CHAPTER 1

INTRODUCTION

1.1 Introduction

For many decades Afrikaans and English were the only official languages in South Africa. After the first democratic elections in 1994 the number of official languages was increased to 11 when Sepedi, Sesotho, Setswana, siSwati, Tshivenda, Xitsonga, isiNdebele, isiXhosa and isiZulu were also granted official status. It cannot be taken for granted that these nine African languages, which previously did not enjoy the status of official languages, will automatically fulfill the requirements of an official language in all spheres of their application. In most cases considerable language development strategies will most likely be required to ensure that these languages can be used at all functional levels, especially in areas such as science, commerce, governmental communication, education, etc.

1.2 Research questions

The primary research question this study will aim to answer is whether Sepedi is able to function comfortably as a medium of communication in all the higher domains of life as listed above.

The second research question emanates from the first, and has to do with the sources consulted to ascertain whether Sepedi does indeed have the lexical capacity to fulfill all the mentioned functions. This question then centres around the quality of the existing English-Sepedi dictionaries; more specifically their adequacy as reference sources, i.e. the

way in which these dictionaries reflect the linguistic and communicative reality of Sepedi as used in higher communicative functions.

1.3 Objectives of the study

The purpose of this research is to establish to what extent the Sepedi language has the potential of expressing those concepts typically found in areas such as academic, commerce, the news media, the civil service, law and education. The words denoting these concepts will be referred to as "high function words". High function words are typically found in academic literature, manuals, newspapers and magazines, advertisements and brochures, religious literature and prose. It will be argued that the ability of the Sepedi language to express such high function concepts will be a first indication of the language development required to equip Sepedi for its various roles as a fully-fledged official language.

In order to determine the capacity of Sepedi to express higher function concepts, a measuring instrument had to be found. The logical first step would have been to compile a Sepedi corpus on the basis of texts used in domains such as science, commerce and education. The analysis of such a corpus by means of certain sophisticated query tools would then have provided answers to many questions regarding the capacity or incapability of the language to express high function concepts. The compilation of such a corpus for Sepedi was indeed attempted in following a genre-based approach. However, due to the lack of written materials in Sepedi, especially in domains such as government communication, advertisements and brochures, manuals, magazines and newspapers in Sepedi, this route was not a viable one.

A second possible way to answer the primary research question could be to translate texts from a language of wider communication (e.g. English) into Sepedi, to record the lexical gaps and other lexical problems to find suitable translation equivalents by either field

research or consultation of bilingual Sepedi-English dictionaries. Although this method is reliable, it would have been very time-consuming and the results could be greatly influenced by factors such as the translator's proficiency in both the source and the target language.

The researcher chose a third option, namely to develop an alternative instrument of measurement/evaluation by using a second language renowned for its ability to act comfortably in higher functions. Languages such as English, Afrikaans, German and French would qualify, but on the basis of the following reasons English was selected as the measuring instrument:

- The researcher knows English better than any of the other languages
- English can be regarded as ideologically neutral
- English is the lingua franca of South Africa
- English has a well developed vocabulary at all functional levels (especially words used in higher functions).

A number of English high-function words, selected on the basis of their frequency and spreading across sources, were identified as denominators of important high function concepts. It was argued that if adequate translation equivalents for these English words could be found in Sepedi-English dictionaries, the result could be a first indication of whether Sepedi was capable of fulfilling the functions typically associated with official languages (i.e. languages used across a wide spectrum of functions and in a variety of contexts). The results could also give an indication of the nature and extent of lexicographical work needed to be done with regard to bilingual Sepedi-English dictionaries.

1.4 Methodology

1.4.1 Theoretical framework

The researcher followed a functional approach which demanded that more than one theoretical paradigm had to be invoked, namely theories of language planning, theories and models of corpus-building and theories of bilingual lexicography.

1.4.2 Data collection

A genre-based approach was followed in the compilation of a Sepedi corpus (using English as a measuring instrument). Similar studies have been undertaken in English in the compilation of the Cobuild Corpus by prominent international lexicographers (Sinclair 1987), the LOB corpus (Hofland and Johansson 1989) etc. but they were mainly based on topics, not genre per se.

A corpus of English - as explained in paragraph 1.3 - consisting of certain text categories normally associated with higher function usage of the English language was compiled. It comprised of the following text categories:

- academic texts consisting of selections from student notes, handbooks and subject manuals
- advertisements and brochures
- Bible texts
- magazines and newspapers
- operating manuals of household appliances, vehicles, etc. and
- a selection of prose and poetry.

The English data was analyzed primarily in two ways, namely

- calculating the word frequency totals, as well as the spreading of words over the different source categories; and
- studying the words in context in a concordance layout.

A number of English words which may be regarded as high function words were selected to form the basis for the subsequent evaluation of Sepedi.

Firstly the treatment of these words, or the lack thereof, in English-Sepedi dictionaries was evaluated. Those words which were *not treated* in Sepedi-English dictionaries were isolated. In order to establish whether suitable translation equivalents could be found or coined for those English high function words with no translation equivalents in the existing Sepedi bilingual dictionaries, a small survey was conducted. It was done by sending out questionnaires to a number of Sepedi mother tongue respondents (see appendix 8 for a copy of questionnaire). Being a mother tongue speaker of Sepedi, the researcher used his intuition to evaluate and augment the responses.

The translation equivalents of those words which were indeed treated in English-Sepedi dictionaries were evaluated in order to determine whether suitable Sepedi equivalents for high function purposes were given.

1.5 Preview

The study consists of 6 chapters which have been arranged in the following way:

Chapter 2 deals with language policy and the revalorisation (development) of the autochthonous languages. It also deals with the way the indigenous languages were developed from the time they were recognized as official languages of the Republic of South Africa. This chapter also touches on constitutional principles which are relevant to language policy and language stipulations as they appear in the new South African

Constitution of 1996. It also discusses different aspects of language planning, namely corpus planning, status planning and acquisition planning.

Chapter 3 deals with electronic corpora as authentic sources of information on the vocabulary of a language. In order to glean information on the principles of corpus-building, three English corpora were studied, namely the Cobuild Corpus, the Lancaster Oslo Bergen (LOB) Corpus, as well as the Longman-Lancaster English Language Corpus. These corpora are compared in order to determine the most important principles of corpus-building, criteria for the selection of text categories and spreading of words across different sources, and the generation of concordance lines. Thereafter a schematic comparison is made between these three corpora in order to determine their similarities and differences.

This chapter also discusses selection criteria for the compilation of an English corpus (i.e. a measuring instrument as explained in 1.2). It aims at the selection of text categories which will be used to determine whether Sepedi is capable of expressing high function concepts. The selection of the text categories is as follows: academic literature, advertisements and brochures, the Bible, magazines and newspapers, manuals and prose. These texts were scanned, analysed and interpreted in terms of especially overall word frequency counts and spreading across sources. This section is followed by a discussion of the high function words with the lowest and highest frequencies in all the categories, for the purpose of analysing their treatment in Sepedi bilingual dictionaries, namely the *New English Northern Sotho dictionary* (NEND), *New Sepedi dictionary* (NSD) and the *Northern Sotho Terminology and Orthography* (NTO).

Chapter 4 focuses firstly on the theory of bilingual lexicography, as the treatment of lexical items in bilingual dictionaries in this chapter demands knowledge of and insight into bilingual lexicography. The main emphasis is on the principle of equivalence.

Different equivalence relationships are investigated to develop diagnostic tools for analysing the treatment of the English high function words in English-Sepedi dictionaries.

Following a tabulated exposition of the high-function words selected for investigation, an in-depth analysis is made of the meaning and use of each of these words, as they appear in the *Concise Oxford Dictionary* and the *Oxford English Dictionary*.

The primary focus of this chapter is the actual treatment of the selected English high function words in Sepedi-English dictionaries. The critical evaluation of the Sepedi translation equivalents and equivalent discriminating information is done against the background of:

- the treatment of the English high function words in the above-mentioned monolingual English dictionaries;
- the linguistic (especially semantic) properties of the English words as demonstrated by their occurrence in concordance lines;
- the mother-tongue competence of the researcher regarding the selection, meaning and use of the Sepedi equivalents.

Chapter 5 deals with lexical gaps in Sepedi at high function levels, focusing on words from the English database which are not treated in Sepedi bilingual dictionaries. It also deals with the responses of subjects (mother-tongue speakers of Sepedi) to a questionnaire on possible Sepedi translation equivalents for those English high function words not entered in the macrostructures of Sepedi bilingual dictionaries and the *Northern Sotho Terminology and Orthography*.

Chapter 6 provides an overview of the study and makes recommendations for further research into lexicological and lexicographic matters related to high function words in Sepedi.

CHAPTER 2

LANGUAGE POLICY AND THE REVALORIZATION OF THE AUTOCHTHONOUS LANGUAGES

2.1 Introduction

This chapter gives an overview of the policies and practices of language planning and their roles in creating a context that is conducive to language development.

Before the democratic elections in South Africa, English and Afrikaans were the only two official languages. The indigenous languages such as Sepedi, Sesotho, Setswana, siSwati, Tshivenda, Xitsonga, isiNdebele, isiXhosa and isiZulu were not officially recognised by the South African government of the time. Sepedi and other indigenous languages became officially recognised after the first democratic elections in April 1994. Sepedi, as one of the autochthonous languages, now needs to fulfill certain requirements in order for it to have the status of an official language. This implies that Sepedi, now being one of the official SA languages, needs to fulfill certain functions. It must for instance be used at all functional levels, especially the "higher levels", such as communication in government, science and technology, commerce and education. Therefore, in order for Sepedi to be revalorized (developed) and to become a fully-fledged official language of South Africa, thorough language planning is necessary.

In sections 2.2, 2.3, 2.4 and 2.5 below, language policy and practices, language principles in the constitution and language planning are discussed in detail.

2.2 Language policy and practices

The new Constitution of the R.S.A. (1996:4) stipulates that there are eleven official languages. There are nine indigenous languages amongst these official languages which need to be developed, namely, isiNdebele, isiXhosa, isiZulu, Sepedi, Sesotho, Setswana, siSwati, Tshivenda and Xitsonga. These nine languages had not enjoyed official status up to 1994, and no effort was made by the previous government to develop them as multifunctional tools.

It is important to note that the development of these languages has now become imperative, if the government sincerely believes in the constructive and empowering role that multilingualism can play, as stated in an Unesco report (Unesco, n.d., p.116):

> To promote African languages is to safeguard national independence and to provide a sounder foundation for the exercise of genuine democracy. It is also a means of liberating creative faculties in general and of giving people, mentally, deep roots in genuinely African culture. This approach to the problem means looking beyond the mere development of culture and considering language policy as a factor in political independence and a requirement for democracy. The experts were unanimously agreed that the political battle was not over until the cultural and linguistic battle had been won.

Democracy and language development go hand in hand with socio-cultural upliftment, a feeling of unity and nationalism among the speakers of a language. It is the language speakers themselves who must revalorize the language before external revalorization can be successful.

In the past, speakers of these marginalised languages were made to believe that their languages were less important than English and Afrikaans. Msimang (1991, in Webb 1995:98) states that:

> Most Blacks in South Africa have come to hate their languages and consider them irrelevant to the education process.

In sections 2.2 and 2.3 below, the language principles and stipulations of the constitution are discussed in detail.

2.3 Language principles in the constitution

The new constitution of South Africa (1996) includes 34 principles, amongst which five are relevant to language policy, namely, sections III, IX, XI, XII and XX. Two of these principles, namely sections XI and XII, make very specific reference to language. The others bear an indirect reference, but are linked to linguistic matters in the sense that they address principles related to discrimination, national unity and cultural diversity.

III. **The Constitution shall prohibit racial, gender and all other forms of discrimination and shall promote racial and gender equality and national unity.**

Among the other forms of discrimination one may include linguistic discrimination. By not discriminating against indigenous languages as the previous government did, the development of these languages will no longer be inhibited. Sepedi will thus eventually reach the position of a high function language.

IX. **Provision shall be made for freedom of information so that there can be open and accountable administration at all levels of government.**

If Sepedi as one of the official languages of South Africa is also used in all provincial government documents in the Northern Province, Gauteng and Mpumalanga, the language will acquire the promotional status it deserves.

XI. **The diversity of language and culture shall be acknowledged and protected, and conditions for their promotion shall be encouraged.**

This entails that Sepedi as one of the eleven official languages of South Africa needs to be promoted in order to be used at all higher communicative levels.

XII. **Collective rights of self-determination in forming, joining and maintaining organs of civil society, including linguistic, cultural and religious associations, shall, on the basis of non-discrimination and free association, be recognised and protected**.

This will decrease the likelihood of linguistic and cultural alienation (cf. Webb 1995:99) which contributed towards a low functional usage of a language like Sepedi.

XX. **Each level of government shall have appropriate and adequate legislative and executive powers and functions that will enable each level to function effectively. The allocation of powers between different levels of government shall be made on a basis which is conducive to financial viability at each level of government and to effective public administration, and which recognises the need for and promotes national unity and legitimate powered autonomy and acknowledges cultural diversity**.

The promotion of Sepedi and other indigenous languages to perform at a high functional level will automatically lead to the promotion of national unity (cf. Webb in preparation:54).

2.4 Language stipulations

Apart from the above-mentioned general principles, the New Constitution of South Africa (1996:66-67) contains the following stipulations which directly refer to the languages of South Africa:

6. (1) **The official languages of the Republic are Sepedi, Setswana, isiSwati, Tshivenda, Xitsonga, Afrikaans, English, isiNdebele, isiXhoza and isiZulu.**

(2) **Recognising the historically diminished use and status of the indigenous languages of our people, the state must take practical and positive measures to elevate the status and advance the use of these languages.**

(3)(a) **The national government and provincial governments may use any particular official languages for the purpose of government, taking into account usage, practicality, expense, regional circumstances, and the balance of the needs and preferences of the population as a whole or the province concerned; but the national government and each provincial government must use at least two official languages.**

This stipulation opens the possibility for Sepedi as an official language to be used in the Northern Province, Gauteng, Mpumalanga and the North West Province.

(b) **Municipalities must take into account the language usage and preferences of their residents.**

Because of the significant number of Sepedi-speaking residents the Northern Province, Gauteng, Mpumalanga and the North West

Province the municipalities of towns and cities in these provinces have to be serious about recognising its status in official and public communication. If, however, the language does not have the capacity for conveying concepts in all such domains, effective and efficient communication cannot take place. These "new" contexts of use necessitate thorough linguistic research and language planning.

(4) **The national government and provincial governments, by legislative and other measures, must regulate and monitor their use of official languages. Without detracting from the provisions of subsection (2), all official languages must enjoy parity of esteem and must be treated equitably.**

It means that Sepedi, as one of the autochthonous languages, must enjoy the same treatment and status as English and Afrikaans, and must be as "visible" in official use.

(5) **A Pan South African Language Board established by national legislation must -**

(a) **promote, and create conditions for the development and use of:**

(i) **all official languages;**

(ii) **the Khoi, Nama and San languages, and**

(iii) **sign language; and**

(b) **promote and ensure respect for :**

(i) **all languages commonly used by communities in South Africa, including German, Greek, Gujarati, Hindi, Portuguese, Tamil, Telugu and Urdu; and**

(ii) **Arabic, Hebrew, Sanskrit and other languages used for religious purposes in South Africa.**

> Although the Bible has already been translated in Sepedi and sermons are conducted in the language, there are other text types concerned with religion that have not been addressed, e.g. The Book of Concord (i.e. a religious book used by Lutherans).

The above stipulations can only be heeded if the government puts in place mechanisms to:

(i) promote the use of all the official languages, particularly the autochthonous languages, at both provincial and national levels;

(ii) develop the autochthonous languages to such an extent that they have the capacity to be used in all the various functions either explicated or implicated by the constitution and its stipulations;

(iii) monitor the progress that is made in terms of language planning and development, as well as in terms of implementing the language stipulations.

It is important to note that government policies will remain sterile theories until they are put into practice. Specific, concrete measures have to be put in place in order to make the policy effective, i.e. to bring about changes in the functional use of languages that will empower their speakers.

2.5 Language planning

As defined above, language planning is initiated and orchestrated by policy-makers. It means that authority lies in the hands of the government. Weinstein (1980, in Cooper 1989:30, 31) therefore defines language planning as:

> [...] a government authorised, long term sustained and conscious effort to alter a language itself or to change a language's functions in a society for the purpose of solving communication problems.

Government on its own, can however, not implement language policies. Experts, often academically trained professionals, have to be invoked to facilitate the process.

Firstly, the language planners appointed have to conduct research in order to identify the language problems which lie in the way of implementing the policy, and secondly possible solutions for each particular problem have to be found. Language planning may thus be regarded as a problem-solving activity, as captured by the following definition by Rubin and Jernudd (1971, in Cooper 1989:30):

> ...language planning is focused on problem-solving and is characterised by the formulation and evaluation of alternatives for solving language problems to find the best.

Once the problems have been identified, solutions to them have to be found. In this sense language planning involves the following:

- Coordinated measures taken to select,
- codify and,
- in some cases, to elaborate orthographic, grammatical, lexical, or semantic features of a language and
- to disseminate the corpus agreed upon.

  (Gorman 1973:73).

Language planning as a **problem-solving** activity therefore firstly concentrates on **WHAT** the problem is, and then on **HOW** to solve it. An example of a language problem might be WHAT language to use as a medium of instruction at primary and secondary schools, or WHAT languages should be used in courts of law.

When the WHAT question has been answered satisfactorily, the language planner should ask HOW the languages identified for these functional uses (e.g. education and law) should be equipped.

One way of facilitating the use of Sepedi as a medium of instruction in courts of law, in commerce, in government communication, etc. is to make sure that the lexicon of this language is capable of expressing all the concepts of these functional domains. The primary sources in which the lexicons of languages are represented, are dictionaries. If language planners should feel that Sepedi dictionaries either do not reflect the lexicon of the language adequately, or that these dictionaries testify to the inadequacy of the lexicon, entry points for the planning process have to be identified. This opens the agenda for a systematic process of investigation: stock-taking of the existing vocabulary of the language, determining the lexical gaps, and identifying or coining vocabulary items that might fill these gaps.

Cooper (1989:31) focuses on the **process** of planning, which partially overlaps with the problem-solving approach. He asks the following research question:

WHO plans **WHAT** for **WHOM** and **HOW**?

The question **WHO** refers to those who initiate and implement the process, namely *policy-makers* and *planners*. For example, some definitions restrict language planning to activities undertaken by governments, government-authorized agencies, or other authoritative bodies, i.e. organizations with a public mandate for language regulation. The **WHAT** refers to the focal point of language planners (compare the exposition above) or the type of planning, which might be *status planning*, *corpus planning* or *acquisition planning*. These types of planning may be done in the interest of a certain group of people – the *beneficiaries*. This will then answer the question **FOR WHOM** this planning is done. The last question to be answered is **HOW** will this planning take place? The HOW question should be answered by paying attention to:

(a) the *needs* of the speech community in whose interest the planning is done;
(b) the explication of the *goals*;
(c) the *means* and how they are tailored to these ends; and
(d) the *monitoring of results* in order to permit the adjustment of means and ends to one another (cf. Cooper 1989 :31, 35, 40).

In the South African context, the question **WHO** refers to the government. The government should refer this process of language development to the language planners as it is their main task to see to it that proper planning is put into place before a language can be developed. The **WHAT** refers to the problem itself, for example, development of Sepedi into a fully-fledged official language of South Africa. The **WHOM** refers to the people whose behaviour is to be influenced, and the **HOW** refers to the procedure to be followed in the promotion of indigenous languages such as Sepedi. This is the reason why Kennedy (1984:5) attests that:

> Language planning is future oriented. That is, the outcomes, policies and strategies must be specified in detail in advance of action taken.

Planning is needed in order to address the future development of the autochthonous languages.

Language planning involves three stages, namely (a) status planning, (b) corpus planning and (c) acquisition planning. (cf. Kloss 1969, in Cooper 1989:31)

2.5.1 Status planning

According to Cooper (1989:32), status planning refers to the allocation of languages or language varieties to given functions.

Different dialects can be used in different situations. The aim of this type of planning is to promote a language so that it may also be used for higher functions. For Sepedi in particular, in terms of the constitution, it means that this language should be promoted so that it can also be used at high function communicative levels. This will be discussed in great detail in chapters 4 and 5.

Status planning also involves the social development of the language and has to do with the attitude of the community. There are four different measures by which status planning can be facilitated, namely:

- statutory and governmental measures
- an increase in the economic value of a language
- educational value and
- the socio-cultural meaning of language

(cf. Webb 1995:104-108)

2.5.1.1 Statutory and governmental measures

There are three ways in which the government can promote a language. Firstly, political leaders can *take control of the functional distribution* of a language such as Sepedi. They can do this through the use of Sepedi in various governmental bodies, for example, courts of law, educational institutions and state controlled schools, etc. According to Cooper (1989:108, 109) the use of a language as a medium of primary or secondary education, either regionally or nationally, is a means of giving a language a high functional status.

The use of Sepedi as a medium of instruction in schools will promote the language regionally, as well as nationally.

Secondly, the government can *enforce its ideology* on the nation as a whole. This usually depends on how strong the current government of the country is. The government can make sure that it reaches this goal through enforcing language laws, policy formulations, policy directives as well as decrees. A typical example are the measures taken by the South African government during the apartheid era when it gave Afrikaans and English superior status above the indigenous languages.

Thirdly, language promotion can take place *through the use of the different official languages* in parliament, for instance, the use of Sepedi in political debates, publications and interpretation as well as translation. English and Afrikaans as the only two official languages of South Africa in the apartheid era were the only languages used in parliament and in governmental publications etc. Due to the fact that the new Constitution of South Africa includes/proclaims Sepedi and other indigenous languages as official languages, it could be expected that Sepedi has to be used in all the governmental institutions. The use of Sepedi at provincial level in parliament in the Northern Province, Mpumalanga and Gauteng will have a positive impact in giving this language higher functional status.

2.5.1.2 An increase in the economic value of a language

According to Webb (1995:106), one of the most important determinants of the fate of a language is its economic value. The only way to make the autochthonous languages economically valuable, is by giving them a higher status in the working fraternity. Alberts (1998:230) states that:

> An increase in the demand for creativity in the African languages resulted from other developments in South Africa such as the

> development of the black media (broadcasting and telecasting), the growth in the consumer market (advertising industry), and mother-tongue education at primary school level.

For Sepedi, this could be achieved by setting fluency in the language as a prerequisite for job opportunities when certain posts are advertised. This implies that knowledge of Sepedi will enable speakers to sell themselves in the market since currently many young mother tongue speakers of this language believe that English is the only means of economic empowerment.

2.5.1.3 Educational value

Webb's (1995:107) view concerning the educational value of the autochthonous languages is that:

> They gradually developed into indispensable instruments of educational development.

These languages must be used in schools and institutions in order to be developed. Hence Webb (1995:107) emphasizes that the educational development of such languages can only happen if:

- they are used as a medium of instruction first at primary school level, and then later in secondary and tertiary education;
- they are offered as school subjects and can be studied and researched at tertiary level;
- new, meaningful language syllabi for pre-tertiary education are designed;
- appropriate teaching materials and textbooks are developed; and
- effective literacy and adult training programmes are available.

This is echoed by Cooper (1989:112) who attests that:

> An excellent way to impart the indigenous languages is to use them as media of instruction.

This study concurs with the above-mentioned scholars with regard to the argument that Sepedi as one of the autochthonous languages will undergo a gradual, natural process of development if it is used as a medium of instruction in primary schools, secondary schools, as well as tertiary institutions. The process of language development will greatly be stimulated if the writing of textbooks in Sepedi is promoted, which in turn will boost the development of Sepedi as a medium of instruction in Further Education (Grades 10-12) and Higher Education (university and college level).

2.5.1.4 The socio-cultural meaning of language

It is not only the instrumental value of a language that stimulates its development, but also the socio-cultural status given to it by its speakers. Born (1992:439) states the following in this regard:

> If a language is spoken by the leading social groups in the country, it becomes a symbol of cultural identity, and if it symbolises people's link with a glorious past, then the language will be held in high esteem by its speakers.

The community plays an important role in as far as the continuous use of the language is concerned. For instance, the community may elect or nominate a committee which will specifically deal with Sepedi events. Such events will result in the community respecting

and having confidence in the language. It is important to note that people can only communicate and know their cultures through the use of their mother-tongue.

Status planning should therefore promote, and not ignore the role of a language in different social situations. The status that a language has in its own community, will to a large extent, determine the status that it will have in the eyes of potential second language learners. This aspect is called acquisition planning, and will be dealt with in 2.4.3 below.

2.5.1 Corpus Planning

Corpus planning is primarily concerned with the **WHAT** question and it involves the creation and redefinition of words. According to Cooper (1989: 31, 32, 33) corpus planning refers to activities such as coining new terms, reforming spelling and adopting a new script.

New terms should be created and coined for those words which do not have translation equivalents for a source language item. For example, if there are some English high function words which have no Sepedi translation equivalents, then new terms need to be created for them. There are also instances where certain normative language rules need to be adapted as the language develops. For instance, certain spelling rules in the *Northern Sotho Terminology and Orthography* might need to be changed as the language develops. This process of coining new terms and reforming spellings will improve the language as new scripts will be adopted.

The aim of corpus planning is to expand a language. This is also cited by Webb (1995:109) who states that:

> The aim of corpus planning is the expansion of a language to enable it to perform the (higher level) functions allocated to it.

Corpus planning is regarded as the initial step to be taken in the development of a language. This is done in order to ascertain that a language acquires the high function status it deserves. According to Webb (1995:109) this generally involves the selection of a variety to be cultivated as a standard language.

A language variety which will ultimately be called the standard language should be selected. Webb's view is in tandem with that of Cluver (1989:75) who says that:

> In a normal complex industrial (or post-industrial) society, a uniform language is needed for the mass media, for education and for government and it also serves to group people into a nation. This uniform language is generally known as the standard language.

The process of corpus planning can only succeed if done through the standard language. According to Hudson (1980:33, 34), a typical standard language will have passed through the following processes:

(i) Selection: the choice of one dialect from the many to be developed into a standard language.

(ii) Codification: the systematised fixing of the grammatical rules of the chosen language in grammar books and dictionaries, after which the members of the relevant speech community will have to learn it.

(iii) Elaboration of function: enlargement of the scope of use of the language which was chosen, so that it is now used in government circles, schools, the media, religious activities, and in literature.

(iv) Acceptance: the variety which was chosen should be accepted by the community and serve as a unifying force. (cf. also Mathumba, 1993:20, 21), (number insertion, mine).

As far as *selection* is concerned, Pedi is a dialect which has been developed into a standard language called Sepedi. In terms of *codification*, it can be argued that Sepedi as a standard language has its own grammatical rules which must be observed at all times on all levels of communication. The Pan South African Language Board has all the powers to see to it that sub-language-committees whose task it is to revise the terminology and orthography of the language are constituted for Sepedi. Thirdly, with respect to *elaboration of function*, the new constitution of the Republic of South Africa (1996:4) recognises Sepedi as one of the eleven official languages of the country, and this is why this language is now being introduced in all government circles, the media, in courts of law etc. Finally, as far as *acceptance* is concerned, Sepedi has been developed into a standard language which is officially accepted by its native speakers.

In the process of corpus planning language variation plays an important role (cf. Hudson 1980:24). Three concepts referring to aspects of variation - and the choices made between alternatives - demand attention, namely dialect, register and style.

(i) Dialect

Crystal (1969:92) defines a dialect as:

> A regionally or socially distinctive variety of a language, identified by a particular set of words and grammatical structures. (cf. Wolfram 1991:2)

Two types of dialects can be distinguished, namely social dialects and regional dialects. Calteaux (1996:39) states that social dialects are dialects which can be distinguished on the basis of non-regional differences such as social class, age, sex, status or social setting, whereas regional dialects are variations based on geographical factors.

Dialects differ from the standard variety of a language in that they are non-standard varieties. This is attested by Wardhaugh (1986:25) who says that a dialect is often equivalent to "non-standard" or even "substandard" language. On the other hand standard varieties of a language have often been derived from nonstandard varieties and the lexicons of the former are often replenished from the stock of the latter. Dialects therefore play a major role in the development of each and every standard language.

Sepedi was developed from the regional dialect Pedi. Other dialects are Lobedu, Tlokwa, Hananwa etc. In the process of corpus planning - to expand the standard language - elements of any of the dialects may be "borrowed" or adapted to fulfill a particular lexical or other need.

(ii) Register

According to Van Wyk (1992:4, 5) a register is a more or less discrete set of lexical items and expressions adapted to specific topics and social situations e.g. law, religion, history etc.

Dembetembe (1982:2) defines register in the following way:

> It is the type of linguistic variation which bears a mutual relation with context in a wide sense of the term, including both textual context and situational context.

This means that the term register refers to the different "types" of a language used in different social situations by different people (cf. Sekhukhune 1988:18, 19).

Sepedi needs to be expanded in order to fulfill the entire spectrum of linguistic functions needed by the speakers of Sepedi. This could involve the creation of new registers depending on the situation in which the speakers find themselves, e.g. a medical register

for use between doctors, nurses and representatives of other medical professions; a bureaucratic/official register for use in government offices, and between government and the public; an academic register for use if/when Sepedi is used as a medium of instruction in institutions of higher learning.

The following examples demonstrate the impact of different registers on the use of lexical items:

Example 1:

English : *Business people sell goods to their customers*
Sepedi : (*Barekiši ba rekišetša bareki diphahlo)*

Example 2:

Medical personnel look after the patients in hospitals
(*Balwetši ba hlokomelwa ke ba tša kalafo dipetleleng)*

Example 3 :

Lawyers defend their clients in the courts of law
(*Boramelao ba emela badirelwa ba bona kgorong tša tsheko)*

Example 1 shows that business people provide services to the *customers* (*bareki*). Example 2 shows that the medical personnel provide services to the *patients* (*balwetši*), and the legal people provide services to the *clients* (*badirelwa*) in example 3. Each one of these three fields has its own term for the people they serve, namely *customers* (*bareki*), *patients* (*balwetši*) and *clients* (*badirelwa*).

Each of these three fields, namely business, medical and legal provide services to the people. This may be in the form of goods, service or knowledge. Business people use the term *customer* (*bareki*) to refer to people who buy goods from them. The medical

personnel (e.g. doctors) earn a living for the services they provide to their *patients* (balwetši) but they use the term *patients* (*balwetši*) to refer to their *clients* (*bareki*). Legal personnel, for example, lawyers use the term *clients* (*badirelwa*) to refer to their *customers*. Therefore, the researcher concludes that different speech registers, namely *customers* (*bareki), patients* (*balwetši*), *clients* (*badirelwa*) are used in different situations (i.e. in business, medical and legal) but they all refer to *customer* which is *moreki in* Sepedi. Therefore, they vary situationally as the two words, *patients* (*balwetši*) and *clients* (*badirelwa*), which belong to different registers have the same meaning. The only difference is that these three speech registers vary situationally due to different situations and ethical rules (cf. Stark 1990:174).

Register and terminology are related concepts in that the registers for certain academic, technical, scientific and professional domains often derive their distinctive character from the subject-field terms used for the important concepts of this field. Terminology as such, however, falls outside the scope of this contribution and will not be dealt with further.

(iii) Style

The situation in which speakers find themselves, determines the *style* of the language. Fromkin & Rodman (1993:299) define style as a situation dialect. Hartmann & James (1998:132) define style as a variety of a particular language associated with different texts, authors, genres and oeuvres.

The style of a language is determined by the relationship between the participants during communication, the topic of the conversation and the formality of the situation. A speaker would, for instance, use a chatty style with friends in an informal situation. On the other hand the speaker would use a formal conversational style in a context where he/she does not know the other speech partner intimately (e.g. with a colleague from a neighbouring university), and when conversing on topics of academic importance such as the new dispensation in higher education.

Style parameters can often be quantified, for example, the language style used by a teacher in a classroom situation will be more formal than the style used on the tennis court.

With regard to the central theme of this thesis style is important in the sense that higher functional uses of language are normally characterised by a formal style, making use of different subject-specific registers. Language planners are under the obligation to determine whether Sepedi does indeed have the capacity to express the concepts of higher domains in a highly formalised style and register.

2.5.2 Acquisition planning

According to Cooper (1989:33) acquisition planning is:

> [...] directed at the spreading of the language. If the language spreads, the number of language users, speakers, writers, listeners or readers will increase.

It is the task of the language speakers to make sure that knowledge reaches the people in different media such as textbooks, public speeches etc., as the language develops. Hence Cooper (1989:33) argues that new users may be attracted by the new uses to which a language is put.

As the language develops, its usage also changes and this will ultimately increase the number of users. The more the language spreads, the more easily it can be acquired. For example, publication of new textbooks, compilation of new dictionaries, writing of articles, etc. will automatically spread knowledge nationally and internationally. The elevation of the status of Sepedi amongst mother tongue speakers may depend on how efficiently it handles various linguistic concepts and functions. This involves the

elaboration of the vocabulary in general, as well as the development of the various technical vocabularies of the languages.

Sepedi as one of the autochthonous languages should be revalorised in order to acquire a similar technological status as English and Afrikaans. This can only be accomplished through the creation and redefinition of Sepedi terms (i.e. certain technological words for example, computer terminology are standard world-wide-English and difficult to translate). As such, translation equivalents for Sepedi need to be created so that the technological status of the language will be the same as that of English and Afrikaans.

This exercise of creating new words and redefining and re-evaluating existing words and terms will automatically lead to the compilation of new Sepedi (monolingual, bilingual, trilingual, multilingual, etc.) dictionaries. Then in turn the language speakers or users will benefit from such an endeavour.

2.6 Conclusion

Language politics of South Africa can, simplistically spoken, be divided into two phases. The *first phase* is represented by the constitution of South Africa in the apartheid era and the *second phase* by the post apartheid constitution (1996). The language principles and stipulations in the constitution of the previous government recognised English and Afrikaans as the only two official languages of South Africa, and indigenous languages such as Sepedi were marginalised. The speakers of these languages were made to believe that their languages were inferior to English and Afrikaans, and as a result they developed a negative attitude towards their mother tongues.

The *second political phase* started with the post apartheid constitution of South Africa. It recognised the indigenous languages as well as English and Afrikaans as official languages of South Africa, thereby officially changing their status. The language

stipulations in the constitution entail that Sepedi and other indigenous languages should be promoted so that they can enjoy the same high functional status as English and Afrikaans. Status-planning for Sepedi has therefore been accomplished to a certain extent. This kind of language planning must, however, be followed up by government through the promotion and sanctioning of the autochthonous languages as languages of further and higher education.

Efficient status planning makes it possible for acquisition planning to take place without any hindrance. Users may acquire this language (more fully) through speaking and reading textbooks written in Sepedi. Lexicographers and terminographers can also elevate the status of the language more successfully amongst mother-tongue speakers by compiling Sepedi monolingual dictionaries, a dictionary type which currently do not exist in Sepedi. The existing Sepedi bilingual dictionaries can also be improved, not only to assist students and translators, but also to ascertain that Sepedi is able to take up its place as a fully fledged official language next to a world language such as English. In addition to this the compilation of bilingual, monolingual and bilingualised learner's dictionaries could be a significant step in making Sepedi more accessible to speakers of other languages, thereby strengthening its position as a second language and making proficiency in this language an imperative for mobility in South African public life.

In order for a language like Sepedi to become a widely used high function language, effective and efficient corpus planning is an imperative. The first step in this process would entail assessment of its functional mobility, i.e. the use of the language across a wide spectrum of social functions, including higher functions.

One way of achieving this goal is to build up a computer corpus of English texts used in higher functions and then use this corpus to determine possible lexical gaps in Sepedi.

In chapter 3, an overview will be given of different approaches to corpus building applied by prominent overseas dictionary houses. A detailed discussion about the English high function Corpus will follow thereafter.

CHAPTER 3

LEXICOGRAPHIC APPROACHES TO CORPUS BUILDING

3.1 Introduction

According to Sinclair (1991:171), a corpus is a collection of naturally-occurring language text, chosen to characterize a state or variety of a language. Hartmann & James' definition is similar. They (1998:30) define **corpus** as a systematic collection of texts which document the usage features of a language or language variety.

One of the main purposes of a corpus is to collect data in order to find evidence for describing particular aspects of a language. Presently, corpora are recognised more and more by research and development groups as the most precious aid in designing systems that respond to user needs, in terms of types of texts and real language to be treated (cf. Calzolari 1996:4). These linguistic corpora are suitable or appropriate scientific departure points and in order to use them effectively, it is necessary to take note of this science and its practices.

The compilation of lexicographic corpora has become a science in its own right. In order to capture the essence of this science and practice, the process of corpus-building by prominent overseas publishing houses is investigated.

In this chapter, three approaches to corpus building are analysed and discussed. Each approach will be discussed with regard to: (a) its main aim and rationale behind its delimitation of text categories, (b) the principles followed with regard to determining frequency and distribution, and (c) its use of concordances (cf. De Schryver & Prinsloo 2000:291-309)

(a) A **text category** is defined by Sinclair (1987:175) as a complete and continuous piece of spoken or written language with a distinct character and function.

(b) Texts can be **classified** or **distributed** into different categories, such as journals, magazines, newspapers etc.

(c) Hartmann & James (1998:27) describe a **concordance** as "a systematic list of the vocabulary which occurs in a text or an author's work, with a minimal verbal context provided for each word". On the same note, Sinclair (1991:32) refers to a concordance as "a collection of the occurrences of a word-form each in its own textual environment".

Thereafter the English corpus which was compiled for the purpose of isolating a number of typical high function words will be discussed in detail.

3.2 The Cobuild Corpus

3.2.1 Aim

According to Sinclair (1987:2) the aim of this corpus is to identify those aspects of the English language which are relevant to the needs of the international user. Moreover, the approach is synchronic, taking texts mainly from 1970.

3.2.2 The composition of the text categories

The general text categories (written and spoken) of this corpus consist of 35,000 words of classroom discourse, one million words of applied science text, seven hundred and fifty thousand words of economic text and an assortment of texts known as the NATLAN COLLECTION, (cf. Sinclair 1987:1). See appendix 1 for a more detailed table.

The written texts mainly comprise of fiction and non-fiction books. The written fiction texts are divided into five categories. The first category is a general category which contains works on human relations with different settings; the second category comprises of historical texts; the third category, namely thrillers, comprises of detective novels like *Jaws* by Peter Benchley. The fourth category is the fantasy category which contains a variety of texts with communal settings and events. One of the subgenres is prose. The final category includes magazines and journals which were published nationally and internationally on a weekly and monthly basis. Major newspapers (national and international) were mainly used to obtain the names of the best-sellers as well as the catalogues from the leading publishers (cf. Sinclair 1987:23-32). These text categories were processed by means of OCR (optical character recognition) text computerisation. No effort was made to include any scientific and/or technical texts in this corpus as subject-field terminology falls outside the scope of this study.

The spoken texts comprise of suitable data on tapes and transcripts from different university departments, spoken material recorded for different research purposes and radio interviews made at the University of Sussex. The BBC also supplied radio batches of transcribed data based on radio broadcasts of reports, interviews and discussions. Lastly, the British Council produced series of transcripts of unscripted and informal conversation, (cf. Sinclair 1987:34-35), (see appendix 1 for more examples).

It seems as if the basic selection principles for the Cobuild Corpus have been text types related in format and medium, such as written books, newspapers, magazines and journals. Books in the written text corpus are categorised primarily by topic/setting. Newspapers are categorised primarily by circulation, secondly by frequency, and thirdly by language variety; and the magazine and journal corpus is categorised primarily by circulation, secondly by frequency, then by language variety and finally by topic. The texts in the spoken corpus are categorised primarily by text type or sub-genre, such as face to face, telephone, TV, video interview etc., and then secondly by general topic.

According to Kennedy (1998:47) the Cobuild Corpus also includes a smaller sub-corpus containing about 1 million words representative of the English used in texts and course books for learners of English (the TEFL Corpus). This corpus was compiled as part of the Cobuild project to be a point of reference for future developments in curriculum design for teaching English as a foreign language.

3.2.3 Frequency counting

The compilation of the Cobuild Corpus is followed by frequency counting, which is the process or result of establishing the frequency of words or other linguistic units in a text or text corpus (cf. Hartmann & James 1998:59). The purpose of frequency counting is to determine how frequently the word is used in different text categories. Example 1 shows an example of frequency word counting as in Pedersen and Zettersten (1996:147).

Example 1

Word	**1**	**2**	**3**	**4**	**5**	**6**	7	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**Total**
Exploits	91	245	224	13	73	150	146	31	116	64	48	197	122	39	211	1770
Exploiters	7	5	15	2	1	8	2	1	0	0	2	8	2	3	5	61
Exploitive	4	0	2	0	2	6	77	0	0	1	0	1	0	0	0	23

In this paradigm, the total number of occurrences of the word *exploit* is 1770. Columns 1 to 15 reflect the spreading of this word over the different text categories (see appendix 2).

3.2.4 Concordances

A concordance shows the meaning of each word in a particular context. In example 2 below, the word *crawl* is given as appearing in different contexts in concordance form in the Cobuild Corpus (Sinclair 1987:36).

Example 2

	Back	Word	Forward
1	she began to	crawl	on the floor
2	the seconds	crawl	past as if they were anchored to
3	future made her skin	crawl	she stripped to her panties
4	she was forced to	crawl	along at a snail's pace

The word as used in example 2, line 1, indicates the *movement across space*. The same word in line 2 has been used figuratively to *suggest slow movement or progress,* and in line 3 the word is used as an idiomatic expression to *show an emotional reaction to a horrible scene or prospect,* while in line 4 it refers to a *slow movement* as in line 2. Compare also appendix 3 for more exhaustive examples from this corpus.

3.3 The Lancaster-Oslo/Bergen (LOB) Corpus

3.3.1 Aim

The aim of this Corpus is to study the grammar and texts, in stylistics as well as in automatic language analysis. The approach used in the LOB Corpus was synchronic, taking texts only from 1961. Hofland & Johansson (1989:2) represent the basic composition of the LOB Corpus (see appendix 4). The LOB corpus is composed of the following text categories:

Example 3

A.	Press: Reportage	44
B.	Press: Editorial	27
C.	Press: Reviews	17
D.	Religion	17

E.	Skills and Hobbies	36
F.	Popular Lore	48
G.	Belles Lettres, Biography, etc.	75
H.	Miscellaneous	30
J.	Learned and Scientific Writings	80
K.	Fiction: General	29
L.	Fiction: Mystery and Detective	24
M.	Fiction: Science	6
N.	Fiction: Adventure and Western	29
O.	Fiction; Romance and Love Story	29
P.	Humor	9
Total		500

(Louwrens 1991:56)

3.3.2 The composition of the text categories

The LOB Corpus, which follows a diachronic approach, consists of fifteen text categories (see example 3). It covers relevant categories and sub-categories of the texts as shown in Example 3(a) below (Hofland & Johansson 1989:2).

Example 3(a)

Categories : A : Press : reportage
B : Press : editorial
C : Press : reviews
D : Religion
E : Skills, trade and hobbies
F : Popular lore
G : Belles-lettres, biography, essays

H : Miscellaneous

I: Learned and scientific writings (see appendix 4)

The press category is divided into three parts, namely report on the press, editorial press and reviewers press. These references were obtained from national Sunday and daily newspapers, as well as provincial weekly and daily newspapers. The type of information obtained from all of the three press categories differs in the sense that Category A contains politics, sports, finance, culture as types of information, whereas Category B includes institutional editorials, personal editorials and letters to the editors. In category C no reference is made to the types of information obtained from the National daily and Sunday weekly newspapers, and the provincial daily and weekly newspapers in Category C. The other remaining categories are religion, miscellaneous, general, fiction etc. (see example 3(a) for more information). The process of categorization was followed by the computerisation of the texts by means of OCR (optical character recognition).

3.3.3 Processing of the LOB Corpus

3.3.3.1 Frequency counting

Frequency counting reflects the total number of words appearing in different columns. Each column shows the total number of occurrences of each word. For example, frequency counting is given in example 4 as in Hofland & Johansson (1989:43).

Example 4

Column 1 below consists of the different words in the corpus, column 2 reflects the total number of occurrences, column 3 indicates the number of text categories in which the form is represented, while column 4 shows the distribution in text samples (cf. Hofland & Johansson 1989:42)

1	2	3	4	5
Word	**Total number**	**Distribution**	**Distribution**	**Distribution**
Abandon	27	11	5	11
Abashed	3	1	1	1

In this table, the frequency and distribution of words such as *abandon*, *abashed,* etc. are given. The total number for the spreading of the word *abandon* and *abashed* is 27 and 3 respectively. (see appendix 5).

3.3.3.2 Concordance

The fact that a word can be viewed in context with a number of words preceding and a number of words following the particular word enables the researcher to see at a glance its meaning, style, syntactic behaviour etc. Kennedy (1998:252) has given the word *on* in concordance form for the LOB Corpus. This word appears in different contextual meanings as shown in example 5.

Example 5

	Backward	Word	Forward
1	mothers help their children	**on**	baking day
2	former freely J73 mounted	**on**	ball races
3	pointed out and compared notes	**on**	beaches

The word *on*, given as an example in 1, refers to a temporal relationship (baking day), in 2 to a spatial relationship (ball) whereas in 3 it refers to a topic-related relationship. (Compare appendix 6).

3.4 Longman Lancaster English Language Corpus

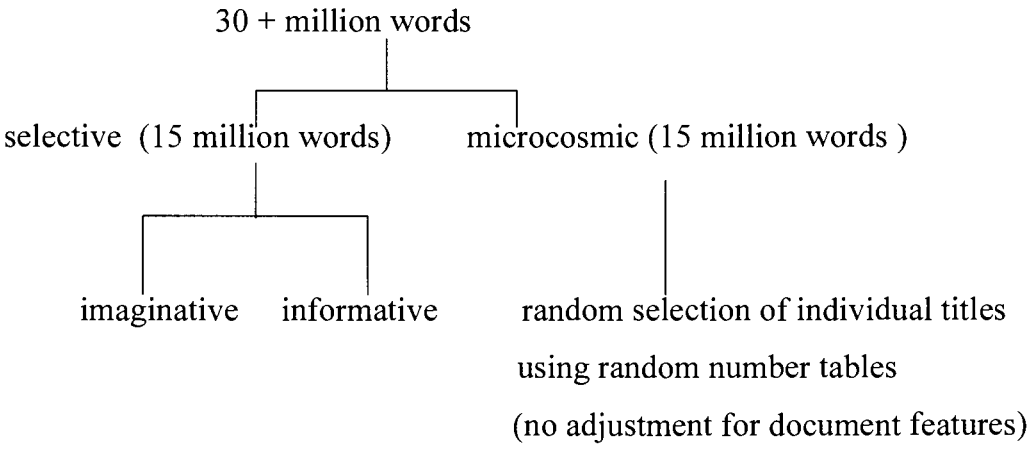
3.4.1 Aim

Each and every data corpus is compiled to fulfill certain objectives. According to Summers (1993:184), the aim of the English Language Corpus is:

> To design and collect a well balanced corpus of between 30 and 50 million words of twentieth century English, covering American and British English predominantly, but also including other major varieties of native English, and including both written and spoken language.

The corpus covers a wide range of written and spoken American and British English of the twentieth century. The approach used here is diachronic, covering language from 1900 onwards. The structures of Summers (1993:201-202) Longman / Lancaster English Language written Corpus and Spoken Corpus are as follows:

Example 6



More information with regard to the imaginative and the informative sources is given in the following paragraph.

3.4.2 The composition of the text categories

The Longman Lancaster English Language Written Corpus contains information from various sources. Those sources are divided into two categories, namely:

- imaginative sources and
- informative sources.

Examples of imaginative sources and informative sources are given in the following table:

Example 6(a)

Written:Microcosmic (15 Million words)

Imaginative sources:
random selection of
individual titles
using random number
tables (no adjustment
for document features)

Examples of imaginative sources:

1. Author
2. No title or subject classifications
3. Subsequent classification into 10 superfields

Examples :

1. Natural and Pure Science
2. Applied Science
3. Social Science
4. World Affairs

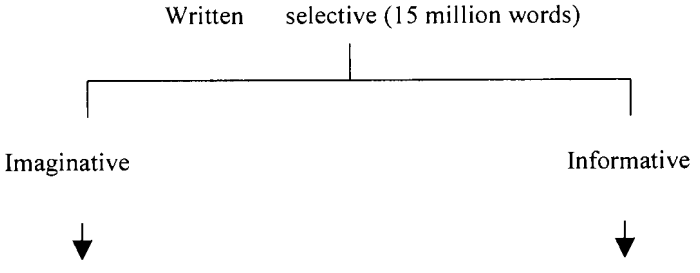
5. Commerce and Finance
6. Arts.
7. Belief and Thought
8. Leisure
9. Fiction
10. Non-fiction
10.1 Poetry
10.2. Drama
10.3. Humor

Four Primary Document Features

Examples:

1. Region
2. Time
3. Level
4. Medium

Example 6(a)



Literary books	Academic textbooks	Newspapers & Journals	Unpublished and Ephimera
1. Achebe, C: Things fall apart 2. Collins, J: Rockstein 3. Clavell, J: Shogun	1. Hawking, S.W. A brief history 2. Periodicals 2.1. Fisherman 2.2. Golf Monthly 2.3. Me etc.	1. Business Week 2. Today 3. Daily Telegraph 4. Financial Times	1.BritishTelecom: How you could save Time and Money 2. Internationalist: 3. Shelter: Who's Homeless Now? 4. Try the New Magazine

Concerning the category microcosmic (see example 6), a random number of tables are used to select titles from books which have already been printed. The criterion used is that of the author's name, and not the title or subject classification. Ten (10) broad subject areas known as Superfields are adopted, namely, natural and pure science, applied science, social science, world affairs etc. (see example 6 (a), example 3 of imaginative sources). Lastly, four primary document features are used for the classification of the texts into document types, namely **Region,** which refers to the varieties of English used in major countries; **Time**, in which a diachronic approach is followed instead of a synchronic one in order to cover the language as from 1900 onwards; **Medium,** of which examples are books, periodicals, newspapers and ephemera, and lastly **Level,** where the emphasis is on the high level of imaginative language rather than the technical.

Example 6(b)

Spoken Corpus

10 million words

selective	Microcosmic
5 million	5 million
predetermined percentages from the following text types	'randomly' selected conversations from a representative sample of population

lectures
demonstrations
speeches
sales talk
interviews
debates
broadcasts (TV, radio, unscripted, commentaries, discussions)

The structure of the English language spoken corpus replicated the written corpus in the sense that it comprises two approaches, namely selective and microcosmic. Examples of selective text types are demonstrations, lectures, speeches, sales talk, interviews,

broadcasting (TV and radio). In as far as the microcosmic approach is concerned, details of the participants including name, gender, age, race, region, occupation, education, social class etc. are taken into consideration. One must remember that text classification according to reference information, details of conversation and details of participants depend mainly on whether one uses a selective as well as a microcosmic approach when conducting such research. This process is followed by the computerisation of the texts by means of OCR.

3.4.3 Processing of the corpus

3.4.3.1 Frequency counting

In total, the Lancaster corpus of written English comprises of 28 million words. These words were extracted from more than two thousand sources. From the two thousand sources, more than six hundred are books. The Lancaster spoken English corpus comprises of only five to ten million words. The relative small size of the corpus is due to lack of spoken material available during the time the data corpora were compiled.

3.4.3.2 Concordance

According to Sinclair (1991:170) the concordance (a word in context) is at the centre of corpus linguistics, because it gives access to many important language patterns in texts. Consider the following example as given by Summers (1993:204) for the Lancaster Spoken English corpus.

Example 7

	Backward	Word	Forward
1	air seems to have dried me	**really**	well
2	quite a good lunch time	**really**	I managed to see people

3	preferably Tuesday or Friday	**really**	…they must…

In 1 and 2, the word *really* has a positive connotation whereas its occurrence in 3 emphasizes an instruction.

This forms the last part of the general treatment of the three corpora, namely, the Cobuild corpus, the LOB corpus and the Longman English Language corpus. These three corpora will now be schematically compared.

3.5 Comparison: schematic representation of the Cobuild Corpus LOB and Longman English Language Corpus

	The Cobuild Corpus	**LOB Corpus**	**Longman/Lancaster English Language Corpus**
1. **Text categories** 1.1 **Approach**	Written and Spoken Books Tapes ↓ ↓ fictional and non-fictional radio Magazines informal Journals conversations Newspapers Synchronic approach	Press Category ↓ report editorial reviews ↓ religion ↓ skills, trades & hobbies ↓ Popular love ↓ belles lettres, biography, essays ↓ Miscellaneous ↓ Learned and Scientific Writings Synchronic approach	Written and Spoken ^ ^ Selective microcosmic Selective microcosmic ↓ ↓ ↓ ↓ ↓ ↓ books 10 lectures conversations newspaper Superfields speeches journals ↓ interviews 1. National etc. Science 2. Applied Science 3. Social Science etc. Diachronic approach
2. **Scanning**	Text were computerized by means of OCR	Texts were computerized by means of OCR	Texts were computerized by means of OCR
3. **Frequency counts**	- Word overall frequency and frequency for separate categories	-Word overall frequency and frequency for separate categories	- Word overall frequency and total of occurrences only
4. **Concordance**	- regular form i.e. word running in the middle with similar or different contexts	- regular form: i.e. same as the Cobuild corpus	- regular form: same as the Cobuild & LOB corpora

3.6 The compilation of an English Corpus of high function words

3.6.1 Introduction

It has been stated in chapter one that the aim of this research is to establish to what extent Sepedi is capable of expressing those concepts found in the higher domains of life, for example, in science, commerce, governmental communication, education, etc.

In order to evaluate the ability of Sepedi to express high function concepts, the logical step would be to build a corpus for Sepedi. This is, however, highly problematic. Although literature is available in some higher function categories such as poetry, prose and religion (e.g. the Bible); other categories such as government communication, advertisements and brochures, manuals, magazines and newspapers are not well represented in Sepedi.

An alternative route towards evaluating the lexical capacity of Sepedi in higher domains was to compile an English Corpus of high function words and then to determine whether Sepedi has translation equivalents for all these words. Following the example of the LOB Corpus it was decided to identify text categories on the basis of genre.

3.6.2 Selection of the text categories

The following six categories were selected and used for the construction of the corpus:

* **academic literature**, for example linguistics, music, economics, sociology, geography etc.
* **advertisements and brochures**, for example advertisements on furniture, jobs, houses etc.
* **the Bible** (mainly the New Testament)

* **magazines and newspapers** (e.g. the *SABC, MetroRail, Scope, Drum* etc. and *Pretoria News, Sowetan, Sunday Times*, *Citizen, Star* etc).
* **manuals** (e.g. for operating and maintaining computers, cellular phones, motor-cars etc).
* **prose** as well as short stories.

The corpus compiled for this study differs from overseas corpora for English, such as the Cobuild Corpus, LOB Corpus and Longman-Lancaster English Language Corpus, as these corpora are intended to represent the entire vocabulary of English, ranging from the spoken vernacular to academic and scientific language, while the limited corpus for this study focuses mainly on genres representative only of higher social functions.

3.6.3 Scanning

Texts were scanned using OCR in order to convert the data to an electronic format that would allow operations such as frequency counting, concordancing, etc.

3.6.4 Organising the data

The data was organised by dividing the entire corpus into the following eight columns:

Column:

1. Word
2. Total number of words
3. Academic literature category
4. Advertisements and brochures category
5. Bible category
6. Magazines and newspapers category
7. Manuals category
8. Prose category

The output was firstly organised in an alphabetical order. Thereafter it was processed in two ways, namely: frequency counts and concordance lines. Thereafter, a number of English high function words were randomly selected with the aim to ascertain whether Sepedi has suitable equivalents or the ability to express such concepts as denoted by these words.

3.6.5 Alphabetic ordering

The English high function words were firstly arranged alphabetically in order to obtain an overall impression of the extent of the alphabetical stretches and typical derivational patterns e.g. *account*, *accountability*, *accounting*, *accounts* etc.

Example 1

Word	Total Number	Academic Literature	Advertisements and brochures	Bible	Magazines and Newspapers	Manuals	Prose
Abhorrence	7	7					
Abilities	7	5	1	1			
Ability	47	11	23	4	7	2	
Able	268	164	24	37	21	7	15
Abnormal	4	2	1	1			
Abnormalities	13	13					
Aboard	10	2	1	7			
Above	80	21	18	12	6	15	8

3.6.6. Word frequency counts

The second output of the corpus to be studied was overall frequency and spreading in a descending order. The importance of the overall word frequency count is to see how frequently the word is used in all of the seven categories. Words reflecting a high frequency count but also words with a low count are of interest to the researcher. Determining which words have significant overall counts is of the utmost importance for this study, since the ability or inability of Sepedi to express high concepts will be an indication to what extent this language can be used as a high function language. Consider the following examples in this regard:

Example 2

Word	Total	Academic literature	Advertisemets and brochures	Bible	Magazines and Newspapers	Manuals	Prose
Disciples	165	164	1				
Systems	165	94	47	12	12		
Reference	144	107	23	1	2	3	
Performance	99	78	2	13	5	1	
Network	81	1	4	7	69		
Lecturers	40	39	1				
Luggage	40	34	5	1			
Mechanisms	40	37	1	1	1		
Media	40	8	15	16	1		
Courses	39	34	1	3	1		
Tragicomedy	4	4					
Tribulation	4	4					
Warrant	4	1	1	1	1		
Workmanship	4	3	1				
Wrapped	4	2	1	1			

Example 2 gives a random selection of high function words from the English Corpus. This example demonstrates high, medium and low frequency counts of high function words from the English Corpus (for more examples, see appendix 7).

3.6.7 Concordance

This study illustrated in Chapter 2 that viewing words in concordance layout is a useful way to determine the different senses of such a word. Consider the following typical examples:

Example 3

	Backward	Word	Forward
1	why mention love, never	**mind**	carry on about how love is
2	I left it in a safe, would you	**mind**	checking it up?
3	I have a good	**mind**	to go without my sister
4	he takes his	**mind**	off by playing soccer

Example 4

	Backward	Word	Forward
1	this dumping	**area**	will affect other people's life
2	houses in Cape Town	**area**	are very expensive
3	much interested in the	**area**	of African art
4	a service branch in your	**area**	needs a technical advisor

The above examples show the words *mind* and *area* in a concordance format.

The study introduced the two types of outputs of the corpus most relevant to this study, namely frequency counts and concordance lines.

3.6.8 Frequently used high function words

As indicated in paragraph 3.6.6 words with high frequency counts as well as words with low frequency counts were investigated.

Random selections of both high and low frequency words were made. Table 3.1 shows the selection of high function words with fairly high frequencies.

Table 3.1

Word	Total	Academic literature	Advertisements and Brochures	Bible	Magazines And Newspapers	Manuals	Prose
Area	94	11	31	5	15	31	1
Assume(d)	20	10	5	2	3		
Assign	17	3	1	6			
Expert(s)	18	5	4	3	3	3	
Policy	61	3	27		29	1	1

The word *area* is frequently used as a high function word as indicated in table 3.1, column three of the academic literature category. The total number of occurrences in the respective categories are: 11, 31, 5, 15, 31 and 1 respectively. The table demonstrates frequency counting for the other remaining words, namely *assume(d)*, *assign*, *expert(s)* and *policy*.

In general, the high function words *area, assume(d), assign, expert(s)* and *policy* have 94, 20, 17, 18 and 61 as the overall number of occurrences in all of the six categories respectively.

3.6.9 Low frequency high function words

Table 3.2 shows the selection of high function words with fairly low frequencies.

Table 3.2

Word	Total	Academic literature	Advertisements and brochures	Bible	Magazines And Newspapers	Manuals	Prose
Creditor	9	9					

Equilibrium	18	18					
Hierarchy	5	3	1	1			
Innovations	4	2	1	1			
Relevance	4	3	1				

The word *creditor* has 9 as the total number of occurrences in table 3.2. It appears 9 and 0 times in the academic literature category and other categories respectively. The next word, namely, *hierarchy* has 3 occurrences in the academic literature category, 1 in both the advertisements/brochures and the Bible categories while 0 in the remaining categories, namely, magazines and newspapers category, manual category and prose category. The total number of its occurrences in all categories is 5. The same process is followed with words such as *innovation, relevance* and *equilibrium*. Although *equilibrium* has a fairly high overall frequency, it was included amongst the lowest frequently used words because it does not have a good spreading.

3.6.10 Conclusion

This chapter gave an overview of the aims, organising principles, data-collection technologies, and data-processing mechanisms of three important English corpora, and the compilation of an English high function Corpus. In conclusion a comparison will be given between the above features of the mentioned corpora and those of the English corpus compiled for the purpose of this study.

3.7 Aim

The main aims of the international corpora are the following:

- Cobuild Corpus: to identify those aspects of the English language which are relevant to the needs of the international user, primarily for compiling learners' dictionaries of English.
- LOB Corpus: to study the grammar and stylistics in texts, as well as in automatic language analysis.

- LLEL Corpus: to design and collect a well balanced corpus of between 30 and 50 million words of twentieth century English, covering American and British English predominantly, but also including other major varieties of native English, and including both written and spoken language. This corpus is also mainly directed at dictionary compilation.

The English Corpus compiled for this study has quite a different purpose and aim, namely to make a computerized collection of English words representing the categories of words typically used in higher linguistic functions in the South African context. The size of the corpus is, of course, much smaller than those of the English corpora discussed here. It comprises modest words.

3.8 Organising principles

In as far as the Cobuild Corpus is concerned, topic seems to have played the most important role in the categorisation process of the written corpus, and genre seems to have played the major role in the spoken corpus. The LOB Corpus has genre as its major organisational principle as indicated by the comparative table in 3.5. Secondary principles for categorisation are distribution (national, provincial etc.) and broad topics such as politics, sports, finance, culture, history, travel etc. The Longman-Lancaster Corpus (written corpus) seems to have adopted linguistic functions as one of its primary principles of organisation, namely imaginative and informative. Secondary principles of organisation are genre (newspapers, books, unpublished and ephemera), topic and document features (such as region, time, level and medium).

For the purpose of this study, however, a genre-based approach was regarded as more suitable as it is easier to correlate genre with communicative function than it is to correlate topic with communicative function. Moreover, a genre-based categorization would produce text categories that are lexically homogeneous (cf. Summers 1993:193),

e.g. text categories such as academic literature (English, Sociology, Economics etc.), manuals (computer, cellular phones etc.), Bible (New Testament) etc.

3.9 Data collection and information processing

According to the comparison table in 3.5 all three corpora went through a similar procedure of scanning texts, counting word frequency and organising words in a concordance form (cf. Prinsloo 1991:56; Lutton 1992:50).

The High Function English Corpus was similarly compiled, namely collecting an archive of appropriate texts and saving the data as text files by making use of scanning and optical character recognition.

The data was analyzed in the same way as the overseas data, namely by

- counting total word frequencies, as well as spreading of words over the different source categories; and
- studying the words in context in a concordance layout.

The English Corpus was compiled by using different text categories. A random selection including words which are highly used, as well as seldomly used ones were made to form the basis of study in the following chapters (i.e. 4 and 5). The discussion about these two categories of words was based on overall frequency counts and spreading.

In the next two chapters (4 and 5), bilingual Sepedi-English dictionaries will be evaluated with regard to their treatment of a number of randomly selected high function words.

CHAPTER 4

THE TREATMENT OF HIGH FUNCTION WORDS IN SEPEDI BILINGUAL DICTIONARIES

4.1 Introduction

In the previous chapter the modus operandi for building a corpus of high function academic words in English was described in detail. The purpose of the corpus is *firstly* to serve as a basis for assessing the ability of Sepedi to be used in all functional domains, especially the higher functions; in other words to determine whether the lexicon of Sepedi has the capacity to express higher order concepts. The second purpose of the corpus is to evaluate existing Sepedi-English dictionaries with regard to their treatment of high function words in Sepedi. The lexical items of the English Corpus served as an entry point for the investigation.

Sepedi has a number of bilingual dictionaries, namely *The New English Northern Sotho dictionary*, *New Sepedi dictionary*, *Pukuntšu* etc. The first two bilingual dictionaries and the *Northern Sotho Terminology and Orthography* were used in the research on the treatment of high function words. In Landau's (1989:8) terminology both are two-way or bi-directional dictionaries.

The analysis will be preceded by a discussion on the theory of bilingual lexicography, with the main emphasis on the concept of equivalence and the diagnostic tools it provides for evaluating bilingual dictionaries. The theory of bilingual lexicography will serve as a matrix for the description of lexical inadequacies of Sepedi if it is to fulfill its role as a high function language.

4.2 Equivalence

Zgusta (1971:312) defines a translation equivalent as:

> a lexical unit of the target language which has the same lexical meaning as the respective lexical unit of the source language.

Equivalence then implies that the meaning of a selected word in the target language item possesses the same meaning and use as the source language item (see example 1).

Three main types of equivalence are distinguished, namely: absolute equivalence, partial equivalence and zero equivalence.

4.2.1 Absolute equivalence

Svensén (1993:143) describes complete equivalence as complete correspondence between words and expressions in two languages as regards content and register. According to Zgusta (1971:312), absolute equivalence requires that the lexical meaning of the two lexical units be absolutely identical, regarding all components, namely designation, connotation and range of application.

Example 1

English	Afrikaans
leap year	*skrikkeljaar*

Example 2

English	Sepedi
amino acid	*aminoesiti*

Afrikaans and Sepedi have "skrikkeljaar" and "aminoesiti" as absolute equivalents for the English words "leap year" and "aminoesiti" respectively.

4.2.2 Partial equivalence

This implies that there is an incomplete correspondence between the semantic content of source language item and target language item. Svensén (1993:143) speaks of an incomplete agreement of the content and register of the words in the respective languages. That means that there is an agreement, but only partial. Compare the following possibilities:

(a) the target language word has fewer semantic features than the source language

Example 3

Source language	Target language
Afrikaans	English
vreet	*eat*

Example 4

Afrikaans	Sepedi
vreet	*ja*

English and Sepedi have no absolute equivalents for *vreet* in Afrikaans. English has *eat* and Sepedi *ja* as partial equivalents to *vreet* in Afrikaans. Additional information concerning the features in the case of the target language must be given in order to bring equality between the source language and the target language.

Example 5

vreet: *eat (of an animal)*

ja (ga phoofolo)

(b) One of the equivalents may be marked for register and such a restricted translation equivalent must be marked by a lexicographic label.

Example 6

Sepedi	English
bolela	*speak, chat (informal)*

In example 6, the translation equivalent *chat* must be marked with the lexicographic label "informal" since it is not used in formal styles, such as writing.

It is important to note that partial equivalence can also occur in one to more than one equivalence relationships. The following types are found: divergence and convergence.

4.2.2.1 Divergence

Divergence occurs when there is a one to more than one equivalence relationship between the source and the target language items. Two subtypes can be distinguished, namely lexical and semantic divergence.

(a) Lexical divergence

Lexical divergence occurs when there is more than one translation equivalent in the target language for one source language term. The translation equivalents are synonymous, and are usually separated by commas in dictionaries (cf. Gouws 1996:17 and 1990:59).

Example 7

impala	*rooibok, impala* (Eng. / Afr.)
atmosphere	*lefaufau, atmosfere* (Eng. / Sepedi)
section	*karolo, kgaolo* (Eng. / Sepedi)

Afrikaans as a target language has *rooibok* and *impala* as equivalent words for the English word *impala* and Sepedi has *lefaufau* and *atmosfere* as translation equivalents for *atmosphere* in the source language. Lastly the word *section* in English can be translated as *karolo* or *kgaolo* in Sepedi.

(b) Semantic divergence

According to Gouws (1996:17) semantic divergence comes into play when the members of the paradigm are not synonyms because the lemma is a polysemous lexical item and different translation equivalents are needed to represent the different polysemous senses of the lemma. In such cases, a number of possible markers are used to separate these translation equivalents. Consider the following examples:

Example 8

(1) Numbering

(a) Bilingual learners' Dictionary \ Tweetalige Aanleerderswoordeboek.

skerp[1]. ❶ sharp [a] *She cut herself with a sharp knife.* Sy het haar met 'n skerp mes gesny. [b] *A thorn has a sharp point.* 'n Doring het 'n skerp punt. [c] *"Slow down - there is a sharp bend in the road."* "Ry stadiger - daar is 'n skrp draai in die pad." [d] *The photograph is so sharp that you can see the hair on the man's arms.* Die foto is so skerp dat jy die hare op die man se arms kan sien. [e] *An alarm clock makes a sharp sound when it goes off.* 'n Wekker maak 'n skerp geluid wanneer dit afgaan. [f] *Mustard has a sharp taste.* Mosterd het 'n skerp smaak. [g] *Cats have sharp eyes and can see well in the dark.* Katte het skerp oë en kan goed in die donker sien. ❷ pointed *A thorn is a pointed growth on the stem of some plants.* 'n doring is 'n skerp groeisel aan die stingel van sommige plante.
□ **skerp** *byvoeglike naamwoord* [*attributief* **skerp**] **skerper. skerpste**

skerp[2] sharply [a] *The road runs straight and then turns sharply to the right.* Die pad loop reguit en draai dan skerp na regs. [b] *"Don't be so rude!" she said sharply.* "Moenie so onbeskof wees nie!" het sy skerp gesê.
□ **skerp** *bywoord*

In this example, *skerp* as an adjective may be translated as *sharp* in seven polysemic senses (as indicated by the example sentences in *skerp* 1 (a) – (g), and as pointed in another sense (*skerp* 2)).

2. Separate entries with bracketed sense distinctions

The *Northern Sotho Terminology and Orthography* is the only bilingual dictionary that provides separate entries for each polysemous sense of a lexical item. Consider the following example:

Example 9

sheet (cloth)	laken	lakane
sheet (of metal)	plaat	lesenke
sheet (paper)	vel papier	letlakala

Example (9) shows that *Northern Sotho Terminology and Orthography* uses two mechanisms for the purpose of equivalent discrimination: bracketed information, separate entries and translation complements (see 4 below). The lemma *sheet,* for instance, appears in three different senses as lemmas and the explanatory information marking those different senses is given next to each lemma in brackets. For example: the first *sheet* refers to *lakane* and the second and third refer to *lesenke* and *letlakala* respectively.

3. Semi-colon

A semi-colon (;) can also be used as a sense marker to separate translation equivalents belonging to different senses of a lemma.

Example 10

The New English Northern Sotho Dictionary

English	Sepedi
barrel	*faki; molomo wa sethunya*

The word *barrel,* as used in example 10, has two senses, namely *a large container* and *the long part of a gun.* Sepedi has separate translation equivalents for these senses, and these are separated by a semi-colon.

4. Translation complements

Translation complements refer to the explication of senses (polysemic senses) whereby explanatory information is given in brackets, after the equivalent it refers to (cf. Carstens 1998:16).

Example 11

Afrikaans	English
duim	*thumb; inch; cam (mining)*

The Afrikaans word *duim* has three polysemic senses, instantiated by the English translation equivalents *thumb*; *inch* and *cam*. The translation equivalent *cam* has *mining* as explanatory information given in brackets.

5. Explication of senses

In some cases, the polysemic senses of the source language lemma are spelt out in the source language. These sense descriptions normally precede the translation equivalent.

Example 12

English	Sepedi
Uncle	*(mother's brother) malome;*
	(father's younger brother) rangwane;
	(father's older brother) ramogolo

For a speaker of Sepedi, the word *uncle* has three separate senses, translated by the words *malome, rangwane and ramogolo.* The phrases *mother's brother*, *father's younger brother' and father's older brother* are explications of these senses.

4.2.2.2 Convergence

Convergence occurs where two or more source language lemmas translate as one translation equivalent in the target language. There are two types of convergence, namely *lexical convergence* and *semantic convergence*. Due to the fact that it is most applicable in Sepedi, lexical convergence is the only type which will be discussed in this research.

4.2.2.2.1 Lexical convergence

Lexical convergence refers to the occurrence of two absolute synonyms which are entered as separate lemmas but which have the same translation equivalent.

Example 13

Afrikaans	English
taalwetenskap	*linguistics*
linguistiek	*linguistics*

Sepedi	English
tagi	*alcohol*
alkoholo	*alcohol*

Words such as *taalwetenskap* and *linguistiek* are entered as separate lemmas in Afrikaans but they both refer to *linguistics* as an English translation equivalent. The second example shows that the words *tagi* and *alkoholo* are entered as separate lemmas in Sepedi but they both refer to *alcohol* in English.

4.2.3 Zero equivalence

complete or partial equivalents in the target language. This phenomenon is known as zero equivalence.

Zero equivalence usually occurs in cases where terms denote culture-specific concepts in the source language. Due to the fact that language is deeply rooted in the culture of different language speakers, the lexicon will reflect the particular way of life of its speakers.

Example 14

English	Sepedi
Lord chancellor	---

In this case there is no equivalent in the target language (Sepedi) for the multiword lexical item *Lord chancellor* in the source language.

Two types of zero equivalence are distinguished, namely, linguistic gaps and referential gaps.

4.2.3.1 Linguistic gaps

Linguistic gaps occur where the concept exists in the minds of both speakers but it is only lexicalised in one language.

Example 15(a)

	Concept	Word
Afrikaans	*Young immature dog*	Ø
English	*Young immature dog*	*puppy*

Example 15(b)

	Concept	Word
English	*Young immature dog*	*puppy*
Sepedi	*Young immature dog*	Ø

Example 15(c)

	Concept	Word
Sepedi	*Cattle herder who refuses to look after the cattle*	*maganagodiša*
English	*Cattle herder who refuses to look after the cattle*	Ø

Afrikaans and Sepedi do not have words denoting an "immature dog" (which is *puppy* in English). These two languages only have diminutives like *hondjie* and *mpšanyana* respectively. The concept "cattle herder who refuses to look after the cattle" is imaginable for speakers of English but there is no one-word equivalent for *maganagodiša* in English.

4.2.3.2 Referential gaps

A lexical item in the source language does not have a translation equivalent in the target language because the concept in the source language is not known to the speakers of the target language (A = source language, B = target language). The emphasis here is on referential meaning (cf. Tourcy 1987:36).

Example 16(a)

	Concept	Word
A = French.	(*Roll with a chocolate stuffed in the middle)*	*pain au chocolat*

B = Eng. "_____" Ø

Example 16(b)

A = Eng. *A type of bean used as a substitute for animal protein in certain foods* *soya*

B = Sepedi "_____" Ø

Pain au chocolat is a French word and is only known to the speakers of the source language in A, but not to the speakers of B. This is also the case with the word *soya*, which is known to the speakers of the language in A, but not to those of B (Sepedi).

In case of zero equivalence created by either a lexical or a conceptual gap, the lexicographer of a bi-or multilingual dictionary has to find a surrogate equivalent.

Svensen (1993: 153) distinguishes the following types of surrogate equivalence:

(a) Expressional aspect

It refers to the word being taken as the last resort and acting as a counterpart in the target language (cf. Svensén 1993:153). The headword can be used as an equivalent for the target language (direct borrowing / transliteration / loan word) but accompanied by some explanation.

Example 17

Source Language	Target Language
French	English
pain au chocolat	*(roll with a chocolate stuffed in the middle)*

Example 18

English	Sepedi
soya	*(Dinawa tša go ba le proteini ya diphoofolo)*

The explanatory information given in English for the French word *pain au chocolat* is "*a roll with a chocolate stuffed in the middle*", and for *soya* as "*dinawa tša go ba le proteini ya diphoofolo*" in Sepedi.

The other possibility is to give a paraphrasal construction for the missing equivalent.

Example 19

German	English
arzthelferin	*doctor's administrative assistant*

Example 20

English	Sepedi
refraction	*kobego ya mahlasedi*

(b) Content aspect

It is only applied when there is no approximate counterpart and the meaning given is the form of a definition or encyclopaedic definition, explanation or notes as stated by Bergenholtz & Tarp (1995:109). One possibility of doing that is to supply the definition in the target language (B) for the word in the source language (A).

Example 21

A	B
Source language	Target language
Telekollege	*series of lectures on television,*

followed by examination for a certificate

Example 22

capillarity	*Ke tlhatlogo goba theogo ya meetse ka gare ga peipi ye e dirwago ke kgogedi magareng ga meetse le peipi.*

Definitions are given in the target languages (English / Sepedi) for the source language words *telekollege* (Afrikaans) and *capillarity* (English).

In rounding up the whole discussion, one could say that the principle of equivalence plays an important role in as far as the theory of bilingual dictionaries is concerned. This theory firstly focused on absolute or complete equivalence as one of three types of equivalence, where the word in the source language has complete equivalence in the target language. Secondly, partial equivalence occurs where there is an incomplete relationship between the semantic content of the target language item and the source language item. Lastly, attention was given to zero equivalence, where lexical gaps are filled by giving explanatory information or paraphrasal information if the speakers of the source language and the target language do not share the same culture.

In conclusion, bilingual dictionaries are in essence based on the principle of equivalence. This basic principle will serve as a guide in describing the lexical inadequacies in the current Sepedi bilingual dictionaries. It will for instance be used where the current Sepedi bilingual dictionaries do not indicate the relevant relationships between some of the English words and Sepedi translation equivalents as well as where there is a total absence of Sepedi translation equivalents for the English words in the existing Sepedi bilingual dictionaries. The latter will be discussed in detail in chapter 5.

4.3 A sample of English high function words

In order to evaluate the quality and comprehensiveness of Sepedi bilingual dictionaries in terms of their treatment of high function words, an empirical survey was conducted. Firstly it had to be established which high function words had been entered in these dictionaries, and which not. Secondly, the treatment of those that had indeed been entered, was investigated. In order to answer these research questions, the focus was placed on the 13 words comprising the random sample mentioned in chapter 3, 3.6.8 and 3.6.9.

Upon investigation it was found that only 5 of these 13 words were indeed lemmatised in Sepedi bilingual dictionaries namely: *area*, *assume(d)*, *assign*, *expert(s)* and *policy* (*see Table 4.1 below*). The other 8 namely: *creditor, equilibrium, hierarchy, innovation, rational, relevance, role and technology* were not entered in any of the dictionaries. The interesting fact was that those which were entered all had overall frequencies of 18 and above whereas most of those which were not entered had frequencies of 10 and below.

In this chapter the quality of the treatment of the first five lemmas will be investigated in detail.

Table 4.1

Word	Total	Academic literature	Advertisements and brochures	Bible	Magazines And newspapers	Manuals	Prose
Area	94	11	31	5	15	31	1
Assume	20	10	5	2	3	6	
Assign	17	3	1	6			
Expert	18	5	4	3	3	3	1
Policy	61	3	27		29	1	1

Firstly, definitions will be given for each of the above-mentioned words from the *Oxford English Dictionary* (1998:CD-ROM) and the *Concise Oxford Dictionary* (1996:CD-ROM). Secondly, the meaning of the word as used in different contexts will be considered, and an analysis of the treatment of these words in the Sepedi bilingual dictionaries will be done. The discussion will be preceded by a comparison table showing English high function words with the translation equivalents provided in the current Sepedi bilingual dictionaries.

4.3.1 English high function words and the translation equivalents given in the existing Sepedi bilingual dictionaries (New English Northern Sotho Dictionary (NEND), New Sepedi Dictionary (NSD) and Northern Sotho Terminology and Orthography (NTO)).

Table 4. 2

English words	New English Northern Sotho Dictionary	New Sepedi Dictionary	Northern Sotho Terminology and Orthography
1. abnormal	sa tlwaelegago	Feta tekanyo, sa tlwaelegago	bjo sa tlwaelegago
2. academic	ya thuto, ya kgopelo; (n) morutegi	–	Akademiki
3. access	botseno, kgoro patametšo, tumelelo, katišo, koketšo	–	–
4. acid	esiti, sedilana, bodila; - soil, mobu wa esiti	Esiti, sedilana, bodila	esiti, sedilana
5. area	area, sekgoba, sekgala; - egion, tikologo, felo	Area, sekgoba, sekgala	area, (1xb) area, sekgoba, sekgala area (part) seripa sa

			sekgoba/ sekgala/ area area (region) tikologo, felo
6. assign	bea, beela, šupa, abela; -ation, kabelo; -ment, kabelo, thoto, tiro	–	–
7. assume	tšea, gopola, itlhoma, ikgantšha, ikgogomoša, hloma; assumption, kamogelo, kgopolo, tlhomo; boikgogomošo		
8. atmosphere	atemosfere, lefaufau	Lefaufau, moya, atemosfere	Lefaufau, atemosfere, sebakeng
9. alcohol	alkoholo, tagi, twatwatwa, senotagi, bjalwa	Alkoholo, tagi	Alkoholo, tagi
10. chapter	kgaolo	Kgaolo	Kgaolo
11. creditor	–	–	–
12. expert	sediri, setswiriri, setsibi, senatla	Setsibi	Setsibi
13. environment	tikologo	–	Tikologo
14. equilibrium	–	–	–
15. hierarchy	–	–	–
16. innovation	–	–	–
17. parliament	kgotlakgolo, palamente	Palamente	Palamente
18. policy	maikemišetšo, kwano ya insuransi	Molawana, maikemišetšo, morero	Policy (insurance) kwano (ya inšoransi) pholisi policy (principle of

			procedure) maikemišetšo
19. rational	–	–	–
20. region	selete, setereke, tikologo ya selete	Selete, tikologo	selete, tikologo
21. relevance	–	–	–
22. role	–	–	–
23. section	karolo, kgaolo	Karolo	Karolo
24. technology	–	–	–

Table 4.2 gives a first impression of the treatment (or lack of treatment) of these English words. In order to properly evaluate the success of such treatment, each word will be evaluated against the background of definitions given in the *Oxford English Dictionary* (1988:CD-ROM) and the *Concise Oxford Dictionary* (1996:CD-ROM).

4.3.1.1 *area*

The *Concise Oxford Dictionary* (1996:CD-ROM) gives the following treatment:

1 the extent or measure of a surface (over a large area; 3 acres in area; the area of a triangle).
2 a region or tract (the southern area).

The *Oxford English Dictionary* (1998:CD-ROM) defines the word 'area' as follows:

area . Pl. **areas**, *rarely* **areæ**.
1. A vacant piece of ground, a level space not built over or otherwise occupied; a clear or open space within a building, such as the unseated part of a church, the arena of an amphitheatre, etc.

2. a. A particular extent of surface, *esp.* of the earth's surface; a space, region, tract.

According to the concordance, the word *'area'* is used in the following contexts:

Example 23

	Back	Word	Forward
1	labour relations director and the	**area**	manager of pollsmoor prison in the
2	the burnt area even a superficial	**area**	can cause very severe shock
3	lies north western namibia vast	**area**	showers each enjoy shady verandahs
4	appliances service branch in our	**area**	technical specifications dimensions
5	press the esc key or click the	**area**	with the mouse when you have

The word *area,* which appears in the concordance in example 23, line 2, refers to a measurable area (e.g. burnt area). The meaning is similar to the meaning given by the *Concise Oxford Dictionary* and the *Oxford English Dictionary.* Line 3, in example 23 refers to a geographical area (e.g. region). In sense 2, the *Concise Oxford Dictionary* and the *Oxford English Dictionary* give the same meaning of *area* as it appears in line 3 of example 23.

The first translation equivalent for the English word *area* given by the NEND, NSD and NTO is *area,* which is a mere borrowing (*see Table 4.2*). From the survey done, the majority of the language speakers spoken to, prefer not to use a borrowed word if Sepedi has its own equivalent. Moreover, the loan word *area* in Sepedi does not cover all the senses of the English word *area.*

The NEND (*as in Table 4.2*) has three translation equivalents, namely *area*, *sekgoba* and *sekgala.* The translation equivalents, namely *sekgoba* and *sekgala* refer to a measurable area whereas the next sense (*as in Table 4.2*) refers to a geographical area *(tikologo and felo).* The translation equivalents *sekgoba* and *sekgala* mean one and the same thing in

example 23 line 2, while line 3 refers to region *(tikologo, felo)*. Consider the following example as a suggestion for adequate treatment of the word *area*:

Example 24

English	Sepedi
area	*(-size), sekgoba, sekgala;*
	(-region), tikologo, felo

By explicating the sense in brackets, the lexicographer shows that the English word *area* is polysemous and for each sense there are different translation equivalents in Sepedi. A semi-colon separates the two equivalent paradigms.

The NSD has three translation equivalents for the lemma *area,* namely *area, sekgoba* and *sekgala.* From the survey I have conducted, most of the language speakers don't accept the use of borrowed words like *area* where Sepedi has its own translation equivalents. The remaining two translation equivalents, namely *sekgoba* and *sekgala* may also confuse the user in the sense that it would seem that the word *area* only refers to *sekgala* and *sekgoba* in Sepedi. This is, however, the case because it can also refer to *tikologo* or *lefelo.* In addition to that, if the user is more knowledgeable about the meaning of the English word *area*, then he may simply conclude that both translation equivalents refer to a measurable and geographical area, which is not the case. The only suitable translation equivalents are those given in example 24, namely *sekgoba, sekgala* (size); *tikologo, felo* (region).

The NTO treats the word *area* as follows:

area (1xb) sekgoba, sekgala.
area (part) seripa sa sekgoba/sekgala/area.
area (region).

The NTO treats the translation equivalents for the word "*area*" in an acceptable way except for the fact that it uses a borrowed word "*area*" as its first entry word which is not acceptable, as already stated before. Unfortunately NTO is not freely available and is not as widely consulted as general commercially available dictionaries.

4.3.1.2 *assume*

The *Concise Oxford Dictionary* (1996:CD –ROM) defines this lemma as follows:

assume v.tr.
1 (usu. foll. by that + clause) take or accept as being true, without proof, for the purpose of argument or action.
2 undertake (an office or duty).

The *Oxford English Dictionary* (1998:CD-ROM) provides the following definitions:

assume , *v.*
I. To take unto (oneself), receive, accept, adopt.
2. To take into the body (food, nourishment, etc.). So in L.; cf. assumption 4. *Obs.*
I. To take upon oneself, put on, undertake.
3. *trans.* To take for granted as the basis of argument or action; to suppose **a.** *that* a thing is, a thing *to be*.

Examples of contexts in which the word *assume* occurs, are as follows:

Example 25

	Back	Word	Forward
1	a senior financial professional who	**assume**	responsibility for the successful
2	department of commerce will sure	**assume**	responsibility for the selection

3	authorised personal computer dealer	**assume**	the entire cost of all necessary
4	motionless priest was done so as to	**assume**	exactly the simple falsehood
5	one simple falsehood that you did	**assume**	it was done to make you take
6	the knowledge was natural that they	**assume**	the leadership boxer and clover

The word *assume*, as it appears in the concordance in example 25, lines 3 and 4, respectively denotes the acceptance of the computer cost for repairs. Line 3 of example 25 demonstrates the same meaning as given by the *Concise Oxford Dictionary* and the *Oxford English Dictionary* (i.e. to accept), and line 4 of the same example also has a similar meaning as given in the definitions of the *Concise Oxford Dictionary* and the *Oxford English Dictionary*.

The word *assume* has not been entered in the NSD and the NTO. The NEND is the only source which has entered it as a headword. The following translation equivalents are given (*see also Table 4.2*): *tšea, gopola, itlhoma, ikgantšha, ikgogomoša, hloma;* assumption, *kamogelo, kgopolo, tlhomo; boikgokomošo*. The first two translation equivalents, namely, *tšea* and *gopola* in the target language (Sepedi) as in *Table 4.2*, would be suitable for conveying the meaning of *assume* in line 3 of example 25. Translation equivalents like *itlhoma, ikgantšha, ikgogomoša* and *hloma* however, variously refer to reflex *(settle etc.*), to have pride (*ikgantšha and ikgogomoša*) and to imagine. These four translation equivalents for the NEND are completely unsuitable. There is no cross-reference from words such as *reflex* back to the lemma a*ssume.* Consequently, the user will become confused or misguided. Consider the following suggested treatment:

Example 26

English	Sepedi
assume	*tšea, gopola;*- assumption, *kgopolo*

The first two translation equivalents, separated by a comma, refer to the source language lemma *assume*, and are followed by the translation equivalent *kgopolo* which is a noun. The semi-colon, as used in example 26, serves as a marker to separate the verb and the noun of the word *assume* in the source language. That means the first two translation equivalents function as verbs and the last one as a noun. All these equivalents refer to the idea of accepting something to be true without any proof. This is also concurred by definitions as given by the *Concise Oxford Dictionary* and the *Oxford English Dictionary*. This illustration shows that the translation equivalents given for the noun *assumption*, namely *kamogelo*, and *tlhomo* and *boikgogomošo* as in *Table 4.2*, have no semantic relationship with the word *assume*. There are two synonymous translation equivalents (in the target language) for one source language term (*tšea* and *gopola*) and this will result in lexical divergence. It was argued in this chapter, section 4.2.2.1, that lexical divergence occurs when there is more than one translation equivalent in the target language for one source language term.

4.3.1.3 *assign*

The *Concise Oxford Dictionary* (1996:CD-ROM) gives the following explanation for *assign*:

assign v. & n.
1 (usu. foll. by to). a allot as a share or responsibility. b appoint to a position, task, etc.
2 (foll. by to) transfer formally (esp. personal property) to (another).
n. a person to whom property or rights are legally transferred.

On the same note, the *Oxford English Dictionary* (1998:CD-ROM) has the following:

assign , *n.*2 Also 5-7 **assigne**.
† **1.** One who is appointed to act for another, a deputy,agent, or representative; = assignee
1. *Obs.*

2. One to whom a property or right is legally transferred; = assignee 2. Esp. in the phrase *heirs and assigns*: see quot. 1865.

According to the concordance, the word *assign* demonstrates the following contextual behaviour:

Example 27

	Back	Word	Forward
1	exercise for individual classes is to	**assign**	practical class tests
2	not know and will punish him and cut	**assign**	his lot with the unfaithful servant
3	long including blank spaces if you	**assign**	a password leave this field blank
4	move to the attribute you want to	**assign**	and select it by pressing the spacebar
5	Some application programs	**assign**	filename extensions automatically
6	keyboard for your computer you	**assign**	the monetary symbol decimal
7	any attempt otherwise to see	**assign**	or transfer any of the rights duties

Assign in line 1 and 5 of example 27 means to give, as also defined by the *Concise Oxford Dictionary*, namely allot as a share or responsibility in 1(a), while in line 6 it means to put. This definition refers to the same concept as described by the *Concise Oxford Dictionary* and the *Oxford English Dictionary* in 1(b) (i.e. appointed to a position) and 1(a) (i.e. one who has been appointed to work for another).

The NSD and NTO do not include the lemma *assign*. The translation equivalents appearing as verbs and nouns in the NEND are marked by a semi-colon.

The first four translation equivalents, namely *bea, beela, šupa, abela* are verbs (*as in Table 4.2*). The first translation equivalent, namely *bea*, means to put (as used in example 27, line 6 of the concordance), the second translation equivalent *beela* means to *assign*. The third and fourth translation equivalents *šupa* and *abela* also means to *assign*. In

order to make this type of dictionary more user-friendly, one can prioritise the translation equivalents starting with the most frequently used words. Consider the following suggested treatment of the word *assign*:

Example 28

English	Sepedi
assign	*(go) fa, šupa, abela, bea, beela;*
	***-ment**, modiro, mošomo*

The first five translation equivalents, used as verbs, are user-friendly because they are entered in the order of frequency of use and they all refer to the word *assign* as a source language item. That means the source language word *assign* has more than one translation equivalent in the target language, and a semi-colon has been used to separate semantically divergent paradigms.

A translation equivalent *kabelo* has been selected by the NEND to represent the English nouns *assignation* and *assignment*. The word *kabelo* is derived from the verb *abela* in Sepedi. The word *kabelo* can also refer to noun *distribution* in English. The other translation equivalents for the word *assignmen*t in Sepedi are *modiro* or *mošomo.*

The translation equivalent *kabelo* for the noun *assignation* in *Table 4.2* can therefore also refer to both the nouns *assignment* and *distribution.* Some of the dictionaries refer the word *assignation* to a meeting, especially a secret, one e.g. with a lover (cf. Hornby 1995:61).

Concerning the word *assignment*, it could be claimed that the given equivalent *thoto* and is totally inappropriate (*compare Table 4.2).* The word *thoto* refers to property and not to *assignment* and the word *tiro* is used by both Setswana and Sepedi speakers. The translation equivalents which are frequently used for the word *assignment* in Sepedi

(target language) are *mošomo* or *modiro*. These translation equivalents are absolute synonyms. Therefore they can be entered as separate lemmas even if they do have the same translation equivalents. In other words, it is a case of lexical convergence. For example:

Example 29

Sepedi	English
mošomo	*assignment*
modiro	*assignment*

The translation equivalents *mošomo* and *modiro* are absolute equivalents of the word *assignment* as in example 29. All omissions and additions in example 29 were made in order to revise the existing dictionaries for standardisation purposes, and to give a data driven account of Sepedi as well as its ability to act as a high function language.

4.3.1.4 *expert*

The *Concise Oxford Dictionary* (1996:CD-ROM) and the *Oxford English Dictionary* (1998:CD-ROM) provide the following definitions for this concept respectively:

expert adj. & n.
adj.
1 (often foll. by at, in) having special skill at a task or knowledge in a subject.
2 (attrib.) involving or resulting from this (expert evidence; an expert piece of work).
n. (often foll. by at, in) a person having special knowledge or skill.

expert , *n.*
1. One who is an expert or has gained skill from experience. Const. *at*, *in*, *with*.

2. One whose special knowledge or skill causes him to be regarded as an authority; a specialist.

The behaviour of the word in a concordance format is as follows:

Example 30

	Back	Word	Forward
1	position requires a well experienced	**expert**	with years experience in
2	profession at the h s r c makes him an	**expert**	when it comes to identifying
3	we would like you as a known	**expert**	to try out this cooker for us in a
4	do what I ask you and I'm not the	**expert**	in murder what do you want having

The word *expert* refers to logistical expertise as in line 1, meaning having an advanced knowledge of logistical problems, and line 2 refers to a person having a high knowledge as far as the identification of students is concerned. Both lines refer to a person having a special skill for a particular area of knowledge. In addition, the *Concise Oxford Dictionary* and the *Oxford English Dictionary* also define an *expert* as someone who is knowledgeable or has gained a special skill for a particular subject.

The NEND, NSD and NTO offer the word *setsibi* as a translation equivalent for the word *expert* in the source language (*see Table 4.2*). The translation equivalent *setsibi* has the same meaning as the translation equivalent *expert* as illustrated in example 30. Consider the following examples:

Example 31 (as in the NEND)

English	Sepedi
expert	*sediri, setswiriri, setsibi, senatla*

The translation equivalent paradigm for the lemma *expert* as in example 31 contains three words which are not appropriate Sepedi translation equivalents for the word *expert.* The word *sediri* refers to a subject (i.e. a person or a thing which performs the action of a verb (cf. Procte et. al. 1995:451), not an *expert,* while *setswiriri* (a Sesotho translation equivalent for the lemma *expert*) and *senatla* refer to a strong man, not an *expert.* The word could only be used as an idiomatic expression to refer to an *expert* person. The following example is a more adequate reflection of the linguistic facts of current-day Sepedi :

Example 32

English	Sepedi
expert	*setsibi, sekgoni, matwetwe*

The implication of separating the translation equivalents by means of commas is that the lemma *expert* has more than one synonymous translation equivalent in the target language (Sepedi), namely *setsibi, sekgoni* and *matwetwe.* This type of translation equivalent instantiates lexical divergence as defined in section 4.2.2.1 (a) of this chapter. The three synonymous translation equivalents, namely, *setsibi*, *sekgoni* and *matwetwe* have been put in order of frequency of use and they all refer to an *expert* as is the case in example 32.

4.3.1.5 *policy*

The *Concise Oxford Dictionary* (1996:CD-ROM) defines the lexical item/word 'policy' as follows:

policy[1] n. (pl. -ies)
1 a course or principle of action adopted or proposed by a government, party, business, or individual etc.
2 prudent conduct; sagacity.

policy[2] n. (pl. -ies)

1 a contract of insurance.

2 a document containing this.

In addition, the *Oxford English Dictionary* (1998:CD-ROM) provides the following:

1.a. An organized and established system or form of government or administration (of a state or city); a constitution, policy. Now rare or Obs.

b. An organized state, a commonwealth. Obs.

2.a. Government, administration, the conduct of public affairs; political science.

3. A course of action adopted and pursued by a government, party, ruler, statesman, etc.; any course of action adopted as advantageous or expedient. (The chief living sense.)

The linguistic conduct of *policy* is explicated by the concordance format below:

Example 33

	Back	Word	Forward
1	it is the	**policy**	of the university not to award supplementary
2	formulating provincial	**policy**	within the national policy framework in
3	life policy to another insurance	**policy**	such as an endowment or whole life policy
4	such endowment of whole life	**policy**	without filling in a medical examination

The word *policy* in lines 1 and 2 of the concordance above refers to a particular rule for a particular university as far as an examination is concerned, and to the formulation of provincial rules respectively. Line 3 and 4 refer to a policy contract which has also been cited by the *Concise Oxford Dictionary* in one of its definitions of the word *policy* (see policy 2 n. (pl.-ies).

The NEND provides *maikemišetšo, kwano ya insuransi* as translation equivalents for the word *policy*. These two translation equivalents are separated by a comma as if they are semantically similar to each other. This is not the case, because the translation equivalent *maikemišetšo* in the target language (Sepedi) refers to a rule or principle and *kwano ya insuransi* refers to a policy contract. It will mean that the lemma "policy" has two different senses, namely a rule or principle and a policy contract.

The NSD has *molawana, maikemišetšo, morero* as translation equivalents for the word *policy*. These three translation equivalents in the target language (Sepedi) are separated by commas, which implies that they are synonymous. This, however, is not the case because the word *morero* refers to a *theme,* not a *policy,* whereas *molawana* can be used as a translation complement for the word *maikemišetšo* in the Sepedi bilingual dictionary. The explanation as given is to show a distinction between a principle or rule and an insurance policy.

The NTO presents two senses for *policy*. These different senses have not been numbered. Very few translation dictionaries use numbering, although it could be a very helpful sense-discriminating device (see section 4.2.2.1 (b) 1 of this Chapter, which states that numbering can also be used to mark the number of different lemmas). The first translation equivalent *kwano* refers to an *insurance contract* as in example 23, lines 3 and 4, whereas the second translation equivalent *maikemišetšo* refers to the principle or procedure as in example 23, line 1 and 2. The NTO has treated the translation equivalents for the word *policy* fairly well.

In order for Sepedi bilingual dictionaries to treat a lemma such as *policy* in a user-friendly way, the following suggested example can be taken into consideration:

Example 34

English Sepedi

policy	*(principle) maikemišetšo;*
	(insurance) pholisi

The translation equivalent *maikemišetšo* (principle) has the same meaning as in example 34, lines 1 and 2. The lemma *policy* has two different senses as marked by a semi-colon (;) and brackets ( ) (see section 4.2.2.1 (b) example 8 (2) and (3)). The second translation equivalent in the target language (Sepedi) is *pholisi* which refers to insurance as in example 34, lines 3 and 4. As of now, the translation equivalent word *pholisi* in the target language is frequently used by language (Sepedi) speakers when they refer to a policy contract. This is a borrowed word but it is frequently used by the majority of the language speakers to distinguish the insurance policy contract from the concepts, principles or rules.

4.4 Conclusion

In this chapter, the concordance lines drawn from the English corpus were used to measure the adequacy of the treatment of English high function words in the existing Sepedi bilingual dictionaries, namely the *New English Northern Sotho Dictionary* , *New Sepedi Dictionary* and the *Northern Sotho Terminology and Orthography* .

In the first instance, definitions were given from the *Concise Oxford Dictionary* and the *Oxford English Dictionary*. The next step was to correlate definitions with meanings of the English high function words as appearing in different verbal contexts. This was done in order to get real meanings of each English word. Thereafter, the treatment of English high function words in existing Sepedi bilingual dictionaries was evaluated.

It was observed that some of the English high function words were treated adequately, and others were treated inadequately; the reason being that, in some cases, borrowed words were given preference above indigenous translation equivalents. The native

speakers prefer to use indigenous translation equivalents where applicable, and borrowed words where there is a semantic or a lexical need.

In some cases, the translation equivalents provided were not semantically and pragmatically equal to the English high function words, but to a concept other than that which the term in the source language refers to. Moreover, some translation equivalents are not used by Sepedi speakers but by other Sotho language groups. Lastly, there were cases where inappropriate Sepedi translation equivalents were used, and in some instances idiomatic expressions were given instead of one-word translation equivalents.

The above discussion of the treatment of high function words in the Sepedi dictionaries paves the way for a detailed discussion of the non-treatment of certain English high function words in the Sepedi Bilingual dictionaries. Chapter 5 deals with this issue.

CHAPTER 5

LEXICAL GAPS IN SEPEDI CONCERNING HIGH FUNCTION CONCEPTS

5.1 Introduction

This chapter deals with words from the English data base which have not been entered in any of the existing Sepedi bilingual dictionaries. One of the aims is to establish whether the dictionaries truly reflect the language situation in the Sepedi-speaking community, or whether the lexical gaps are merely a symptom of inadequate dictionaries. However, the primary objective is to obtain lexical data directly from mother-tongue speakers, so as to make responsible recommendations with regard to improving the quality of current Sepedi bi- and multilingual dictionaries.

A small-scale survey was conducted in order to establish whether suitable translation equivalents could be found or coined for those English high function words with no translation equivalents in the existing Sepedi bilingual dictionaries. A questionnaire was distributed among a number of mother-tongue speakers. The respondents were required to suggest translation equivalents for those English high function words which have not been entered in existing Sepedi bilingual dictionaries (*see Table 4.2 in Chapter 4*).

The questionnaire is divided into two parts. Part one requires the personal details of the respondents, namely name, age and occupation. Part two requires suggestions regarding possible translation equivalents for a selection of English high function words, some of which have been entered as headwords in bilingual Sepedi dictionaries, and some which have not. The questionnaire is open-ended in the sense that comments by the respondents are invited (see appendix 8).

The responses could assist the researcher in various ways:

- It could serve as a confirmation of existing practice in dictionaries;
- It could prove the current lexicographic treatment to be wrong or misguided;
- It could provide valuable information on the use of lexical items in the community, of which standardizing organisations have not taken cognizance.

The outcomes of the survey could play an important role in corpus planning, and the concrete results (new, revised dictionaries) could serve as educational tools to familiarize speakers of the language with words that denote important concepts of higher domains of public life. An example is the word *creditor*:

creditor (n)
a person, company to whom money is owed
e.g. His creditors are demanding to be paid

---------- (Sepedi translation equivalent)

comment(s) ----------

5.2 English high function words

It was found that only eight high function words in the corpus (*see Table 4.2 in Chapter 4*) did not have translation equivalents in Sepedi bilingual dictionaries (the *New English Northern Sotho Dictionary*, the *New Sepedi Dictionary* and the *Northern Sotho Terminology and Orthography*) namely, *creditor, equilibrium, hierarchy, innovation, rational, relevance, role* and *technology* (*as in Table 4.2, chapter 4, section 4.3.1*).

Table 5.1

	NEND	NSD	NTO
Creditor	_	_	_
Equilibrium	_	_	_
Hierarchy	_	_	_
Innovation	_	_	_
Rational	_	_	_
Relevance	_	_	_
Role	_	_	_
Technology	_	_	_

Most of the 8 had an overall frequency of 10 and fewer in the English high function Corpus:

Table 5.2

Word	Total Number	Academic literature	Advertisements and brochure	Bible	Magazines and Newspapers	Manuals	Prose
Creditor	9	9					
Equilibruim	18	18					
Hierarchy	5	3	1	1			
Innovation	4	2	1	1			
Rational	4	3	1				
Relevance	4	3	1				
Role	104	75	8	20	1		
Technology	35	21	5	7	2		

Only five of the above mentioned words will be evaluated in this study, namely *equilibrium, creditor, hierarchy, innovation* and *relevance*. Firstly the meaning of the word will be given as defined by the *Concise Oxford Dictionary* and the *Oxford English Dictionary*, and secondly, the meaning of the word as instantiated by the occurrences in the concordance will be analysed. Thirdly, the translation equivalents suggested by the

target language speakers (Sepedi) will be considered (see a questionnaire as appearing in appendix 8 for each translation equivalent as suggested by respondents).

5.2.1 *Creditor*

The *Concise Oxford Dictionary* (1996:CD-ROM) defines the word as follows:

creditor n.
1 a person to whom a debt is owing.
2 a person or company that gives credit for money or goods (cf. debtor).

The *Oxford English Dictionary* (1998:CD-ROM) defines the same word as follows:

creditor
1. One who gives credit for money or goods; one to whom a debt is owing; correlative to *debtor*.
2. *Book-keeping. Creditor* (or *Cr.*) being written at the top of the right-hand or credit side of an account (originally in personal accounts, in apposition with the name of the person whose account it is), is hence applied to that side of any account, or to what is entered there.

Example 1

	Back	Word	Forward
1	contract lie breach by debtor on the	**creditor**	positive nonperformance and
2	debtor should offered to perform and the	**creditor**	discuss the consequences of
3	of more debtors has effect on the	**creditor**	to conserve the object of the
4	is of the essence of the contract and the	**creditor**	has obtained the study objectives

According to Example 1 the word *creditor* appears 4 times in the concordance. All of its occurrences, as already defined by the *Concise Oxford Dictionary* and the *Oxford English Dictionary*, refer to a person or a company to whom money is owed. The word *creditor* belongs to the speech register of economists.

The translation equivalent suggested by 80 percent of respondents was *mokolotwa*. 10 percent of the respondents suggested *mokolotiši* while the remaining 10 percent suggested the words *moadimi* and *mokoloti* (see appendix 8). The most suitable translation equivalents for the word *creditor* as used in example 1, will therefore be *mokolotwa* and *mokolotiši*. The following treatment is suggested:

Example 2

English	Sepedi
creditor	*mokolotwa, mokolotiši*

According to example 2, Sepedi has more than two translation equivalents for *creditor*, namely *mokolotwa* and *mokolotiši* for one source language term. These translation equivalents are synonymous, and like other instances of lexical divergence, they are separated by a comma (cf. chapter 4, section 4.2.2.1 (a)). One way of introducing these translation equivalents in the Sepedi-speaking communities is through the compilation of new Sepedi bilingual dictionaries. By means of acquisition planning, mother-tongue speakers can be convinced to start using the new words.

The translation equivalents as suggested by the remaining ten percent of the respondents could refer to *debtor* (i.e. *moadimi* and *mokoloti,* meaning people who owe money). Since *moadimi* and *mokoloti* can be used to refer to debtors, it is better not to use them as translation equivalents for *creditor*, since the language already has two equivalents for *creditor*, namely *mokolotwa* and *mokolotiši*. It is suggested that *moadimi* and *mokoloti* be entered as translation equivalents for *debtor*.

5.2.2 *Equilibrium*

The *Concise Oxford Dictionary* (1996:CD-ROM) defines this word as follows:

equilibrium n. (pl. equilibria or equilibriums)
1 a state of physical balance.
2 a state of mental or emotional equanimity.
3 a state in which the energy in a system is evenly distributed and forces, influences, etc., balance each other.

The *Oxford English Dictionary* (1998:CD-ROM) defines *equilibrium* as follows:

equilibrium Also 7-9 **equilibrium**.
1. a. In physical sense: The condition of equal balance between opposing forces; that state of a material system in which the forces acting upon the system, or those of them which are taken into consideration, are so arranged that their resultant at every point is zero.
A body is said to be in *stable* equilibrium, when it returns to its original position after being disturbed; in *unstable* when it continues to move in the direction given to it by the disturbing force; in *neutral*, when it remains stationary in its new position.
b. ***equilibrium of temperature***: see quot.
2. a. The state of equal balance between powers of any kind; equality of importance or effect among the various parts of any complex unity.
c. Well-balanced condition of mind or feeling.

Consider the following example as an illustration of the contextual conduct of the word:

Example 3

	Back	Word	Forward
1	industry make use of the short run	**equilibrium**	positions in the two market
2	short run equilibrium determine	**equilibrium**	price and quantity show
3	demand curve for a rational consumer	**equilibrium**	in the utility approach the price
4	demand and supply of motorcars has an	**equilibrium**	price and equilibrium quantity
5	equilibrium price will increase while	**equilibrium**	quantity in the market for cars

The word *equilibrium*, as used in the above example, refers to a state of balance in all of its different contextual appearances in Example 3. This definition goes along with the one given by the *Concise Oxford Dictionary* and the *Oxford English Dictionary*. For example, line 2 refers to a state of balance in price and quantity, and line four refers to a balance of price and quantity being affected by a supply and demand of motor cars.

The translation equivalents suggested by the majority of the respondents (40 percent) were *tekatekano* or *tekatekanelo,* followed by 33 percent for *tekatekanyo*, 20 percent for *boemotekanelo* and *tekatekanelo* (maemo) and 6,6 percent for *ekhwilibriamo*.

The word *ekhwilibriamo,* as suggested by some respondents, would not be accepted by the majority of Sepedi speakers due to the fact that Sepedi has its own suitable translation equivalents, which are not partially borrowed, namely *tekatekano* or *tekatekanelo* and *tekatekanyo,* as suggested by the majority of respondents. These three translation equivalents all refer to the state of balance as in example 3, lines 1 to 4. The following treatment is suggested:

Example 4

English	Sepedi
equilibrium	*tekatekano, tekatekanelo, tekatekanyo*

The word *equilibrium* in the source language therefore has more than one translation equivalent in the target language, namely *tekatekano, tekatekanelo* and *tekatekanyo,* (see section 4.2.2.1 of chapter 4), which implies a relationship of lexical divergence. It is not necessary to add *boemo* in brackets for the word *tekatekanelo* as suggested by some respondents because the words *tekatekano*, *tekatekanelo* and *tekatekanyo* are all suitable translations for *equilibrium.* Explanatory information in brackets is only necessary where there is a possibility that the meaning of the word may not be clear to the reader or user or where there is a need for meaning discrimination. The translation equivalent *boemotekanelo* is a compound word of *boemo,* which refers to position, and *tekanelo,* which refers to the state of balance (equilibrium).

5.2.3 *Hierarchy*

hierarchy n. (pl. -ies)

1. A system in which grades or classes of status or authority are ranked one above the other (bottom of a hierarchy). b a hierarchical system (of government, management, etc.). c (foll. by of) a range in order of importance (hierarchy of values).

2. A priestly government. b a priesthood organized in grades (cf. *Concise Oxford Dictionary* 1996:CD-ROM)

According to the *Oxford English Dictionary* (1998:CD-ROM) the word *hierarchy* has the following meanings:

hierarchy

1. Rule or dominion in holy things; priestly rule or government; a system of ecclesiastical rule.

2. The collective body of ecclesiastical rulers; an organized body of priests or clergy in successive orders or grades.

3. A body of persons or things ranked in grades, orders, or classes, one above another; *spec.* in *Natural Science* and *Logic*, a system or series of terms of successive rank (as *classes*, *orders*, *genera*, *species*, etc.), used in classification.

Consider the following examples in the concordance format:

Example 5

	Back	Word	Forward
1	the student must be able to follow the	**hierarchy**	of the courts in south africa
2	drawn from facts according to	**hierarchy**	of needs the highest level
3	advantages and disadvantages and use of	**hierarchy**	of data and illustrate it with the
4	public affairs executive and chief in	**hierarchy**	had produced documentaries
5	of the authority and the whole	**hierarchy**	want to see one of the men

The meaning of the word *hierarchy,* as instantiated by Example 5, line 1 in the concordance form, coincides with definition 1 of the *Concise Oxford Dictionary* as well as with sense 1 of the *Oxford English Dictionary*, namely to the levels of authority. The appearance of the same word in line 3 refers to the various levels or ranks of importance, which is also the case as defined by the *Concise Oxford Dictionary* in sense 1 (b) and (c) as well as sense 3 of the *Oxford English Dictionary*. The word *hierarchy* in all the examples in 5, refers *to rank or position;* i.e. various levels or ranks of importance.

The translation equivalents suggested by the respondents for the word *hierarchy* are as follows: *tlhatlamano* or *tatelano* (60 percent), *tlhatlamano* (bogolo / maemo) or *tatelano* (bogolo / maemo) (30 percent) and *hieraki* (10 percent). It is not necessary to give explanatory information in brackets (such as *bogolo/maemo*) for the translation equivalent *tlhatlamano* because the word itself is an absolute translation equivalent for *hierarchy* in the source language. Explanatory information given in brackets is only

necessary for cases where there is a possibility that the meaning of the word may not be clear to the reader or user or where there is a need for sense discrimination.

The only explanatory information that needs to be given, is for *tatelano* (bogolo/maemo). *Tatelano* is not an absolute equivalent of *hierarchy*. It has two prominent senses, namely "chronological occurrences" (i.e. one after the other) and "hierarchy". This implies that there is semantic divergence which comes into play where the members of the paradigm are not synonymous, because the lemma is a polysemous lexical item and different translation equivalents are needed (see Chapter 4, section 4.2.2.1 (b)). Thus if *tatelano* is given as a translation equivalent for the English word *hierarchy,* its relevant sense must be marked. Lastly, *hieraki* as suggested by ten percent of the respondents cannot be included in the translation equivalent paradigm since transliterations should only be used in cases where suitable Sepedi equivalents do not exist, or where the transliteration has a semantic value which differs from that of the indigenous equivalent.

5.2.4 *Innovation*

According to the *Concise Oxford Dictionary* (1996:CD-ROM), innovation means to:

1 bring in new methods, ideas, etc.
2 (often foll. by in) make changes.

On the same subject, the *Oxford English Dictionary* (1998:CD-ROM) provides the following explanations:

innovation

The action of innovating; the introduction of novelties; the alteration of what is established by the introduction of new elements or forms.

A change made in the nature or fashion of anything; something newly introduced; a novel practice, method, etc.

3. The action of introducing a new product into the market; a product newly brought on to the market.

Example 6

	Back	Word	Forward
1	that influenced what ever changes	**innovations**	have been introduced
2	theatres closed down although many	**innovations**	were introduced in the theatres
3	that take advantage of the logical	**innovations**	for example, could be a direct
4	stock breeder in the farm was full of	**innovations**	and improvements about field

The word *innovations,* which appears in the concordance form in example 6, line 1, refers to the introduction of new plays and new versions in English literature. All the usages of the word in the concordance refer to the introduction of new things, ideas or techniques. Both the *Concise Oxford Dictionary* and the *Oxford English Dictionary* reflect the meaning of *innovation* as it appears in the different contextual occurrences in Example 6.

The translation equivalents suggested by the majority of respondents were *tlholo* 50 percent, followed by 43,3 percent, for *boinaganelo*, *boithomedi* and *boithomelo*. The following lexicographical treatment is suggested:

Example 7

English	Sepedi
innovation	*tlholo, boinaganelo, boithomedi, boithomelo*

The target language (Sepedi) in the above example demonstrates lexical divergence as described in Chapter 4, section 4.2.2.1 (a). Some of the informants (6,6 percent) suggested *mokgwa wo moswa* (a new method) as a translation equivalent for *innovation.*

Since up to four single word equivalents, namely *tlholo, boinaganelo, boithomedi* and *boithomelo* in example 7 are suitable translations, it is not necessary to give a multiword phrase such as *mokgwa wo moswa* as well. It has already been stated earlier in this study that surrogate equivalents should only be considered for cases where no suitable single word equivalents are available (see Chapter 4).

5.2.5 *Relevance*

The *Concise Oxford Dictionary* (1996:CD-ROM) and the *Oxford Engish Dictionary* (1998: CD-ROM) provide the following explanations respectively:

1.

relevant adj. (often foll. by to)

bearing on or having reference to the matter in hand.

relevance n.

2.

'relevance.

Relevancy; *spec.* in recent use, pertinency to important current issues (as education to one's later career, etc.); social or vocational relevancy.

An example of the concordance form of the word *relevance* is as follows:

Example 11

	Back	Word	Forward
1	texts recognized as classics lessens	**relevance**	to issues of todays justice system
2	literature have been chosen for their	**relevance**	to students lives and careers
3	the images of the night candle for what	**relevance**	might this have in the past era

The word *relevance,* which appears in the concordance lines in example 11, line 2, emphasises the significance of reading. In the *Oxford English Dictionary*, the word *important* is not given as a synonym but forms an important part of the descriptive definition of the word *relevance*. Line 3 of the concordance of the word *relevance* and the *Concise Oxford Dictionary* emphasise the suitable date or period for a particular thing.

90 percent of the respondents suggested *nepišo* as the primary translation equivalent for *relevance*, 53 percent suggest *tebanyo* as one of the possible translation equivalents, and 46,6 percent suggested also *tebano* as a possible translation equivalent for the same word. In the opinion of the researcher all the above Sepedi words are suitable translation equivalents for the word *relevance; i.e.* meaning anything being connected with what is happening or discussed. The following treatment is suggested:

Example 12

English	Sepedi
relevance	*nepišo, tebano, tebanyo*

The translation equivalents, *nepišo*, *tebano* and *tebanyo* as suggested by the respondents, are synonymous, and therefore separated by commas as motivated in chapter 4, section 4.2.2.1.(a).

5.3 Conclusion

This chapter has demonstrated the importance of combining different scientifically motivated methods of data collection and analysis. Where lexical gaps exist in a language it is not sufficient to rely only on the intuition of the lexicographer. It is of the utmost importance to involve mother tongue speakers of the particular language when considering the treatment of a source language item with no apparent translation equivalent. Field work and user surveys do not only serve the purpose of verifying or

refuting the intuitions of the lexicographer on the basis of frequency, but provides invaluable insight into the preferences of users. Affective responses are of sociolinguistic importance, and cannot be ignored. By comparing the results of this kind of empirical research with evidence from systematically organised sources such as concordances and dictionaries of languages with well established, data-driven dictionaries, the researcher ensures that the end product (a revised dictionary) will not only be a reliable reflection of actual usage, but will also be a socially and educationally relevant and useful tool.

CHAPTER 6

CONCLUSION

The democratic elections in 1994 gave birth to a new constitution in South Africa. Amongst these changes in the constitution, is the official recognition of the indigenous languages of South Africa, namely Sepedi, Sesotho, Setswana, siSwati, Tshivenda, Xitsonga, isiNdebele, isiXhosa, and isiZulu. It was clearly stated in the discussion of the language principles and stipulations in chapter 2 that every official language should be protected and promoted. Sepedi as one of the eleven official languages benefits from this new dispensation.

The promotion of Sepedi can only succeed if it is preceded by proper language planning. This has to be done in line with the constitutional principles which are relevant to language policy and language stipulations as they appear in the new South African Constitution of 1996. It would mean that language problems need to be identified first and followed by possible solutions as discussed in chapter 2 of this study.

In this study, the primary research question was to investigate whether Sepedi is capable of functioning comfortably as a medium of communication in all higher domains of life such as government communication, health communication, medium of instruction in schools and at tertiary levels, commerce, law, science and technology, etc.

To achieve this, the logical step was to build a Sepedi corpus consisting of different types of data that represent communication in higher domains. This was, however, not possible due to the fact that the literature which is available for high function categories in Sepedi comprises mainly of poetry, prose and religion. The remaining categories, such as those mentioned in the paragraph above, are not well represented in Sepedi.

An alternative option was to compile a corpus of high function English words as a measuring instrument. In order to do that, one had to be acquainted with the principles and practice of corpus-building. Three internationally renowned corpora of English, namely the Cobuild Corpus, the Lancaster Oslo Bergen (LOB) Corpus, as well as the Longman-Lancaster English Language Corpus were studied. The main emphasis for these corpora was on text categorisation, frequency counting and the use of concordances.

The main aim of these international corpora was to study the grammar and stylistics in texts, including automatic language analysis and to compile English dictionaries. The English corpus in this study differs in terms of its aim and purpose from international corpora in the sense that it comprised of categories of words used in higher social functions in South Africa. The collection of the data went through a similar procedure of scanning texts, counting word frequency and organising words in a concordance format.

Words reflecting a high frequency count and also words with a low count were of interest to the researcher. A random selection of this type of words was made to form the basis of this study. The whole discussion about these two categories of words was based on overall frequency counts and spreading. The main purpose of selecting these two categories was to determine whether Sepedi as an official language is capable of expressing these types of concepts or not.

In order to evaluate the treatment of high function words in bilingual Sepedi-English dictionaries, the theory of bilingual lexicography was invoked.

A pivotal aspect of this theory is equivalence, and different types of equivalence were considered in order to analyse the treatment of English high function words in the Sepedi bilingual dictionaries.

What does this imply for the central issue, namely the ability of Sepedi to express high function concepts?

It is not the occurrence of these equivalence relationships that is important for this study, but the information these dictionaries supply on the lexical capacity of Sepedi.

A detailed analysis was made of the treatment of five high function words in Sepedi-English dictionaries. This was done to gain an impression of the quality of bilingual lexicography in dictionaries for the African languages, and to assess the lexical capacity of Sepedi to account for the range of meanings that English high function words have. The following words were randomly selected: *area, assume, assign, expert* and *policy.*

In order to delimit the meaning/concept for which Sepedi needs a lexical item the meaning of each of the above English words was studied by comparing its conduct in the concordance lines of the English High Function Corpus and the representation of its meaning by two prominent English dictionaries.

It was found that all five had not been adequately treated in the existing Sepedi-English bilingual dictionaries, namely the *New English Northern Sotho Dictionary* (NEND), the *New Sepedi Dictionary* (NSD) and the *Northern Sotho Terminology* and *Orthography* (NTO). The findings were as follows:

- Some of the Sepedi-English bilingual dictionaries do not have translation equivalents for these words at all.
- Some of the translation equivalents were incorrectly represented.
- Some of the translation equivalents given were not Sepedi words but Sesotho and words.
- In the translation equivalent paradigm, the translation equivalents were not arranged in an order of frequency of use.

- In some cases, idiomatic translation equivalents were given.

Suggestions were made with regard to a more realistic treatment (in Sepedi) of the English high function words under scrutiny. These suggestions were based on the mother tongue intuition of the researcher as well as on the responses from other mother-tongue speakers.

Firstly, the meanings of the English high function words were established on the basis of definitions given by the Concise Oxford Dictionary and the Oxford English Dictionary. Secondly, a questionnaire was compiled to establish whether mother-tongue speakers knew and/or used Sepedi translation equivalents for these words.

In most cases more than one possible translation equivalent were given, and in some cases at least one of the equivalents was a transliteration, e.g. *ekhwilibriamo* and *hieraki*. Suggestions for revision of the dictionaries under scrutiny were made on the basis of the responses to the questionnaire.

The two most important findings of this study are that:

(a) For the majority of high-function words in English there are equivalents in Sepedi. However, the treatment of the equivalent paradigms by Sepedi-English dictionaries is far from satisfactory.

(b) Among the 300 English high function words investigated, only 8 were not entered in the macrostructures of bilingual Sepedi-English dictionaries. The omission of these 8 words did however not mean that they lacked translation equivalents. This fact was corroborated by the response to the questionnaire.

What has been achieved in this research, marks only the beginning of a process of lexical stock-taking in Sepedi. Although it has been established that Sepedi has the general high function words to be used in any domain of life, it still has to be determined whether the

language is capable of expressing the concepts of scientific and technical domains, such as law, medicine, the human sciences, etc. This type of research will indicate to language planners how much terminological work needs to be done in order to place Sepedi alongside the other scientific languages of the world.