already shown remarkable results in many countries elsewhere. South Africa is one of the best examples where the implementation of several lexicographic units in a multilingual society has succeeded tremendously. With 11 official languages (Afrikaans, English and nine South African Bantu languages), the country has managed to give all languages the opportunity to develop further. Community members, well trained in both metalexigraphy and practical lexicography, demonstrate this through the compilation of dictionaries.

In South Africa each of the 11 national languages has its own lexicographic unit. Listed below are the official national lexicography units (NLUs)<sup>6</sup> as they are recognised by the Minister of Arts, Culture, Science and Technology:

- The Bureau of the Woordboek van die Afrikaanse Taal in Stellenbosch.
- The Dictionary Unit of South African English in Grahamstown.
- Sesicha zamazwi SesiZulu National Lexicography Unit in Isipingo.
- isiXhosa National Lexicography Unit in Fort Hare.
- Sesotho sa Leboa National Lexicography Unit in Polokwane.
- Setswana National Lexicography Unit in Mafikeng.
- isiNdebele National Lexicography Unit in Pretoria.
- Sesotho National Lexicography Unit in Bloemfontein.
- Silulu Se Siswati National Lexicography Unit in Nelspruit.
- Tshivenda National Lexicography Unit in Thohoyandou.
- Ngula ya Xitsonga National Lexicography Unit in Tzaneen.

Some of these NLUs are well established and others are still at the initial stage. These units are hosted in tertiary institutions around the country in accordance with the government policy on language research. Many African countries in need of such

---

<sup>6</sup> Some lexicographic units in South Africa have changed their name. Unfortunately by the time we completed this work, we did not have any information about the new names.

units could learn from this example but the problem resides in the specificity of each country. The lexicographic unit, when empowered with all the necessary means, plays an important role in the development of the language in the community for which the project is established. In accordance with the needs of the community, the lexicographic unit has the mission to do the following:

- Participate in the planning of the compilation of a national corpus of the language and project the production of comprehensive explanatory dictionaries in that language as a long-term goal.
- Provide a research facility for the study of the language in general and lexicology and lexicography in particular.
- Be responsible for translation services and compilation of dictionaries for immediate use as a short-term goal, which can help in solving terminological issues and can be used as reference sources for researchers and language users.
- Review all existing dictionaries dating from the colonial era.
- Monitor, report on and advise on the viability of the language as a medium of communication at all levels in society, both within its mother-tongue community and beyond.
- Play a significant role in the marketing and sales of both larger and smaller dictionaries in a cost-effective manner.

For the Gabonese languages, the South African situation could serve as example and be adapted to fit the particular Gabonese context. In Gabon, French is not only an official language used in administration and business but also a first language for some Gabonese people when the local languages are only recognised by the State in the Constitution as to be promoted. Kwenzi Mikala (1987; 1998) in his investigation identified around 62 different speech forms (including languages and dialects) that he grouped into 10 language units. Other researchers such as Guthrie (1953), Jacquot (1978) and Raponda-Walker (1960) classified these languages. The intension here is not to go through all the classifications but to focus on the oldest and the latest. It is going to be difficult to adopt all these languages as official languages.

### **3.6.1 Guthrie's classification**

The classification of Guthrie (1953) presented here is focused on the languages that belong to zones A, B and H.

#### **Zone A**

A.31 Bubi (Benga)

A.75 Fang

#### **Zone B**

##### **B.10 Myene Cluster**

B.11a Mpongwe

B.11b Orungu

B.11c Galwa

B.11d Adjumba

B.11e Nkomi

##### **B.20 Kele Group**

B.21 Sekyani

B.22 Kele cluster

- B.22a Kele
- B.22b Ngom
- B.22c Bubi

B.23 Mbangwe

B.24 Wumbvu

B.25 Kota (Shake, Mahongwe)

##### **B.30 Tsogo Group**

B.31 Tsogo (Mitsogo, Apindji)

B.32 Kande (Okande)

**B.40 Shira-Punu Group**

B.41 Sira (Shira)

B.42 Sangu

B.43 Punu

B.44 Lumbu

**B.50 Njabi Group**

B.51 Duma (Adouma)

B.52 Nzebi

B.53 Tsaangi

**B.60 Mbede Group**

B.61 Mbeté

B.62 Mbaama (Mbamba)

B.63 Nduumo (Mindumbu)

**B.70 Teke Group**

B.71 N Teke

B.72 N.E. Teke

B.73 W. Teke

B.74 Central Teke

B.75 Bali (Teke)

B.76 E. Teke

B.77 S. Teke

B.78 Wuumu

## **Zone H**

H.12 Vili

### **3.6.2 Kwenzi's classification**

The 62 languages identified by Kwenzi (1998) were classified as follows in 10 *unités-langues* (language units), according to the interunderstanding/intercomprehension criterion:

#### 1. L'unité-langue mazona

- Atsi
- Meke
- Mven
- Okak
- Ntumu
- Nzaman

#### 2. L'unité-langue myene

- Enenga
- Galwa
- Mpongwe
- Nkomi
- Orungu
- Okoa

### 3. L'unité-langue mekena-mena

- Akele
- Ungom
- Lisigu
- Mbangwe
- Metombolo
- Seki
- Tumbidi
- Shake
- Wumpfu
- Lendambomo

### 4. L'unité-langue mekona-mangote

- Ikota
- Benga
- Shamayi
- Mahongwe
- Ndas
- Bakola

### 5. L'unité-langue membe

- Getsogo
- Gepinzi
- Kande
- Gevhovhe
- Gehimbaka



- Geviya
- Ebongwe
- Kota kota

#### 6. L'unité-langue merye

- Gisira
- Givharama
- Givungu
- Yipunu
- Yilumbu
- Ngubi
- Civili
- Yirimba
- Yigama

#### 7. L'unité-langue metye

- Inzebi
- Itsengi
- Imwele
- Iveli
- Liduma
- Liwanzi
- Ibongo

#### 8. L'unité-langue membre

- lembaama
- lekaningi

- lindumu
- latege
- latsitsege

#### 9. L'unité-langue makena

- Bekwil
- Shiwa (makina)
- Yesa

#### 10. L'unité-langue baka

- Baka

These two classifications can be combined and adjusted in such a way that 10 main languages can be chosen to be adopted as national languages. The rest of the languages, fall within the scope of dialects or variants of the major ones. Nothing can be done, however, if the government does not take the decision in that direction. Researchers can play an important role in this regard by propositioning ideas that can help solving such linguistic problems in Gabon.

In Gabon the major research projects are conducted by the CENAREST to which the government provides finances for research projects of any kind including research pertaining to language. Within this research group, a lexicographic unit can be established. The proposal in this regard is to have one lexicographic unit that will deal with all the Gabonese languages since the number of researchers in the field of lexicography is still limited. That unit will plan and schedule the compilation of dictionaries in these languages according to a well-designed plan of action that will suit the government policy. It will be difficult to have one lexicographic unit for one language in one university because the situation in Gabon is such that almost all the universities belong to the state and they are few. It would be beneficial as a long term objective if there are enough staff members to establish a lexicographic unit for each language. For the effective management of the Gabonese languages, it is necessary to limit the number of lexicographic units per university instead of concentrating these

lexicographic units in the same university. This will allow all the languages to be established and developed.

The lexicographic unit will be in charge of research in all the languages, provided that the national language policy be well defined. With the plan of the government in mind, different dictionaries can then be compiled after huge amounts of written and spoken materials have been gathered in each language. Decisions of a lexicographic nature can be taken in accordance with the policy of the research unit.

The classifications presented above clearly show different dialects in each language group. As already mentioned, by taking into account all the dialects of a given language during the investigation of data, the lexicographer will reach optimal results covering the vocabulary of that language in all the domains. The introduction of all the dialects in the corpus will potentially provide the lexicographer with different variants and polysemous senses of words. The lexicographic unit will have the role of guiding the lexicographer in the identification of the needs of the users of the dictionaries to be compiled for each language. This will be achieved through cooperation between government language policies, the lexicographers' expertise and the desire of the lexicographic unit to see a real development of all the languages.

In the study of geographical variation, corpora have long been recognised as a valuable source of comparison between language varieties as well as for the description of those varieties themselves. In the case of Gabon, the use of corpora in the determination of dialects of the same language has to follow as far as possible the same sampling procedures as other corpora in order to take full advantage of the degree of comparability, as suggested by McEnery and Wilson (1993). This will enable the identification of illustrative examples of the different word variants, synonyms or translation equivalents of the different dialects of the language.

All these elements will help in the determination of the norm of a language and the direction in which the lexicographer should go in order to achieve her or his goals. Dialectal data are very important in dictionaries, depending on the type of dictionary. This data can demonstrate the source and the usage of a given word or expression in a dictionary. One will know the region that a lexical item was taken from.

### **3.7 Language status and corpus building**

It is important for the lexicographer when building corpora in multilingual and multicultural societies to be aware of the status of the language environment. Worldwide, many languages are composed of different dialects and idiolects in which several language categories (informal and formal languages, for instance) can be found. When designing a corpus for those languages, the lexicographer has to take into consideration the existence of all varieties in order to accomplish her or his task properly. One can start by dealing with the main variety, or the most spoken dialect, as it can serve as a standard form and later complete the research into the language material by extending it to other speech forms. This can easily be done in communities where languages have a well-defined status in the sense that the language policy makes a clear distinction between official languages and languages that have a regional communicative function.

Once the identification of the different language levels is done, the lexicographer can proceed with the comparison of the different lexical items drawn from corpora of those dialects. The comparison of the macrostructural data will enable the lexicographer to amalgamate the lemma lists of the different dialects in such a way that they will help to distinguish the common data from those that are different from one dialect to another. The lexicographer will have to work with the assumption that all the dialects belonging to a certain group have words in common, and the corpus will assist in capturing the degree of resemblance of the lexical items. The next phase will allow the lexicographer to construct the microstructural programme that will help in the compilation of dictionaries in a particular language.

The situation described here is applicable to the Gabonese and all African languages sharing the same linguistic environment. This could help to solve problems such as language conflicts in those societies where a specific variety is seen as superior by the speech community using that dialect. This sometimes results in a situation where people refuse to learn any word from the so-called better variety because of the fact that their languages have been regarded as inferior. In this regard, the lexicographer should not take part in the conflict but rather work towards the accommodation of everyone's ideas in order to fill the needs of the entire community.

Many Gabonese languages such as Fang, Omyene, Yipunu, Yinzebi, Teke, Yilumbu and Civili are spoken in different areas, both locally and in neighbouring countries. Both Fang and Omyene have several dialects while the remaining is known as representing single varieties. Some of these languages have spread to neighbouring countries. Medjo-Mvé (1997a; 1997b; 1997c) conducted studies about Fang dialectology by showing the geographical environment where it is spoken, namely Cameroon, Equatorial Guinea and Gabon. The concluding analysis confirmed that all varieties are dialects of the same language.

Yipunu, Yinzebi, Teke, Yilumbu and Civili are spoken in Gabon and the Republic Congo. The use of corpora in the comparison and analysis of all these languages will constitute a very important tool for the formulation of hypotheses that can explain the historical relations between different varieties within the same country or area.

In the study of geographical variation, corpora have long been recognised as a valuable source of comparison between language varieties as well as for the description of those varieties. In the case of Gabon, the use of corpora in the determination of dialects of the same language has to follow as far as possible the same sampling procedures as other corpora in order to take full advantage of the degree of comparability, as suggested by McEnery and Wilson (1993).

Thorough research on different dialectal forms has to be undertaken, for there is a tendency among speakers to reject certain words that they regard as borrowed. This impedes the growth of a language. Speakers should stop looking upon the so-called standard dialect as the only correct dialect. All languages are, at one stage or another influenced by languages with which they come into contact. Dialectal forms contribute to the growth of a language.

These speech differences that arise within a language may be due to various factors; for example, one quite often finds that the speech of a group is influenced by the language of an adjoining area and that certain sounds or words of that language begin to creep in over a long period of time. These innovations then become established within that group and become part of their own language. When compilers take other dialects into consideration, monolingual dictionaries that define concepts pertaining to dialects in different areas will be made possible.

### 3.8 On alphabets

Many scientific discussions have been held in Gabon with regard to the standardisation of an orthographical writing system. None of them have been adopted until now since the government is still hesitating with regard to the introduction of the Gabonese languages in the educational system. All the writing systems proposed up to now are based on the IPA and the African Languages Alphabet (ALA). The presentation of these alphabets is given for a better understanding of the orthographic problem of the Gabonese languages and to justify the choice of the one used in the present work. The proposals of alphabets of Carpentier de Changy and Voltz (1990), Hombert (1990) Idiata (2002), Mayer (1989) and Ndinga-Koumba-Binza (forthcoming) presented in the tables below are given here as a summary of the different alphabets.

a) Alphabet of the Gabonese languages according to André Raponda-Walker (1932)

**Table 3.1: Alphabet of the Gabonese languages by André Raponda-Walker**

Vowels		Simple consonants			Consonants with diacritics
a	[a]	b	l	v	ê [ʔ]
e	[e]	c [k]	m	w	ɛ [tʃ]
è	[ɛ]	d	n	x	ġ [ʒ] or [ʝ]
i	[i]	f	p	y	ñ [ɲ]
o	[o]	g	q [k]	z	n' [ŋ]
ö	[ɔ]	h	r		r' [ʀ]
u	[u]	j [dʒ]	s		ÿ [ʏ]
ü	[y]	k	t		

## b) The Scientific Alphabet for the Gabonese Languages (SAG)

**Table 3.2:** Scientific Alphabet for the Gabonese Languages

Vowels		Consonants				Tons		
<i>Minuscules</i>	<i>Majuscules</i>	<i>Mini</i>	<i>Maj</i>	<i>Mini</i>	<i>Maj</i>	<i>Signs</i>	<i>et denominations</i>	
a	A	b	B	n	N	∨̇	infra-low	
ə	Ě	c	C	ŋ	Ŋ	∨̇	low	
e	E	d	D	p	P	∨̄,	mid	
ɛ	ɛ	ð	Ð	r	R	∨̇	high	
i	I	f	F	ɸ	Ɔ	∨̇		
ɔ	ɔ	g	G	s	S		descenda	
o	O	ɣ	Ƴ	ʃ	ʃ		nt	
u	U	h	H	t	T	∨̇	montant	
u̇	U̇	ʒ	Ʒ	v	V	∨̇	extra	
		j	J	β	β		high (TH)	
		k	K	w	W	∨̇	haut	
		ʔ	ʔ	ẉ	Ẉ		abaissé	
		l	L	x	X		∨̇ TH	
		m	M	y	Y		descendant	
				z	Z			
<b>Diacritics</b>						<b>Digraphs</b>		
<i>Signes et Utilisations</i>						ny	mb	ts
ɥ	tilde for nasals					ty	nd	dz
ɥ̣	extra-short vowel					dy	ŋg	aa
ɥ̣̣	palatalisation					kw	kp	oo
ɥ̣̣̣	centralisation					tw	gb	ee

c) Alphabet of the Gabonese languages session of April 1999

**Table 3.3:** Alphabet of the Gabonese languages

vowels		Les tons		consonants		
<i>Signs</i>	/	<i>Signs</i>	/	<i>Monographs</i>	<i>Digraphs</i>	
<i>realisations</i>		<i>denominations</i>				
a		´	high	b	m	gh
e		`	low	c	n	[ʎ]
<u>e</u>	[ɛ]	-	mid	[tʃ]	<u>n</u> [ŋ]	jh
ə		^	descendant	d	p	[ʒ]
i		˘	montant	<u>d</u> [ð]	r	sh
o				f	s	[ʃ]
<u>o</u>	[ɔ]			g	t	vh
u				h	v	[β]
<u>u</u>	[y]			j	w	ny
				k	y[j]	[ɲ]
				l	z	

To these alphabets can be added the one in use by Rapidolangue, published by the Raponda-Walker Foundation and inspired by the alphabet proposed by Raponda-Walker. This alphabet presents the following characters:

**Vowels**

- a
- e
- e
- ë
- i
- o
- o



u

u

### Consonants

b	f	k	mp	ny/gn	t
d	g	l	n	p	v
dy	h	m	nd	r	w
dj	j	mb	ng	s	z

In this last presentation one can note that special characters and diacritics are not taken into consideration. This is certainly to avoid complications if one has to type texts into a computer and it is also time-saving. This alphabet is more or less what has been used in the corpus of this study except for the fact that in this corpus there is no underlining of any character. To these characters some specific Yipunu phonemes are added. The researcher's take on the choice of orthography for any language is to stay as simple as possible. The African languages have many difficult diacritics and the more one wants to reproduce all of them, the more problems one is going to deal with on a daily basis. For instance, the current study does not take into account long vowels in the orthographic system; they will be represented in phonetic transcriptions when necessary.

All these alphabetic characters are important indicators in the retrieval of words in the corpus. For instance, one has to look under a specific letter to find the word one is looking for in a frequency list or lemma candidate list. The characters will help in the identification of word structures that are very important for the segmentation process. The structure of a complete word in Yipunu is prefix + stem/root + final (vowel).

### 3.8.1 Graphemes of Yipunu

Table 3.4 and Table 3.5 give the different graphemes of Yipunu, partly in combination with the proposal of Kwenzi-Mikala (1980) and Rittaud-Hutinet (1980), as summarised here. All these graphemes are more or less based on both the IPA and the ALA.

#### 3.8.1.1. Vowels

Yipunu has five short vowels, i, e, a, o and u, that have their corresponding long vowels, as shown here. Long vowels are doubled in the presentation below and in the final orthography; the long vowels are not taken into consideration. However, in a dictionary where the phonetic transcription is needed, the lexicographer should consider their representation.

**Table 3.4:** Yipunu vowels

<b>i</b>	<b>I</b>	[i]	Ditoghu	Mat
<b>ii</b>	<b>Ii</b>	[i:]	Diimbu	Village
<b>e</b>	<b>E</b>	[e]	Temu	Epoque
<b>ee</b>	<b>Ee</b>	[e:]	Nzwengi	Colibri
<b>a</b>	<b>A</b>	[a], [ə]	Taji	Father
<b>aa</b>	<b>Aa</b>	[a:]	Ejabi	He knows
<b>o</b>	<b>O</b>	[o]	Kodu	Nuque
<b>oo</b>	<b>Oo</b>	[o:]	Dilongi	Advice
<b>u</b>	<b>U</b>	[u]	Mutu	Man
<b>uu</b>	<b>Uu</b>	[u:]	Nungi	Plantation

### 3.8.1.2. Consonants

As shown in the following table, Yipunu has 23 consonants, both simple and combined.

**Table 3.5:** Yipunu consonants

<b>p</b>	<b>P</b>	[p]	Putu	Maize
<b>b</b>	<b>B</b>	[b]	Batu	Men
<b>mb</b>	<b>Mb</b>	[mb]	Mbata	Slap
<b>t</b>	<b>T</b>	[t]	Taba	Sheep
<b>d</b>	<b>D</b>	[d]	Diisu	Eye
<b>nd</b>	<b>Nd</b>	[nd]	Ndasi	Show me
<b>k</b>	<b>K</b>	[k]	Kaki	Thunder
<b>g</b>	<b>G</b>	[ɣ]	Gaba	Share
<b>ng</b>	<b>Ng</b>	[ŋ]	Ngandu	Caiman
<b>f</b>	<b>F</b>	[f]	Ifumba	Family
<b>v</b>	<b>V</b>	[β]	Vatsi	On the floor
<b>mv/mf</b>	<b>Mv</b>	[mv]	Mvumbi	Dead body
<b>s</b>	<b>S</b>	[s]	Disala	Shower
<b>nz</b>	<b>Nz</b>	[nz]	Nzyembu	Blame
<b>ts</b>	<b>Ts</b>	[ts]	Tsana	Seat down
<b>ny</b>	<b>Ny</b>	[ɲ]	Nyambi	God
<b>l</b>	<b>L</b>	[L]	Dilolu	Papaye
<b>r</b>	<b>R</b>	[r]	Burela	Hunting
<b>w</b>	<b>W</b>	[w]	Bwali	Disease
<b>y</b>	<b>Y</b>	[y]	Yotsi	Cold
<b>j</b>	<b>J</b>	[dʒ]	Jotsu	Whole
<b>m</b>	<b>M</b>	[m]	Matuji	Hears
<b>n</b>	<b>N</b>	[n]	Nandi	With him/her

### **3.8.2 The chosen orthography of Yipunu**

For this corpus the following orthography has been chosen to write Yipunu. The motivation is to try to stay as simple as possible, and while compiling this corpus, the researcher has noted some difficulties with regard to the usage of the computer program for the query of the corpus and the typing of some material that was added to the scanned material. Those characters can be finalised in the following manner in this chronological order; thus a Yipunu alphabet of 28 characters is proposed:

A, B, D, E, F, G, I, J, K, L, M, Mb, Mf, N, Nd, Ng, Ny, Nz, O, P, R S, T, Ts U, V, W,  
Y

This will also be the access alphabet of the envisaged dictionaries. The problems regarding standardisation still remain and researchers along with communities and the government have the important task to agree on the final alphabet of the Gabonese languages. No language in the world can be standardised without a clear orthography; researchers of the Gabonese languages will have to create their own way out of this problem and the proposals will increase the number of existing alphabets. The corpus proposed in this dissertation reflects the actual language use and all gathered materials are edited in order to fit the standardising approach.

### **3.8.3 Conjunctive vs. disjunctive**

The opposition between agglutinative, inflectional and isolating languages, as already explained, gives a clear idea in principle about how a specific language should be written. Conjunctive writing applies to those languages whose morphemes are combined with the stem to form a block of letters while disjunctive writing refers to those languages in which morphemes are separated from the stem or root. Any language can be written either conjunctively or disjunctively. It is simply a matter of convention. A given community should then decide according to which of the two systems it would like to write its language, in order to standardise the spelling or writing system.

It is also possible for any language to integrate conjunctive writing in a language principally using disjunctive writing and vice versa. This is the case for Yipunu, which is predominantly a conjunctively written language but also uses a disjunctive method in particular cases. The following is an example of Yipunu text:

Yika bivunda bamasila nana tsi jetu.

Mba me nyisamala ne bivunda bye.

Na bamaburu na bamabanguga.

Bamena bura bamama.

Bamama bamenatura.

Me nymaburulu o Mukaba Disimu.

Me dina dyami Majinu ma Kombila.

Yibandu yami jungu.

Yike dusavu dwedudu bivunda bamadusavanga.

Bibunda byenyi bajaji bami bakaduwaka.<sup>7</sup>

### **English translation**

It is the elders who transmitted what follows not me.

Besides I never knew those elders.

They were born and grown up before us.

And they gave birth to our mothers.

Our mothers gave birth to us.

I was born in Mukaba Disimu.

My name is Madjinou ma Kombila.

My clan group is Djoungou.

---

<sup>7</sup> This text is an extraction of an orthographical transcription made by the researcher of the *Mumwanga a Punu* epic published by Kwenzi in a phonological transcription.

Thus, this story was told by the elders.

Those elders were also taught about.

Once the corpus has been finalised, any user of that corpus can judge by looking at word combinations which of the conjunctive or disjunctive writing systems is dominant. The choice of either is arbitrary and it can happen that after opting for the conjunctive system, the language board decides to go for the disjunctive system because of some reason. Which system is used is very important in determining the number of running words in a corpus. The number of tokens will be more if the language is written disjunctively and less, if written conjunctively.

### **3.9 Standardising an orthography**

The writing system of any language can be influenced by many factors, such as the integration of loan words in the language. When a language borrows a word or expression, it can take it as it is, thus including the word with its original structure and making no changes, like in the case of the word *parking* in French that is borrowed from English. The word has been loaned with its orthographical, phonological and phonetic implications.

Some examples can be taken from Yipunu. For instance, words such as *dila*, in French *lait* (milk), and *kambini*, in French *camion* (truck), are borrowed from French since these words do not have a direct translation equivalent in Yipunu. The language board should consider such phenomena in the language since they are crucial.

For centuries the lexicographer was seen as judge or ruler over the language who has to prescribe correct usage and, to point out unacceptable use, as in the case of Samuel Johnson's *A Dictionary of the English Language*, published in 1755, as recognised by Al-Kasimi (1977), Gouws (1989), Landau (2001), Wells (1973) and many others quoted by Nong *et al.* (2002). In accordance with this vision, the dictionary should rectify and cleanse the language, preserve its purity, lengthen its duration, correct and ban improprieties and absurdities, sensor faulty usage and repress anomaly.

In opposition to the prescriptive approach stands the descriptive approach with the focus on actual language usage, according to Gove (1961:13), quoted by Al-Kasimi

(1977:84), quoted by Nong, De Schryver and Prinsloo (2002) when he wrote a letter to *Life Magazine*:

The responsibility of a dictionary is to record the language, not set its style. For us to attempt to prescribe the language would be like *Life* reporting the news as its editors would prefer it to happen.

This opinion is in line with that of Prinsloo (1992:10), quoted by Nong *et al.* (2002), as he rightfully emphasises that the Northern Sotho Language Board made an invaluable contribution towards the clarification, systematisation, standardisation and coining of new terms for Northern Sotho in, for example, religion, news broadcasting, mathematics and general science. The last Terminology and Orthography for Northern Sotho, T&O for short, produced by this Language Board was published in 1988 (Department Northern Sotho Language Board 1988(4)). The Language Board adopted a sensible approach in being prescriptive in the coinage and approval/disapproval of terminology on the one hand while still placing a high premium on actual usage as criterion for acceptability on the other hand. The Language Board (T&O: 3) allows for more than one option rather than attempting to enforce just one term while suppressing others. There exists a new terminology in a 2005 Edition called: Spelling and Orthography Rules. Sesotho sa Leboa. 2005. National Language Body. Pan South African language Board.

African languages, such as the Gabonese ones, that are still fighting for the standardisation of their language, by taking into consideration all the factors, should adopt this attitude. Zgusta (1971:187) rightly points out that lexicographers can coin new expressions, they can normalise their form and meaning and they can systematise and clarify the old ones; they can help in an endless number of such exceedingly useful and necessary tasks. The real life of a language, however, is in its use, and the definitive, fully fledged stabilisation of the standard national language is brought about by its being really and extensively used in literature and in oral communication of all types.

The lexicographer computerises the corpus in a certain format and orthography; by doing so she or he is implicitly standardising that language. One can say without any doubt that the corpus helps the standardisation of a language as it contains material

written in such a way that it gives the spelling of lexical items that are going to be used as part of the macrostructure and dictionary articles in the central list. This can clearly be observed in languages that lack a writing system that is commonly used by researchers and the speech community. Such a situation leaves the lexicographer without alternative but must come up with something that can give a head start to the compilation of dictionaries and other books.

### **3.10 Fonts**

The choice of orthography must be in line with the font that one uses to save one's files. There are many existing fonts for different languages. There are fonts for European languages and some for African languages. It is therefore easy to obtain computer software that suit English and French. Accommodating fonts that will suit a particular writing system, such as the one for African languages, will be difficult, however. Up to now most African languages have had no standard writing systems and are still commonly using special characters in their orthography, and special fonts need to be installed in computers for the characters to be read.

The researcher believes that Microsoft economically and strategically is not yet ready to consider the required software for African languages when designing computers. It is then to the benefit of nonrecognised languages to adapt their alphabet accordingly. Imagine a Yipunu speaker writing a letter with characters like the ones given in tables 3.1 and 3.2. How would the text look like to someone who does not have those special characters on his or her computer? Surely, the recipient of the letter will face difficulties in opening and reading it correctly.

Fonts are also available for some European languages. If one, for example, tries to include a quotation in Russian in an English document, one will find that one has no Cyrillic characters available. The same will happen if one sends a Spanish document in electronic format to someone in Greece; one will be told that Greek characters have replaced the accented Latin characters.

The solution is to leave behind the assortment of eight-bit fonts with their limit of 256 characters, where the same character number can represent a different character in different alphabets, and move on to a system that assigns a unique number to each character in each of the major languages of the world. Such a system has been



developed and is known as Unicode.<sup>8</sup> It is intended for use in all computer systems. Otherwise, the challenge for African languages would be to negotiate with Microsoft for their different characters to be immediately included, or the existing special characters must be made available to the general public. It is better for African languages to get as close as possible to the Latin languages and all other languages that use similar characters. That is the reason why that alphabet was chosen for Yipunu by the researcher.

### **3.11 Corpus structure**

It would be more convenient if all the files are saved in the same environment to avoid losing any of them. The development of the SYC project is still at its initial stage. This corpus needs to be enlarged with more material. It is too early to give the overall structure of the corpus. The bottom line is that all the files will be saved by topic or genre. The corpus at this stage has the following structure since few materials have been available and the researcher thinks that the corpus is sufficient for the present dissertation. One author can publish books in different domains; each book will be introduced in the relevant category.

#### **Religious Works**

1 Bible

#### **Traditional Stories**

2. Mumbwanga

3. Dilengi na Dilengi

#### **Literature**

Proverbs

#### **Cultural Aspects**

Cooking

---

<sup>8</sup> This information was taken from Alan Wood's Unicode Resources, American Indian Studies Research Institute, Unicode and Multilingual Support in HTML, Fonts, Browsers and Other Applications. AISRI Northern Caddoan Linguistic Text Corpora.

Hunting and fishing

Traditional building

Food

Riddles

As presented above, the Yipunu corpus is based on these listed domains that with further investigation will need to be enlarged in order to include all fields not covered by the present research. The researcher believes that a corpus can be considered as a garden that needs to be filled with plants. As long as there is not sufficient material within a specific category, one can fill these empty spaces if there is any need. One can remove unnecessary materials and replace them with more needed ones from time to time. Building a corpus is an ongoing process that needs patience and time in order to ensure a product that will suit the needs of the researcher, whose main aim is to satisfy the needs of the users of dictionaries.

There is still a great deal of work to be done in Yipunu with regard to lexicography in order to bring the language on a par with other languages in the world. The introduction of additional domains into the corpus will ensure the compilation of technical dictionaries. This will allow the compilation of different specialist dictionaries such as medical, agricultural, economical, accounting, commercial and legal dictionaries.

### **3.12 Concluding remarks**

This section has explained how the corpus can be sorted after gathering different sources. It has dealt with the way in which the data should be stored for better management of the files in order to retrieve the information needed. It has presented the different advantages of annotated texts versus non-annotated texts and the different reasons for keeping texts in a special format. The lexicographer has the mission to work with the community in order to standardise the language by choosing the appropriate alphabet if the language is still looking for one, like in the case of Yipunu. The international requirements should guide the choice. It is also crucial to realise the important role a language board can play in a specific language. The

establishment of a lexicographic unit for the compilation of dictionaries is an essential enterprise.

## **CHAPTER 4:       COMPUTERISING A CORPUS**

### **4.1. Introduction**

Computers have speeded up and improved research in languages. It is important to have the right equipment if one decides to gather huge amounts of language material. The storage and the retrieval of these materials require not only bytes in computers but also the necessary tools to handle text files in plain texts or rich texts. In this regard, there are several corpus query tools that have been developed across the world. In this research, WordSmith Tools was used for analysis of the corpus.

This section essentially deals with the way in which different sources have been processed onto computer. It explains the methods used to digitalise and transfer the material into different databases: Attention is paid to the keyboarding and scanning methods regarding the way in which sources were downloaded from the Internet. This is because all the materials have been edited and transformed using the orthography proposed in this research. A focus on a general analysis of the corpus via a quantitative and qualitative method takes the reader through statistics, frequencies, concordances, collocations and lemmatisation. Attention is also paid to the way in which inconsistencies in the corpus were taken care of in order to enable it to be effectively and easily utilised.

### **4.2. Data transfer methodology**

There are many ways of storing material on computer after gathering it from all the sources one needs for the corpus. The storage of material in a computer file can be done both manually and automatically. The transfer is possible via three different methods: keyboarding, scanning and retrieval from existing electronic sources, e.g downloading from the internet, as will be explained below in the manner it was done for the SYC.

#### **4.2.1. Keyboarding**

When dealing with language material, it is unavoidable to keyboard data for the material to be processed into digital form. Both spoken and written material can be

keyboarded to suit the aim of a given corpus. For the COBUILD project, Renouf (1987) explains that almost all spoken data were keyboarded, except for the BBC transcripts that could be printed more effectively. Most of the newspapers and some journals were also keyboarded. Newspaper material was initially sent to a commercial agency for reasons of speed, but the expenses involved forced the managerial team of the project to later undertake the work in-house. The cost factor, combined with the time involved in keyboarding material, made the Kurzweil Data Entry Machine (KDEM) the preferred alternative, if the material was suitable.

As far as the SYC is concerned, keyboarding was used to orthographically retype all the printed spoken material published in phonological or phonetic transcriptions. This has been done to correct the characters used in the original version of some existing material included in the corpus. The frequent use of special characters in African languages makes the process of material management very complicated, as can be seen in figure 4.1.

- 1 ' yíkə, biβúndə b́ámási:lə nánə tsí jétu.  
alors / PN pl. (8) + anciens / PV pl. (2) ils / pas. loin. / laisser / ainsi # nég. / nous //
- 2 mbá mé<sup>1</sup> nyisámalá<sup>2</sup> né biβúndə byē: <sup>3</sup>  
ensuite / moi / je / nég. / pas. loin. / voir / même PN pl. (8) + anciens-ceux-là //
- 3 na b́ámabúru, na b́ámabaŋgúya <sup>4</sup>;  
comme / PV pl. (2) ils / pas. loin. / être né // comme / PV pl. (2) ils / pas. loin. / grandir //
- 4 baménə<sup>4</sup>burə b́ama:ma <sup>5</sup>;  
PV pl. (2) ils / prés. ac. + aller-mettre au monde / PN pl. (2) + mamans //
- 5 bam:amə baménə<sup>4</sup>túburə.  
PN pl. (2) + mamans / PV pl. (2) elles / prés. ac. + aller / nous / mettre au monde //
- 6 mé nyímaburulu o Mukábə Dísimu.  
moi / je / pas. / être né / à / Mukabə Disimu //
- 7 mé dínə dyā:mi Majínu ma Kómbilə.  
moi / PN sg. (5) + nom – mon – Majinu – con. – Kombilə //
- 8 yibandu yā:mi jū:ŋgu.  
PN sg. (7) + clan – mon – ju:ŋgu //
- 9 yíkə, dusáβu dwedúdu, biβúndə b́amadusaβā:ŋgə.  
alors / PN sg. (11) + conte - celui-ci // PN pl. (8) + anciens / PV pl. (2) ils / pas. loin. / PN sg. (11) / conter + durée //
- 10 biβundə byē:nyi, ba jā:ji b́ā:mi bakədúwákə  
PN pl. (8) + anciens-ceux-là // PN pl. (2) + aînés - mes / PV pl. (2) ils / nar. / PN sg. (11) / revoir //

**Figure 4.1:** Example of the phonological transcription extract from Kwenzi-Mikala (1997)

As it can be seen, the text was full of unnecessary elements that needed particular attention for it to be usable. This can justify why the researcher opted to keyboard the whole book. The following text is the keyboarded version of the above text:

Yika bivunda bamasila nana tsi jetu.

Mba me nyisamala ne bivunda bye.

Na bamaburu na bamabanguga.

Bamena bura bamama.

Bamama bamenatura.

Me nymaburulu o Mukaba Disimu.

Me dina dyami Majinu ma Kombila.

Yibandu yami jungu.

Yike dusavu dwedudu bivunda bamadusavanga.

Bibunda byenyi bajaji bami bakaduwaka.<sup>9</sup>

The above example emphasises the importance of using the keyboard when transforming written material into a machine-readable version. Short sections of texts, especially if numerous, are best suited to keyboarding, as suggested by Svensén (1993:252), and have to be taken from a large number of sources with varying typography and layout. When keyboarding the material it is necessary to provide as many data categories as possible with unique codes so that they are accessible, for example, for verification and processing.

#### **4.2.2. Scanning**

There are many types of scanning systems in use for machine-readable text processing. Among others can be mentioned the KDEM, which has been stated to be capable of reading any typeface. The use of this machine requires for each new typeface a running-in period during which the system interacts with the operator and gradually learns to recognise the various characters. There is also optical character recognition (OCR) whereby the computer uses a light-sensitive scanning device to read the printed text and converts the dark parts of each character into ones and zeros in its memory. There are certain limits to its power of recognition, though. Some OCR software can be trained to recognise special characters such as Omnipage.

The original version of the above text of the Yipunu corpus could have complicated things even more if this text was scanned by the OCR system that was used to scan the existing Yipunu Bible. The scanning of the Bible was less difficult because it contained characters that the scanner could read and transfer in order to allow the edition of the texts. OCR was made available to the researcher by an HP scanner 1310 that scanned the entire Bible in a reasonable time.

---

<sup>9</sup> This text has already been translated in Chapter 3.

The Bible pages were placed face-down one by one on the scanner and the light of the machine crossed the page from one side to the other and photocopied the image and displayed it on the computer screen. Consider the following text as an excerpt of a scanned text taken from the Yipunu Bible:

Musalitsi aguvu na malubu edili maf, nli rnakaga tumba elasi kabu andi ombu

Ijlllramusunda;"tlmba ulasa yingeba ombu

la dtiyitsu.

Mutumubi amabetsu na mbivulu tsialuli tumba mutu usungamaedili'yisuemunu

Iit'si va dufu duandi..

Ijllllnakambulu Diela divu na bango dujabu tsiadivu

VII gari bidu.

Dusungugu duibangusi bulongu tumba

Illlsumu meburombisi yisonyi.<sup>10</sup>

The phoneme /m/ was in some cases scanned as /rn/ and both sequences /ab/ and /be/ were recognised by the programme as /OO/ in a number of words throughout the scanned text. All these mistakes needed to be corrected via certain methods, as explained in 4.2.4. The lexicographer has to be careful and vigilant while dealing with material such as that presented above in order to give him or her a format that can be beneficial to the utilisation of the corpus first and the compilation of the dictionary as a final product.

#### **4.2.3. Downloading**

The researcher received an electronic version of a textbook used by the Fondation Raponda Walker in the teaching of some of the Gabonese languages to extract some Yipunu texts. Those sources were sent by email and automatically transferred into a

---

<sup>10</sup> The text cannot be translated because it is an original scan that does not display all the characters in such a way that the meaning of all the sentences can be recognised. It is therefore given here only to show a sample of a scanned text.



text file format. To be once again utilised, the material needed a transformation of certain characters to suit the orthography used in the SYC. The sent file actually previously had to be transformed into an appropriate format via a special programme that was downloaded from the Internet. Microsoft Word then automatically changed the material into a readable format.

#### 4.2.4. Cleaning the corpus

Keyboarding and scanning texts are not 100% accurate since many errors can be encountered in the texts. One has to be careful and always verify all the material in the corpus.

When keyboarding texts, mistakes can be made by confusing some vowels and consonants. The researcher was forced to proofread the whole text by looking at the original form in the book, and in a couple of days all the misspellings in 30 pages of the corpus were corrected. Since a non-mother tongue speaker of Yipunu to whom indications of the changes of the alphabet were given did the typing of the text, the researcher was quite sure of finding errors in the text. The researcher could then observe that some words were written in a certain way instead of the other, as can be noted in the following:

Mulomi instead of mulumi (husband)

Nama instead of nami (with me)

Kiyuna instead of kuyina (only dancing)

Akemulongambura instead of akemulonga mbura (he/she went to show him/her the place)

The researcher had to use the *find and replace* function for the substitution of the incorrect spellings for the correct ones. It was also necessary to correct the double spacing as well as the semicolons, commas, question marks and other punctuation.

As far as the correction of the scanned material is concerned, a custom dictionary to correct all the mistakes was strategically compiled. The procedure was to generate a word list that was run with an existing French or English dictionary, and all the incorrect words were automatically underlined in red, as it is done if an English word

is misspelled. The words underlined in red were corrected and added to the dictionary. The procedure should be understood as taking Yipunu words one by one and by right clicking adding them to the custom dictionary active at the time for capturing additions. The compilation of the spelling checker is a shortcut that facilitates the procedure. By doing so, one can correct several texts in a short time and consequently speed up the work.

### 4.3. Corpus analysis

Analysis of the corpus was possible thanks to WordSmith Tools, which is a corpus query program, as mentioned in the previous sections. A corpus has limited value if one cannot easily extract the information needed for different purposes. Hanks (2004) rightfully defines corpus pattern analysis (CPA) as a new technique for mapping meaning onto words in a text. For him, a powerful aid in doing CPA is the Waspbench Word Sketches program of Kilgarriff and Tugwell (2001). According to Hanks, this exploits the concept of mutual information as a measure of statistical significance (Church & Hanks 1989), listing the words that are most associated (in terms of statistical significance) with the target word in different clause roles (subject, verb, object, adverb, etc.).

Similarly, Bergenholtz and Schaefer (1978:136), as quoted by Smit (1996), illustrate by means of the lexical item *Angst* (fear) how one could go about when processing the data in a lexicographic text corpus. According to them the citations in the corpus should be sorted and classified according to specific criteria. In the case of the above-mentioned lexeme, the authors explain that they searched the corpus for all occurrences of the sequences -ANGST- and -ÄNGST- as well as -angst- and -ängst-. After having printed out all the instances of occurrences, they sorted them into three piles:

- those containing nouns
- those with verbs
- those with adjectives and verbs

In the same way, Bergenholtz and Schaefer (1978:137) think that a lexicographer could easily land in a situation, especially in the case of frequently used lexemes, of

having too much material or giving too much attention to one lexeme and too little to another. They suggest that lexicographers should avoid such pitfalls by posing certain questions regarding the use of the lexemes in question. The questions must be formulated so precisely that another lexicographer could check the outcomes and arrive at the same conclusions or use the questions on other material to obtain the same outcomes.

One can then discuss some outcomes of certain questions towards the citations extracted from the corpus as well as the occurrences for attribute adjectives, for instance verbs in predicative function when used and the plural form of nouns, and so on. Bergenholtz and Schaeder (1978:154–165) list some suggestions about the contents a dictionary article for the lemma sign *Angst* should have in view of the insights gained from the citations in the corpus, as shown in the following list:

- The representation of the lemma: The lemma sign and its various morphological forms are presented.
- The explication: This may include, among others, a short narration or an iconic presentation to serve as an explanation. These could be followed by, for example, a paradigmatic explanation of meaning, which could include synonyms, collocations and other indications of usage. Then a syntagmatic explanation could follow. This might include examples that indicate the combination possibilities with attribute adjectives, paratactic nouns, verbs combining with *Angst* as grammatical subject and with *Angst* as grammatical object, and the occurrences of *Angst* in prepositional phrases.
- Examples could be given of the occurrence of the lemma in texts. These could be followed by statistics on the frequency with which the items occur in the corpus. References to literature that analyses or discusses the lemma sign could also be listed.

What has been said here has direct implications in this chapter and Chapter 5, and the researcher will attempt to clearly demonstrate this. The researcher believes that even if there is too much data at the disposal of the lexicographer, he or she should not be snowed under because the lexicographer has the final decision in choosing the data needed for inclusion or exclusion. It is also an advantage to have abundant material available because one can carefully and rightfully analyse the item in different

contexts of usage and choose the one that best suits the type of dictionary being compiled.

#### **4.3.1. General statistics**

In Table 4.1 below it can be noted that the overall sources in the SYC give a token total or number of running words of 1 352 905 and 64 660 types or number of different words. Source 1 contains 15 820 tokens for 5 147 types; Source 2 stands at 418 084 tokens and 24 014 types; Source 3 stands at 261 926 tokens and 12 260 types; and Source 4 has 652 004 tokens and 39 335 types. The Yipunu corpus stands at 1.3 million words, which is a good starting point when dealing with an African language such as Yipunu in which materials are rare and one is building an electronic corpus from scratch.

One can also note that the type/token ratio is 4.78 and the number of types in the whole corpus is far lower than the number of tokens. It is also obvious in Table 4.1 that there is a huge difference between the number of tokens and the number of types in the corpus, and the impact can be observed on the different ratios provided in the table. The sources display different ratios, when the gap between types and tokens is huge, the ratio is very low and if the gap is small, the ratio is big. This is quite explicit when comparing the type/token ratio of Source 1, which is 32.53, with the ratio of Source 4, which is 6.03, and the ratio of Source 2, which is 5.74, with the ratio of Source 6, which is 50.73.

The standardised type/token is 36.3, and the average word length is 5.28. There are 1 025 paragraphs and no sentence has been recognised. Furthermore, the length of words varies from one-letter words to words with 12 letters and more.

The fact that the language is written conjunctively has an impact on both token and type numbers in the corpus and the number could increase significantly if the disjunctive writing system is chosen.

TEXT FILE	OVERALL	SOURCE 1	SOURCE 2	SOURCE 3	SOURCE 4	SOURCE 5	SOURCE 6
<b>Bytes</b>	10 365 301	117 998	3 159 860	1 988 616	5 064 242	31 792	2 793
<b>Tokens</b>	1 352 905	15 820	418 084	261 926	652 004	4 659	412
<b>Types</b>	64 660	5 147	24 014	12 260	39 335	1 418	209
<b>Type/token ratio</b>	4.78	32.53	5.74	4.68	6.03	30.44	50.73
<b>Standardised type/token</b>	36.3	59.89	35.41	28.19	38.12	41.72	
<b>Ave. word length</b>	5.28	6.14	6.27	6.28	6.47	5.51	5.52
<b>Sentences</b>	0	0	0	0	0	0	0
<b>Sent. length</b>							
<b>Sd. sent. length</b>							
<b>Paragraphs</b>	1 025	184	230	84	522	4	1
<b>Para. length</b>	1 316.51	85.98	1 812.12	3 093.04	1 248.90	1 164.75	412
<b>Sd. para. length</b>	1 921.62	110.17	2 147.97	3 627.29	1 395.80	1 988.54	
<b>Headings</b>	0	0	0	0	0	0	0
<b>Heading length</b>							
<b>Sd. heading length</b>							
<b>1-letter words</b>	12 966	158	3 012	2 102	7 568	122	4
<b>2-letter words</b>	142 250	1 769	42 831	24 328	72 559	705	58
<b>3-letter words</b>	164 955	753	53 099	36 690	74 201	199	13
<b>4-letter words</b>	194 634	2 633	65 719	42 979	82 501	709	93
<b>5-letter words</b>	176 781	2 023	52 597	31 568	89 963	576	54
<b>6-letter words</b>	126 048	2 193	38 459	20 356	64 279	702	59
<b>7-letter words</b>	101 136	1 941	31 010	19 040	48 354	736	55
<b>8-letter words</b>	133 621	1 264	41 362	28 053	62 541	375	26
<b>9-letter words</b>	38 935	1 084	9 044	3 764	24 752	273	18
<b>10-letter words</b>	90 287	698	31 463	20 166	37 839	105	16
<b>11-letter words</b>	50 160	477	16 011	10 464	23 141	65	2
<b>12(+)-letter words</b>	6 176	275	1 814	832	3 216	32	7

**Table 4.1:** Overall statistics of sources used in the SYC

### 4.3.2. Word frequencies in corpora

It is obvious that frequencies in corpora have a direct consequence for the total count in terms of tokens and types. Regarding the above statistics, if there is only a 4.78% types/token ratio in the whole corpus, it is because some word frequencies are very high. Texts in the corpus give lists of words in terms of types and frequencies and some statistical profile of the relation between types and tokens by indicating the distribution of types across the text categories. They also give summaries of lists in a form that can be assimilated by the user of the corpus.

The corpus is investigated for the frequency of words and word combinations and the result is used to ensure that data selected to be included in the dictionary are as representative as possible. If the material is not sufficient, the result could be less useful for practical lexicography. This is one of the reasons why it is important to take account of the number of types and the number of tokens.

Frequencies in the corpus can cause confusion at a morphological level in the sense that the text will give the frequencies of words that have the same form when used in different contexts. In Yipunu, for instance, *ndeju* (you) can also be used in its diminutive form *nde*, *mama* (mother) can become *ma* and *jaji* (older brother or older sister) can be used in some contexts as *ja*. These words will then be counted separately for frequency because the program used for the calculations is not aware of all the variations in the language. It is up to the lexicographer to pay attention to the different forms that a given word can have.

This situation is identical with the one that can be faced when dealing with homographs in the corpus. Words with the same spelling will occur in the same frequency count. The same can be seen in words that belong to several parts of speech as highlighted by Svensèn (1993) when he says that, the syntactic occurrence of a given character string, can provide important clues in a language such as English: compare the use of *light* in the typical expressions ‘put on the light’ (noun), ‘a light room’ (adjective) and ‘light up the room’ (verb). This problem can be solved by the lexicographer when dealing with the senses and parts of speech of words during the analysis of concordances of a specific lexical item. Frequencies drawn from the SYC can be seen in the following top 200 hundred words:

**Table 4.2:** Frequency word list of the 200 top frequencies in the SYC

Rank	Word	Frequencies	%
1	NA	27 647	2.06
2	VA	5 433	0.40
3	RIE	5 243	0.39
4	FUMU	4 918	0.37
5	MU	4 785	0.36
6	NSAMBI	3 299	0.25
7	NESI	2 810	0.21
8	OMBU	2 222	0.17
9	BATU	2 206	0.16
10	ANDI	2 081	0.15
11	MUANA	2 010	0.15
12	KAGA	1 985	0.15
13	U	1 909	0.14
14	AVA	1 766	0.13
15	MUTU	1 705	0.13
16	TE	1 672	0.12
17	MBANA	1 624	0.12
18	YIRI	1 619	0.12
19	NANA	1 598	0.12
20	BISI	1 587	0.12
21	NDAGU	1 478	0.11
22	BAANA	1 450	0.11
23	BOTSU	1 435	0.11
24	AMI	1 429	0.11
25	YIARI	1 373	0.1
26	BULONGU	1 341	0.1
27	PA	1 336	0.1
28	TUMBA	1 324	0.1
29	YI	1 323	0.1
30	DIAMBU	1 297	0.1
31	MBARI	1 253	0.09
32	MOSI	1 245	0.09
33	BAISRAEL	1 242	0.09
34	DIBAALA	1 217	0.09
35	NE	1 139	0.08
36	NDERI	1 121	0.08
37	JULU	1 105	0.08
38	MBURA	1 096	0.08
39	A	1 090	0.08
40	GUSU	1 089	0.08
41	PAGU	1 058	0.08
42	BANDI	1 036	0.08
43	JANDI	998	0.07
44	YIKA	980	0.07
45	MENU	959	0.07
46	FU	950	0.07
47	GARI	947	0.07
48	MUNA	947	0.07
49	LA	940	0.07

50	TINDI	917	0.07
51	YILUMBU	912	0.07
52	AGUVU	901	0.07
53	AMA	867	0.06
54	NDE	864	0.06
55	JOGU	849	0.06
56	MOTSU	835	0.06
57	BIOTSU	828	0.06
58	DAVID	825	0.06
59	DIBANDU	792	0.06
60	ENI	787	0.06
61	TEMU	786	0.06
62	BABABAALA	773	0.06
63	MOISE	771	0.06
64	AYI	758	0.06
65	BA	745	0.06
66	MANGOLU	741	0.06
67	AMAVOSA	738	0.05
68	NO	737	0.05
69	O	725	0.05
70	JENU	698	0.05
71	MAMBU	674	0.05
72	BENI	648	0.05
73	KABOGU	648	0.05
74	AVU	629	0.05
75	WANDI	626	0.05
76	MUSIENGI	615	0.05
77	DEDI	606	0.05
78	TAJI	602	0.04
79	AJI	601	0.04
80	JERUSALEM	600	0.04
81	VANA	587	0.04
82	ENU	579	0.04
83	ADI	557	0.04
84	BOGU	555	0.04
85	TSI	550	0.04
86	ABI	548	0.04
87	BAGU	521	0.04
88	DIANDI	494	0.04
89	AMABA	487	0.04
90	NANDI	486	0.04
91	ISRAEL	484	0.04
92	BATAJI	481	0.04
93	YIMOSI	480	0.04
94	DIINA	475	0.04
95	MUGETU	471	0.04
96	BAMI	469	0.03
97	MALONGU	459	0.03
98	KA	456	0.03
99	MAMBA	451	0.03
100	MBE	449	0.03
101	EGYPTE	446	0.03



102	MAGU	441	0.03
103	KAMA	432	0.03
104	BU	426	0.03
105	OGU	424	0.03
106	MABOTI	420	0.03
107	MA	418	0.03
108	MOGU	416	0.03
109	VO	413	0.03
110	DIELA	412	0.03
111	BUTAMBA	405	0.03
112	MANDI	405	0.03
113	WISI	404	0.03
114	ABA	401	0.03
115	BAGORA	400	0.03
116	UPAGU	393	0.03
117	MUISI	389	0.03
118	MISIENGI	387	0.03
119	MURIMA	376	0.03
120	MIGAGA	375	0.03
121	YIANDI	375	0.03
122	BAPAGU	374	0.03
123	UBUEJI	369	0.03
124	NAGU	366	0.03
125	BENU	364	0.03
126	YIPAGU	360	0.03
127	BATEGULA	357	0.03
128	BAVU	356	0.03
129	KALALA	355	0.03
130	AMATSINGULA	353	0.03
131	YIGUMI	353	0.03
132	MIOGU	351	0.03
133	TSIRI	349	0.03
134	BUANDI	342	0.03
135	MUKONGU	341	0.03
136	YILIMA	341	0.03
137	BAKAGA	339	0.03
138	MAGUGA	339	0.03
139	MAMI	338	0.03
140	PANGINI	336	0.03
141	YISAMBUALI	333	0.02
142	KABU	332	0.02
143	MABI	331	0.02
144	YIFUMBA	331	0.02
145	MUTUBU	329	0.02
146	ANANA	328	0.02
147	AMARUMA	326	0.02
148	NSANGU	326	0.02
149	DIWENDI	321	0.02
150	LEVI	321	0.02
151	LORA	321	0.02
152	YISALU	321	0.02
153	SAUL	320	0.02

154	TSIOTSU	318	0.02
155	DI	317	0.02
156	GU	317	0.02
157	YIENI	317	0.02
158	MUJI	315	0.02
159	NGUJI	315	0.02
160	AGUMABA	311	0.02
161	BIGUJI	308	0.02
162	BISALU	308	0.02
163	BAMABA	307	0.02
164	UBA	303	0.02
165	ETU	302	0.02
166	ABAMABA	299	0.02
167	NAMUNYI	299	0.02
168	ANYI	298	0.02
169	DIBI	296	0.02
170	JACOB	296	0.02
171	BIMA	293	0.02
172	BIVUNDA	293	0.02
173	BI	292	0.02
174	KUMU	292	0.02
175	ABAVU	291	0.02
176	DIVITA	291	0.02
177	NSILA	291	0.02
178	BIOGU	287	0.02
179	MANDAGU	287	0.02
180	MISU	286	0.02
181	BIANDI	285	0.02
182	MUVAGULITSI	284	0.02
183	MUKUTA	282	0.02
184	BAPUELA	279	0.02
185	UOTSU	279	0.02
186	BEJI	276	0.02
187	MIINA	276	0.02
188	BASUSU	274	0.02
189	AARON	272	0.02
190	BAJUDA	266	0.02
191	DIVU	264	0.02
192	AMAVAGA	263	0.02
193	NYIVU	263	0.02
194	NSIMA	261	0.02
195	MUA	260	0.02
196	BABAALA	259	0.02
197	MULUMI	258	0.02
198	BIBULU	257	0.02
199	YIOTSU	257	0.02
200	JA	256	0.02

It is easily seen from these frequencies that the most frequently used word in the corpus is *na* (and or with) ranking at 1 and its frequency count 27 647 but with a very high percentage of 2.06 overall, i.e. more than once in every 50 words in a given paragraph. The rest of the frequencies are under 1% of the total count. This proves that in a corpus, there are words that occur more frequently than others but all of them are relevant to helping the lexicographer to make major decisions on the usage of items.

One cannot say enough about the importance of frequency counts in lexicographic research. Gouws and Prinsloo (2005:32) point out that frequency counts obtained from the corpus assist the lexicographer in solving one of the crucial problems in relation with the compilation of dictionaries: what to include in and what to exclude from the dictionary. According to Gouws and Prinsloo, with the help of the corpus lexicographers should be able to motivate the reasons for every inclusion in or exclusion from the dictionary.

In a corpus some words occur only once; they are called hapex legomena. They usually constitute almost half of the total word count in the corpus. In the present corpus, words with a frequency superior or equal to two constitute 30.88% of the total and words occurring with a frequency equal to one constitute 69.12%. In the same way, Clear (1987) mentions that for the COBUILD project, the corpus contains a very large number of strings with a frequency of one. The majority of these strings do not have to be considered as headwords or lexical items and typically are typographical errors, formulae, initials, roman numerals, proper names, KDEM errors, and so on. These single items are automatically presented in alphabetical order by the program itself.

Consider in the following table some typical hapex legomena:

**Table 4.3:** 15 Hapex legomena extracted from the SYC

Word	Frequencies
ABEFUTI	1
ABYEBI	1
AGANPALA	1
AGEGARU	1
AGEGIMI	1
AGEGULU	1
AGEJI	1
AGEKIPI	1
AGOBA	1
AGOGARU	1
AGOLA	1
AGOLABA	1
AGOMWERETSAMA	1
AGOMWEWAGULA	1
AGORUNGULA	1

Frequency counts have other important functions as they help the lexicographer in both the macro- and the microstructural treatment of the different items in dictionaries. A more detailed explanation of this will be given in Chapter 5.

Nevertheless, it should be emphasised here that frequency counts are also helpful in the construction of relevant word lists, also known as lemma lists or lemma candidate lists. These items are the ones from which the lexicographer has to decide which words to include in the dictionary and which to leave out. Frequency counts of specific items can be drawn across different sources of the corpus in order to evaluate the spreading rate of those sources. More on that is to be said in the following section.

### 4.3.3. Word lists

From any given list of words that gives the frequency counts of different words, a lemma candidate list can be generated. That word list is actually a lemma list or lemma candidate list. The problem mentioned earlier in connection with the lexicographer's facing difficulties in deciding which of the different words to include in or exclude from the dictionary is addressed in more detail here. When constructing the lemma candidate list, the lexicographer must make some important decisions about the frequencies to be considered as qualification for inclusion in the lemma list of the dictionary.

Word lists are very useful for the production of a basic list of lemmata. Lemma lists can be taken from other dictionaries and vocabularies. They can also be compared with the corpus frequency list. A word list can also be given in alphabetic order with their frequencies in the corpus or as a simple lemma list.

Any dictionary that is compiled in Yipunu has to use a corpus. Thus frequencies, concordance lines and lemma candidate lists should be the base for the compilation of such a dictionary. As for now, the emphasis is on giving an example of a word list drawn from the corpus, as shown in the following two tables:

**Table 4.4:** Section of a simple alphabetical word list from the Yipunu corpus

ABYEBI	
ADI	
ADIDI	
ADINA	
ADYEDI	
AGA	
AGABUGA	
AGAJABA	
AGANA	
AGANPALA	
AGEGIMI	
AGEGULU	
AGEJABI	
AGEJI	
AGEKIPI	
AGELA	
AGESUNDUGI	
AGOBA	
AGOGARU	
AGOLA	

**Table 4.5:** An alphabetic frequency word list from the Yipunu corpus

Word	Freq.
A	37
ABA	2
ABAVHU	2
ABEFUTI	1
ABYEBI	1
ADI	3
ADIDI	2
ADINA	5
ADYEDI	2
AGA	2
AGABUGA	2
AGAJABA	2
AGANA	2
AGANPALA	1
AGEGARU	1
AGEGARWE	1
AGEGIMI	1
AGEGULU	1
AGEJABI	2
AGEJI	1
AGEKIPI	1
AGELA	2
AGESUNDUGI	2

The different items in the lemma candidate list can be checked in the corpus for their occurrences in context and cotext by studying a number of concordance lines for the particular word. The concordance lines taken from the corpus serve multiple purposes, as shown below. These lemmas, presented in strict alphabetical ordering, highlight the way the macrostructure could look like in a dictionary. The lexicographer has to be guided by the results drawn from the corpus in order to take important decisions on both the macrostructure and microstructure dimensions.

#### 4.3.4. Concordance lines in the SYC

Observation has clearly shown that concordance lines depend on the frequency count of a particular lexical item in the corpus. The more frequently the item occurs in the corpus, the more concordance lines it will have. This can help to solve the problem of examples and sense limitation, polysemic values, hierarchies of meaning in a polysemic structure, etc. in the corpus (but see page 126). The lexicographer has to

pay attention to these phenomena in the corpus to avoid confusions and unclarity in the microstructural treatment of the relevant lexical items in dictionaries. It is therefore worth it to indicate that concordance lines play an important role in the corpus. Consider here the concordance lines of the following items: *na* and *nana*:

**Table 4.6:** Concordance lines for the associative *na* (with)

bye bimakeji Atsipala	na	bitsatsu bye bimakeji
a mwana dibala bonawe	na	tsisiga Akavhungila
alila o dituji we Misopu	na	bambombu na masan
ganga jojimaresa nguji	na	dimi Aji vhana Aji o
angula Atsiganga kodu	na	kodu Tata nzmabi at
a vhavha ka ja masyala	na	nguji me ja nguji tum
andi Yika vhana ustana	na	nguji Yilumbu yimos
kunga yisigu Akapala	na	yisigu Amabura ta le
batu botsu Pwela batu	na	abavhu mu mugula N
jenu dwirungi o pungu	na	yiliba yika tsyatsi na
o nzaji Atsigoka o muji	na	bakwaga bandi mbe
a katsi Atsileliga bigari	na	tandu Marundu ageki
ndu Aji vhana Manyiki	na	bayinnyi baka na mu
gulinga Mutu amapalana	na	wandi mwana Amay
Yika esundugi nana Ja	na	mwana katsi sobana
walangu Atsibengunu	na	murela Murele nana
mukongu kayi gumata	na	nzambi vhavha yika t
na banumba Na bafudu	na	banumba twendyanu
gomina bitsiga Numba	na	wandi mufudu Numb
la amatindanga kwanga	na	bamabanga nana bisi
mana biguji Atsiwenda	na	yibuku yandi yi musy
Na bafudu na banumba	na	bafudu na banumba
a maponzi na manyabi	na	koku na taba na yot
na abavhu mu mugula	na	abavhu mbura tsyots
guna ha dina Banzagu	na	batsiesi bawa guna h
Nana vhavha no dugasu	na	Mwabye Yi o yi me
a usogama nana vhavha	na	vhana Sunga diyebul
mbunga Eruyi Mbata	na	taji mbata na taji Ta
ye dina dilengi Bakoku	na	bataba bawa guna ha
ina kola dilengi Bakoku	na	bataba bawa guna
adina yivhunda yi dibala	na	yivhunda yi mugetu
Misopu na bambombu	na	masangu bi vhana dul
e dibala akawela digumi	na	bagetu baranu Baget

**Table 4.7:** Concordance lines for the comparative *nana* (like)

Yika tsye ukendengila	nana	nde winsabye Yay
abila mikanzu Agajaba	nana	atsibanzila muji Yi
mupumeguna unosuka	nana	Mupuma mosi nob
engila Ja ayina yi evha	nana	Amakunga vhana
ba Akapala vha dimbu	nana	yigumbi yimaweru
muna amatsoka bapela	nana	Mumbamba umati
a nganga wa amavhosa	nana	Bilongu bimawend
a Ta lembu yika emati	nana	Atsitola o matayi
itsatsu yimosi yikipala	nana	Ayina yitsisuka Y
magasa amagasa Aka	nana	vhavha Mba agom
u Ayepala vha mulandu	nana	Atsibo bitsatsu by
epala vha mwila unenyi	nana	vhevha masanga
ungili banyabi Akebura	nana	mwa vhavha Ne ka
muji Yike nzaji emuvhe	nana	Atsibo nzaji Atsig
tsisika atsisika atsisika	nana	tumasikila matsuvh
syeta ne utsyeta Ndubi	nana	nderi batsoka Dik
a dimbu Aketola dimbu	nana	vhevha mbu kasa y
unda yi dibala Dimanyi	nana	vhavha ngomfi esigi
elyenyi Mba baketsapa	nana	Bakevhindama na
mumbari yi ustivhagila	nana	mumbwanga asa
mapala Atsunyi yisalu	nana	yitimba Ya la yiti
akakotisa bana ba nguji	nana	bisi tsugu Bana b
u Mangala amatebuga	nana	vhavha tuvhu mu m
dibaga tuka reba mama	nana	tumarebila jami Ts
tsisingiga nana Atsiwe	nana	vhana Atsetsana n
yebi bikeki Mungili uka	nana	kongudu Atsitinda
u Bana bakakotisa taji	nana	sugu Mba bana b
u bi makeji Atsisingiga	nana	Atsiwe nana vhan
kwanga na bamabanga	nana	bisi tsugu Bana b
u di ja rundu Nzingulya	nana	nyuwendila Ngang
iguna Marundu agajaba	nana	atsieryabila mikan
umbumbu Bana abefuti	nana	Amabura amabura
ali vha kodu dimbu Aka	nana	mwana dibala bona
a Botsu yika nana yika	nana	yika nana Yivhem
Amweyeba mwa nyanyi	nana	mwa vhavha Batu
na amabongila musomu	nana	Akafwamina biguji
a Dina diko mwila Aka	nana	solye Yika kwang
yika didi dikenkwela Yi	nana	Ukawela marundu
akeyisunza nana Yike	nana	nguji akaruga Yi n
be jitsitba Jo bako julu	nana	mandagu amema j
unduga Yika esundugi	nana	Ja na mwana kat



ji mbe jike kumu Kumu	nana	yivhevhye Tsya m
a dimosi dimayenapala	nana	makulu na myogu
utsu nodibuga Atsikota	nana	Sunga mu dilaban
pa nana Bakevhindama	nana	Akenasunza jandi
gunu na murela Murele	nana	musavhu kombila
ye Mumba duka pwela	nana	kadi nyiregiminyi J
etu aguna etabuli dimbu	nana	awelu Yi nesi aga
mbu vha dulombi ebegi	nana	mwa vhavha mwa
anyine Ya twendi tuvha	nana	tumavhala Batsye
umba Jandi jandi kodu	nana	o mwila Avhavha g
gu Dina Dina Ya dina	nana	tumavhanganyine
bwanga Odi musyengili	nana	Odi musyengili na
nzagu akamurambusa	nana	vhavha no fwala aji
giji kunyi ka gunu Nguji	nana	bila o ndeki Sunge
ikake Dikaka ayenavha	nana	A dikake me nyiv
ayibonga Bakeyisunza	nana	Yike nana nguji a
gutsitabuga Botsu yika	nana	yika nana yika nan
enu me mulumi atsipala	nana	kabogu Bayepala
tsi Batsyebala mulaku	nana	vhana jetu twidubili
ili nana Odi musyengili	nana	Tsya mbolwanu a
eburu nekuru yi mindubi	nana	mwa vhavha Ndubi
u Na yiliba yika tsyatsi	nana	ayivhu nenu 0 pung
i ugovhyoga Ovhyo yari	nana	sefu yi ugovhyoga
nga yirungi yi tsana Aji	nana	mwa vhavha Sung
e taji osyala o bulongu	nana	vhavha no dugasu
i labya make ma mama	nana	makavhu ka batu
u Aletana Ovhyo yari	nana	sefu yi ugovhyoga
letana Atsivha kongudu	nana	mwa bipela abyebi
e Bakabanda usogama	nana	vhavha na vhana S
nyine Ya ka tsye dina	nana	tumavhanganyine
a amabongila mulembu	nana	Yabutsu nodibuga
a Ka ja mwana mugeyi	nana	mwa vhavha yisalu
okasana Atsibatandiga	nana	dupelwe Ya Yayi
usavhu amatoba vhana	nana	vhamabila penyun
ila tata malam Dimbu	nana	dyetu didi di maba
ga Yika tsye wimpagili	nana	Me uvhyo nyivhyo
ibusi Taji yi Yina kabo	nana	Yika taji akavha y
tsye A mwane dibala	nana	musavhu kombila a
ika nana yika nana yika	nana	Yivhembra dina dy
nu Bakapala vha dimbu	nana	yigumbi yimaweru

One can note that these concordance lines give important information about the occurrences of the items. A concordance is meant to create a list of all occurrences of each word in a text. Presently, these two items were focused on just to give examples of occurrences from the Yipunu corpus. Normally, each occurrence is supposed to be accompanied by data on its position in the text, the text from which the line was drawn.

In the above-mentioned concordances, to avoid filling up the text with too much data and not to confuse things, the researcher opted to limit the data on the environment of the word. This means that the occurrences give five words before and five words after the search word. That will be discussed in more detail in the section on collocations. These keywords-in-context concordance (KWIC concordance), are words or phrases extracted from a text and listed in alphabetical, frequency or other order, together with the words occurring in its immediate environment (Hartmann & James, 1998: 79) and are useful in many ways. This has a direct link with the context of the words that is known as the part of a text where a particular word, phrase or term occurs (also called ‘verbal context’ or ‘cotext’) (Hartmann & James, 1998: 29).

Concordances can be utilised for the investigation of the combinational properties of words, in other words their construction. They can also enable the identification of collocations and idioms. In a concordance, the occurrence and use of words can be used as a basis for the division of dictionary entries according to meaning and as sources of dictionary examples. A comprehensive analysis of this will be given later in this chapter and in Chapter 5 as well.

One can note the way in which the examples containing *nana* are sorted according to the first and the second words following *nana*: *yika nana*, *mwa vhavha*, *bila o*, and *nguji akaruga*. This kind of occurrence can be sorted for each word in the text as long as one chooses the preceding word.

In a polysemous lexical item like *crawl* or *read*, a single look at the concordance lines will help the lexicographer to detect the main senses. In a verb like *crawl*, one will be able to identify senses such as ‘moving hands and knees’, ‘time moving slow’ etc. These different senses will help the lexicographer in writing the definitions of a polysemic word.

#### 4.3.5. Collocations

Yet again, there is a relation between frequencies and collocates in the corpus in such a way that the occurrence of words in collocations is higher or more significant when the frequency of the occurrence of a given word is high. Makkai (1971) mentions that it follows, from the foregoing observations dealing with frequency, those collocational ranges are themselves subject to frequency and geo-sociolinguistic stratification. The table below is used here only as an illustration. The instruction given to the corpus query tool to calculate and list the collocates of a given word in order of their frequency, looking up to five places to the left and up to five places to the right, one gets tremendous results. If one looks at a specific lexical item, one will know the number of times the items collocate with each other in the lines in the horizon L5-R5 (hence in the range from 5 places to the left up to 5 places to the right). It will also give the number of time these collocates occur to the left of word and to the right of the same word, a breakdown of occurrences to the right (times one, two, three, four and five to the right). All these observations cannot be imagine by any mother tongue speaker using her/his intuition. Consider Table 4.8 for collocations with *na*:

**Table 4.8:** Collocation range for *na* within the horizon L5-R5 in SYC

WORD	TOTAL	LEFT	RIGHT	L5	L4	L3	L2	L1	*	R 1	R 2	R 3	R 4	R 5
DINA	16	5	11	0	1	2	1	1	0	3	1	1	1	5
TAJI	16	3	13	0	0	1	2	0	0	6	0	2	2	3
YIKA	14	11	3	3	2	4	1	1	0	0	1	2	0	0
DUFU	13	5	8	2	0	2	1	0	0	5	0	0	1	2
BAFUDU	12	4	8	1	0	1	0	2	0	4	0	3	0	1
BANUMBA	12	4	8	0	0	1	0	3	0	4	0	2	0	2
NYUNDU	12	7	5	2	0	5	0	0	0	0	0	0	5	0
UBWEJI	12	7	5	2	0	0	5	0	0	0	1	4	0	0
VHANA	12	7	5	2	0	1	4	0	0	1	0	2	1	1
YISONYE	12	7	5	0	2	0	5	0	0	0	0	0	0	5
MAMA	11	9	2	1	1	1	0	6	0	0	0	2	0	0
MBATA	11	8	3	0	2	0	2	4	0	0	3	0	0	0
NANA	10	6	4	0	3	0	3	0	0	0	1	1	1	1
YIMATSYEMUGILA	10	5	5	0	5	0	0	0	0	0	0	5	0	0
YITSIBA	10	5	5	5	0	0	0	0	0	0	5	0	0	0
DIBALA	9	7	2	0	0	3	3	1	0	0	0	2	0	0
MBOLWANU	9	6	3	1	3	0	2	0	0	0	0	2	1	0
MUMBWANGA	9	6	3	2	1	1	2	0	0	1	1	0	1	0
ALETANA	8	6	2	2	1	2	0	1	0	0	0	0	1	1
DILENGI	8	6	2	2	0	1	3	0	0	0	1	0	0	1
GUNA	8	2	6	1	1	0	0	0	0	0	0	4	0	2
MBWANGA	8	4	4	0	1	0	0	3	0	0	0	0	4	0
MUNA	8	2	6	0	2	0	0	0	0	4	0	0	0	2
MWANA	8	4	4	0	3	1	0	0	0	2	1	0	0	1
TSYA	8	5	3	3	0	2	0	0	0	0	2	1	0	0
WANDI	8	3	5	0	1	2	0	0	0	3	0	1	0	1
NGUJI	7	3	4	1	0	1	0	1	0	2	1	0	1	0
NYIGANA	7	4	3	1	0	0	3	0	0	0	2	0	0	1
UGOVHYOGE	7	4	3	1	1	0	0	2	0	0	2	0	0	1
YANDI	7	2	5	0	2	0	0	0	0	3	1	1	0	0
YIBUSI	7	2	5	1	0	1	0	0	0	1	4	0	0	0
MUFUDU	6	3	3	1	0	1	1	0	0	0	2	0	0	1
NZAMBI	6	3	3	2	0	0	0	1	0	1	2	0	0	0
TATA	6	0	6	0	0	0	0	0	0	0	4	0	0	2
YIKE	6	3	3	0	2	0	1	0	0	0	1	1	1	0
BAWA	5	1	4	1	0	0	0	0	0	0	4	0	0	0
EPALI	5	3	2	0	0	1	0	2	0	0	0	0	2	0
KOKU	5	2	3	0	0	1	0	1	0	1	0	1	0	1
MANYABI	5	3	2	1	0	1	0	1	0	1	0	1	0	0
MARUNDU	5	1	4	1	0	0	0	0	0	0	2	1	1	0
MBWANGE	5	4	1	1	1	0	0	2	0	0	0	1	0	0
NUMBA	5	4	1	2	0	0	0	2	0	0	0	1	0	0

One can observe that *na* co-occurs 16 times with *dina* and *taji*, 14 times with *yika* and 13 times with *dufu* but also in low frequencies, for example five times with *epali* and *numba* and seven times with *yandi* and *nguji*.

Such occurrences are usually invisible in citation files and can only be discovered through the use of corpora. This is the reason why the identification of collocations if available is very important in corpus lexicography, as observed by Landau (2001). By applying statistical norms to them, the relative significance of their association is indicated. They play an important role in LSP and bilingual dictionaries when readers are unfamiliar with the common associative patterns of the language they are trying to learn.

#### 4.3.5.1. Word clusters

From the concordance lines of any word in the corpus, WordSmith Tools can search for clusters of that word in the corpus. In this regard, one can consider clusters as words that occur together in a certain position and in a certain way in the corpus with a frequency superior or equal to three. Consider in the following table the cluster drawn from the concordance lines for the Yipunu locative *vha*:

**Table 4.9:** Example for Yipunu locative cluster *vha*

Cluster	Freq.
ngunge vha mataye	7
ja lembwe ngunge vha	6
vha mataye nyimalunga ja	6
nyimalunga ja lembwe vha	6
vha mataye nyimalunga	4
dibenyi ne vha	3
katsi barata vha	3
lembwe ngunge vha	3
munu ne vha	3
nana yigumbi yimaweru	3
ne vha julu	3
ne vha mbasu	3
ne vha munu	3
vha julu dibenyi	3
vha mbasu ne	3
yigumbi yimaweru maranu	3

As can be observed from Table 4.9, the locative *vha* is a cluster of 16 different lexical items at any position in the corpus. In fact, the program brings out the number of

times the search word was used in context with other words. The locative *vha* here shows that it can be used in central position and in first position, as one can note in the following examples:

*ne vha mbasu* (even on the nose)

*vha julu dibenyi* (on the breast)

In the same way, clusters, according to Scott (2000), are words that are found repeatedly in each other's company. They represent a tighter relationship than collocates, more like groups or phrases. Scott calls them clusters because this term has already been used in grammar.<sup>11</sup>

All these examples provide valuable data to be used in the dictionary on the microstructure level, as translation equivalents, for instance. The problem could reside in the fact that it is not always easy to find a typical translation equivalent of a lexical item between the source language and the target language. The context in which the cluster is used will determine its sense in the sentence. The position of the cluster can change the meaning of the whole sentence. In the next chapter this will be shown in more detail.

#### ***4.3.5.2. Sorting definitions and examples***

Corpora are of great help to the lexicographer when it comes to defining a word and retrieving examples for microstructural treatment of a specific item. These will then be used for the translation equivalent as well and for the illustration of usage in the specific context of a given lexical item. This is possible only with a well-designed and representative corpus, containing large numbers of words. Those millions of words will furnish enough evidence that the lexicographer can use as examples or to determine the paraphrases of meaning.

By retrieving examples from the corpus, the lexicographer is departing totally from the traditional way of practicing lexicographical work whereby researchers were self-listing words to construct the macrostructure and also used self-made examples and definitions. A dictionary has to reflect the language of a given community. That is

---

<sup>11</sup> This information is from a handout given during a lecture by Professor Prinsloo.

why the involvement of the speakers of a language is vital when gathering material. This once again shows the revolution brought about by corpora in linguistic and lexicographic research. The typical example retrieval is facilitated by the use of concordance lines that give the lexicographer enough material for his or her theoretical and practical work.

For instance, in Yipunu the lexical item *dina* has many senses:

- *Name*: ka nde dina dyagu yi Ya dina dyami dilengye (and you what is your name, he answered my name is dilengi).
- *Namesakes*: dilengi tukuba dina na dina mugetwe ya yino (dilengi we will be namesakes, the woman said yes).
- *The other*: divhavha diko mayumba dina diko mwila Aka na (one is far like from here to Mayumba and the other from here to Mouila<sup>12</sup>). In this example *dina* can be also be considered as a homonym.

For the sake of disambiguation, the researcher checked the concordance lines and found all the necessary evidence to distinguish the three senses. The lexical item presented above has polysemous senses. Each sense should be treated in its own subcomment on semantics. The illustrations are presented as cotext entries and are relevant evidence of the usage of the translation equivalents.

The same methodology will be implemented in the next chapter when demonstrating how a corpus can help with a specific type of dictionary. There are many contextual elements about definitions that are not given here. However, it is important to mention that corpora contain a variety of data that can be used to enhance the quality of definitions in a dictionary. The lexicographer has to be vigilant in the analysis and extraction of data as she or he has the objective of not losing focus.

#### **4.3.6. The ruler as instrument to measure the macrostructure stretch**

With corpora, many applications are possible, as pointed out in the previous paragraphs. The lemma lists drawn from the corpus help the lexicographer to

---

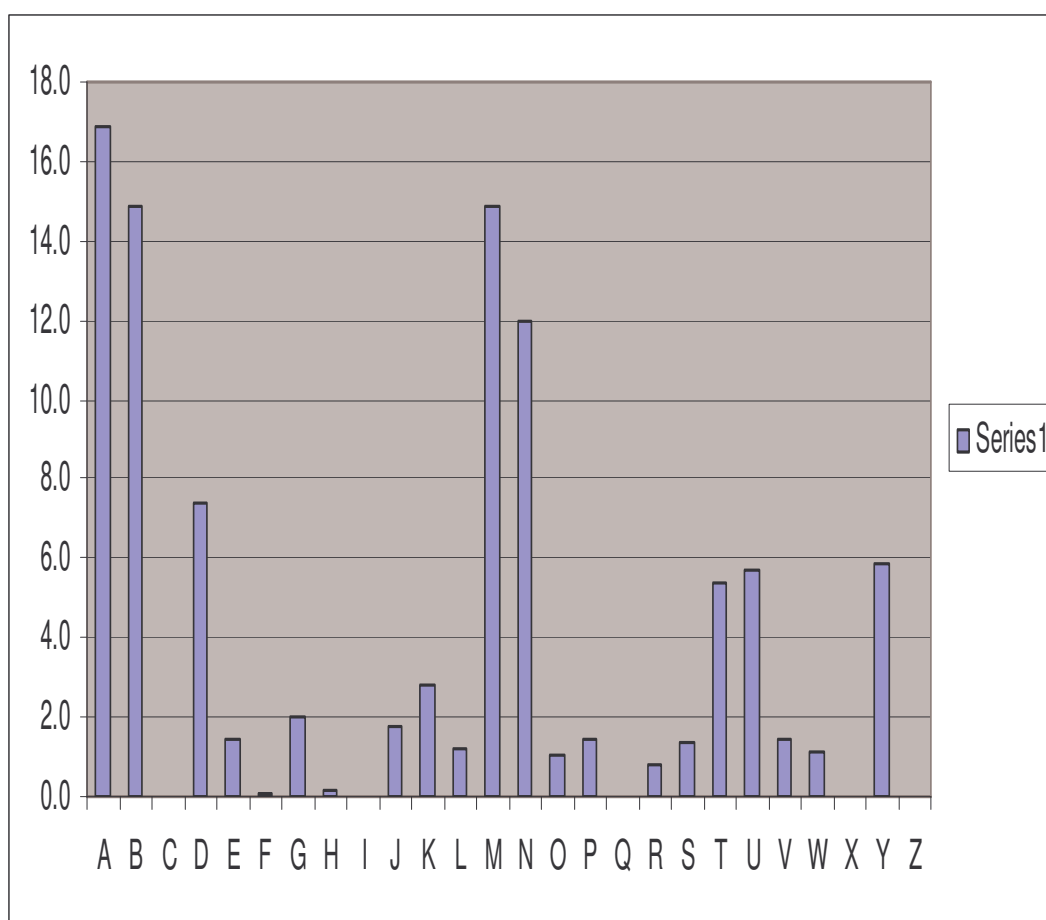
<sup>12</sup> Mouila is a capital province in Gabon.

calculate at the macrostructural level in order to know more or less the percentage of the alphabetical stretch of each letter in the dictionary.

In the following ruler, 1 550 lemmas were used for the calculation. One can note the absence of the letters C, I, Q, X and Z in Yipunu. That explains why these letters appear with a value equal to zero. If one has to compile a dictionary in Yipunu, one has to bear in mind the limitation of the alphabet in the language. A comprehensive discussion on the use of rulers can be found in Gouws and Prinsloo (2005). In order to address the inconsistencies on the macrostructural level, Prinsloo and De Schryver (2002; 2005) studied the balance between alphabetical categories for English, Afrikaans and the African languages in existing dictionaries and electronic corpora of these languages. They designed practical instruments of measurement (and even prediction) for the relative length of alphabetical stretches in alphabetically ordered dictionaries according to the generally accepted principle that alphabetical categories in any given language do not contain an equal number of words or lexical items. This is also done for the ruler based on the Yipunu corpus. For Prinsloo and De Schryver, the question would be whether a specific balance, preferably one that could accurately be measured, exists between the different categories in a given language. The ruler can be constructed for each language in Gabon and the balance between the different alphabetical categories will depend on each language.

By looking at Figure 4.2, one notes that the letter A has the top value of 17%, followed by B and M with 15% each. The letter N is in fourth position with 12%, D has a value of 7.5%, Y stands at 6%, U and T stand at more or less 5% each, K stands at 3% and G and J each has a value of 2%. The letters E, P, S and V have approximately the same value: 1.5%. Finally, F and H rank at the bottom with values close to zero: 0.1 and 0.2% respectively. Consider Figure 4.2 as the ruler that allows the measurement of the alphabetic stretch for the lemma candidate list of Yipunu, based on two sources of the whole corpus.





**Figure 4.2:** Alphabetical stretch measurement of the SYC

One can then take these different values into consideration when deciding on the number of lemmas to include in a dictionary for each alphabetical letter. Obviously, here the letter A will have a larger number of lemmas, followed by the rest proportionately with their different values.

#### 4.3.7. Projecting Yipunu databases

The researcher believes that at a later stage, for the ease of data manipulation there will be a need for compiling a database for Yipunu and for each one of the Gabonese languages. In order to accomplish this, a database structure will need to be designed. There are two broad database-structuring approaches: a paper-based system and a computerised system. Both of them are useful for lexicographic works.

The researcher does not think that lexical databases can be used in the Gabonese environment for now because the project is still at the initial stage and needs to be improved for better usage. Each language in Gabon has yet to compile comprehensive corpora before the construction of databases. In order to succeed in a project such as that of creating databases, as mentioned earlier, the first step should be the compilation of large corpora in the Gabonese languages, which is an urgent task awaiting lexicographers because without language materials no publication of dictionaries and productive databases can be envisaged. A database is a sophisticated enterprise that in this case can use the existing and completed corpora. Svensén (1993) points out that as yet, there is no consensual strategy as to the structure of such a lexical database, let alone how to annotate corpus data to fit into it so that it might be of use to the lexicographer.

However, the paper-based system used before and still used nowadays by lexicographic units such as the Bureau of the WAT can take the form of cards on which the material the system contains is written by hand, or typed data can also be included. The system also accommodates pasted cuttings from publications. The structuring of data in such a system would include an indexable entry, for example an entry arranged alphabetically or a date and cross-reference to other entries. It can also include information such as translations, parts of speech, plurals, synonyms, source references, excerptor details and comments on usage, for example regional, slang, archaic and etymological detail.

As far as the computerised system is concerned, it may contain different forms of data records, text files containing data markers or tags, or concordances. The structure of data records is similar to the item types on cards, for example keyword, part of speech, citation, source reference, translation and synonym.

Both the computerised and the paper-based system can be suitable for the Gabonese context because they are easy to design and to manage. They can also be used simultaneously by more than one editor and will facilitate the search for any data in the corpus. In an environment where people are still getting used to systems such as these, it is better to keep things as simple as possible to avoid confusion and chaos.

#### **4.4. Corpora as means of enhancing the quality of the lexicographic representation**

The analysis of corpora is a phase of great importance as it gives the lexicographer an overall view of the data included in the corpus and she or he could judge whether this data is sufficient for the compilation of the dictionary. This data will be used for the macro- and microstructure in the central list of dictionaries. This is explained meticulously in the next chapter. The user of the dictionary who consults the final product will judge the compiler in the end. This urges the lexicographer to be efficient when building the corpus.

If built adequately a corpus could be used to obtain semantic, pragmatic, grammatical and stylistic data in order to enhance the quality of the lexicographic representation. In fact, many dictionaries lack appropriate data that users are looking for.

##### **4.4.1. Corpora and semantic data**

In previous chapters, the benefits of corpora in helping in meaning and sense disambiguation were touched on slightly. It is clear that disambiguation is only possible if one uses material in context rather than isolated material, otherwise the meaning will not be apparent or obvious. This gives credit to an assumption raised by Firth (as quoted by Landau 2001: 281)<sup>13</sup> who believes that language study should have its points of departures in actual occurrences of language use in different texts rather than in isolated sentences made up on the basis of intuition.

In the same way, both Halliday (1991) and Sinclair (1987) estimate that no language can be meticulously analysed if the material used does not reflect the usage of language in context. To fulfil the mission of decontextualisation, the description should allow the demonstration of the near mutual relation between meaning and grammar. Sinclair believes that before one could really say what words mean, one must first draw the relation entertained by those words when used in sentences. The description of the same words when used in sentences is subject to the identification of their meanings.

---

<sup>13</sup> Landau (2001: 281) does not give any bibliographic reference of Firth that he quotes.

In line with the above, it is worthwhile to emphasise that in this regard, corpora as containers of text types from different sources can help with the identification of the meaning of words and sentences as semantics tend to be the study of both, as expressed by Saeed (2003).

In fact, there is a close relation between meaning and form in such a way that the two cannot be separated. This is a foundation of corpus linguistics with regard to the study and language description. Mahlberg (2005) estimates that corpora can allow the observation of collocations and patterns of words that make visible the meanings that are common to the members of a discourse environment. Consequently, the author shares the view that the meaning of a given text is based on the interaction of conventional patterns and novel or creative combinations of words. Literary stylistics is concerned with creative or unusual collocations; therefore, the relationship between meaning and form is analysed with regard to the effects that can be achieved by deviating from linguistic norms.

In this regard, Mahlberg (2005) continues arguing that corpora can play an important role. They can make available useful tools to identify textual features that characterise the style of a particular text. They can also enable comparisons of typical patterns with features foregrounded in particular text(s) by a particular author, and corpus linguistic techniques can suggest quantitative approaches to the analysis of texts.

Taking into account example sentences taken from different texts, the lexicographer is in a favourable position to better illustrate the different usages of lexical items. Provided with abundant material he or she will choose the examples that are suited best to the microstructural treatment of these lexical items. Words are often used in different contexts and the meaning is not always easy to grasp by simply defining the items. Examples will help in this regard. To illustrate particular linguistic features in dictionaries, the same example can be used to describe different usages in the microstructural treatment of another lexical item in another dictionary article. This means that there is no restriction on the number of times an example can be used in the same dictionary. What is important is the ability of the example to clearly and rightfully express the intention of the lexicographer by providing the users with the information needed.

Sharing the same view, Hori (2004) and Stubbs (2001) believe that the use of a corpus in stylistic analysis may, for instance, help to describe the development of a narrative by looking at distributions of content words or the analysis may focus on collocations that reflect the way in which characters are portrayed.

Semino and Short (2004) rightfully stipulate that corpus linguistic methodology can be used to test existing theories for the stylistic analysis of texts. With the help of corpora one can identify features that a reader may not be fully aware of when she or he arrives at a particular interpretation of a text. It can be clearly affirmed that unusual instances of collocations may turn out to be part of a textual pattern that is created by a set of semantic prosodies spanning large passages of text.

Another issue that can be raised here concerns spoken language, which is a latent problem in computational linguistics and artificial intelligence. These two domains underestimate the usage of oral data because of their tradition of focusing on written language. Despite the described situation, semantic processing of speech is necessary for the understanding of natural language. In societies where written material is insufficient, oral material is successfully used for the semantic illustration of words. Sometimes the lexicographer has to deal with what is available in the language. Lexicographic data can be found in spontaneous speech and rightfully enable its semantic representation. Fundamentally, research on the semantic representation of speech can provide one with a suitable basis for further analysis of related linguistic levels, such as prosody or pragmatics.

For Ronald (2003) corpus linguistics is much more than just lexical bean-counting; it provides a new perspective on language use, one in which probability is more central than possibility, through which previously unavailable types of data become available and previously invisible word patterns become visible. As such, it enables the investigation of collocation, connotation and phraseology from a perspective that is not otherwise possible. More than this, corpus linguistics has opened up language study to exciting advances in theories of language competence, of *langue* and *parole*, of lexis-grammar relations, of connotation and of textual cohesion.

Mindt (1991) demonstrates how a corpus can be used in order to provide objective criteria for assigning meanings to linguistic terms. This is quite often observed in dictionary articles of monolingual dictionaries in which lexical items are treated in a

way that gives the user the meaning of the item he or she is looking for in order to satisfy his or her particular communication need. In this regard, corpora assist the lexicographer in conveying the right meaning of the items by providing the necessary data. This is done by extracting from a corpus example sentences that help the lexicographer to achieve his or her goal. Corpora will be even more efficient in identifying meaning in polysemic lexical items via texts that characterise their syntactic and morphological usage in context. The meaning can be caught in text corpora by considering the environments of the linguistic entities as an empirical objective indicator for a particular semantic distinction.

Corpora allow the study of semantic relations in texts between words and their collocations, including antonymy and near-synonymy. These features of language create lexical cohesion over a large proportion of text. By taking into consideration the results of lexical items studied in context, the corpus is the right source that helps one to obtain semantic data in order to improve the quality of the lexicographic representation.

#### **4.4.2. Corpora and pragmatic data**

Pragmatics is known as a field of research that studies language in context, according to Myers (1991) and Fasold (1990); it is the use of context to make inferences about meaning. Pragmatics and semantics are related to some extent. The intension here is not to enter into the debate some scholars have introduced about whether or not pragmatics overlap with semantics; the researcher rather considers them as separate disciplines that share some features to a certain extent. This is why semantics has been treated in another paragraph above. A pragmatic data will be every data that expresses the contextual use of a given lexical item that has to be presented and treated in the different dictionary articles.

The study of a language in a community, therefore, is an asset that no doubt will ensure the analysis of the language material needed by the pragmatist or lexicographer in a certain context. The researcher believes that in dictionary articles, context can ensure good results if one takes into consideration natural languages in the way they are respectively used in texts and sentences. Researchers in lexicography have everything to benefit by using corpora in their studies because they contain enough of

the contextual materials that are necessary for such studies. In Chapter 5, the researcher will show with regard to the dictionary how material helps to capture the contextual usage of words.

Lexicographers are encouraged to adopt the tendency of putting data into dictionaries that reflect the usage of words (in context) by the speech community instead of the way they portray themselves the use of words in the language. Studies that do not take into consideration pragmatic data will not yield conclusive results certainly because of the non-usage of precise language material that reflects the real usage of that language in context. To fulfil her/his objectives, the lexicographer is once again asked to use a corpus in order to reach better conclusions. In this regard, the use of corpora will help in the enhancement of pragmatic data in the microstructural treatment of lexical items in dictionary articles. All pragmatic studies use language material in their assumptions, even if they are limited to a context of what needs to be demonstrated. Besides, this statement is applicable to almost all disciplines that use language as evidence.

As an illustration, it is important to reveal that in a series of lectures at Harvard University in 1967, the English language philosopher H.P. (Paul) Grice (1967) outlined an approach to what he called conversational implicature, which is how hearers manage to work out the complete message when speakers mean more than what they say. As an example of what Grice meant by conversational implicature, the following utterance can be looked at:

“Have you got any cash on you?”

Hereby the speaker really wants the hearer to grasp the following meaning:

“Can you lend me some money? I don’t have much on me.”

In fact, the conversational implicature is a message that is not found in the plain sense of the sentence. It is implied by the speaker in the way he or she expresses him- or herself. The hearer is able to understand or imagine that message in the utterance by appealing to the rules governing successful conversational interaction. Grice proposed that implicatures like the second sentence could be deduced from the first, by understanding three things:

- The usual linguistic meaning of what is said.
- Contextual information (shared or general knowledge).
- The assumption that the speaker is obeying what Grice called the cooperative principle.

One might not see the direct link between some ideas expressed in this subsection because they point out some general issues. The researcher believes that to fully understand the concept of pragmatics it is necessary to point out some issues. The fundamental benefit of this comprehensive analysis of pragmatics is provided at the end of the section.

All these observations are possible when a researcher is busy gathering language material needed for the study in accordance with his or her principles. The assumption is that much of the work that has been carried out in this area has used only real conversational corpus material, as seen with Grice's theory of conversational implicature presented above. It is of no less importance to indicate that all the studies in pragmatics have used limited samples of data for particular purposes. If one uses real language data, one can encounter real linguistic problems that will grab one's attention and enable better analysis and conclusions, as already pointed out.

According to the understanding of Liu (2000) most of the aspects of language studied in pragmatics are in relation with the following:

- Deixis: This means 'pointing to' something. In verbal communication, however, deixis in its narrow sense refers to the contextual meaning of, for example, pronouns and in its broad sense what the speaker means by a particular utterance in a given speech context. It must also be pointed out that in verbal communication deixis has a much wider influence.
- Presupposition: This refers to the logical meaning of a sentence or meanings logically associated with or entailed by a sentence.
- Performative: This implies that by each utterance a speaker not only says something but also does certain things: giving information, stating a fact or displaying an attitude. In addition, the study of performatives led to the speech act theory that holds that a speech event embodies three acts: a locutionary act,



an illocutionary act and a perlocutionary act, as remarked by Austin (1962) and Searle (1969).

- Implicature: This refers to an indirect or implicit meaning of an utterance that is derived from the context and is not present in its conventional use.

Furthermore, Liu estimates that such studies also allow exploration of why interlocutors can successfully converse with one another in a conversation. Normally, the principal conception is that interlocutors obey certain principles in their participation so as to maintain the exchange. For Grice (1975) one such principle is the cooperative principle, which assumes that interactants cooperate in the conversation by contributing to the ongoing speech event. In the same way, Leech (1983) identifies the politeness principle that maintains that interlocutors behave politely to one another, since people respect each other's faces (Brown & Levinson 1978).

Many pragmatists share the assumption that a referential cognitive description of social interactive speech events was provided by Sperber and Wilson (1986). They rightfully state that in verbal communication people attempt to be attached to what they intend to say and to the receptor of the utterance. As a result, there has been an increase in curiosity about how people in different languages observe a certain pragmatic principle. Cross-linguistic and cross-cultural studies reported that what is considered polite in one language is sometimes not seen as polite in another.

It is observable to some extent that contrastive pragmatics is not restricted to the study of certain pragmatic principles. It must be indicated here that elements such as cultural breakdowns and pragmatic failure, to name a few, are indeed also part of cross-cultural pragmatics. It is also important to reveal that research in pragmatics has taken into account learner language or *interlanguage*. In accordance with interlanguage pragmatics, the emphasis is on the analysis of the way in which non-mother tongue speakers understand and generate a speech act in a target language and how their pragmatic competence develops over time (Kasper 1995; Kasper & Blum-Kulka 1993;).

Stenstöm (1987) tried to analyse 'carry-on signals' such as *right*, *right-o* and *alright*. He categorised these signals conclusively according to the typology of their various functions. He found out that

- *right* was used in all functions but especially as a response to evaluate a previous response or terminate an exchange;
- *alright* was used to mark a boundary between two stages in discourse;
- *that's right* was used as an emphasiser; and
- *it's alright* and *that's alright* were responses to apologies.

With the growth of oral and text corpora around the world, it is important that this material be available and at the disposal of researchers dealing with works of this kind. The contexts explained above show once more that all these analyses are only possible when one has large numbers of texts available. All the described situations can be found in text corpora that are gathered in different contexts and contain several sources. Analysing them will give the lexicographer enough evidence and encouragement to extract sentences for the illustration of specific situations that are important for the understanding of a lexical item in a special context. The lexical item may have a certain sense in a particular context and a different sense in another situation. One of the challenges for the lexicographer is to find always a suitable sentence example to illustrate the usage of a lexical item. The use of corpora in lexicographic studies is unavoidable in developing the quality of examples in the microstructural treatment of dictionaries.

#### **4.4.3. Corpora and grammatical data**

In the previous chapters, the researcher has tried to show and explain how annotated corpora for part of speech can play an important role in the analysis of the material included. The annotation also allows grammatical recognition of lexical items. If well designed and annotated, a corpus query programme can help the researcher to call up different word categories in the corpus. Thus, it can be possible to retrieve nouns, verbs, adjectives, adverbs, conjunctions, demonstratives, and so on. Corpora contain different parts of speech patterns. In order to fulfil its mission, the corpus needs to be representative enough by containing material from different sources and language varieties. In this dissertation, the researcher presents different word categories that are part of the grammar of Yipunu, as the SYC has allowed the extraction.

Interestingly, as referred to by Landau (2001), it must be mentioned that Quirk in 1959 set out to collect 200 samples of British English, each consisting of 5 000 words, to form a corpus of a million words. The corpus was to be used to study the grammar and usage of educated British speakers and writers. The corpus provided comprehensive detailed annotations of the grammatical features of each word as well as a rising inflection or a pause for spoken material.

Jan Svartvik of the University of Lund in Sweden continued Quirk's work in 1975. His aim was to digitalise the spoken portion of Quirk's survey. After adding new material, he reached a total of more than 500 000 words, to which was given the name of London-Lund Corpus, published in 1980. It is known as one of the largest computerised corpora of spoken material. The corpus was subsequently used for careful grammatical analysis and later it was used along with the Brown Corpus and the LOB Corpus as a foundation for the largest grammar of English. The work was then imitated worldwide by researchers and has since been an important resource for the study of grammar in particular. In addition, it was concluded that because of the identification of grammatical features this corpus represented the point of automatic tagging and parsing for corpora.

It should be mentioned that in the study and analysis of grammar for quantitative analysis based on corpora, the work of Schmied (1993), who studied relative clauses, and that of Oostdijk and De Haan (1994a), who also endeavoured to analyse the frequency of the various English clause types, is typical. To these two examples can be added the study initiated at the University of Nijmegen, as reported by Aarts (1991), in which principally rationalist formal grammars are tested on real-life language contained in computer corpora. The way in which the grammar is analysed is in line with the way in which the data were gathered. At a later stage, the grammar was transferred into a computer parser and run over a corpus to test how far it accounts for the data in the corpus.

As reported by Antaki and Naji (1987), Quirk *et al.*'s (1985) reviews of grammar of causal connectives have taken into consideration the following: *because, since, so, for, as, in order to, by virtue, in the light of* and *due to*. Some other terms such as *in case* and *seeing that* have been identified along with other circumstantial clauses. The study showed that *because* is by far the most frequently observed causal link in the

corpus of about 200 000 spoken and 100 000 written English words, according to Altenberg (1984).

The importance of corpora is such that examples extracted from them display different types of sentence of different degrees of difficulty that a user can come across during the consultation process. It is when looking for specific information that is not found in a dictionary that users would label the dictionary as being not good because they did not find what they were looking for.

In line with what has been said earlier, concordance lines and the word frequency list will give indications of the content of the corpus in terms of grammar and text genres in that language. As far as grammar is concerned, corpora will provide different tenses in which a verb or a sentence is used in a specific context. Different types of sentence should be presented in dictionaries to illustrate different tenses of verbs via an example sentence. The same is applied to text production situations where different examples are used to specify different usages. This means that there will be situations where the present, past and future tenses are clearly distinguished. In the same way, one can distinguish verbs (transitive and intransitive for instance) from nouns, adjectives, pronouns and adverbs, to name a few. In African languages, causatives, ablatives, possessives, passives, perfectums, applicatives and also locatives, which are grammar phenomena, can be extracted from a corpus.

In the SYC the following elements generated by WordSmith Tools can be displayed.

As elements belonging to the category of nouns, items such as the following are found:

**Table 4.10:** List of different nouns and their frequencies from the SYC

Items	Meanings	Frequencies
1 Mwana	Child	2 010
2 Kaga	Grandparent	1 985
3 Mutu	Person	1 705
4 Ndagu	House	1 478
5 Bulongu	World	1 341
6 Mbari	Palm tree	1 253
7 dibala	Man	1 217
8 Mbura	Place	1 096
9 Pagu	Fine/fee/contribution	1 058
10 Yilumbu	Day	912

All the above items are in the singular form and, as can be noted from the top to the bottom, one can identify the prefixes *Mu-* (class 1), *Ka-*, *N/M-* (class 9), *Bu-* and *Di-* (class 5). For instance, the rules of plural formation show that almost all the words starting with the prefix *Mu-* (class 1) will form their plural with *Ba-* (class 2) – for *mutu* and *mwana* one will have respectively *batu* and *bana*. Words like *ndagu* and *mbari* will become *mandagu* and *mambari* (class 10) in the plural. The lexical item *yilumbu* will become *bilumbu* (class 6) in the plural. Not all these lexical items are tagged as nouns in the corpus. The lexicographer who is supposed to know the language makes the identification.

As far as verbs are concerned, particles such as morphemes and affixes can help to identify the different tenses. This is possible with the help of the usage context to distinguish the meaning of the items, as can be noted in the following examples:

**Table 4.11:** List of some verb forms and their frequencies from the SYC

Items	Meanings	Frequencies
Uba (infinitive)	To be	303
Agumaba	The one who was	311
Bamaba	They were	307
Abamaba	Those who were	299
Ubagengila	You (will) look at them	2
Ubaganga	You (will) catch them	2
Ubalisila	You (will) feed them	2
Tuvhayianu	Let's do it	2
Tutsibusa	We do not want	2
Tumuboka	We will kill him	2
Tumuyitsyanu	Let's worship him	2

Another illustration will be the case of words expressing the localisation of places and objects and some introducing possession, as in the following:

**Table 4.12:** List of some locatives and possessives with their frequencies from the SYC

Items	Meanings	Frequencies
Ombu	To/at	2,222
Gari	Middle/inside	947
Jandi	Him	998
Bandi	His/her (plural form)	1,036
Aguvhu	The one who is	901
Jogu	Them	849
Jenu	You (plural form)	698
O	At	725
Bogu	Their	555
Enu	Your (plural form)	579
Bagu	Your/yours ( one person)	521
Bami	My/mine	469
Mandi	His/hers	405
Mogu	Theirs	416
Vha	On	413
Nagu	With you	366
Benu	For you/your (plural form)	364
Mami	Mine	338

All the items presented above are important materials that one can use for the treatment of lexical items in a Yipunu dictionary, and the same procedure can be implemented in the rest of the Gabonese languages. The gained experience can be extended to the other African languages only if there are corpora available in those languages in order to avoid constructed examples by promoting authenticity in dictionaries. The forms and positions of the lexical items given above are based on their contextual usage with the other items in the syntagm or sentence.

The use of corpora in lexicographic representation allows the researcher to use more language evidence than self-made assumptions, examples and nonrepresentative lemmata that was practiced by certain scholars even if the the lexicographer remains the only to make final decisions especially when it comes to choosing the right definition of a given word . This is the reason why, when comparing the quality of dictionaries, one will find that some use materials that reflect the way in which the language is used authentically while others use examples that were invented by the researcher and that express his or her own opinion. The difference will be obvious in the way lexical items are presented in dictionaries

The use of concordance lines will help the lexicographer to provide users with the data needed for special purposes when they are consulting dictionaries to satisfy their different needs. For instance, the lexicographer can clearly explain and analyse the sentences that the user will face in the day-to-day usage of the language. Using a corpus as an aid, the extracted materials will help in the compilation process of dictionary texts. It is of great importance to understand that corpora allow the segmentation of words in terms of their structures in order to distinguish morphemes from phonemes, lexemes and syntagmas, which are significant for linguistic and lexicographic studies.

In line with the illustrations given above, it is also relevant to mention that the multifunctional character of corpora could open ways to the use of corpora in language teaching for the Gabonese languages since textbooks are still rare. That is possible because around the world many corpora have been used as illustrative material for the teaching of some languages, as was the case for English. Early in 1921, as mentioned by Landau (2001), the so-called first modern comprehensive corpus, i.e. large corpus of English compiled for lexical studies as opposed to literary analysis was Edward L. Thorndike's word count of 4.5 million words. It was published as *The Teacher's Word Book* and enlarged over the following decades in several stages to a corpus of 18 million words. It was republished as *The Teacher's Word Book of 30 000 Words*, based on texts from magazines and juvenile reading. It consisted of several lists of words showing their relative frequency. The book was designed to help educators and teachers determine which words are common enough to be used at particular grade levels.



In accordance with the above, it is of great importance to mention that Holmes (1988), Kennedy (1987a; 1987b) and Mindt (1992) have made considerable contributions to the usage of corpus materials. Kennedy in his analysis looked at different ways of expressing quantification and frequency in English as second language (ESL) textbooks, while Holmes was interested in the manner of expressing doubt and certainty in ESL textbooks. Mindt looked at future time expressions in German textbooks of English.

#### **4.5. Concluding remarks**

When collecting material for inclusion in the corpus, whatever method is used, it remains the lexicographer's responsibility to effectuate the necessary transformation of the material in order to fulfil his or her needs and objectives. It is more comfortable to work with a cleaned corpus even if a dirty corpus can also be of much value to the lexicographer.

It can be said without any doubt that the corpus is very important for the improvement of both the macro- and the microstructural treatment of lexical items. The section has also helped in constructing a ruler that can assist the lexicographer in measuring the relative alphabetical length of the alphabetical stretches in a specific dictionary. Corpora supply the lexicographer with data that will be presented in dictionaries in order to better serve the users.

## **CHAPTER 5: CORPORA AND THE COMPILATION OF DICTIONARIES**

### **5.1. Introduction**

It is generally acknowledged in both practical lexicography and in metalexicography that a dictionary, as a container of text types, is divided into three different components, namely the front matter texts, the central list and the back matter texts. All these dictionary parts display important data needed by the user of a specific dictionary. The quality of dictionaries is judged by the data they provide, and researchers have found ways to improve the performance of dictionaries. This has been accomplished through the introduction of corpora into lexicographic theory and practice.

On the one hand, researchers have already demonstrated the way in which corpora help lexicographers to improve the macro- and microstructural treatment of the dictionary article. On the other hand, the identification of dictionary functions has also played an important role in the compilation of dictionaries. In this regard, there is no doubt about taking into consideration these theoretical issues when dealing with a particular dictionary. In this chapter, the researcher tries to stretch the statement that gives credit to the compilation of corpora according to dictionary types. The researcher also endeavours to demonstrate the relation between the corpus and the knowledge-oriented function and the communication-oriented function. In order to achieve this aim, it is necessary to give a background of decisions regarding dictionary typology, the macrostructure and the microstructure. This will help with the lemmatisation and the treatment of the lexical items on different levels.

### **5.2. About different typologies**

Many researchers have focused on the identification of dictionary types. When looking at the typologies favoured by theoretical and practical lexicographers, one finds an inadequacy in these typologies. One must then agree with Gouws and Prinsloo (2005) that without a doubt none of these classifications can be regarded as absolute. This statement points out the idea that lexicographers must propose a

common classification that will suit everyone. This could also help in solving the problem of standardisation in metalexigraphy.

As far as the macrostructure and the microstructure are concerned, the problem of classification also needs to be tackled carefully in order to be resolved, and all the researchers are quite dissatisfied with the way things are dealt with on both the theoretical and practical levels. One has to believe that a corpus helps to improve the macro- and microstructural presentation of a specific dictionary type. The intention here is not to give an exhaustive analysis of all the types but to focus on the main features. These will be illustrated in the following paragraphs.

### **5.2.1. Different classifications of dictionaries**

As pointed out earlier, several classifications of dictionary typologies have been proposed by lexicographers, among them Geeraerts (1984), Geeraerts and Janssens (1982), Gouws (1989), Hausmann (1989), Kukenheim (1960), Landau (1984), Malkiel (1962), Rey (1970), Sebeok (1962), Svensén (1993), Swanepoel (2003) and Zgusta (1971). By giving these different classifications of dictionaries, the researcher does not have the intention of judging the quality of the research. The classifications are given here to show the importance of the topic in lexicography and explain the need for a definite agreement among lexicographers once again.

Nevertheless, all these classifications of dictionaries are valuable material for dictionary compilation. When one decides to compile a dictionary, one always thinks about the users and their needs before deciding on the type of dictionary one has to compile. It is counterproductive if a lexicographer is not clear about where he or she is going to start and what he or she is going to write about. This makes it clear that one has to make choices when planning a dictionary project. Choosing a specific type of dictionary to be compiled in advance helps the lexicographer to solve many problems, such as uncertainty, ambiguity and even confusion. What would a dictionary look like if one just decides to compile a dictionary for the sake of doing things? That dictionary will not even have the form of what is called a hybrid dictionary. One has to plan the compilation of any type of dictionary in such a way that one knows the envisaged objectives of the dictionary project in advance.

The activity of dictionary compilation involves many steps such as the definition of a target group, the purposes and the functions of the dictionary and the material that will be used. That is followed by establishing the structure of the dictionary in terms of the data distribution procedure. This, once again, goes towards deciding which data should be strategically used in the macrostructure and in the microstructure. The corpus enables the whole process from the starting point. Without a well-designed, representative and balanced corpus, this lexicographic process could yield limited results. Ultimately, the product is going to be judged by the general public; this will depend on who the intended target users are. A successful dictionary must be eminently suitable for utilisation by people for their different objectives. These statements will be dealt with throughout this chapter.

The classification of dictionaries into different types goes along with the classification of types of corpus in such a way that the material of the different corpus types is used for the compilation of the corresponding dictionary type. A bilingual corpus would best suit the compilation of a bilingual dictionary but can also provide valuable data for other dictionary types. Both dictionaries and corpora overlap in their features and functions. One can rightfully use a monolingual corpus of two different languages as a basis for the compilation of a bilingual dictionary. It must be emphasised that the 1 000 most frequent words in one language are not necessarily the equivalent of the 1 000 most frequent words in another language. It must also be emphasised that a corpus containing two languages with one being the translation of the other can play an important role in the process.

It is generally recognised that dictionaries are different from each other in many ways. The differences can be seen in their aims, for example, whether it is a learner's dictionary or a school dictionary; in their scope, in other words the size of the dictionary, for example a comprehensive dictionary; and in the subjects they deal with (limited or general). In a dictionary, the lexicographer presents a part of the lexicon of a language. The material included is sorted in accordance with linguistic and typological principles. Consequently, dictionaries differ from one another in their representation of data with regard to this lexicon. Three different dictionary classifications can be seen in the following paragraphs.

In his classification, Landau (1984) identifies the different types of dictionary. He takes into consideration the following criteria: the number of languages, the age of the

users, the scope, subject and size of the dictionary as well as the manner of financing.

He proposes the following types:

- Monolingual dictionary
- Bilingual, trilingual or multilingual dictionary
- Synonym dictionary or thesaurus
- Scholarly dictionary
- Descriptive dictionary
- Commercial dictionary
- School dictionary
- Desk dictionary
- College dictionary
- Unabridged dictionary
- Elementary dictionary
- Special-field dictionary
- Slang dictionary

Landau (1984) presents a nonformal typology and merely a convenient way to highlight significant differences among dictionaries. The categories are not exclusive according to him.

As for Malkiel, the criteria of range, perspective and presentation have guided his classification. He suggests the following dictionary types:

- Monolingual
- Bilingual
- Multilingual
- Diachronic
- Synchronic
- Didactic

Zgusta (1971) gives a comprehensive classification of dictionaries in the following manner. He makes a definite distinction between encyclopaedic dictionaries and linguistic dictionaries:

- Encyclopaedic dictionaries (encyclopaedia)
- Linguistic dictionaries
  - Diachronic dictionaries (historical and etymological)
  - Synchronic dictionaries
  - Restricted dictionaries
  - General dictionaries (standard-descriptive dictionaries, overall-descriptive dictionaries)
  - Monolingual and bilingual dictionaries

As can be noted, there is a strong similarity between the different types of dictionary classification presented above. The researcher did not find it necessary to give the defining features of the listed types. This will be done in the rest of the dissertation as the researcher intends to show once again the implications of corpora for dictionary typologies via the macro- and microstructure. This has already been demonstrated by Nielsen (1995), Prinsloo (1991; 1994; 2004), Prinsloo and De Schryver (2000; 2001; 2002; 2005) and Prinsloo and Gouws (1996; 2000), De Schryver and Prinsloo (2000; 2000a).

Most of these analyses are pioneering works that have already shown how the use of corpora could improve the macro- and microstructure of a specific dictionary. In this dissertation, a special emphasis has is put on the demonstration of how a corpus can be used in the macro- and microstructure of different types of dictionary. Before doing so, however, it is necessary to something about the macrostructure and the microstructure.

Dictionaries also differ with regard to the types of macrostructure and microstructure they present to users. Depending on the type of dictionary, the lexicographer can use the same macrostructure for the compilation of both monolingual and bilingual dictionaries. The major difference between these two types will be the microstructural treatment of the lemmas. These differences can be shown from one type of dictionary

to another according to the size of the dictionary, the target users and the function that each dictionary type would achieve.

The researcher believes that, overall, there are basically two types of dictionary: the monolingual dictionary and the multilingual dictionary.

The term *monolingual dictionary* refers to dictionaries dealing with one language; in other words, the presentation and treatment of lexical items are done in that unique language. This category includes all dictionaries from the desk dictionary, to the explanatory dictionary, to the special-field dictionary.

The term *multilingual dictionary* refers to dictionaries dealing with more than one language; in other words, the presentation and treatment of lexical items are done in two or more languages with one being the source language and the other(s) the target language(s).

### **5.2.2. Different types of macrostructure**

It is of importance to mention that in this section the relation between the macrostructure and the lemmatisation strategies is not treated in detail. The macrostructure leads to the way in which the lemmas will be presented in the central list of a given dictionary. The strategies of presentation will differ from one language to another according to the internal structure of the language. The presentation is limited to the main features of the macrostructure.

The macrostructure is known as the collection of lemmata included as part of the central list. It is, therefore, the selection of lexical items to be included in the dictionary as lemma signs. They become the primary treatment units of the lexicographic process, as stated by Gouws and Prinsloo (2005).

In Chapter 4, the researcher shows the representativeness of the macrostructure of an envisaged dictionary as a list of words drawn from a corpus by means of WordSmith Tools; it could also be called a collection of words. All the lexical items of a given language come from the lexicon of that language. The lexical items can be words, elements larger than words and some other elements smaller than words, in other words multiword and subword lexical items; they are all constituents of the lexicon.

In African languages, a single phoneme or morpheme such as /u/ or /be/ can play an important role in the whole word or phrase by being a suffix or a prefix of another word. In Yipunu, the conjugated form of the verb *uwenda* (to go) is *amawenda* (he has gone, he had gone in different contexts). The two different forms *uwenda* and *bewendi* in Yipunu respectively mean you will go and they go. These different prefixes taken out of context will definitely not appear independently; therefore, they will always be part of another word. The same can be said about multiword lexical items such as particle verbs. All these differences can be observed in many languages. It is clear that the corpus representing the lexicon or the vocabulary of a language will help the lexicographer in the identification of the different lexical items. This is why lexicographers are urged to make use of different texts for the compilation of lexicographic corpora.

According to Gouws (1991), lexicography in the past has often been dominated by a word bias. This situation has led to the compilation of dictionaries to focus on the presentation and treatment of words instead of lexical items. The development in metalexicography has changed this situation and today the macrostructure of dictionaries endeavour to reflect the lexicon and not only the words of a given language. Thus, different types of lexical items are included as different types of lemmata, words as lexical lemmata, subword lexical items as sublexical lemmata and multiword lexical items as mutilexical lemmata. Yet again, the corpus has to be representative enough to allow the distinction of these different categories, and the disjunctively written languages are the ones that can benefit because this system displays a clear distinction between the lexical items contained in the corpus. This brings out the question of how they should be included in the dictionary.

There are many ways of entering lemmas in a dictionary. A significant difference exists between the alphabetical ordering of lemmata, displayed by the majority of dictionaries, and the thematic approach, which apparently entails a nonalphabetical ordering of the macrostructure. In the last category can be included a thesaurus and in the first category standard, comprehensive and learner's dictionaries as typical examples. In an alphabetical macrostructure, ordering does not mean that all lemmata should keep a strict alphabetical arrangement. For practical reasons deviation from a strict alphabetical ordering is possible. In this regard, in both monolingual and translation dictionaries different types of macrostructural ordering can be found.



The lexicographer is then forced to give a clear explanation of the macrostructural ordering of a specific dictionary in the front matter text. Within a strict alphabetical approach, a straight alphabetical ordering of the lemmata displays a vertical macrostructural arrangement. A discussion of these macrostructures can also be found in Gouws and Prinsloo (2005) and Hausmann and Wiegand (1989:336–337).

Another distinctive feature of a strict alphabetical ordering could be what is known as a sinuous lemma file, presenting a system of niching and nesting. The researcher will not extensively discuss sinuous lemma files; the focus will be on the theoretical distinction of the two concepts nesting and niching. A sinuous lemma file contains clusters of lemmata that display a horizontal ordering and are attached to the article of a vertically ordered main lemma. A niche is formed through a process of niching and is characterised by an alphabetical clustering of the lemmata, which may or may not be semantically related. A nest is formed through a process of nesting and is characterised by a clustering that stretches the rules of alphabetical ordering in order to exhibit morphosemantic relations between words (cf. Hausmann & Wiegand 1989; Wolski 1989, as referred to by Gouws & Prinsloo 2005).

Malkiel (1962) distinguishes three contrasting patterns of arrangement. For him, the basic arrangement of dictionary items may be conventional (alphabetical), semantic (ordering by part of speech or provinces of life) or entirely arbitrary (chaotic).

One has to argue that the corpus displays different data types needed for better presentation of the different lemmata in the dictionary. In both the horizontal and vertical ordering the lexicographer will use data contained in the corpus in order to better achieve the objectives and functions of the dictionary in the speech community. This is possible if the corpus is able to display an alphabetical ordering that brings out all the different lexical items that the lexicographer will have the task of grouping into their semantic relations.

A new approach or terminology can then come to the fore in the sense that both the thematic and strict alphabetical approaches can be introduced into what is called here syntagmatic and paradigmatic ordering. Paradigmatic ordering should be understood as the way in which lemmata are presented vertically and syntagmatic ordering as the arrangement of lemmata in a horizontal way. These two concepts can also be applied in thematic presentation of the macrostructure. In fact, the researcher believes that in

dictionaries that display a thematic presentation, the alphabetic ordering of themes can be followed. The same alphabetic ordering can be applied inside given treated paradigmatic themes, in other words lemmata presented under the main theme in a vertical order.

In order to achieve optimal presentation of the different lexical items in the central list, alphabetically ordered or thematically ordered, lexicographers in the Gabonese languages have to make use of corpora. These corpora will display the necessary potential lemmas that one will have to present separately in an alphabetical dictionary or group in themes or topics in a thematic dictionary. For the organisation of any macrostructural presentation in any dictionary, the corpus remains an indispensable and reliable source that will help in the retrieval of the items from which the lexicographer will have to select the lemma candidates for any dictionary. Regarding the alphabetical presentation of the central list, in Chapter 4 a frequency list of 200 items is presented alphabetically.

Such frequency lists could be used as a basis for the selection of a lemma candidate list for a given dictionary in Yipunu. When the frequency count of a corpus displays a limited number of lexical items, these items can only be used for the compilation of a restricted dictionary. When the corpus displays a large number of lexical items, comprehensive or bigger dictionaries can be envisaged. In either case, the corpus will assist the lexicographer in the selection of items for the specific type of dictionary, even in a situation where different types of dictionary have to be combined, for example a learner's dictionary and a dictionary for special purposes. In these two specific types, the lexicographer could use some items suitable for inclusion in one type in the other type and vice versa and the corpus will have to assist in the selection of these items.

### **5.2.3. Different types of microstructure**

It must be clearly indicated that the corpus does not play a role in choosing a specific type of microstructure to be used in a dictionary. Rather, the importance of the corpus is more obvious in the microstructural treatment of the different lexical items. Each dictionary project will use a corpus in order to be assisted in a specific microstructural treatment of a specific dictionary type. This is why lexicographers of the Gabonese

languages are urged to focus primarily on the compilation of comprehensive and balanced corpora in order to make available language material essential to the formulation of microstructural programmes. In the same way, the available material will help lexicographers in the selection of data necessary for specific microstructures chosen for optimal treatment of lexical items in the multilingual environment of Gabon. This will also assist lexicographers to achieve the goals set for the compilation of dictionaries for the particular speech communities of Gabon, all of which have their own needs.

In the previous chapters, some important points on microstructure have already been dealt with. This has been done in relation with the presentation of data types in terms of the comment on semantics and the comment on form. The researcher will focus the presentation on the integrated and nonintegrated microstructures in Wiegand (1996) and the obligatory microstructures and extended obligatory microstructures in Gouws (1999; 2003), as recommended by Gouws and Prinsloo (2005).

The above-mentioned lexicographers are in favour of the fact that the distinction between integrated and nonintegrated microstructures is made on the grounds of the proximity and the directness of the relation between each entry representing a paraphrase of meaning in a monolingual dictionary or each entry representing a translation equivalent in a bilingual dictionary and the relevant cotext entries. The difference can also be distinguished with regard to the supporting cotext entries representing illustrative examples in the specific article. The type of microstructure will be determined according to the relative positioning (the proximity) between the cotext entries and the respective core entry of each subcomment on semantics, represented by the paraphrase of meaning/translation equivalent given for each polysemous sense of the lemma sign.

The close relation between a cotext and context entry and the relevant paraphrase of meaning/translation equivalent is shown by an integrated microstructure. It should also be mentioned here that when a dictionary article displays an integrated microstructure the integrate that accommodates the translation equivalents/paraphrases of meaning also contains the relevant cotext entries. In contrast, a nonintegrated microstructure does not have cotext or context entries in the same integrate as the respective translation equivalents/paraphrases of meaning to which they are addressed.

It follows that as far as the nonintegrated microstructure is concerned, some dictionaries would include all the different subcomments on semantics in a text block and follow this with a separate text block that contains cotext entries. In spite of the fact that the cotext entry and translation equivalent/paraphrase of meaning are not presented in the same integrate, the presentation and ordering of the cotext entries in the text block are done in such a way that the relation between a given cotext entry and the item at which it is addressed is clearly marked (cf. Gouws & Prinsloo 2005). A user who is familiar with the use of the nonintegrated microstructure would easily match up the cotext entries with their corresponding addresses in the dictionary articles.

According to these two researchers, the dictionary-specific lexicographic process of each project has to instruct lexicographers with regard to the type of microstructure to be employed in the dictionary. The decision should not be taken in an arbitrary way or be isolated from other decisions regarding the specific dictionary. After the determination of the type of microstructure to be used in a specific dictionary, the lexicographer's next task is to select from the corpus data to be used in the chosen microstructure, in other words the different data categories to be used for the presentation and treatment of lemmas in obligatory and extended microstructures. When treating a lexical item, all the data used in the microstructure will allow the lexicographer to present the article in a specific way in accordance with the different data provided. This takes one to the identification of the difference between obligatory microstructures and extended obligatory microstructures. These two types of microstructure have direct links with the data distribution structure and the extent of the data categories to be included in the articles.

Consequently, the obligatory microstructure refers to the microstructural items that will be found in every article of the dictionary. The extended obligatory microstructure is more or less the presentation of all the data categories needed to ensure an adequate treatment of the lexical items represented by the lemma sign in terms of the minimum requirements of the specific dictionary. The extended obligatory microstructure contains data used in the obligatory microstructure and additional data. In other words, the microstructural data in the article of a lemma sign representing a preposition will differ from the microstructural data in the article of a

lemma sign representing a noun, and that is applicable to all the other types of lemma, as suggested by Gouws and Prinsloo (2005).

If one uses a corpus as evidence for the treatment of lexical items on the microstructural level, one will definitely eliminate inconsistencies, ambiguity and even unclear definitions in dictionaries. In this regard, the dictionary can better play its role of assisting the users in their text production and text reception. This can optimally be achieved if both the selected lemmata and their different treatment in dictionaries are done in a way that render clearly either the meaning or the different senses a given lexical item can cover in one context or the other. A lexicographer's never-ending battle is to try to reach perfection in the product that he or she is producing. In the microstructural treatment of lexical units, the lexicographer gives a paraphrase of the meaning or translation equivalents of the lexical items or the lemmata extracted from a corpus. The corpus will help one to choose the best presentation and strategies in the microstructural treatment of the lemmata in any dictionary. This is applicable to both the obligatory and the extended obligatory microstructures.

### **5.3. Lexicographic functions and corpora**

The last two decades have shown a remarkable development in the theory of dictionary functions. From Tarp and Bergenholtz to Wiegand, lexicographic functions have played a big role in dictionary compilation. These authors all believe that a dictionary, to be able to serve the needs of the community it is compiled for, must be capable of supplying users with the information they are looking for. In a language, the user will perform well if he or she is able to produce and receive text needed for a specific purpose. That is why dictionaries should be compiled in a way that would not only enable them to perform knowledge- and communication-oriented functions but that would also take into consideration the different target users and their reference skills.

The lexicographer has to bear these functions in mind when compiling a dictionary for the specific needs of a specific community. In certain communication situations, the speaker of a given language will be judged by her or his ability to produce and perform in the language she or he is using to express her- or himself. This applies to

both speaking and writing, when the user consults the dictionary for a specific translation equivalent, for instance.

In this regard, some lexicographers have written extensively on the development of the theory of lexicographic functions, amongst them Bergenholtz and Tarp (2002), Tarp (1994; 2000; 2002a; 2002b; 2004), Tarp and Gouws (2004) and Wiegand (1996a; 1996b; 2001), as referred to by Gouws and Prinsloo (2005). Despite their differences in the approach to the problem, they all encourage the compilation of dictionaries that are based on lexicographic functions.

These functions are observable in monolingual and bilingual dictionaries when users are looking for the meaning of a lexical item or for a translation of items, sentences and text from a foreign language to the local language or vice versa. Researchers in that theory of lexicographic functions should favour the idea that the dictionary must be planned in consideration of the fact that different users might consult the dictionary for different reasons at different moments.

In accordance with this, Tarp (2002) states that lexicographic functions represent the assistance that a dictionary provides to a particular type of user to fulfil the needs of that user in a specific user situation. Cognitive functions, according to Bergenholtz and Tarp (2002), assist the user by providing the following:

- general cultural and encyclopaedic data
- specific data about the subject field
- data about the language

Communication-oriented functions will assist the user who needs to achieve the following:

- text production in the native language
- text production in the foreign language
- text reception in the native language
- text reception in the foreign language
- translation of texts from the foreign language to the native language
- translation of texts from the native language to the foreign language

All these functions can be performed if the corpus used by the lexicographer contains sufficient material to satisfy the needs of the users and the aim of the dictionary. A corpus that contains a lexical item *marathon* will have to contain data defining the lemma, for example *the marathon is a long-distance race*, for text reception. The same corpus also must contain data specifying the distance of that marathon, for example *the marathon covers 42.195 kilometres*, for achieving a cognitive function.

The bottom line here is that the dictionary cannot achieve the above-mentioned objectives without using a well-balanced and representative corpus. Both spoken and written languages play an important role in helping the lexicographer to fulfil the assigned mission of the dictionary because as rightfully mentioned by Sinclair (2003) with regard to the transposability of the corpus, it is important to remember the ease with which the mode of text can be changed:

- Any written text can be read out.
- Any spoken text can be written down.
- Any written text can be put into electronic form.
- Any spoken text can be put into electronic form.
- Any electronic text can be printed out.
- Any electronic text can be read out.

All these different data types in the dictionary will help the user in speech situations and in the writing of any kind of text in a given language, being it a native language or a foreign language. Both written and spoken material contained in the corpus can be used in order to integrate the above circumstances presented above.

The representativeness and balance of the corpus are two important features that qualify the corpus as a container of data from different conversational situations and text types that the lexicographer will use in the presentation and the treatment of the items included in the dictionary. Both the items presented in the macrostructure and their treatment in the microstructure of dictionaries play an important role in the achievement of the functions of dictionaries. If lexicographers in Gabon are to compile functional dictionaries that would really help communities in solving their different lexicographic problems, dictionaries should be compiled with the help of well-designed corpora. These corpora should effectively assist the lexicographer in

comprehensive treatment of the lemmas selected for inclusion. In this regard, the different consultation processes would be meaningful and dictionaries compiled for the languages in Gabon could fully play their role and perform their function and in the end satisfy the target users.

#### **5.4. Corpus for monolingual dictionaries**

The term *monolingual* will be used to qualify all those dictionaries that are compiled in a single language. That means that the microstructural treatment of lemmata is done in the same language as the language of the lemma sign. This will include encyclopaedic dictionaries even if they are different from linguistic dictionaries because they concentrate on the extra-linguistic treatment of words. The dichotomy between synchronic and diachronic dictionaries will also be taken into consideration. These two types can also be divided into subcategories that will be mentioned later since these paragraphs concentrate on the ways in which a corpus can assist the lexicographer in the macrostructural selection and the microstructural treatment of lexical items in monolingual dictionaries in general.

##### **5.4.1. Encyclopaedic dictionaries**

It is generally admitted by lexicographers that encyclopaedic dictionaries differ from linguistic dictionaries. The first type concentrates on the description of extra-linguistic aspects of words and the second gives the linguistic aspect of words. The objective here is not to elaborate on the distinction between these two types but rather to show how the general corpus can contain data to be used in the macrostructure and the microstructure of encyclopaedic dictionaries. By doing so, the distinctive presentation among all the types of dictionary could help one to better understand the treatment of the macro- and microstructure of each dictionary type. All the characteristics and principles of encyclopaedic dictionaries also apply to encyclopaedias. These two terms are sometimes used alternatively.

###### ***5.4.1.1. The macrostructure of encyclopaedic dictionaries***

The fact that the corpus is intended to be multifunctional, representative and balanced is an asset that helps the lexicographer to have at his or her disposal material from



every branch of knowledge that can be contained in encyclopaedic dictionaries. The lemmas of the macrostructure of encyclopaedic dictionaries also display an alphabetical ordering. The macrostructure is mainly composed of nouns, both common nouns and proper nouns; it includes lemmas referring to things and cultural features.

According to Zgusta (1971), encyclopaedic dictionaries are arranged in order of the words or lexical items by which the segments of this extralinguistic world are referred to when spoken about. This could imply that all the lexical items included in any dictionary could be treated in the encyclopaedic dictionary provided that the treatment primarily be directed at nonlinguistic facts. As a result, the macrostructure has to include different lexical items and the corpus will help in the selection of lemma candidates in the same way it does for a general dictionary. However, encyclopaedic data cannot always be retrieved from a general text; the lexicographer has to compile a special corpus in this regard.

The lemma list extracted from the SYC contains lexical items that can be candidates for inclusion not only in general linguistic dictionaries but also in an encyclopaedic dictionary if one decides to compile such a dictionary in Yipunu. Consider the items in Table 5.1 as potential lemma candidates that could be included in the macrostructure of a Yipunu encyclopaedic dictionary. All these lexical items have been randomly selected just for the purpose of illustration. For their inclusion in the dictionary as such, the choice should be guided not only by the frequency counts presented by the corpus but also by the function(s) of the dictionary, the dictionary type and the needs and reference skills of the target users of the dictionary. This is applicable to all types of dictionary. Each one of these items can rightfully constitute the main or the subtheme in the macrostructure by representing a specific domain or area of knowledge. In the case of the items presented below, the difference between their occurrences in a linguistic or an encyclopaedic dictionary is determined by the microstructural treatment. In fact, the macrostructure of an encyclopaedic dictionary should present more lemmas that deal with, for example, country names, place names and proper names.

**Table 5.1:** Ten restricted alphabetical lemma lists from the SYC

Lexical units	Frequencies
Bategula (grandsons)	357
Malongu (countries)	459
Marundu	24
Migaga (officials/court)	375
Murima (heart)	376
Ndagu (house)	1 478
Ubueji (beauty)	369
Uloba ( to fish)	16
Uramba (make a trap)	27
Yibandu (clan group)	331

The treatment of some of the items presented in Table 5.1 will be beneficial for the cultural perspectives of Yipunu-speaking and non-Yipunu-speaking people. It will enable users to capture some of the cultural realities of Yipunu society if the lexicographer focuses on the accommodation of cultural data in the dictionaries. It will then enhance the knowledge of the users. As the items are presented, they can educate the different users about specialised vocabulary in a specific domain such as family tree, proper names, housing, fishing, and so on. The use of a thematic presentation of these items in the macrostructure could display in the main theme a specific cultural reality that the subthemes will have to represent in the different categories. If a given lexical item is used as a theme, it will be treated first and the rest of the identified items belonging to the same theme will be treated under the main one. In this regard, both the main and the subthemes could display a strict alphabetic ordering of the macrostructure. Thus, the lexical items will follow a main thematic alphabetical ordering and a subthematic alphabetical ordering.

The lemmas *Bategula*, *Malongu*, *Marundu*, *Uloba*, *Uramba* and *Yibandu* can be treated in a Yipunu encyclopaedic dictionary because they can convey data about specific activities and the lifestyle of the Bapunu people. The inclusion of these items in this section aims to show how items that can be candidates for a linguistic dictionary are treated differently in an encyclopaedic dictionary.

If the lexicographer chooses the lexical item *Yibandu* as an element to be treated and presented as a main theme, he or she will have to make sure that the different clan groups are also treated and presented as subthemes. Mupunu-speaking and non-

Mupunu-speaking people will learn that *Ndinga*, *Basumba*, *Dijaba*, *Bupeti*, *Dikanda* and *Jungu* are some of the clan groups of the Bapunu people.

Additionally, the compilation of an encyclopaedic dictionary in Yipunu could include the following typical lexical items in the macrostructure:

- Names of cities: *Mwila*, *Masanga*, *Moabi*, *Pungu*.

Users will learn about the different cities of Gabon and the treatment of each city in the macrostructure should highlight the features of each city.

- Birds: *muvhiji*, *dibembi*, *mukongusuli*, *mambi*, *ngwali*
- Fishing: *mbusa*, *mbaga*, *uvhusa*
- Farming: *mumbu*, *musaga*, *nungi*

The lexicographer has to explicitly describe the different ways of farming and fishing and the different birds.

The lexicographer will have to design criteria according to which certain lexical items will only be selected for inclusion in an encyclopaedic dictionary but not any other dictionary. These criteria should provide for the exclusion of linguistic items such as verbs, nouns, adjectives, adverbs, pronouns, and so on because they are already the focus of linguistic dictionaries. The criteria will rather have to make provision for the selection of items that are likely to be treated in encyclopaedic dictionaries and items that can easily be grouped into themes and topics according to the context of their applications, such as the ones presented above. It is not always easy to draw a strict line between the items that should be selected for inclusion in encyclopaedic and linguistic dictionaries, respectively; some of the items will be treated in both types of dictionary. The major distinction, however, will be on the microstructural level, in other words the way in which the lexical items will be treated in the dictionary articles.

#### **5.4.1.2. *The microstructure of encyclopaedic dictionaries***

The microstructural treatment of encyclopaedic dictionaries provides definitions and descriptions that go beyond the data given in linguistic dictionaries. The articles of encyclopaedic dictionaries are essentially topical, dealing with the entire subject represented by the title of the article (cf. Landau 1984).

The corpus should contain data useful for the microstructure in such a way that if one deals with an article on religion, for instance, one will focus on the systematic description of the religions of the world: their histories, doctrines and practices. To achieve this, the lexicographer has to be helped by the corpus as it contains material of a general and restricted nature. This will help in explaining the origin, the development and other distinctive features of each religion. The lexicographer will have the final say in selecting relevant data for the treatment and the compilation of the necessary data in the dictionary article. The amount and type of data will depend on the material available to her or him. The corpus will help the lexicographer to interpret and define the word in a specific field it applies to from the use of the term in a specific sense in the corpus.

As an illustration, the treatment of the lemma sign *garden* will give the history and geography of gardening, an explanation of its techniques. Typical microstructural treatments can be found in the following sample items. The material presented below can assist the lexicographer in the compilation of the microstructural entries but the quality will have to be improved in order to be used successfully. This given data provide an indication of what can be done when dealing with encyclopaedic dictionaries in which a more detailed microstructural treatment must be presented for each lemma included in the macrostructure. The lexicographer has to pay more attention to the definition of the items as this is a very important data entry in encyclopaedic dictionaries.

All definientia in dictionaries must be based on actual language usage. The definition consists of a definiendum and a definiens. The lemma acts as the definiendum and the explanation of the meaning as the definiens. Therefore, definitions can be compiled from evidence presented in the corpus if the lexicographer has a complete and reliable database. The definition includes both the lemma and the explanation of the meaning. The corpus will provide the lexicographer with at least the basic material that can help

him or her to compile the definitions of the items in dictionary articles. In other words, in certain circumstances the lexicographer will have to make some choices in defining the items, based on his or her linguistic intuition and his or her knowledge of the language. In some cases, the lexicographer can, of course, use his or her linguistic intuition to compile a dictionary when there is a lack of necessary data. The ability of the lexicographer to combine both the existing material and his or her expertise will determine the quality of the definition in the dictionary.

The lexicographer has to make realistic decisions on both the strategy of presentation of the lemmas and the type of definition to be used in the microstructure of the dictionary for the treatment of the different lemmas. These decisions have to be taken according to the lemma the lexicographer is dealing with because both the included data and the definition of specific lexical items will depend on the availability of the language material. This decision making must be done in accordance with the cognitive function that the encyclopaedic dictionary is intended to serve. The completeness and the correctness of the definition in a dictionary will always be ensured by the use of a corpus that presents all the contexts in which a given item is used. The corpus has the illustrative role of providing the lexical item with an accurate example that allows capturing the environment and the context in which the item is used.

It is important to note that the encyclopaedic data presented in the following lexical items are not sufficient to fully treat the lemmas.

### *Malongu*

*Mambura ama batu betsanini, mavhu na migagu, mabanombi, bashinu, baraba. Mu malongu motsu ma butamba batu baminongu na minongu myotsu borungusa utsana vhana batsirondila. Malongu mebi na mavhula ma pwela na makeki na maneni. (The places inhabited by men are governed by laws, among Blacks, Asians, and Arabs. Each country has a constitution, a map, a flag and a national anthem. In all the countries, people of different races can live together. In one country you can find different provinces and towns.)*

### *Marundu*

*Dina di mavhasa bedivheyi mwana mugetu, marundu divhasa adirerili upala mba divhasa adi mubeji dina diandi mbumba. Marundu mwana tata nsambi mba mwana ngwandi diyevharakesa, abe jabi dusavhu du mumbwanga. (Marundu is a cultural name that is given to one of twin girls. This name is for the first-born twin and the second one is called Mbumba. Marundu is the daughter of father Nsambi and her brother is Diyevharakesa; it will be familiar to those who know Mumbwanga.)*

### *Ndagu*

*Mbura aji batu betsanini, vhaji minongu na minongu mimandagu. Vhaji ndagu milomba, ndagu mabaji, ndagu butamba na mandagu mabirika bicutamba na bisima. Mu urunga ndagu besyebi mbamba na urunga kunsu mukubiga o julu. Mu bibaga urungula ukibiga na makuku, mamosi na tsimbi mba na butamba bu badiga. (The place inhabited by men are governed by laws. You can find many houses in one place; there are many types of houses: houses made with planks, soil houses and brick houses. The way those houses are covered is different to one another; the walls can be covered with bark or planks.)*

All these items can also be treated in linguistic dictionaries and the articles could look different, considering the inclusion of labels and other illustrations. Their treatment gives encyclopaedic data for each lemma. The first topic provides data about the types of country and the types of people one can find in a country as well as racial distinction. The distinction is made on the basis of European, African, Asian and Arabic countries. The second topic gives an account of the cultural environment of the the proper name *Marundu* that is given to one of the twin girls and usually to the one who is born first. Finally, the third topic is about the different types of house in the African tradition. It describes the way those houses are built. The treatment of these items gives their formal features, such as spelling, and can also present the pronunciation, inflexion and so on. All these features should be shown in a dictionary.

Including names in dictionaries is important when the name has become part of the lexicon of the language, meaning that it has changed from being a proper noun, a word denoting a unique phenomenon, to being an appellative or common noun, a

word denoting one or several phenomena of the same kind. The lexicographer should take into account the transferred meaning taken by a name and the associations that the name evokes, in other words its connotations. A proper name can also be important when it can form derivatives, since any connotations the name may have often allude to a certain reality. When dealing with proper names the lexicographer should take into account all the different forms that relate to the name and the different meanings they display.

The three items used as illustration above convey cultural knowledge of the Yipunu people from the way denomination is done in a specific event, for example when naming twins, to the kinds of house existing in this society as well as the traditional material used to build those houses. Data such as these describe the style of living and the behaviour of that particular society.

The lexicographer has to decide on the way in which the treatment of cultural data will best suit the users' requirements. According to Gouws (2002), in the central list the lexicographer should distinguish between culturally bound lexical items and lexical items with one or more culturally bound uses or senses. This should be clearly indicated and is also of extreme importance for cross-referencing purposes to guide a user from an outer text entry to the relevant venue in the central list. Gouws (2002) also indicates that the inclusion and the treatment of cultural data are relevant in both monolingual and bilingual dictionaries. In this regard, the researcher believes that the strategies of treatment should be common to both these two types to a certain extent. In the dictionary type concerned in this paragraph, the selection, inclusion and treatment of this type of lexicographical data need a much more comprehensive account, as argued by Gouws.

An encyclopaedic dictionary seems to be the best place to treat cultural data because of the comprehensive coverage that a dictionary article can provide, and this depends on the typical cultural spectrum of each speech community. The article stretch will accommodate a huge amount of data for the microstructural treatment of such items. This is why Gouws (2002) remarks that the central list treatment of cultural data will primarily be of encyclopaedic and pragmatic nature and could be accommodated within the comment on semantics. The lexicographer will have to make the right decisions (before commencing the compilation of the dictionary) on the way this data will be treated in the dictionary in order to satisfy the users' needs. The use of

pictorial illustrations will, in a certain way, help in the transfer of meaning and the description of the cultural event the dictionary article intends to present.

In this regard, the researcher believes that the lexicographer should go beyond word corpora to picture corpora as both can be managed alternatively. The term *corpus* should not be reduced to the selection of lexical items but should be extended to the selection of pictures. To be able to choose the right picture to illustrate any lexical item, the lexicographer needs to have at his or her disposal as many pictures as possible about a given event, concept or lexical item in order to pick the best. This will help to put an end to the use of pictures that are not expressive enough in dictionaries. A good picture enables a good understanding of the definition of a word because sometimes a picture is more expressive than a verbal explanation. If one intends to use pictorial illustrations in a dictionary, one should in advance select the words that need pictures and select the best ones from the pictorial corpus that one has already gathered in advance or for the specific situation. Pictorial corpora can be extended and improved from time to time in order to update the corpus. This will help to improve the quality of pictorial illustrations in dictionaries. A good picture leads to a good semantic interpretation of the word to capture the message behind the image.

#### **5.4.2. Diachronic dictionaries**

Diachronic dictionaries deal with the history of the changes in the meaning and form of lexical items. Thus, they describe the way in which the language has developed over a long period of time, even centuries; they also focus on the origin of lexical units and their prehistory. Two subtypes can be identified, namely historical and etymological dictionaries. This is why etymology and historical data are very important in this kind of dictionary.

##### ***5.4.2.1. The macrostructure of diachronic dictionaries***

The corpus for a diachronic dictionary needs to include a wide-ranging selection of texts that represent the period covered by the diachronic approach. In this regard, the corpus should contain material representing the specific time period in the development of the language. To be able to convey any information about the history



and the etymology of the different lexical items in dictionaries, the lexicographer should make use of a corpus that covers both a selective vocabulary and a vocabulary of a broader nature. The corpus should be gathered from a restricted period of time and cover a time span from a particular period to the present. A corpus of this kind will help partially or totally to retrace the history of the changes that occurred in the course of the development of the language.

Diachronic dictionaries display an alphabetical arrangement of lexical units in the macrostructure and they include a more general vocabulary. Historical linguistics has provided some basics on language families and their origins. In Europe one talks about Indo-European and in Africa about Proto-Bantu. The Proto-Bantu forms will be determined by the lexicographer or linguist who is familiar with the historical development of language forms. This is one of the reasons why dictionary compilation is a team work that involves staff of different backgrounds and specialities, and in this regard the linguist and the lexicographer have to join hands.

Proto-Bantu in Africa has given birth to the Bantu languages of which the present lexical items can be linked to their Proto-Bantu counterparts. Consider these 20 (randomly chosen) lexical items from Proto-Bantu, reconstructed by Guthrie. However, one has to be careful in treating Proto-Bantu as a historic language because Proto-Bantu is a synchronic abstraction of presumed forms. Unlike languages such as English and French there are no historic texts for the Bantu languages.

### **Proto-Bantu**

*-peep-	(blow)
*-pot-	(twist)
*-paagu	(junction of two branches)
*-dut-	(tear)
*-taata	(father)
*-tatu	(trios)
*-kata	(balls)
*-gego	(molar, tooth)

*-kot-	(enter)
*-koko	(chicken)
*-keedo	(morning)
*-patid-	(tighten)
*-pika	(slave)
*-kida	(tail)
*-pidi	(viper)
*-dagu	(house)
*-cende	(thorn)
*-jino	(tooth)
*-joga	(mushroom)
*-jungu	(pot)

All these Proto-Bantu reconstructed items can be part of the macrostructure of the diachronic dictionary of Yipunu. The presentation of these items can be alphabetical and they will not contain the prefix because the reconstructions are given in these standard forms of the Proto-Bantu. This, once again, brings out the problem of lemmatisation strategies that the lexicographer has to take into consideration when dealing with Bantu languages. Each African language can retrieve the current form of each word in the language as used today. The use of the star shows the historicity of the forms and their appurtenance to the Proto-Bantu forms. The corpus does not make a distinction between the Proto-Bantu forms and the forms used in the actual language. The identification will have to be made in the dictionary. A transcribed list of all the Proto-Bantu forms can be included in the corpus and labelled as such by the lexicographer so that the corpus can help with the identification of these forms. Otherwise, if the corpus contains only the forms of a more recent period, the lexicographer's expertise will be the only reliable source that can help to relate the actual language forms with the Proto-Bantu ones.

Two different ways of introducing the above-presented lexical items can be implemented. The first option could be to use them as main lemmata in the macrostructure of the dictionary as they are to show the difference between the diachronic form, which will be presented as above, and the synchronic form, as the ones contained in the Yipunu forms given in the next subsection. The second option will be to introduce them as sublemmata in the dictionary article of the present-day variant form.

African languages in general pose problems when it comes to the lemmatisation of the lexical items of each language. After the identification of the lexical categories of all the words, it is important for the lexicographer to determine the strategy to follow. The decision should be guided by identification of the state of the dictionary culture and accommodation of the user-friendliness principle of a good dictionary.

A decision needs to be taken on the lemmatisation of either singular or plural forms for lexical items displaying the two categories. Ambivalence can also be noted when deciding between lemmatisation of nouns on the first or the third letter and the stem lemmatisation strategy. The success of the chosen strategy will reside in the capability of the lexicographer to present users with a product that will fulfil their needs. The choice of writing a language disjunctively or conjunctively will also play a role in the problem-solving strategies. What is said here is applicable to all dictionary types with slight differences based on the specificity of each type. The fact that most of the reconstructed forms of Proto-Bantu seem to be represented by the stem or the root implies that the lexicographer will have to take important decisions in relation with how they are going to be introduced in the macrostructure of the dictionary.

#### ***5.4.2.2. The microstructure of diachronic dictionaries***

As already mentioned, the microstructure of diachronic dictionaries indicates the historical transformation of the words both on semantic and morphological levels. In the case of the microstructural treatment of the above-presented lexical units, the morphological changes of words are given and the semantic transformation of some of the items is referred to when possible while quite the majority of them remain with the original senses. In the following list the Proto-Bantu items are presented on one side and the Yipunu on the other.

The corpus will present data that will help the lexicographer in decision making on the microstructural level. This justifies the importance of including texts from different times in corpora for diachronic dictionaries as they impact not only on the ordering of the meaning but also on how the microstructure treats the way in which some lexical items have changed over time, in other words how their plural and singular forms have changed. The corpus can help to identify where the changes occurred in a language at different levels. For the purpose of this presentation, the focus will be on polysemy and homonymy as these semantic phenomena hold implications for the microstructural arrangement of semantic values.

<b>Proto-Bantu</b>		<b>Yipunu</b>
*-peep-	(blow)	<i>upepa</i>
*-pot-	(twist)	<i>upota</i>
*-paagu	(junction of two branches)	<i>dipaku</i>
*-dut-	(tear)	<i>uduta</i>
*-taata	(father)	<i>tata</i>
*-tatu	(trios)	<i>bitatu</i>
*-kaaka	(grand parent)	<i>kaga</i>
*-gego	(molar, tooth)	<i>dikeku</i>
*-kot-	(enter)	<i>ukota</i>
*-koko	(chicken)	<i>koku</i>
*-keedo	(morning)	<i>keedi</i>
*-patid-	(tighten)	<i>uvharila</i> (exaggerate)
*-pika	(slave)	<i>muvhiga</i>
*-kida	(tail)	<i>mugila</i>
*-pidi	(viper)	<i>pili</i>
*-dagu	(house)	<i>ndagu</i>
*-cende	(thorn)	<i>dusyendi</i>

*-jino	(tooth)	<i>dinu</i>
*-joga	(mushroom)	<i>bogu</i>
*-jungu	(pot)	dwengu

If the corpus presents polysemic lexical items, all the senses covered by the item should be contextualised and clearly expressed. This will allow the presentation of the different distinctions in meaning in a dictionary article for a polysemic lemma in separate entries. These distinctions in meaning are not arranged arbitrarily but are arranged according to fixed criteria. These arrangements should be done in accordance with the dictionary types and the needs of the potential users. The preferences of lexicographers also play a decisive role when choosing a specific arrangement. One of the main features of a diachronic dictionary is its chronological arrangement of distinctions in meaning.

As already mentioned, this type of dictionary reflects the changes in form and meaning that a particular lexical item undergoes over a given period of time. When treating an item, the different polysemic values should follow one another chronologically and the first recording of a polysemic value is provided by a date. This is possible as far as dated written or spoken language data are available in order to regard the principle of historical arrangement as objective and reliable. All this, once again, should be helped by the use of adequate language material provided by the corpus. The chronological arrangement simplifies the lexicographer's task considerably. In these types of dictionary, the example sentences supporting each distinction in meaning should also be arranged chronologically. Thus in a diachronic approach, the oldest polysemic value is regarded as the primary value and placed first.

As far as homonyms are concerned, they make provision for different, unrelated meanings being lexicalised as formally identical lexical items. In keeping with lexicographical conventions, homonyms are dealt with as separate lemmas. They are not single lexical items with different meanings but are different lexical items that coincidentally have the same form. Homonyms differ from one another etymologically in the sense that one homonym can belong to an old variety of a language while the other may be a word borrowed from another language.

Even if the treatment is limited, one can note that from the Proto-Bantu forms to the Yipunu ones, the changes regarding some phonemes are remarkable. The different rules that can be applied to those changes are not going to be discussed but at least some phonemic changes that occurred in the transformation process can be shown. It must also be mentioned that transformation can occur in the vowels and in the tone. These changes can be observed as follows:

**D** becomes **d** or **l**

**P** becomes **p** or **vh**

**T** becomes **t** or **r**

**K** becomes **k** or **g**

All these changes or reflexes are noticeable in all Bantu languages and the difference will rely on the specificity of each language. If one decides on the compilation of dictionaries of a diachronic nature, one should look at these issues meticulously to enable a more comprehensive analysis of the observations. Once again, the corpus cannot assist in the distinction of these reflexes; the lexicographer should be aware of all the changes in order to be able to analyse the situation correctly.

#### **5.4.3. Synchronic dictionaries**

Compared to diachronic dictionaries, synchronic dictionaries are restricted to a much narrower range of time and attempt to set out the lexicon or a part of it as it appeared at a specific moment. Most synchronic dictionaries concentrate on the language as it is being used at the time of the compilation of the dictionary concerned. The objective of these dictionaries is to describe the vocabulary of a particular language at a particular stage or period in the development of that language. The corpus should then contain texts from a given period in order to fulfil the requirements of a synchronic dictionary. The corpus should be of a more general nature by containing a representative and balanced selection of material of the chosen period.

Synchronic dictionaries can be divided into two major subtypes, namely (1) monolingual descriptive dictionaries, which include comprehensive descriptive, standard descriptive and pedagogically descriptive dictionaries, within which learner's and school dictionaries can be identified, and (2) translation dictionaries. The last category will be discussed in the coming paragraphs. The focus in this section will not be on a specific subtype of synchronic dictionaries but it will deal with some general characteristics of the whole category.

Another major distinction in synchronic dictionaries is between LGP and LSP dictionaries.

#### ***5.4.3.1. The macrostructure of synchronic dictionaries***

Synchronic dictionaries usually do not include archaic and obsolete words in their macrostructures, but if one decides to compile a comprehensive dictionary that is a multivolume and multidecade project, all these words should be taken into account. These dictionaries can target the entire vocabulary or a specialised subsection of the vocabulary to be included in the macrostructure. They can also include the standard variety and varieties from a broader range by including items from different domains and areas. The corpus in this regard is of great help in achieving the purposes of these dictionaries because a very large corpus contains a large number and diversity of lexical items on which the lexicographer will base the macrostructure.

It must also be pointed out that both oral and written materials are necessary here. That is why synchronic dictionaries need more comprehensive source material. It is important to have different texts, usually coming from a longer period of time, to be reasonably sure that the lexicon of the language is representatively manifested. Consider the items listed below that are found in the SYC as candidates for inclusion in the lemma list. These lexical items have been randomly chosen and because they are in use in the language today, other items could have been chosen too with the same motivation. However, for the treatment of the items in a synchronic dictionary, the lexicographer should be guided by the frequency counts provided by the corpus to make sure that the items are still frequently used by the speech community of the language. The corpus should reproduce the actual language usage as the success of the

microstructural treatment will depend on the frequency of usage of the selected items and the register in which they are used.

Here are the items from the SYC:

- *Dibala* (man)
- *Dikonsi* (pineapple)
- *Dina* (name; already referred to in the previous chapter)
- *Mambi* (kind of bird)
- *Miogu* (arms)
- *Muguma* (cassava)
- *Mwisi* (ripe banana)
- *Nguji* (pig)
- *Udunda* (to leave)

Furthermore, the macrostructure should be enlarged sufficiently to accommodate different lexical items, words needed by different user groups, regional words, idioms, proverbs, abbreviations, standard varieties, high-frequency words, high-register and low-register words, slang words or informal-register words. The inclusion of these words should be based on analysis of the corpus in such a way that if two words appear in the corpus, the one with the highest frequency count will be preferred for inclusion in the dictionary because it will be considered as the most used word by the speech community. By choosing one word and not the other, the lexicographer will help in the standardisation of the language.

#### **5.4.3.2. *The microstructure of synchronic dictionaries***

The microstructural treatment will focus on giving the meanings of words by providing the definitions or paraphrases of meaning, and they can also be abundantly provided with authentic examples that would extensively supplement the semantic description. In other words, the dictionary articles can include a variety of entries representing a large range of data types (pronunciation data, morphological data, stylistic data, and so on) to be used in the treatment of the lemma. The treatment will



differ from one specific subtype to another; in other words, comprehensive dictionaries will be more detailed than standard descriptive ones that are also different from desk/college dictionaries or school as well as learner's dictionaries. The type of dictionary will have an influence on the nature of the corpus. A comprehensive dictionary will require a more general corpus while restricted and special dictionaries will necessitate a corpus based on a special-field lexicon. This study will not be looking at a specific type but some main features will be shown. The use of a corpus will help in many ways to retrieve the data needed for the microstructural treatment.

The items presented below show the morphological resemblance. They are, of course, homonym items and should be treated separately in dictionaries. The distinction between these items is made at the macrostructural level. The corpus via the concordance lines will help to disambiguate the meanings and the usage of these words by revealing the following:

1) **Mwisi** (ripe banana) singular, **misi** (plural).

*Dinekiri adi dikana misi mipwela* (these 'small' bananas are ripped).

**Mwisi** (somebody from) singular, **bisi** (plural).

*Mwana ami aji mwisi masanga* (my child is from Tchibanga).

2) **Mambi** (kind of bird) singular, **bamambi** (plural).

*Nyitsi boka mambi mu igudiga* (I got this bird (mambi) in a trap).

**Mambi** (faeces) singular; invariable

*Mwana ngebi atsi bivhisa ikutu yami na mambi* (this baby has messed up my clothes with his shit).

3) **Dibala** (man) singular, **babala** (plural).

*Dibala dyeni abe vhavha* (that man was here).

**Dibala** (kind of tree) singular, plural **mabala**.

*Mabala ma mgumuga dibandu mangala* (these trees (mabala) are dry because of the season).

4) *Nguji* (pig) singular, plural *banguji*.

*Nguji aji tutsiboka mbe jimagwela* (the pig we killed was mature).

*Nguji* (mother) singular, plural *banguji*.

*Nguji ami ana magumi masyamunu mabilima* (my mother is 60 years old).

5) *Miogu* (arms) plural, singular *gogu*.

*Gugu ami gumaniengi na mamba ma muji* (my arm burnt with hot water)

*Miogu* (for them) plural.

*Mikansu myogu mi masuka* (their fire wood is finished).

These words, in fact, look the same in writing because a simple writing system that does not take into consideration long vowels and marked tones has been adopted. It must be mentioned briefly that in Bantu languages a semantic distinction between words such as the ones presented above can be possible when the words are used in context but also according to tonal identification. A word containing a low tone will differ from one having a high tone even if they are written alike. The lexicographer taking cognisance of the lemma list and the concordance lines presented by the corpus will use his or her semantic knowledge of the language to identify the semantic relations or the lack thereof that exist between the items.

The treatment of the lexical units, as can be seen in the examples listed below, will require the use of labels to point out the region where the item is mostly used and the register to which it belongs. One should mention that labels play a very important lexicographical role in the presentation of data and the identification of the usage of the lexical items in the dictionary article. Here, the greatest emphasis is directed at geographic labels. All the synonyms are listed separately in the frequency counts and they function as different lexical items. The treatment of each one should be done in different dictionary articles with cross-references that will link them with each other. It is clear that the lexicographer will analyse the results provided by the frequency counts and the concordance lines together with her or his linguistic knowledge of the language to determine, once more, the synonymic relation between these items.

*Dikonsi* (pineapple), name of fruit used in Tchibanga, is synonymous with *Difubu*, used in Moabi, and can be used alternatively. It is also called *dilanga*.

*Muguma* (cassava), frequently used in Moabi, is synonymous with *ikwanga*, used in Tchibanga or Ndende, and to *Mulemba*, occurring in Mouila.

*Udunda* (to leave), used in Tchibanga, is a synonym of *usila*, used in Ndende; *ubumina* mostly from the region of Moabi and Manga used in Mouila.

For the identification of the different meanings of the above lexical items, the corpus will display concordance lines for each lexical item appearing in the frequency counts. As it is done with synonyms, the lexicographer will make use of the statistical results generated by a specific program used as tool (WordSmith Tools in the present case). The lexicographer will also use his or her knowledge of the language to identify different meanings conveyed by the example sentences in which these different items are used by making sure that they illustrate the usage of the items in question in different contexts.

The lexicographer will group all the lexical items that are synonymic and treat them separately using a system of cross-referencing to relate them to one another. The example sentences and the geographical context of usage can be provided as microstructural treatment of the items, as is the case for the items listed above where *Tchibanga* and *Moabi* are used as areas to specify the region where a given lexical item is used. For a dictionary of this kind to be compiled for the Gabonese languages with success, the lexicographer will have to take a good look at the corpus results in order to avoid mistakes and confusion in both the contextual use and the definitions of the items. The following items can be seen as polysemous:

### ***Ubweji***

1 – tasty

*Yamba ayi utsilamba masiga ibtsi ibweji.*

The soup you prepared yesterday was tasty.

2 – beautiful

*Nitsi benguna na mwana mugetu avhu tindi ubweji.*

I met a very beautiful lady.

### ***Dina***

1 – name

*Dina dyami Marundu me mwana bayema.*

My name is Marundu and I belong to the clan bayema.

2 – namesake

*Dina dyami atsi nsumbila sambi.*

My namesake bought me a radio.

3 – the other (homonym)

*Tuna makalu mabeji dimosi a dyami dina dimwana ami.*

We have two bicycles; one is mine, the other is for my son.

The treatment of the polysemous items presented above is not given in a comprehensive way. There is a great deal of data that should have been included in the treatment. However, the presentation provided has the objective of expressing the difference in terms of the period that should be considered when dealing with the material needed for a synchronic dictionary. The data included in the dictionary article are not presented in a disorganised way. The entries must be presented according to the principle of empirical arrangement adopted especially in synchronic dictionaries and it is done in relation to the general lexicographic principle that dictionaries must reproduce actual language usage. For instance, polysemic values are

arranged hierarchically on the basis of the frequency of their use and the register in which they are used. In general, such an arrangement should be determined statistically with the help of frequency counts.

Distinctions in meaning can be based on different categories of usage and register and arranged hierarchically to determine their sequence in the dictionary article. In the treatment of polysemic items in a dictionary article, the following elements could influence the arrangement of values:

- general and regular distinctions in meaning
- archaic and obsolete distinctions in meaning
- colloquial language and dialect forms
- slang or argot
- technical distinctions in meaning

If a corpus can present these different distinctions for a given lexical item, the data should be arranged according to this model. While dealing with polysemic lexical items, lexicographic prominence should be given to those distinctions in meaning that are the best known and that the typical user will wish to consult. The microstructural treatment of the items displaying such differences should be done in a way that will ensure immediate access by users to the information about the meaning that is required. The treatment should be done in a manner that participates in the achievement of the user-friendliness, the genuine purpose and the easy access to the dictionary article that the dictionary has the mission to accomplish.

The treatment of synonyms should also obey the lexicographic principles that will allow a clear identification of all the synonyms of a given lexical item. In this regard, the corpus will present the lexical items separately and the lexicographer will have to identify their semantic relations as being synonymous to one another. The lexicographer should bear in mind that not all synonyms are ‘equally synonymous’. In the treatment of synonyms it is also necessary to make a distinction between absolute and partial synonyms. Absolute synonyms stand in a one-to-one relationship to each other and one can be substituted for the other in all contexts. Partial synonyms cannot replace each other in all contexts as there is a difference in the range of application between the two items. The material contained in the corpus should display enough

evidence to the lexicographer that will allow her or him to draw limits between the different synonyms. The lexicographer before being able to give an account of synonymy needs to be aware of two things:

- He or she should ascertain whether the supposed synonyms can replace each other in all contexts and are thus absolute synonyms or whether they are partial synonyms.
- He or she should determine which member of the synonym paradigm (the collection of lexical items that are in synonymous relationship) has the highest frequency of usage and the corpus can provide the counts for making the choices.

Normally, the synonym with the highest frequency of usage is provided with a definiens. The other members of the synonym paradigm are assigned a synonym definition that refers users by means of a cross-reference to the member of the paradigm that has already been explained. The article containing the explanation of the meaning may not only have the definiens as the sole information on the meaning. In such an article, the explanation of the meaning must be followed by a list of all the members of the synonym paradigm. The purpose is to ensure that the user who begins with the synonym that is explained will also be given an indication of the synonymic relationship of this item to other lexical items.

The sequence in which the members of a synonym paradigm are listed is important because the language can convey implicit information to the user in this way. It is also desirable that the sequence in this list be determined by the frequency of usage of the different synonyms. Thus, the members of a synonym paradigm do not have to be listed alphabetically. If adopted, this approach should be clearly indicated by the lexicographer in the explanatory notes in the front matter texts.

Even if the frequency of usage seems to be the primary criterion that determines which member of the synonym paradigm must be given the definiens, and in what order the other members must be arranged, elements such as specialised language as opposed to general usage, differences in style and register, and standard language as opposed to substandard language can also be taken into consideration. There are some other issues that need to be dealt with when treating synonyms in dictionaries; the lexicographer should pay attention to them when necessary. The corpus planned with

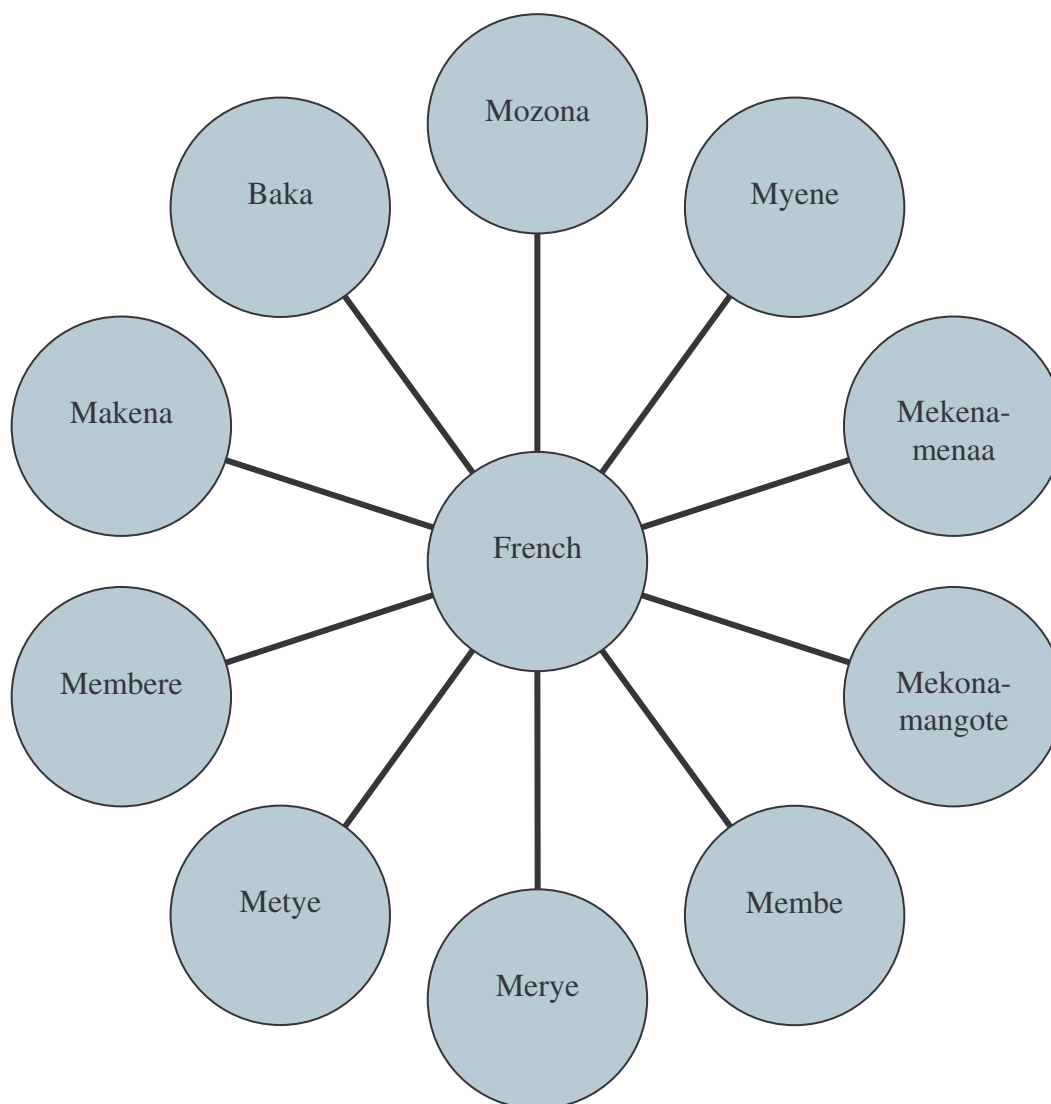
the representative, balanced and hybrid features must contain data that could help the lexicographer in the treatment of all the synonyms in the dictionary. The concordance lines, once again, will display language material that the lexicographer will compare and base his or her choices on. The corpus will display concordance lines of the different lexical items that the lexicographer will have to recognise as synonyms in the end and cross-reference with each other where they are treated in different dictionary articles.

A corpus contains a range of items and data that the language can benefit from in the sense that it allows the different contexts in which these items are used. A language such as Yipunu needs a corpus to improve the lexicographic quality of the language in such a way that several lexicographic activities and interpretations can be envisaged.

### **5.5. Corpus for translation dictionaries**

The term *translation dictionary* is applied to any dictionary that includes more than one language. This includes bilingual and multilingual dictionaries. They translate words from one language into another language, when possible. Trying to translate African languages into European languages or vice versa sometimes can become very difficult because some linguistic realities can occur on one side but not on the other. In this situation other strategies should be applied in the microstructural treatment. The most frequently compiled translation dictionary is a bilingual one that can be monoscopal, for example French-Yipunu going in one direction, or biscopal, for example French-Yipunu, Yipunu-French going in both directions.

It is up to the lexicographer to decide on a specific direction of the dictionary he or she intends to compile, following the identification of the needs of the users and the function of the dictionary. In this case, Yipunu will be the source language and French the target language. Bilingual dictionaries are suitable for multilingual societies such as Gabon where the lack of a dictionary culture can be noted and French is the only official language. The process can be carried out as follows:



**Figure 5.1:** Representation of bilingual/multilingual dictionaries for the Gabonese languages

Corpora developed in these languages will help to compile bilingual dictionaries from the Gabonese languages to French and vice versa. Each language in the diagram can be alternatively used as source language or target language. The lexicographer can use a corpus compiled for each language to generate lemma candidates for each source and target language to allow different translations and treatments according to the principles of translation dictionaries. The translation between two languages will be facilitated by French, which is the language most used by communities because it is the only official language used in the public and private sector. This is a typical



situation that faces the majority of multilingual African countries with a dominating European language.

This proposal is in line with the view of Martin (1996) who presents the hub-and-spoke model, inspired by air traffic organisation in which certain airports (hubs) act as centres to and from which flights from other airports (spokes) operate. The similarity here is that several bilingual dictionaries can be compiled by linking the lexical items of the spoke languages (source languages) to those of a common hub language (serving as a target language in a language pair). In the case of the Gabonese languages, French can form a hub-and-spoke configuration with Merye and Membe or with Makena and Mekenamena. In fact, Martin specifies that this consists in providing for and making explicit the kind of relationship there exists between the lexemes of the spoke language and their equivalent hub lexemes. In order to achieve the compilation of dictionaries using these different languages, corpora in all these languages have to be compiled so that the different lexemes can be generated from the different corpora. That will help the lexicographer extract the macrostructure of each spoke and hub and the different microstructural programmes. Different parallel corpora could be used for each language pair for the compilation of each bilingual dictionary.

The use of this model in such an environment could save time and energy for compilers and reduce the number of bilingual dictionaries to be compiled. With a central-hub lexical database, the spoke and hub can be reversed and spoke 1 and spoke 2 can also be linked and reversed. To ensure the compilation of these bilingual dictionaries, one must invest in the gathering of language material in each language representing a hub or a spoke. The corpus will then supply the lemma candidates necessary for the compilation of the dictionaries following either one direction or both according to the language pair involved. Each lemma candidate list of any hub or spoke language used as macrostructural items for a source language will have to be systematically translated into the target language. In this regard, each language should have a corpus that will be used in parallel with another corpus to enable different combinations on both the macrostructure and the microstructure levels. The parallel corpora will ensure the compilation of dictionaries between a spoke and a hub. In the Gabonese context the situation will allow one to have parallel corpora between French and Yipunu, Fang, Yinzebi and Omyene, to name but a few.

### **5.5.1. The macrostructure of translation dictionaries**

The macrostructure of bilingual or multilingual dictionaries displays an alphabetical arrangement of the lemmas. The macrostructure could contain both standardised and nonstandardised lexical items, depending on the type of dictionary to be compiled. The arrangement is not different from the one used in monolingual dictionaries. The following two texts, already used in Chapter 3, can rightfully provide potential lemmas to be included in a bilingual or multilingual dictionary. The choice of the lemmas to be included in the macrostructure of a translation dictionary is also indicated by the frequency counts provided by the corpus. The restriction and the extension of the macrostructure as well as the positioning of entries are influenced by semantic aspects. A polysemic lexical item will be presented as one lemma while synonyms will be included as separate lemmas.

#### **Yipunu**

Yika bivunda bamasila nana tsi jetu.

Mba me nyisamala ne bivunda bye.

Na bamaburu na bamabanguga.

Bamena bura bamama.

Bamama bamenatura.

Me nymaburulu o Makaba Disimu.

Me dina dyami majinu ma Kombila.

Yibanadu yami jungu.

Yika dusavu dwedudu bivunda bamadusavanga.

Bibunda byenyi bajaji bami bakaduwaka.

#### **French**

Ce sont les anciens ont laissé comme ça pas nous.

D'ailleurs je ne les ai jamais connu.

Ils sont nés et ont grandi avant nous.  
Ils ont mis nos mamans au monde.  
Nos mamans nous ont mis au monde.  
Je suis né à Mukaba-Disumu.  
Je m'appelle Madjinou Kombila.  
Je suis du clan Djoungou.  
Ainsi, cette histoire était racontée par les vieux.  
Et les vieux l'ont aussi appris des autres.

### **English translation**

It is the elders who transmitted what follows not us.  
Besides I never knew those elders.  
They were born and grown up before us.  
And they gave birth to our mothers.  
Our mothers gave birth to us.  
I was born in Mukaba-Disumu.  
My name is Madjinou ma Kombila.  
My clan group is Djoungou.  
Thus, this story was told by the elders.  
Those elders were also taught about.

From these texts, two different lemma lists can be drawn out separately for the two languages and can be used commutatively as target and source languages. The lemma candidates can be listed as follows in the two corpora:

## **Yipunu**

*Bivhunda*

*Byeni*

*Dusavhu*

*Dyami*

*Ne*

*Uburu*

## **French**

Vieux (elders)

Moi (me)

Jamais (never)

Histoire (story)

Ceux (those)

Naître (to be born)

In a bilingual dictionary using bidirectional translation, in other words, in which a lemma sign of a source language has to be translated into a target language, and the one for the new source language and new target language requires the same in a reversible way, the lexicographer cannot always find a translation equivalent in the target language for each source language item. When this happens, the missing equivalent in a target language can always be added to the lemma candidate list of that language as a macrostructural element. This is also applicable in the reverse situation when a source language becomes a target language and a target language a source language.

According to Gouws (1989:162), the lexicographer has to give users the opportunity to test the communicative aptness of a given translation equivalent. There are many ways of achieving this. For him, one of these, which is regarded as the most obvious method and one that every lexicographer has to accept as part of his or her

responsibility and lexicographic assignment, is the application of the ‘reversibility principle’. According to the reversibility principle, a lexical item A included as translation equivalent of a lemma B in the X-section of a bidirectional translation dictionary has to be included as a lemma in the Y-section of the dictionary with at least the lexical item B, the relevant lemma from the X-section, as one of its translation equivalents. Each lexical item included as translation equivalent in the Y-section of a bidirectional translation dictionary has to be included as lemma in the X-section of the dictionary with at least the respective lemma from the Y-section as a translation equivalent. In practice this is feasible in a situation where the identification of a translation equivalent in the target or the source language is not obvious and in an environment such as Gabon where good bilingual dictionaries are needed. The reversibility principle is more or less similar to the hub-and-spoke theory in the sense that both are applicable in a bilingual environment, in other words a situation where bilingual dictionaries are needed.

As it has already been noted with regard to the hub-and-spoke theory, the reversibility principle has everything to benefit from using corpora to generate first the macro- and the microstructural components of each language. The lemmas of the source or the target language will have to be translated and used as different lemmas and translation equivalents for the new lemmas for the other language(s). All the different transformations, that is, from the lemma of a source language to a translation equivalent of a target language and vice versa, will equally benefit all the languages in Gabon in both the improvement of the number of words in the corpus of each language pair and the data quality.

### **5.5.2. The microstructure of translation dictionaries**

The treatment of these items on the microstructural level could give each lemma of the source language a translation equivalent in the target language(s) for a bilingual or multilingual dictionary. The translation could be preceded by data on pronunciation and followed by examples. When possible the treatment should make provision for polysemous senses. As far as the lemmas given above are concerned, the treatment will be limited to providing a translation equivalent of the Yipunu lemmas in French followed by an example.

Yet the treatment illustrated below is not presented in a more comprehensive way; the objective here is to show how a corpus can help in the retrieval of a lemma sign in a source language together with the example sentence to illustrate the usage of the specific lexical item. In bilingual dictionaries, the meaning is also regarded as an important data type to be included in the microstructural treatment of a lexical item, as it is for descriptive dictionaries. However, data on, for example, pronunciation, part of speech, the form of the item (singular or plural, if applicable) and the different contexts of usage as well as a cross-reference to a synonym lemma in the central list or in the back matter texts are very important information that dictionary users can look for in bilingual dictionaries. The lexicographer must pay attention when compiling a dictionary so that the users' needs are considered first. The treatment and the presentation of the different data types should be done in such a way that users can easily retrieve the necessary information in a text production or text reception situation. Both the comment on semantics and the comment on form should display the most important data that the average user of a bilingual dictionary would be looking for when consulting such a dictionary. The treatment should negotiate the access structure, the search area structure and the mediostructure to assist the user in the dictionary consultation process.

The examples provided below have been randomly chosen from the concordance lines of the SYC and they have not been edited. For their adequacy and effectiveness the lexicographer will have to compare a large number of concordance lines from which the best examples have to be chosen. It is also up to the lexicographer to make sure that the chosen examples illustrate the treated lemma with precision. Choosing examples from a corpus helps to avoid the use of self-constructed ones, which is often practised by lexicographers instead of reproducing the way in which the language is authentically perceived by the speech community.

*Bivhund* , vieux (elders)

*Yika dusavu dwedudu bivunda bamadusavanga.*

Ainsi, cette histoire était racontée par les vieux (thus, this story was told by the elders.)

*Byeni*, ceux (those)

*Bibunda byenyi bajaji bami bakaduwaka*

Ces vieux aussi l'ont appris des autres (the elders also learned the story from others.)

*Dusavhu*, histoire (story)

*Yika dusavu dwedudu bivunda bamadusavanga.*

Ainsi, cette histoire était racontée par les vieux (thus, this story was told by the elders.)

*Dyami*, my

*Me dina dyami Majinu ma Kombila.*

Je m'appelle Madjinou Kombila (my name is Madjinou ma Kombila.)

*Ne*, never

*Mba me nyisamala ne bivunda bye.*

D'ailleurs je n'ai jamais vu ces vieux (besides I never knew those elders.)

*Uburu*, naître (to be born)

*Me nymaburulu o Mukaba Disimu.*

Je suis nait à Mukaba-Disumu (I was born in Mukaba Disimu.)

These items can be used in the compilation of trilingual or multilingual dictionaries with a slight difference in the treatment that will provide only a translation equivalent of the target languages. The Gabonese languages displayed in Chapter 3 can play an important role in the compilation of a well-devised Gabonese multilingual dictionary(ies). The corpus of the target language can render a lemma list in which the translation equivalent(s) to be used as part of the treatment of a lexical item from a source language are included. That means that a corpus containing material from each

language will be used for the illustration of the lexical items in the source language and the target language by giving both the translation equivalent and the sentence example.

The microstructural treatment of the lexical items in bilingual or translation dictionaries involves many theoretical and practical principles. Not all these matters can be discussed here but the lexicographer has to take cognisance of these phenomena when dealing with these types of dictionary. The data provided by the corpus will play an important role in both the treatment and the presentation of the lemmata. The frequency of usage of a lexical item, which can be indicated by the data provided in a corpus, will always be of prime importance to support the lexicographer's choices in the treatment of, for example, polysemous lexical items and synonyms in dictionaries.

The presentation of translation equivalents can be regarded as the most important entries in the article of a translation dictionary from a semantic point of view. The average user of the translation dictionary consults the dictionary primarily to obtain a target language form for a given source language form. The lexicographer should make sure that the presentation of translation equivalents reflects the total spectrum of semantic values. In this regard, the corpus should be able to provide data that prove the contextual and the cotextual use of the selected lemmas included as a microstructural treatment in a bilingual or multilingual dictionary. These evidences extracted from the corpus must reflect or render one or different meanings that a lexical item can cover when used in a given situation. These elements constitute different data types that can be necessary to illustrate specific phenomena in the treatment of the items. Each potential lemma and translation equivalent in bilingual and multilingual dictionaries should be fully treated when possible and the lexicographer should make sure of that by using a corpus to extract all the necessary data.

The respective translation equivalent paradigms, in other words the presentation of one or many translation equivalents in a single article, will render different types of equivalent relation. The nature of the language will determine the translation and the treatment of each lemma sign. One will not always have a situation of one-to-one translation in the sense that a certain lexical item will be translated by only one item in a target language. It often happens that some lexical items in the source language



can be translated by more than one translation equivalent in the target language; in other words, a lemma displaying monosemy in the source language can be translated by a polysemic lexical item in the target language. The comment on semantics in the dictionary article of a bilingual dictionary should include the different polysemous senses of a given lexical item in the treatment through the translation equivalent. It is not always possible to deal easily with polysemy because for a certain polysemous item in the source language, one will not necessarily find a target language translation equivalent with the same polysemous senses. In this regard, the lexicographer is supposed to present a translation equivalent for each polysemous sense of the lemma sign. In some instances the translation equivalents represent target language synonyms while some represent target language forms for the different polysemous senses of the source language form. The lexicographer always has to present data in a way that will assist users to retrieve the information needed for a specific lemma in the dictionary article.

When talking about different types of equivalent relation one has to make a distinction between full equivalence, partial equivalence and zero equivalence. In this research, these different equivalent relations are not elaborated on extensively. A more comprehensive discussion can be found in Gouws (1989; 1996; 2000; 2002).

Full equivalence exists when a source language item, represented by a lemma sign, is related to a single target language item, represented by a translation equivalent, and this one-to-one relation prevails on both a lexical and a semantic level. This type of equivalence is also known as congruence. The source language item and the target language item have exactly the same meaning, function on the same stylistic level and represent the same register.

Partial equivalence occurs when there is no one-to-one relation between source and target language items. This can be only on the lexical or on the semantic level or on both levels. Partial equivalence also includes an equivalent relation of divergence, which prevails in a one-to-more-than-one relation between source and target language. The article displaying an equivalent relation of divergence contains two subtypes: lexical divergence and semantic divergence. Lexical divergence exists when a monosemous lexical item, functioning as lemma sign, has more than one translation equivalent, being target language synonyms, whereas semantic divergence can be observed when the lemma sign represents a polysemous lexical item.

Zero equivalence will be observed where the target language has no item to be coordinated as translation equivalent with a lemma representing a source language item.

The corpus will have to contain data that will help the lexicographer to render the different equivalent relations that prevail between the source and the target language, if possible. The lexicographer dealing with the corpus of each language pair should interpret the different contexts and limits of the equivalence adequately. Bilingual and multilingual dictionaries compiled for the Gabonese languages should obey the basic principles if they are to play a key role in the multilingual Gabonese environment. The different corpora compiled for these languages should be able to assist lexicographers in the treatment of both the potential lemmas and the translation equivalent of the different languages.

## **5.6. Corpus for restricted dictionaries**

Restricted dictionaries can be bilingual or monolingual, depending on the purpose of the dictionary to be compiled. By extension, specialised dictionaries can be included in this category because they all deal with special matters or special fields of research.

### **5.6.1. The macrostructure of restricted dictionaries**

The macrostructure of these dictionaries can display an alphabetical arrangement of lemmas in the central list if dealing with special-field lexical items. Otherwise, when dealing with dictionaries such as one for proverbs, alphabetical ordering is not compulsory. The lexicographer could order entries according to the frequency of usage of each one so that the most used and best known will be treated first. The lexicographer has to present the different proverbs in a manner that will facilitate data accessibility in a systematic way. The lexicographer could also set up his or her own criteria of ordering; the use of typography (bold, italics, etc.) in a separate position can play an important role in this regard. As far as the proverbs presented below are concerned, the ordering was done according to the order of appearance in the corpus. Consider these proverbs as potential candidates for inclusion in a Yipunu dictionary of proverbs.

- *Ba ditotu me ngaba mungeli.*
- *Dingenda katsyagu mugumbi katsi ngana.*
- *Dilongi asamabasa ponzi.*
- *Koku busa ndagu mutu yivhuni.*
- *Ngebi busa ndongi yikota matsugu.*
- *Ngebi busa dilongi yitsukma diambu.*
- *Muana tsiana ageboli nzagu.*

The proverbs contained in the corpus will also have to be identified by the lexicographer in the corpus. Failing to extract proverbs in their totality will impede the transfer of meaning of some material. The lexicographer will end up with portions of sentences that cannot be used effectively. Proverbs can also be included in a general dictionary and their treatment at all levels should obey the requirements and principles of such a dictionary. The lexicographer will have to make sure of this. Other restricted dictionaries are not the focus of this dissertation.

The lexicographer dealing with a specific restricted dictionary will have the mission of compiling it in a manner that will take into account the target users and their reference skills, the functions of the dictionary and the nature and extent of the lexicographic treatment. Focusing here on restricted synchronic dictionaries, one sees that the selection of items usually represents a well-defined subsection of the lexicon of the given language, for example the lexicon of a specific hobby or specialised field. A lexicographic dictionary, a dictionary of soccer terminology or an aeronautics dictionary would be included in this category; only a subsection of the lexicon is selected for inclusion. However, the restriction in the selection of lexical items can also be determined by other criteria; for example, a dialect dictionary reflects a social grouping and a dictionary of idioms treats a specific type of lexical item.

Furthermore, if one deals with a pronunciation or a spelling dictionary, the lexical items will have to be selected from the general vocabulary but their treatment should be restricted to the presentation of a single type of linguistic data. Another type of restricted dictionary is the thesaurus. A thesaurus reflects semantic data, especially different types of semantic relation between lexical items. The ordering of lemmata in a thesaurus differs from that of a general dictionary because a thesaurus does not

reflect an alphabetic ordering but rather an ordering according to semantic fields. Restricted dictionaries can be either bilingual or monolingual.

In order to successfully compile restricted or specialised dictionaries in Gabon or elsewhere, the corpus will have to contain lexical items pertaining to a special field. The data will have to be composed of different lexical items or words that come from, for instance, mathematics, in which the different theorems and vocabulary are more or less described and explained. The selected items will require of the lexicographer to approach a qualified person in that field to help with the definitions. The lexicographer may also be obliged to create new terms in a given language as translation equivalents for terms from another language that is well-established, with the aid of the speech community.

#### **5.6.2. The microstructure of restricted dictionaries**

The lexicographer has to compile a microstructural programme that will help the different users in the retrieval of the variety of information needed in the dictionary consultation process. For example, in a dictionary of proverbs, a clear distinction should be made between the macrostructural presentation of the proverbs and their microstructural treatment in order to avoid confusion. As far as the microstructural treatment is concerned, it can be limited to giving the explanation of the Yipunu proverb in French followed by its translation equivalent. It is important to mention that for the microstructural treatment of these items, the focus will be on the translation equivalent, if possible, or a paraphrase of meaning in French. When possible, example sentences will be provided to illustrate the usage of the items in context, for a clear understanding of the proverbs.

*1 Ba ditotu me ngaba mungeli*

Il faut laisser aux autres le soin de reconnaître notre valeur ou notre talent

In life it is better to be down to earth

*2 Dingenda katsyagu mugumbi katsi ngana*

Quel qu'il soit l'oncle doit être respecté; Faute de grive on mange les merles

Your uncle remains your uncle no matter what

*3 Dilongi asamabasa ponzi*

Un conseil n'a jamais rempli une hotte, il ne faudra pas donner des conseils à quelqu'un tous les jours pour qu'il comprenne

No advice can fill up the basket; that is, you do not need much advice to understand

*4 Koku busa ndagu mutu yivhuni*

Une poule qui ne veut pas être enfermée est la proie des prédateurs

A pet that does not stay in a safe place risks to be killed

*5 Ngebi busa ndongi yikota matsugu*

L'enfant qui refuse les conseils risque de se trouver dans les problèmes

A youngster who refuses advice can end up in trouble

*6 mbatsi atsigudi yatsi ndegwa mudi ngumba*

Œil pour œil dans pour dans

An eye for an eye and a tooth for a tooth

*7 Muana tsiana ageboli nzagu*

Il ne faut pas créer des problèmes quand on ne peut compter sur un soutien

If you are poor do not pretend to be rich

If an idiom in a source language is translated by an idiom in a target language, the lexicographer must make sure that there is a transfer of meaning and that the metaphor is the same in the two languages. This means full equivalence so the

lexicographer will not need to provide additional comments in the treatment of such idioms, especially when the agreement is close. The presentation of the idioms can also be done in the same way it is for proverbs, using typeface and labels to indicate them.

A restricted dictionary does not necessarily mean a restricted treatment of the lexical items. The treatment should cover all the contexts of definition and application of the lexical items and the corpus should always play an important role in the extraction of the necessary data. When compiling such dictionaries, the lexicographer will have to provide the users with the data needed to catch the meaning of the concept or terminology the users would like to be informed about by respecting the principles.

### **5.7. Data distribution: A corpus view**

In whatever way distribution is being done by the lexicographer, the corpus remains the key element of the whole process. All the material at the disposal of the lexicographer cannot be used in the central list of a dictionary. That is why to avoid confusing the users, some material needs to be used in the outer texts: the front matter texts and the back matter texts. The lexicographer has the freedom of choosing which data he or she would like to present in one of these text types. In fact, when planning the compilation of any dictionary, lexicographers have to emphasise the functionality of the eventual product that exceeds the limit of the central list.

Both the front and the back matter can contain texts that play a functional role in the presentation of the lexicographic data. A dictionary that displays a frame structure increases the options of the lexicographer when planning the lexicographic presentation. Apart from assisting the user to ensure successful dictionary consultation procedures and to obtain an optimal retrieval of information, outer texts also play a prominent role in the data distribution structure of the dictionary by enabling the lexicographer to accommodate the lexicographic data in more than one text (cf. Gouws & Prinsloo 2005). This justifies why the dictionary is referred to as a carrier of text types.

The corpus contains a variety of data that are needed for the distribution process. This data help the lexicographer when choosing data for the outer texts and the central list. In fact, the data in the outer texts are included to serve a specific purpose along with

the material in the central list. The relation between the data included in the three areas that is the front matter texts, the central list and the back matter texts, is explicitly indicated by means of cross-references, if necessary.

The management of this data is part of the formulation of the dictionary conceptualisation plan with regard to the decisions to be taken on the structure of the dictionary and the possible use of the outer texts. The lexicographer has to decide on the material to be part of the outer texts and whether he or she prefers integrated or nonintegrated outer texts. These two concepts will not be fully discussed here; they are referred to just to help in the explanation of data distribution.

The corpus contains the spoken and written materials that are used for the distribution strategies of the dictionary. These materials will help the lexicographer to extract data to be included in the front and back matter texts and in the central list.

According to Gouws and Prinsloo (2005), the aspects of the formulation of the dictionary conceptualisation plan fall within the scope of the data distribution programme of the dictionary. It is in the same way the programme that organises the distribution of all the lexicographic data between the different texts presented in the dictionary.

Some outer texts contain data from which users can retrieve information about the subject matter of the specific dictionary, for example data regarding the meaning, grammar and spelling of lexical items in a general dictionary or technical data in a language for special purposes dictionary (cf. Gouws & Prinloo 2005:59). Deciding on what data to include in the outer texts is done in accordance with users' needs and the desire to satisfy the needs of those users. The dictionary type as well as the skills of the users should guide the choice regarding data included in the outer texts.

As a guide to the use of dictionaries, one of the texts in the front matter will describe every part of the dictionary article: lemma, syllabication, pronunciation and inflected forms, various kinds of label, cross-references, variants, etymologies, synonyms and usage notes. The purpose of the user guide in the front matter text is to describe as clearly as possible all the kinds of data included in the dictionary. In other words, it shows the reader how to interpret the data given or how to understand the style of the dictionary and provides clues about locating particular items of data as quickly as possible (cf. Landau 1984).

In the back matter texts, some dictionaries have sections listing bibliographical and geographical names not included in the central list. These texts will also give a variety of practical guides to writing, covering punctuation, grammar, style, form of address and proofreading marks. In some cases, tables of weights and measures, signs and symbols, lists of abbreviations, foreign words and phrases as well as given names are also included. In a learner's dictionary, back matter texts contain various linguistic aids, such as lists of irregular verbs and tables of ordinal and cardinal numbers.

During the lexicographic process, special attention should be given to the outer texts in order to deliver a quality dictionary to users. Some lexicographers are not rigorous in the selection of significant data to be included in their dictionaries. This situation can negatively affect the value of the dictionary and one has to bear in mind that any dictionary is compiled in order to be appreciated and utilised. The product is not for the compiler but for the community that the dictionary aims to serve. For the Gabonese languages this should be taken seriously, and equal attention should be paid to the outer texts and the central list. This will ensure the fulfilment of the genuine purpose of the dictionary. The more material there is available, the more data there are at the disposal of the lexicographer to help her or him in the distribution of data in all the component parts of the dictionary as carrier of text types.

If a frequency list of a corpus presents 20 000 lemmas to be included in the dictionary, it is clear that all the lexical items will not be treated in the central list of the dictionary. For some practical reasons, many of the items will be included in the back matter text of the dictionary. Among these elements to be treated in the back matter text can be found cultural data and other elements such as country names, capital cities, idioms, abbreviations and measures and weights as well as different formulae, for example for condolences, congratulations, commands, greetings, rebukes, thanksgiving, warnings, and so on. These back matter items have to be related to the ones treated in the central list by means of a cross-reference. A comprehensive explanation can be found in Gouws (2002).

The corpus can be used to retrieve those words that can be arranged thematically in the back matter text in secondary macrostructures. The use of back matter text that makes provision for inclusion of lists qualifies the dictionary as a poly-accessible source because it has more than one macrostructure. Thus, after deciding on which data to include in the outer texts, more precisely in the back matter texts, the



lexicographer should not present the data in an arbitrary way. The data that are chosen to be presented in the back matter texts should be grouped according to their lexical relation and alphabetically ordered in text blocks. Each text should accommodate lexical items of a given category.

This should be done in a way that will make it easier for users when looking for specific information in the dictionary or accessing different text types. In this regard the way in which the items are selected, arranged and presented in the back matter text participates in fulfilling the genuine purpose of dictionaries. The use of this system of presenting lexical items in the different texts, namely front and back matter texts and central list, is suitable for dictionaries adhering to a frame structure to help the lexicographer to find a place to accommodate data that do not need to be treated in the central list. The data distribution structure has many implications with regard to the positioning of the data, the nature and the extent of the presentation and the integration of several textual components.

With regard to the use of outer texts to accommodate extra data in dictionaries, it is important to mention the existence of integrated and nonintegrated outer texts (cf. Kammerer & Wiegand 1998, as referred to by Gouws & Prinsloo 2005). Nonintegrated outer texts function alongside the central list and are not needed to retrieve information presented in the articles of the central list or to contain data relevant to achieving the genuine purpose of the dictionary. Integrated outer texts function in conjunction with the central list and are needed to ensure optimal and full retrieval of the data distributed in the dictionary with regard to the subject matter of that dictionary, in order to achieve the genuine purpose thereof. Different data can be allocated to and treated in one of the outer texts, and the lexicographer must make sure that this is done in a way that does not mislead users when looking for specific information during the dictionary consultation process. Additional analysis of the different types of item and the way in which they are treated in the outer texts can be found in Gouws (2001; 2004) and Gouws and Prinsloo (2005).

The compilation of representative and hybrid corpora will make provision for the inclusion of different data types that will assist the lexicographer in the distribution of data in the different texts in the dictionary. Based on the language material available in the corpus, the lexicographer during the dictionary conceptualisation plan of the

dictionary specific lexicographic process will be able to decide exactly where the different categories of lexicographic data should be accommodated.

In addition to the different data types that are commonly used in some dictionaries and presented in this research for the benefit of the Gabonese languages, the back matter texts of the Gabonese dictionaries can accommodate specific data from the Gabonese cultures. Where necessary, lexicographers could present in the back matter texts data regarding traditional wedding ceremonies, hunting techniques, Gabonese food and cooking, funeral and burial events, farming and fishing systems, traditional courts, astral representations, cultural dances and rites, stories and tales, clan groups, and so on. This data must complement the cultural data that should be treated in the central list of the planned dictionaries.

### **5.8. Concluding remarks**

All these different strategies are only possible if one uses a well-designed corpus that represents the language in different elocutional situations. The quality of a dictionary is normally judged by the quality of the corpus used to extract the necessary data to base the analysis and the compilation process on. All the macrostructural options and microstructural treatments are possible only if one has a corpus that contains all the necessary data.

The way in which the data are organised and structured in the corpus is of great help in the whole dictionary compilation process. It is very important to plan the compilation of any dictionary properly so that with the data contained in the corpus, the lexicographer has time to manage the material in order to fit the process. In the process of the compilation of the dictionary, the lexicographer remains the final architect who organises the material. A corpus can show in different concordance lines multiple senses of a given lexical item, but the lexicographer has the final say in identifying those senses and compiling them in order to be utilised efficiently.

The development of corpora for the Gabonese languages will allow speeding up the compilation of dictionaries and will facilitate the macrostructural presentation and microstructural treatment of the lexical items of the different dictionaries. A corpus enables the lexicographer to acquire in a short period of time all the material needed for the presentation of lemmas in dictionary articles. With the use of corpora, the

lemma candidates, definitions, example sentences, clusters, collocations and other important data types are identified more easily than before. A corpus contains various data types that can be used in different dictionary types, and the lexicographer will have to select the ones needed for the compilation of a specific type of dictionary. In this regard, when the main corpus or the hybrid corpus is not able to assist as expected, the lexicographer should compile a dedicated corpus in order to improve the available data.

## **CHAPTER 6: GENERAL CONCLUSION**

This dissertation has helped the reader to realise that compiling a dictionary is a practical activity that requires practical material from which the necessary data have to be selected. The nonusage of language material as evidence for decision making about the material to be included in or excluded from a dictionary has led lexicographers to the compilation of dictionaries that leave out important data frequently used by the speech community for which the dictionary is compiled. The use of mechanical methods to compile dictionaries was such that finishing the compilation of a single dictionary was laborious and time consuming.

Positions have been taken up on some existing theoretical issues to clarify understanding of or give further insight into the theoretical concepts. The development of corpora presupposes many challenges that the lexicographer will encounter and will have to overcome throughout the research. This study has demonstrated that the gathering of language material is preceded by identification of the material and of the places where different sources will be found before formulating any data collection policy.

Over time, with the development of corpora and appropriate programmes and query tools, the completion of dictionaries has sped up and the lexicographer's task has become relatively unproblematic. For a project of this kind to be completed with success, one has to invest time, energy and finances in the gathering of representative language material. These different elements have to be considered by lexicographers of languages such as the Gabonese ones where the lack of sound written materials is still persistent, and urgent actions need to be taken in order to solve this problem. It is furthermore imperative for lexicographers to start recording and transcribing material in the Gabonese languages in order to fill the gap. These oral materials are an extremely rich source of data.

The above-described context explains why in this work the main focus was directed at spoken sources: because of the lack of a written tradition in Gabonese languages. Spoken materials were recorded among mother-tongue speakers and together with the available written sources they constitute the materials for the Yipunu corpus used in this dissertation. These recordings were transcribed and the process is still going on because many materials have not yet been processed. In these materials are included

public or official material, radio and television scripts as well as natural, spontaneous or private material such as transcripts of conversations, dialogues and interviews. A Detailed discussion of this situation is provided in chapters 2 and 3. As far as written sources are concerned, materials such as linguistic and religious texts were the focus in this dissertation and constitute the domains in which such materials were available.

Throughout this study, the researcher has come to realise that when gathering language material in a certain community, the lexicographer is working with the potential users of the dictionary that will be compiled from that material. This is ensured by the involvement of speakers from different age groups, level of education and background. During the research, attention should also be paid to the inclusion of both female and male informants who are sometimes speaking different varieties of the same language. This explains the presentation of two classifications of Gabonese languages in order to get a picture of the diversity that the country presents. When the research is completed, the lexicographer has to organise the material in order to start the lexicographic process. This goes along with the early identification of the needs of the potential targets users of the dictionary being compiled.

In this research the focus has not really been on the development of databases but rather on the research of language material for the compilation of corpora from which materials to be used in the macrostructure and the microstructure during the compilation of dictionaries could be extracted. The gathered material could be stored in a database at a later stage. Many corpus compilers have shown and described the relation between corpus type and dictionary compilation. They have then distinguished between monolingual and bilingual or multilingual corpora, synchronic and diachronic corpora, spoken and written corpora, samples and reference corpora, monitor and parallel corpora, and comparable and multilingual corpora.

The different corpora are compiled to serve certain functions and must follow certain principles. A clear distinction is made between corpora for general studies in the language and corpora for the compilation of dictionaries. The latter category, which is the focus here, has to be able to really assist the lexicographer in the dictionary conceptualisation plan.

Apart from the material acquisition phase, the material preparation phase and the material processing phase that have been discussed in detail in this dissertation, the

publishing preparation phase must be seen as the final activity for the lexicographer before the realisation of the dictionary. It is during these phases that most important decisions are taken and they describe step by step the whole process, from the gathering of the materials to the compilation of a dictionary.

It is noticeable that each corpus type is helpful in the compilation of a specific type of dictionary. However, the researcher believes that a corpus should be compiled in a way that allows the compilation of many dictionary types. In order to better achieve the goals of the lexicographer when compiling a very specific type of dictionary, special-field dictionaries for instance, special corpora can be envisaged. In contrast with this and to better serve the purpose of a multilingual corpus and allow dictionary compilers to use their corpora for different dictionaries, the researcher suggests the compilation of hybrid corpora, which should be as representative as possible and should contain materials from as many different sources as possible. This will help to provide Yipunu with a huge volume of data to be included in a database. Databases as such will be developed at a later stage.

Up to now, language materials have been regarded as the sole elements that constitute the corpus, certainly because of the important role that they play in day-to-day conversations. But in this research, the researcher appeals to a broader understanding of the concept *corpus*. The researcher believes that one should consider a paradigm of different pictorial illustrations as a corpus of pictures because, as explained in Chapter 5, they participate in the transfer of meaning of the lexical items that they are used for. That is why special attention should be paid to them, both regarding their collection and their inclusion in different dictionaries.

In addition, this research has made provision for the formulation of a well-designed data collection policy that will guide the collection of the necessary material to be included in the corpus. The present research has given attention to the formulation of such a policy whereby the importance of the collection of the material has been emphasised. The researcher has come to realise that the identification of sources into what is called, by many lexicographers, primary sources, secondary sources and tertiary sources can be reduced simply to written sources and oral sources. The reason is that primary, secondary and tertiary sources are composed of either written and/or oral material. All these sources are containers of data types that are important in helping the lexicographer to identify which of these data types are candidates for

inclusion in a dictionary. The different types of data, for example pragmatic data and morphological data, are useful for the enhancement of the lexicographic representation of both the comment on form and the comment on semantics.

Many dictionary users have formulated criticisms of the way in which the comment on form of some dictionary articles is presented. In order to enhance that presentation on the lexical item level, the researcher has found out that the research in HLT has something more to give to lexicography. In this regard, Roux and Bosch (2002) have already shown how HLT can be applied in the compilation of electronic dictionaries on the one hand and on the other hand how through corpora these technologies are useful for the creation of interactive systems such as multilingual telephone-based information systems, multilingual multimedia information systems and multilingual automatic/machine-aided translation systems. Based on these technologies, the researcher has constructed text grids from some oral material to generate spectrograms. This was done in order to explain the way in which printed dictionaries can also benefit from new technologies in phonetic word segmentation, segment identification and spectrographic segment identification. All of these are very important for the phonetic and morphological presentations of lexical items in dictionaries.

In order to fully achieve its objectives and facilitate the lexicographer's work, the language needs to have a stable orthographic system, even if adjustments are made from time to time according to the needs of the language. In a multilingual society such as Gabon, as presented in this dissertation, the development of all the languages is certainly possible if a unique orthography is compiled for all the languages with some differences according to the specificities of each language. The reason is that the majority of the languages share the same morphological features because they are all Bantu languages.

That is why the researcher has presented the different orthographic alphabets that have already been compiled for these languages and implemented them in this research. It is clear that each language will have to choose between disjunctive and conjunctive writing. This is possible with the establishment of lexicographic units and a language board with a multitask function in each language. To avoid too much work because of the number of languages that need dictionaries, it is important that the language policy of Gabon allow choosing some major languages and giving them the

status of official languages. This will help to eliminate some linguistic issues that the lexicographer can encounter while working in these languages.

Furthermore, this research has led to the development of the SYC. The transfer of written and oral materials into digital formats was made possible by using three different methods: scanning, keyboarding and downloading. The way in which the material is made ready by the correction of errors of different kinds before enabling the manipulation of the corpus with WordSmith Tools is also explained in detail. The analysis of this corpus has produced interesting results that can be used in the compilation of Yipunu dictionaries. This was done through quantitative and qualitative methods that revealed that the corpus contains 1 352 902 tokens and 64 660 types. From the corpus frequency list from which lemma candidates lists can be chosen, the lexical item *na* appears to have the highest frequency. Concordance lines of some items have been generated in order to sort example sentences, definitions, collocations and clusters. These outcomes of the corpus have an effect on the quality of lexicographic representations of pragmatic, semantic and grammatical data in dictionary articles. Thus, the corpus has allowed the measurement of the alphabetical stretch of the macrostructure.

In terms of compiling dictionaries, this research has focused on demonstration of how the corpus can assist the lexicographer to complete his or her task. More specifically, the researcher has presented how the corpus can help to improve the macro- and microstructural treatment of lexical items in dictionaries. This was done in accordance with what has been regarded as principal dictionary types. The principle that qualifies dictionaries as containers of text types has been used to portray how a corpus can help to improve data distribution in allocating data to a specific part of the dictionary. That takes into consideration the front matter texts, the central lists and the back matter texts. The use of both the primary macrostructure in the central list and the secondary macrostructure in the back matter text has played a prominent role in the process. All of the above are made possible by keeping in mind the communicative-oriented function and the cognitive-oriented function that any dictionary has to perform to be able to satisfy the needs of the users.

The selection of the different lexical items to be included in the macrostructure of any dictionary as well as the microstructural treatment of these items should not be done arbitrarily but should be done following the principles of each dictionary type. In this



regard, the corpus will assist the lexicographer in decision making as it will display different data types needed for the arrangement and the presentation of the items in the dictionary articles of each type of dictionary. The arrangement and the presentation of data in historic dictionaries will be different from that in synchronic dictionaries and bilingual dictionaries, with some minor resemblances.

Overall, it is possible to state that the goals of this research were achieved because with regard to any lexicographic activity that can take place today, the researcher is sure that there are more than one million running words or tokens at the disposal of lexicographers to compile dictionaries for Yipunu. The size of a corpus should not matter as such. Even if the corpus is not as large as one could expect the lexicographer should not hesitate to commence the compilation of any dictionary because even pocket dictionaries can be a starting point for a language that does not have a larger amount of language material. Furthermore, it is advisable that the priority be to start the compilation of dictionaries that consume less time and finances. The researcher will be totally satisfied when the SYC will be used to realise the first Yipunu dictionary and also if this model can motivate the development of corpora for the rest of the Gabonese languages. The remaining written and oral materials need to be transcribed and saved in a computer database in the manner explained in Chapter 3.

The researcher recommends the following:

- The establishment of multitask national lexicographic units for each Gabonese language.
- The gathering of material in all the Gabonese languages.
- The establishment of a language body for each speech community.
- The gathering of additional material to enlarge the SYC.
- The involvement of the different speech communities for the successful completion of all the projects.
- The equal release of funds by the government to avoid imbalanced progression of the different projects.
- The establishment of a clear language policy that will help to promote local languages by giving them an official status together with French.

- The adoption of a writing system for each selected official language.
- The introduction of national languages into the educational system to give meaning to all projects.

The results of the research presented in this dissertation are essentially applicable to the compilation of printed dictionaries but some of the principles are common and can rightfully be used for the conceptualisation of electronic dictionaries. The researcher did not elaborate much with regard to the development of electronic dictionaries. This is probably because the corpus is not tagged and at this stage, it offers some basic data more suitable for printed dictionaries. This is why the next stage of this research could be to tag the corpus in order to enable sophisticated manipulations so that all data types are retrieved automatically and lexical items are sorted into different parts of speech together with example sentences. However, in order to be able to use the corpus for the compilation of electronic dictionaries, the researcher has to overcome many technical challenges. One of them would be to have a computer skilled staff member(s) that could assist in the conceptualisation of complicated issues such as spellcheckers, n-gram analysis, lexical and error recall, text to speech, etc. For future applications of Yipunu corpora, these aspects of electronic dictionaries need to be addressed carefully. In this research with our limited knowledge on these issues, we are attempting to give some indication and analysis of what could be done in future if one plans any compilation of electronic dictionary in Yipunu. The objective here is only to indicate the future research venues that could be followed.

One has to investigate and implement the different techniques and muster the terminology useful for the compilation of electronic dictionaries in the Gabonese languages. The compilation process should integrate international standard and requirements. We believe that after assigning a label to each word and lexical item in the corpus manually or automatically, the corpus can be manipulated in a way that will allow N-gram analysis, in other words breaking up words into sections of permissible or impermissible combinations. The N-gram analysis of the Yipunu corpus could allow the researcher to segment the number of sequences of letters that exist in the language. This is important, alongside word lists and morphological boxes, for the construction of spellcheckers. The division of words into blocks of three letters would help the spellchecker to recognise the permissible combinations and enable it to point out the impermissible ones.

If the SYC is annotated it can enable the construction of a mouse-language, which is a cursor-based dictionary that automatically opens links to synonyms for the monolingual and bilingual dictionaries when one clicks on a lexical item. In other words, in a multilingual and multicultural environment where different languages are equipped with comprehensive databases, the lemma lists of each language can be used for the compilation of a cursor-based dictionary. This will be possible if routes are opened between the different lexical items of each source language and the corresponding translation equivalent(s) of the target language(s).

In the same way, the translation equivalent of a lexical item can be provided in different languages if previously related to the main database. In other words, for the dictionary in use is Yipunu, if one would like to automatically translate an item into one of the other Gabonese languages, one has just to click on the item pointed at by the cursor. The programming system will have to be designed in such a way that a direct link to these languages is provided. The lexicographer will then only have to choose the language of which the translation is needed. This will be the ideal development for multilingual countries such as Gabon, where the compilation of printed bilingual dictionaries for the existing languages could be difficult to achieve.

Corpora in the Gabonese languages will have to be developed and tagged in order to allow the following, among others, as mentioned by Prinsloo (2001) and De Schryver (2003):

- pop-up access
- bringing together related items
- new routes to the data
- less dependency on alphabetical order
- fuzzy spelling
- intelligent extrapolation of characters keyed in
- audible pronunciation

One should investigate the way the SYC could be tagged in the following ways:

- speech to speech
- text to speech

- speech to text

Morphological to syntactic parsing must be done on a single part of speech to a sentence level. It can happen that one has to do morphological and syntactical parsing before annotating the part of speech in a conjunctively written language. In fact, the lexicographer will have to opt for the most convenient method in a manner that will enable the SYC to be manipulated and effectively used for the compilation of electronic dictionary.

For machine translation, one has to take into account all the rules of the two languages involved by syntactically parsing as follows:

- text to text
- text to speech
- speech to text

The difference between subject concord and object concord has to be considered and one has to adapt existing programmes to design new ones. These are very important steps to take in order to fulfil the international aim of being present in a global world and in communication with other countries or languages. Electronic dictionaries are the target of the lexicographic units of most languages, maybe because they involve less time and finances or perhaps they are easier to compile and very sophisticated. In order for such projects to take place effectively in the Gabonese languages, the emphasis should be placed on the development of annotated corpora.

## BIBLIOGRAPHY

- Aarts, J. 1991. Intuition-based and observation-based grammars, in K. Aijmer and B. Altenberg (eds.). *English corpus linguistics: Studies in honour of Jan Svartvik*. London: Longman. 44–62.
- Aarts, J. & Meijs, W. (eds.). 1986. *Corpus linguistics II*. Amsterdam: Rodopi.
- Aijmer, K. & Altenberg, B. (eds.). 1991. *English corpus linguistics: Studies in honour of Jan Svartvik*. London: Longman.
- Al-Kasimi, A. 1977. *Linguistics and bilingual dictionaries*. Leiden: E.J. Brill.
- Altenberg, B. 1984. Causal linking in spoken and written English. *Studia Linguistica*, 38:20–69.
- Antaki, C. & Naji, S. 1987. Events explained in conversational “because” statements. *British Journal of Social Psychology*, 26:119–126.
- Atkins, B.T.S. (ed.). 1994. *Computational approaches to the lexicon*. Oxford: Oxford University Press.
- Atkins, B.T.S., Clear, J. & Ostler, N. 1992. Corpus design criteria. *Literary and Linguistic Computing*, 7(1):1–16.
- Atkins, B.T.S. & Levin, B. 1995. Building on a corpus: A linguistic and lexicographical look at some near-synonyms. *International Journal of Lexicography*, 8(2):85–114.
- Atkins, B.T.S., Rundell, M. & Weiner, E. 1997. *Salex '97: A training course in the compilation of monolingual dictionaries*. Unpublished course material of a tutorial held at the Dictionary Unit for South African English, Rhodes University, Grahamstown, 15–26 September 1997.
- Austin, J.L. 1962. *How to do things with words?* New York: Oxford University Press.
- Béjoint, H. 1989. Codeness and lexicography, in G.C. James (ed.). *Lexicographers and their works*. Exeter: University of Exeter. 1–4.
- Bell, A. 1984. Language style as audience design. *Language in Society*, 13(2):145–204.

- Bergenholtz, H. & Schaefer, B. 1978. Ausblicke auf eine deskriptive Lexicographie, in H. Henne *et al.* (eds.). *Interdisziplinäres Wörterbuch in der Diskussion*. Düsseldorf: Pädagogischer Verlag Schwann.
- Bergenholtz, H. & Tarp, S. 2002. Die moderne lexicographische Funktionslehre. Diskussionsbeitrag zu neuen und alten Paradigmen, die Wörterbücher als Gebrauchsgegenstände verstehen. *Lexicographica*, 18:253–263.
- Bergenholtz, H., Tarp, S. & Wiegand, H.E. 1999. Datendistributionsstrukturen, Makro- und Mikrostrukturen in neueren Fachwörterbüchern, in L. Hoffmann *et al.* (eds.). *Fachsprachen. Languages for special purposes: An international handbook of special-language and terminology research*. Berlin: De Gruyter. 1762–1832.
- Bessieux, J.R. 1847 *Dictionnaire francais-mpongoué/mpongoué-francais*. 2 volumes. Amiens: Lenoël-Hérouart. 52 pages
- Brown, P. & Levinson, S. 1978. Universals in language usage: Politeness phenomena, in E. Goody (ed.). *Questions and politeness: Strategies in social interaction*. Cambridge: Cambridge University Press. 56–311.
- Bureau of the WAT. 2001. *Training course in lexicography of the Woordeboek van die Afrikaanse Taal (WAT)*. Stellenbosch: Bureau of the WAT.
- Calzolari, N. *et al.* 1987. The use of computers in lexicography and lexicology, in A. Cowie (ed.). *The dictionary and the language learner*. Papers from the Euralex seminar at the University of Leeds, 1–3 April 1985. *Lexicographica Series Maior* 17. Tübingen: Niemeyer. 55–77.
- Carpentier de Changy, H. & Voltz, M. 1990. Alphabet scientifique du Gabon (ASG): Liste alphabétique. *Revue Gabonaise des Sciences de l'Homme*, 2:113–115.
- Čermák, F. 2003. Source materials for dictionaries, in P. van Sterkenburg (ed.). *A practical guide to lexicography*. Amsterdam/Philadelphia: John Benjamins 19-25.
- Church, K.W. & Hanks, P. 1989. *Word association norms, mutual information and lexicography*. Proceedings of the 27<sup>th</sup> Annual Meeting of the Association

- for Computational Linguistics. Reprinted in *Computational Linguistics*, 16:1, 1990.
- Clear, J. 1987. Computing: Overview of the role of computing in Cobuild, in J.M. Sinclair (ed.). *Looking up: An account of the COBUILD Project in lexical computing and the development of the Collins COBUILD English Language Dictionary*. London/Glasgow: Collins ELT.
- Clyne, M. (ed.). 1997. *Undoing and redoing corpus planning*. Berlin/New York.
- Cole, R., Mariani, J., Uszkoreit, H. & Battista Varile, G. (eds.). 1998. *Survey of the state of the art in human language technology* (Studies in Natural Language Processing). Pisa: Cambridge University Press/Giardini Editori.
- De Schryver, G-M & D.J Prinsloo. 2000. Electronic corpora as a basis for the compilation of African-language dictionaries, Part 1: The *macrostructure*. *South African journal of African Languages* 20/4: 291-309.
- De Schryver, G-M & D.J. Prinsloo. 2000a. Electronic corpora as a basis for the compilation of African-language dictionaries, Part 2: The *microstructure*. *South African journal of African Languages* 20/4: 310-330.
- Doneux, J.L. 1967. *Questionnaires d'enquêtes linguistiques (Greenberg-Tervuren-Welmers)*. Université de Dakar.
- Eichhoff, J. 1982. Erhebung von Sprachdaten durch schriftliche Befragung, in W Besch et al. (Hrsg.): *Dialektologie. Ein Handbunch zur deutschen und allgemeinen Dialektforschung*. Band 1.1. Berlin : Walter de Gruyter, 549-554.
- Eglise Evangélique du Sud Gabon. 1966. *Dictionnaire Français-Yipounou/Yipounou-Français*. Mouila: CMA.
- Fail, L. 2004. *Corpus linguistics (2): The corpus approach* [Online]. Available: <http://www.proz.com/howto//174>.
- Fasold, W.R. 1990. *The sociolinguistics of language: Introduction to sociolinguistics*. Basil Blackwell.
- Ferreira, F. & Anes, M. 1994. Why study spoken language?, in A.M.Gernsbacher (ed.). *Handbook of psycholinguistics*. Academic Press.

- Finegan, E. & Biber, D. 1994. Register and social dialect variation: An integrated approach, in D. Biber & F. Finegan (eds.). *Sociolinguistic perspectives on register*. New York: Oxford. 315–347.
- Furfey, P.H. 1944. Men's and women's language. *The American Catholic Sociological Review*, 5:218–223.
- Galley, S. 1964. *Dictionnaire fang-français et français-fang : Suivi d'une grammaire fang*. Neuchâtel: Henri Messeiller.
- Garnham, A., Shillock, R., Brown, G., Mill, A. & Cutler, A. 1981. Slips of the tongue in the London-Lund corpus of spontaneous conversation. *Linguistics*, 19:805–817.
- Gautier, G. 1998. *Preliminary reflections for the constitution of a national corpus of Kurdish language building a Kurdish language corpus: An overview of the technical problems*. Paper presented at the Sixth International Conference and Exhibition on Multilingual Computing (ICEMCO 98) held by the Centre for Kurdish Studies, Cambridge, UK, April 1998.
- Geeraerts, D. 1984. Dictionary classification and the foundation of lexicography. *ITL Review*, 63:37–63.
- Geeraerts, D. & Janssens, G. 1982. *Wegwijs in woordenboeken: Een kritisch overzicht van de lexicografie*.
- Gouws, R.H. 1989. *Leksikografie*. Pretoria: Academica.
- Gouws, R.H. 1991. Towards a lexicon-based lexicography. *Dictionaries*, 13:75–90.
- Gouws, R.H. 1996. Bilingual dictionaries and communicative equivalent for a multilingual society. *Lexikos*, 6:14–31.
- Gouws, R.H. 1999. A theoretically motivated model for the lexicographic processes of the National Lexicographic Units. Unpublished report submitted to the Pan South African Language Board.
- Gouws, R.H. 2000. Strategies in equivalent discrimination, in A. Zettersten *et al.* (eds.). *Symposium on Lexicography IX*. Tübingen: Max Niemeyer: 99–111.



- Gouws, R.H. 2001. Lexicographic training: Approaches and topics, in J.D. Emejulu (ed.). *Eléments de lexicographie gabonaise*. Tome 1. New York: Jimacs-Hillman Publishers. 58–94.
- Gouws, R.H. 2002. Using a frame structure to accommodate cultural data, in J.D. Emejulu (ed.). *Eléments de lexicographie gabonaise*. Tome 2. New York. Jimacs-Hillman Publishers. 54–69.
- Gouws, R.H. 2003. Aspekte van mikrostrukturele verskeidenheid en inkonsekwentheid in woordeboeke. *Lexikos*, 13:92–110.
- Gouws, R.H. 2004. Outer texts in bilingual dictionaries. *Lexikos*, 14:264–274.
- Gouws, R.H. & Prinsloo, D.J. 2005. *Principles and practice of South African lexicography*. Stellenbosch: SUN Press.
- Gove, P.B. 1961. Letter to the editor. *Life Magazine*, 17 November:13.
- Greenberg, J. 1955. *Studies in African linguistic classification*. New Haven: Compass Publishing Company.
- Grice, H.P. 1975. Logic and conversation, in P. Cole and J. Morgan (eds.). *Syntax and semantics 3: Speech acts*. New York: Academic Press.
- Guthrie, M. 1948. *The classification of the Bantu languages*. London: Dawsons.
- Guthrie, M. 1953. *The Bantu languages of western equatorial Africa*. Oxford University Press.
- Halliday, M. 1991. Corpus studies and probabilistic grammar, in Aijmer and Altenberg (eds.). 30–43.
- Hanks, P. 2004. *Corpus pattern analysis*. Proceedings of the 11<sup>th</sup> Euralex International Congress, Le Paquebot/Lorient/France, 6–10 July 2004, Volume 1.
- Hartmann, R.R.K. (ed.). 1984. *LEX'eter'83*. Papers from the International Conference on Lexicography at Exter, 9–12 September 1983.
- Hartmann, R.R.K. & James, G. 1998. *Dictionary of lexicography*. London: Routledge.
- Hausmann, F.J. 1977. *Einführung in die Benutzung des neufranzösischen Wörterbücher*. Tübingen: Niemeyer.

- Hausmann, F.J. 1986. Grundprobleme des zweisprachigen Wörterbuchs. *Fremdsprachen Lehren und Lernen*, 23:206–220.
- Hausmann, F.J. 1989. Wörterbuchtypologie, in F.J. Hausmann *et al.* (eds.). 968–981.
- Hausmann, F.J. & Wiegand, H.E. 1989. Components parts and structures of general monolingual dictionaries: A survey, in F.J. Hausmann *et al.* (eds.) 328–360.
- Hess, H. *et al.* 1983. *Maschinenlesbare Deutsche Wörterbücher: Dokumentation, Vergleich, Integration*. Tübingen: Niemeyer.
- Hofland, K. & Johansson, S. 1982. *Word frequencies in British and American English*. Bergen: Norwegian Computing Centre for the Humanities.
- Holmes, J. 1988. Doubt and certainty in ESL textbooks. *Applied Linguistics*, 9:21–44.
- Holmes, J. 1994. Inferring language change from computer corpora: Some methodological problems. *ICAME Journal*, 18:27–40.
- Hombert, J.M. 1990. Présentation de l’alphabet scientifique des langues du Gabon. *Revue Gabonaise des Sciences de l’homme*, 2:105–111.
- Hori, M. 2004. *Investigating Dickens’s style : A collocational analysis*. Basingstoke: Palgrave Macmillan.
- Hubert, J. 1995. *Rapidolangue Niveau 1: Méthode d’apprentissage des langues du Gabon*. Volume 1 (Fang, Inzebi, Lembaama, Omyene, Yipunu). Libreville: Editions Raponda-Walker.
- Idiata, D.F. 2002. *Il était une fois les langues gabonaises*. Libreville: Editions Raponda-Walker.
- Jaquot, A. 1978. Le Gabon, in D. Barreteau (ed.). *Inventaires des études linguistiques sur les pays d’Afrique noire d’expression française et sur Madagascar*. Paris: CLIF. 493–503.
- Johansson, S. & Norheim, E. 1988. The subjunctive in British and American English. *ICAME Journal*, 12:27–36.
- Johansson, S. & Stenström, A-B. (eds.). 1991. *English computer corpora: Selected papers and research guide*. Berlin: Mouton de Gruyter.

- Kammerer, M. & Wiegand, H.E. 1998. Über die textuelle Rahmenstruktur von Printwörterbüchern: Präzisierungen und weiterführende Überlegungen. *Lexicographica*, 14:224–237.
- Kasper, G. 1995. Interlanguage pragmatics, in J. Verschueren, J-O Östman and J. Blommaert (eds.). *Handbook of pragmatics*. Amsterdam: John Benjamins. 1–7.
- Kasper, G. & Blum-Kulka, S. (eds.) .1993. *Interlanguage pragmatics*. Oxford: Oxford University Press.
- Kennedy, G. 1987a. Expressing temporal frequency in academic English. *TESOL Quarterly*, 21:69–86.
- Kennedy, G. 1987b. Quantification and the use of English: A case study of one aspect of the learner’s task. *Applied Linguistics*, 8:264–286.
- Kennedy, G. 1998. An introduction to corpus linguistics. London/New York: Longman.
- Kilgarriff, A. & Tugwell, D. 2001. *WASP-bench: An MT lexicographers’ workstation supporting state-of-the-art lexical disambiguation*. Proceedings of MT Summit VII, Santiago de Compostela.
- Kirk, J. 1994. Teaching and language corpora: The Queen’s approach, in Wilson, A. and McEnery, A. (eds.). *Corpora in language education and research: A selection of papers from TALC94*. Lancaster: Unit for Computer Research on the English Language. Technical Papers 4: 29–51.
- Kjellmer, G. 1986. The “lesser man”: Observations on the role of women in modern English writings, in Aarts and W. Meijs (eds.). *Corpus linguistics II*. Amsterdam: Rodopi. 163–76.
- Kromann, H.P., Riiber, T. & Rosbach, P. 1984a. Active and passive bilingual dictionaries: The Scerba concept reconsidered, in R.R.K. Hartmann (ed.). *LEX’eter ’83 proceedings*. Tübingen: Max Niemeyer. 207–215.
- Kromann, H.P., Riiber, T. & Rosbach, P. 1984b. Überlegungen zu Grundfragen der zweisprachigen Lexicographie, in H.E. Wiegand (ed.). *Studien zur neuhochdeutschen Lexicographie*. Hildesheim: Georg Olms Verlag. 159–238.

- Kukenheim, L. 1960. Van glossarium tot thesaurus: Lexicologische theorieën en lexicographische realiteiten. *Levende Talen*, 203:15–34.
- Kwenzi Mikala, J.T. 1980. Esquisse phonologique du punu, in F. Nsuka-Nkutsi (ed.). *Éléments de description du punu*. Travaux du Centre de Recherches Linguistiques et Sémiotiques. Lyon: Université Lumière de Lyon. 7–18.
- Kwenzi Mikala, J.T. 1987. Contribution à l’inventaire des parlers bantu du Gabon. *Pholia*, 2:103–110.
- Kwenzi Mikala, J.T. 1988. L’identification des unités-langues bantu gabonaises et leur classification interne. *Muntu*, 8:54–64.
- Kwenzi Mikala, J.T. 1997a. *Mumbwanga: Récit épique*. Libreville: Editions Raponda-Walker.
- Kwenzi Mikala, J.T. 1997b. *Parémies d’Afrique centrale: Proverbes et sentences*. Libreville: Editions Raponda-Walker.
- Kwenzi Mikala, J.T. 1998. Parlers du Gabon: Classification du 11-12-97, in A. Raponda-Walker (ed.). *Les langues du Gabon*. Libreville: Editions Raponda-Walker. 217–221.
- Kytö, M., Rissanen, M. & Wright, S. (eds.). 1994. *Corpora across the centuries*. Amsterdam: Rodopi.
- Labov, W. 1966. *The social stratification of English in New York City*. Washington, DC: Centre for Applied Linguistics.
- Lalljee, M., Watson, M. & White, P. 1983. Some aspects of explanations of young children, in J. Jaspars, F. Finschman and M. Hewstone (eds.). *Attributions theory and research: Conceptual, development and social dimensions*. London/New York: Academic Press.
- Landau, S.I. 1984. *Dictionaries: The art and craft of lexicography*. Cambridge: Cambridge University Press.
- Landau, S.I. 2001. *Dictionaries: The art and craft of lexicography*. 2<sup>nd</sup> edition. Cambridge: Cambridge University Press.
- Laver, J. 2006. Speech, in K. Brown (ed.). *The encyclopedia of language and linguistics*. Volume 11. Oxford: Elsevier. 636–647.

- Leech, G. & Fallon, R. 1992. Computer corpora – what do they tell us about culture? *ICAME Journal*, 16:29–50.
- Leitner, G. 1990. Corpus design: Problems and suggested solutions. *ICE Newsletter* 7. London: University College.
- Leitner, G. 1991. The Kolhapur corpus of Indian English: Intravarietal description and/or intervarietal comparison, in S. Johansson and A-B. Stenström (eds.). *English computer corpora: Selected papers and research guide*. Berlin: Mouton de Gruyter. 215–32.
- Macaulay, R.K. 1991. *Locating dialects in discourse: The language of honest men and bonnie lasses in Ayr*. Oxford: Oxford University Press.
- Makkai, A. 1971. Theoretical and practical aspects of an associative lexicon for 20<sup>th</sup> century English, in L. Zgusta (ed.). *Manual of lexicography*. The Hague: Mouton.
- Mahlberg, M. 2005. *Dickensian patterns: Meaning and form in literacy text*. Abstracts of Presentation for Corpus Approaches to Literature, pre-conference workshop at CL2005.
- Malkiel, Y. 1962. A typological classification of dictionaries on the basis of distinctive features, in Housholder, W. and Saporta, S. (eds.). *Problems in lexicography*. Bloomington, IN: Indiana University Press 3-24.
- Martin, W. 1996. Lexicographical resources in a multilingual environment: An orientation. *Lexikos*, 6:199–214.
- Mayer, R. 1989. Histoire de l'écriture des langues du Gabon. *Revue Gabonaise des Sciences de l'Homme*, 2:65–92.
- McEnery, A., Baker, P. & Wilson, A. 1995. A statistical analysis of corpus-based computer vs. traditional human teaching methods of part of speech analysis. *Computer-Assisted Language Learning*, 8(2/3): 259–74.
- McEnery, A. & Wilson, A. 1993. The role of corpora in computer-assisted language learning. *Computer-Assisted Language Learning*, 6(3):233–48.
- Medjo-Mvé, P. 1997a. *Dialectologie fang*. Séminaire de Dialectologie held by LASCIDYL de l'ENS, Libreville, 1–6 December 1997.

- Medjo-Mvé, P. 1997b. *Essai sur la phonologie panchronique des parlers fang du Gabon et ses implications historiques*. PhD dissertation. Lyon: Université Lumière (Lyon II).
- Medjo-Mvé, P. 1997c. Interaction ton et quantité vocalique dans le parler fang de la région de Cocobeach. *iBoogha*, 1:151–165. Libreville: Les Editions du Silence.
- Meijs, W. (ed.). 1987. *Corpus linguistics and beyond*. Amsterdam: Rodopi.
- Menge, H.H. 1982. Erhebung von Sprachdaten in ‘künstlicher’ Sprechsituation (Experiment und Test), in W. Besch *et al.* (eds.). *Dialektologie : Ein Handbuch zur Deutschen und allgemeinen Dialektforschung*. Berlin: Walter de Gruyter. 195–231.
- Mesthrie, R., Swann, J., Deumert, A. & Leap, W.L. 2000. *Introducing sociolinguistics*. Edinburgh University Press.
- Mindt, D. 1991. Syntactic evidence for semantic distinctions in English, in K. Aijmer and B. Altenberg (eds.). *English corpus linguistics: Studies in honour of Jan Svartvik*. London: Longman. 182–196.
- Mindt, D. 1992. *Zeitbezug im Englischen: Eine didaktische Grammatik des Englischen Futures*. Tübingen: Gunter Narr.
- Mutaki, N.N. & Tamanji, P.N. 2000. *An introduction to African linguistics*. LINCOM EUROPA.
- Myers, G. 1991. *Pragmatics and corpora*. Paper presented at Corpus Linguistics Research Group, Lancaster University.
- Ndinga-Koumba-Binza, H.S. (forthcoming). Alphabet et écriture: Approche historique et cas des langues du Gabon, in J. Hubert and P.A. Mavoungou (eds.). *Ecriture et standardisation des langues gabonaises*. Stellenbosch: SUN Press.
- Nielsen, S. 1995. Alphabetic macrostructure, in H. Bergenholtz and S. Tarp (eds.). *Manual of specialized lexicography*. Amsterdam: John Benjamins. 190–195.

- Nong, S., De Schryver, G.M. & Prinsloo, D.J. 2002. Loanwords versus indigenous words in Northern Sotho: A lexicographic perspective. *Lexikos*, 12:1–20.
- Nza, M. 2005. *Pratiques culturelles au village*. Libreville: Editions Raponda-Walker.
- O'Connor, J. & Arnold, G. 1961. *Intonation of colloquial English*. London: Longman.
- Oostdijk, N. & De Haan, P. 1994a. Clause patterns in modern British English: A corpus-based (quantitative) study. *ICAME Journal*, 18:41–79.
- Oostdijk, N. & De Haan, P. (eds.). 1994b. *Corpus-based research into language*. Amsterdam: Rodopi.
- Prinsloo, D.J. 1992. Towards computer-assisted word frequency studies in Northern Sotho. *South African Journal of African Languages*, 11(2):54–60.
- Prinsloo, D.J. 1994. Lemmatization of verbs in Northern Sotho. *South African Journal of African Languages*, 14(2):93–102.
- Prinsloo, D.J. 2004. Revising Matumo's Setswana-English-Setswana dictionary. *Lexikos*, 14:158–172.
- Prinsloo, D.J. & De Schryver, G-M. 2000. *The concept of simultaneous feedback as applied to the South African context, with special reference to the Sepedi Dictionary Project (SeDiPro)*. Fifth International Conference of the African Association for Lexicography held at Stellenbosch University, South Africa.
- Prinsloo, D.J. 2001. The compilation of Electronic Dictionaries for the African Languages. *Lexikos* 11: 139-159.
- Prinsloo, D.J. & De Schryver, G-M. 2001. Monitoring the stability of a growing organic corpus, with special reference to Sepedi and Xitsonga. *Dictionaries*, 22:85–129.
- Prinsloo, D.J. & De Schryver, G-M. 2002. Designing a measurement instrument for the relative length of alphabetic stretches in dictionaries, with special reference to Afrikaans and English, in A. Braasch and C. Povlsen (eds.). *Proceedings of the 10th EURALEX International Congress*. Copenhagen: Center for Sprogteknologi, Københavns Universitet. 483–494.

- Prinsloo, D.J. & De Schryver, G-M. 2005. Managing eleven parallel corpora and the extraction of data in all South African languages, in W. Daelemans, T. du Plessis, C. Snyman and L. Teck (eds.). *Multilingualism and electronic language management*. Pretoria: J.L. van Schaik. 100–122.
- Prinsloo, D.J. & Gouws, R.H. 1996. Formulating a new dictionary convention for the lemmatization of verbs in Northern Sotho. *South African Journal of African Languages*, 16(3):100–107.
- Prinsloo, D.J. & Gouws, R.H. 2000. The use of examples in polyfunctional dictionaries. *Lexikos*, 10:138–156.
- Quirk, R., Greenbaum, S., Leech, G. & Svartvik, J. 1985. *A comprehensive grammar of the English language*. London: Longman.
- Raponda-Walker, A. 1932. L'alphabet des idiomes gabonais. *Journal de la Société des Africanistes*, 2(2):139–146. Reprinted in Raponda-Walker (ed.). 1998:7–15.
- Raponda-Walker, A. 1960. *Notes d'histoire du Gabon*. Memoire IEC9.
- Raponda-Walker, A. 1995. *Dictionnaire mpongwè-français*. Paris/Libreville: Classiques Africains/Editions Raponda-Walker.
- Renouf, A. 1987. Corpus development, in J.M. Sinclair (ed.). *Looking up: An account of the COBUILD Project in lexical computing and the development of the Collins COBUILD English Language Dictionary*. London/Glasgow: Collins ELT. 1–40.
- Rey, A. 1970. Typologie génétiques des dictionnaires, in Rey-Debove (ed.). *Languages*, 19:48–68.
- Rickford, J.R. & McNair-Knox, F. 1994. Addressee- and topic-influenced style shift: A quantitative sociolinguistic study, in D. Biber and F. Finegan (eds.). *Sociolinguistic perspectives on register*. New York: Oxford. 235–276.
- Rissanen, M. 1989. Three problems connected with the use of diachronic corpora. *ICAME Journal*, 13:16–19.
- Rissanen, M., Kytö, M. & Palander-Collin, M. (eds.). 1993. *Early English in the computer age*. Berlin: Mouton de Gruyter.



- Rittaud-Hutinet, C. 1980. Lexique du punu, in F. Nsuka-Nkutsi (ed.). *Eléments de description du punu*. Travaux du Centre de Recherches Linguistiques et Sémiotiques. Lyon: Université Lumière de Lyon.
- Ronald, J. 2003. Words and phrases, in M. Stubbs (ed.). *Corpus studies of lexical semantics*. Oxford: Blackwell Publishers.
- Roux, J.C. & Bosch, S.E. 2002. *Human language technologies and the national lexicography units*. Paper presented at the Seventh International Conference of the African Association for Lexicography held at Rhodes University, Grahamstown, 8–10 July 2002. Abstract contained in G-M de Schryver (ed.), 2002:26–27. Pretoria: (SF) Press.
- Saeed, J.I. 2003. *Semantics*. Blackwell Publishing.
- Schaeder, B. 1979. Maschinenlesbare Textkorpora des Deutschen und des Englischen, in H. Bergenholtz and B. Schaeder (eds.). *Empirische Textwissenschaft: Aufbau und Auswertung von Text-Corpora*. Königstein. 356–370.
- Schmied, J. 1993. Qualitative and quantitative research approaches to English relative constructions, in C. Souter and E. Atwell (eds.). *Corpus-based computational linguistics*. Amsterdam: Rodopi. 85–96.
- Schreuder, R. & Kerkman, H. 1987. On the use of a lexical database in psycholinguistic research, in Meijs (ed). *Corpus linguistics and beyond*. Proceedings of the Seventh International Conference on English Language Research on Computerized Corpora. Amsterdam: Rodopi. 295–302.
- Scott, M. 2000. Home page [Online]. Available: <http://www.liv.ac.uk/~ms2928/wordsmith/screeshots>>.
- Searle, J. 1969. *Speech acts*. Cambridge: Cambridge University Press.
- Sebeok, T.A. 1962. Materials for a typology of dictionaries. *Lingua*, 11:363–374.
- Sebeok, T.A. (ed.). 1971. *Linguistics in sub-Saharan Africa*. CTL 7. The Hague: Mouton.
- Semino, E. & Short, M. 2004. *Corpus stylistics: Speech, writing and thought presentation in a corpus of English writing*. London: Routledge.

- Sinclair, J. 1996. *Preliminary recommendations on corpus typology*. EAGLES Document EAG--TCWG--CTYP/P, May, 1996.
- Sinclair, J.M. 1987. The nature of the evidence, in J.M. Sinclair (ed.). *Looking up: An account of the COBUILD Project in lexical computing and the development of the Collins COBUILD English Language Dictionary*. London/Glasgow: Collins ELT.
- Sinclair, J.M. 2003. Corpora for dictionaries, in P. van Sterkenburg (ed.). *A practical guide to lexicography*. Amsterdam/Philadelphia: John Benjamins.
- Smit, M. 1996. *Wiegand's metalexigraphy as a framework for a multilingual multicultural explanatory music education dictionary for South Africa*. Unpublished doctoral dissertation. Stellenbosch, Stellenbosch University.
- Soami, L.S. 2000. *Les prénoms à l'HPO: Étude sociolinguistique*. Unpublished MA thesis. Libreville, Omar Bongo University.
- Soami, L.S. (forthcoming). Orthographe du yipunu, in P.A. Mavoungou and H. Guerinaud (eds.). *Ecriture et standardisation des langues gabonaises*. Stellenbosch: SUN Press.
- Sperber, D. & Wilson, D. 1986. *Relevance: Communication and cognition*. Oxford: Blackwell.
- Stenstöm, A-B. 1984. *Discourse items and pauses*. Paper presented at the Fifth ICAME Conference, Windermere. Abstract in *ICAME News*, 9:11.
- Stenstöm, A-B. 1987. Carry-on signals in English conversation, in W. Meijs (ed.). *Corpus linguistics and beyond*. Amsterdam: Rodopi. 87–119.
- Strassel, S. et al. 2003. *SLX corpus of classical sociolinguistics interviews*. Philadelphia: Linguistic Data Consortium.
- Stubbs, M. 2001. *Words and phrases: Corpus studies of lexical semantics*. Oxford: Blackwell.
- Svensén, B. 1993. *Practical lexicography: Principles and methods of dictionary-making*. Oxford/New York: Oxford University Press.

- Swanepoel, P. 2003. Dictionary typology: A pragmatic approach, in P. van Sterkenburg (ed.). *A practical guide to lexicography*. Amsterdam/Philadelphia: John Benjamins. 44–69.
- Tarp, S. 1994. Funktionen in Fachwörterbüchern, in B. Schaeder and H. Bergenholtz (eds.). *Das Fachwörterbuch: Fachwissen und seine Repräsentation in Wörterbüchern*. Tübingen: Narr. 229–246.
- Tarp, S. 2000. Theoretical changes to LSP lexicography. *Lexikos*, 10:189–208.
- Tarp, S. 2002a. Functions in De Gruyter Wörterbuch Deutsch als Fremdsprache, in H.E. Wiegand (ed.). *Perspektiven der pädagogischen Lexicography des Deutschen II: Untersuchungen anhand des De Gruyter Wörterbuch Deutsch als Fremdsprache*. Tübingen: Max Niemeyer Verlag. 609–619.
- Tarp, S. 2002b. Translation dictionaries and bilingual dictionaries – two different concepts. *Journal of Translation Studies*, 7:59–84.
- Tarp, S. 2004. Reflections on dictionaries designed to assist the users with text production in a foreign language. *Lexikos*, 14:299–325.
- Tarp, S. & Gouws, R.H. 2004. Wie leer wat uit Afrikaanse (aan)leerderwoordeboeke? *Tydskrif vir Geesteswetenskappe*, 44 (4):276–298.
- Tottie, G. 1991. *Negation in English speech and writing: A study in variation*. San Diego: Academic Press.
- Wells, R.A. 1973. *Dictionaries and the authorisation tradition: Studies in English usage and lexicography*. Berlin: Mouton de Gruyter.
- Westermann, D. 1911. *Die Sundansprachen*. Hamburg.
- Wiegand, H.E. 1984a. Germanistische Wörterbuchforschung nach 1945: Eine einführende Übersicht für Deutschlehrer. *Der Deutschunterricht*, 36(5):10–26.
- Wiegand, H.E. 1984b. On the structure and contents of a general theory of lexicography, in R.R.K. Hartmann (ed.). *LEX'eter '83 proceedings*. Tübingen: Max Niemeyer. 13–30.
- Wiegand, H.E. 1984c. Prinzipien und Methoden historischer Lexikographie, in W. Besch et al. (eds.). *Sprachgeschichte: Ein Handbuch zur geschichte der*

- deutschen Sprache und ihrer Erforschung*. Berlin: Mouton de Gruyter. 557–620.
- Wiegand, H.E. 1989a. Arten von Mikrostrukturen im allemeinen einsprachigen Wörterbuch, in F.J. Hausmann *et al.* (eds.). *Wörterbücher. Dictionaries. Dictionnaires. Ein internationales Handbuch zur Lexikographie*. Berlin: Mouton de Gruyter. 462–501.
- Wiegand, H.E. 1989b. Der Begriff der Mikrostrukturen: Geschichte, Probleme, Perspektiven, in F.J. Hausmann *et al.* (eds.) 1989–1991: 409–462.
- Wiegand, H.E. 1996a. Das Konzept der semiintegrierten Mikrostrukturen: Ein Beitrag zur Theorie zweisprachiger Printwörterbücher, in H.E. Wiegand (ed.). *Wörterbücher in der Diskussion II*. Tübingen: Max Niemeyer Verlag: 1–82.
- Wiegand, H.E. 1996b. Zur Einführung, in H.E. Wiegand (ed.) *Wörterbücher in der Diskussion II*. Tübingen: Max Niemeyer Verlag: vii–xv.
- Wiegand, H.E. 1998. *Wörterbuchforschung*. Berlin: Mouton De Gruyter.
- Wiegand, H.E. 2001. Was eigentlich sind Wörterbuchfunktionen? Kritische Anmerkungen zur neueren und neuesten Wörterbutchforschung. *Lexicographica*, 17:217–248.
- Williamson, K. 1971. The Benue-Congo languages and Ijo. CTL 7, T. Sebeok (ed). The Hague: Mouton.
- Wilson, A. 1992. The usage of since: A quantitative comparison of Augustan, modern British and modern Indian English. *Lancaster Papers in Linguistics* 80.
- Wodak, R. 1982. Erhebung von Sprachdaten in natürlicher oder smuliert-natürlicher Sprechsituation, in W. Besch *et al.* (eds.). *Dialektologie: Ein Handbuch zur deutschen und allgemeinen Dialektforschung*. Band 1.1. Berlin: Walter de Gruyter. 539–544.
- Wolski, W. 1989. Formen der Textverdichtung im allemeinen einsprachigen Wörterbuch, in F.J. Hausmann *et al.* (eds.) *Wörterbücher. Dictionaries. Dictionnaires. Ein Internationales Handbuch zur Lexikographie*. Berlin: Mouton de Gruyter. 956–967.

Zang-Bié, Y. 2002. Le corpus lexicographique dans les langues à tradition: Le cas du dialecte fang-mekè. *Lexikos*, 12:211–225.

Zgusta, L. (ed.). 1971. *Manual of lexicography*. The Hague: Mouton.

Zgusta, L. (ed.). 1980. *Theory and method in lexicography: Western and non-Western perspectives*. Columbia: Hornbeam Press.

[http://encarta.msn.com/encyclopedia\\_761565449/African\\_Languages.html](http://encarta.msn.com/encyclopedia_761565449/African_Languages.html)

<http://bowland-files.lancs.ac.uk/monkey/ihe/linguistics/corpus4/4fra1.htm>

<http://www.sal.tohoku.ac.jp/ling/corpus3/3qual.htm>

## APPENDIX: The top 3027 words extracted from the SYC

1	NA	27,647	2.06
2	VA	5,433	0.4
3	RIE	5,243	0.39
4	FUMU	4,918	0.37
5	MU	4,785	0.36
6	NSAMBI	3,299	0.25
7	I	2,943	0.22
8	NESI	2,810	0.21
9	BATU	2,206	0.16
10	ANDI	2,081	0.15
11	MUANA	2,010	0.15
12	KAGA	1,985	0.15
13	U	1,909	0.14
14	AVA	1,766	0.13
15	MUTU	1,705	0.13
16	TE	1,672	0.12
17	MBANA	1,624	0.12
18	YIRI	1,619	0.12
19	AGU	1,617	0.12
20	NANA	1,598	0.12
21	BISI	1,587	0.12
22	NDAGU	1,478	0.11
23	BAANA	1,450	0.11
24	BOTSU	1,435	0.11
25	AMI	1,429	0.11
26	YIARI	1,373	0.1
27	BULONGU	1,341	0.1
28	PA	1,336	0.1
29	TUMBA	1,324	0.1
30	YI	1,323	0.1
31	DIAMBU	1,297	0.1
32	MBARI	1,253	0.09
33	MOSI	1,245	0.09
34	BAISRAEL	1,242	0.09
35	DIBAALA	1,217	0.09
36	NE	1,139	0.08
37	NDERI	1,121	0.08
38	JULU	1,105	0.08
39	MBURA	1,096	0.08
40	A	1,090	0.08
41	GUSU	1,089	0.08
42	PAGU	1,058	0.08
43	BANDI	1,036	0.08
44	JANDI	998	0.07
45	YIKA	980	0.07
46	MENU	959	0.07
47	FU	950	0.07
48	GARI	947	0.07
49	MUNA	947	0.07
50	LA	940	0.07
51	TINDI	917	0.07
52	YILUMBU	912	0.07
53	AGUVU	901	0.07

54	AMA	867	0.06
55	NDE	864	0.06
56	JOGU	849	0.06
57	MOTSU	835	0.06
58	BIOTSU	828	0.06
59	DAVID	825	0.06
60	DIBANDU	792	0.06
61	ENI	787	0.06
62	TEMU	786	0.06
63	BABABAAL	773	0.06
	A		
64	MOISE	771	0.06
65	II	770	0.06
66	AYI	758	0.06
67	BA	745	0.06
68	MANGOLU	741	0.06
69	AMAVOSA	738	0.05
70	NO	737	0.05
71	O	725	0.05
72	JENU	698	0.05
73	TQR	686	0.05
74	MAMBU	674	0.05
75	TQC	671	0.05
76	BENI	648	0.05
77	KABOGU	648	0.05
78	AVU	629	0.05
79	WANDI	626	0.05
80	B	624	0.05
81	MUSIENGI	615	0.05
82	DEDI	606	0.05
83	TAJI	602	0.04
84	AJI	601	0.04
85	JERUSALE	600	0.04
	M		
86	QC	598	0.04
87	VANA	587	0.04
88	ENU	579	0.04
89	ADI	557	0.04
90	BOGU	555	0.04
91	TSI	550	0.04
92	ABI	548	0.04
93	BAGU	521	0.04
94	DIANDI	494	0.04
95	AMABA	487	0.04
96	NANDI	486	0.04
97	ISRAEL	484	0.04
98	BATAJI	481	0.04
99	YIMOSI	480	0.04
100	DIINA	475	0.04
101	MUGETU	471	0.04
102	BAMI	469	0.03
103	QR	460	0.03
104	MALONGU	459	0.03
105	KA	456	0.03
106	MAMBA	451	0.03
107	MBE	449	0.03

108 EGYPTE	446	0.03
109 MAGU	441	0.03
110 KAMA	432	0.03
111 BU	426	0.03
112 OGU	424	0.03
113 MABOTI	420	0.03
114 N	420	0.03
115 MA	418	0.03
116 MOGU	416	0.03
117 AA	415	0.03
118 VO	413	0.03
119 DIELA	412	0.03
120 BUTAMBA	405	0.03
121 MANDI	405	0.03
122 WISI	404	0.03
123 ABA	401	0.03
124 BAGORA	400	0.03
125 UPAGU	393	0.03
126 MUISI	389	0.03
127 MISIENGI	387	0.03
128 MURIMA	376	0.03
129 MIGAGA	375	0.03
130 YIANDI	375	0.03
131 BAPAGU	374	0.03
132 UBUEJI	369	0.03
133 NAGU	366	0.03
134 BENU	364	0.03
135 YIPAGU	360	0.03
136 BATEGULA	357	0.03
137 BAVU	356	0.03
138 KALALA	355	0.03
139 AMATSING ULA	353	0.03
140 YIGUMI	353	0.03
141 MIOGU	351	0.03
142 TSIRI	349	0.03
143 BUANDI	342	0.03
144 MUKONGU	341	0.03
145 YILIMA	341	0.03
146 POSNEGY	340	0.03
147 BAKAGA	339	0.03
148 MAGUGA	339	0.03
149 MAMI	338	0.03
150 PANGINI	336	0.03
151 Y	335	0.02
152 YISAMBUA LI	333	0.02
153 KABU	332	0.02
154 MABI	331	0.02
155 YIFUMBA	331	0.02
156 MUTUBU	329	0.02
157 ANANA	328	0.02
158 AMARUMA	326	0.02
159 NSANGU	326	0.02
160 DIWENDI	321	0.02
161 LEVI	321	0.02



162 LORA	321	0.02
163 YISALU	321	0.02
164 SAUL	320	0.02
165 TSIOTSU	318	0.02
166 DI	317	0.02
167 GU	317	0.02
168 YIENI	317	0.02
169 MUJI	315	0.02
170 NGUJI	315	0.02
171 FSWISS	312	0.02
172 AGUMABA	311	0.02
173 BIGUJI	308	0.02
174 BISALU	308	0.02
175 BAMABA	307	0.02
176 UBA	303	0.02
177 ETU	302	0.02
178 ABAMABA	299	0.02
179 NAMUNYI	299	0.02
180 ANYI	298	0.02
181 DIBI	296	0.02
182 JACOB	296	0.02
183 BIMA	293	0.02
184 BIVUNDA	293	0.02
185 BI	292	0.02
186 KUMU	292	0.02
187 ABAVU	291	0.02
188 DIVITA	291	0.02
189 NSILA	291	0.02
190 BIOGU	287	0.02
191 MANDAGU	287	0.02
192 MISU	286	0.02
193 BIANDI	285	0.02
194 MUVAGULI TSI	284	0.02
195 MUKUTA	282	0.02
196 BAPUELA	279	0.02
197 UOTSU	279	0.02
198 BEJI	276	0.02
199 MIINA	276	0.02
200 BASUSU	274	0.02
201 AARON	272	0.02
202 III	272	0.02
203 BAJUDA	266	0.02
204 DIVU	264	0.02
205 AMAVAGA	263	0.02
206 NYIVU	263	0.02
207 NSIMA	261	0.02
208 MUA	260	0.02
209 BABAALA	259	0.02
210 MULUMI	258	0.02
211 BIBULU	257	0.02
212 YIOTSU	257	0.02
213 JA	256	0.02
214 NAMI	256	0.02
215 YINYENSU LU	256	0.02

216	JETU	255	0.02
217	YISONYI	254	0.02
218	BABEJI	253	0.02
219	UMOSI	253	0.02
220	BASALITSI	250	0.02
221	DIMOSI	250	0.02
222	MATSI	250	0.02
223	NIL	250	0.02
224	YIBENGUN	245	0.02
	A		
225	DUTALU	244	0.02
226	KALA	243	0.02
227	SALOMON	243	0.02
228	MABADI	242	0.02
229	ABU	240	0.02
230	MENI	240	0.02
231	MUSALITSI	240	0.02
232	TSIOGU	239	0.02
233	DIAGU	237	0.02
234	MIOTSU	237	0.02
235	UVIOLILA	236	0.02
236	MILAKU	233	0.02
237	BAVAGULI	231	0.02
	TSI		
238	BAVINITSI	231	0.02
239	JE	231	0.02
240	MONYU	231	0.02
241	DU	230	0.02
242	MUGATSI	230	0.02
243	MIULA	229	0.02
244	NOGU	229	0.02
245	YITSIGA	228	0.02
246	ADINA	226	0.02
247	BUOTSU	225	0.02
248	GOMA	225	0.02
249	TOSINI	225	0.02
250	BUFUMU	223	0.02
251	DIUFI	222	0.02
252	GUVU	222	0.02
253	DIKABU	220	0.02
254	YIVUNDA	220	0.02
255	CGRID	219	0.02
256	DIAMI	219	0.02
257	DUANDI	219	0.02
258	UVU	218	0.02
259	JOSUE	217	0.02
260	NENU	217	0.02
261	JIVEMA	216	0.02
262	MALUNGU	216	0.02
263	MUSUGA	216	0.02
264	JEREMIE	215	0.02
265	COLUMN	214	0.02
266	MARUMA	214	0.02
267	UL	214	0.02
268	AMAMA	213	0.02
269	PLAIN	213	0.02

270	SECTD	213	0.02
271	SECTDEFA ULTCL	213	0.02
272	SFTNBJ	213	0.02
273	KEDI	212	0.02
274	SECT	210	0.02
275	GULUANU	209	0.02
276	TIE	209	0.02
277	BASEFU	207	0.02
278	AMAWAKA	206	0.02
279	BIDUARU	204	0.02
280	DUIMBU	204	0.02
281	ATSI	203	0.02
282	PHARAON	203	0.02
283	JUDA	202	0.02
284	ME	202	0.02
285	MUBIYITSI	202	0.02
286	AMAWEND A	201	0.01
287	PANOSE	201	0.01
288	UKU	200	0.01
289	BANSAMBI	199	0.01
290	BILIMA	199	0.01
291	MAMANYI	199	0.01
292	MUGULA	198	0.01
293	UBANDILA	197	0.01
294	BUNDUMB A	196	0.01
295	MBELA	196	0.01
296	BUNSONSI	195	0.01
297	DIBULI	194	0.01
298	J	194	0.01
299	GUNA	192	0.01
300	MABAANA	192	0.01
301	BABI	191	0.01
302	MUINSI	191	0.01
303	MIRIMA	190	0.01
304	NGOMBI	190	0.01
305	DUSUNGU GU	189	0.01
306	JI	189	0.01
307	MAVU	188	0.01
308	YINGEBA	188	0.01
309	AMAWIVUL A	187	0.01
310	BILUMBU	187	0.01
311	MAGUMAB EJI	187	0.01
312	YIGARA	187	0.01
313	YIPANGINI	187	0.01
314	FROMAN	186	0.01
315	RI	186	0.01
316	JIBOTI	185	0.01
317	MUKOLU	185	0.01
318	UDIBAALA	185	0.01
319	DIULEBA	184	0.01
320	MUBU	182	0.01

321 YILU	182	0.01
322 DIKOTULU	181	0.01
323 NYIINA	180	0.01
324 POSULU	180	0.01
325 YIAGU	180	0.01
326 JOSEPH	179	0.01
327 BIFU	178	0.01
328 DIONYI	178	0.01
329 NSALA	178	0.01
330 BABAGETU	177	0.01
331 DIKAKA	177	0.01
332 GAGU	176	0.01
333 MIKUTA	176	0.01
334 MIS	176	0.01
335 DUFU	175	0.01
336 SAMUEL	175	0.01
337 YIMA	175	0.01
338 TSIAGU	174	0.01
339 BIBEJI	173	0.01
340 MUNYONG	171	0.01
U		
341 TU	171	0.01
342 AVANA	170	0.01
343 OBA	170	0.01
344 BABYLONE	169	0.01
345 DIBUMI	169	0.01
346 DIENI	169	0.01
347 BABINDOM	168	0.01
BU		
348 EPALILI	168	0.01
349 MUNONGU	168	0.01
350 YINDOMBU	168	0.01
351 MISAMU	167	0.01
352 YIBUKU	167	0.01
353 YIFUMU	167	0.01
354 BAKA	166	0.01
355 KEEPEN	166	0.01
356 ABIVU	165	0.01
357 ABRAHAM	165	0.01
358 AKABURA	165	0.01
359 BAPHILISTI	165	0.01
N		
360 BUEGYPTE	165	0.01
361 YIAMI	165	0.01
362 BANA	164	0.01
363 JIVU	164	0.01
364 NYI	164	0.01
365 ANA	163	0.01
366 BIAGU	161	0.01
367 YIFU	161	0.01
368 EVOSI	160	0.01
369 NSORURU	158	0.01
370 AYIVU	157	0.01
371 BAMAVAG	157	0.01
A		
372 UVAGA	157	0.01
373 BALSRAEL	156	0.01

374 DUJABU	156	0.01
375 JINENI	156	0.01
376 MUJAMBA	156	0.01
377 DIMANYI	155	0.01
378 YIRANU	155	0.01
379 KOKOLU	154	0.01
380 NUNGI	154	0.01
381 DISIALA	153	0.01
382 YIRIERU	153	0.01
383 MIGT	152	0.01
384 UBANDA	152	0.01
385 BAKOSI	151	0.01
386 MI	151	0.01
387 SUI	151	0.01
388 BAMAWEN	150	0.01
DA		
389 YITU	150	0.01
390 BOBA	148	0.01
391 DIEDIDI	148	0.01
392 DIFUMU	148	0.01
393 RIN	148	0.01
394 SEFU	148	0.01
395 T	147	0.01
396 MIKONGU	146	0.01
397 UJI	146	0.01
398 MALAMU	145	0.01
399 OA	144	0.01
400 UMABA	144	0.01
401 GA	143	0.01
402 MUNGO	143	0.01
403 BAGETU	142	0.01
404 MASUMU	142	0.01
405 BIVU	141	0.01
406 BUAGU	141	0.01
407 DIWELIMIN	141	0.01
A		
408 EGUGU	141	0.01
409 TSUNGI	141	0.01
410 YIVINGA	141	0.01
411 AMAFU	140	0.01
412 BETU	140	0.01
413 DIBUKU	140	0.01
414 DURUMU	140	0.01
415 GULU	140	0.01
416 UTOGA	140	0.01
417 EJJI	139	0.01
418 PVPARA	139	0.01
419 UWENDA	139	0.01
420 BINDOMBU	138	0.01
421 NYIVOSI	137	0.01
422 VALA	137	0.01
423 DIBOTI	136	0.01
424 DIJENGA	136	0.01
425 BEBELI	135	0.01
426 AMAVEGA	134	
427 NYIMBU	134	

428	USUSU	134
429	MANA	133
430	ADIVU	132
431	DILONGU	132
432	DUVU	131
433	LNDSCPSX N	131
434	BIENU	130
435	L	130
436	MAPA	130
437	SAMBI	130
438	AGUNA	129
439	BIRIERU	129
440	AMALABA	128
441	YIVU	128
442	DIOTSU	127
443	TSIANDI	127
444	D	126
445	BUSINA	125
446	DUAGU	125
447	YA	125
448	DINA	124
449	TSIENU	124
450	VAGAA	124
451	VAMABA	124
452	AMABONG A	123
453	DUVOTSU	123
454	GAMI	123
455	NYIRONDI	123
456	TSIUNYEN SA	123
457	WENDA	123
458	BAMA	122
459	MAGUMARI ERU	122
460	SAM	122
461	YIBUSUSU	122
462	YIMABA	122
463	PESI	121
464	PISAMA	121
465	BIENI	120
466	MANDILU	120
467	MFULA	120
468	UNENI	120
469	ELISEE	119
470	TSIE	119
471	VIA	118
472	BANGOMBI	117
473	DIGUMBA	117
474	DUBUNGU	117
475	DUNENI	117
476	MURU	117
477	YIANSA	117
478	ISAAC	116
479	MABEJI	116
480	NGA	116

481 BANGO	115
482 DUAMI	115
483 USAKAMA	115
484 BAVIGA	114
485 BIMBUNGA	114
486 BINDAGU	114
487 JIJUDA	114
488 JOAB	114
489 UVOSA	114
490 AJIVU	113
491 DISIMU	113
492 JIISRAEL	113
493 MABILIMA	113
494 MIGAMU	113
495 NYINA	113
496 AMAVU	112
497 EZEKIEL	112
498 GANDI	112
499 MIAGU	112
500 MUBI	112
501 MUNU	112
502 Z	112
503 BANDA	111
504 BIAMI	111
505 DIKASI	111
506 DUNANGU	111
507 MIENU	111
508 VAVA	111
509 KADI	110
510 URUGA	110
511 ABIMABA	109
512 DIOGU	109
513 MBU	109
514 MIGAMBA	109
515 S	109
516 TSISIGA	109
517 BIARI	108
518 DIRUMA	108
519 EZEKIAS	108
520 YIJALALA	108
521 BUOGU	107
522 AMANSING ULA	106
523 BANGANA	106
524 DURONDU	106
525 MAFUMU	106
526 MOAB	106
527 NYAMA	106
528 MIAMI	105
529 BOU	104
530 BURANGA	104
531 MBA	104
532 YISUSU	104
533 MAPUELA	103
534 MAGUMAR ANU	102

535 MBONGU	102
536 NETU	102
537 AMABURA	101
538 DUOGU	101
539 MIRI	101
540 MUSIRU	101
541 BABIYITSI	100
542 BANENI	100
543 JIUNYENS	100
A	
544 METERA	100
545 NYIMALAB	100
A	
546 TSIAMI	100
547 YINYUNYI	100
548 AMAWAGU	99
LA	
549 BUAMI	99
550 DIB	99
551 MAGUMAN	99
A	
552 UDILA	99
553 BUPAGU	98
554 MANENI	98
555 ORUNGUL	98
A	
556 UROGA	98
557 BEBABA	97
558 MBIGULU	97
559 NN	97
560 PUNGA	97
561 BILEV	96
562 DIMABA	96
563 DULOMBILI	96
564 JIBABYLON	96
E	
565 BAKABALA	95
566 NOMBU	95
567 USIGAMA	95
568 YINDAGU	95
569 ABSALOM	94
570 BAKOLI	94
571 DINENI	94
572 IS	94
573 JIOTSU	94
574 YINSAMBI	94
575 YIUBA	94
576 BAMOSI	93
577 BIPUELA	93
578 DITASA	93
579 GOGU	93
580 DANIEL	92
581 NYUBA	92
582 SION	92
583 YINENI	92
584 YIVAGA	92
585 BALUMI	91



586 DINDAGU	91
587 FETA	91
588 UGULU	91
589 VENGU	91
590 YITSINGU	91
591 AMAGULU	90
592 AMAMUTSI NGULA	90
593 BISALULU	90
594 DUSAMBU	90
595 KANA	90
596 MILUNDA	90
597 SAMARIE	90
598 YINGITSI	90
599 AMANENGI LA	89
600 BIVAGA	89
601 DUMABA	89
602 NGU	89
603 YISUEMUN U	89
604 MALONGI	88
605 MIMBINI	88
606 TSIVU	88
607 TSOMI	88
608 UJABA	88
609 MANUNGI	87
610 V	87
611 YINANA	87
612 BIBAGUNU	86
613 DIUBA	86
614 EWENDISI NI	86
615 JIUTSINDU LA	86
616 LU	86
617 MAKULU	86
618 MFUMBI	86
619 VAJULU	86
620 YIOGU	86
621 AGAA	85
622 BINA	85
623 BISUSU	85
624 CANAAN	85
625 DIVINDUG ULU	85
626 ELIE	85
627 GALAAD	85
628 MIANDI	85
629 MINONGU	85
630 MUM	85
631 PILA	85
632 TSIMABOTI	85
633 YIBULU	85
634 BAMAVOS A	84
635 JINA	84

636 KINGU	84
637 MEMAMA	84
638 MUTEGULA	84
639 YISIAMUNU	84
640 DISUKUSU	83
LU	
641 EDOM	83
642 R	83
643 TANDU	83
644 YINSABI	83
645 AMAMUWI	82
VULA	
646 BIPAGU	82
647 BUVU	82
648 DE	82
649 MAKALU	82
650 MUGAGA	82
651 NU	82
652 PESU	82
653 USALA	82
654 YIBUFUMU	82
655 AMATINDA	81
656 BATSOLI	81
657 BENYI	81
658 BIBUKU	81
659 MATASA	81
660 MBANGU	81
661 VU	81
662 MAKAKA	80
663 MANSILA	80
664 TSONGU	80
665 YIBUSI	80
666 YIKAMA	80
667 BAGATSI	79
668 DUNSALA	79
669 JIUBA	79
670 MATSANDA	79
671 MIENI	79
672 AMAVAYIL	78
A	
673 BAALA	78
674 BARANU	78
675 CE	78
676 DUBA	78
677 E	78
678 NYILABI	78
679 TSIMBI	78
680 TUR	78
681 BET	77
682 EJABI	77
683 ERONDI	77
684 IE	77
685 NYANGU	77
686 YIBILIMA	77
687 AMAVOSIL	76
A	
688 BAMBATSI	76

689 BILUMBI	76
690 DIFUFUND U	76
691 DIMBU	76
692 MASUSU	76
693 NGUANGU	76
694 USUNGAM A	76
695 YIVUNGA	76
696 AMASINDI GA	75
697 BIUBA	75
698 BONGA	75
699 DISUSU	75
700 DUNGOSU	75
701 JIMABA	75
702 TSANDA	75
703 AMABUELA	74
704 BAJACOB	74
705 BANSIALUT SU	74
706 K	74
707 MBANDA	74
708 MBATSI	74
709 MBWANGE	74
710 MUENYI	74
711 NSA	74
712 UFUMU	74
713 ABRAM	73
714 BARIERU	73
715 BIISRAEL	73
716 BIRANU	73
717 DINGENSA	73
718 MATUJI	73
719 MOU	73
720 UFU	73
721 YIASI	73
722 YIDUKA	73
723 AMAVANG ANA	72
724 BALAAM	72
725 BALTIC	72
726 BAMALABA	72
727 BAMFUMBI	72
728 BIMABA	72
729 BIVISI	72
730 BUVIGA	72
731 CYR	72
732 GREEK	72
733 LABA	72
734 BUEBUBU	71
735 NDI	71
736 ACHAB	70
737 ESAU	70
738 KODU	70
739 MABATU	70
740 MBUNGA	70

741 MURAMBU	70
742 AKA	69
743 BIFUMBA	69
744 BITOGULU	69
745 BUENI	69
746 DIMUNGI	69
747 JOURDAIN	69
748 MAKUTU	69
749 OSI	69
750 POSNEGX	69
751 SANS	69
752 TSIENI	69
753 UMUGETU	69
754 VAVU	69
755 AMARUMA	68
NGA	
756 BAAL	68
757 BIMAGA	68
758 MAMMA	68
759 TANGI	68
760 YIBILUMBU	68
761 YIDUNA	68
762 BANSONSI	67
763 BITAYI	67
764 MIJAMBA	67
765 MUVIGA	67
766 NGILA	67
767 OOTSU	67
768 ARIAL	66
769 BAFUDU	66
770 BAMARAM	66
BUYILA	
771 BILORA	66
772 BUNENI	66
773 H	66
774 JIASSYRIE	66
775 JONATAN	66
776 MUGANDA	66
777 NDOSI	66
778 TSIA	66
779 TSI LORA	66
780 AMAGAA	65
781 BAAMMON	65
782 BLE	65
783 MAMABA	65
784 PURULU	65
785 YIBANDA	65
786 BABENJAM	64
IN	
787 DUDAVID	64
788 DUYITSU	64
789 KERI	64
790 MAGUMIDU	64
SAMBUA+	
791 MBEMBU	64
792 METU	64
793 NYIMAVOS	64

A	
794 SI	64
795 TSUGU	64
796 UBENGA	64
797 UTSIRA	64
798 YIONYI	64
799 YITEGA	64
800 YIYIGUMI	64
801 AGUGAA	63
802 AM	63
803 BIOU	63
804 DUFUNU	63
805 MISUSU	63
806 MUDUDU	63
807 MUVINITSI	63
808 NEW	63
809 UONYI	63
810 VII	63
811 YIETSU	63
812 AYIMABA	62
813 ELABI	62
814 KAKA	62
815 NDUKU	62
816 YIMBUNGU	62
817 BAFUMU	61
818 BASIGAMA	61
819 BASUNGA	61
MA	
820 F	61
821 GUIRONDI	61
822 HEBRON	61
823 JIONYI	61
824 JITSANGU	61
825 KULINI	61
826 MUFUDU	61
827 MUKILU	61
828 NABUCOD	61
ONOSOR	
829 SA	61
830 TSIDUYITS	61
U	
831 UDAVID	61
832 UDUGUSU	61
833 YIBUVAGU	61
LITSI	
834 DUGANU	60
835 MUIRI	60
836 NSUNGISIN	60
I	
837 VHANA	60
838 YIBAGORA	60
839 YINSI	60
840 ABAKAVU	59
841 BUCANAAN	59
842 DIGUMI	59
843 FUNDU	59
844 JOSAPHAT	59

845 MOOA	59
846 UTSANA	59
847 ABABA	58
848 AMABILA	58
849 AMASOLA	58
850 DIBADI	58
851 ELAZAR	58
852 JIFUMU	58
853 MIURU	58
854 NGAA	58
855 PINSÁ	58
856 TSYA	58
857 UKAGA	58
858 UVEGA	58
859 UYITSA	58
860 ASA	57
861 BADAVID	57
862 BIBAGORA	57
863 BIBI	57
864 BIBIANSA	57
865 BIJAGA	57
866 BOJABA	57
867 DIBAISRAE	57
L	
868 EFRAIM	57
869 FMODERN	57
870 GENU	57
871 JISUSU	57
872 LABAN	57
873 MUGOBUT	57
SI	
874 OUT	57
875 TSINGULA	57
876 UGABUGA	57
877 UNSAMBI	57
878 YILIMBA	57
879 AMABATSI	56
NGULA	
880 AMAGABU	56
GA	
881 AMAMURU	56
MA	
882 DA	56
883 DUT	56
884 GENGILA	56
885 JEROBOA	56
M	
886 KUNGA	56
887 MAMUNGI	56
888 NGENSA	56
889 NYIJABI	56
890 YIJUNGI	56
891 ABAGAA	55
892 ASSYRIE	55
893 BEVOSI	55
894 BUBA	55
895 BULUMBI	55

896 DUFUNUS	55
U	
897 GUBA	55
898 JEHU	55
899 NGUBU	55
900 TANGUJI	55
901 TULU	55
902 TUVU	55
903 UKELISA	55
904 YERI	55
905 YITOSINI	55
906 ADU	54
907 AMADILA	54
908 BO	54
909 BUENU	54
910 DIJOURDAI	54
N	
911 FNIL	54
912 IN	54
913 MATOSINI	54
914 MIBEJI	54
915 MIVU	54
916 NDA	54
917 URUMA	54
918 AI	53
919 AMAROND	53
A	
920 BAGAA	53
921 BETHEL	53
922 BOJI	53
923 BURU	53
924 DIENU	53
925 HEB	53
926 MBOLWAN	53
U	
927 MFUGA	53
928 MOBA	53
929 MUBOTI	53
930 NYIMABA	53
931 OYIDIVOSI	53
932 YIRIARIAM	53
U	
933 ALETANA	52
934 AMABUSA	52
935 AMARUMIN	52
A	
936 BAMARINA	52
937 BIONYI	52
938 BIS	52
939 DIMI	52
940 IT	52
941 LETASATIA	52
NU	
942 MAWELIMI	52
NA	
943 NDUNGA	52
944 NSIENGA	52

945 NYIMAVAN	52
GANA	
946 SYRIE	52
947 TSIMIKON	52
GU	
948 WENDANU	52
949 AMABOKA	51
950 AMAGAMU	51
GA	
951 AMIVU	51
952 BAJUIF	51
953 BOLABA	51
954 DIMUSIEN	51
GI	
955 GEDEON	51
956 GUIJABI	51
957 MAMA	51
958 MASIYI	51
959 MASUKA	51
960 MIJUDA	51
961 MIKUDU	51
962 MIONYI	51
963 MULOQU	51
964 RUGA	51
965 YITSANU	51
966 YIUVUMA	51
967 BAMABON	50
GA	
968 BAMATOLA	50
969 BIYINGITSI	50
970 BUDUKA	50
971 DUYIVUND	50
A	
972 KEMBU	50
973 KUMBU	50
974 MABA	50
975 MALORA	50
976 MUMBWAN	50
GA	
977 NYUVAGA	50
978 SICHEM	50
979 TAVULA	50
980 ULILA	50
981 URINA	50
982 YIBAVAGU	50
LITSI	
983 YIJAGA	50
984 BABURU	49
985 BUISRAEL	49
986 DIMABI	49
987 DUJABA	49
988 JOB	49
989 MABANDU	49
990 MAGUMASI	49
AMUNU	
991 NSONSI	49
992 NYIDU	49



993 SAMSON	49
994 ULABA	49
995 AGORUNG ULA	48
996 AMATABUL A	48
997 BAMBUELI LI	48
998 BIDURUMU	48
999 BISINGA	48
1000 BUMBEMB U	48
1001 DIKALU	48
1002 DINSAMBI	48
1003 LIA	48
1004 MADIAN	48
1005 MBANSA	48
1006 MIMBU	48
1007 MS	48
1008 MUMAPALI LA	48
1009 TSIKABUS ULU	48
1010 UBUSA	48
1011 ABETSANI	47
1012 ABNER	47
1013 AYINA	47
1014 BAMABEG A	47
1015 BANUMBA	47
1016 BILU	47
1017 HI	47
1018 JIMOSI	47
1019 MAKOTULU	47
1020 MIL	47
1021 NYIMAVAG A	47
1022 PURA	47
1023 SBKNONE	47
1024 UYIONYI	47
1025 YIABUTSU	47
1026 YITANDU	47
1027 ABIMELEK	46
1028 ABUVU	46
1029 AJINA	46
1030 AMATOLA	46
1031 BAISRAEI	46
1032 BAMAYEVA GA	46
1033 BASYRIE	46
1034 BIVARU	46
1035 DIGAA	46
1036 GUIBEA	46
1037 HLL	46
1038 IRI	46
1039 LI	46
1040 MAPAPI	46

1041 MARDOCH EE	46
1042 MINENI	46
1043 NYAMBI	46
1044 RUGANU	46
1045 YIBAGA	46
1046 YIRANGUM UNU	46
1047 ABAMA MBULU	45
1048 BAKELISI	45
1049 BE	45
1050 BINSABI	45
1051 ERUYI	45
1052 MALUBU	45
1053 MUGAMBA	45
1054 NSIMBULU	45
1055 NYIJI	45
1056 NYIMAGUL U	45
1057 TSIPUELA	45
1058 YIMANASS E	45
1059 YL	45
1060 ABANA	44
1061 AMARELA MA	44
1062 BIBAGA	44
1063 BIRUITSI	44
1064 BIWELIMIN A	44
1065 BOBOKU	44
1066 GUILABI	44
1067 IUKELISILA	44
1068 MAJENGA	44
1069 MANGUAN GU	44
1070 MIFUMU	44
1071 MUGESA	44
1072 MULU	44
1073 NYU	44
1074 PEYI	44
1075 UBUNGA	44
1076 UMU	44
1077 YILALA	44
1078 YILUMBI	44
1079 YIMUSAMB UALI	44
1080 YINOMBU	44
1081 AMALUGA	43
1082 AN	43
1083 BAMATSIN GULA	43
1084 BIDUKA	43
1085 BUALI	43
1086 JIABAFU	43
1087 JIEGYPTE	43

1088 JOAS	43
1089 KULA	43
1090 MANASSE	43
1091 MBIVULU	43
1092 MILUNGA	43
1093 MONYI	43
1094 MWANA	43
1095 NYULASA	43
1096 PITA	43
1097 ROBOAM	43
1098 SEDECAS	43
1099 ULIOMA	43
1100 BAAMOR	42
1101 BARUMITSI	42
1102 BIBOTI	42
1103 DIDUMA	42
1104 DUGOBUT SU	42
1105 ESTHER	42
1106 G	42
1107 GULABA	42
1108 JIUNU	42
1109 LABANU	42
1110 LIE	42
1111 MABATAJI	42
1112 MBWANGA	42
1113 MIKANSU	42
1114 MINA	42
1115 MUNENI	42
1116 PAQUE	42
1117 PENYI	42
1118 PUELA	42
1119 TSIBAGOR A	42
1120 ABAMATSA NANGA	41
1121 AJIMABA	41
1122 AMANA	41
1123 AMANYOYI LA	41
1124 BABEYITSI	41
1125 BAMATSAN A	41
1126 BIETSU	41
1127 BUJUDA	41
1128 GABUGA	41
1129 GETU	41
1130 IIIU	41
1131 LABATI	41
1132 MITATU	41
1133 MUNGUDI	41
1134 NGULA	41
1135 NYIM	41
1136 TSU	41
1137 UBOKA	41
1138 UKUMUGA	41
1139 ULABANA	41

1140	YITSUNGI	41
1141	ABOKU	40
1142	AMAKOTA	40
1143	AMAMUVE GA	40
1144	BAAARON	40
1145	BABUNGUL ITSI	40
1146	BAEFRAIM	40
1147	BAM	40
1148	BILIMBA	40
1149	BIMUNONG U	40
1150	BUKETI	40
1151	DIBATU	40
1152	DUMBULU	40
1153	JERICHO	40
1154	JIBATU	40
1155	KOSI	40
1156	LABATIANU	40
1157	MBEJI	40
1158	MIRAMBU	40
1159	NSAMBL	40
1160	NYOGA	40
1161	OFU	40
1162	RUTH	40
1163	ULASA	40
1164	YIMAMBA	40
1165	ZACHARIE	40
1166	AMABEGA	39
1167	AMAMUEV OSA	39
1168	BABABAAI A	39
1169	BILEKA	39
1170	DIBAAIA	39
1171	DIMALABA NA	39
1172	DUITSINGI NGI	39
1173	FSCRIPT	39
1174	GOTHIC	39
1175	HADAD	39
1176	JER	39
1177	JITEGA	39
1178	JIWELIMIN A	39
1179	KAOOGU	39
1180	MIUNU	39
1181	NAMAPAPA	39
1182	NI	39
1183	NYUWEND A	39
1184	ODUVEGA	39
1185	PUILA	39
1186	RACHEL	39
1187	WAMI	39

1188	YIENU	39
1189	YIVUMU	39
1190	ADIDI	38
1191	AMABOKIS A	38
1192	AMAMABA MBANA	38
1193	AMARINA	38
1194	AMASSIA	38
1195	AMAWAKIS A	38
1196	AMAYENA WENDA	38
1197	BABANGO MBI	38
1198	BIBAKAGA	38
1199	BINENI	38
1200	BINYENSU LU	38
1201	BISIAMUNU	38
1202	DILIMA	38
1203	EVOSILI	38
1204	GUNU	38
1205	IOU	38
1206	ISMAEL	38
1207	JIMAGAAN GA	38
1208	JOSIAS	38
1209	MAKABU	38
1210	MIMBANDA	38
1211	NOE	38
1212	YIBAWIMBI TSI	38
1213	YIEYIYI	38
1214	AMAGUKIG A	37
1215	AMAPALA	37
1216	AMAPANSA	37
1217	AMASAMBI LA	37
1218	BABAKAGA	37
1219	BASAMO	37
1220	BIMUSIRU	37
1221	DUDILA	37
1222	KABUSULU	37
1223	LIMBA	37
1224	MAPAGU	37
1225	MAPURA	37
1226	MUNGONG U	37
1227	MUSONITSI	37
1228	NANGA	37
1229	NATAN	37
1230	NOVOSA	37
1231	NYURUNG ULA	37
1232	TSIBATU	37
1233	UWELIMIN	37

A		
1234	AJIGAA	36
1235	AMABARU	36
	MA	
1236	AMATSANA	36
1237	ANOFU	36
1238	AZARIA	36
1239	BAMADILA	36
1240	BAMAKOT	36
	ULU	
1241	BAMAVAYI	36
	LA	
1242	BAMAWAK	36
	A	
1243	BENJAMIN	36
1244	BETHLEEM	36
1245	DUBATU	36
1246	DULMBU	36
1247	FUFU	36
1248	GUKAVAG	36
	A	
1249	JITSANDA	36
1250	KANU	36
1251	MABABAAL	36
	A	
1252	MATT	36
1253	MICROSOF	36
	T	
1254	MOJI	36
1255	MUKATA	36
1256	MULAMBU	36
1257	MULEMBU	36
1258	NSIEMBU	36
1259	NZAMBI	36
1260	UJIONYI	36
1261	UNU	36
1262	USAUL	36
1263	UYETOGA	36
1264	YIMETERA	36
1265	ABEPINI	35
1266	ADUVU	35
1267	BAGAD	35
1268	BAMABOK	35
	A	
1269	BAMIKUTA	35
1270	BIBATU	35
1271	BILUNDUL	35
	U	
1272	BUNSAMBI	35
1273	CAIN	35
1274	DL	35
1275	ESDRAS	35
1276	GORA	35
1277	GUJI	35
1278	IA	35
1279	MAKOKOK	35
	U	

1280 MANSAMBI	35
1281 MAVITA	35
1282 MBI	35
1283 MILENDU	35
1284 MISPA	35
1285 MUBUNGU	35
LITSI	
1286 NDEJU	35
1287 ON	35
1288 SARA	35
1289 TAMAR	35
1290 TSIBABAAL	35
A	
1291 TSIRAYISA	35
N	
1292 UAGU	35
1293 UNA	35
1294 UVIOGA	35
1295 VOSA	35
1296 YIJUDA	35
1297 ABIMAKIND	34
U	
1298 ADIMABA	34
1299 AJIUSIELU	34
SUNU	
1300 AMABAVE	34
GA	
1301 BAJOSEPH	34
1302 BAMAWIVU	34
LA	
1303 BAMAYENA	34
WENDA	
1304 BANGOGA	34
GA	
1305 BARUBEN	34
1306 BEN	34
1307 BIABUTSU	34
1308 BIKAGA	34
1309 DIISRAEL	34
1310 DISUMU	34
1311 DITELI	34
1312 DUOTSU	34
1313 JISYRIE	34
1314 JORAM	34
1315 KATSI	34
1316 KU	34
1317 MAGAJI	34
1318 MARIERU	34
1319 MINSANGU	34
1320 MIRAYISAN	34
1321 NYIWENDY	34
E	
1322 NYURU	34
1323 OYADA	34
1324 SU	34
1325 TSIBLE	34
1326 UDUROND	34
U	

1327 UTSIENGIL A	34
1328 VAGA	34
1329 VEGA	34
1330 YILORA	34
1331 AGUAMI	33
1332 AHAZIA	33
1333 AMAVAGA NGA	33
1334 BAMANGO LU	33
1335 BOOZ	33
1336 CHIMEI	33
1337 DAN	33
1338 DIBISI	33
1339 DIGUKI	33
1340 DIKONGU	33
1341 FUIRI	33
1342 IIN	33
1343 JIKABUSUL U	33
1344 LUCIDA	33
1345 MIKANDU	33
1346 MURUMITS I	33
1347 NGEBI	33
1348 SADO	33
1349 SERIF	33
1350 TA	33
1351 TSOMU	33
1352 YITSONA	33
1353 AHAZ	32
1354 AMABAWIV ULA	32
1355 AMABURIL A	32
1356 AMAVAGUL A	32
1357 APALILA	32
1358 BELEETI	32
1359 BIETU	32
1360 BITUMBA	32
1361 BUETU	32
1362 CHEMAYA	32
1363 DIMIGAGA	32
1364 DIWERU	32
1365 DUEDUDU	32
1366 DUFUMU	32
1367 DUGABA	32
1368 ERUMI	32
1369 EVEYI	32
1370 GULA	32
1371 GUMAVAG A	32
1372 INA	32
1373 JIZREEL	32
1374 KUERU	32



1375 MABISI	32
1376 MAMATSA NDA	32
1377 MAMETER A	32
1378 MATSIENDI	32
1379 MINSAMBI	32
1380 MO	32
1381 NGI	32
1382 NGORUNG ULA	32
1383 OBONGA	32
1384 PAGA	32
1385 RA	32
1386 TSANGU	32
1387 TSIMABA	32
1388 UBASA	32
1389 UBOKU	32
1390 UKOTA	32
1391 USION	32
1392 UVAGULA	32
1393 VHA	32
1394 YE	32
1395 YILIMBU	32
1396 YIYIFUMBA	32
1397 ABAMASAN SUMU	31
1398 AMABANDI LA	31
1399 AMAGUDA MA	31
1400 ASAMABA	31
1401 BABINSABI	31
1402 BALAE	31
1403 BAMAGUL U	31
1404 BAMAMUT SINGULA	31
1405 BANONGU	31
1406 BERUYI	31
1407 BODILA	31
1408 DIBISALU	31
1409 DUGAA	31
1410 DUKA	31
1411 DUWAMUS U	31
1412 MABULON GU	31
1413 MABUTAM BA	31
1414 MAGEYI	31
1415 MAGUYI	31
1416 MICHEE	31
1417 MISONGU	31
1418 MUKUDU	31
1419 NDASI	31
1420 QUIRIAT	31
1421 REBECCA	31

1422 TSIUBA	31
1423 TYR	31
1424 UNENGILA	31
1425 VOBA	31
1426 YECHOUA	31
1427 YIBOTI	31
1428 ABANDI	30
1429 ABERONDI	30
1430 AGUGU	30
1431 AGUITSANI	30
1432 AGUMUBEJ	30
I	
1433 AMAKOTIS	30
A	
1434 AMAMABA	30
1435 AMASAMAL	30
ASA	
1436 AMATUMA	30
1437 AMINA	30
1438 BANGUJI	30
1439 BELABI	30
1440 COURIER	30
1441 DILE	30
1442 DIMABAAL	30
A	
1443 DIMUANA	30
1444 DIPA	30
1445 DIYIFUMBA	30
1446 DUJANGAN	30
GA	
1447 FALT	30
1448 GAT	30
1449 GUILGAL	30
1450 GUIVAYI	30
1451 GURUNGU	30
LA	
1452 IICSI	30
1453 IIU	30
1454 IU	30
1455 KEBANANU	30
1456 KILU	30
1457 LONGU	30
1458 MIRIERU	30
1459 MUGELI	30
1460 MULAKU	30
1461 MUYITSA	30
1462 NANGULA	30
1463 NINIVE	30
1464 NYIMA	30
1465 NYIMAVEG	30
A	
1466 ODILA	30
1467 PHPG	30
1468 PVPG	30
1469 RANGIMIN	30
A	
1470 ROMAN	30

1471 RUMA	30
1472 TAHOMA	30
1473 TSietu	30
1474 UBI	30
1475 UeFRAIM	30
1476 UWARISA	30
1477 UYETSING ULA	30
1478 YIBI	30
1479 AGA	29
1480 AGUNYI	29
1481 AMARUNGI SA	29
1482 AMAYENA GABUGA	29
1483 ANOMANA	29
1484 BABAKABA LA	29
1485 BANU	29
1486 BENAYA	29
1487 BINOVIog A	29
1488 CALEB	29
1489 DIGEYI	29
1490 DIKAGA	29
1491 DIMOYI	29
1492 DINGASA	29
1493 DUMBU	29
1494 ELASI	29
1495 GULUATI	29
1496 GUYALABA	29
1497 MAKUMBI	29
1498 MAMUNON GU	29
1499 MFUNGA	29
1500 MIFUNA	29
1501 MIMASONU GU	29
1502 MUFI	29
1503 MUKELISI	29
1504 MUMBAND A	29
1505 NUMBA	29
1506 NYUNDU	29
1507 OTAM	29
1508 OVAGULA	29
1509 SILO	29
1510 TUNA	29
1511 UDUKA	29
1512 UJOURDAI N	29
1513 UNOUN	29
1514 UWAKA	29
1515 UWE	29
1516 YIBATU	29
1517 AMAGAAN GA	28

1518	ATSIVU	28
1519	BAAMALEC	28
1520	BAMABUSA	28
1521	BAMAKEVE GA	28
1522	BAMAKOTA	28
1523	BANYAMA	28
1524	BASAMAR UNGULA	28
1525	BUKANYI	28
1526	DAMAS	28
1527	DIKUMBI	28
1528	DUJI	28
1529	DUKELISA NGA	28
1530	DURUNGU LA	28
1531	DUTSONS UGU	28
1532	DUYALABA	28
1533	EM	28
1534	GAD	28
1535	HIRAM	28
1536	LILA	28
1537	LOT	28
1538	MADUFUN U	28
1539	MANDOSI	28
1540	MBUTA	28
1541	MUKUILI	28
1542	NSANUNU	28
1543	NYIDUVEYI	28
1544	NYUGUVE GA	28
1545	NYUVOSA	28
1546	OSUSA	28
1547	ROM	28
1548	TSIGARAM A	28
1549	TSISUSU	28
1550	UBEGA	28
1551	VHAVHA	28
1552	WOGU	28
1553	YIETU	28
1554	YIKADI	28
1555	YINA	28
1556	YITSIBA	28
1557	ADONIA	27
1558	AMAGABIL A	27
1559	AMALASA	27
1560	AMAMUWA KA	27
1561	AMATSIGU	27
1562	BAKUILI	27
1563	BAMAGAA NGU	27
1564	BAMAGUK	27

AMA	
1565 BAMAJI	27
1566 BAMAPALIL	27
A	
1567 BAMAVEG	27
A	
1568 BAVAYITSI	27
1569 BILA	27
1570 BISIANA	27
1571 DIBU	27
1572 DISALA	27
1573 DIURUNGU	27
LA	
1574 DUIRONDI	27
1575 EDILI	27
1576 GABUGAN	27
U	
1577 GUKABON	27
GA	
1578 HAMAN	27
1579 HELVETICA	27
1580 IIII	27
1581 JILORA	27
1582 JO	27
1583 KAGAJIJUD	27
A	
1584 LENZA	27
1585 MAGAA	27
1586 MAMFURU	27
LU	
1587 MUFUNA	27
1588 NYIMASOL	27
A	
1589 OLABA	27
1590 PNHANG	27
1591 PNTXTA	27
1592 RELAMA	27
1593 SUSU	27
1594 T'WENDI	27
1595 TATA	27
1596 TIMES	27
1597 TSOLI	27
1598 VASAMABA	27
1599 WA	27
1600 WAGU	27
1601 YAYIYE	27
1602 YIUSILA	27
1603 ABIGAA	26
1604 ABUBA	26
1605 AMAINGUL	26
A	
1606 AMAJABA	26
1607 AMARETSA	26
MA	
1608 AMAWETSI	26
GA	
1609 ANOLABA	26
1610 BADIELA	26

1611	BAMAGUD AMA	26
1612	BANGANG A	26
1613	BAYIFUMB A	26
1614	BIFUMU	26
1615	BIKUNGA	26
1616	BIRIARIAM U	26
1617	BUMBINI	26
1618	DIA	26
1619	DIBAGORA	26
1620	DINGUJI	26
1621	DUBILA	26
1622	DUI	26
1623	EN	26
1624	ES	26
1625	FUTA	26
1626	GUEDALIA	26
1627	HOTSU	26
1628	JIGAA	26
1629	JISAKAMA	26
1630	LEA	26
1631	MAGUMIDU NANA	26
1632	MAKILU	26
1633	MANSINSI MA	26
1634	MIENSILI	26
1635	MIETU	26
1636	MIFIGA	26
1637	MIISRAEL	26
1638	NGANA	26
1639	NOEMI	26
1640	OVAGA	26
1641	P	26
1642	PIASU	26
1643	SAKU	26
1644	UANA	26
1645	UBATU	26
1646	UKANDU	26
1647	UOU	26
1648	UTAJI	26
1649	UYETOYIS A	26
1650	YIBAANA	26
1651	YIDUARU	26
1652	YIDUJEGU SU	26
1653	YIKUMBU	26
1654	YIKUTU	26
1655	YILINGA	26
1656	YINUNU	26
1657	YIOU	26
1658	YISINA	26
1659	YISUISU	26

1660 AB	25
1661 ABINA	25
1662 ADUDU	25
1663 AMAFUILA	25
1664 AMAMU	25
1665 AMAMURU MINA	25
1666 AMASILA	25
1667 AMASUND A	25
1668 AMATASA	25
1669 AMAWELA	25
1670 BABA	25
1671 BABATU	25
1672 BAGEYI	25
1673 BAKUNSA MA	25
1674 BAMAFU	25
1675 BAMALON GU	25
1676 BAMAVAG ULILA	25
1677 BAMBEJI	25
1678 BAPONDU	25
1679 BERCHEBA	25
1680 BIKUNDUB UGA	25
1681 BIRIRI	25
1682 BOGABUG A	25
1683 BORUNGU LA	25
1684 BOVOSA	25
1685 BUKAGA	25
1686 CADES	25
1687 DIKUTU	25
1688 DITODI	25
1689 DULABA	25
1690 DUYAJI	25
1691 GUMAVOS A	25
1692 ISI	25
1693 JEFTE	25
1694 JIYILUMBU	25
1695 MADUMBU LU	25
1696 MARANU	25
1697 MIEMIMI	25
1698 MIGANDA	25
1699 MIKA	25
1700 MOO	25
1701 N'OMBU	25
1702 NDU	25
1703 OBANGAN GA	25
1704 OVAGULIL A	25
1705 OVOSA	25

1706 RAMA	25
1707 SIDON	25
1708 TUGAA	25
1709 UBAMBAN A	25
1710 UBONGA	25
1711 UGAMUGA	25
1712 UGENGILA	25
1713 UNYOGA	25
1714 UVAGULIL A	25
1715 UWAMUSA	25
1716 YIA	25
1717 YIDIBUTSU	25
1718 YILEKA	25
1719 YIMATSYE MUGILA	25
1720 YISONYE	25
1721 AGO	24
1722 AJIMUBEJI	24
1723 AMABELIS A	24
1724 AMABUAG A	24
1725 AMABUESA	24
1726 AMAGAAN GU	24
1727 AMALABAN A	24
1728 AMAPALIS A	24
1729 AMUNA	24
1730 ANOGULU	24
1731 BACHAN	24
1732 BAEDOM	24
1733 BAMABILA	24
1734 BAMAGAM UGA	24
1735 BAPANGA	24
1736 BAWELIMI NA	24
1737 BEWENDI	24
1738 BIBINA	24
1739 BOJANGAN GA	24
1740 BOKELISA NGA	24
1741 BUJI	24
1742 BUKUILI	24
1743 BUMBEJI	24
1744 DIULABAN A	24
1745 DUBANGA NGA	24
1746 JESSE	24
1747 KILUMETE RA	24
1748 MABAALA	24



1749 MAMAMBA	24
1750 MARONDA	24
1751 MARUNDU	24
1752 MBENGI	24
1753 MFURULU	24
1754 MIBAANA	24
1755 MILONSI	24
1756 MIPUELA	24
1757 MISEDERA	24
1758 MUSAMU	24
1759 MUYITSIAN	24
U	
1760 MYRIAD	24
1761 NDUNSULU	24
1762 NYIVOSILI	24
1763 NYUVEGA	24
1764 OOGU	24
1765 RANGIMIN	24
ANU	
1766 TUVAGA	24
1767 UBOTI	24
1768 UGABILA	24
1769 ULORA	24
1770 UMUTUBU	24
1771 URANGA	24
1772 UTUMUNA	24
1773 VIETNAME	24
SE	
1774 WEB	24
1775 YILA	24
1776 YIMUSIEN	24
GI	
1777 YISAKU	24
1778 YITUMBA	24
1779 ABEL	23
1780 AMARUNGI	23
LA	
1781 AMASONA	23
1782 AYIMURIER	23
U	
1783 BAGEJABI	23
1784 BAJEBUSIE	23
1785 BAJERUSA	23
LEM	
1786 BAMABEN	23
GUNA	
1787 BAMAGAB	23
UGA	
1788 BAMAPANS	23
A	
1789 BAMATSAN	23
ANGA	
1790 BAMUTETI	23
1791 BARUCH	23
1792 BEVAYI	23
1793 BIJERUSAL	23
EM	
1794 BIMBURA	23

1795 BISIGU	23
1796 BISUNSUL	23
U	
1797 BUMABA	23
1798 BUONYI	23
1799 DIBUFU	23
1800 DILONGI	23
1801 DISU	23
1802 EVAYI	23
1803 GUIVOSI	23
1804 HANANIA	23
1805 HAZAEL	23
1806 JIMAMANYI	23
1807 JINSAMBI	23
1808 JONAS	23
1809 KANDA	23
1810 MADILU	23
1811 MADUBIAT	23
SU	
1812 MIBI	23
1813 MIBOTI	23
1814 MUVAGULII	23
1815 MUVUMU	23
1816 NADAB	23
1817 NDUNGUL	23
U	
1818 NGANGA	23
1819 NGO	23
1820 NYUTSIEM	23
USA	
1821 SIELA	23
1822 TSIONYI	23
1823 UBIVA	23
1824 UDIONYI	23
1825 UFUNA	23
1826 UPALA	23
1827 UPANSA	23
1828 UTABULA	23
1829 UTAALA	23
1830 UTU	23
1831 YIDAN	23
1832 YIDUYITSU	23
1833 YIKURU	23
1834 YISANGA	23
1835 ABAMI	22
1836 AGUBA	22
1837 AJIUGULU	22
NU	
1838 AMADUKA	22
1839 AMAFUBAN	22
A	
1840 AMANYOG	22
A	
1841 AMASAMB	22
A	
1842 AMASUNS	22
A	

1843	ATSIMABA	22
1844	ATSINA	22
1845	BAFETA	22
1846	BAHITT	22
1847	BAMABASA	22
1848	BAMABOK	22
	U	
1849	BAMADUK	22
	A	
1850	BAMATSAN	22
	INANGA	
1851	BASALII	22
1852	BASIELA	22
1853	BAY	22
1854	BERONDI	22
1855	BIDUNGA	22
1856	BIVONGU	22
1857	BIYINYENS	22
	ULU	
1858	BL	22
1859	BUBOTI	22
1860	BUVUYISA	22
1861	DIDIBI	22
1862	DILORA	22
1863	DIWAVI	22
1864	DUENI	22
1865	DUKAVAG	22
	ULANGA	
1866	DUNA	22
1867	DUNYURU	22
1868	GABAON	22
1869	GUEHAZI	22
1870	GUGAA	22
1871	GUKATSIN	22
	GULA	
1872	HL	22
1873	IIIIU	22
1874	ILA	22
1875	ILL	22
1876	JIMUGETU	22
1877	JIYIARI	22
1878	JOAQUIM	22
1879	LIGA	22
1880	LIMA	22
1881	MABENGA	22
1882	MABILUMB	22
	U	
1883	MANGA	22
1884	MBUSA	22
1885	MIBATU	22
1886	MIBISI	22
1887	MIKAL	22
1888	MILIBAN	22
1889	MILOGU	22
1890	MINDUBI	22
1891	MIOU	22
1892	MITSUNDU	22

1893 MUMBINI	22
1894 MUSUEGA	22
1895 MUTETI	22
1896 MUVA	22
1897 NAAMAN	22
1898 NGOLU	22
1899 NSI	22
1900 OBED	22
1901 ONATAN	22
1902 OTUBULA	22
1903 RAMOT	22
1904 SIHON	22
1905 SUNGA	22
1906 TS	22
1907 TSIBAISRA EL	22
1908 TSIMUNON GU	22
1909 TSYE	22
1910 TUGA	22
1911 TUMAVAG A	22
1912 UACHAB	22
1913 UFUBANA	22
1914 USANGILA	22
1915 UTOYISA	22
1916 UVIVA	22
1917 YIBAGUNU	22
1918 YIBENJAMI N	22
1919 YIDUMBITS I	22
1920 YIDURUMU	22
1921 YIGAD	22
1922 YIISRAEL	22
1923 YIKE	22
1924 YISAGULU	22
1925 YISUNSUL U	22
1926 YIVIOVI	22
1927 ABAMA	21
1928 ABIATAR	21
1929 ABIBI	21
1930 AMAMATA	21
1931 AMANYEN SA	21
1932 AMAPASA	21
1933 AMAVAGUL ILA	21
1934 ANDL	21
1935 ATSIGAA	21
1936 AYIMUBEJI	21
1937 BABABYLO NE	21
1938 BABAPHILI STIN	21
1939 BABOTI	21

1940 BAMAGUL ULU	21
1941 BAMATABU LA	21
1942 BAMOAB	21
1943 BAN	21
1944 BANDAGU	21
1945 BANGEBI	21
1946 BASAUL	21
1947 BAWIMBIT SI	21
1948 BIEBIBI	21
1949 BIMBUNGU	21
1950 BOTSANAN GA	21
1951 BUKIDI	21
1952 BULU	21
1953 BUNDU	21
1954 BUNGEBI	21
1955 CENTURY	21
1956 DIBALA	21
1957 DIKULU	21
1958 DUBISI	21
1959 DURU	21
1960 DUSIMBI	21
1961 GAMUGAN U	21
1962 GUFU	21
1963 HEBREW	21
1964 JABA	21
1965 JIKAGA	21
1966 JIYINGITSI	21
1967 KABALA	21
1968 KOBULI	21
1969 KWAKI	21
1970 LINOTYPE	21
1971 LUE	21
1972 MABAISRA EL	21
1973 MAMOSI	21
1974 MARC	21
1975 MARU	21
1976 MATSANG A	21
1977 MIBIJI	21
1978 MIYOMBU	21
1979 MUENDU	21
1980 MUWAMUS I	21
1981 NEHEMIE	21
1982 NOBUEJI	21
1983 NONGU	21
1984 NYIMAGEN GILA	21
1985 NYUBUNG A	21
1986 OHANAN	21

1987 OTABULA	21
1988 PALATINO	21
1989 PANGA	21
1990 RIA	21
1991 SUSE	21
1992 TSIBISI	21
1993 TSIMAMBA	21
1994 TSIMBURA	21
1995 TSOKUDU	21
1996 UBOKISA	21
1997 UFUTA	21
1998 UGEYI	21
1999 UNICODE	21
2000 UREGAMA	21
2001 URUNGA	21
2002 UVAGANG	21
A	
2003 UYIGUMI	21
2004 VAGOBA	21
2005 VERDANA	21
2006 YIBIMA	21
2007 YIEFRAIM	21
2008 YIMAGUMA	21
BEJI	
2009 YIMUBU	21
2010 YIMUGUMI	21
2011 YIRELA	21
2012 YIVOSI	21
2013 ABAMATSA	20
NA	
2014 ABAMAWE	20
NDA	
2015 ADUMAWI	20
MBULU	
2016 AFU	20
2017 AJIJI	20
2018 AKICH	20
2019 AMABAWA	20
KA	
2020 AMALABILA	20
2021 AMARUNG	20
A	
2022 AMATOLIL	20
A	
2023 AMMON	20
2024 ANGA	20
2025 BAKUANIT	20
SI	
2026 BALEYITSI	20
2027 BAMAMUW	20
IVULA	
2028 BAMARAS	20
UNA	
2029 BAMATOLI	20
LA	
2030 BAMAWAG	20
ULA	
2031 BANYINGI	20

2032 BILANDILA	20
2033 BILEVI	20
2034 BIMINONG U	20
2035 BIMUSIEN GI	20
2036 BISUEGUL U	20
2037 BITSUNGI	20
2038 BUKULU	20
2039 BUMOAB	20
2040 BURA	20
2041 BUSINGAN YI	20
2042 CHEMECH	20
2043 DIFU	20
2044 DUGOBA	20
2045 DURANGU	20
2046 DUYAVAGA	20
2047 GUYAFUBA NA	20
2048 HILQUIA	20
2049 IFUMU	20
2050 JABANU	20
2051 JIMUTU	20
2052 JITULU	20
2053 JOSH	20
2054 KAGAJIISR AEL	20
2055 KELISA	20
2056 MAKUNGA	20
2057 MANGANA	20
2058 MASUNGA MA	20
2059 MAYINYUR U	20
2060 MBUALI	20
2061 MIMATUBU LA	20
2062 MISIRU	20
2063 MUANGUL U	20
2064 NGOGAGA	20
2065 NL	20
2066 NOKIPANG A	20
2067 NSUMBILI	20
2068 NYUBOKIS A	20
2069 OSOLA	20
2070 OVEGA	20
2071 PENDA	20
2072 PI	20
2073 RAMBUGA	20
2074 SANTIMET ERA	20
2075 TSATSABI	20
2076 TSIMUTU	20

2077 TSISEDER A	20
2078 TSIVEMA	20
2079 TSIWELIMI NA	20
2080 TUBA	20
2081 TUJI	20
2082 UBURA	20
2083 UMBU	20
2084 UNEBAT	20
2085 UROMBA	20
2086 USU	20
2087 USUNDA	20
2088 UTSIEMUG A	20
2089 VIAGA	20
2090 VIRI	20
2091 WATSI	20
2092 WE	20
2093 WIMBILAN U	20
2094 YIEVI	20
2095 YIFUFU	20
2096 YIMBEMBU	20
2097 YIMUKONG U	20
2098 YISIEYI	20
2099 ABAGEKIPI	19
2100 ABEKELISI	19
2101 ABICHAJ	19
2102 ABOBA	19
2103 ABUENU	19
2104 ADIGAA	19
2105 ADIUBA	19
2106 AGULU	19
2107 AMABAND A	19
2108 AMADUAR A	19
2109 AMANDI	19
2110 AMARUGA	19
2111 ANNE	19
2112 BACHA	19
2113 BADUKITSI	19
2114 BAESAU	19
2115 BALORA	19
2116 BAMAMUT SIGA	19
2117 BAMANASS E	19
2118 BAMAPALA	19
2119 BAMAYENA GABUGA	19
2120 BIBULONG U	19
2121 BIBUVAGU LITSI	19
2122 BIDUNA	19



2123 BISA	19
2124 BISIENGU	19
2125 BOFU	19
2126 BUNUNU	19
2127 COR	19
2128 DIBAANA	19
2129 DIMASUMU	19
2130 DUILABI	19
2131 DUKAGA	19
2132 EWENDI	19
2133 GANGA	19
2134 GUJABA	19
2135 GUKABATS	19
INGULA	
2136 GUMALABA	19
2137 IGA	19
2138 IIC	19
2139 IL	19
2140 JILSRAEL	19
2141 KOPA	19
2142 KUMFU	19
2143 LASA	19
2144 LUC	19
2145 LULU	19
2146 LUMBU	19
2147 MABULI	19
2148 MAKONGU	19
2149 MALOOGU	19
2150 MAMBI	19
2151 MASAKAM	19
A	
2152 MBIJI	19
2153 MBILA	19
2154 MIGULA	19
2155 MIOLIVA	19
2156 MIUBA	19
2157 MUANSA	19
2158 MUGAI	19
2159 MULONSI	19
2160 NAMOOYI	19
2161 NGANSI	19
2162 NOLABA	19
2163 NYIGULU	19
2164 NYIMATAS	19
A	
2165 NYUBAKILI	19
SA	
2166 ORA	19
2167 OWENDA	19
2168 PINHAS	19
2169 RU	19
2170 SALEM	19
2171 SIENGI	19
2172 SISRA	19
2173 TI	19
2174 TSIEMI	19
2175 TSIFUMU	19

2176	UBINDOMB U	19
2177	UKUNSA	19
2178	ULU	19
2179	UMUANA	19
2180	UMUTU	19
2181	UNGI	19
2182	USAMBILA	19
2183	USINAI	19
2184	UTSINDUL A	19
2185	UTSIYISA	19
2186	YIBURU	19
2187	YIIGA	19
2188	YIMABAAL A	19
2189	YIMUNANA	19
2190	YINEFTALI	19
2191	YIRUBEN	19
2192	YIYINGEBA	19
2193	ABIA	18
2194	ADAM	18
2195	AGAR	18
2196	AGUTEGA	18
2197	AMADUVE GA	18
2198	AMAGANIN A	18
2199	AMAKALUG A	18
2200	AMAMBEG A	18
2201	AMARAMB UYILA	18
2202	AMASILAM A	18
2203	AMAWARIS A	18
2204	ANATOT	18
2205	ANOBURA	18
2206	ANTIQUA	18
2207	ARABIC	18
2208	BAHIVIE	18
2209	BAMADUA RA	18
2210	BAMAFUBA NA	18
2211	BAQUEHAT	18
2212	BASAMBI	18
2213	BATEGA	18
2214	BATYITSI	18
2215	BIBUSUSU	18
2216	BLACK	18
2217	BOOK	18
2218	BOOKMAN	18
2219	BUGALAAD	18
2220	COMIC	18
2221	CORSIVA	18

2222 DIETU	18
2223 DIGOBA	18
2224 DIMAMBA	18
2225 DIMUTU	18
2226 DUBAISRA EL	18
2227 DUBURU	18
2228 EGENGILI	18
2229 EWARISI	18
2230 FRANKLIN	18
2231 GARAMON D	18
2232 GEORGIA	18
2233 GUMA	18
2234 HAETTENS CHWEIL+	18
2235 HECHEBO N	18
2236 HELL	18
2237 HITT	18
2238 IIIII	18
2239 IMPACT	18
2240 JEZABEL	18
2241 JIMUSAMB UALI	18
2242 JINDAGU	18
2243 LENSANU	18
2244 MADULOM BILI	18
2245 MAJERUSA LEM	18
2246 MAMAVIND AMA	18
2247 MAMBANG U	18
2248 MAMOYI	18
2249 MAPENGA NA	18
2250 MAYIBULU	18
2251 MAYINYEN SULU	18
2252 MBIRA	18
2253 MEDIUM	18
2254 MIBAISRAE L	18
2255 MIBINSABI	18
2256 MILORA	18
2257 MONOTYP E	18
2258 MUKASA	18
2259 MULENDU	18
2260 NARROW	18
2261 NDONSI	18
2262 NOWENDA	18
2263 NSORUM	18
2264 NY	18
2265 NYIMAJAB A	18

2266 NYIMAVAYI LA	18
2267 NYUFUNIS A	18
2268 NYUJABA	18
2269 NYULABA	18
2270 NYUPANSA	18
2271 OLD	18
2272 OTSANAN GA	18
2273 SANGILAN U	18
2274 SODOME	18
2275 STYLE	18
2276 SYLFAEN	18
2277 TREBUCHE T	18
2278 TSAMBU	18
2279 TSANA	18
2280 TSIUNU	18
2281 UAARON	18
2282 UANDI	18
2283 UBUNGAN A	18
2284 UGU	18
2285 UMBOKA	18
2286 URIE	18
2287 URONDA	18
2288 USALILA	18
2289 UTSIEMUS A	18
2290 UWAGULA	18
2291 VAGANU	18
2292 YAYA	18
2293 YIBIATSI	18
2294 YIGU	18
2295 YIKONGU	18
2296 YIMBONGIL A	18
2297 YIMUBEJI	18
2298 YIPELA	18
2299 YISA	18
2300 YITAJI	18
2301 YIZABULO N	18
2302 ZOROBABE L	18
2303 ABAGU	17
2304 AEE	17
2305 AG	17
2306 AGEJABI	17
2307 AGEKIPI	17
2308 AGUBEMB A	17
2309 AGUNENI	17
2310 AHIA	17
2311 AHIMAAS	17
2312 AMAGABU	17

	SA	
2313	AMAGABU	17
	SILA	
2314	AMAGULUL	17
	U	
2315	AMALENSA	17
2316	AMAMUBU	17
	RILA	
2317	AMARU	17
2318	AMARUYIL	17
	A	
2319	AMASIALA	17
2320	ANI	17
2321	ANOTOLA	17
2322	ATU	17
2323	BABIFUMB	17
	A	
2324	BAJI	17
2325	BALEVI	17
2326	BAMAKAKA	17
	NA	
2327	BAMAKOTI	17
	SA	
2328	BAMANEN	17
	GILA	
2329	BAMANYE	17
	NSA	
2330	BAMASILA	17
2331	BAMAWEN	17
	DANGA	
2332	BANOMAN	17
	A	
2333	BIBATOSIN	17
	I	
2334	BIBUGA	17
2335	BILI	17
2336	BILILA	17
2337	BIMOSI	17
2338	BINGANA	17
2339	BINSI	17
2340	BOTSIEMU	17
	GA	
2341	BUEDOM	17
2342	CB	17
2343	CHELA	17
2344	CYRUS	17
2345	DIGUGA	17
2346	DILENGI	17
2347	DIMABOTI	17
2348	DINOVIQG	17
	A	
2349	DUIVOSI	17
2350	DUKAVAGA	17
2351	DUKULA	17
2352	DUMAVAG	17
	A	
2353	DUNDAGU	17
2354	DUNSIENSI	17

2355 EGULU	17
2356 EWIVULI	17
2357 FURU	17
2358 GUDILA	17
2359 GUITASI	17
2360 GUKA	17
2361 GUMABA	17
2362 GUMAVAN	17
GANA	
2363 GUVAGA	17
2364 HA	17
2365 JOEL	17
2366 LILANU	17
2367 MAAKA	17
2368 MABE	17
2369 MAKULINI	17
2370 MAMUSIEN	17
GI	
2371 MBIKA	17
2372 MBISU	17
2373 MIRIAM	17
2374 MONU	17
2375 MUBUNGIT	17
SI	
2376 NDEN	17
2377 NYUVAYIL	17
A	
2378 OBADIA	17
2379 OBUILA	17
2380 OLASA	17
2381 ONYENSA	17
2382 OOU	17
2383 OPALA	17
2384 RIC	17
2385 RIEFUMU	17
2386 SAMBA	17
2387 SENNAKER	17
IB	
2388 SICLAG	17
2389 SILA	17
2390 SINSÁ	17
2391 TSIKULINI	17
2392 TSITONDIN	17
I	
2393 TUIRONDI	17
2394 UBIGA	17
2395 UDEDA	17
2396 UDEDILA	17
2397 UDUMA	17
2398 UGA	17
2399 UGAGANA	17
2400 UJOB	17
2401 UN	17
2402 UNDAGU	17
2403 UNIL	17
2404 UPHARAO	17
N	

2405 USIELASA NA	17
2406 UTSINGUL A	17
2407 UTUBULA	17
2408 UVAYILA	17
2409 UVURA	17
2410 UYE	17
2411 VHINGWE	17
2412 WAKA	17
2413 WENDANG A	17
2414 YIBULONG U	17
2415 ABAMAFU	16
2416 ABEGUVINI	16
2417 ABEJABI	16
2418 ABESALI	16
2419 AGUMAKA MBULU	16
2420 AMABAWA KISA	16
2421 AMADIMBA	16
2422 AMAGU	16
2423 AMAJI	16
2424 AMAMUET SINGULA	16
2425 AMANSINDI GA	16
2426 AMAPEGA	16
2427 AMAROND ANGA	16
2428 AMASUND UGA	16
2429 AMATSANI NANGA	16
2430 AMAYETSA NA	16
2431 AMOBA	16
2432 ASSAEL	16
2433 AYIKAVU	16
2434 BABAAL	16
2435 BAEGYPTE	16
2436 BAGEKIPI	16
2437 BAGORUN GULA	16
2438 BAMALENS A	16
2439 BAMAMUW AKA	16
2440 BAMARUN GA	16
2441 BAMASOLA	16
2442 BAMATIND A	16
2443 BAMAVOS ANGA	16
2444 BAMAYETS INGULA	16

2445	BAMONDI	16
2446	BATCHEBA	16
2447	BEGAMUYI	16
2448	BEJANGIN GI	16
2449	BEVAYINGI	16
2450	BIANSA	16
2451	BIUMANGA	16
2452	BLL	16
2453	BUBISI	16
2454	BUDILU	16
2455	BUNGUNU	16
2456	CORE	16
2457	DFFIANDU	16
2458	DIMA	16
2459	DIMUNONG U	16
2460	DIN	16
2461	DISIYI	16
2462	DIWELA	16
2463	DMANDU	16
2464	DULABANG A	16
2465	DULIOMUS U	16
2466	DUNDA	16
2467	DUNSAMBI	16
2468	DUPASU	16
2469	DUTSANAN GA	16
2470	EKILISI	16
2471	ELAM	16
2472	ELCANA	16
2473	ETHIOPIE	16
2474	EWAKI	16
2475	EZEK	16
2476	GAZA	16
2477	GUGEJABI	16
2478	HANAN	16
2479	HARAN	16
2480	ICHEBAAL	16
2481	IH	16
2482	IOGU	16
2483	JIGONA	16
2484	JIMALUNG U	16
2485	JIMOAB	16
2486	JISINAI	16
2487	JOAKIN	16
2488	LIBAN	16
2489	MAGALA	16
2490	MB	16
2491	MIBINDOM BU	16
2492	MIG	16
2493	MIKATA	16
2494	MIPIGU	16



2495 NAHOR	16
2496 NDOMBILI	16
2497 NENGILA	16
2498 NONU	16
2499 NSINGISI	16
2500 NYUDUVE	16
GA	
2501 NYUWAKIS	16
A	
2502 NYUWARIS	16
A	
2503 OBEGA	16
2504 OG	16
2505 ONADAB	16
2506 OVAYILA	16
2507 QUEILA	16
2508 SANGA	16
2509 SANSAMA	16
2510 SARAI	16
2511 SINGA	16
2512 T'WENDIAN	16
U	
2513 TAMBUSA	16
2514 TIRSA	16
2515 TSIOOYEN	16
SA	
2516 TSITOSINI	16
2517 TSONA	16
2518 TUWENDA	16
2519 UBINGA	16
2520 UGABUSA	16
2521 UKILISA	16
2522 URUMANG	16
A	
2523 USOLA	16
2524 UTEGA	16
2525 UTOLA	16
2526 UYI	16
2527 VAGULITSI	16
2528 VULA	16
2529 XERXES	16
2530 YABECH	16
2531 YIBAISRAE	16
L	
2532 YIBILU	16
2533 YIFETA	16
2534 YII	16
2535 YIKABU	16
2536 YIKUMU	16
2537 YIMUINA	16
2538 YIYIJUNGI	16
2539 ABAMARU	15
YILA	
2540 ABAMASOL	15
U	
2541 ABETS	15
2542 ABEVOSI	15

2543	AGELABI	15
2544	AGOTAMB A	15
2545	AJIMUGUM I	15
2546	AJIUBA	15
2547	ALA	15
2548	AMABASA	15
2549	AMABURA NGA	15
2550	AMADUGU SU	15
2551	AMAGUSU GA	15
2552	AMAGUSU LA	15
2553	AMAKELAS A	15
2554	AMAKEVE GA	15
2555	AMAPATUL A	15
2556	AMAROMBI SA	15
2557	AMASINGA	15
2558	AMATUNG A	15
2559	AMAVIOGA	15
2560	AMIMABA	15
2561	AMNON	15
2562	AMON	15
2563	ASSAF	15
2564	ATSANA	15
2565	AYAJI	15
2566	BAA	15
2567	BABAANA	15
2568	BABISI	15
2569	BABULONG U	15
2570	BADAN	15
2571	BAHEBREU	15
2572	BALAMBI	15
2573	BAMAKIND A	15
2574	BAMALABIL A	15
2575	BAMASOG AMA	15
2576	BAMATOG A	15
2577	BAMAYETS ANA	15
2578	BANOLABA	15
2579	BANS	15
2580	BAPERIZIE	15
2581	BAPITA	15
2582	BASAMAG ULU	15

2583	BASIAMUN U	15
2584	BASIMEON	15
2585	BATOYITSI	15
2586	BAZABULO N	15
2587	BESANGILI	15
2588	BIBISI	15
2589	BIDUMI	15
2590	BIFETA	15
2591	BIGAA	15
2592	BIKADI	15
2593	BINUNU	15
2594	BISANGA	15
2595	BITAJI	15
2596	BITSIGA	15
2597	BIVIOVI	15
2598	BOBEGU	15
2599	BOKOTA	15
2600	BORUGA	15
2601	BUM	15
2602	BUSUSU	15
2603	CHALLOUM	15
2604	CONSOLE	15
2605	DARIUS	15
2606	DIDIMUNGI	15
2607	DIGA	15
2608	DIGUYI	15
2609	DIMUJAMB A	15
2610	DINONGA	15
2611	DITUJI	15
2612	DIVINA	15
2613	DUENU	15
2614	DUMA	15
2615	DUMUTU	15
2616	DUONYI	15
2617	ECRON	15
2618	GUKAVOS A	15
2619	GUMAVOSI LA	15
2620	HAMAT	15
2621	IONA	15
2622	IVU	15
2623	JALALA	15
2624	JIMUTUBU	15
2625	KEBANA	15
2626	KELISANU	15
2627	KWAKU	15
2628	MABASEFU	15
2629	MABENYI	15
2630	MABIVUND A	15
2631	MAGUMIDU FU	15
2632	MAKAGA	15

2633	MAKASA	15
2634	MAMASUK A	15
2635	MAMIGAGA	15
2636	MANGILA	15
2637	MASU	15
2638	MECHOULL AM	15
2639	MIKILU	15
2640	MILEMBU	15
2641	MIMI	15
2642	MULONGA	15
2643	MURELA	15
2644	MUSAYI	15
2645	MUTSORIT SI	15
2646	NOBA	15
2647	NYILABING I	15
2648	NYIVINI	15
2649	NYUGABU SA	15
2650	NYUTINDA	15
2651	OBILA	15
2652	OBOKU	15
2653	ONGU	15
2654	ORUGA	15
2655	OTSU	15
2656	OYIRONDI	15
2657	PNTXTB	15
2658	PUGU	15
2659	RUBEN	15
2660	SERAYA	15
2661	SIBA	15
2662	SOLANU	15
2663	TABA	15
2664	TERA	15
2665	TLDOT	15
2666	TSIBANGO MBI	15
2667	TSIMAMAN YI	15
2668	TSINDAGU	15
2669	TSUVA	15
2670	UBELA	15
2671	UDUVEGA	15
2672	UJANGANG A	15
2673	UKADI	15
2674	UMATA	15
2675	USEROUIA	15
2676	UVANGAN A	15
2677	UYINYENS ULU	15
2678	VAGOMUE BA	15
2679	VI	15

2680 W	15
2681 YETSINGU LA	15
2682 YIASSER	15
2683 YIBALUMI	15
2684 YIDIBADI	15
2685 YIISSAKAR	15
2686 YIKUNGA	15
2687 ABAMASIA LA	14
2688 ABEWENDI	14
2689 ABIGAIL	14
2690 ABIHOU	14
2691 ADUGARI	14
2692 AGUMA	14
2693 AGUMABU RA	14
2694 AHITOFEL	14
2695 AMABAMB ANA	14
2696 AMABUNG A	14
2697 AMADIBUL A	14
2698 AMAGAMU YILA	14
2699 AMAKWIVU LA	14
2700 AMALEC	14
2701 AMALILA	14
2702 AMAMUNE NGILA	14
2703 AMASUNDI LA	14
2704 AMATSIEM USA	14
2705 ATHALIE	14
2706 AYEPALA	14
2707 AYIGAA	14
2708 BAASSER	14
2709 BAGUERC HON	14
2710 BAMAGUS UGA	14
2711 BAMAKAM BULU	14
2712 BAMAMUB EGA	14
2713 BAMASALA	14
2714 BAMASALA NGA	14
2715 BAMATIGA NGA	14
2716 BAMERARI	14
2717 BANGUBU	14
2718 BASA	14
2719 BASAMABA	14
2720 BAVINII	14

2721	BEBUSI	14
2722	BEGA	14
2723	BIDUFUNU	14
2724	BIFUFU	14
2725	BIGEYI	14
2726	BIKUTU	14
2727	BIMAWETS UGU	14
2728	BIO	14
2729	BIRODI	14
2730	BITSI	14
2731	BITUNGU	14
2732	BOSALANG A	14
2733	BOTI	14
2734	BUBAPHILI STIN	14
2735	BUGA	14
2736	BUMUTU	14
2737	CHEBA	14
2738	DALILA	14
2739	DIBAGA	14
2740	DIJERUSAL EM	14
2741	DIMUBU	14
2742	DIMUJI	14
2743	DIPANGINI	14
2744	DOOA	14
2745	DUBANSA	14
2746	DUGUSU	14
2747	DUITASI	14
2748	DUIVAYI	14
2749	DUL	14
2750	DUMALABA	14
2751	DUWENDA	14
2752	EDEN	14
2753	ELEK	14
2754	ENENGILI	14
2755	ERONDILI	14
2756	ETSANI	14
2757	ETUBULI	14
2758	GOLIATH	14
2759	GUIRONDIL I	14
2760	GUMABUS A	14
2761	GUMAVAN GININA	14
2762	GUMAVAYI LA	14
2763	HEBREU	14
2764	IBAGORA	14
2765	JEAN	14
2766	JIBAANA	14
2767	JIBISI	14
2768	JIKUNGA	14
2769	JIMAPA	14

2770	JIMBELA	14
2771	JIMUKUTA	14
2772	JITAJI	14
2773	JUIU	14
2774	KILINGU	14
2775	KOLI	14
2776	LITSI	14
2777	MAALI	14
2778	MAGETU	14
2779	MAN	14
2780	MBAMI	14
2781	MILUTSU	14
2782	MINU	14
2783	MOKALUG	14
	A	
2784	MUSUNDIT	14
	SI	
2785	ND	14
2786	NDEMBU	14
2787	NGEBA	14
2788	NGERONDI	14
2789	NON	14
2790	NS	14
2791	NYIFU	14
2792	NYIMAGUS	14
	OLA	
2793	NYIMARUM	14
	A	
2794	NYIMAVOSI	14
	LA	
2795	NYIRANGI	14
	MINI	
2796	NYIVEYI	14
2797	NYUBALAS	14
	A	
2798	NYUBANG	14
	ANGA	
2799	NYUDUGU	14
	SU	
2800	NYUDULAS	14
	A	
2801	NYUKALUS	14
	A	
2802	OESI	14
2803	OLABANA	14
2804	OOI	14
2805	OONI	14
2806	ORSA	14
2807	OTABULILA	14
2808	POLU	14
2809	Q	14
2810	QUEHAT	14
2811	SIKIRI	14
2812	SOLA	14
2813	TSA	14
2814	TSIBI	14
2815	TSIMFURU	14
	LU	

2816	TSIMIURU	14
2817	TSIRL	14
2818	TSIUKELISI LA	14
2819	TULABA	14
2820	TUMALABA	14
2821	TURUNGUL A	14
2822	UBAAL	14
2823	UBALASA	14
2824	UFURA	14
2825	UGABUSIL A	14
2826	UGUKIGA	14
2827	UGUMBA	14
2828	UKANDA	14
2829	UKEVEGA	14
2830	ULA	14
2831	ULINGA	14
2832	UMULUMI	14
2833	UPALISA	14
2834	UROYISA	14
2835	URUNGULI LA	14
2836	UWARILA	14
2837	UYEJI	14
2838	VAYILA	14
2839	VUNDA	14
2840	YEARIM	14
2841	YIBABAALA	14
2842	YIBISI	14
2843	YILEVI	14
2844	YIMUTU	14
2845	YISIMEON	14
2846	ABAMABE GANGA	13
2847	ABAMABO KU	13
2848	ABAMARO NDA	13
2849	ABAMATSA NINANG+	13
2850	ABASA	13
2851	ABIMAVEG U	13
2852	AGUISIELIS INI	13
2853	AGUMARU NGISA	13
2854	AGURUNG ULA	13
2855	AKI	13
2856	AMABAGA BILA	13
2857	AMABANG ANGA	13
2858	AMABENG USANA	13



2859	AMAKELIS A	13
2860	AMALABILI LA	13
2861	AMAMANIN A	13
2862	AMANDAS A	13
2863	AMARAMB UGA	13
2864	AMARASU NA	13
2865	AMATSANA NGA	13
2866	AMATUBUL A	13
2867	AMAYE	13
2868	ASAMARU NGULA	13
2869	AYIYI	13
2870	BAFENETE RA	13
2871	BAI	13
2872	BAKUNGA	13
2873	BAMAJANG ANGA	13
2874	BAMALILA	13
2875	BAMARELA MA	13
2876	BAMARON DA	13
2877	BAMARUM UNGU	13
2878	BAMASON A	13
2879	BAMASUN SA	13
2880	BAMATALA	13
2881	BAMFURUL U	13
2882	BANSIENG A	13
2883	BANYI	13
2884	BASAMADE DILA	13
2885	BAT	13
2886	BEDILI	13
2887	BELA	13
2888	BETSANINI	13
2889	BIBUFUMU	13
2890	BIDIONYI	13
2891	BIJIONYI	13
2892	BILOMBI	13
2893	BILSRAEL	13
2894	BIPUNGA	13
2895	BOKA	13
2896	BOLASU	13
2897	BOVAGA	13
2898	BOWENDA	13

2899 BOWIVULA	13
2900 BULO	13
2901 C	13
2902 CHADRAC	13
2903 CHAFAN	13
2904 DIBATEGU LA	13
2905 DILANGI	13
2906 DILENSA	13
2907 DIU	13
2908 DIVIOGULU	13
2909 DUASA	13
2910 DUBAANA	13
2911 DUGEJABI	13
2912 DUGEYI	13
2913 DUJEGUSU	13
2914 DUTOLILA	13
2915 DUVAGA	13
2916 DUVAGULI LA	13
2917 DUVIAKUS UNU	13
2918 DUYASALA	13
2919 EBER	13
2920 EJI	13
2921 ELA	13
2922 ELICHAMA	13
2923 EPALI	13
2924 EUPHRATE	13
2925 FUM	13
2926 GULITSI	13
2927 GUVOSA	13
2928 HASSOR	13
2929 HOUCHAI	13
2930 HULONGU	13
2931 IANDI	13
2932 IENU	13
2933 IOTSU	13
2934 ISU	13
2935 IULA	13
2936 JANDU'GU	13
2937 JETHRO	13
2938 JIBAISRAE L	13
2939 JISUNGAM A	13
2940 KAGAJIBAB YLONE	13
2941 KENGILI	13
2942 KOTANU	13
2943 KWAMUSI	13
2944 LAKICH	13
2945 LAMBI	13
2946 MADIBA	13
2947 MAEGYPTE	13
2948 MAFU	13
2949 MAGONA	13

2950	MAKA	13
2951	MAKOJU	13
2952	MALI	13
2953	MALIMA	13
2954	MAMAKUT U	13
2955	MAMANGO LU	13
2956	MAMBONG U	13
2957	MANNA	13
2958	MANOA	13
2959	MANOVIOG A	13
2960	MASAMBA KA	13
2961	MATEGA	13
2962	MATODI	13
2963	MIGA	13
2964	MIGERANA DA	13
2965	MILANSA	13
2966	MISUNDA	13
2967	MONDI	13
2968	MUSEDER A	13
2969	MUVAYITSI	13
2970	NABOT	13
2971	NDILA	13
2972	NESL	13
2973	NGEJABI	13
2974	NOJABA	13
2975	NYIGUVEYI	13
2976	NYILABILI	13
2977	NYIMABAV EGA	13
2978	NYIMARON DA	13
2979	NYL	13
2980	NYUGUYIT SANGA	13
2981	ODE	13
2982	OFUILA	13
2983	OGUSULU	13
2984	OUZA	13
2985	PUASASA	13
2986	SIMEON	13
2987	SOUKOT	13
2988	TAKA	13
2989	TOBIA	13
2990	TSIAKASIA	13
2991	TSIGAA	13
2992	TSIGUMUG A	13
2993	TSIMOAB	13
2994	TUILABI	13
2995	TWENDIAN U	13

2996 UA	13
2997 UBABOKA	13
2998 UHOR	13
2999 UISRAEL	13
3000 UJESSE	13
3001 UJUDA	13
3002 UKAMBULU	13
3003 UKIPA	13
3004 UKUNSAM	13
A	
3005 ULUBUSA	13
3006 UMUISI	13
3007 UPATULA	13
3008 URUNGILA	13
3009 USABUGA	13
3010 UTALU	13
3011 UTSANANG	13
A	
3012 UVIOYILA	13
3013 UWIMBILA	13
3014 UYETSANA	13
3015 UYETUMU	13
NA	
3016 VIOGA	13
3017 WAKISA	13
3018 WENU	13
3019 WIVULA	13
3020 YIKAGA	13
3021 YILEMONDI	13
3022 YILI	13
3023 YIMAGETU	13
3024 YIMUGETU	13
3025 YINGUMA	13
3026 YITAYI	13
3027 YIVARU	13