# The representation and distribution of conjunctions in selected Sepedi home language textbooks:

## a corpus-based Investigation

**by**

**Mpho Gift Mahlobogoane**

**Submitted in fulfillment of the requirements for the degree**

**MAGISTER EDUCATIONIS**

**in the Faculty of Education**

**at the**

**UNIVERSITY OF PRETORIA**

**Supervisor: Dr MC Makgabo**

**Co-supervisor: Dr EN Nzimande**

**June 2024**

## Declaration

I declare that the dissertation/thesis, which I hereby submit for the degree of Master in Education at the University of Pretoria, is my own work and has not previously been submitted by me for a degree at this or any other tertiary institution.

.....

Mpho Gift Mahlobogoane

5 June 2024

Dr MC Makgabo (Supervisor)

12 June 2024

# Ethical clearance certificate


FACULTY OF EDUCATION
Ethics Committee

**RESEARCH ETHICS COMMITTEE**

| | |
|---|---|
| **CLEARANCE CERTIFICATE** | CLEARANCE NUMBER: **EDU039/23** |
| **DEGREE AND PROJECT** | MEd<br>The representation and distribution of conjunctions in selected Sepedi home language textbooks: a corpus-based investigation |
| **INVESTIGATOR** | Mr Mpho Mahlobogoane |
| **DEPARTMENT** | Humanities Education |
| **APPROVAL TO COMMENCE STUDY** | 02 May 2023 |
| **DATE OF CLEARANCE CERTIFICATE** | 03 June 2024 |
| **CHAIRPERSON OF ETHICS COMMITTEE**: | Prof Funke Omidire |
| **CC** | Mr Simon Jiane<br>Dr Connie Makgabo |

This Ethics Clearance Certificate should be read in conjunction with the Integrated Declaration Form (D08) which specifies details regarding:

- Compliance with approved research protocol,
- No significant changes,
- Informed consent/assent,
- Adverse experience or undue risk,
- Registered title, and
- Data storage requirements.

# Declaration of authorship

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Education

## Copyright declaration

I hereby certify that, where appropriate, I have obtained and attached hereto a written permission statement from the owner(s) of each third-party copyrighted matter to be included in my thesis, dissertation, or project report ("the work"), allowing distribution as specified below. I certify that the version of the work I submitted is the same as that which was approved by my examiners and that all the changes to the document, as requested by the examiners, have been effected. I hereby assign, transfer and make over to the University my rights of copyright in the work to the extent that it has not already been effected in terms of a contract I entered into at registration. I understand that all rights with regard to copyright in the work vest in the University who has the right to reproduce, distribute and/or publish the work in any manner it may deem fit.

**Kopieregverklaring**

Hiermee sertifiseer ek dat, waar toepaslik, die skriftelike toestemming verkry is van elke derdeparty wat die eienaar is van materiaal wat aan outeursreg onderhewig is en wat ingesluit is in my proefskrif, verhandeling of projekverslag ("die werk"), waardeur verspreiding op die wyse hieronder gemeld, ook toegelaat word. Ek sertifiseer dat die weergawe van die werk wat deur my ingedien is, dieselfde is as dié wat deur my eksaminatore goedgekeur is en dat alle veranderinge soos deur hulle versoek, aangebring is. Ek dra en maak hiermee my outeursregte in die werk aan die Universiteit oor insoverre dit nie reeds ingevolge 'n kontrak by registrasie deur my gedoen is nie. Ek begryp dat alle regte met betrekking tot outeursreg van die werk dus by die Universiteit berus en dat die Universiteit die reg voorbehou om na goeddunke die werk te reproduseer, versprei en/of publiseer.

04 June 2024

SIGNATURE/HANDTEKENING DATE/DATUM

## Dedication

I express deep gratitude to my beloved mother, Cathrine Mmamma Phalane, ngwan'a Mahlobogoane my esteemed father, Silas Mbuti Phalane, and my cherished young brothers, Delron Mashole and Nong Junior Phalane. It is with great honour that I dedicate this dissertation.

In fond memory of my late and strong-willed grandmother, Elizabeth Dikeledi Mahlobogoane, ngwan'a Aphane, now residing in the celestial realm, I joyfully announce the completion of my master's degree — a milestone influenced by your enduring inspiration. I kindly ask that you share this news with your son, Fana Cheer Mahlobogoane, along with your brothers-in-law, Stephan's and James Mahlobogoane, and our esteemed Segodi ancestors. Through this dissertation, your first grandson proudly upholds the esteemed value of education within our clan.

Also, I am proud of myself for the effort, patience and dedication that I invested in this study.

Lastly, I offer profound gratitude to the God of Mount Zion, whose unwavering presence and divine love continually guide my path. With unwavering faith and deep reverence, I acknowledge His blessings, growing in love and gratitude with each passing day.

## Acknowledgements

To have achieved this milestone in my life, I would like to express my sincere gratitude to the following people:

- My supervisors, Dr MC Makgabo and Dr EN Nzimande, whose guidance and scholarly insights were instrumental throughout the stages of research, composition, and finalisation of this dissertation. Dr Makgabo, your steadfast encouragement and sagacity are deeply appreciated. May your endeavours be blessed abundantly. Dr Nzimande's scholarly and intellectual supervision throughout the process is acknowledged with profound gratitude. Your constructive critique significantly contributed to the fruition of this dissertation. May the divine guidance of Almighty Shembe uNyazi LweZulu continue to fortify and bestow further wisdom upon you.
- My former Sepedi Lecturer, Mr MS Mabule, who emersed me with knowledge and love for Sepedi grammar. I am who I am because of you. I thank you.
- Prof E Taljard, for her invaluable support during the preliminary phases of this study.
- Ms B Strydom, for generously sharing pertinent information pertaining to the Sepedi textbooks.
- The ethos of Ubuntu permeates the lives of individuals of African heritage, and I am no exception. Numerous individuals, both directly and indirectly, played integral roles in manifesting this dissertation. The support, affection, and resources bestowed upon me have been indispensable in attaining this milestone.
- Lastly, heartfelt appreciation is extended to all the ancestors of the Mahlobogoane family for their enduring influence and guidance.

# Abstracts

Different views exist concerning the function of conjunctions in Sepedi language. The use of a comma with some of Sepedi conjunctions is the most controversial aspect, regarding the usage of conjunctions in this language. While several studies have been conducted on various linguistic aspects within the South African context, there is a lack of research on the usage and meaning of conjunctions in the Sepedi language. Therefore, the aim of the present study was to compare the syntactic and semantic features of Sepedi conjunctions between Specialised Sepedi Corpus and General Sepedi Corpus. The study investigated the syntactic and semantic features of six selected Sepedi conjunctions as observed between Specialised Sepedi Corpus and General Sepedi Corpus. Furthermore, the study sought to determine whether there are similarities and differences in the usage and meaning of Sepedi conjunctions between scholarly sources and the corpora. The study employed corpus-based approach for data analysis and interpretation, and a corpus-query software called 'LancsBox X' was used for querying the corpora. This study was grounded in Noam Chomsky's seminal work on generative grammar. Generative grammar is conceived as a structured system of statements and rules aimed at describing and defining grammatically correct utterances within a language, while excluding those that are not well-formed. The findings revealed that the similarities in the syntactic and semantic features of Sepedi conjunctions between the Specialised Sepedi Corpus and General Sepedi Corpus outweigh the differences. The results further indicated that the similarities in the syntactic and semantic features of Sepedi conjunctions between the scholarly sources and the corpora surpass the differences. This, therefore, indicated that the corpus-based approach to investigating linguistic phenomena provides valid, credible and invaluable information that can benefit linguists, language educators, researchers, as well as students.

***Key terms:*** *Sepedi, intuition, corpus, conjunctions, Specialised Sepedi Corpus, General Sepedi Corpus, syntactic meaning, semantic meaning, corpus-based approach*

# Senaganwa

Mo polelong ya Sepedi, go bonagala dikgopolo tše di fapanego ge go tliwa mo tabeng ya mohola wa makopanyi. Ye nngwe ya tšeo di tlogago di gakantšha le go hlola ngangišano kudu mo polelong ya Sepedi, ke tšhomišo ya fegelwana le makopanyi. Le ge go bonagala dinyakišišo tše mmalwa tšeo di dirilwego mo dikamanong tša Afrika Borwa ka mo go thutamaleme, go tloga go bonagala tlhaelelo ye ntši ya go nyakišišwa ga tšhomišo le tlhalošo ya makopanyi mo polelong ya Sepedi. Ka gona go realo, maikemišetšo a nyakišišo ye ke go no bapetša dibopego tša popopolelo le tša semanthiki tša makaponyi magareng ga khophase yeo e kgethegilego le khophase ya go tlwaelega. Mo nyakišišong ye go nyakišišitšwe dibopego tša popopolelo le tša semanthiki tše tshela tšeo di kgethilwego tša makopanyi a Sepedi bjalo ka ge go lemogilwe magareng ga khopase ya go ikgetha le khophase ya go tlwaelega. Go tlaleletša seo, nyakišišo ye e be e nyaka go lekola go swana goba diphapano tabeng ya tšhomišo le tlhalošo ya makopanyi a Sepedi magareng ga methopo ya diithuti le go khophora. Nyakišišo ye e šomišitše mokgwa wa khopora go sekaseka le go hlatholla tshedimošo, gape go šomišitšwe lenaneo la khomphutha la '*corpus-query*' leo le bitšwago 'LancsBox X' go nyakišiša le go botšiša ka khophora. Nyakišišo ye e be ikepetše le go ithekga ka mošomo wa Noam Chomsky wa seminale wa tšweletšo ya popopolelo. Tšweletšo ya popopolelo e tsebja bjalo ka tshepetšo yeo e rulagantšwego ya ditatamentele melawana, yeo e ikemišeditšego go hlaloša polelo yeo e napagetšego ya popopolelo, go sa balwe tšeo di fošagetšego. Dikhwetšo di utollotše gore, go swana ga dibopego tša makopanyi a Sepedi go popopolelo le semathiki magareng ga khophase ya go ikgetha le ya go tlwaelega go feta diphapano. Dipoelo di tloga di tšweleditše gape gore go swana ga dibopego tša makopanyi a Sepedi ka mo go popopolelo le go semanthiki magareng ga methopo ya diithuti le go khophora, e feta diphapano. Ka gona go realo, go tloga go bile le taetšo ya gore mokgwa wo wa khophase ge o šomišwa go nyakišiša tiragalo ya polelo, o tloga o fana ka tshedimošo ya go nweša a mokgako yeo e ka holago ditsebi tša maleme, barutiši ba polelo, banyakišiši gammogo le baithuti.

***Mareo:*** *Sepedi, tsebo ya tlhogo, khophase, makopanyi, khophase yeo e kgethegilego ya Sepedi, khophase yeo e tlwaelegilego ya Sepedi, tlhalošo ya popopolelo, tlhaloša ya semanthiki, mokgwa wa khopora*

## Language editor

Below is the letter from language editor indicating that language editing has been done.

Krista G. Verster

C/o Klip Street & Rossouw Crescent
Prince Albert, 6930

Cell: 082 499 7844
Email: kgverster@gmail.com

23 May 2024

Client: Mpho Gift Mahlobogoane

**The representation and distribution of conjunctions in selected Sepedi home language textbooks: a corpus-based investigation**

**by**

**Mpho Gift Mahlobogoane**

This letter serves to certify that I edited the above document in the English language to correct grammar, spelling, abbreviations and the list of references.

No content changes were made, track changes were used to make suggestions and the onus rests upon Mr Mahlobogoane to accept recommended changes.

Yours faithfully

K Verster

Krista G. Verster
*B.Bibl (UP)* *B.Bibl (Hons) (Unisa)* *PEF Associate Member*

# List of abbreviations

| | |
|---|---|
| ACE | English Australian Corpus |
| ARCHER | A Representative Corpus of Historical English Registers |
| BNC | British National Corpus |
| BONSE | Bilingual Oxford Northern Sotho-English dictionary |
| BoE | The Bank of English |
| CANCODE | Cambridge and Nottingham Corpus of Discourse in English |
| CAPS | Curriculum and Assessment Policy Statement |
| CBE | Cobuild Bank of English |
| CEFR | Common European Framework of Reference for Languages |
| CL | Corpus linguistics |
| COCA | American National Corpus and the Corpus of Contemporary American English |
| COHA | Corpus of Historical American English |
| DBE | Department of Basic Education |
| EFL | English as a Foreign Language |
| ESL | English as a Second Language |
| FET | Further Education and Training |
| GSC | General Sepedi Corpus |
| HL | Home Language |
| ICE | International Corpus of English |
| ICLE | International Corpus of Learner English |
| LGP | Language for General Purposes |
| LGSWE | Longman Grammar of Spoken and Written English |
| LLC | London–Lund Corpus of Spoken English |
| LOB | Lancaster Oslo-Bergen |
| LSP | Language for Special Purposes |
| MICASE | Michigan Corpus of Academic Spoken English |
| MICUSP | Michigan Corpus of Upper-level Student Papers |

| | |
|---|---|
| OCR | Optical Character Recognition |
| POS | Part-of-Speech |
| PSC | Pretoria Sepedi Corpus |
| PZC | Pretoria Zulu Corpus |
| RP | Received Pronunciation |
| SBCSAE | Santa Barbara corpus of spoken American English |
| SEC | Spoken English Corpus |
| SSC | Specialised Sepedi Corpus |
| TEC | Translational English Corpus |
| WSC | Wellington Corpus of Spoken New Zealand English |

# Table of contents

# List of addenda

# List of figures

# List of tables

# CHAPTER 1: INTRODUCTION

## 1.1 Background to the study

Conjunctions are one of the parts of speech that are found in most languages of the world. Regarding their function, conjunctions mainly link sentences to each other. Lombard (1993) points out that they indicate the association between sentences. Furthermore, Lombard (1993) states that in Sepedi, there are only three basic conjunctions, *ge* 'while'; 'when'; 'if' and *ga* 'of'; at the place /homestead' and *kapa* 'or'. On the other hand, Louwrens, Kosch and Kotzé (1995: 38), contend that, 'there are hardly any "basic" conjunctions (cf. *ge* 'if') in Sepedi, but most conjunctions have been derived from other word categories, such as verbs, e.g., *gore* 'that', *fela* 'but' etc.'. Therefore, it appears that conjunctions are a part of speech that is not clearly defined and understood in Sepedi language.

As a part of speech, conjunctions also form part of the formal curriculum in South African schools. They form part of the content of textbooks prescribed at schools. According to the Curriculum and Assessment Policy Statement (CAPS), language teachers should have the prescribed language textbook used by learners in order to successfully conduct teaching of the language (Department of Basic Education, 2012). Language textbooks are used as core resources in the classroom. The content in these textbooks shape the learners understanding of language conventions.

The Sepedi Home Language (HL) textbooks treat conjunctions in a comprehensive way, meaning the definitions as well as usage examples are provided. However, the definitions and usage examples in these textbooks are typically based on authors' intuition of how conjunctions are defined and how they function. Given the fact that conjunctions are a highly controversial part of speech especially in Sepedi, it is really doubtful if material developers provide accurate information on the meaning and function of conjunctions based purely on their intuition. Gabrielatos (2005: 4) remind us that:

> The discrepancy between intuitions and attested use indicates that when the linguistic information learners are given is based only on intuitions, and when the examples and texts used in class are chosen to reflect these intuitions, then teachers and material writers may unwittingly present their personal informal

> observations about language as the true and full picture of language structure and use, or present their own preferred usage as the only 'correct' or 'acceptable' one.

Therefore, in order to give an empirically sound and justifiable account of meaning and usage of conjunctions in Sepedi, it is necessary to base claims on a representative sample of language. A corpus could provide such a representative sample. The corpus-based approach is an acceptable approach in investigating linguistic phenomena. Researchers make use of corpus to study language patterns. Within the South African contexts, existing views on linguistic aspects can also be tested and possibly revised in light of corpus evidence. Kennedy (1998: 4) contends that:

> A corpus can be analysed and compared with other corpora or parts of corpora to study variation. Most importantly, it can be analysed distributionally to show how often particular phonological, lexical, grammatical, discoursal or pragmatic features occur, and also where they occur.

What can be noted in the above quote is that two or more corpora can be compared depending on the nature of the research inquiry. Furthermore, it can be observed from above quote that the distribution of linguistic phenomena in a corpus can be analysed to see how often and where does it appear.

The present study, therefore, seeks to explore and compare the usage and function of conjunctions in a Specialised Sepedi Corpus (SSC) and General Sepedi Corpus (GSC). The SSC is composed of Sepedi Home Language textbooks for the Senior Phase. The Senior Phase is selected for this study since it is the introductory phase to Further Education and Training (FET), Grades 10-12. This is the phase that begins to teach language structures and conventions that are more complex and therefore would be the best phase to investigate how conjunctions are defined and used. Therefore, the present research seeks to explore how conjunctions are represented and frequently distributed in the Senior Phase textbooks. Furthermore, the research endeavours to compare the distribution, function and usage of the conjunctions between SSC and GSC.

The GSC consists of general texts gathered from internet and supplemented by general texts received from the University of Pretoria, Department of African Languages and is

used as a reference corpus in order to establish the actual usage of these conjunctions in general language. The aim of the comparison between the SSC and the GSC is to establish whether there are commonalties in the manner in which conjunctions are defined and used between the Language for Special Purposes (LSP) and Language for General Purposes (LGP). The SSC represents LSP since it comprises school textbooks which may be said to belong to the field of education. It can also be argued that it is encompassed under the field of linguistics when considering the fact that the study has its focus on conjunctions, which form part of parts of speech. The GSC, on the contrary, represents LGP since it is made up of general texts collected from the internet and supplemented by general texts received from the University of Pretoria, Department of African Languages. Therefore, the GSC, which is used as reference corpus, will confirm whether or not the intuitions of developers of the school textbooks are in line with the general usage and meaning of Sepedi conjunctions. In addition, the definitions as well as usage examples given by scholars to conjunctions will also be explored to determine whether or not they coincide with the findings obtained from the two corpora. This will, in a way, demonstrate how the corpus-based method can be used together with the traditional manual method of analysis to further strengthen the findings of the study. McEnery, Xiao and Tono (2006: 7) rightly advised that:

> The corpus-based approach can offer the linguist improved reliability because it does not go to the extreme of rejecting intuition while attaching importance to empirical data. The key to using corpus data is to find the balance between the use of corpus and the use of one's intuition.

Therefore, the present study also seeks to determine whether there are commonalties between the findings obtained from the corpora and those obtained from data outside the corpora.

### 1.2 The research problem

Although within the South African context there are several studies that have been conducted on a variety of linguistic aspects, studies on the usage and meaning of conjunctions in Sepedi language are lacking. Previous researchers focused on investigating other parts of speech, such as nouns, verbs, ideophones, etc. while others focused on exploring formatives, such as locatives, prefixes, suffixes, etc (Taljard, 2012;

Gauton and de Schryver, 2002; Gauton, de Schryver, and Mohlala, 2004; Van Olmen, Breed, and Verhoeven, 2019; Taljard and de Schryver, 2016). Furthermore, previous research also focused on linguistic phenomena such as grammaticalisation and pronominalisation, as well as other aspects such as the noun class system characteristic of the African Indigenous languages. Currently, no study has investigated the function and meaning of conjunctions in Sepedi, especially in their authentic occurrence. It is, therefore, necessary that the meaning and function of Sepedi conjunctions are investigated, especially given the fact that their meaning and function are not clearly defined. Therefore, the present study seeks to address this gap in existing literature by investigating and comparing the usage and meaning of Sepedi conjunctions in the SSC and GSC. The aim of comparing the two corpora is to determine whether the meanings that are attached to conjunctions by developers of the textbooks forming part of the SSC, as well as usage examples provided, coincide with meaning and usage of conjunctions found in the GSC. The study further incorporates definitions of conjunctions and usage examples provided by various scholars in order to strengthen the findings obtained from the two corpora.

### 1.3 The rationale

The foregoing discussion highlighted that currently no study investigated the usage and meaning of Sepedi conjunctions, especially in their authentic occurrence. Therefore, the present study will make an enormous contribution in this regard to the body of knowledge, on the meaning and function of parts of speech in general and conjunctions in particular. Furthermore, Taljard (2012) contends that within the South African context, there is a notable absence of corpus-based studies on teaching of South African Indigenous languages. Therefore, the present study will also make a contribution to the field of corpus-based studies by looking at the function and meaning of conjunctions as found in the corpus of school textbooks that are used for teaching and learning at schools. The study also compares the function and meaning of conjunctions in the SSC to those found in the GSC. This will be the first study in Sepedi thus far to compare linguistic features between a specialised and a general corpus.

Furthermore, the corpus-based approach is still developing within the South African context. Taljard (2012) echoes this sentiment when she posits that it was only in 2002 that the first fully-fledged corpus-based study on a South African Indigenous language was conducted. Moreover, studies that employed this approach within the educational context are very limited. Therefore, the present study will expand the body of knowledge on corpus-based studies within the South African context. It will further contribute to the on-going body of knowledge pertaining to the application of corpus-based approach to the educational context.

### 1.4 The research questions

In the present study, the following research questions will be addressed:

- What is the frequency of occurrence of Sepedi conjunctions in the Specialised Sepedi Corpus and the General Sepedi Corpus?
- What are the commonalities and differences in the usage and meaning of Sepedi conjunctions between Specialised Sepedi Corpus and General Sepedi Corpus?
- Do the meaning and function given by scholars to Sepedi conjunctions coincide with that found in the corpora?

### 1.5 Aim and objectives

The aim of the present study is to investigate and compare the syntactic and semantic features of Sepedi conjunctions between the Specialised Sepedi Corpus and General Sepedi Corpus.

The above aim can be broken down into the following objectives:

- To investigate the frequency of occurrence of Sepedi conjunctions in the Specialised Sepedi Corpus and General Sepedi Corpus.
- To establish whether the syntactic and semantic features of Sepedi conjunctions are similar or different between the Specialised Sepedi Corpus and General Sepedi Corpus.

- To determine if there are similarities and differences in the usage and meaning of Sepedi conjunctions between the scholarly sources and the corpora.

By achieving the aforementioned aim and objectives, the present study will provide some insight into the syntactic and semantic features of Sepedi conjunctions.

### 1.6 Delineation of the study

The present study investigates and compares the syntactic and semantic features of Sepedi conjunctions between the Specialised Sepedi Corpus and General Sepedi Corpus. The study is limited to only learners' textbooks used in the Senior Phase (i.e., Grade 7-9) and only one tittle was selected. This was obviously due to the limited scope of the research. Furthermore, the Bilingual Oxford Northern Sotho-English Dictionary (BONSE) was used in identifying the conjunctions to form the focus of the study. This dictionary was selected based on the fact that it was compiled based on a corpus. It is more likely that the selection of lemmas for inclusion in the dictionary was based on frequency of occurrence. Therefore, the utilisation of the dictionary in the present study was for the purposes of selecting conjunctions that have high frequency of occurrence in Sepedi. Moreover, it was discovered that 20 conjunctions are treated in the BONSE and only six of them were sampled based on their high frequency of occurrence in both corpora. Due to limited scope of the present study, it would not be feasible to analyse all the conjunctions found in the dictionary.

### 1.7 Organisation of the study

In Chapter 2, research done using the corpus-based method as the foundation for analysis is reviewed and highlighted. Corpus-based studies conducted globally serve as the starting point. Furthermore, research that used corpus-based methodology which was carried out in various regions of Africa is highlighted. Moreover, corpus-based studies conducted within the South African context are discussed. The chapter then proceeds to corpus-based studies on teaching and learning material development. Lastly, theoretical framework underpinning the present study is discussed.

The methods used in the present study are highlighted in Chapter 3. Firstly, the corpus-based method within the field of corpus linguistics is discussed. Thereafter the discussion

on qualitative and quantitative methods used in the present study is provided. The chapter continues to discuss the corpus-based approach and the intuition-based approach to linguistic analysis. Furthermore, the corpus-based approach versus the corpus-driven approach as types of approaches that employ corpora as the basis for analysis are explored. A general description of the corpus design process is provided. Following this, is a definition of a corpus and a discussion of the various types of corpora.

The chapter proceeds to the discussion of aspects that are related to corpus design, including balance, representativeness and corpus size. The discussion of different types of corpus-query software follows, which is followed by the discussion of the whole process of designing the SSC as well as the GSC for the present research. This is followed by a discussion on corpus-query tools available for usage to analyse a corpus. Lastly, the tools offered by LancsBox X, the corpus-query tool used in analysing the corpora in the present study, constitute the last part of the chapter.

Chapter 4 presents and interprets findings obtained from comparative analysis between the SSC and the GSC. The chapter commences with the discussion of findings on the distribution of Sepedi conjunctions in the SSC, followed by the discussion of findings on frequency of occurrence of Sepedi conjunctions between the SSC and the GSC. Furthermore, the comparative analysis of syntactic and semantic features of Sepedi conjunctions in the SSC and the GSC then follows. Finally, a comparison of the function and meaning of the conjunctions between the two corpora and scholarly sources is performed.

Chapter 5 restates the aim and objectives of the present research. It also gives a brief recap of the previous chapters before explaining how the study's aim and objectives were achieved. After summarising the findings related to each objective, the chapter discusses the potential impact of the study, as well as limitations of the study observed during the research. It concludes by outlining implications for future research and recommendations based on the findings.

# CHAPTER 2: LITERATURE REVIEW

## 2.1 Introduction

Linguistic studies on the South African Indigenous languages and other African languages across the continent have made significant contribution to our understanding of the linguistic diversity in Africa. These studies have helped to uncover the complex grammatical structures of these languages and have shed light on previously unknown aspects of their phonetics, phonology, morphology, and syntax, to name a few.

The use of large-scale linguistic corpora in these studies have been particularly valuable, as it has allowed linguists to analyse language use across different genres and time periods (De Schryver & Nabirye 2010). Ongoing studies on South African Indigenous languages and other African languages continue to build on this work, exploring new aspects of these languages and deepening our understanding of their linguistic systems. As linguistic research in African languages continues to advance, it has the potential to make important contributions to a range of fields, including education.

This chapter aims to provide a comprehensive literature review of studies relevant to the present research. It presents various scholars' perspectives on the subject matter. It begins with studies conducted on an international scale, followed by studies on the African continent and then studies within South Africa.

## 2.2 Corpus-based studies on a global scale

Numerous studies have been carried out on an international scale to explore the use and function of different parts of speech. Since English is the most widely spoken language in the world, a significant amount of research has been done on it. As a result, only the studies conducted on this language are discussed in this section. One of the studies was done by Roslim, Aziz, Abdullah, and Nimehchisalem (2021), which focused on the challenging nature of English prepositions and aimed to provide insights into their usage, teaching, and learning. The methodology employed in the study drew from various sources, including the British National Corpus (BNC) and the Companion Website for word frequencies in written and spoken English. The Longman Grammar of Spoken and Written English (LGSWE) also contributed to the study's corpus-informed materials

(Roslim *et al.* 2021). Texts from diverse corpora, including the Cambridge English Corpus, BNC, English of Malaysian School Students Corpus, and Malaysian English Language Textbooks Corpus, were examined.

The analysis phase involved extracting learner examples from the Cambridge English Corpus and categorising prepositions according to their Common European Framework of Reference for Language (CEFR) levels. These CEFR-listed prepositions were then compared with those in the BNC, and their ranks were determined. Concordance programs were employed to create analysis sheets for prepositions of place (e.g., in, on, at). Additionally, the study incorporated the Malaysian Textbook Corpus, processed through Optical Character Recognition (OCR) software for analysis using WordSmith Tools 4.0.

The study further conducted a qualitative analysis of preposition-related tasks and exercises found in textbooks. This analysis aimed to identify the types of exercises that students are required to complete, shedding light on textbook writers' preferences and teaching decisions regarding prepositions. The findings of the study yielded several insights into preposition usage and education. Firstly, the distribution of prepositions in the CEFR Grammar Profile was found to be helpful in guiding what to teach and learn. Secondly, the study identified a shift in the frequency of prepositions as CEFR levels progress, highlighting the increased use of complex prepositions at higher levels. Thirdly, the research suggested that students can achieve mastery in preposition usage through frequent and repetitive practice. Finally, the study recommended a more communicative approach to teaching prepositions, advocating for a variety of exercises evenly distributed among prepositions to enhance students' learning experiences.

The second study was conducted by Simpson and Mendis (2003) and it aimed to investigate the presence and functions of idiomatic expressions within the context of academic spoken language. A specialised corpus called the Michigan Corpus of Academic Spoken English (MICASE), containing recordings of contemporary speech made at the University of Michigan between 1997 and 2001, was used in the study. Initially, the researchers attempted to identify idiomatic expressions by examining English as a Second Language (ESL) textbooks, but found this approach to be less effective. As

such, they went back to a basic definition of idioms and searched the corpus directly. In order to gather idiom lists for their exploratory phase of research, they examined three university-level ESL textbooks that were released around the same time as MICASE (that is, between 1997 and 2001). The total frequency counts for each recognised idiom were determined using WordSmith Tools. Random transcripts from various academic divisions and speech event categories were selected and comprehensively reviewed.

The findings of the research revealed several noteworthy patterns. Firstly, the study demonstrated that idiomatic expressions are indeed present in academic speech and are not as infrequent as might be initially assumed when examining the language as a whole. Secondly, the distribution of idiomatic expressions within different sub-genres of academic speech did not exhibit predictable patterns based on either the degree of interactivity or academic divisions. Instead, the researchers inferred that the use of idioms appeared to be more reflective of individual speakers' idiolects (i.e., unique linguistic characteristics of individuals) rather than conforming to specific linguistic or content-related categories.

The third study was conducted by Gabrielatos, Davies, Rayson, Hunston, and Danielsson (2007), who investigated the modal load of *if* conditionals in the written BNC. The study aimed to analyse words or grammatical categories expressing modality as indicators of modalisation and compared the degree of modalisation in the *if* clause and main clause. The methodology involved selecting a random sample of 1 000 instances of *if* from the BNC, manually analysing the sample for the frequency of modalisation, and carrying out keyword analyses to establish the statistical significance of modality in the sample and sub-corpus of the BNC. The results showed that *if* conditionals have a significantly higher modal load than average, providing evidence for the claim that they attract modality more frequently than non-conditional sentences.

In the preceding discussion, three corpus-based studies were discussed that made significant contribution on an international scale. In the section which follows, attention is directed at other studies in some parts of the African continent, particularly Luganda language, a language spoken in Uganda, Kiswahili spoken in some parts of east Africa and Igbo spoken in the south eastern parts of Nigeria.

## 2.3 Corpus-based studies on the African continent

Several studies on the function of parts of speech have been conducted in different parts of the African continent. These include a study by Kawalya, de Schryver and Bostoen (2019) which focused on the analysis of Luganda language. In this study, the researchers aimed to examine the more grammaticalised markers, such as the modal auxiliaries -*téekw*- 'must, be obliged, be bound' and *lina*- 'have', and the verbal prefix -*andi*- 'would'. They analysed data from a four-million-word Luganda corpus spanning 13 decades and 18 topics/genres, using the WordSmith Tools to query specific items. Of the three markers, they demonstrated that -*téekw*- had the widest semantic range, encompassing the more arbitrary category of epistemic necessity. While -*andi*- solely communicates deontic necessity, -*lina* is exclusively connected to lexical usage and has a semantic range that does not extend past deontic necessity. The findings suggested that -*téekw*- probably originated as a passive of the verb -*téek*- 'make a law, bind (by law)' and has undergone a process of lexicalisation to acquire its current use.

Toscano and Sewangi (2005) also conducted research to investigate the difficulties in teaching Kiswahili learners to use the *amba*- locative relatives (*ambapo* for 'where-specific', *ambako* for 'where-general', and *ambamo* for 'where-inside') correctly. While grammars and dictionaries provide basic information, learners struggle to use the correct form in their communication, especially in writing. This study used a corpus-based approach in analysing a selection of contemporary KiSwahili literature to identify usage patterns for each *amba*- relative. It used the Concordance 3.00 tool for text analysis and UWAZO, a structured KiSwahili-Italian lexical database, to store the data.

The findings of this research indicated the emergence of behavioural patterns within the structured concordance data, which can be converted into valuable guidance for learners of Swahili as a second language. One significant suggestion arising from the study is to avoid using the term 'locative' when referring to forms of class 16, 17 and 18 of *amba*. Instead, it is recommended to use the terms 'place' and 'time' class, as they align better with actual usage, aiding learners in intuitively acquiring the language and producing more acceptable Swahili utterances.

Furthermore, the research by Okeke and Okeke (2022) explored the intricate interface between semantics and pragmatics, with a particular focus on the sensory verbs in the Igbo language. This investigation adopted a multifaceted approach by integrating the principles of cognitive semantics and neo-Gricean pragmatics, aiming to bridge the divide between linguistics and speaker's meaning. The central objective of the study was to unravel the semantic-pragmatic interface and its role in the construction of meaning. To accomplish this, the researchers employed a corpus-based methodology utilising the AntConc software to extract and examine instances of perception verbs within a wide range of conversational contexts. The study's corpus consisted of meticulously transcribed recordings of interactions involving 500 Igbo language speakers affiliated with the University of Nigeria. The analysis was particularly focused on the various standard Igbo variants of perception verbs, including *hụ́* (see), *lé* (look), *nụ́* (hear/smell/taste), *gé* (listen), and *mètụ́* (feel/touch). The findings of the research illuminated two fundamental aspects. Firstly, they revealed that utterances drawn from the extensive corpora are composed of discrete words. Secondly, they showed that these utterances carry both physical and metaphorical meanings, thereby shedding light on the intricate interplay of semantics and pragmatics in the construction of meaning within the Igbo language.

In the foregoing discussion, the studies conducted on various African languages were discussed, including studies on Luganda, Kiswahili and Igbo languages. In the following section, the corpus-based studies in South African Indigenous languages that made significant contribution are discussed.

### 2.4 Corpus-based studies within South Africa

South Africa is home to a diverse range of languages, many of which are part of the Bantu language family. However, despite the linguistic richness of the region, there has been a notable absence of corpus-based studies on these languages, until recently (Taljard, 2012). Taljard (2012) adds that the first comprehensive corpus-based study on a South African Indigenous language was only carried out in 2002. This is a study carried out by Gauton and de Schryver (2002), and it focused on the Zulu language. Through the use of a corpus-based approach, the study shed light on the use of the Zulu locative prefix *ku-*, and opened the door for further research into the Indigenous languages of South Africa.

The study by Gauton and de Schryver (2002) aimed to investigate the use of the class 17 locative prefix *ku-* in deriving locatives from nouns, and compared it to the *e-/o-...-ini* locativisation strategy. The study used the Pretoria Zulu Corpus (PZC), an organic five-million-token Zulu corpus, and found that the use of *ku-* (and to a lesser extent *kwi-*) is much more frequent with nouns other than those in classes 1/2, 1a/2a and [+human] nouns in class six than previously indicated in Zulu grammars. The study also revealed changes in the use of the prefixes over time, with *ku-* becoming more popular and *ko-* falling into disuse in the 1990s. Furthermore, the use of *ku-* (and to a lesser extent *kwi-*) was found to be much more common in certain genres. The study highlighted the importance of using structured corpora in linguistic studies to uncover previously overlooked aspects of language features.

Furthermore, Gauton, de Schryver, and Mohlala (2004) also conducted a study to highlight how the five-million-word Zulu corpus, i.e., PZC, was queried to investigate the use of the nominal suffix *-kazi* in Zulu. The study used WordSmith Tools to analyse the data and compared the findings with the views expressed by scholars. The study confirmed that the primary significance of the suffix *-kazi* is the expression of the feminine form, with the augmentative significance as secondary. The corpus data provided more evidence to corroborate the claims that nouns referring to domestic animals are typically the source of the feminine form, which is then added to the masculine form by suffixing *-kazi*. Nevertheless, the corpus data also demonstrated that, in contrast to certain traditional sources, *-kazi* is also utilised to create feminine forms from nouns representing wild animals.

Studies as the ones mentioned above show how corpus linguistics can be utilised to evaluate, support, or invalidate opinions and assertions made by scholars and grammarians studying the Zulu language. In the case of the locative prefix *ku-*, there is a general agreement among grammarians that it is used to locativise specific types of nouns and pronouns.

Moreover, Van Olmen, Breed, and Verhoeven (2019) conducted a study on the grammaticalisation of the new '*man*' pronoun in Afrikaans. They examined this process from the perspective of a fully grammaticalised '*man*' pronoun in Dutch. A diachronic

corpus of Afrikaans was used and compared it with two corpora of present-day Afrikaans and Dutch. The study found that Afrikaans is similar to other Germanic languages in having a '*man*' pronoun, and Dutch '*men'* is a suitable comparison. However, there are functional dissimilarities between the two, and Afrikaans '(*'n) mens'* need not develop the same features as Dutch '*men'*. The study also found that the referential definite interpretation of '(*'n*) *mens'* in Afrikaans and Dutch is infrequent and pragmatic.

The study by Taljard and de Schryver (2016) is another important research in the field of corpus linguistics. Through a comparison with data extracted from a sizable electronic corpus of Northern Sotho, the study sought to determine how accurate current descriptions of the language's noun class structure were. The 6,9 million running words in the Pretoria Sepedi Corpus (PSC), which includes over 155 000 distinct orthographic terms, served as the study's corpus. This investigation produced startling and enlightening results. The findings appeared to offer evidence against speakers' intuition as well as against the hierarchy of [+HUMAN] nouns and the idea of topicality in language. The Northern Sotho noun gender system was also revealed to be two-directional, with many single-class genders, rather than one-directional, singular-plural. The study also demonstrated that Northern Sotho's noun class system is dynamic, with nouns continuously being reinterpreted and assigned to new noun classes.

In the foregoing discussion, studies that have sought to explore the use of different parts of speech in different South African Indigenous languages have been highlighted. In the following section, attention is paid to the use of corpus within the educational setting, particularly the content that students are exposed to.

### 2.4.1 Corpus-based studies on teaching and learning material development

Corpus can be a valuable resource for material writers in the development of language teaching materials. By analysing the corpus data, material developers can get information on which words and phrases are most frequently used in a given language, as well as information on how to correctly use them.

Some scholars have raised concerns about the authenticity of language materials used in universities and other educational settings. They used corpus data to check for

discrepancies and ensured that learners were presented with appropriate language structures. For instance, Kennedy (1987), Holmes (1988) and Mindt (1995) have questioned the authenticity of existing language materials used in teaching Indo-European languages. Similarly, Römer (2004, 2005) as quoted in Granath (2009), has shown that German EFL textbooks differ significantly from authentic language use. Römer's (2004) comparison of corpus data and textbook materials reveals that the latter presents learners with inappropriate grammatical structures.

In the context of Indo-European languages, the use of corpora has become a key aspect of compiling and developing language teaching materials McCarten (2010). The application of corpora in the development of language materials reflects a commitment to finding more efficient and effective ways of developing materials that cater to the specific needs of language learners. This approach is particularly evident in the development of English language teaching materials, where the use of corpora has become a central aspect of material development. As such, it is clear that the incorporation of corpora in the development of language materials is an ongoing process that seeks to continually enhance the quality of language teaching and learning.

In recent years, corpus linguistics has made significant inroads into investigation of language teaching and learning in African languages, specifically South African Indigenous languages. Taljard's (2012) study on critical evaluation of existing pedagogical material for Northern Sotho exemplifies this trend. The study used a corpus-based approach to evaluate existing pedagogical materials, revealing patterns of language usage that are not reflected in Sepedi grammar books and other reference materials. This study, and others, demonstrates the potential of corpus linguistics to inform the development of pedagogical materials and improve language teaching and learning.

Taljard (2012) showed that improved learning material selection and sequencing can result from the use of corpus data in the development of pedagogical materials for teaching Northern Sotho as a second language. She discovered, in her research, that the majority of the existing materials are built around the structural model of grammatical description, with less focus on actual language use and the communicative significance of grammatical structures. By employing the Pretoria Sepedi Corpus (PSC), which

comprises 7,5 million words, Taljard (2012) managed to ascertain the language's most prevalent and communicatively significant structures. This data can guide material writers in selecting and sequencing content and examples to demonstrate content, leading to more effective language teaching.

When it comes to the use of corpora in language teaching, corpus linguistics can play a vital role in investigating how language is actually used in real-life situations. For instance, Rescki (2006) conducted a study on the use of corpus data in addressing grammatical questions posed by EFL/ESL teachers in four Orkut communities dedicated to teaching and learning English. The study found that using corpus data led to the discovery of patterns and meanings that were not easily found in other reference materials, such as grammar books and dictionaries. By providing real examples of language usage, corpus data can help teachers and students better understand how language works in context. Therefore, incorporating corpus data into language teaching can provide a more authentic learning experience and enhance students' ability to use English effectively in real-life situations.

The extensive literature review on corpus-based studies across various languages, ranging from English to African languages such as Luganda, Kiswahili, Igbo and South African Indigenous languages, such as Sepedi, Afrikaans, etc., underscores the valuable contribution these studies have made to the fields of linguistics and language education. However, it is evident that the area of conjunction usage or their function within the Sepedi language, especially within the context of corpus-based analysis, has remained largely unexplored. This study, therefore, endeavours to bridge this gap in existing literature and contribute to our understanding of conjunctions in Sepedi.

Now that a survey of literature has been provided on corpus-based studies that have sought to explore the use and function of parts of speech across the languages of the globe, it may be valuable to proceed to the discussion of theoretical framework underpinning the present study.

## 2.10 Theoretical framework

The theoretical framework underpinning this study draws from Noam Chomsky's seminal work on generative grammar, as articulated in his 1957 publication. Halle (1962) defines generative grammar as a structured collection of statements, rules, and axioms that serve the dual purpose of describing and defining all grammatically correct utterances within a given language, while explicitly excluding those that are not well-formed. Central to Chomsky's theory of generative grammar are the abstract conditions that govern the permissible form of statements within these grammars, offering a systematic framework for selecting the most appropriate description when faced with varying linguistic data (Halle 1962).

Yule (2010) extends this concept by positing that sentences are generated through subconscious procedures or rules, thereby motivating linguists to seek ways of modelling these underlying processes. Additionally, as noted by Brown and Miller (1991), generative grammar maintains a principled neutrality between the production and analysis of sentences, allowing for comprehensive linguistic analysis.

The principles of generative grammar theory were applied to a wide range of linguistic studies, such as semantic studies and corpus linguistics studies (Hausser, 2011 and Schiffer, 2015). The present study also contributes to the application of the theory to real-life linguistic studies.

Wasow (2003) summarises six principles that are shared by the vast majority of proponents of generative grammar theory in explaining the theory:

- Descriptive grammar is preferable to prescriptive grammar.
- Grammars should characterise competence, not performance.
- Grammars should be fully explicit.
- Analyses of language should be as broad as possible.
- There should be generalisations in grammar theory.
- Grammars should be psychologically relevant.

The principle that descriptive grammar is preferable to prescriptive grammar refers to the fact that the purpose of systematic language description is to replace traditional grammars' more anecdotal method (Wasow 2003).

The principle that grammars should characterise competence not performance means that generative grammar, across its various forms, consistently emphasises the characterisation of linguistic competence as a central feature. While some generative grammarians express an interest in constructing models for linguistic performance, the majority assert that a theory of competence is an essential element in such models. In other words, there is a widespread agreement that explaining the practical use of language necessitates a thoughtful comprehension of speakers' knowledge of the language itself. Thus, the focus on linguistic competence remains a fundamental aspect within the various frameworks of generative grammar (Wasow 2003).

The principle that grammars should be fully explicit means that traditional grammars assume a certain level of familiarity with the language being described and typically emphasise variable or altered aspects. In contrast, generative grammars aim to establish precise rule systems that define the entire language, avoiding the need for the reader to possess any prior knowledge of the language in question. The focus of traditional grammars is on elements that may vary or have undergone changes, while generative grammars strive for comprehensive and rule-based linguistic characterisation accessible to readers without pre-existing language knowledge (Wasow 2003).

The principle that analyses of language should be as broad as possible means that generative grammarians consider a more concise grammar superior when it covers the same data range as another with two distinct rules. This supports the idea that simplicity in rule formulation is indicative of grammatical superiority (Wasow 2003).

The principle that there should be generalisations in grammar theory meaning that in analysing individual languages, prioritise deriving facts from predominant linguistic principles, thereby minimising reliance on language-specific grammars. This, therefore, entails that more emphasis should be put no general principles applicable across languages, for a comprehensive understanding (Wasow 2003).

Concerning the sixth principle stipulating that grammars should be psychologically relevant, generative grammarians characteristically take their theories to be relevant to psychological questions in this tenet. Chomsky (1986:3) emphasises the significance of 'a particular generative grammar' as a theory delving into the cognitive aspects of individual's proficiency in a specific language. He contends that a robust universal grammar theory is essential to explain effortless language acquisition. This implies a genetic predisposition in humans, characterised as a 'mental organ', facilitating the innate acquisition of particular language types (Wasow 2003).

Furthermore, Chomsky (quoted in Chapel and Clause 2021) reasoned that a child must be born with the ability to learn a language. This perspective maintains that biology has predetermined the process, since language is already ingrained in the neural circuits of the human brain, which has evolved over time. The brain can interpret what it hears in light of fundamental principles or structures it already possesses. Hearing speech stimulates the child's intrinsic capacity to learn language. The term 'language acquisition' means this innate skill. Obviously, Chomsky did not mean to imply that a native English speaker is born with any particular knowledge of the language. There are universal principles shared by all languages spoken by humans, he said. For instance, all humans have nouns and verbs to describe the objects and activities in their world. The onus is on the child to figure out how the particular language s/he encounters conveys these general ideas.

Furthermore, Stanborough (2019) argues that human beings are the only species born with the capacity to use language. Stanborough (2019:1) further echoes Chomsky's argument that 'language is innate faculty'- in other words, what he refers to as 'universal grammar' is the collection of laws about language that humans are born with. In his argument, he further highlights that we acquire our native languages.

This theoretical framework is particularly pertinent to the present study, which focuses on the examination of conjunctions in a corpus of Grade 7-9 Sepedi textbooks as well as that of general language. The theory is relevant for the present study in the sense that it concerns itself with how individuals acquire and construct grammatical structures in various languages and the present study seeks to investigate how conjunctions function

within Sepedi language as well as how learners in this language acquire this function or usage of conjunctions. To ascertain the actual usage of conjunctions, the study utilises a specialised Grade 7-9 Sepedi textbook corpus, which is compared against a general Sepedi corpus used as a reference corpus. The integration of generative grammar into this research holds promise in contributing to the reconstruction of conjunction usage within the Sepedi language.

## 2.11 Conclusion

This chapter provided an overview of corpus-based studies conducted on an international scale, offering valuable insights into linguistic research spanning multiple languages. A notable proportion of these studies have been conducted on the English language, reflecting the prominence of English as a global lingua franca and a focus of linguistic analysis. These investigations have contributed to our understanding of English language structure, usage, and variation. Subsequently, corpus-based studies extend their scope to include languages such as Luganda, Kiswahili, and Igbo, which are spoken in diverse regions of Africa. These studies have enriched our knowledge of language phenomena specific to these languages, highlighting the linguistic diversity present on the African continent. Furthermore, a segment of corpus-based research delves into South African Indigenous languages, shedding light on the unique features and characteristics of these languages. Additionally, this chapter explored corpus-based studies focusing on the development of teaching and learning materials, emphasising the practical application of corpus linguistics in educational context. Finally, the chapter highlighted the theoretical framework that underpins the present study, drawing from Noam Chomsky's seminal work on generative grammar published in 1957. Chomsky's generative grammar theory has had a profound influence on the field of linguistics.

This chapter presented a comprehension literature review of studies relevant to the present study. The next chapter provides a discussion of the research methodology employed in the present study.

# CHAPTER 3: METHODOLOGY

## 3.1 Introduction

The advent of technology has brought about new and innovative ways of presenting and treating parts of speech based on the frequency and context in which they appear in a language. The use of corpus data provides language users with a better understanding of words in their contextual usage. As noted by Conrad (2010), querying a corpus using software tools like WordSmith Tools, AntConc, Sketch Engine, LancsBox X and others, can provide a wide range of information on the behaviour, meaning, and grammatical patterns of words in a language.

This chapter concerns itself with a discussion of the research methodology proposed for this study, as well as the discussion of corpus linguistics as methodology for the study of language. Furthermore, a discussion on corpus-based vs intuition-based approach and on corpus-based vs corpus-driven approach follows. The chapter further defines corpus and different types of corpora. Additionally, a discussion of the software particularly used in this study as well as other corpus query software applications that can be used to query a corpus is provided. The process of creating the corpus for the current study, which serves as the foundation for the analysis, is covered in more detail in this chapter. Finally, some of the corpus query tools provided by LancsBox X, the corpus query software chosen for this investigation, are presented. Some of these tools are employed in the data analysis for this study.

## 3.2 The proposed methodology

The approach to data analysis used in this study is corpus-based. For data analysis and interpretation, the study uses both the qualitative and quantitative methods. McEnery and Wilson (1996) argue that both qualitative and quantitative analyses offer valuable insights in corpus research. Qualitative analysis contributes depth and precision, while quantitative analysis yields statistically reliable and broadly applicable findings. In order to test objective hypotheses, quantitative research, according to Creswell & Creswell (2018), entails analysing correlations between variables. In this study, the quantitative

method is applied to count Sepedi conjunctions and uses statistical data in an effort to explain observed patterns.

The aim of qualitative methods, according to Babbie and Mouton (2001), is to understand and describe human behaviour rather than to interpret it. By employing this technique, researchers aim to comprehend the study from the viewpoint of the participants (Babbie and Mouton 2001). In this research, the qualitative method is employed to describe the function and usage of Sepedi conjunctions in their authentic occurrences. In contrast to quantitative research, in the qualitative method data is used to identify and describe the usage of conjunctions in their authentic appearances.

It has already been highlighted in the foregoing discussion that the present research uses both the qualitative and quantitative method. In research, this is referred to as 'triangulation'. Triangulation refers to the use of multiple research methods in a single study (Greene 2005 and Babbie 2010). According to Saldanha and O'Brien (2014), triangulation is the basis of reliable and superior research. Every research method has pros and cons of its own, and using a 'mixed method' strategy is the best way to improve study outcomes while minimising the drawbacks of using distinct approaches. Lewin (2005: 215) rightly expresses this by suggesting that:

> The use of mixed methods has become increasingly popular as a means to harness the strengths of both approaches, triangulate data and illuminate statistical findings with, for example, case studies and/or vignettes.

Therefore, the incorporation of both qualitative and quantitative methods in examining the representation and distribution of Sepedi conjunctions within the Sepedi corpora will prove to be of immense value to the current study. Both corpora used in the present study are raw corpus. This means it is not possible to search for only Sepedi conjunctions. Therefore, BONSE compiled dictionary is used to identify the conjunctions to form the focus of the present study. In the BONSE, only 20 Sepedi conjunctions are treated. Therefore, based on the frequency of occurrence of the conjunctions in the two corpora, only six conjunctions, namely *ebile*, *ge*, *goba*, *gomme*, *gore* and *mola* form the focus of the study.

In the preceding discussion, the methodology followed in the present study has been expounded. In the following section, the focus is on Corpus linguistics as a methodology.

### 3.3 Corpus linguistics as a methodology for the study of language

Corpus linguistics (CL) is a complex field or discipline within linguistics. CL concerns itself with the systematic analysis of language based on the authentic collection of naturally occurring texts, known as corpus. Corpus linguists have different perspectives on this issue, contributing to nuanced discourse within the academic community. Bennet (2010) states that when analysing language using corpora, there is a methodology that should be employed when analysing data. Furthermore, Tognini-Bonelli (2001) asserts that CL has evolved into an independent discipline and has surpassed its methodological role. It has changed the way in which we look at language, and has resulted in a new philosophical approach to linguistic enquiry. However, the general consensus is that it is a methodology rather than an independent discipline of linguistics, such as phonetics, semantics, syntax, etc. CL is not limited to any one area of language study; rather, McEnery *et al*. (2006) assert that it can be used to investigate nearly any area of language study. According to McEnery *et al*. (2006), CL should be viewed as a methodology with a broad variety of applicability across numerous linguistic theories and fields.

It is possible to identify four main characteristics of CL;

- Firstly, it is empirical since it examines real language use patterns in authentic texts.
- Secondly, it makes use of real language as the foundation for analysis; any real-world situation involving language communication can serve as a corpus.
- Thirdly, the foundation for CL analysis is a corpus, which is a sizable, logically arranged collection of natural texts.
- Fourthly, CL uses computers and software to store data and analyse texts.
- Finally, CL relies on both qualitative and quantitative analysis methods.

The following section performs a comparison of the corpus-based approach and the intuition-based approach to linguistic analysis.

## 3.4 Corpus-based approach vs. intuition-based approach

Intuition-based language research has been, and still is, an accepted approach to investigating language phenomena. Researchers make use of their intuitive knowledge of the language. This is also called an introspective approach. An intuition-based approach is convenient, since it is readily available. However, there are some issues to be considered:

### *3.4.1 Idiolect, sociolect and dialect effect*

- Variation in personal language (idiolect), societal language norms (sociolect) and regional idiolect pose challenges to the generalisation of findings.
- What is acceptable for speaker/researcher A is not necessarily acceptable for speaker/researcher B. The corpus offers evidence of what speakers consider to be acceptable utterances, assuming that what we observe in a corpus is generally grammatical and/or acceptable.

### *3.4.2 Subjectivity in language monitoring*

- The act of monitoring one's language production introduces subjectivity.
- Speakers keep an eye on their language production when they make up an example to support or refute a claim. Even if the researcher determines that a given sample satisfies grammar rules, it might not accurately reflect common language usage.

### *3.4.3 Verifiability challenges*

- Since introspection cannot be observed, results based on it are difficult to verify.
- Lack of verifiability undermines the credibility of research findings.

Intuition-based approach largely ignores corpus data, but a corpus-based approach does not exclude or reject intuition. A corpus-based approach uses empirical data, but also accepts the value of the researcher's intuition. The ideal is to find a balance between corpus-based and intuitive-based approach to language research.

However, not every research question can be addressed using a corpus-based approach. The next section discusses two approaches that use corpora as the basis for analysis: the corpus-based approach and the corpus-driven approach.

### 3.5 Corpus-based approach vs. corpus-driven approach

The corpus-based approach and the corpus-driven approach are the two branches of the corpus approach. When using the corpus-based approach, corpora are mostly utilised to test, elaborate, or provide examples of theories and descriptions that were developed prior to the availability of sizable corpora that could support language research (Tognini-Boneli 2001). On the other hand, according to Tognini-Boneli (2001), the corpus-driven approach is allegedly fully dedicated to maintaining the integrity of the data in its entirety.

McEnery *et al*. (2006) discuss four basic distinctions between the corpus-based and corpus-driven approach:

- type of corpora used,
- attitude towards existing theories and intuition,
- focus of research, and
- paradigmatic claims.

See **Table 3.1** below, where the differences between the two approaches are explicated.

**Table 3. 1:** Corpus-based approach versus corpus-driven approach

| | **Corpus-based approach** | **Corpus-driven approach** |
|---|---|---|

| | | |
|---|---|---|
| **Type of corpus used** | In the corpus-based approach, there is a concerted effort to achieve balance and representativeness in the corpus, as well as to annotate the corpus with objective criteria. | The corpus-driven approach does not make serious efforts to achieve balance and representativeness in its corpus data, and there are no objections to annotating the corpus data. |
| **Attitude towards existing theories** | In the corpus-based approach, linguists have existing theories as a starting point to analyse language. | The corpus-driven approach relies solely on data from the corpus to generate linguistic descriptions. |
| **Different research focus** | Makes distinction between lexis, syntax, pragmatics, sematic and discourse | Makes no distinction between lexis, syntax, pragmatics, sematic and discourse |
| **Paradigmatic claims** | The corpus-based approach is considered to be less radical than the corpus-driven approach. | The corpus-driven approach is considered to be a new paradigm for language description that claims to be capable of describing an entire language. |

**Table 3.1** above indicates that the corpus-based approach and the corpus-driven approach are distinct from each other and that the approach chosen can have a significant impact on the findings of a study on language use. In the section that follows, the focus is on designing a corpus.

## 3.6 Designing a corpus

The process of constructing a corpus involves several stages, which can vary depending on the specific project's requirements. The multitude of steps in corpus compilation can rather be labour-intensive and time-consuming. It becomes more difficult for researchers wishing to use a corpus-based approach because of the apparent lack of funding and attention from relevant organisations for corpus compilation and corpus studies overall (Saldanha and O'Brien 2014). One essential component of any corpus-based research is the building of a corpus. In essence, the corpus is the foundation of any corpus-based research because without it, an inquiry of this kind cannot be conducted (Dlamini 2021). However, before getting into the nuances of corpus creation, it might be prudent to begin by providing a description of a corpus and outlining the many types of corpora.

## 3.7 Defining a corpus and types of corpora

Loosely defined, a corpus (the plural, corpora) is a large sample of database of specific language consisting of either written or transcribed spoken language (Gabrielatos 2005). A corpus is considered to be 'a collection of (1) *machine-readable* and (2) *authentic texts* (including transcripts of spoken data) which is (3) *sampled* to be (4) *representative* of a particular language or language variety' (McEnery *et al*. 2006:5). Additionally, Bowker and Pearson (2002) assert that a corpus is defined as a sizable collection of real texts that have been systematically compiled in an electronic format according to specific criteria.

Different types of corpora exist, and below is a description of most, if not the entirely, of the corpora that can be used when performing various linguistic activities. Since corpora are built for different purposes, they come in many shapes and sizes (Gabrielatos 2005). Various types of corpora and the purposes they are intended to fulfil are described below, namely general corpus, specialised corpus, written corpus, spoken corpus, synchronic corpus, diachronic corpus, learner corpus, and monitor corpus.

### *3.7.1 General corpus*

A **general corpus** is a corpus that attempts to be representative of a particular language as a whole. The British National Corpus (BNC) is a widely recognised general corpus, as noted by McEnery *et al*. (2006). The BNC, which is organised into 4 124 written texts and

spoken scripts in contemporary British English, contains 100 106 008 words. The corpus aims to capture the broadest spectrum of contemporary British English. Additionally, Davies (2008) offers the American National Corpus and the Corpus of Contemporary American English (COCA) as examples of general corpora. Bennett (2010) highlights that it is crucial to emphasise that no single corpus has the capacity to capture the entirety of conceivable language; its primary function is to furnish users with comprehensive insights into language usage.

### *3.7.2 Specialised corpus*

A **specialised corpus** is a corpus that attempt to represent a particular kind of language. Simply put, it is a corpus that is domain-specific. McEnery *et al.* (2006: 60) echo this view when they contend that a specialised corpus 'can be domain or genre specific to represent a sub-language'. The Guangzhou Petroleum English corpus, which comprises 411 612 written English words from the petrochemical domain, is an example of such a corpus. Additionally, the MICASE is one of the specialised corpora that Simpson, Briggs, and Swales (1999) present as examples. Spoken language that originates from an academic environment is included in MICASE. Another illustration is the CHILDES corpus, which is devoted to children's language and is currently incorporated into the TalkBank system's children language project. Moreover, the Michigan Corpus of Upper-level Student Papers (MICUSP) is a medical corpus. MICUSP not only captures the language employed by hospital workers and nurses, but also houses a collection of papers spanning diverse university disciplines.

### *3.7.3 Written corpus*

A **written corpus** is a type of a corpus that comprises only words of written texts. McEnery *et al*. (2006: 61) point out that:

> The first corpus of English was a corpus of written American English, the Brown University Standard Corpus of Present-Day America English. The corpus was compiled using chunks of approximately 2,000 words of written texts. These texts were sampled from fifteen categories. All were produced in 1961.

A few well-known examples of this type of corpora are the written American English Brown corpora, the written British English Lancaster Oslo-Bergen (LOB) corpus, the

written New Zealand English Wellington Corpus (WSC), the written English Australian Corpus (ACE), and the written Indian English Kolhapur Corpus. One thing unites them all: they are all roughly one million words long and purposefully designed to be somewhat similar. Consisting of 500 texts, each about 2 000 words in length, these corpora are meticulously designed to provide a representative sampling across a diverse range of written genres (Lee 2010).

### *3.7.4 Spoken corpus*

A **spoken corpus** is a type of a corpus that comprises transcribed spoken texts, either from radio broadcasts, or dialogue, or monologue. McEnery *et al*. (2006: 62) give examples of spoken English corpora that are available:

> These include, for example, the London-Lund corpus (LLC), the Lancaster/IBM Spoken English Corpus (SEC), the Cambridge and Nottingham corpus of Discourse in English (CANCODE), the Santa Barbara corpus of spoken American English (SBCSAE) and Wellington Corpus of Spoken New Zealand English (WSC).

The British Received Pronunciation (RP), the SEC; (53 000 words), which is primarily composed of radio broadcasts between 1984 and 1991, and the LLC are other well-known examples of this type of corpus (Lee 2010).

### *3.7.5 Synchronic Corpus*

A **synchronic corpus** is a type of a corpus that is mainly used to compare varieties of a particular language. Kennedy (1998: 20) defines this type of a corpus as, ‘an attempt to represent a language or a text type at a particular time’. An example of a synchronic corpus is the Brown corpus. One collection of written American English texts is the Brown corpus, which was released in 1961 (Kennedy 1998).

Specifically created for the synchronic study of international English, the International Corpus of English (ICE) is a well-known example of this type of corpus. In nations or regions where English is a first or official language, ICE seeks to construct 20 corpora of one million words each, comprising of spoken and written English created between 1990 and 1994 (Xiao 2009).

### *3.7.6 Diachronic/Historical corpus*

A **diachronic/historical corpus** is a type of corpus that contains texts from the same language from different periods and used to track the changes of language evolution (Kennedy 1998). Kennedy (1998: 22) points out that, in contrast to synchronic corpora, diachronic corpora:

> Represent a language over a period of time. The diachronic part of the Helsinki corpus of English texts, for example, contains English texts covering the period from AD 700 to AD 1700 and can be used, among other things, for studying language change.

The Helsinki Corpus of English, ARCHER (A Representative Corpus of Historical English Registers), and COHA (Corpus of Historical American English) are well-known examples of this type of corpus. The 1,6-million-word Helsinki Corpus includes Old English (413 300 words), Middle English (608 600 words), and early modern (British) English (551 000 words), covering the period from around 750 to 1700. In contrast, the 1,8-million-word multi-genre corpus known as ARCHER spans both British and American English from the early modern era (1650–1990) to the present. These corpora serve as valuable tools for studying language evolution and usage across different historical periods and geographical regions (Lee 2010).

### *3.7.7 Learner corpus*

A **learner corpus** is a type of a corpus that contains written and/or spoken texts produced by learners in a classroom. McEnery *et al*. (2006: 65) point out that 'a type of a corpus that is immediately related to the language used in classroom is the learner corpora'. The renowned International Corpus of Learner English (ICLE) serves as an example of this category (Granger 2003). This corpus encompasses essays crafted by learners of the English language.

### *3.7.8 Monitor corpus*

A **monitor corpus** is a type of a corpus that is not constant in size as it is expandable. Simply put, texts are consistently and constantly being added. McEnery *et al*. (2005: 67) state that a monitor corpus:

> Is consistently (e.g., annually, monthly or even daily) supplemented with fresh material and keeps increasing in size, though the proportion of texts type included in the corpus remains constant. Corpora of this kind are typically much larger sample corpora. The Bank of English (BoE) is widely acknowledged to be an example of a monitor corpus. It has increased in size progressively since its inception in the late 1980s and is around 524 million words at present.

According to Atkins, Clear and Ostler (1992), texts selected for monitor corpora undergo careful and ongoing examination, with data extracted for database incorporation, although not retained indefinitely. Examples of this corpus category encompass the BoE and COCA, both of which offer online search capabilities (Wynne and Berglund 2012 and Davies 2008).

Corpora are also classified in terms of design criteria. Corpora that are classified in this manner include, but not limited to, the monolingual corpus, bilingual or multilingual corpus, parallel corpus and comparable corpus. Let us discuss each of these.

### *3.7.9 Monolingual corpus*

A **monolingual corpus** is a type of corpus that comprises an extensive compilation of texts all written in a single language (Kruger 2002). Examples of well-known corpora of this type exist worldwide and include the 100-million-word BNC and the 200-million-word Cobuild Bank of English (CBE) (Kruger 2002 and Kenny 2009). The PZC, which has five million words, and the PSC, which has 10 million words, are two of the South African Indigenous language corpora that the University of Pretoria has created (Prinsloo 2015 and Gauton *et al*. 2004). The fact that these corpora are constantly expanding and might grow now or in the future is important to note.

### *3.7.10 Bilingual or multilingual corpus*

A **bilingual or multilingual corpus** is a collection of two or more separate monolingual corpora, which can be assembled either within the same institution or different institutions, all following similar design principles (Baker 1995). A corpus is classified as bilingual when it consists of only two languages, and as multilingual corpus when it consists of many languages. According to Kenny (2009), these corpora are made up of authentic texts in each language—translations excluded. Baker (1995) notes that the Council of Europe Multilingual Lexicography Project is a well-known example of this kind of a corpus.

### *3.7.11 Parallel corpus*

A **parallel corpus** comprises original texts in language A and their corresponding translations in language B (Baker 1995). Such a corpus can primarily be in one of two forms, according to Kruger (2002): a bilingual parallel corpus, which includes original texts and their translations into a single additional language, or a multilingual parallel corpus, which includes original texts and their translations into multiple languages. According to Baker (1995), the Hansard Corpus—which includes the Canadian Parliament's proceedings in both English and French—is an illustration of a generally acknowledged parallel corpus.

### *3.7.12 Comparable corpus*

As stated by Baker (1995: 234), 'two distinct sets of texts in the same language make up a **comparable corpus**: one set consists of original texts in that language, and the other set consists of translations into that language from a specified source language or languages'. An ordinary monolingual corpus is what the original texts are basically. It is stated by Kruger (2002) that these two sets are frequently referred to as the 'translational corpus' and the 'non-translational corpus', alike. The linguistic diversity, domain, length, and time period of the two corpora should be comparable. They ought to reflect the diversity of both original authors and translators, as noted by Baker (1995). One well-known example of such a corpus, according to Saldanha and O'Brien (2014), is the Translational English Corpus (TEC), created by the University of Manchester's Center for Translation and Intercultural Studies.

The present study will use two corpora: a specialised corpus, which solely includes Sepedi home language textbooks, and a general corpus, which serves as a reference corpus. Given that they are both collections of texts in the same language, i.e. Sepedi, they are classified as monolingual corpora.

The preceding discussion has offered a thorough description of a corpus and different types of corpora. It may now be prudent to refocus the discussion on corpus design, specifically addressing the concerns of balance, representativeness and size.

### 3.8 Balance and representativeness

Creating a corpus involves a series of considerations that the corpus compiler must address. Among these are representativeness and balance, two important aspects to take into account when building a corpus. According to Gouws and Prinsloo (2005), a well-balanced corpus usually includes texts from a wide range of text types and genres. As stated by Saldanha and O'Brien (2014), a corpus is usually taken to be representative of the language or of a particular subset of it. A representative corpus should cover all of the basic features of a language, as well as a wide range of linguistic variations and text genres. Furthermore, Gouws and Prinsloo (2005) note that it should include a sufficient number of instances of words and phrases to enable researchers to draw valid and accurate conclusions regarding lexical patterns.

As highlighted by, Kenny (2009a), in Gouws and Prinsloo (2005) and Saldanha and O'Brien (2014), defining the population from which to collect the sample is essential to the process to develop a representative corpus. However, this choice is contingent upon the type of corpus one intends to develop and its intended purpose. As mentioned previously the two corpora used will be employed: specialised corpora, comprising selected Sepedi learners' textbooks, and a general corpus. The specialised corpus consists of one tittle of the Sepedi Home Language learners' textbooks utilised in the Senior Phase. The selection was made from Sepedi learners' textbooks conforming to the CAPS guidelines, published between 2012 and 2022. This decision aligns with the implementation of the CAPS curriculum in 2013, that outlines the teaching guidelines (Ojo and Mathabathe 2021). As for the general corpus, it is constructed from internet sources and supplemented with plain texts received from the University of Pretoria, Department of African Languages and its design will be elaborated on in **Section 3.11.1**.

### 3.9 Corpus size

The considerations of balance and representativeness are intricately linked to the debated issue of corpus size, as extensively discussed by Saldanha and O'Brien (2014). There is a prevailing notion that a larger corpus is generally deemed more favourable, as noted by Gouws and Prinsloo (2005). In the early stages of corpus studies, a 'standard' corpus size was established at one million words. Over time, however, much larger corpora were

created; these include the 100-million-word BNC and the 200-million-word CBE, respectively. Reliability and the type and purpose of the corpus continue to influence the size question. For example, a smaller corpus might be more suitable when concentrating on specialised language such as financial texts, legal documents, or weather forecasts. A larger corpus, however, might be required for studies utilising broad language or a greater variety of text types (Saldanha and O'Brien 2014). In the present study, the GSC comprises 5 198 098 running words (tokens), which is deemed sufficiently large for the current research. The SSC has 311 821 running words (tokens), which is also large enough for a specialised corpus. The corpus sizes are shown at the bottom left-hand side corner of **Figure 3.1** below.

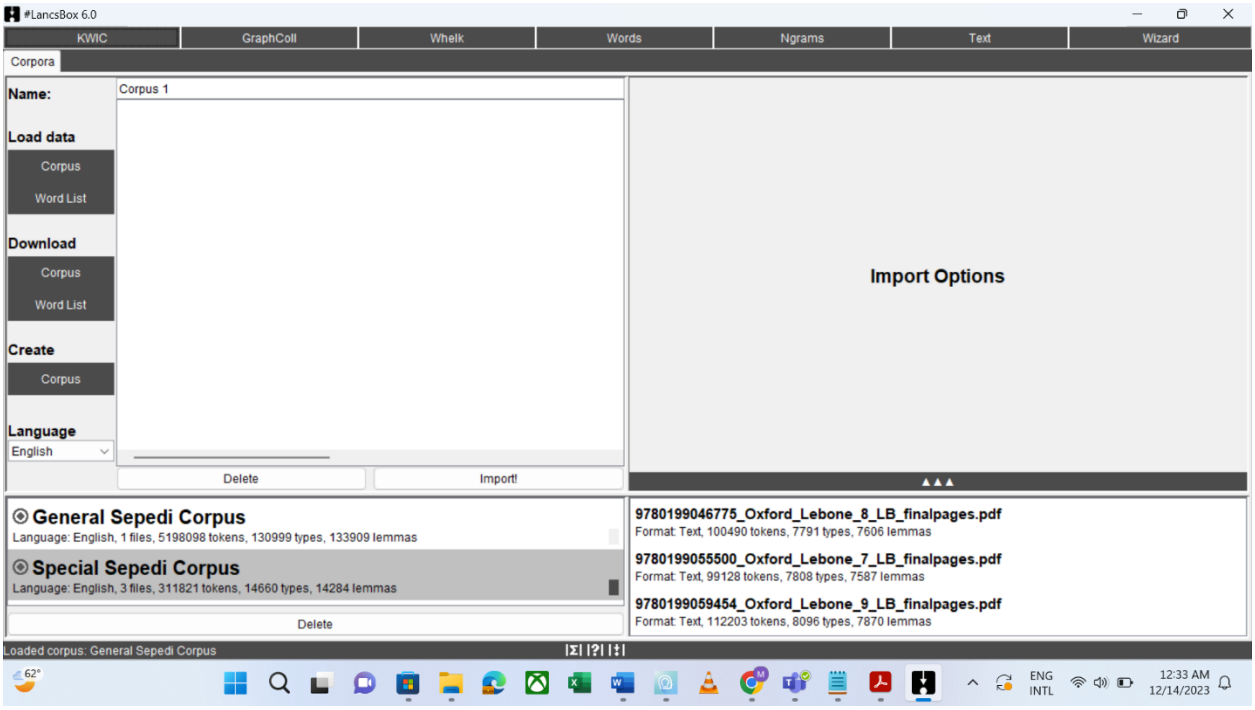


***Figure 3. 1**: LancsBox X screen showing word counts (tokens) for GSC and SSE*

Now that we have discussed the aspects of balance, representativeness, and size, it is time to delve into the discussion of available software options for analysing a corpus.

### 3.10 Corpus query software

As Anthony (2013) notes, corpora are frequently described as the essential elements of Corpus linguistics. Nonetheless, it is critical to acknowledge that corpora are

fundamentally linguistic data, and that specialised software tools are required for their effective analysis and retrieval (Anthony 2013). According to Gouws and Prinsloo (2005: 25), 'Corpora, in themselves, have limited utility unless tools are accessible to manipulate the data in various ways'. There are several software programs that can be used to analyse corpora for different types of research objectives. These programs include MonoConc, AntCoc, WordSmith Tool, Sketch Engine, ParaConc and LancsBox X.

### *3.10.1 MonoConc*

**MonoConc** is a specialised software used for searching monolingual corpora and for text analysis aimed at revealing formal language patterns. Among other language patterns, this entails determining the most commonly used terms in a corpus, words that only occur once (hapax legomena), and collocations (Barlow 2003). Moreover, the application provides the ability to conduct sophisticated searches with part-of-speech tags and regular expressions (Barlow 2003). The program's installation process, which is as simple as copying it from a CD-ROM or floppy disk and pasting it into the computer, is what really makes it stand out (Barlow 2003).

### *3.10.2 AntConc*

**AntConc** is a notable software tool described as a 'freeware, multi-platform, multi-purpose corpus analysis toolkit' (Anthony 2004: 7). It was created in Japan by Laurence Anthony (Anthony 2004 and Froehlich 2015). Being a free tool is one of its unique features that distinguishes it from other software programs. Anthony (2004) highlights that this particular feature makes the software extremely appropriate for individuals, organisations, or institutions that are operating under financial constraints. Similarly, to MonoConc, AntConc it is designed specifically for use with single-language (monolingual) corpora.

### *3.10.3 WordSmith Tool*

**WordSmith Tool** is an integrated suite of software tools designed for analysing the behaviour of words within texts. It was developed by Mike Scott (Scott 1998; Gouws and Prinsloo 2005). As stated in Scott's words: 'You will have access to tools for investigating how words are employed in your own texts, or in texts authored by others' (Scott 1998: 7). The software, according to Mncwango (2017), accommodates both inexperienced and

experienced users and provides a variety of tools appropriate for corpus linguistics analysis. Unlike AntConc, WordSmith Tools can be purchased at a reasonable cost from the Oxford University Press website (Mncwango 2017). Gouws and Prinsloo (2005) pointed out that WordSmith Tools is the most widely used software in South Africa and is thought to be the best option for lexicographic tasks, such as working with dictionaries. WordSmith Tools, like MonoConc and AntConc, is made expressly to work with monolingual corpora.

### *3.10.4 Sketch Engine*

The UK-based company, Lexical Computing Ltd., created the corpus analysis program '**Sketch Engine'** (Kunilovskaya and Koviazina 2017; Kilgarriff 2008). The original developers of this software are Pavel Rychlý, a Czech programmer, and Adam Kilgarriff, a renowned British lexicographer and corpus linguist (Kunilovskaya and Koviazina 2017). The software itself and the online service are the two components that are included in the name 'Sketch Engine'. As a result of operating online and requiring an internet connection to work, this program offers users a web service (Kilgarriff, Baisa, Bušta, Jakubíček, Kovář, Michelfeit & Rychlý 2014). One of the software's primary functions is to generate 'word sketches,' which are essentially brief summaries of a word's grammatical usage and related collocations (Kilgarriff, *et al*. 2014). This is where the term comes from.

### *3.10.5 ParaConc*

For a simultaneous corpus analysis, there is also another program called **ParaConc**. 'A simple software tool designed to facilitate the analysis of translated texts', is how Barlow (2008: 12) characterises ParaConc. The primary purpose for its creation, was as a search engine for parallel texts. Its main goal is to let linguists and researchers evaluate translated literature objectively and investigate the translation process itself (Barlow 2001). ParaConc is similar to MonoConc in that they were both developed in the United States by Michael Barlow (Barlow 2008). The names of these two software programs seem to correspond to the kind of corpus that they may work with: 'ParaConc' is associated with parallel corpora, whilst 'MonoConc' is associated with monolingual corpora. Like MonoConc, ParaConc can be quickly and simply copied from a flash drive or CD-ROM and pasted onto the user's computer, eliminating the need for a traditional

installation and expediting the corpus analysis process. These two software tools differ from the others due to this specific feature.

### *3.10.6 LancsBox X Tool*

The **LancsBox X Tool** is an open-source program created by Lancaster University, that may be downloaded for free from Lancaster University's website. It serves as a useful tool for language analysis and visualisation. Notably, users, whether linguists or individuals from other fields, have the flexibility to input their own data in any format or utilise the data provided with the tool. The software efficiently searches through the specified data, referred to as corpora, generating graphs and conducting statistical analyses. LancsBox X is well-equipped to analyse monolingual corpus, allowing the researcher to explore and extract linguistic patterns within a single language (Brezina, Timperley and McEnery 2015). What sets this program apart is its user-friendly installation process, which involves downloading it from the internet and after installation, it does not require internet connection for querying the corpus or uploading the corpus (Brezina *et.al*. 2015).

Each of the aforementioned corpus query programs offers a variety of tools for carrying out corpus-based analysis. However, the 'Concordance' and 'Word/Frequency list' capabilities are the most common and widely used in these programs.

MonoConc, AntConc, WordSmith Tool and LancsBox X are all well-suited for the manipulation of monolingual corpora. Notably, MonoConc, AntConc, and LancsBox X are freely available tools, in contrast to WordSmith Tools, which requires the purchase of a license, incurring significant costs. With the nature of the current study as well as its aim and objectives, LancsBox X was selected as the preferred software. Persuasive factors in this decision include LancsBox X 's remarkable user-friendliness, a characteristic that becomes evident in the section dedicated to the querying process in the present research. Furthermore, the software does not require internet connection for querying the corpus or uploading the corpus and is adept at handling corpus in diverse formats, including, but not limited to, .txt, .xml, .doc, .docx, .pdf, .odt, .xls, .xlsx, and .zip (Brezina *et al*. 2015).

## 3.11 Designing the GSC for present research

In the design of the GSC, a software called BootCaT toolkit was employed. BootCaT toolkit is a specialised software designed for corpus creation. This tool facilitated the systematic assembling of the corpus, aligning with the established guidelines and procedures outlined in **section 3.8** and **3.9** above. In the section which follows, the process of designing the general corpus is explained in detail.

### *3.11.1 Using BootCaT to create GSC*

The GSC was built using BootCaT toolkit, a software used for corpus creation. The software is freely available for everyone who wishes to create a corpus for themselves.

The purpose for using this software was to create a general language corpus that can be used as a reference corpus, to fulfil the aim and objectives of the study. When using a BootCaT toolkit, the corpus builder has to provide seeds, i.e., words that will identify generic lexical items. The BONSE dictionary was consulted to get a list of Sepedi words, particularly Sepedi words that contain diacritics, such as the '*š*' sound, so that the material retrieval could be documents that are in Sepedi language and not Setswana or Sesotho. The dictionary was consulted for the reason that its micro and macro structure were compiled using a corpus. This implies that the words that are treated in it are the most frequently used words in Sepedi. Their treatment in the dictionary is based on real-life language usage. From the BONSE dictionary, 350 words were identified as seeds to be used in the search engine and were grouped into six, since the toolkit allows five or more unique seeds (see **Figure 3.2** below).

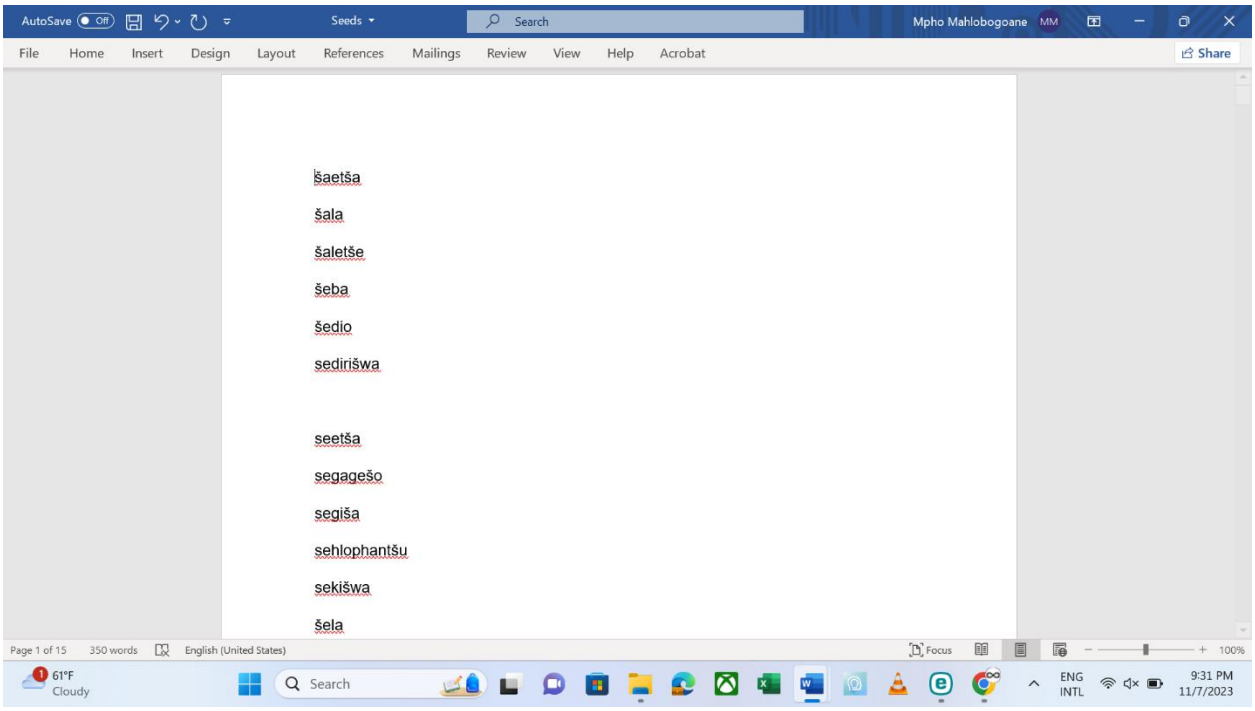


***Figure 3. 2****: Seeds to be used in the search engine grouped into six*

All six seeds (*šaetša*, *šala*, *šaletše*, *šeba*, *šedio* and *sedirišwa)* are entered as a query in the search engine, as shown in **Figure 3.3** below.

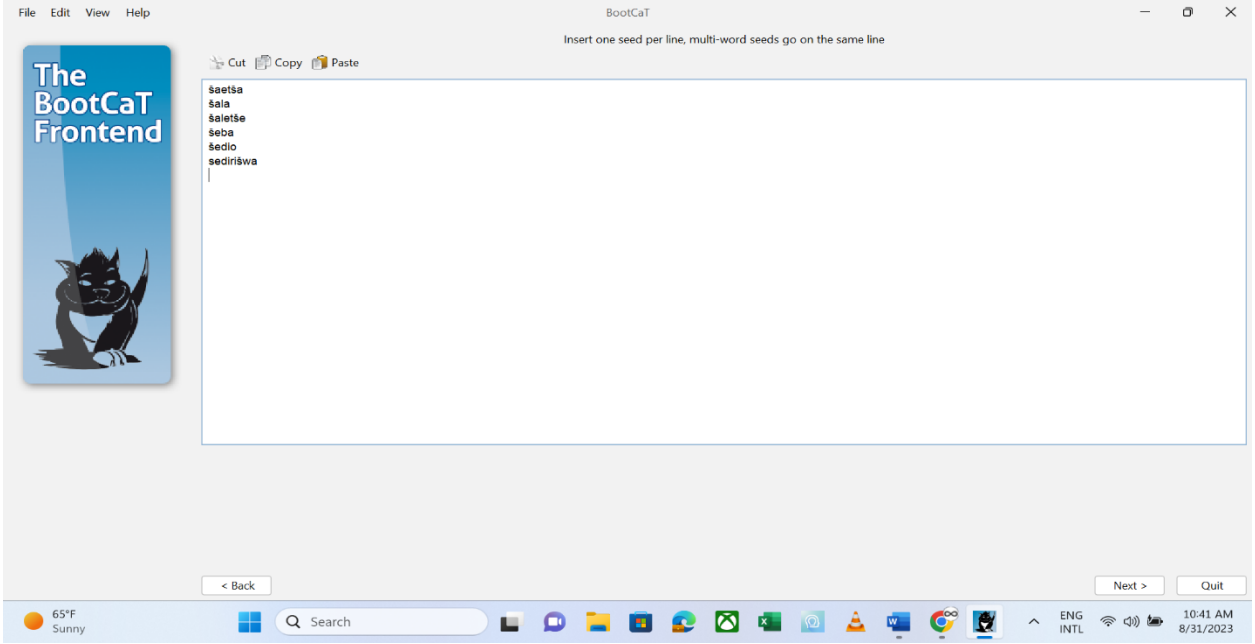


***Figure 3. 3**: BootCat screen showing six entered seeds*

The toolkit will combine these seed words in random tuples, and then query the search engine to find webpages and documents that contain the generated tuples. See **Figure 3.4** below for this.

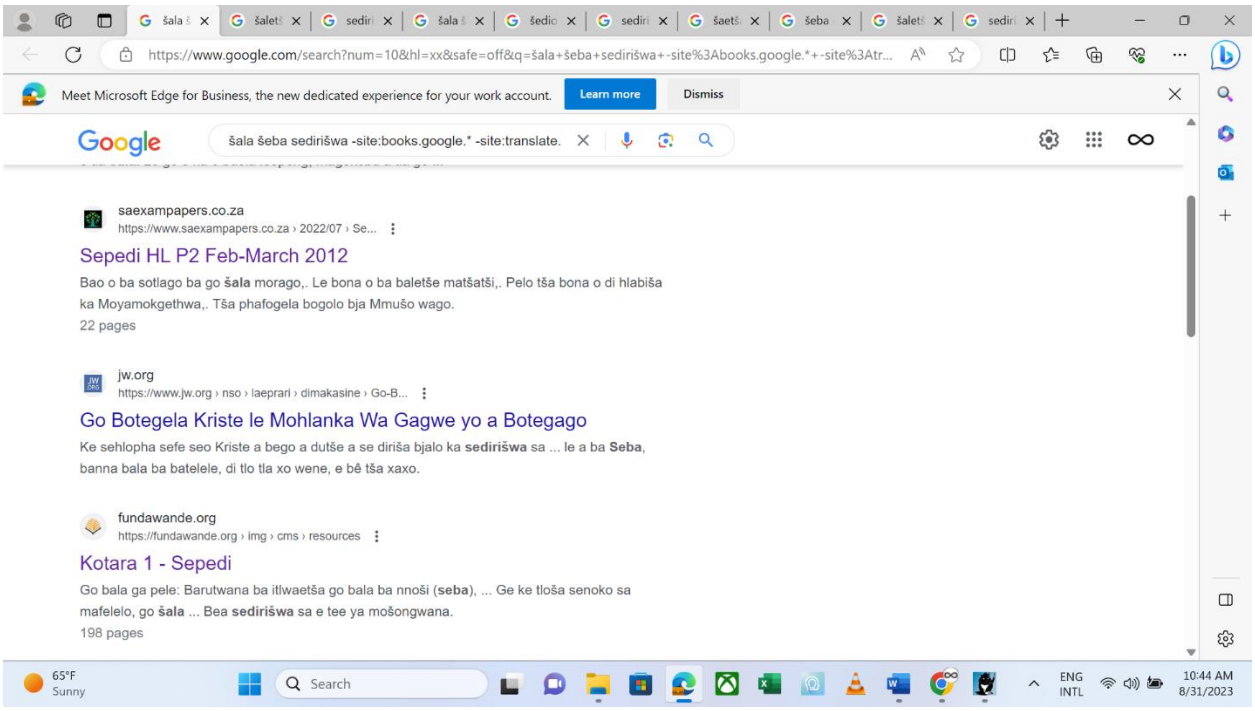


***Figure 3. 4**: Webpages and documents that contain the generated tuples*

In the **Figure 3.4** above, search hit pages are retrieved, which then allows the user to download all the relevant documents. The user can then copy and paste texts into the Word document from downloaded files (see **Figure 3.5**).

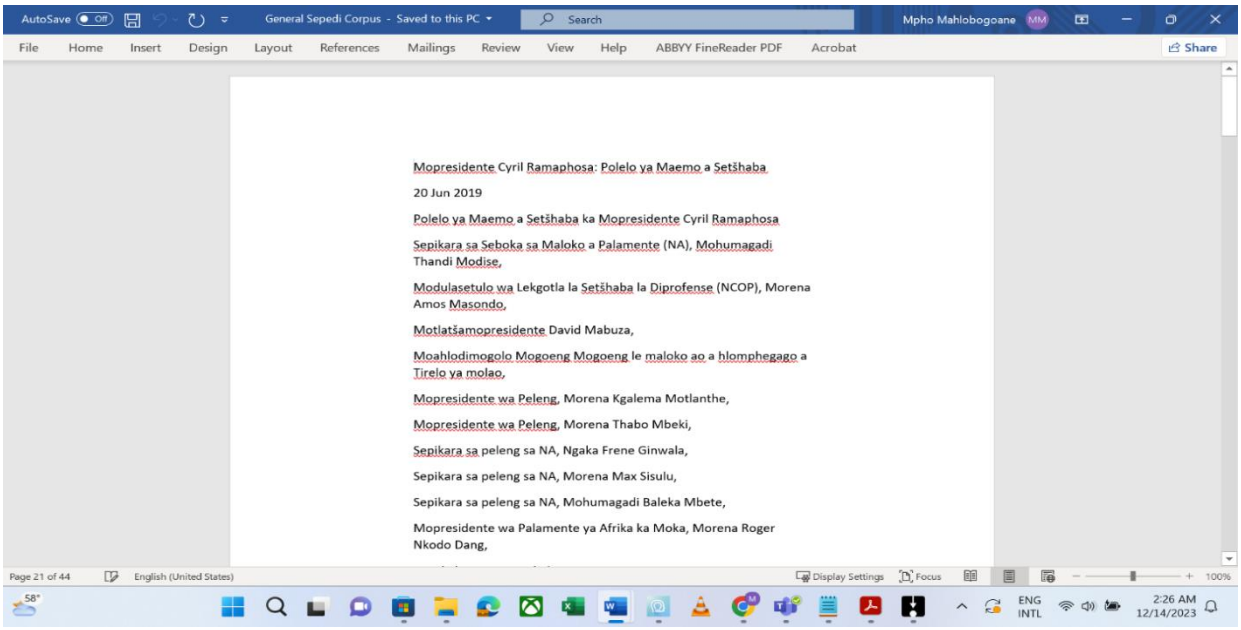


***Figure 3. 5**: GSC in Word format*

The texts in word format were then converted into plain texts format (see **Figure 3.6**). The files were then ready to be uploaded into LancsBox X. It was then possible to upload the files into LancsBox X.

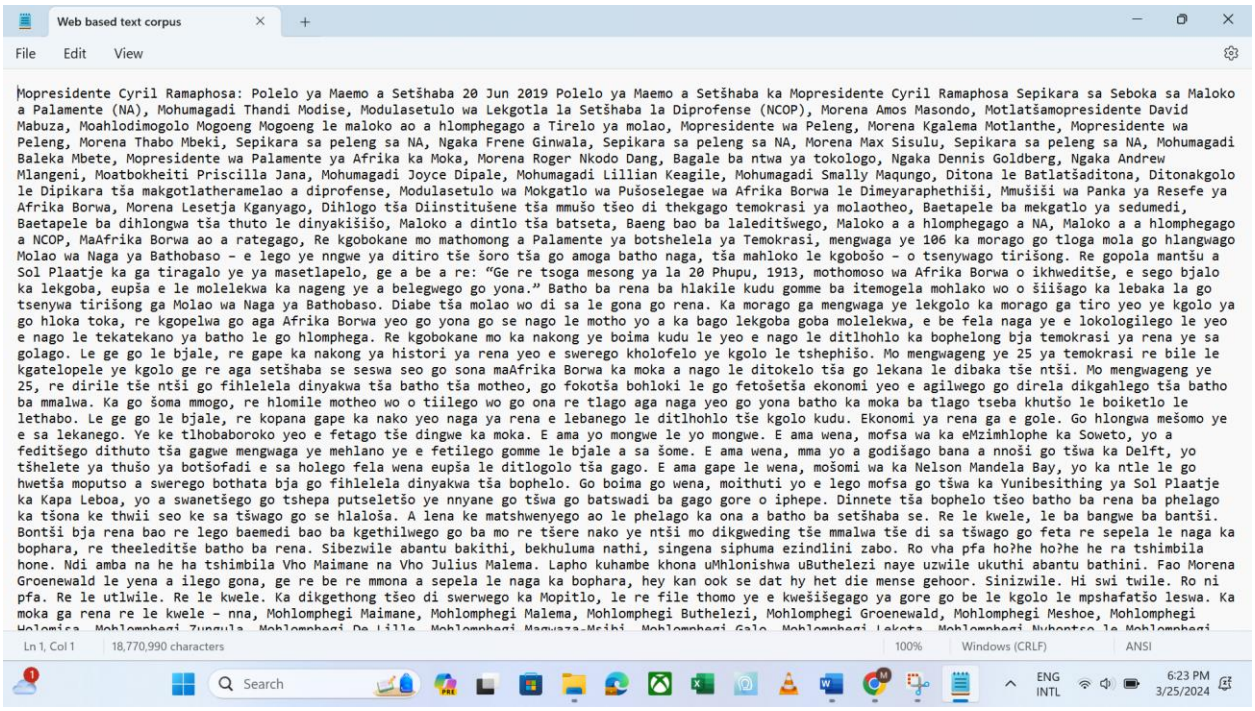


***Figure 3. 6**: The original text in plain text format*

In comparison with English, Sepedi's relevant documents are scarce. Therefore, the corpus that was created from webpages was supplemented with electronic plain texts received from Department of African Languages at the University of Pretoria (see **Figure 3.7**). The purpose was to create a corpus of over five million running words (see **Figure 3.8**). Corpus creators argue that 'bigger is best'. Besides, a corpus of five million running words is among the biggest corpora, especially is the South African context since the field of Corpus linguistics is still very new and undergoing developments. Therefore, the size of this corpus should be sufficient to serve the purpose of the present study.

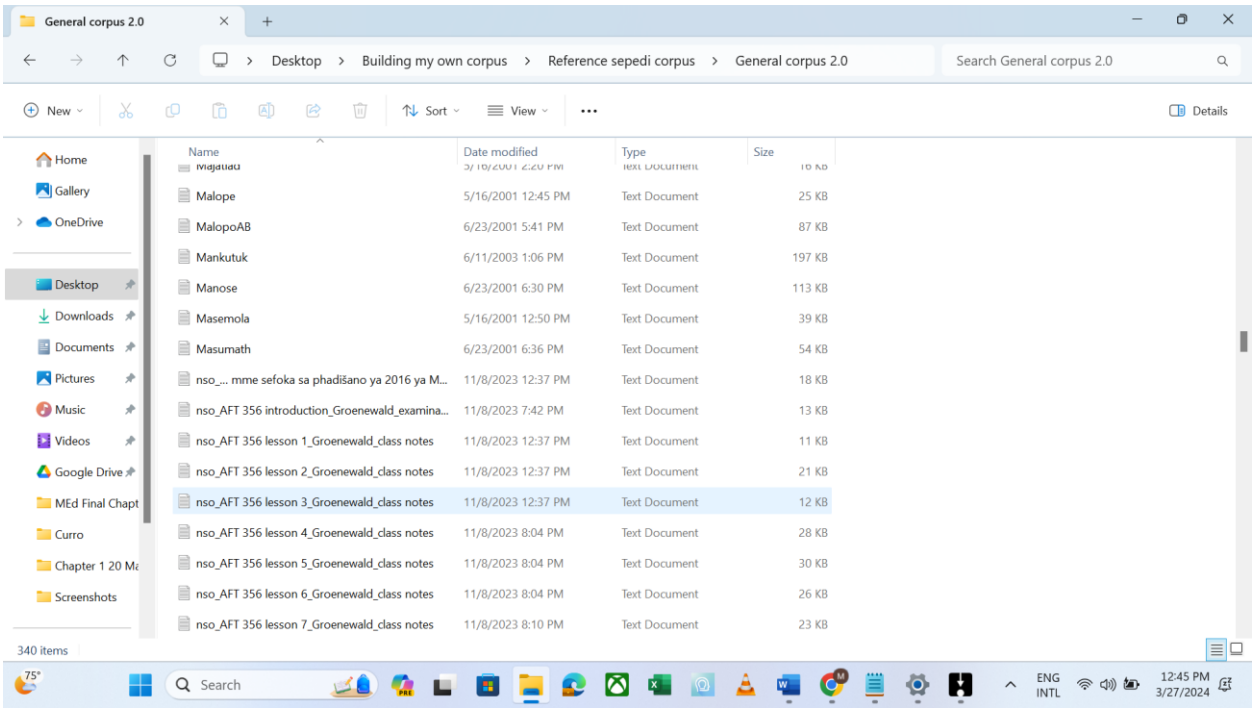


***Figure 3. 7**: General Corpus plain texts received from the Department of African Languages at the University of Pretoria*

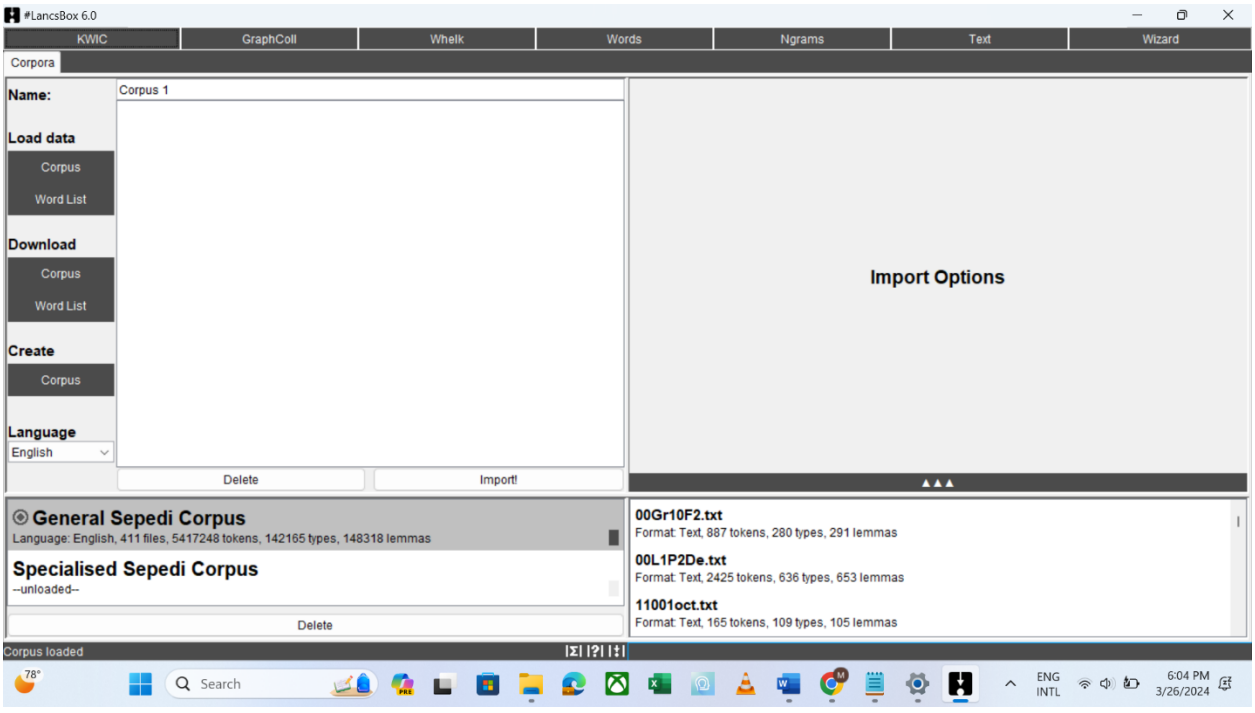


***Figure 3. 8**: LancsBox X displaying running words in a GSC at the bottom left*

### 3.12 Creating specialised Sepedi corpus

It was highlighted in the foregoing discussion that the present study uses learners' textbooks for the SSC. To acquire these textbooks electronically, a formal letter had to be sent to the Oxford University Press, the publisher of these books. The purpose of the letter (see **Addendum A**) was to seek permission from the Oxford University Press to use the books for outlining the aim and objectives of the study and the significance of the books to it.

The Oxford University Press accepted the request and provided the electronic version of the textbooks (See **Addendum B**) This process ensured that the necessary material was gathered in order to develop the SSC. Three textbooks were received in PDF format (see **Figure 3.9**).

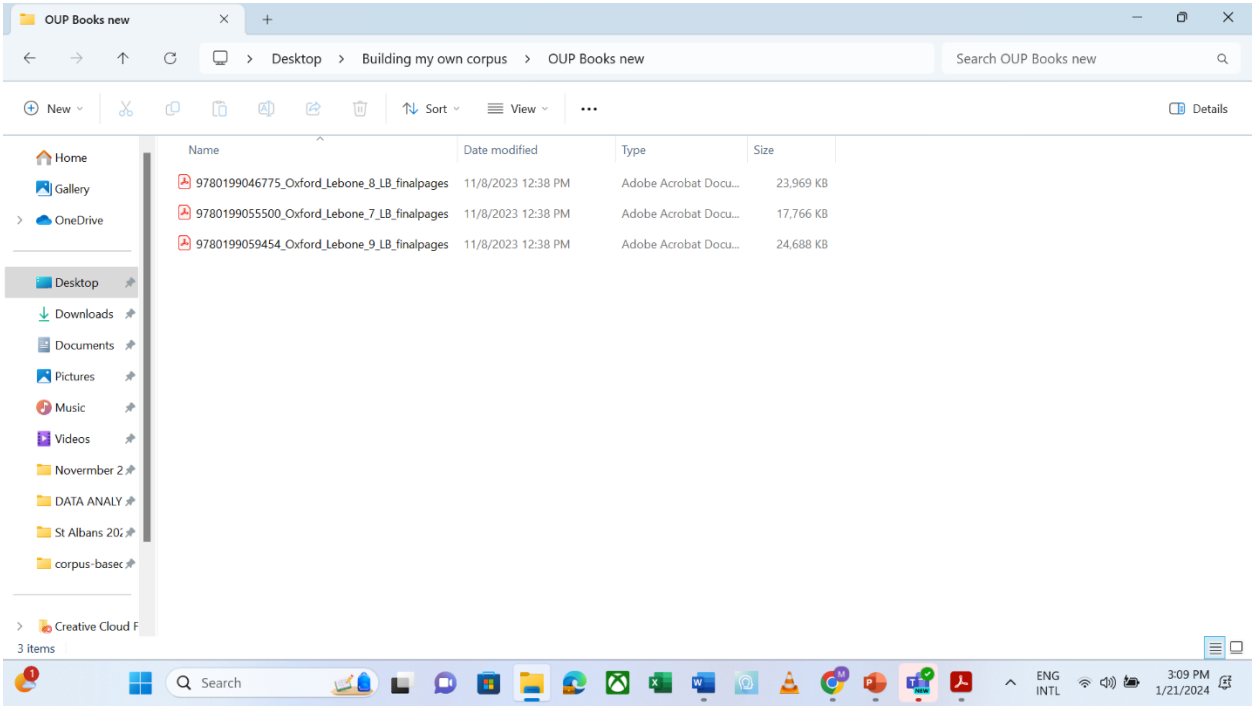


***Figure 3. 9**: Sepedi textbooks in PDF format*

It has already been indicated that LancsBox X software is adept at handling corpora in diverse formats, including, but not limited to, .txt, .xml, .doc, .docx, .pdf, .odt, .xls, .xlsx,

and .zip (Brezina *et.al*. 2015). Therefore, files for the GSC were uploaded in .txt and those for the SSC were uploaded in pdf. format.

### 3.13 Uploading the corpora onto LancsBox X

When uploading files onto LancsBox X, multiple procedures must be followed. It is implicit that the first step is to open the program. **Figure 3.10** below shows the first window that appears when you open LancsBox X. The next step is to click on 'Corpus' from the Corpora tab (see **Figure 3.11**) to upload the various files onto the program. To access the location (folder) where the corpus is saved, select this option, which will open a window (see **Figure 3.12**). Once the corpus folder has been found, you can either hold down Ctrl and left-click on the texts you want to upload, or you can hold down Ctrl + A to pick every file in the folder (see **Figure 3.13**). Then, to load your files, left-click 'Open'. After step, import your corpus onto LancsBox X by left-clicking 'Import' (see **Figure 3.14**). **Figure 3.15** shows the files that are loaded onto the software. The bottom panel and bottom-right part display the imported corpus. The individual text files that make up the corpus or the corpus structure can be viewed (see **Figure 3.16**). The corpus can be accessed again once the software has been closed (see **Figure 3.17**). To open the corpus again, left-double-click on the corpus on the software (reload).

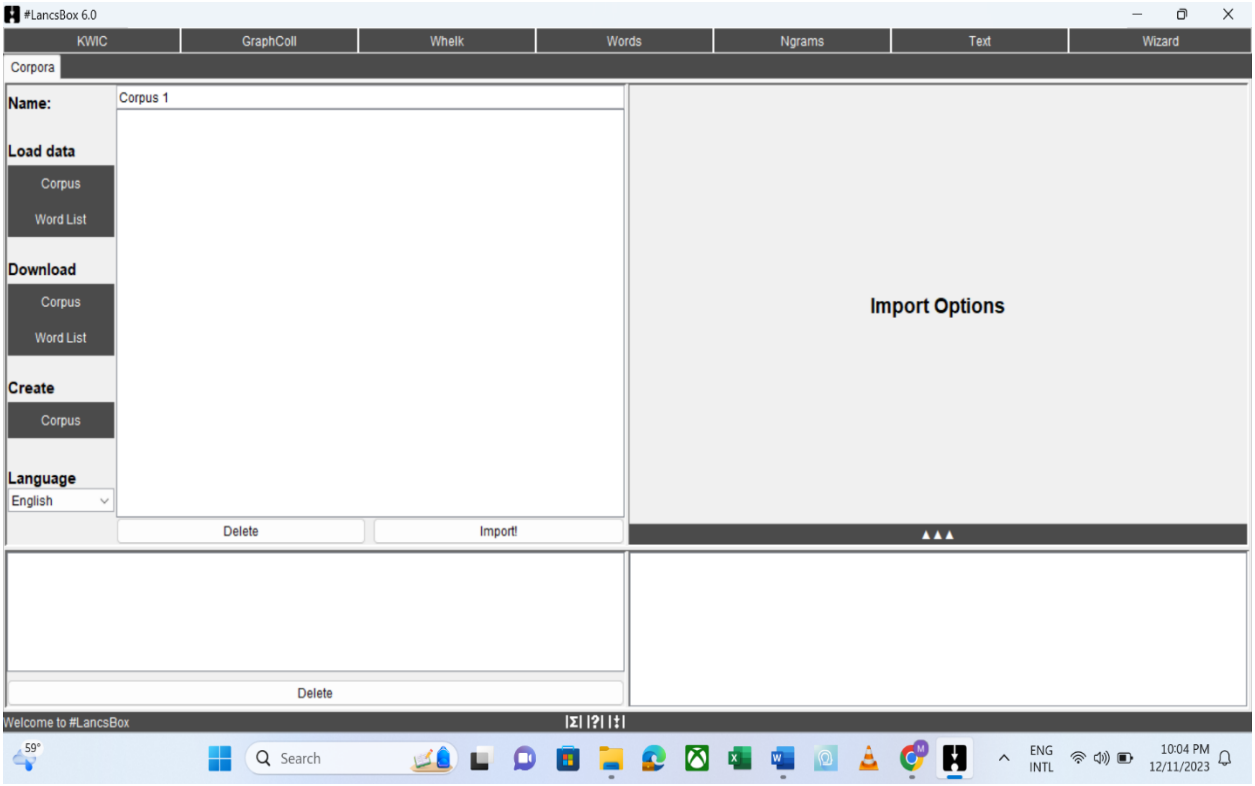


***Figure 3. 10**: The main screen of LancsBox X*

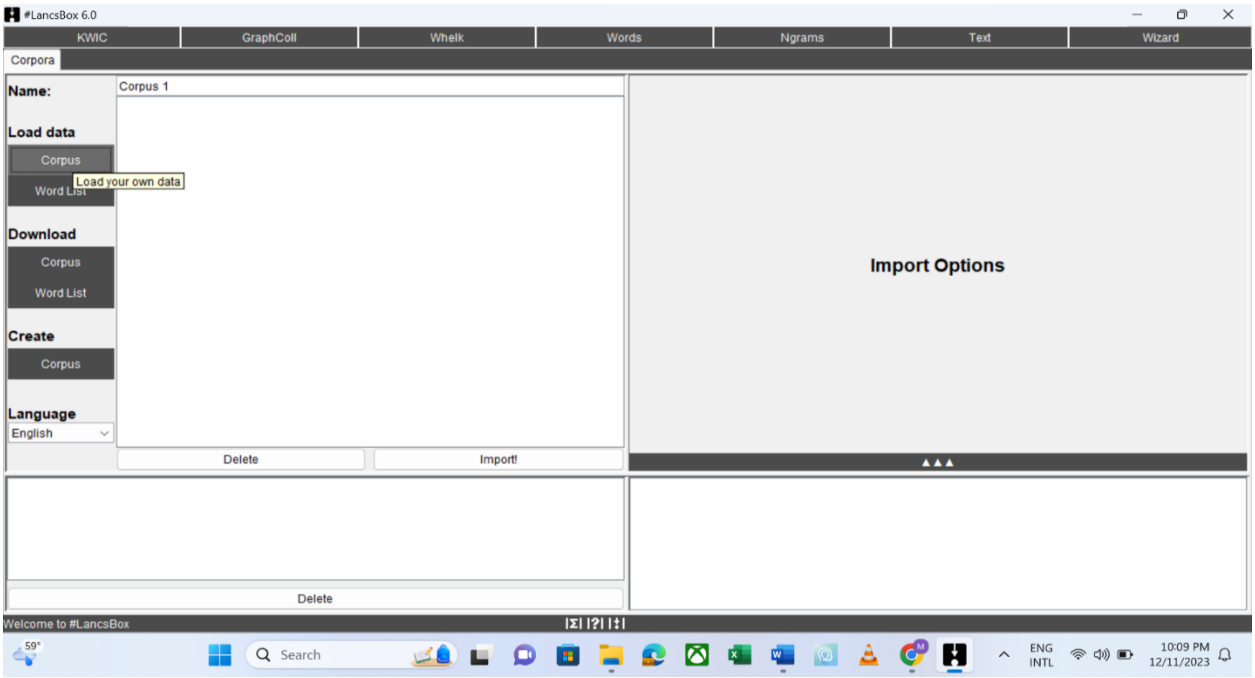


***Figure 3. 11**: LancsBox X window showing the option to upload file(s)*

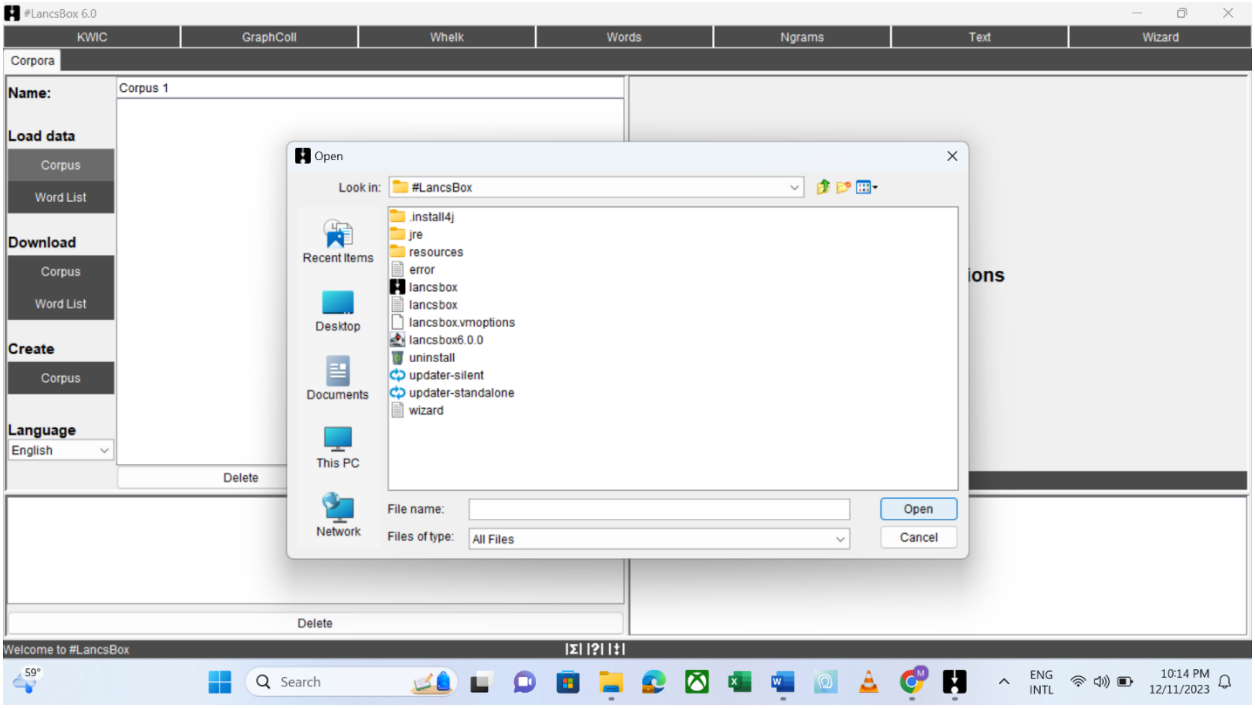


***Figure 3. 12**: LancsBox X window displaying location (folder) where the corpus is stored*

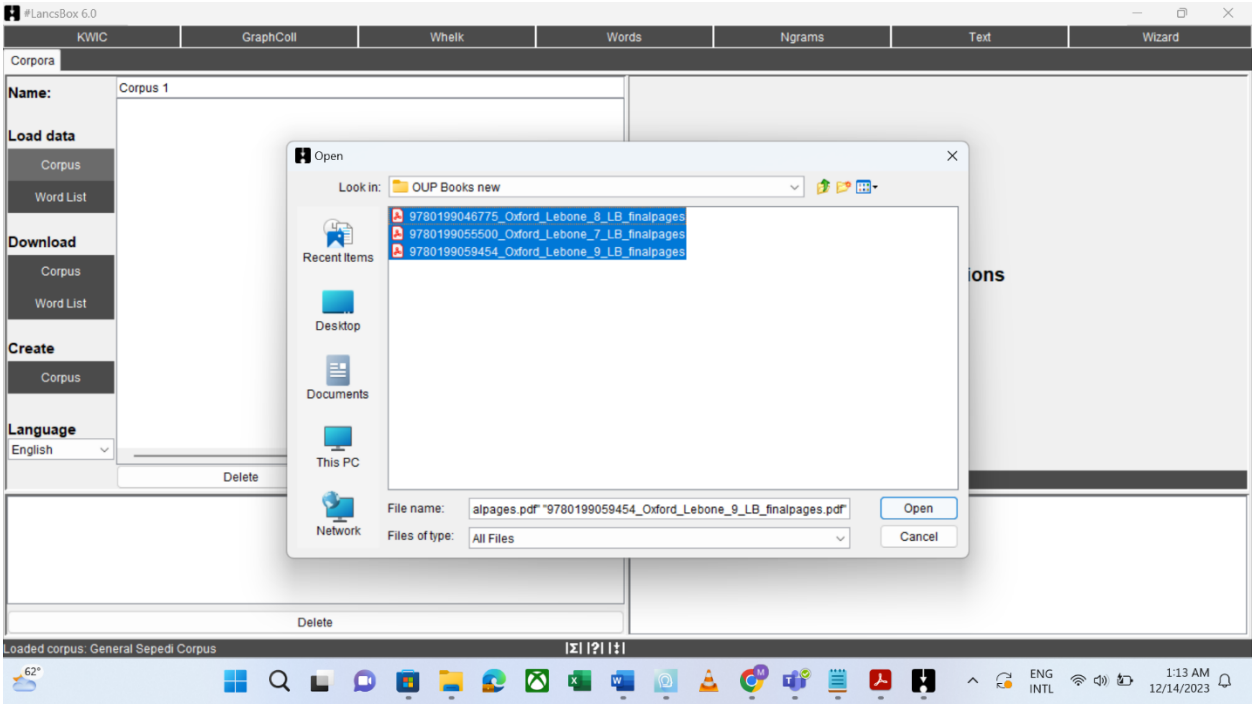


***Figure 3. 13**: LancsBox X window showing selected special corpus files in PDF format*

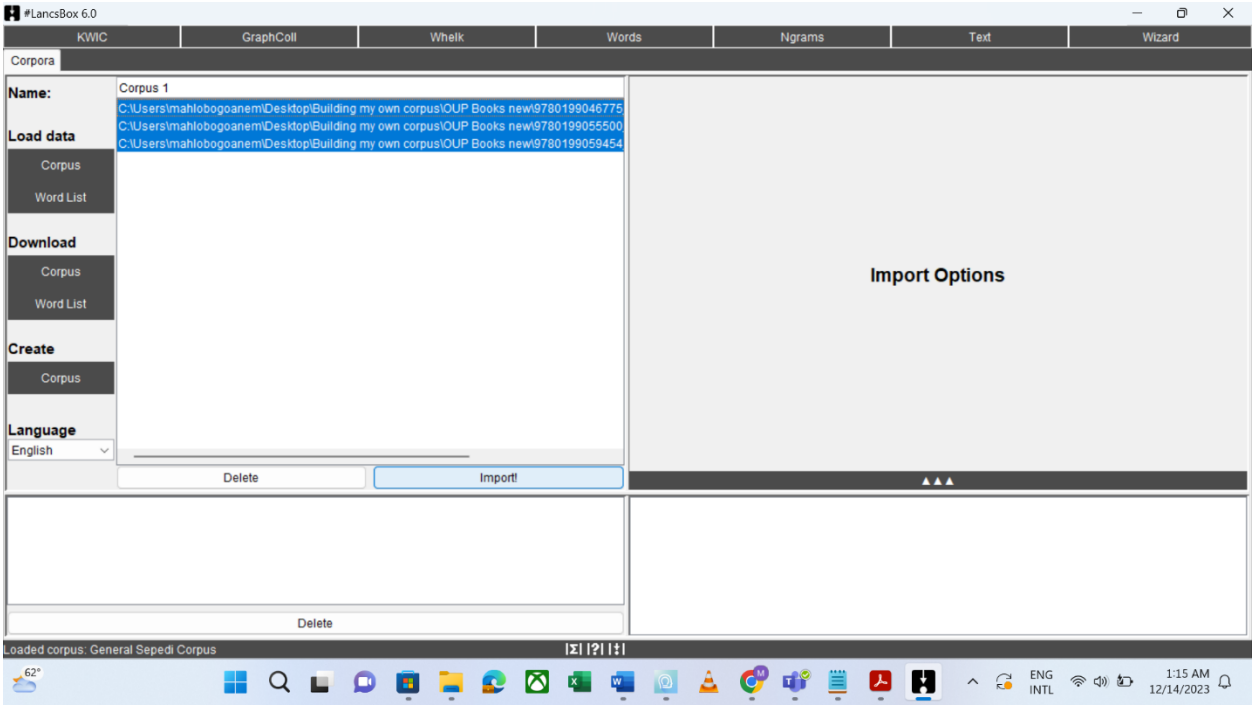


***Figure 3. 14**: LancsBox X window showing imported corpus into LancsBox X*

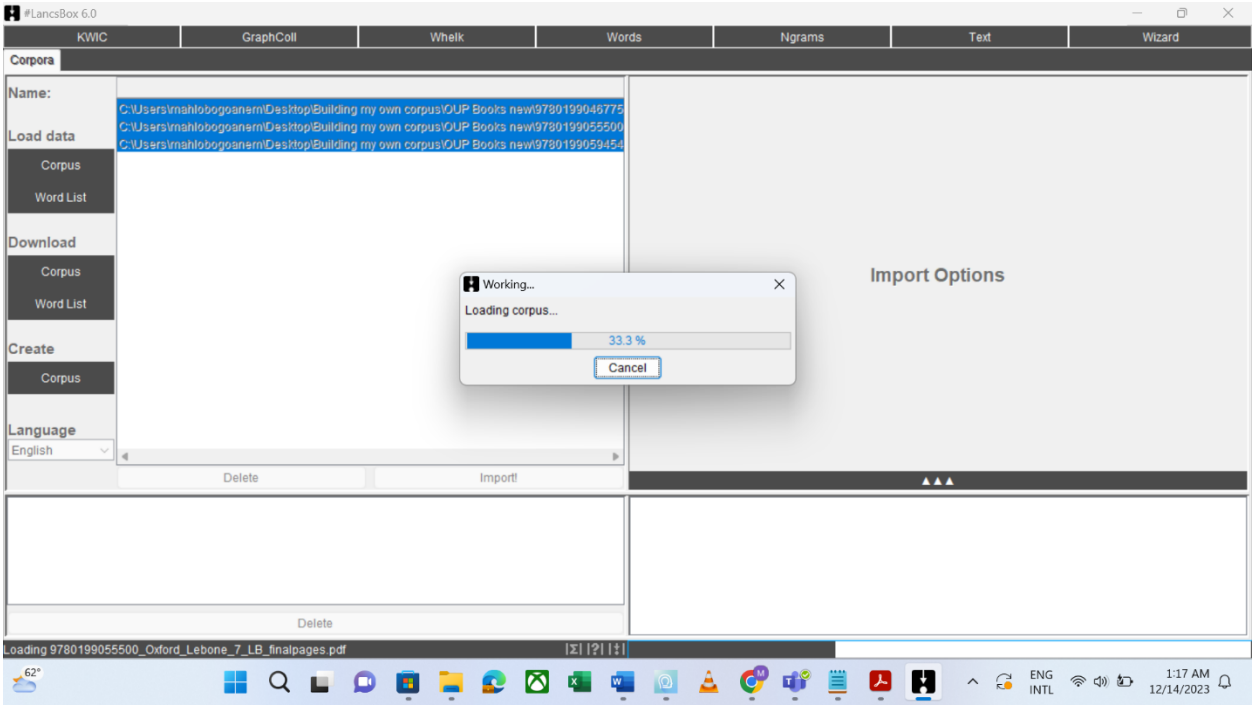


***Figure 3. 15**: LancsBox X window showing files uploaded*

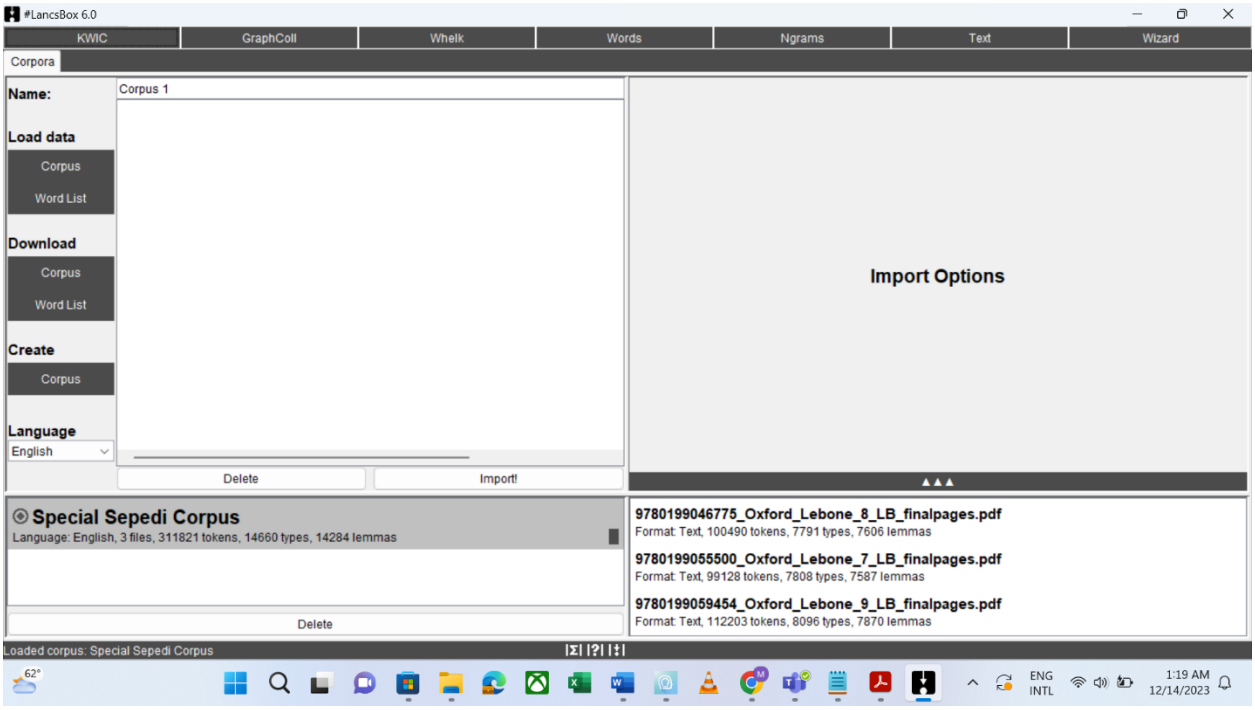


***Figure 3. 16**: LancsBox X displaying imported corpus structure on 'Corpora' tab*

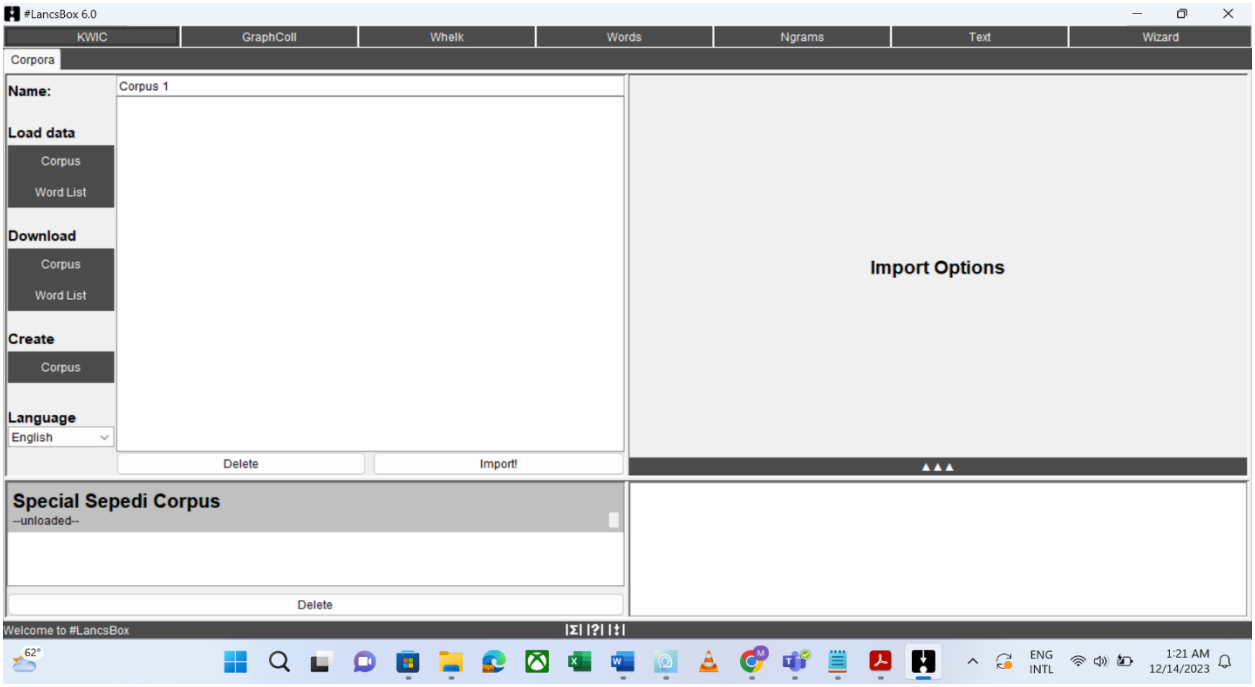


***Figure 3. 17**: LancsBox X window displaying the corpus that is unloaded after opening the software*

LancsBox X software offers researchers the flexibility to upload as many files as possible. In the current study, the process of uploading the SSC mirrors that of uploading the GSC. Therefore, the same process was followed in uploading GSC. To visualise this integration, the interface conveniently displays the two corpora at the bottom left corner of the window (see **Figure 3.18**).

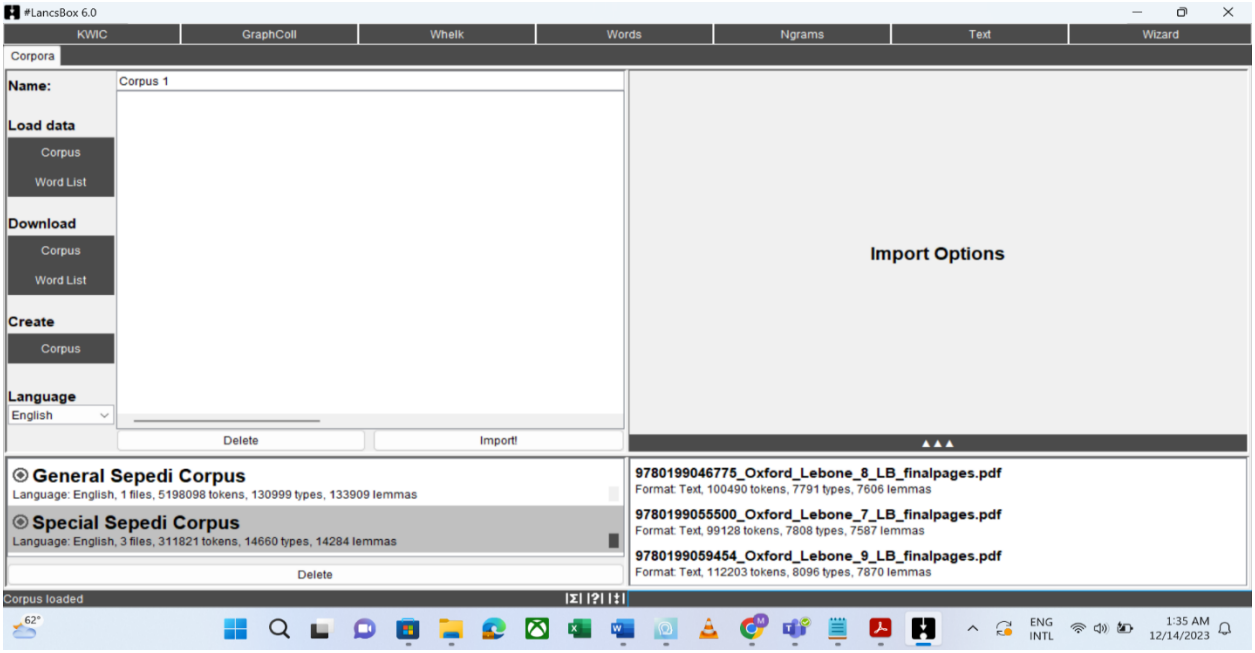


***Figure 3. 18**: LancsBox X Window displaying the two corpora at the bottom left corner of the window*

Once the corpus has been uploaded, the researcher can use the software's various corpus-query tools to start manipulating the data. In the section that follows, these tools are discussed.

## 3.14 LancsBox X tools

As previously mentioned, there is a wide range of programs available for querying corpora. The corpus manipulation tools available in the LancsBox X software include KWIC, Whelk, GraphColl, Words, Ngrams, and Text tools. Let us discuss these in detail.

### *3.14.1 KWIC*

The **KWIC (Key Word In Context)** or **Concordance** tool stands out as one of the most frequently employed and widely utilised instruments for querying a corpus. Brezina *et al*. (2015) define KWIC as a tool that is used to produce concordance lines. As Baker (1995) points out, the search term (node) is positioned in the middle of each concordance line, surrounded by terms that co-occur on its left-hand and right-hand side.

To utilise the KWIC tool on LancsBox X, click the 'KWIC' tab located at the top of the tool. You will then need to enter the desired word or phrase in the upper left-hand corner search box and select it with a left-click (refer to **Figure 3.19**).

**Figure 3.20** below shows an example of a KWIC line using the search term *'ge'*. The search term, known as the 'node', is highlighted in orange and placed in the centre. Words are shown to the left- and right-hand side. In the present study, the KWIC is essential for searching and presenting Sepedi conjunctions in context.

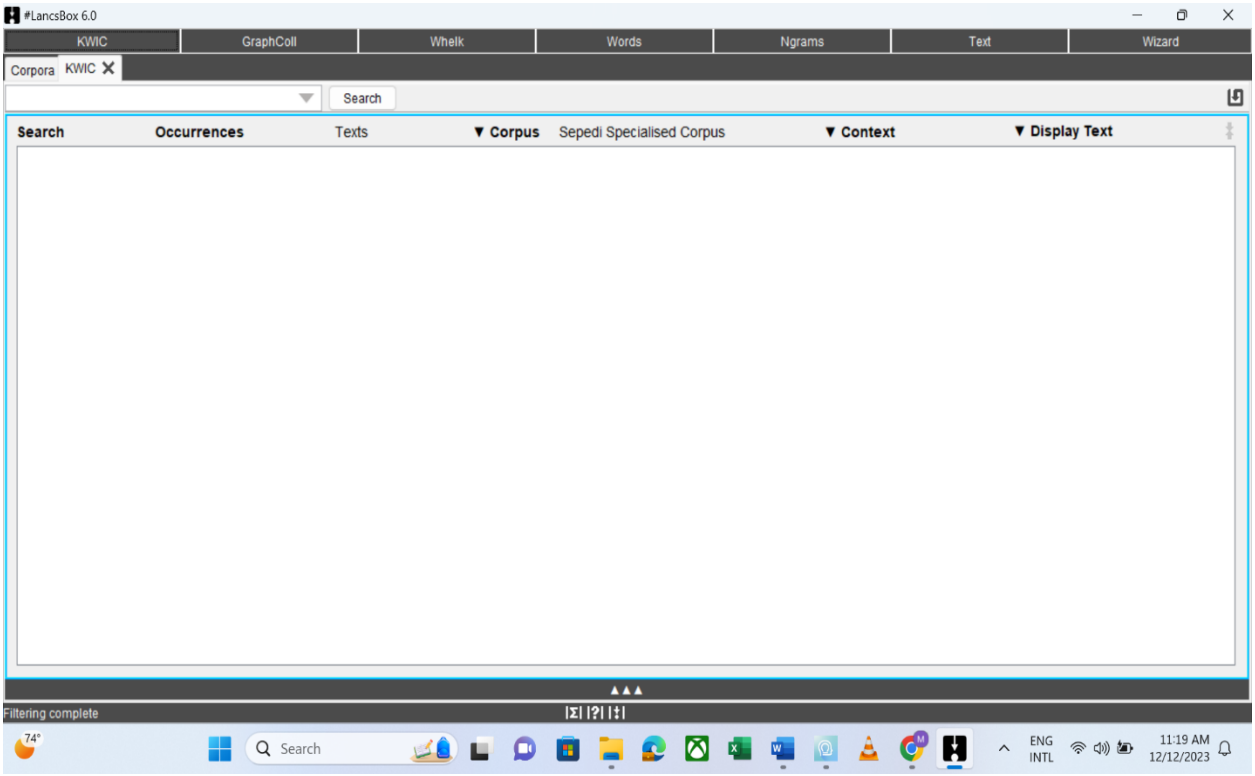


***Figure 3. 19**: KWIC tab to search for any word or phrase*

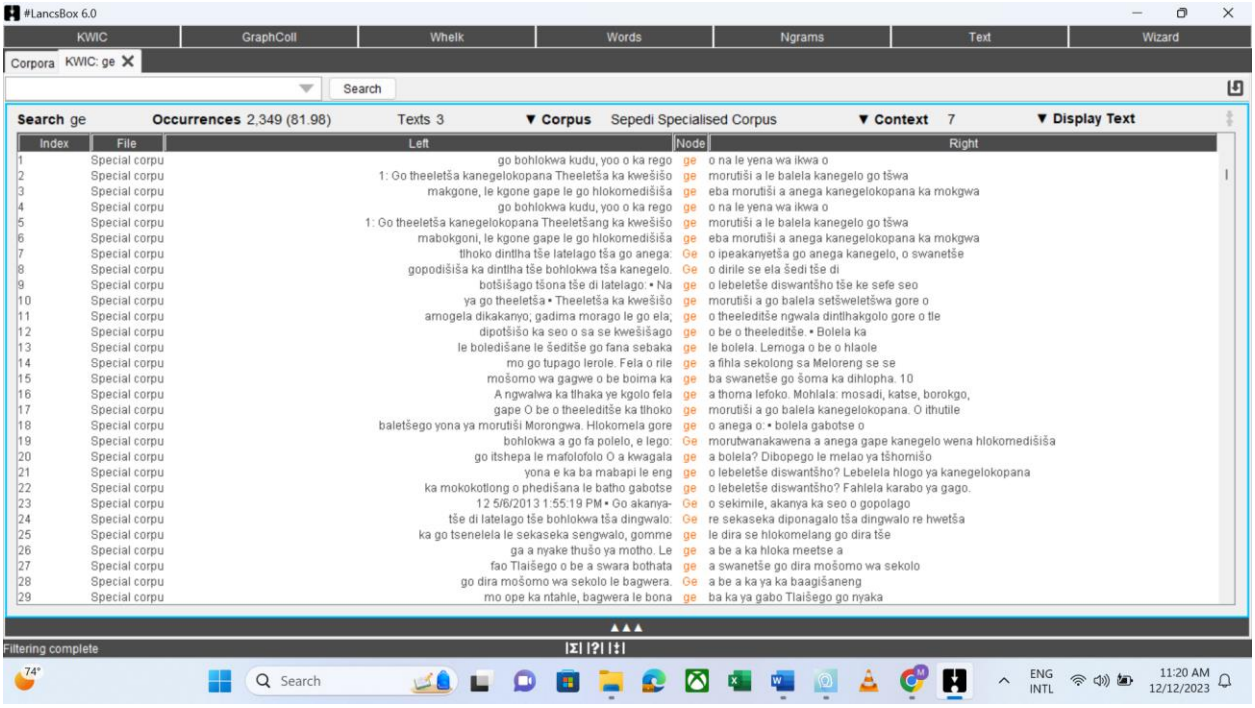


***Figure 3. 20**: Concordance lines for the word ge*

### *3.14.2 Whelk Tool*

Brezina *et.al*. (2015) define the **Whelk Tool** as a tool that provides information on how the search term is distributed across corpus files. Researchers can control the tool to discover both absolute and relative frequencies of the search term within the corpus files, aiding a nuanced understanding of its prevalence. Furthermore, the Whelk tool provides the capability to filter results based on various criteria, enhancing the precision of the analysis by allowing researchers to modify their inquiries. Additionally, the tool facilitates the organisation of files by enabling sorting based on the absolute and relative frequencies of the search term. This multifaceted functionality enhances the researchers' ability to extract meaningful insights from the corpus data (Brezina *et al*. 2015).

In order to access the Whelk tool on LancsBox X, one needs to click on the 'Whelk' tab at the top of the tool. One can then type in the word or phrase of interest in the search box in the top left-hand corner and left-click 'Search' (see **Figure 3.21**).

An example of how the search term is distributed across corpus files is provided in **Figure 3.22** below, with '*ge*' as the search word. The Whelk tool plays a pivotal role in the present study in displaying the distribution of Sepedi conjunctions across SSC files, offering insightful features for full analysis.

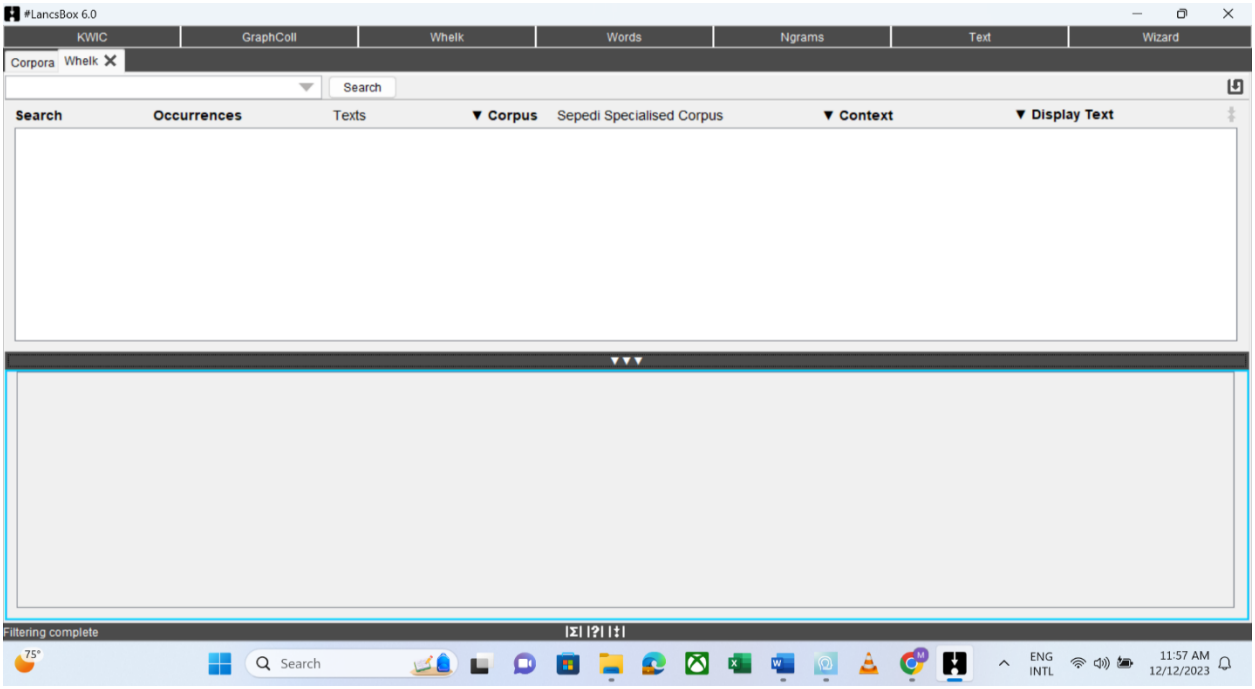


***Figure 3. 21**: Whelk tab to search for any word or phrase*


***Figure 3. 22**: Frequency distribution of ge in separate files in specialised corpus*

In **Figure 3.22** above, the Whelk tool provides detailed information on the distribution of the search term '*ge*' in the Specialised Sepedi Corpus. The '**File**' column lists the names of the individual files in the corpus. The '**Tokens'** column provides the information on the size of each file in running words (tokens). The '**Frequency**' column provides absolute frequencies of the search term *ge*, i.e., how many instances the search term *ge* appears in each file. The '**Relative frequency per 10k**' provides relative frequency normalised to the basis of 10,000 tokens; this value is comparable across files and corpora (Brezina *et al*. 2015).

### *3.14.3 GraphColl Tool*

Brezina *et al*. (2015) define **GraphColl** as a tool that identifies collocations and displays them in a table and as a collocation graph or network. Its usage by the researcher's ranges from tasks such as discovering the collocates of a given term, to identifying colligations involving the co-occurrence of grammatical categories, and visualising collocations and colligations for enhanced comprehension. Additionally, the tool enables the identification

of shared collocates among different words or phrases, contributing to a more comprehensive analysis (Brezina *et al*. 2015).

In order to access the GraphColl tool on LancsBox X, one needs to click on the 'GraphColl' tab at the top of the tool. One can then type in the word or phrase of interest in the search box in the top left-hand corner and left-click 'Search' (see **Figure 3.23**).

An example of how collocations are identified and displayed in table and as a collocate graph is provided in **Figure 3.24** below, with *ge* as the search word. The GraphColl tool is not directly applicable to the nature of the study, as well as its aim and objectives. It is not part of the study to look at collocations and colligations (i.e., co-occurrence of grammatical categories) of searched Sepedi conjunctions.

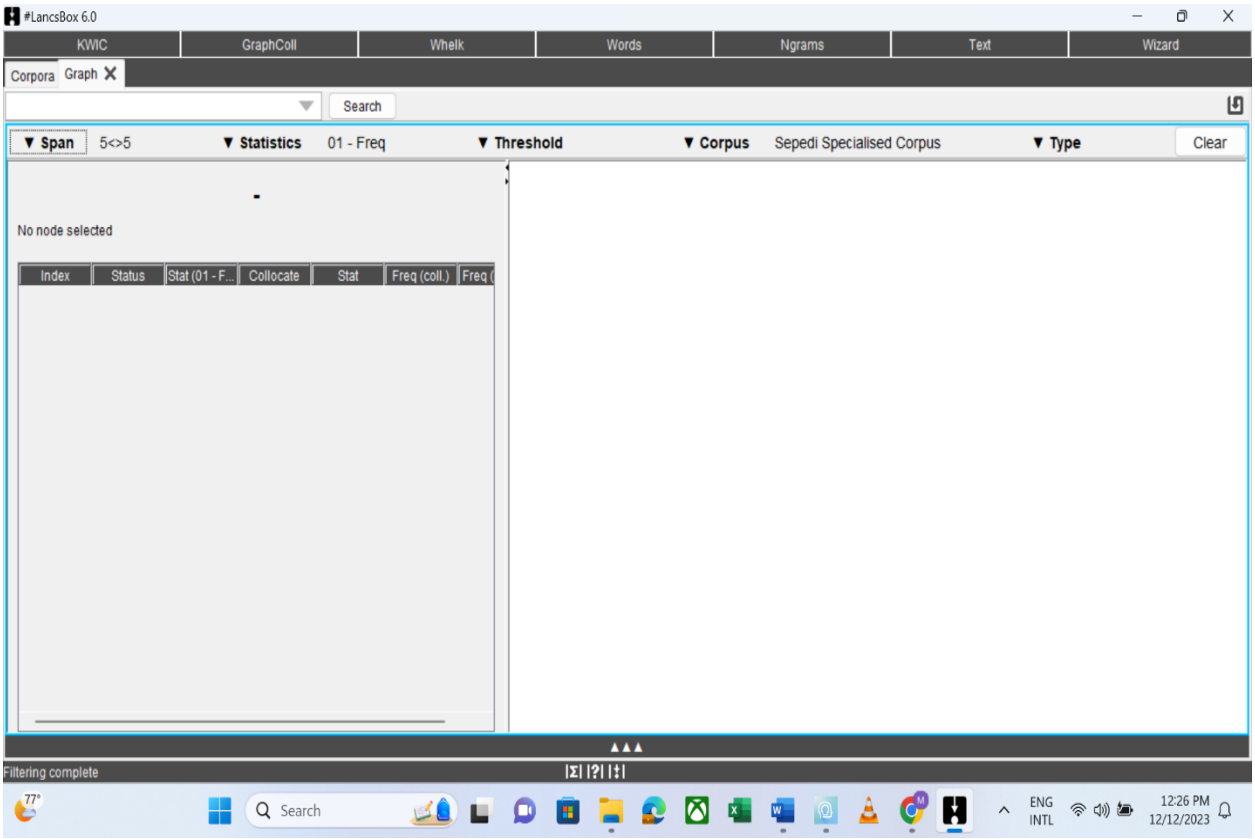


***Figure 3. 23**: GraphColl tab to search for any word or phrase*

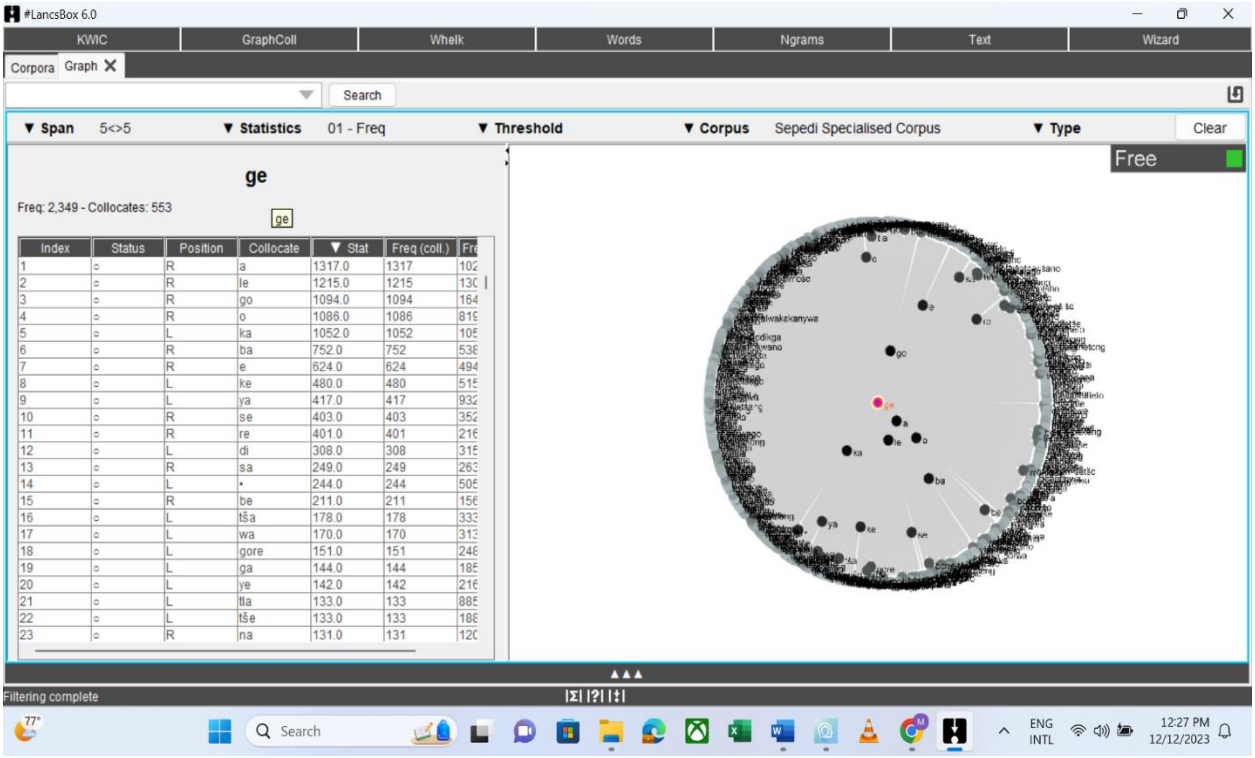


***Figure 3. 24**: Collocations in a table and a collocation graph or network*

A collocation table is a traditional way of displaying collocates. In GraphColl, the table shows the following pieces of information for each collocate:

- index,
- status,
- position,
- collocation,
- stat,
- freq (coll) and
- freq (corpus).

By default, the table is sorted according to the selected collocation statistics (largest-smallest). The graph displays three dimensions:

- strength of collocation,
- collocation frequency and
- position of collocates.

To find out more about a collocate, right-click on it to obtain concordance lines (KWIC), in which the collocates co-occur with the node (Brezina *et al*. 2015)

The **strength** of collocation, as measured by the association measure, is indicated by the distance (length of line (in orange colour)) between the node and the collocates (See **Figure 3.25 and 3.26**). The closer the collocate is to the node, the stronger the association between the node and the collocate (‘magnet effect’) (Brezina *et al*. 2015).

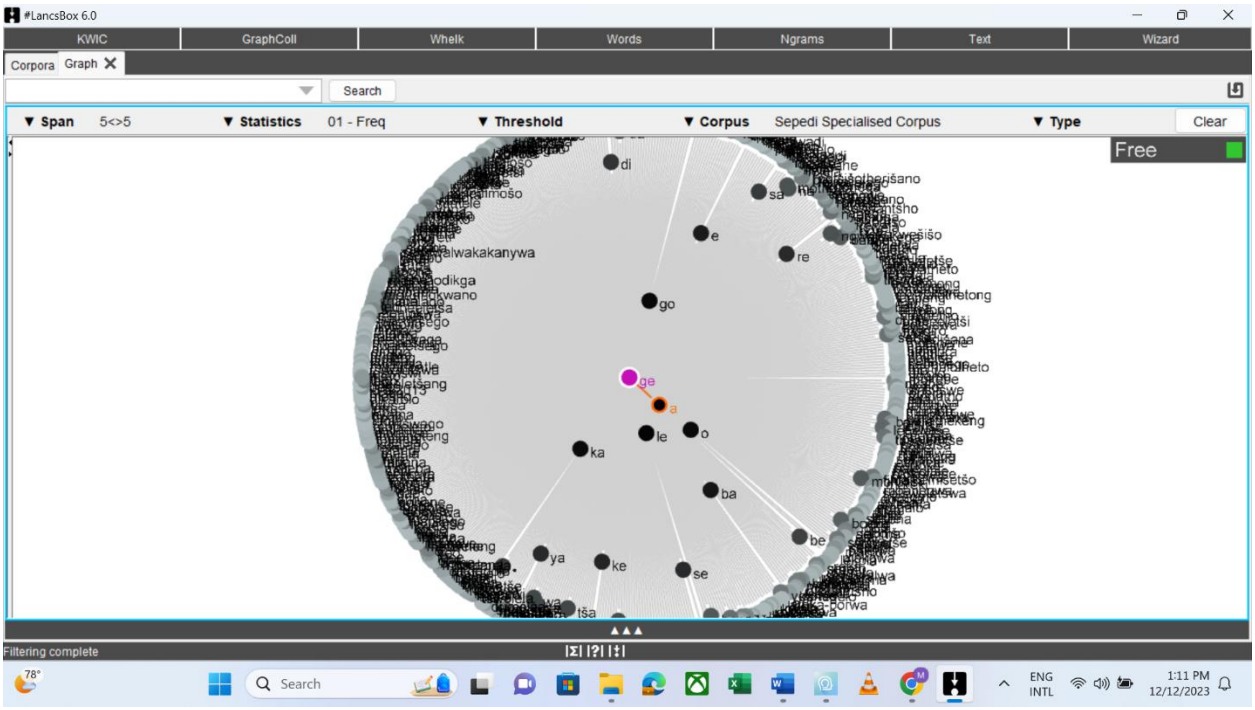


***Figure 3. 25****: The collocate ‘a’ is closer to the search node ‘ge’*

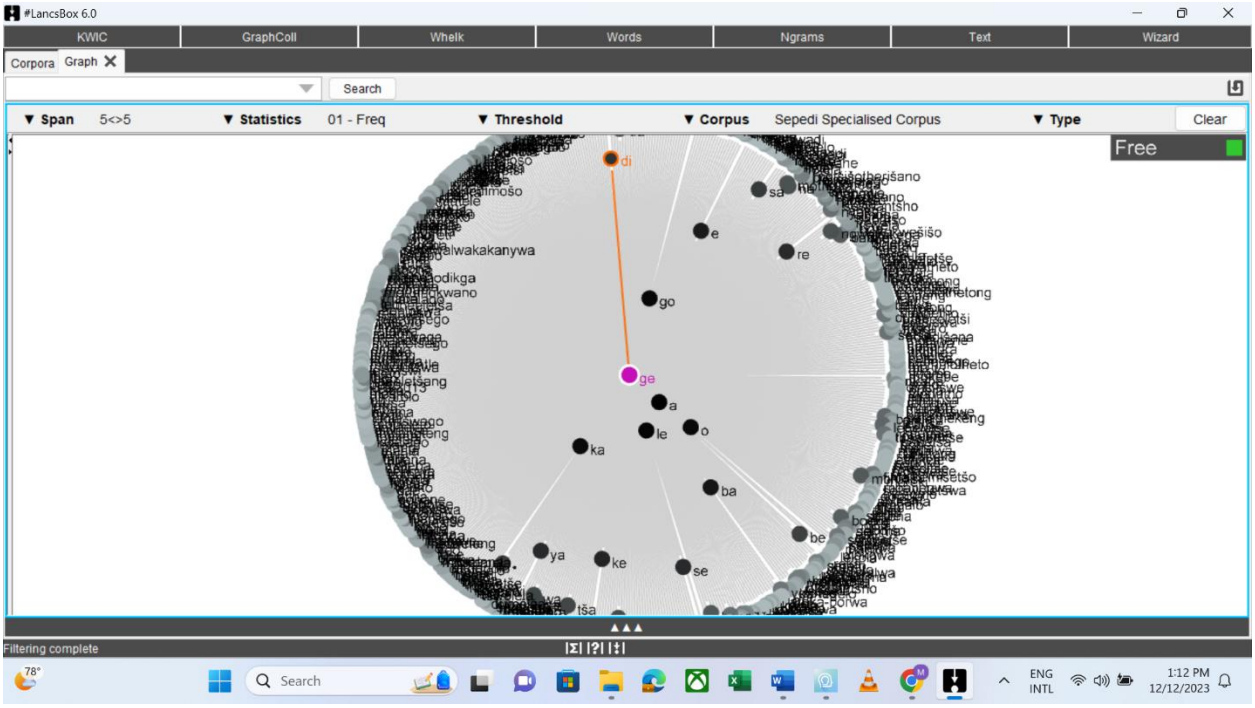


***Figure 3. 26**: The collocate 'di' is further from the search node 'ge' compared to collocate 'a'*

Collocation **frequency** is indicated by the intensity of the colour of the collocate. The darker the shade of colour, the more frequent the collocation (see **Figure 3.27** below) (Brezina *et al*. 2015).

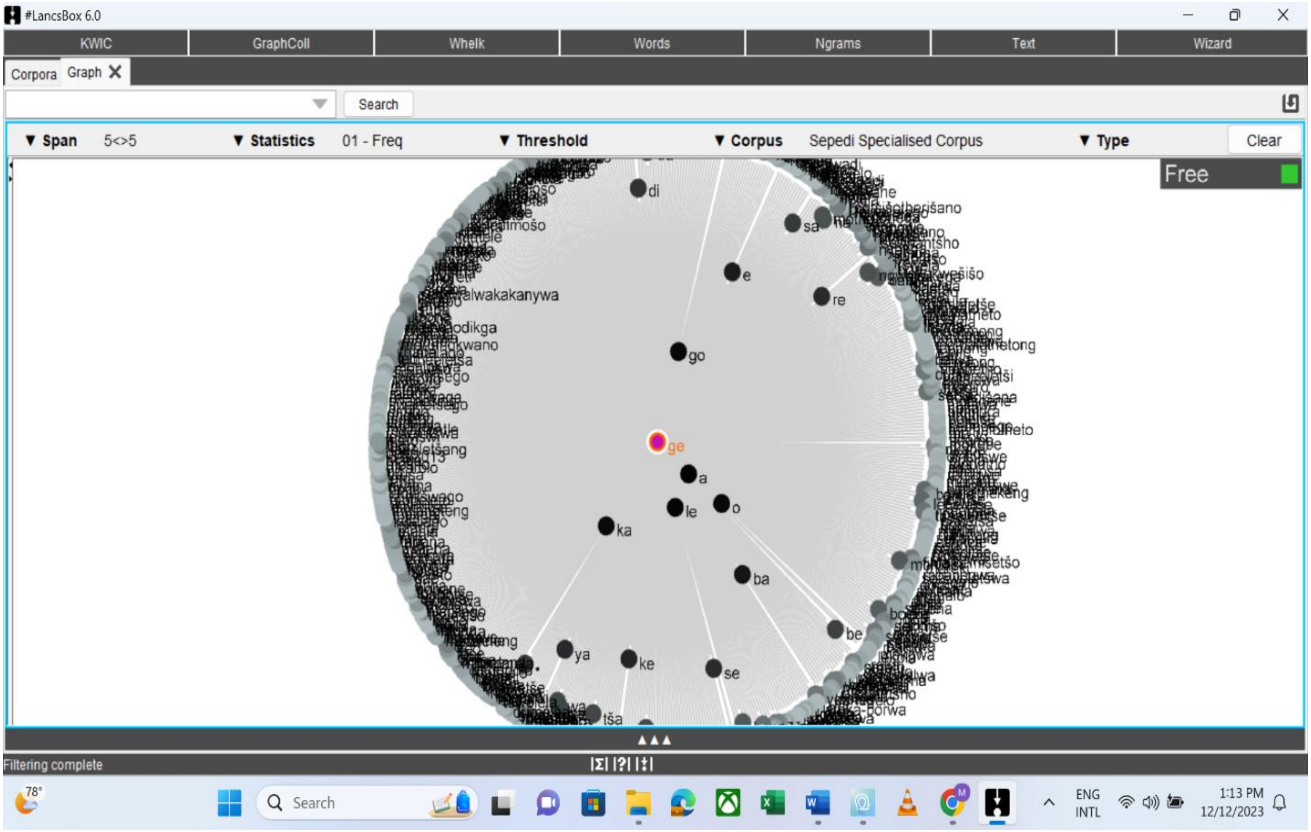


***Figure 3. 27**: Darker colour (Black) frequent collocate*

The **position** of collocates around the node in the graph reflects the exact position of the collocates in the text. Some collocates appear (predominantly) to the left of the node *ge*, others to the right; others still appear sometimes left and sometimes right (middle position in the graph). For the ease of display (if multiple collocates appear in a similar position and hence overlap), the tool allows 'spreading out' of collocates evenly around the node *ge*. This is done by clicking on the 'Spread out' button (top right). When this is done, the collocates are dispersed evenly around the node *ge* with a 'L' or 'R' index displayed above the collocate circle, indicating their original position to the left and to the right respectively (Brezina *et al*. 2015) (see **Figure 3.28** below).

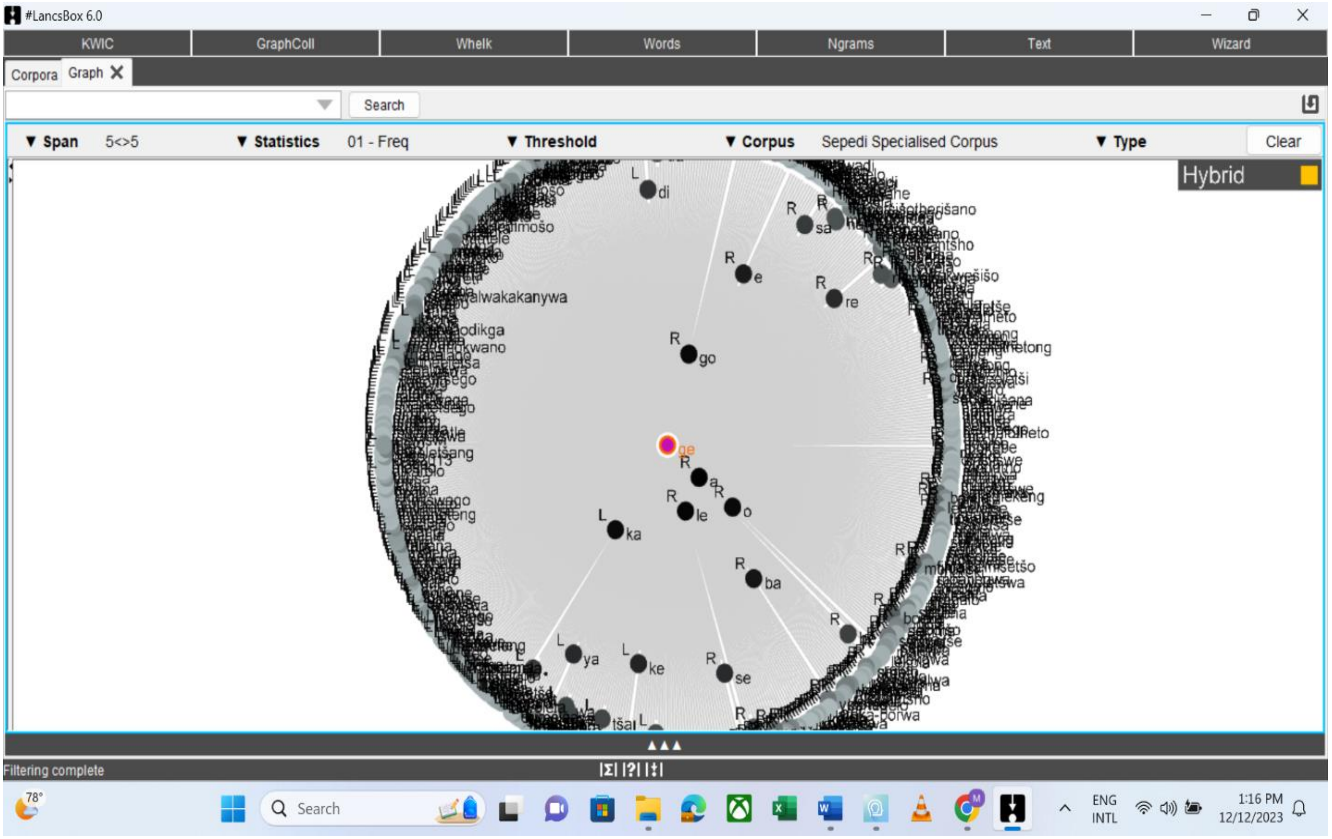


***Figure 3. 28**: Collocate positions*

### *3.14.4 Words Tool*

Brezina *et al*. (2015) define **Words Tool** as a tool used to facilitate a comprehensive examination of type, lemma, and Part-of-Speech (POS) category frequencies, enabling a detailed analysis of textual patterns. Moreover, it supports the comparison of different corpora through the application of the keyword technique. This tool offers a range of functionalities, including the visualisation of frequency and dispersion within corpora, providing researchers with a nuanced understanding of the distribution of linguistic elements. Additionally, it empowers users to conduct comparative analyses between corpora using the keyword technique, facilitating insights into variations and similarities. Furthermore, the Words Tool allows for the graphical representation of keywords, enhancing the visualisation of significant linguistic features within the analysed text (Brezina *et al*. 2015).

In order to access the Word Tool on LancsBox X, one needs to click on the 'Word' tab at the top of the tool. Upon opening, Words Tool produces a frequency list (table) based on the default corpus (see **Figure 3.29**) and default settings. These settings can be changed and a different frequency list is produced.

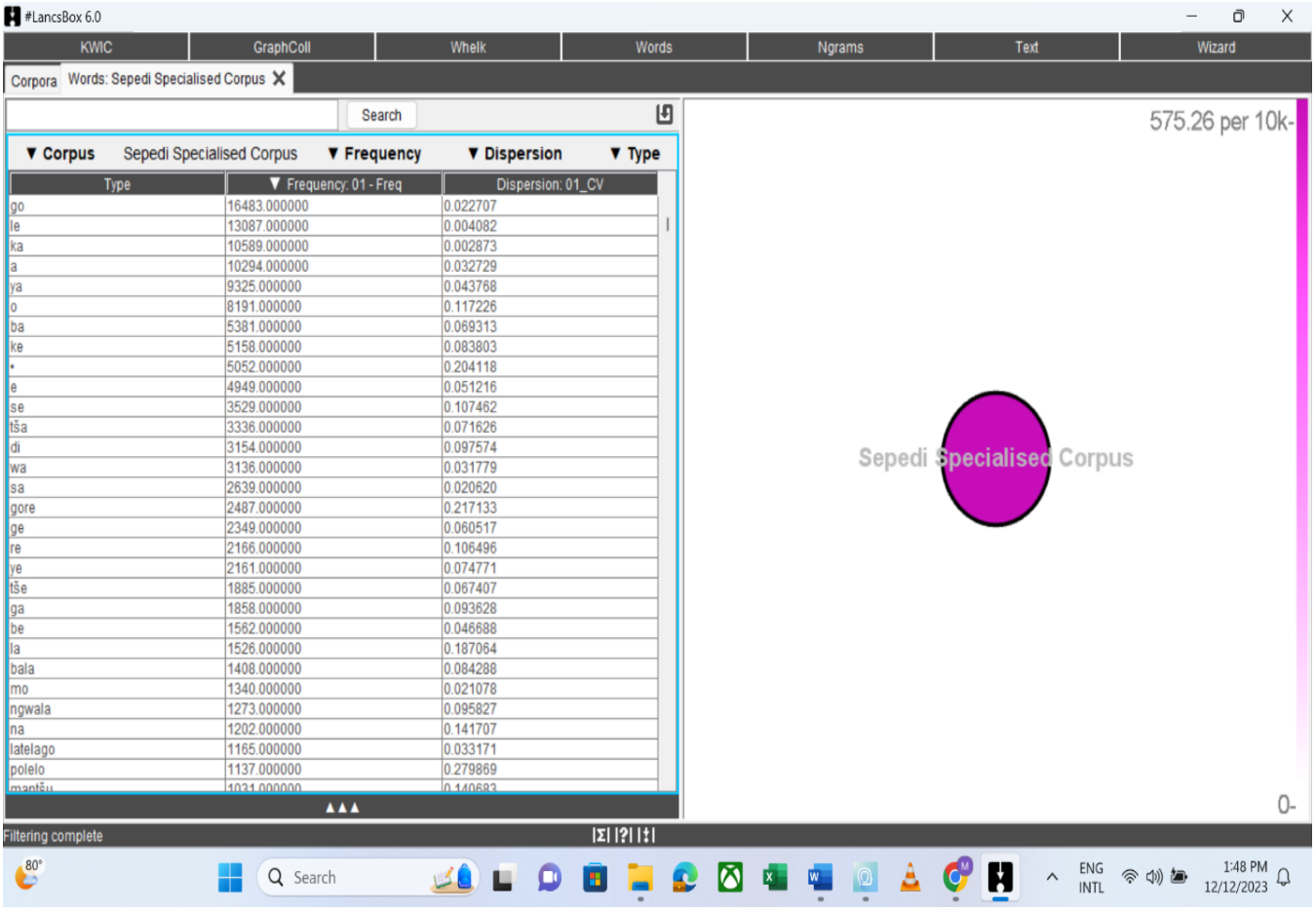


***Figure 3. 29**: Frequency list (table) based on the default corpus and default settings*

The Words Tool displays corpora and corpus files (when a corpus is left-double clicked). It visualises frequency and dispersion of words using intensity of colour and position of individual files displayed as circles. The size of the circle indicates the relative size of the corpus/file (see **Figure 3.30**).

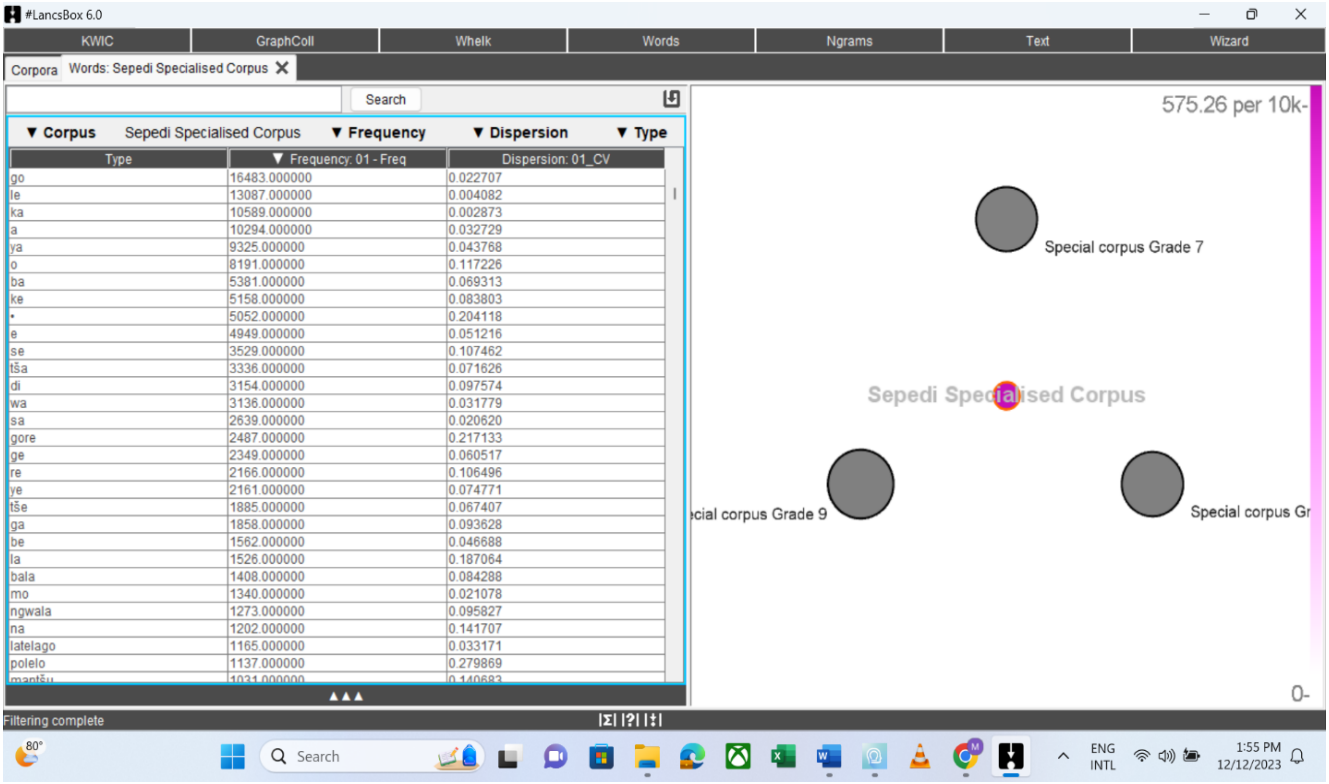


***Figure 3. 30**: Display of corpus when left-double-clicked.*

Left-click on any item in the frequency table to see its frequency visually. The corpus's colour tone will vary based on how frequently this item appears. An interpretive point of

reference is provided by the scale on the right. For this, see **Figure 3.31** below.

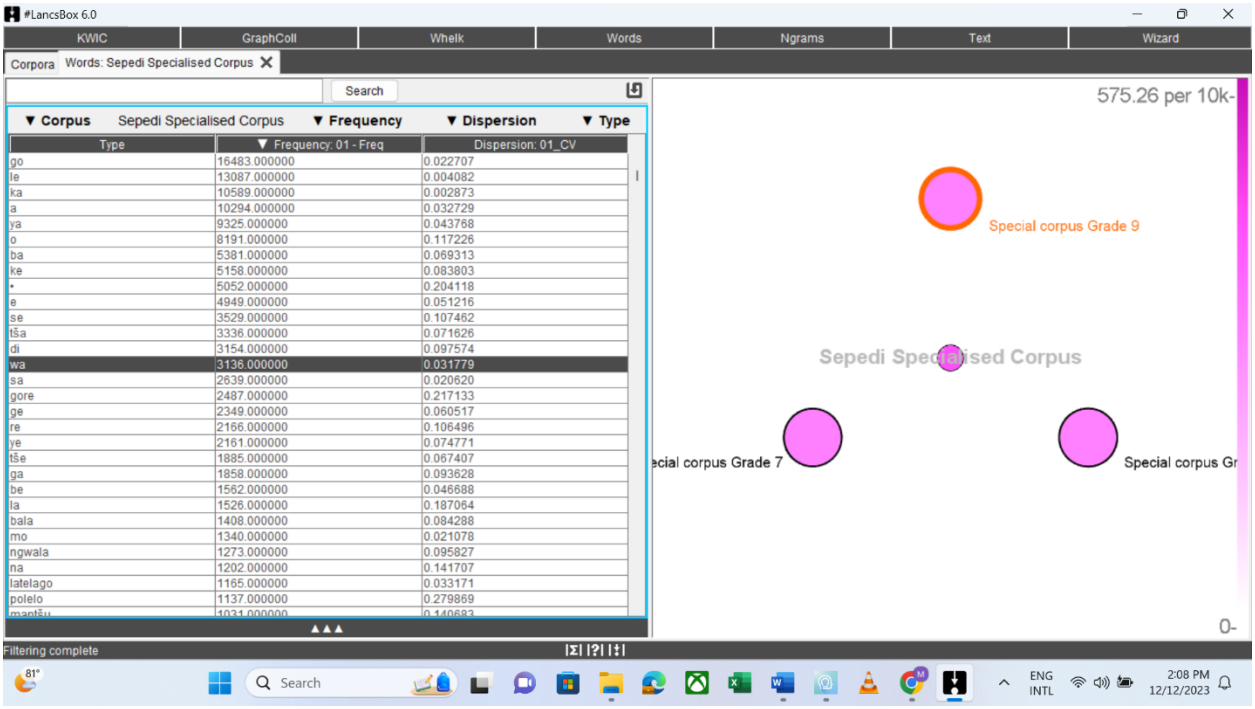


***Figure 3. 31**: Visualisation of frequency of an item in the table*

Furthermore, the Words Tool computes essential corpus statistics:

- Complexity stats and
- Lexical stats.

This is done by right-clicking on corpus, then on the pop-up table toggle between Complexity stats and Lexical stats. See **Figures 3.32** and **3.33** below for this.

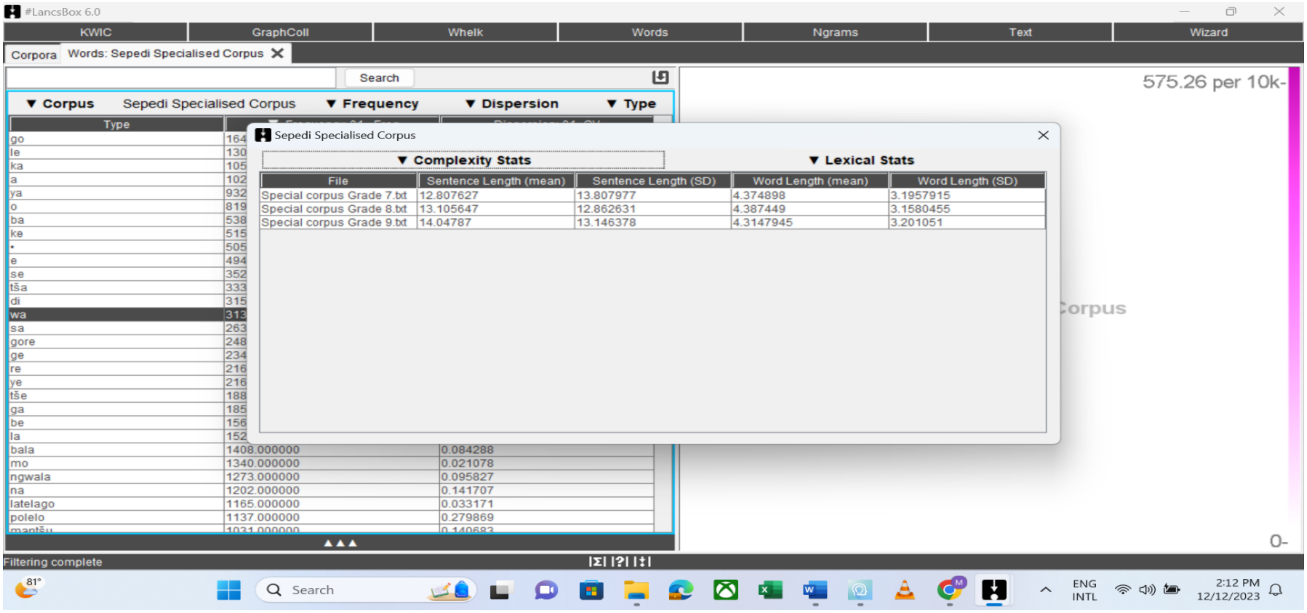


***Figure 3. 32**: Complexity stats*

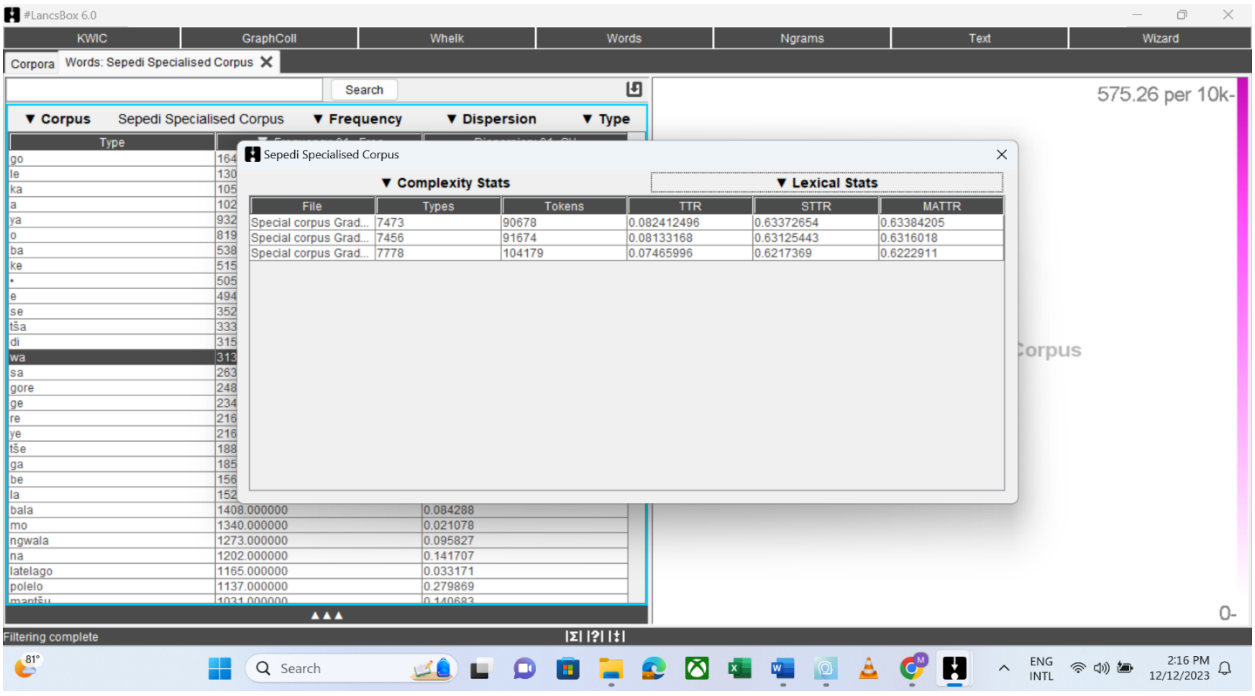


***Figure 3. 33**: Lexical stats*

Since the KWIC tool offers an option to see the frequency of the search item (node) in a corpus, the Word Tool is not employed in the present study. It would simply be redundant using different tools for the same purpose, unless if another offers something different.

### *3.14.5 Ngram Tool*

As delineated by Brezina *et al*. (2015), the **Ngram Tool** provides a detailed examination of the frequencies of n-grams, encompassing bigrams, trigrams, and so forth. N-grams are characterised as contiguous combinations of types, lemmas, and POS categories within the text. This tool enables researchers to delve into the intricate patterns of these linguistic sequences. In addition, the Ngram tool employs a technique akin to keywords, allowing for the identification and comparison of key n-grams between two corpora. This comparative approach enhances the tools utility in uncovering significant linguistic patterns and variations across different datasets.

In order to access the Ngram tool on LancsBox X, one needs to click on the ‘Ngram’ tab at the top of the tool. Upon opening, Ngram produces a frequency list (table) based on the default corpus (see **Figure 3.34**) and default settings.

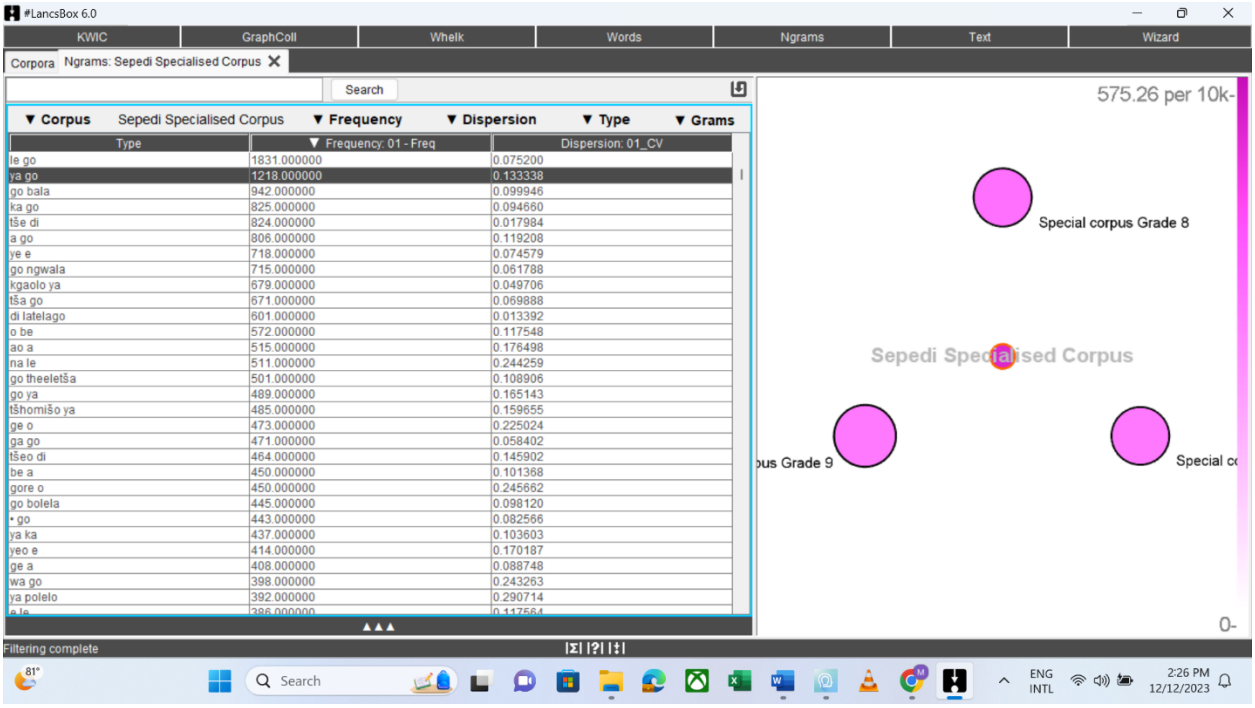


***Figure 3. 34**: Creating frequency list, computing dispersion and key Ngrams*

In the present study, there is no need to compare the Ngrams in two corpora using keyword technique. Therefore, this tool is not directly applicable to the nature of the study.

### *3.14.6 Text Tool*

As expounded by Brezina *et al*. (2015), the **Text Tool** offers a profound exploration of the contextual usage of a word or phrase, providing researchers with a nuanced understanding of its surrounding linguistic environment. This versatile tool facilitates various functionalities, including the ability to scrutinise a search term within its complete context, affording a comprehensive view of its usage. In addition, users can preview individual texts or entire corpora as run-on texts, promoting a seamless and continuous examination of linguistic content. Furthermore, the Text Tool supports the verification of different levels of annotation within a text or corpus, contributing to a detailed analysis of linguistic features and structures. In overall, it serves as a valuable resource for researchers seeking to unravel the intricate nuances of language use in diverse contexts.

In order to access the Text Tool on LancsBox X, one needs to click on the ‘Text’ tab at the top of the tool (see **Figure 3.35** below).

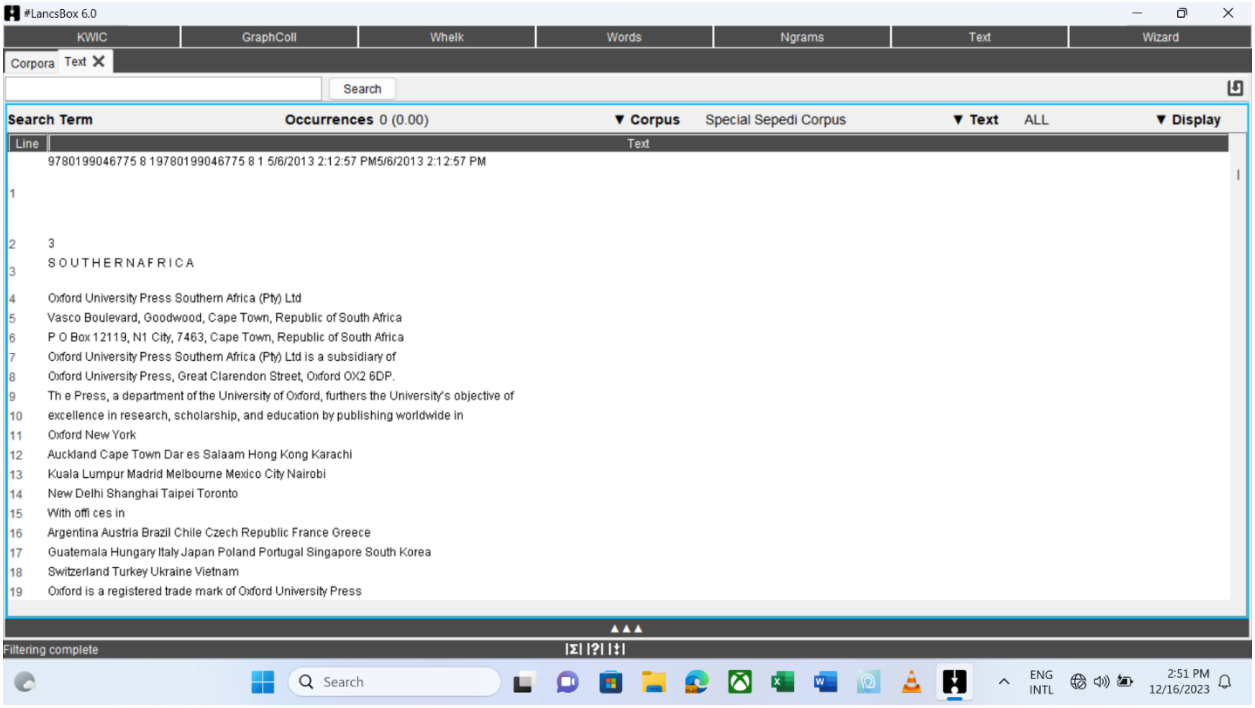


***Figure 3. 35**: Text tab to search for any word or phrase*

The user can then type the search term *ge* into the search box (top left) and left-click ‘Search’. This will highlight all lines in the text where the search term ‘*ge*’ appears in dark grey, with the search term itself in orange. Frequency information (both an absolute and relative frequency per 10 000 tokens) will appear under ‘Occurrences’. A single line can be highlighted by left-clicking on the line. To highlight multiple lines, Ctrl (Command) + Left-click the desired lines.  Refer to **Figure 3.36** below in order to view this.

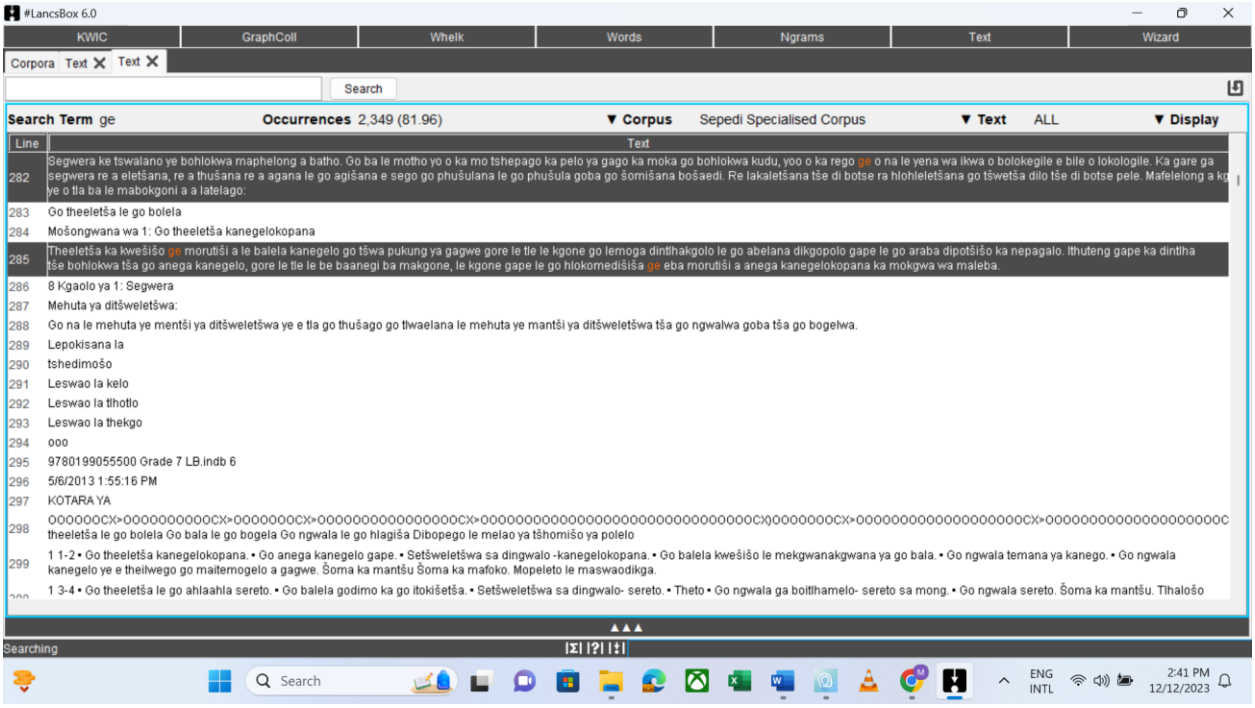


***Figure 3. 36**: Searched ge term in full contexts*

Similarly, KIWC tool offers search word in contexts, and therefore it would simply be redundant using different tools for the same purpose. Therefore, Text Tool will not be employed in the present study.

### 3.15 Conclusion

The research methods used for this study are covered in this chapter. It was stated that the study used both qualitative and quantitative methodologies for data analysis, thus adopting the triangulation approach. It was mentioned that whereas a corpus-based approach neither accepts nor rejects intuition, an intuition-based approach largely ignores corpus data. The use of a corpus-based approach for data analysis and interpretation in the present study was also emphasised. In addition, a definition of a corpus was provided and various corpus types were explained. Furthermore, the issues of balance, representativeness and corpus size when designing a corpus were explicated. The chapter also delved into different software applications for querying a corpus. The selected software for this study is LancsBox X. The chapter proceeded to outlining the

process of creating a general as well as specialised corpus for the present study. The process of uploading the corpus onto the selected software was then expounded. Finally, the chapter detailed the corpus query functionalities offered by LancsBox X, which are essential for data analysis in this study.

The aim of this chapter was to present the research methodology employed in the present study. The following chapter presents analysis and interpretation of data for the present research.

# CHAPTER 4: DATA ANALYSIS AND INTERPRETATION

## 4.1 Introduction

The current study comprises of a small, SSC collected from Sepedi learners' textbooks and a larger GSC compiled from general internet texts and supplemented with text received from the Department of African Languages at the University of Pretoria. For the purposes of this research, it was imperative that the Sepedi learners' textbooks are utilised, specifically the 'Oxford *Lebone* learners textbooks for Grade 7, 8 and 9'. These textbooks are authored by: Bapela, Mphela and Ratshivhambela (2013); Koshane, Mpe and Mphela (2013) and Makhalemele, Mpe and Mphela (2013) respectively. They were selected based on their ease of accessibility and also on the fact that they do provide information on the use and function of Sepedi conjunctions.

Utilising each book as a distinct component of the SSC, the Oxford *Lebone* series contributes the following token counts:

- Grade 7 comprises 90 678 tokens,
- Grade 8 contains 91 674 tokens, and
- Grade 9 has 104 178 tokens.

This is a specialised corpus on which the distribution, usage and meaning of conjunctions is looked into to see how the material writers employ conjunctions.

The GSC is used as a reference corpus to provide comprehensive information on the usage and meaning of conjunctions in general language. This corpus serves as a benchmark for comparing conjunction usage and meaning against the trends observed in the SSC. A detailed exposition of the compilation process for both corpora is provided in Chapter 3.

This chapter concerns itself with a discussion of the process of identifying Sepedi conjunctions from the BONSE dictionary, as well as employing the Whelk Tool of LancsBox X to compare the distribution of Sepedi conjunctions across SSC files. This chapter further details the syntactic and semantic features of selected conjunctions for inclusion in the study, beginning with results from SSC. Moreover, findings from the SSC

are compared with those from the GSC. Lastly, general discussion concerning findings obtained on usage and meaning of Sepedi conjunctions is provided.

### 4.2 Identifying Sepedi conjunctions from Bilingual Oxford Northern Sotho-English dictionary

In the BONSE dictionary, a total of 20 Sepedi conjunctions are treated. Since lemma selection for this dictionary was frequency-based, it can be assumed that conjunctions treated in this dictionary are of high frequency. The conjunctions treated in the BONSE are provided in **Table 4.1** below.

**Table 4. 1**: Total number of Sepedi conjunctions treated in BONSE

| **Conjunctions from the BONSE** |
|---|
| *anthe* 'and yet' |
| *ebile* 'then; even' |
| *efela* 'but' |
| *eitše* 'as; while' |
| *empa* 'but' |
| *erile* 'when' |
| *ešita* 'even; as well as' |
| *eupša* 'but' |
| *ge* 'when; while; if' |
| *goba* 'or' |

| |
|---|
| *ka baka la* ‘because (of)’ |
| *ka fao* ‘therefore’ |
| *kapa* ‘or’ |
| *mohla* ‘while’ |
| *Mola* ‘while; whereas; since; even though’ |
| *nkane* ‘why’ |
| *nke* ‘as if’ |
| *ya ba* ‘it was then that; after/ following that; then' |
| |
| *gomme* ‘and ‘ |
| *gore* ‘(so) that’ |

In the preceding discussion, Sepedi conjunctions were identified from the BONSE dictionary. In the following section, the focus is on employing the Whelk tool to record the distribution of conjunctions across SSC files.

### 4.3 The distribution of Sepedi conjunctions across SSC

The Whelk Tool (a tool that provides information on how the search term is distributed across corpus files) is used to record distribution of each conjunction in individual files. As a first step, the Whelk Tool is employed to search for each conjunction to note its distribution across the individual files in the SSC. The number of incidences per grade are recorded (see **Table 4.2** below). There are two panels in the Whelk Tool (see below **Figure 4.1**), the top one displays the concordance lines, including search term, while the bottom panel shows the distribution of a search term in the individual files. The concordance lines at the top display information on the occurrence of the search terms, and at the bottom, the tool provides information on the number of tokens and the frequency in each file.

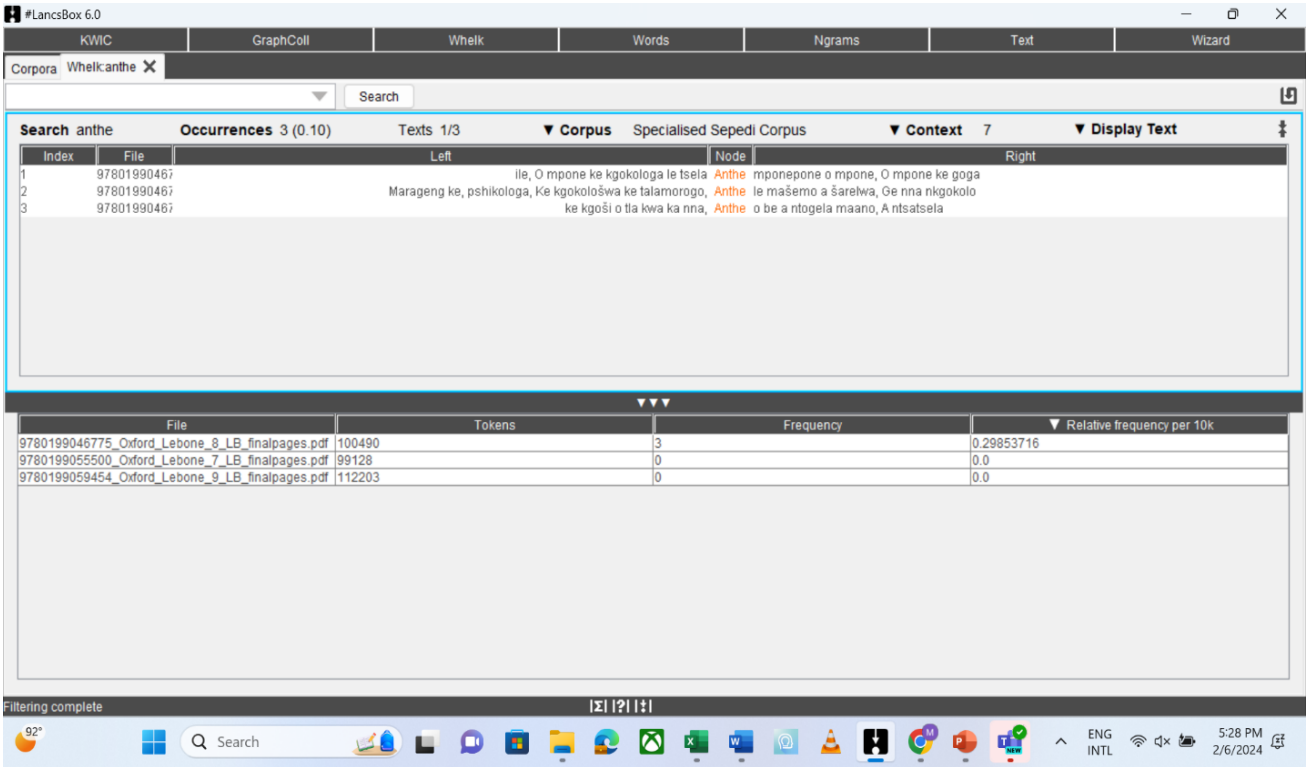


***Figure 4. 1**: Whelk tool displaying the distribution of searched conjunction anthe in the SSC*

**Figure 4.2** below displays the button panel showing the distribution of the conjunction *ebile* in the SSC files. It becomes apparent that the conjunction *ebile* in the Grade 7 file

appears 28 times, in the Grade 8 file it appears 70 times and in the Grade 9 file it appears 96 times. Due to space limitations, distribution of other conjunctions is summarised in a tabular format (see **Table 4.2** below).

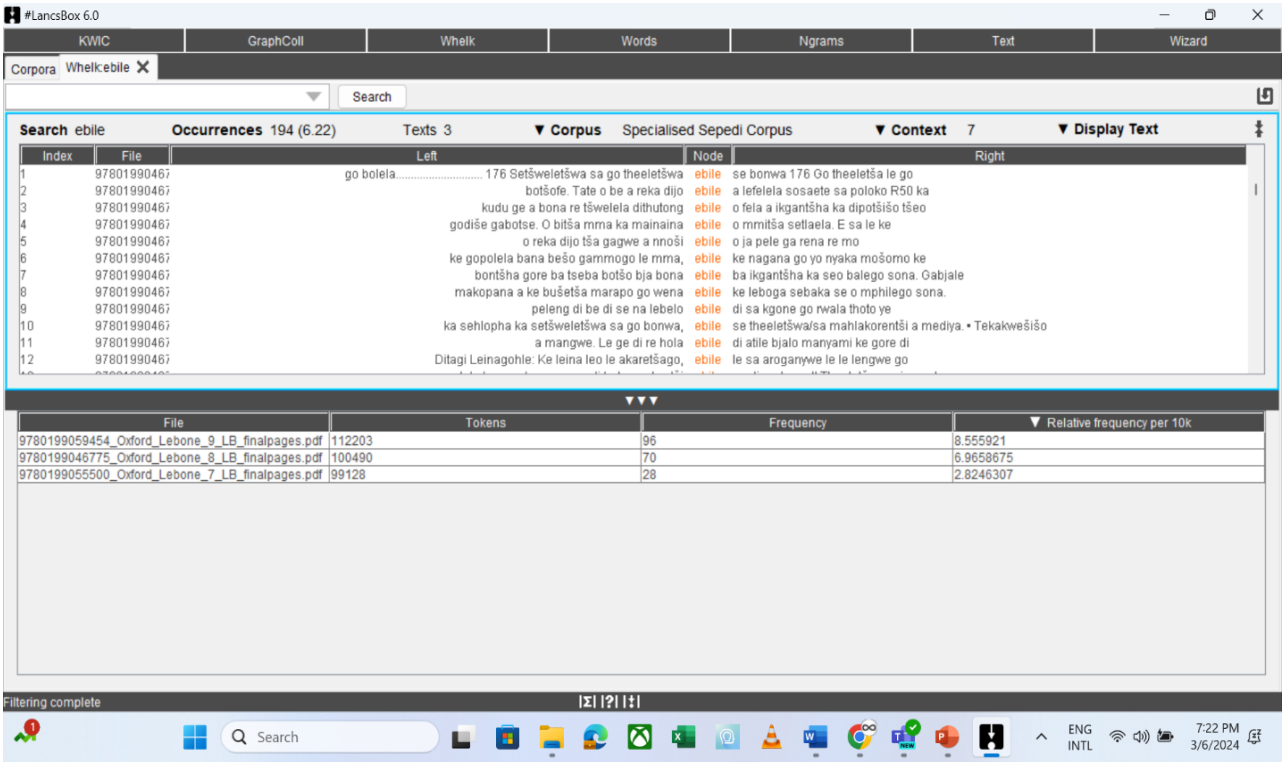


***Figure 4. 2**: Whelk tool displaying the bottom panel showing the distribution of conjunctions ebile across SSC files*

***Table 4. 2:** Distributions of Sepedi conjunctions across the SSC files*

| ***BONSE*** | ***Grade 7*** | ***Grade 8*** | ***Grade 9*** |
|---|---|---|---|
| *anthe* 'and yet' | 0 | 3 | 0 |
| *ebile* 'then; even' | 28 | 70 | 90 |
| *efela* 'but' | 5 | 1 | 7 |
| *eitše* 'as; while' | 0 | 0 | 0 |

| | | | |
|---|---|---|---|
| *empa* 'but' | 6 | 0 | 1 |
| *erile* 'when' | 0 | 0 | 0 |
| *ešita* 'even; as well as' | 0 | 2 | 6 |
| *eupša* 'but' | 20 | 25 | 32 |
| *ge* 'when; while; if' | 824 | 789 | 1030 |
| *goba* 'or' | 244 | 228 | 362 |
| *gomme* 'and ' | 174 | 286 | 458 |
| *gore* '(so) that' | 757 | 689 | 1249 |
| *ka baka la* 'because (of)' | 12 | 43 | 28 |
| *ka fao* 'therefore' | 21 | 20 | 34 |
| *kapa* 'or' | 0 | 1 | 0 |
| *mohla* 'while' | 4 | 4 | 6 |
| *mola* 'while; whereas; since; even though' | 60 | 45 | 65 |

| | | | |
|---|---|---|---|
| *nkane* 'why' | 1 | 0 | 0 |
| *nke* 'as if' | 47 | 17 | 20 |
| *ya ba* 'it was then that; after/ following that; then' | 30 | 31 | 21 |

When examining **Table 4.2** above, it becomes evident that the Sepedi conjunctions *eupša, ebile* and *gomme* display a substantial increase in usage as learners' progress through Grade 7 to 9. Notably, there is a significant escalation in the frequency of occurrences of these conjunctions. The conjunction ***eupša*** increases from 20 occurrences in Grade 7 to 25 in Grade 8 and further to 32 in Grade 9. The conjunction *ebile* demonstrates a rise from 28 occurrences in Grade 7 to 70 in Grade 8 and a remarkable 96 appearances in Grade 9. *Gomme* displays a progression from 174 occurrences in Grade 7 to 286 in Grade 8 and even more pronounced 458 appearances in Grade 9. The conjunctions *ge, goba* and *gore* also exhibit an ascending frequency trend across different grades. The frequency of *ge* rises progressively from 824 occurrences in Grade 7 to 789 in Grade 8 and peaks at 1 030 in Grade 9. Similarly, *gore* shows an increasing trend, starting at 757 in Grade 7, decreasing slightly to 689 in Grade 8 and then surging to 1 249 in Grade 9. Likewise, the conjunction *mola* demonstrates an upward trajectory, with occurrences increasing from 60 in Grade 7 to 45 in Grade 8 and reaching 65 in Grade 9.

Furthermore, *goba* experiences a notable increase from 244 occurrences in Grade 7 to 228 in Grade 8, and ultimately reaching 362 in Grade 9. Conversely, *ya ba* and *ka fao* exhibit consistent frequencies across grades. An interesting observation is found in the usage of ***ešita***, which is absent in Grade 7, then records two occurrences in Grade 8,

followed by a notable increase to six in Grade 9. This noteworthy trend implies an elevated usage of ***ešita*** in the higher grades, i.e., Grade 9 of the Senior Phase.

Moreover, the distribution of conjunctions, *efela* and *empa*, shows notable differences across different grades. For instance, *efela* experiences a decline from Grade 7 (five instances) to Grade 8 (one instance) before surging to seven instances in Grade 9. Similarly, *empa* demonstrates a decrease in frequency, registering six instances in Grade 7, diminishing to zero instances in Grade 8, and re-emerging with one instance in Grade 9.

The conjunctions *kapa* and *anthe* exhibit exclusive occurrences in Grade 8, with frequencies of 1 and 3, respectively. Conversely, *nkane* is observed only in Grade 7, with a singular occurrence, and notably, absent in other grades. The conjunctions *erile* and ***eitše*** are clearly absent across all grades. These findings strongly indicate that *kapa*, *anthe*, *nkane*, *erile*, and ***eitše*** are not favoured conjunctions in the context of the selected Sepedi textbooks.

The aim of the preceding discussion is to shed some light on the frequency of usage of Sepedi conjunctions found in the BONSE dictionary. This will come in handy when we have to determine the conjunctions to be considered for inclusion in the comparative analysis.

The above observations reveal not only the quantitative shifts but also the qualitative nuances on how material writers employ conjunctions as they progress through different grades. In the section which follows, the focus is on performing comparison of frequency occurrences of conjunctions between SSC and GSC.

### 4.4 The frequency of occurrence of Sepedi conjunctions between SSC and GSC

A corpus-based analysis implies that conjunctions should be harvested from the corpus. This is typically done by means of the KWIC tool to see the frequency of occurrence of each conjunction in the textbooks and also to see the context in which the conjunctions appear in the textbooks. As a second step, analysis of these conjunctions is done to establish whether specific syntactic patterns around the usage of these conjunctions in the textbooks can be identified. The comparison is then carried out between the SSC and

GSC to see whether there are any similarities and differences in the meaning and usage of conjunctions in these two corpora. The purpose is to understand the specific contexts in which certain conjunctions are used in textbooks as compared to general language.

Both the SSC and GSC used in the investigation are raw corpus, which means they are not POS tagged. It is, therefore, not possible to search for specific parts of speech, such as conjunctions, verbs, nouns, etc. It is only possible to search for specific lexical items.

It has already been highlighted in the foregoing discussion that the KWIC tool (a tool that is used to generate concordance lines for a specific word) of LancsBox X is used for purposes of the comparative analysis. By using the KWIC tool, each conjunction is used as a search word. By studying the concordance lines, searched conjunctions can be identified. The number of incidences of each conjunction from each corpus is recorded in **Table 4.3** below.

**Figure 4.3 and 4.4** below display the KWIC tool showing the frequency of occurrence (at the top left corner) and contexts in which the conjunction *ebile* appears in the GSC (see **Figure 4.3**) and SSC (see **Figure 4.4**). The same procedure was followed to search for each conjunction in the individual corpora. Due to space limitations, other conjunctions in both corpora are summarised in a tabular format (see **Table 4.3** below).

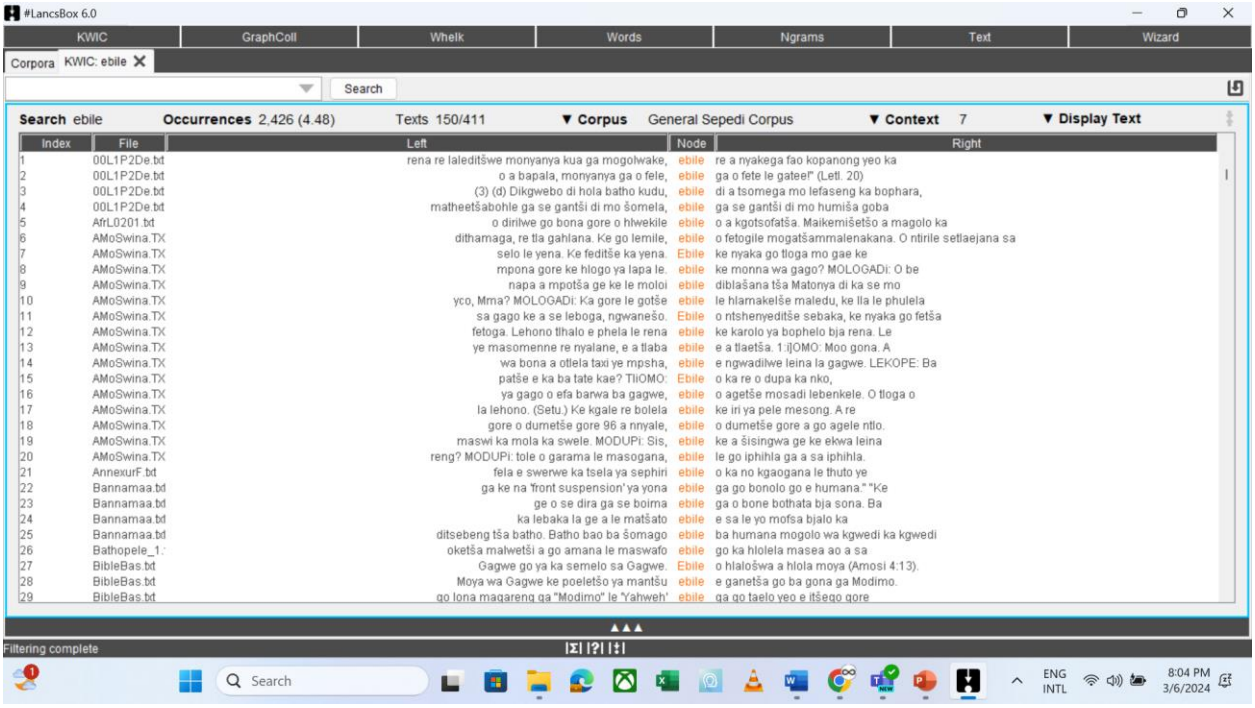


***Figure 4. 3:*** *KWIC tool showing frequency of occurrence and contexts of conjunction ebile in the GSC*

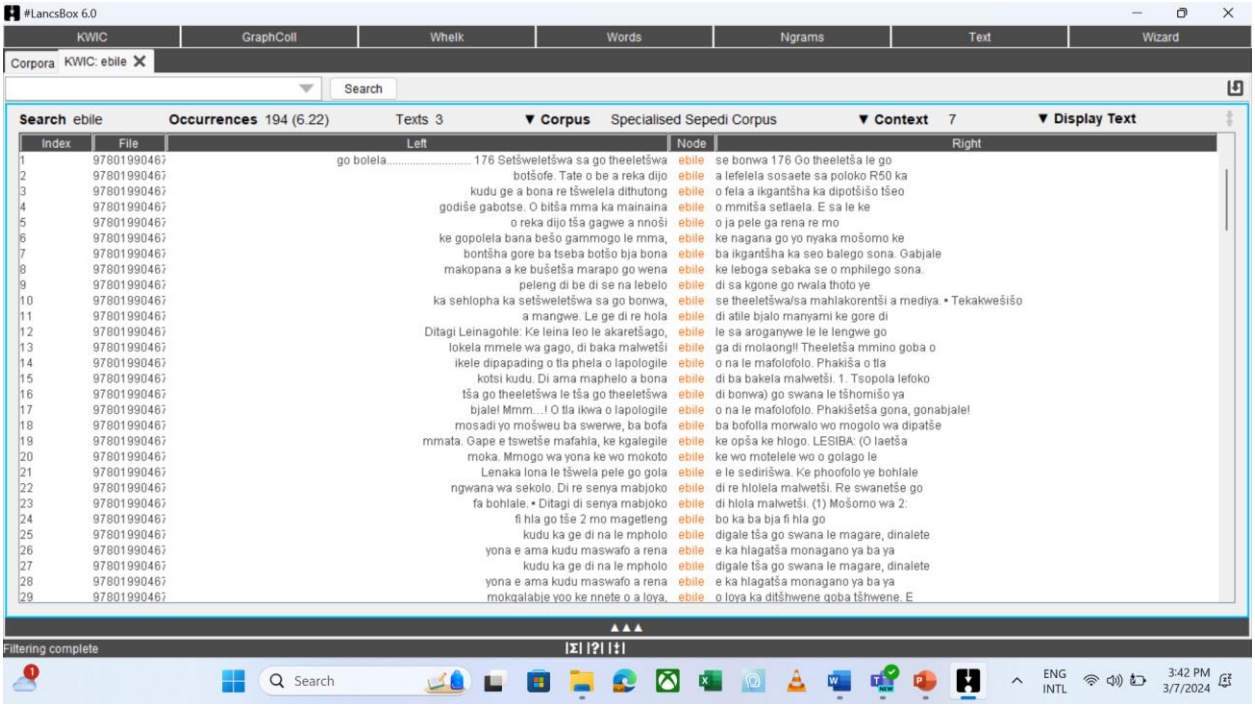


***Figure 4. 4**: KWIC tool showing frequency of occurrence and contexts of conjunction ebile in the SSC*

***Table 4. 3**: Frequency occurrence of Sepedi conjunctions between SSC and GSC*

| **BNSOE** | **Number of hits for each conjunction** | |
|---|---|---|
| | **Specialised Sepedi Corpus** | **General Sepedi Corpus** |
| *anthe* 'and yet' | 3 | 247 |
| *ebile* 'then; even' | 194 | 2 426 |
| *efela* 'but' | 13 | 485 |
| *eitše* 'as; while' | 0 | 25 |
| *empa* 'but' | 7 | 334 |
| *erile* 'when'' | 0 | 191 |
| *ešita* 'even; as well as' | 8 | 318 |

| | | |
|---|---|---|
| *eupša* 'but' | 77 | 4 592 |
| *ge* 'when; while; if' | 2 643 | 74 694 |
| *goba* 'or' | 834 | 18 919 |
| *gomme* 'and ' | 918 | 13 891 |
| *gore* '(so) that' | 2 695 | 63 444 |
| *ka baka la* 'because (of)' | 83 | 1 572 |
| *ka fao* 'therefore' | 74 | 3 132 |
| *kapa* 'or' | 1 | 376 |
| *mohla* 'while' | 14 | 1 565 |
| *mola* 'while; whereas; since; even though' | 170 | 6 604 |
| *nkane* 'why' | 1 | 43 |
| *nke* 'as if' | 84 | 1 888 |
| *ya ba* 'it was then that; after/ following that; then' | 82 | 4 054 |

The **Table 4.3** shows contrast in usage of conjunctions between the two corpora. Notably, conjunctions ***eitše*** and *erile* exhibit zero occurrence within the SSC, contrasting sharply with their prevalence in the GSC. Similarly, *kapa* and *nkane* manifest infrequent appearance within the SSC, with only one occurrence for each, whereas their utilisation is noticeably more prominent within the GSC.

The aim of showing this comparison of frequencies was to help the researcher to arrive at the choice of conjunctions to be included in the comparative analysis, as analysing all of them would not be feasible within the scope of this study. Therefore, only conjunctions with high frequency of occurrence in both corpora are considered, as the aim of the study

is to compare usage of Sepedi conjunctions in specialised language and general language.

These corpus-based results can assist material developers notice that incorporating these conjunctions in the study material can foster a more comprehensive grasp of Sepedi conjunctions and enhance language proficiency among learners.

Now that we have compared frequency of occurrence of each conjunction between the SSC and GSC, it is time to delve into the syntactic and semantic features of Sepedi conjunctions in the SSC.

## 4.5 Syntactic and semantic features of Sepedi conjunctions from SSC

As indicated in the foregoing discussion, it was decided that six conjunctions should form part of the comparative analysis, i.e*. ebile, ge*, *goba*, *gomme*, *gore* and *mola*. The rationale behind this selection was based on the frequency of occurrence of these conjunctions in both corpora (see **Table 4.3** above). Some of them have high frequency in both corpora while others are occurring frequently in one corpus and not the other. For instance, *ebile* appears 194 times in the SSC and 2 426 times in the GSC. *Ge* occurs 2 643 times in the SSC and 74 694 times in the GSC. The rest of the frequencies of occurrence for each of the conjunctions selected can be seen in **Table 4.3** above.

In the sections which follow, the focus is on usage and meaning of conjunctions. The KWIC tool is used to retrieve concordance lines for each conjunction in the Specialised Sepedi Corpus. The conjunction to be considered first is *ebile*.

### *4.5.1 The conjunction: ebile*

Using '*ebile*' as a search word in the SSC, the KWIC tool produced 194 lines (see **Figure 4.5**). A manual analysis of this number of KWIC lines is not feasible and therefore, only a representative sample of 100 lines is selected for analysis. The 100 lines are randomly selected, and this is achieved by making use of the Left-double-click on the header of the 'Index' column. The sample of 100 KWIC lines is subsequently analysed.

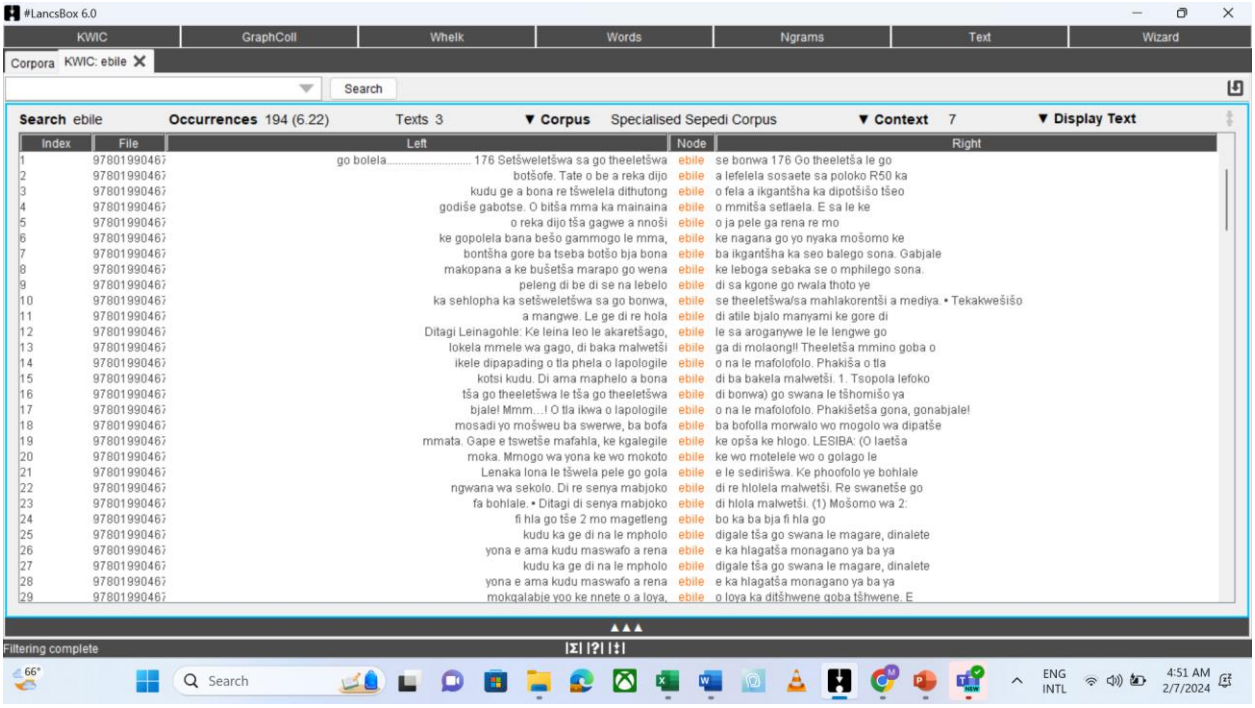


***Figure 4. 5**: KWIC tool displaying ebile as a search word in the SSC*

The data from KWIC lines show that *ebile* appears 90 times positioned between clauses with no comma usage (see **Figure 4.6**), five times positioned between clauses with comma usage before (see **Figure 4.7**) and twice positioned at the beginning of a sentence (see **Figure 4.8**) and three times unspecified (see **Figure 4.9**).

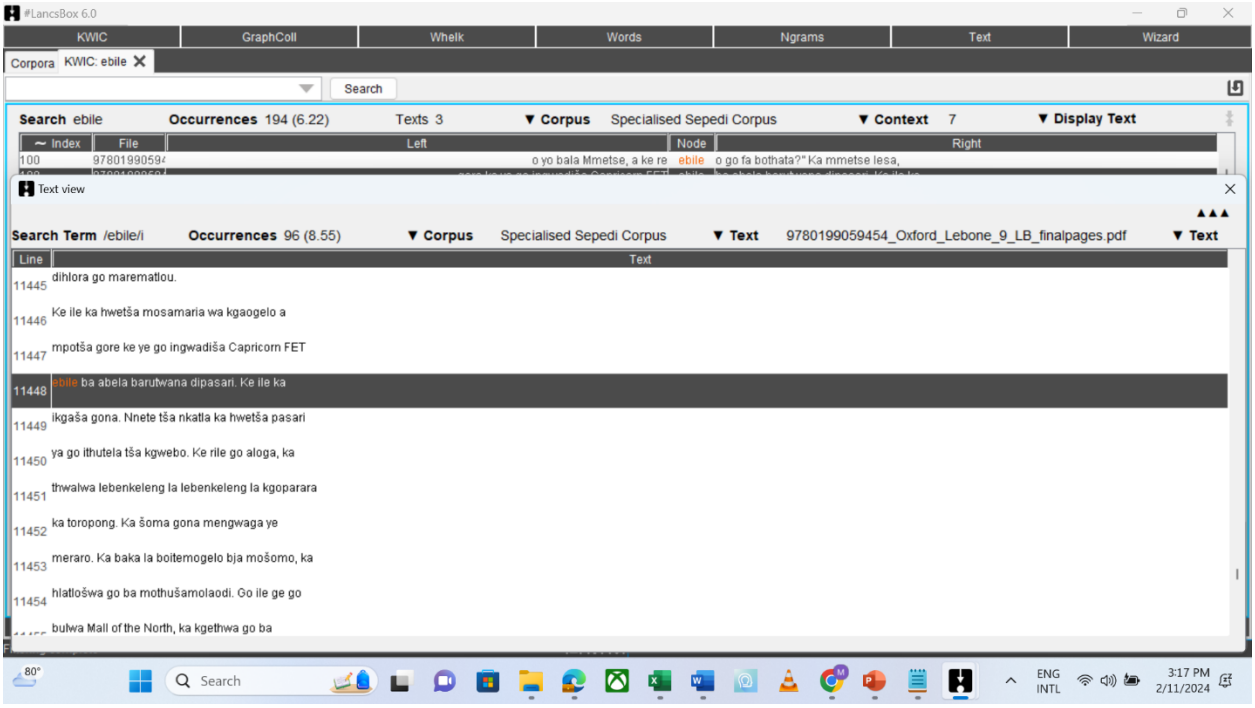


***Figure 4. 6**: Conjunction ebile positioned between clauses with no comma usage in the SSC*

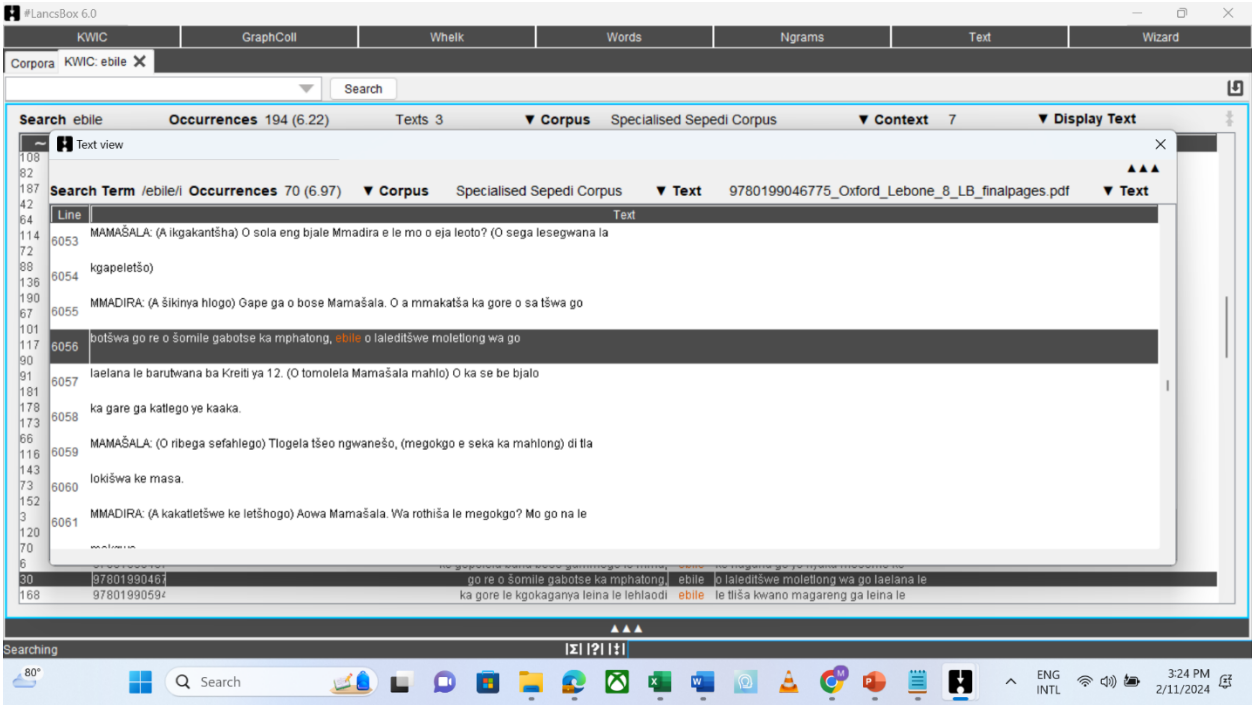


***Figure 4. 7**: Conjunction ebile positioned between clauses with comma usage before in the SSC*

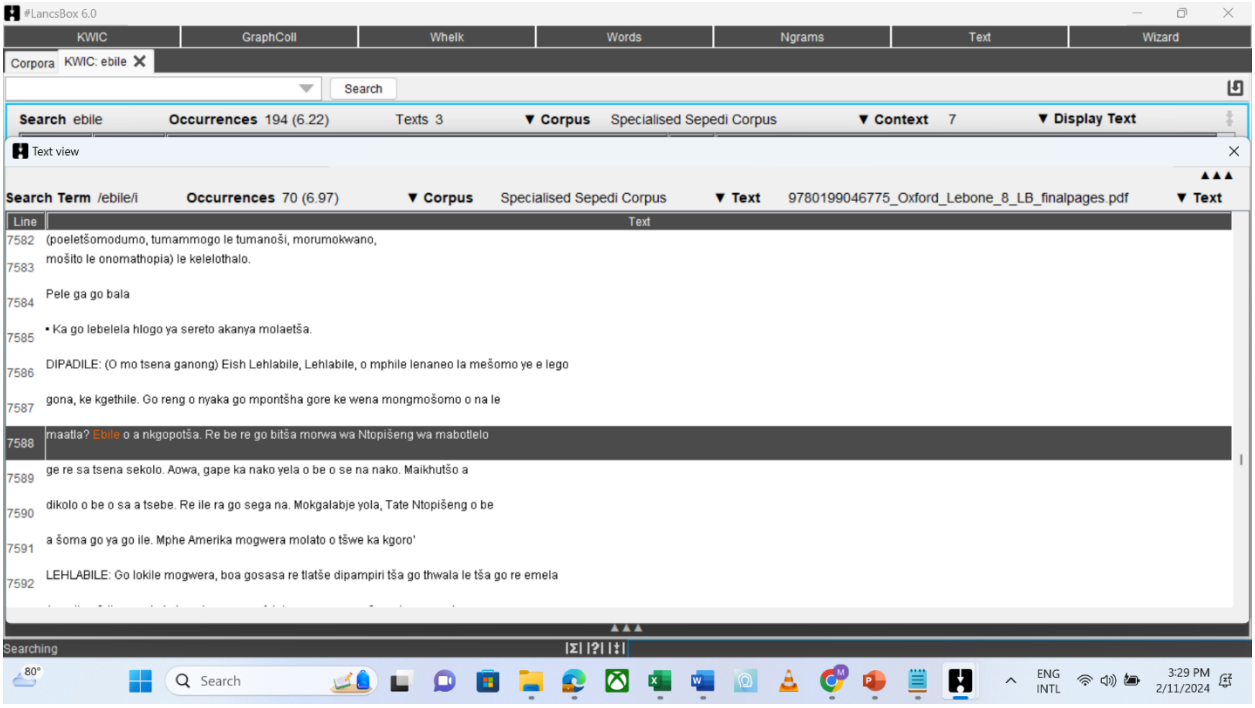


***Figure 4. 8**: Conjunction ebile positioned at the beginning of sentence in the SSC*

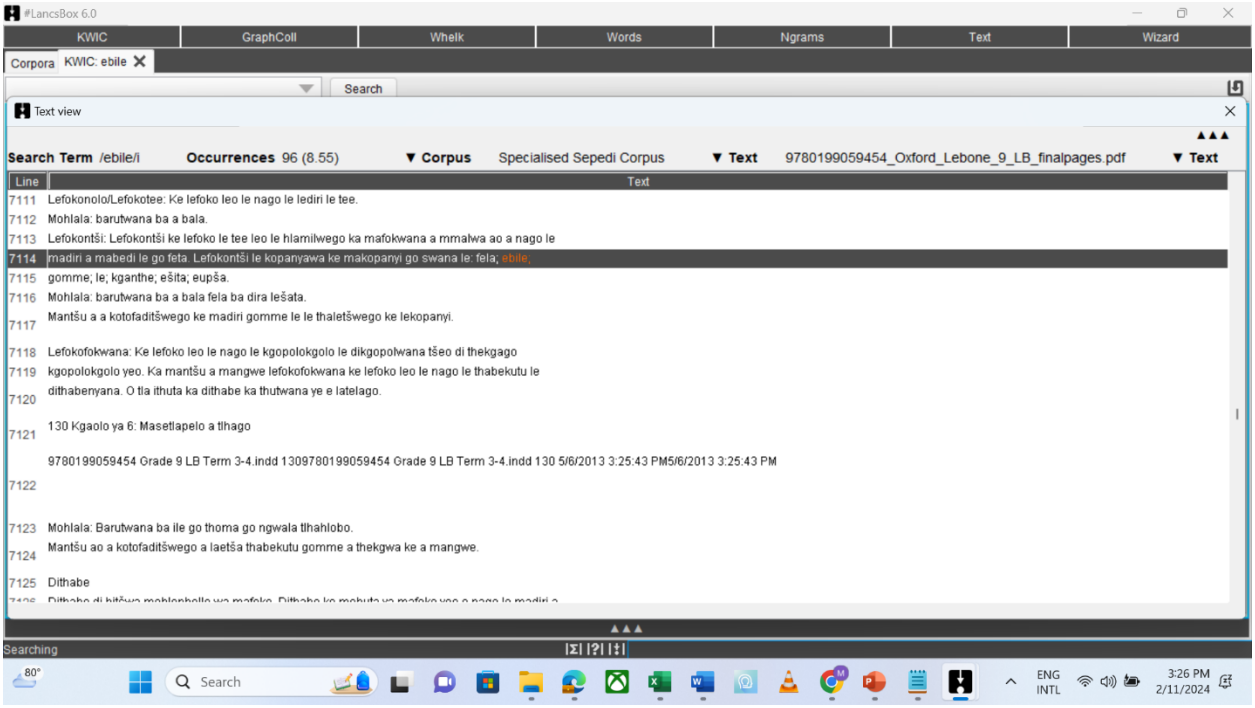


***Figure 4. 9**: Sepedi conjunction ebile appear unspecified in the SSC*

The results show that the usage of *ebile* positioned between the clauses with no comma usage is more frequent than its usage with a comma before and at the beginning of a sentence. The results imply that usage of this conjunction positioned between clauses with no comma usage is a much more prominent discourse compared to its usage with a comma before as well as positioned at the beginning of a sentence.

It is also necessary to highlight what these findings entail in terms of the theoretical framework underpinning the present study. Since Generative Grammar theory is concerned with the rules and laws governing production of grammatically correct sentences or utterances, the conditions highlighted above under which the conjunction *ebile* is used in Sepedi are indicative of the rules or laws governing the use of this conjunction in grammatically correct sentences. Proponents of Generative Grammar further contend that these laws governing the production of grammatical sentences are innate, that is, they are not acquired but children are born with them. This entails human beings are born with the ability in their minds to speak the language. When they hear utterances or read sentences, this innate ability gets activated in the mind to enable them to judge whether or not the utterance or sentence is grammatically correct. What this entails in terms of these findings is that children are born with knowledge of the conditions under which the conjunction *ebile* functions. When a child is taught these conditions at school or at home, the innate knowledge of them in the mind gets activated and the child is able to easily grasp the conditions.

The semantic features of the conjunction *ebile* were checked by going through the concordance lines. When the conjunction is used in the SSC, it expresses the following senses:

- expresses addition which can be translated as ‘furthermore’, ‘and’, and
- expresses contrast or change which can be translated as ‘even’.

Let us consider the following examples gleaned from the SSC concordance lines:

- *Mogale o kwagetše a sa tsena seferong gore o bo gašitše, ka ge a tsene ka go hlaba mašata* ***ebile*** *a hlapola mosadi wa gagwe.* ‘Mogale was heard upon

entering the gate that he is drunk, because upon his arrival he made noise **furthermore** swore at his wife'.

- *Leinagohle: Ke leina leo le akaretšago,* ***ebile*** *le sa aroganywe le le lengwe go amangwe.* 'A common noun: it is a generalising noun, **and** it does not distinguish one from others'.
- *Go reng o nyaka go mpontšha gore ke wena mongmošomo o na le maatla?* ***Ebile*** *o a nkgopotša. Re be re go bitša morwa wa Ntopišeng wa mabotlelo ge re sa tsena sekolo.* 'Why do you want to show me that you are the employer and you have power? You **even** remind me. We used to call you son of Ntopišeng of bottles when we were at school'.

From the above examples, it is clear that the Sepedi conjunction *ebile* can be translated as 'furthermore', 'and' and 'even' in English. This is clear from the examples above. Let us move on to the second conjunction that was sampled, namely *ge*.

### *4.5.2 The conjunction: ge*

The search term/conjunction *ge* occurs 2643 times in the SSC (see **Figure 4.10**).

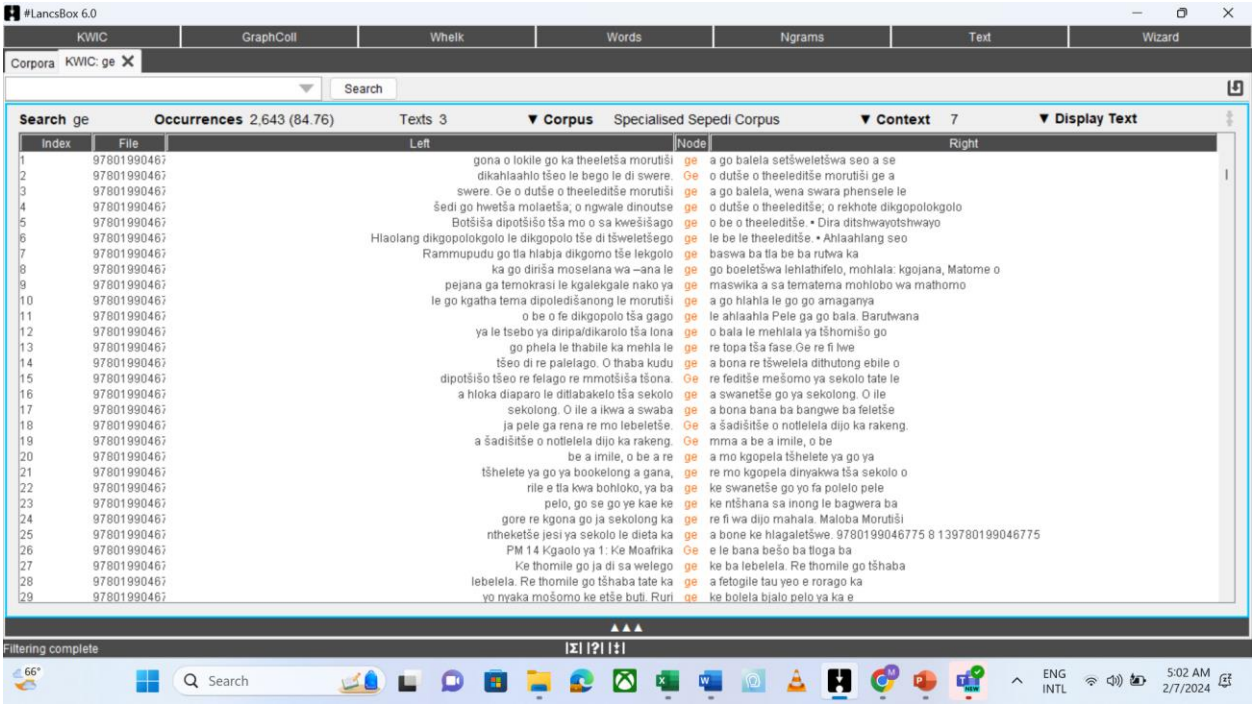


***Figure 4. 10**: KWIC tool displaying ge as a search word in the SSC*

The KWIC lines for *ge* in the 100 lines sampled show that *ge* is positioned 71 times at the beginning of a sentence, (see **Figure 4.11**), positioned 21 times between clauses with a comma usage after the conjunction (see **Figure 4.12**), positioned five times between clauses with comma usage before the conjunction (see **Figure 4.13**) and it appears three times unspecified (see **Figure 4.14**).

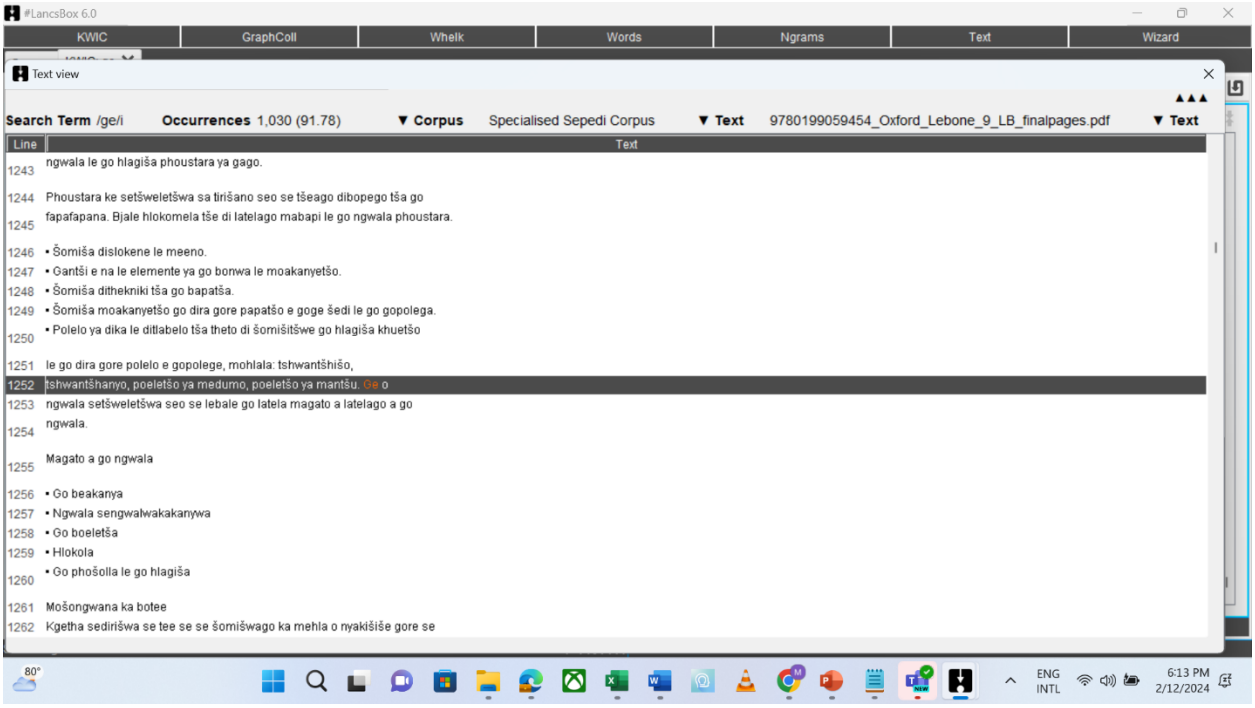


***Figure 4. 11**: The conjunction ge positioned at the beginning of a sentence in the SSC*

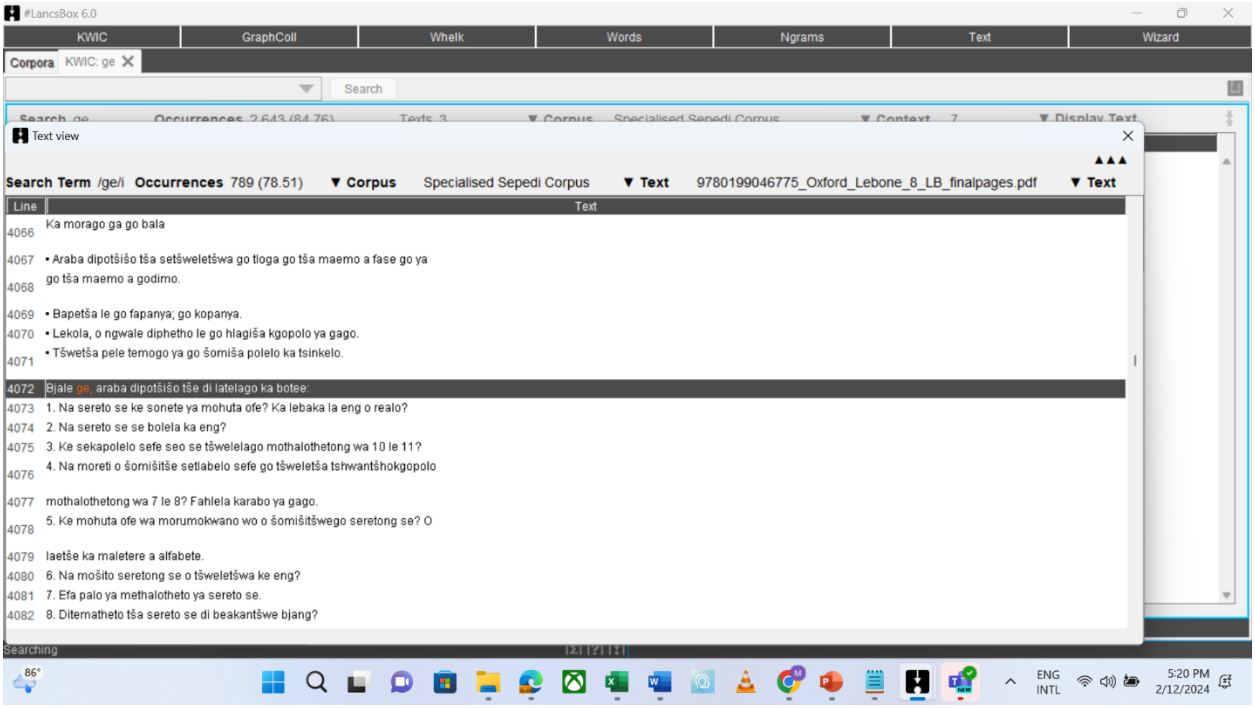


***Figure 4. 12**: The conjunction ge positioned between clauses with comma usage after in the SSC*

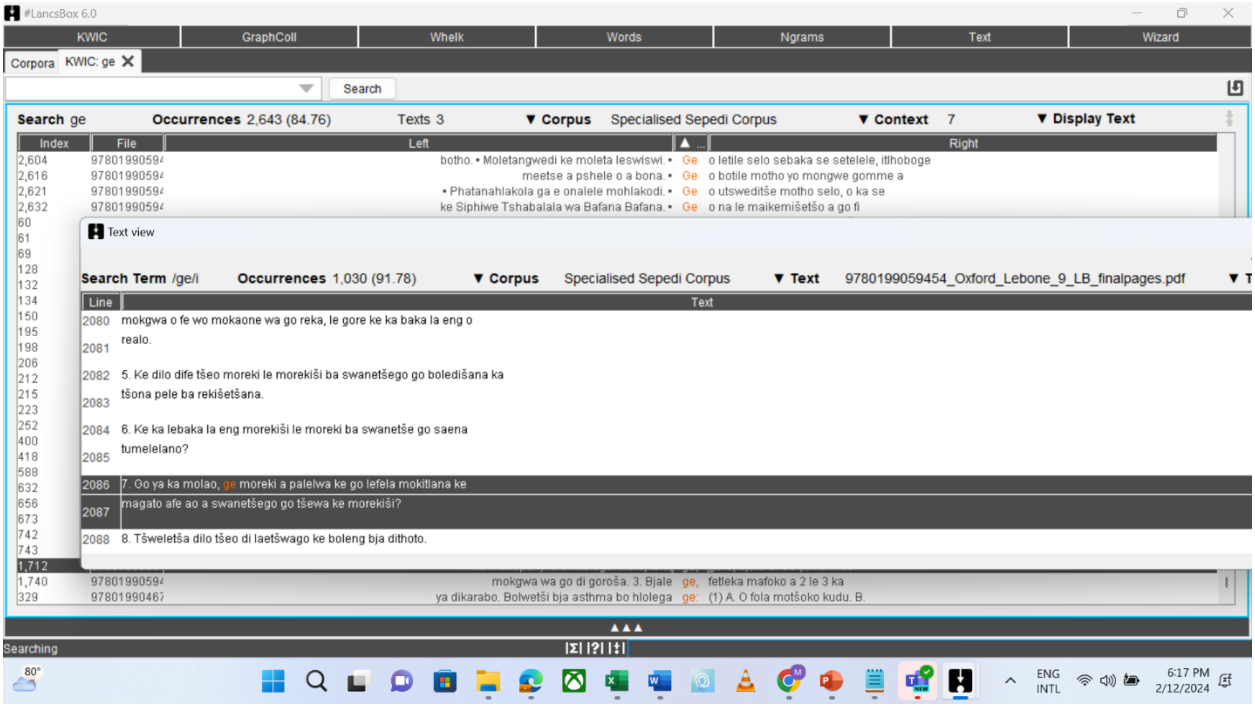


***Figure 4. 13**: The conjunction ge positioned between clauses with comma usage before in the SSC*

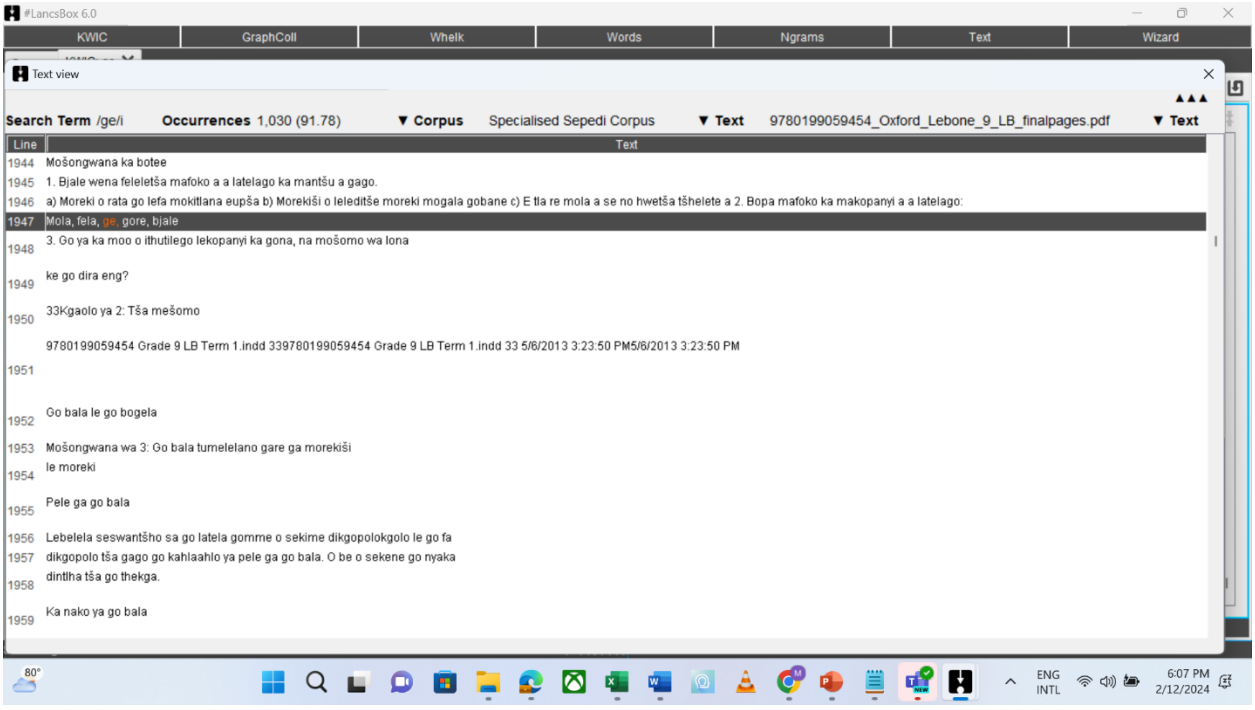


***Figure 4. 14**: The conjunction ge appearing unspecified in the SSC*

The results show that the usage of *ge* positioned at the beginning of a sentence is more frequent than its usage positioned between clauses with a comma usage before or after. The results demonstrate that usage of conjunction *ge* positioned at the beginning of a sentence is much more predominant than its usage with comma before or after. It terms of Generative Grammar approach, the usage of the conjunction *ge* positioned at the beginning of a sentence as well as its usage before and after a comma are part of the universal laws governing the production of grammatical sentences. When Sepedi children encounter these laws for the first time, the innate knowledge they have about them gets activated and they are able to understand how the conjunction functions.

The sematic features of the conjunction *ge* were also checked by going through the concordance lines. When the conjunction *ge* is used in the SSC, it expresses logical order or time which can be translated as 'after', 'when' and 'then'. Consider the following examples gleaned from the SSC concordance lines:

- *O tloga o le sehlogo wena patše.* ***Ge*** *o šetše o fitile bana ba fetoga manaba setšhabeng,* ***ge*** *o šetše o fitile bana ba lahlagelwa ke bokamoso,* ***ge*** *o šetše o fitle bana ba lahlelwa kgolegong.* 'You are cruel marijuana. **After** passing, children become enemies of the community, **after** passing children lose their future, **after** passing children become imprisoned'.

- *….thaba kudu* ***ge*** *a re bona re tšwelela dithutong ebile a fela a ikgantšha tšeo re felago re mmotšiša tšona.* ***Ge*** *re feditše mešomo ya sekolo mma le tate ba re anagela dinonwane.* 'He/she becomes happy **when** seeing us academically succeeding and sometimes proud of him/herself on things we ask him/her. **After** we have finished our school works, mother and father narrate folktales for us'.

- *Barutwana ba swanetše go laetša mabokgoni a go beakanya, go nyakišiša le go rulaganya pele ba ka hlagiša ka bomolomo: bjale* ***ge****, šomang ka dihlopha le swere dipotšišotherišano ka go araba dipotšišo tše di latelago gore ka morago le kgone go itlhamela tša lena le go kgatha tema ka mafolofolo thutong ya lehono.* 'Learners must exhibit organising skills, investigation skills and editing skills before they can present orally: **then,** now work in groups and have interviews with the aim of

answering the following questions so that after you can create yours and participating in today's lesson'.

From the above examples, it is clear that the Sepedi conjunction *ge* can be translated as 'after', 'when' and 'then' in English. Let us now proceed to the third conjunction that was sampled, namely *goba*.

### *4.5.3 The conjunction: goba*

The conjunction *goba* occurs 834 times in the SSC, as shown in **Figure 4.15** below.

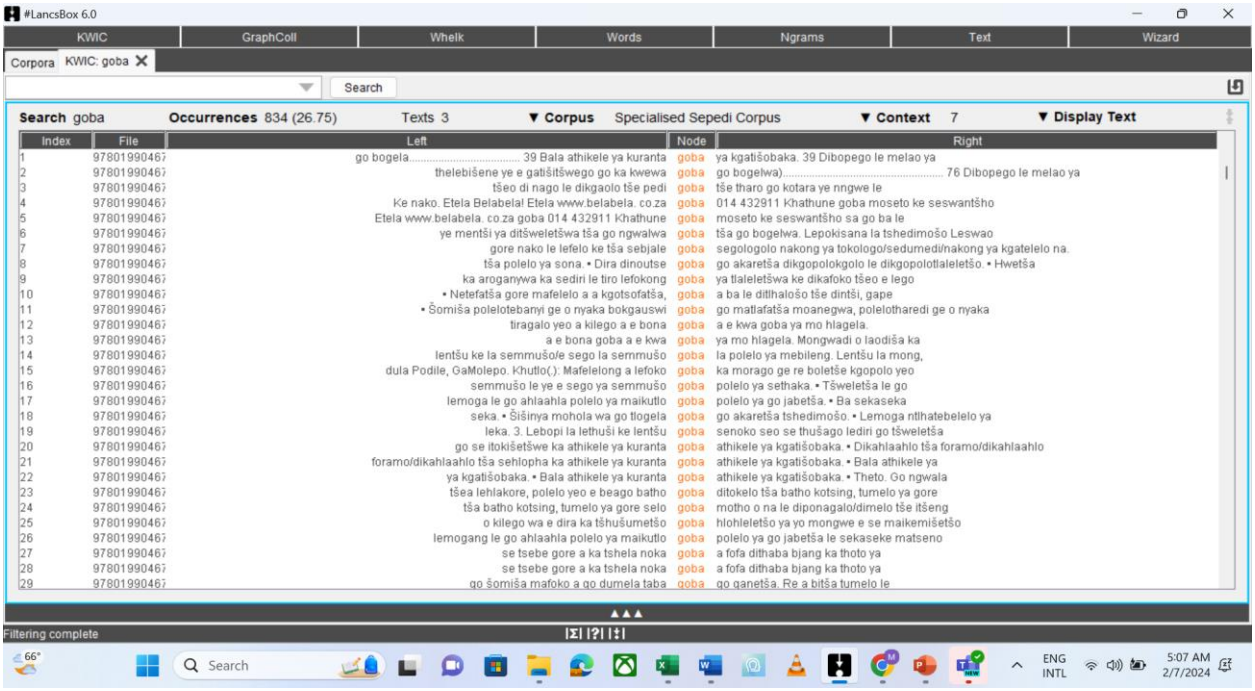


***Figure 4. 15**: KWIC tool displaying goba as a search word in the SSC*

The 100 KWIC lines for the conjunction *goba* reveal that it appears 96 times positioned between clauses with no comma usage (see **Figure 4.16**) and four times positioned between clauses with comma usage before (see **Figure 4.17**).

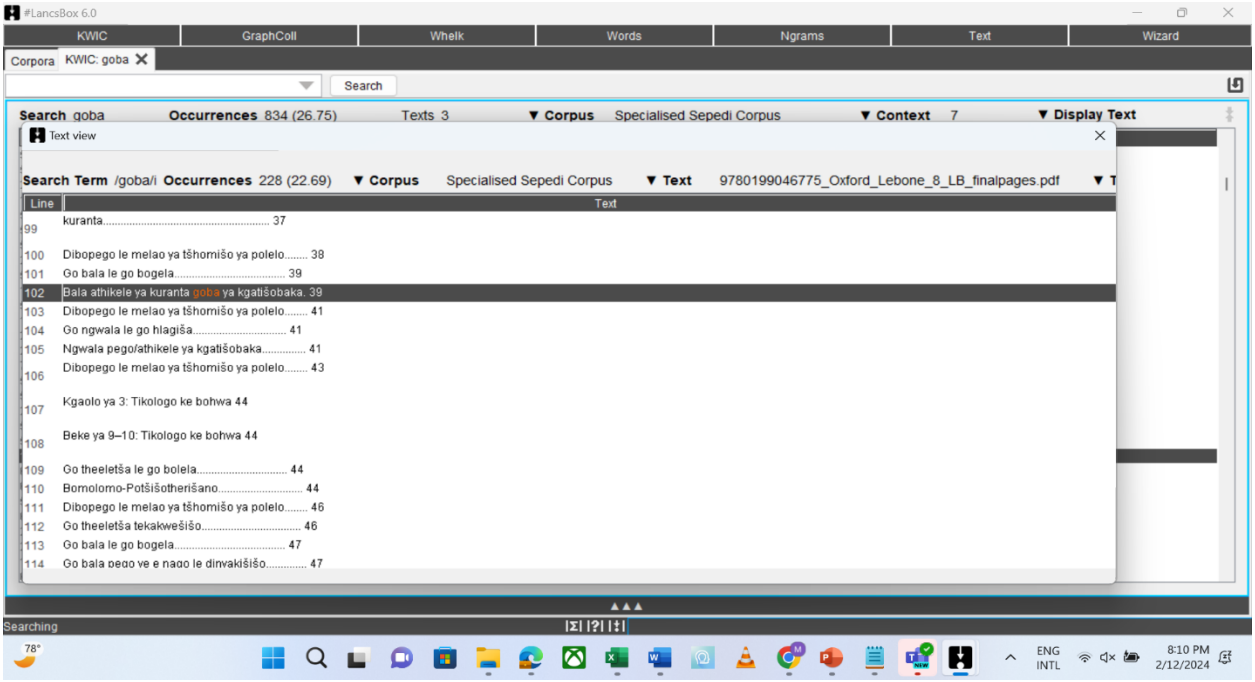


***Figure 4. 16**: The conjunction goba positioned between clauses with no comma usage in the SSC*

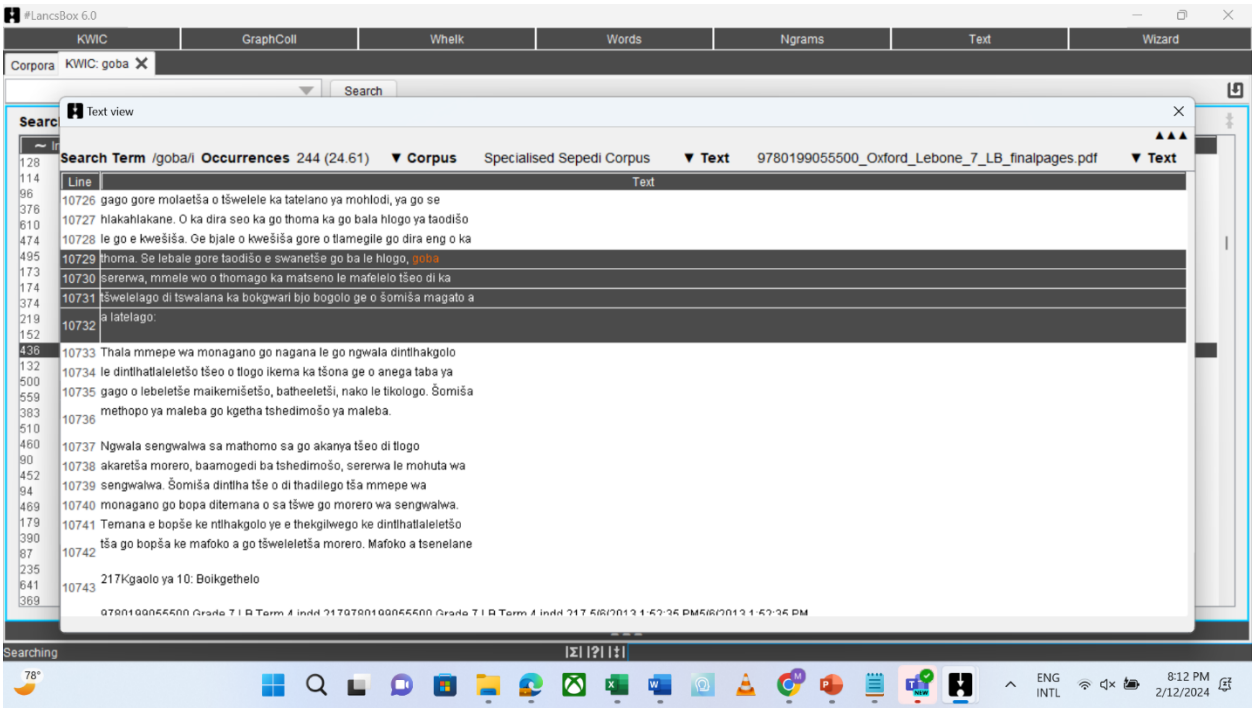


***Figure 4. 17**: The conjunction goba positioned between clauses with comma usage before in the SSC*

The results demonstrate that the usage of *goba* positioned between the clauses with no comma usage is more frequent than its use with a comma before. The results entail that the usage of the conjunction positioned between clauses with no comma usage is a much more prominent discourse than its usage with comma before. In terms of Generative Grammar theory, the conditions of the usage of the conjunction *goba* positioned between clauses without a comma and usage with comma before constitute the general laws governing production of grammatically correct sentences. Everyone, especially Sepedi speakers, is born with these conditions in mind.

The results further demonstrate that when the conjunction *goba* is used in the SSC, it expresses alternatives or choices, which can be translated as 'or'. Consider the following examples gleaned from the SSC concordance lines:

- *Se lebale gore taodišo e swanetše go ba le hlogo,* ***goba*** *sererwa, mmele wo o thomago ka matseno le mafelelo tšeo di ka tšwelelago di tswalana ka bokgwari bjo bogolo ge o šomiša magato a a latelago.* 'Do not forget that an essay has to have a title, **or** topic, body that contains an introduction and conclusion that display their relation in a skillful manner when you use the following steps'.

- *Tše di latelago ke diponagalo tša go ngwala lengwalo la segwera* ***goba*** *leo e sego la semmušo.* 'The following are elements of writing a friendship letter **or** an informal letter'.

The above examples show that this conjunction can be translated as 'or' in English. Let us move on to the fourth conjunction that was sampled, namely *gomme*.

### *4.5.4 The conjunction: gomme*

The conjunction *gomme* occurs 918 times in the Specialised Sepedi Corpus, as demonstrated in **Figure 4.18** below.

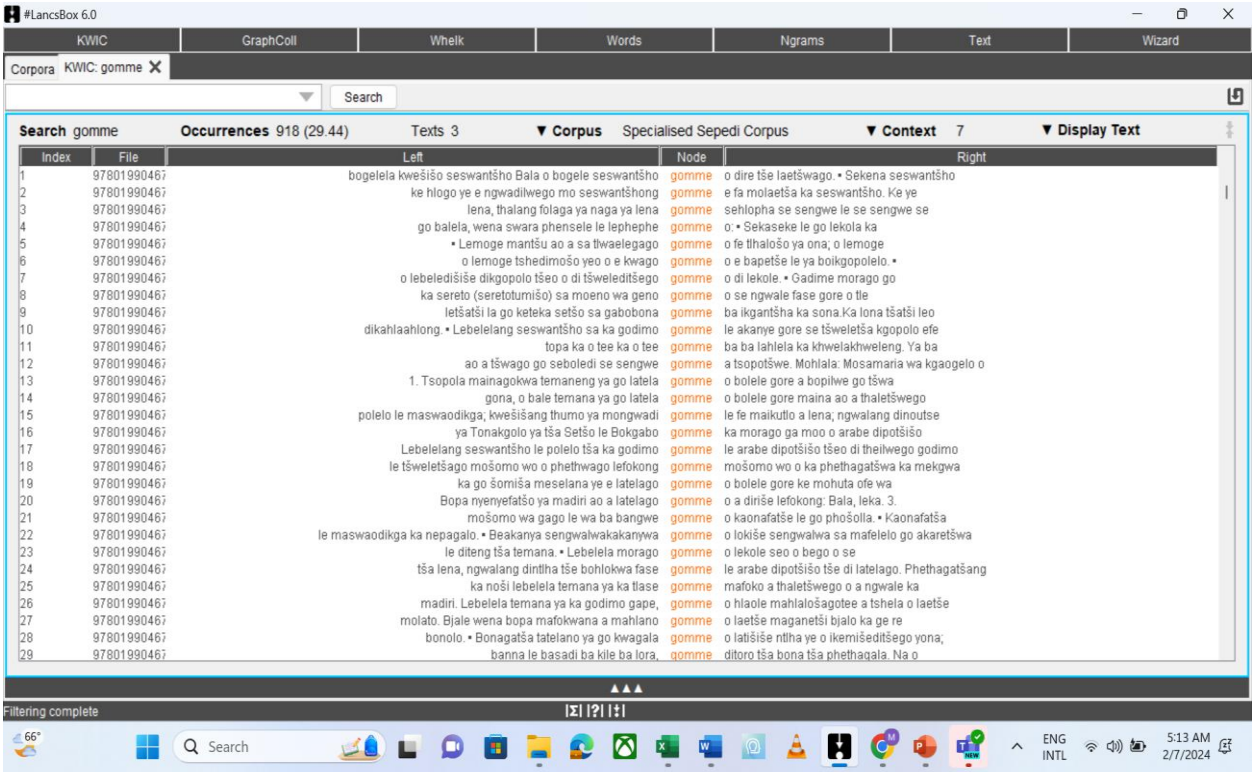


***Figure 4. 18**: KWIC tool displaying gomme as a search word in the SSC*

The 100 KWIC lines indicate that the conjunction *gomme* appears 90 times positioned between clauses with no comma usage (see **Figure 4.19**), eight times positioned between clauses with comma usage before (see **Figure 4.20**) and twice unspecified (see **Figure 4.21**).

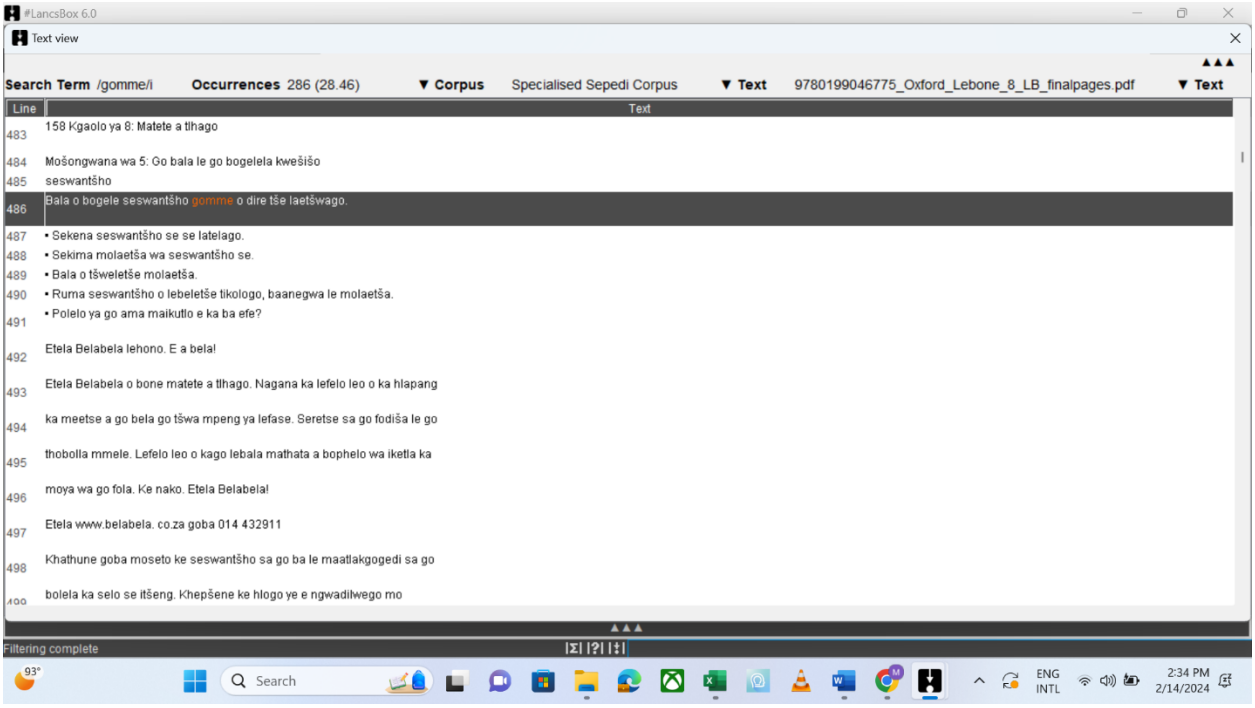


***Figure 4. 19**: The conjunction gomme positioned between clauses with no comma usage in the SSC*

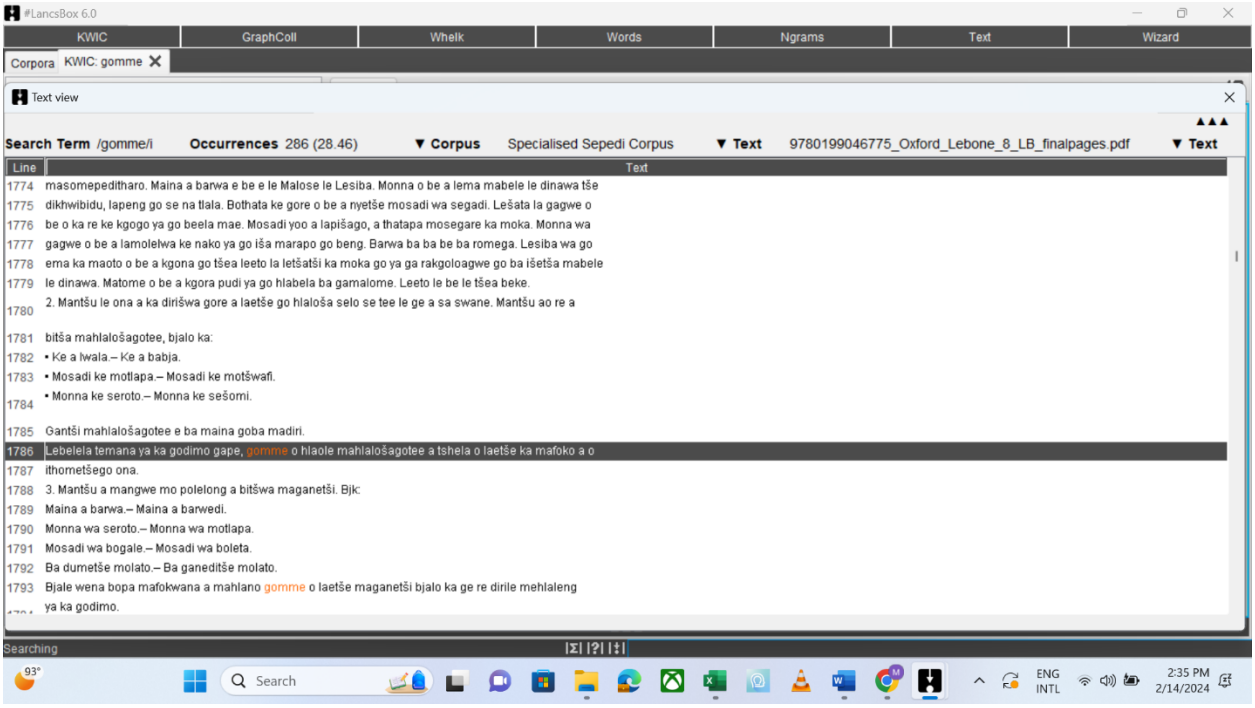


***Figure 4. 20**: The conjunction gomme positioned between clauses with comma usage before in the SSC*

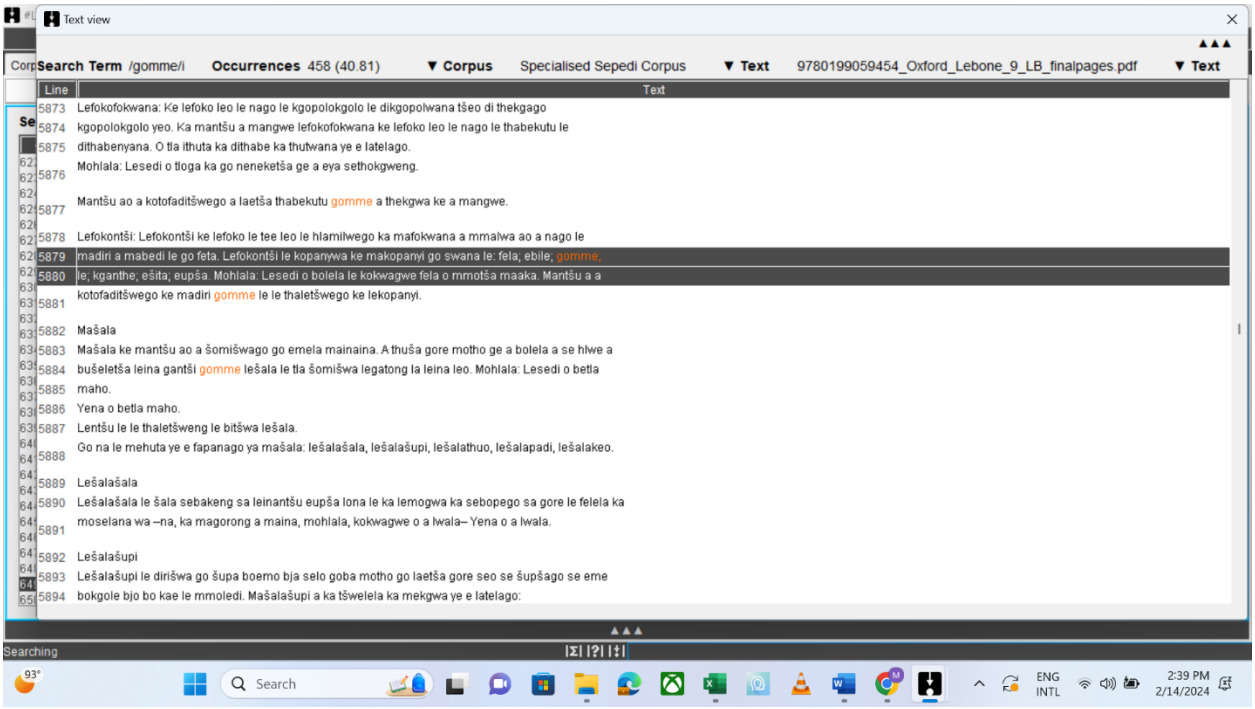


***Figure 4. 21**: The conjunction gomme appears unspecified in the SSC*

These results show that the use of *gomme* positioned between clauses with no comma usage is more frequent than its use with comma before. The findings, therefore, demonstrate that *gomme* positioned between clauses with no comma usage is a much more predominant discourse than its usage with comma before. In terms of these findings, Generative Grammar proponents will argue that the condition of the usage of the conjunction *gomme* positioned between clauses with no comma usage and that of its usage positioned between clauses with a comma before the conjunction form part of the universal rules governing production of grammatical sentences.

The results further indicate that when the conjunction *gomme* is used in the SSC, it expresses addition which can be translated as 'and'. Let us look at the following examples gleaned from the SSC:

- *Kgetha dintlha tšeo o di ratago **gomme** o fe mabaka gore o di ratela eng. '*Select your favorite points **and** give reasons why they are your favorites'.
- *O be a šoma ka go latela molao, a tšhaba go tsena marageng. Ba ile ba mmotša gore o kgethilwe ke bona le gore ba tla mo tloša setulong. Lesiba o ile a gana go tšhošetšwa, **gomme** a tšwela pele go thatafiša hlogo.* 'He/she used to work in accordance to the law, being afraid of being implicated in unlawful acts. They told him/her that they voted him/her and they can remove him/her from the position. Lesiba did not like to be threatened, **and** he/she went on being stubborn'.

The above examples reveal that the conjunction *gomme* means 'and' in English. Let us carry on and consider the firth conjunction that was sampled, namely *gore*.

### *4.5.5 The conjunction: gore*

The conjunction *gore* occurs 2695 times in the SSC (see **Figure 4.22**).

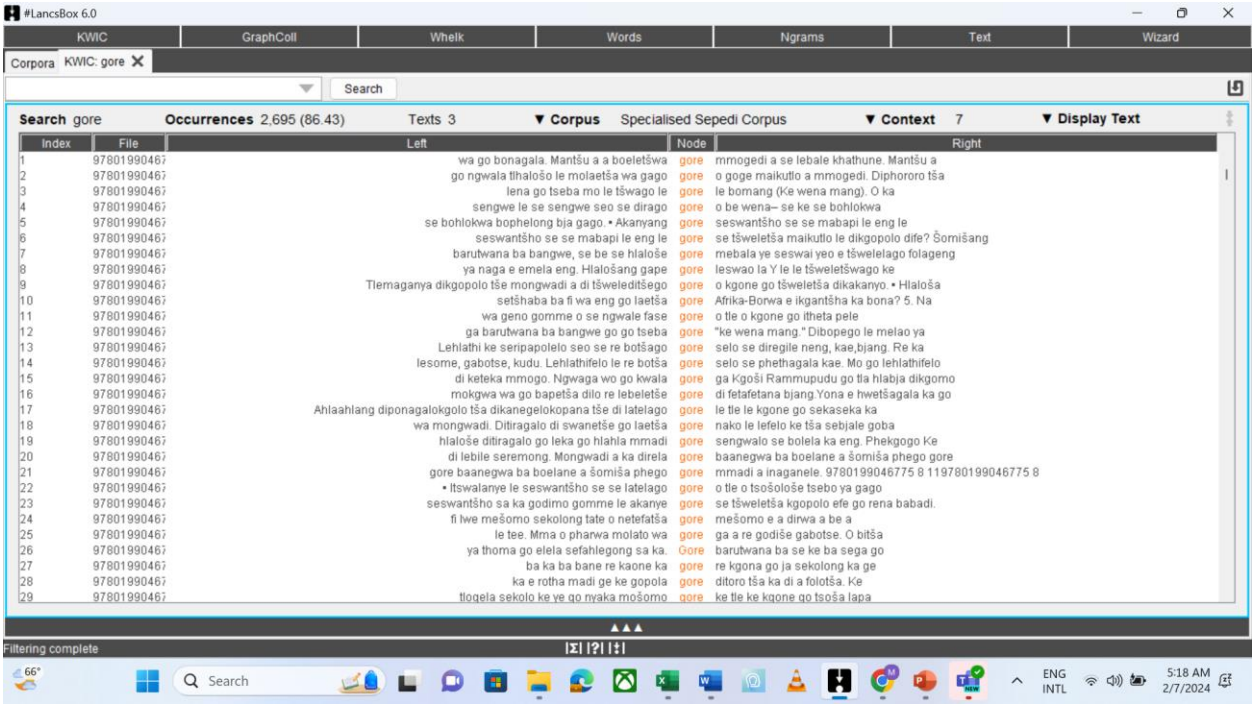


***Figure 4. 22**: KWIC tool displaying gore as a search word in SSC*

The 100 KWIC lines reveal that *gore* appears 94 times positioned between clauses with no comma usage (see **Figure 4.23**), four times positioned between clauses with comma usage before (see **Figure 4.24**) and twice at the beginning of a sentence (see **Figure 4.25**).

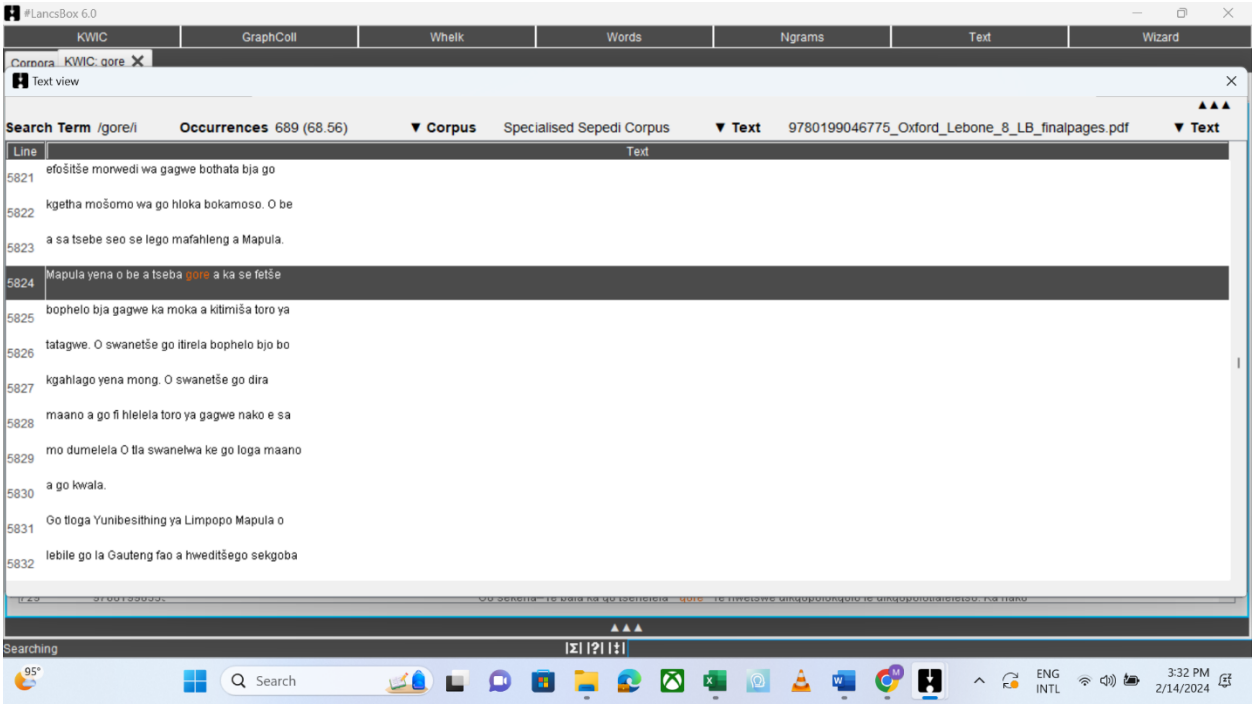


***Figure 4. 23**: The conjunction gore positioned between clauses with no comma usage in the SSC*

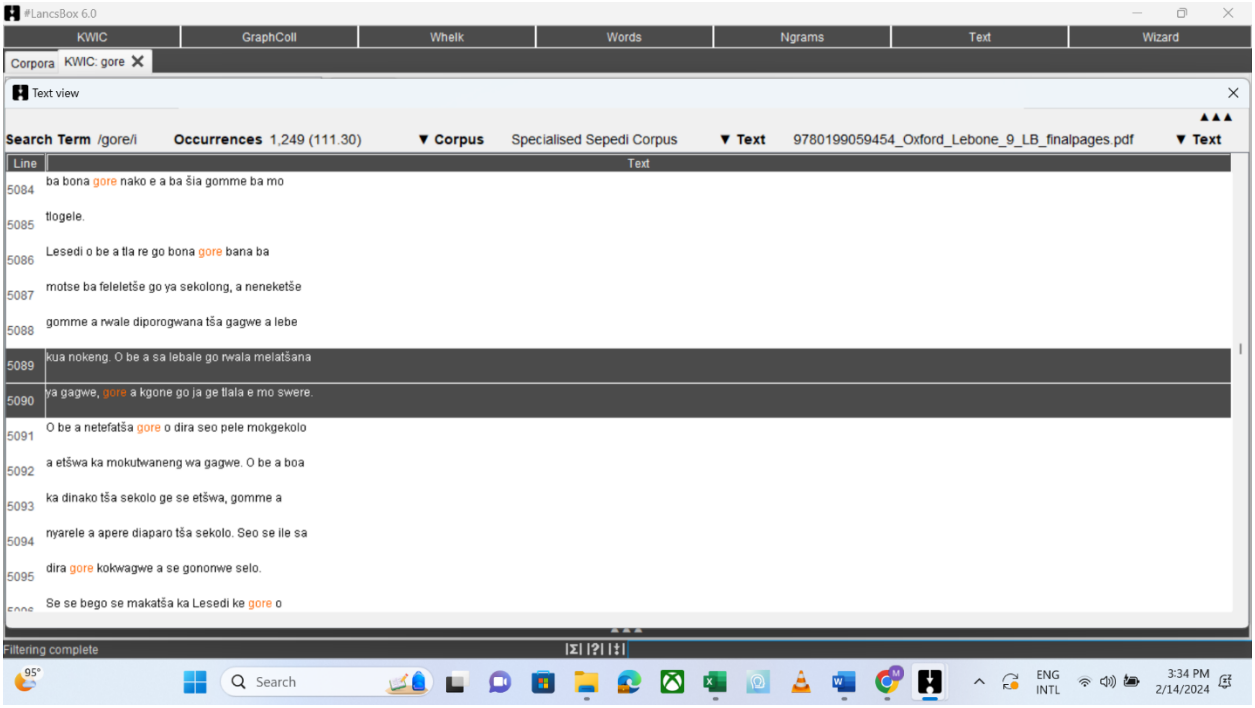


***Figure 4. 24**: The conjunction gore positioned between clauses with comma usage before in the SSC*

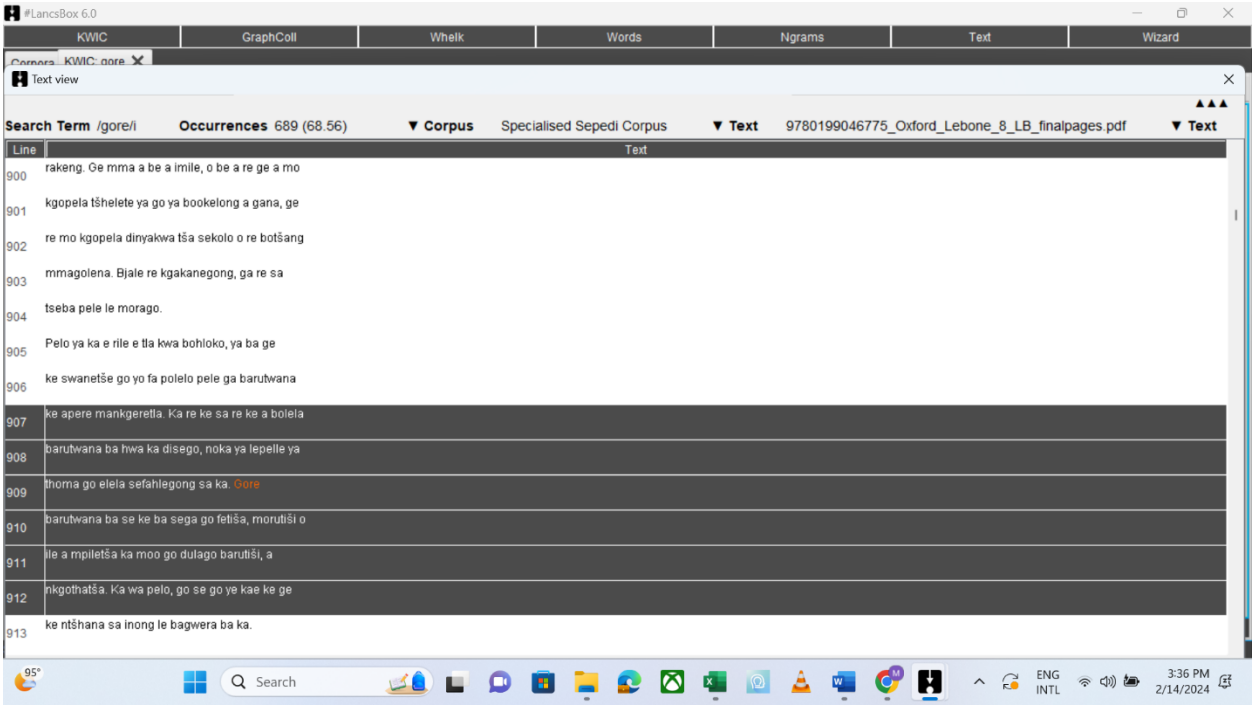


***Figure 4. 25**: The conjunction gore positioned at beginning of a sentence in the SSC*

These results clearly show that the use of *gore* positioned between clauses with no comma usage is more frequent in the SSC than its use with comma before and positioned at the beginning of a sentence. What this entails in terms of Generative Grammar model is the usage of the conjunction *gore* positioned between clauses with no comma usage, its usage positioned between clauses with comma before and its usage positioned at the beginning of a sentences are part of the innate laws and rules engrained in the minds of Sepedi speakers concerning how the conjunction is used.

When the conjunction *gore* is used in the SSC, it expresses cause and effect and can be translated into ‘that’, ‘so that’ and ‘in order’ in English. Consider the examples below from the sampled concordance lines:

- *Ngwaga wo go kwagala **gore** ga Kgoši Rammupudu go tla hlabja dikgomo tše legolo ge baswa ba tla be ba rutwa meetlo ya seetšo sa Bapedi.* ‘This year it was

heard **that** at King Rammupudu's place one hundred cows will be slaughtered during inculcating youth the *Bapedi* cultural customs'.

- *Ikgopotšeng ka diponagalo tša athikele ya kuranta go tšwa thutong ya go feta pele o ka bala mohlala wa athikele ye e latelago,* ***gore*** *ka morago o kgone go araba dipotšišo go laetša kwešišo.* 'Remind yourself of the elements of a newspaper article from the previous lesson before reading example of the following article **so that** later you can show your understanding by answering questions that follow'.
- ***Gore*** *naga e kgone go godiša ekonomi, e swanetše go ba le baeng go tšwa dinageng tša ka ntle.* '**In order** for a country to grow its economy, it must have visitors from foreign countries'.

From the above examples, it is clear that the Sepedi conjunction *gore can* be translated as 'that', 'so that' and 'in order' in English. Let us move on to the sixth conjunction that was sampled, namely *mola*.

### *4.5.6 The conjunction: mola*

The conjunction '*mola'* occurs 170 times in the Specialised Sepedi Corpus (see **Figure 4.26**).

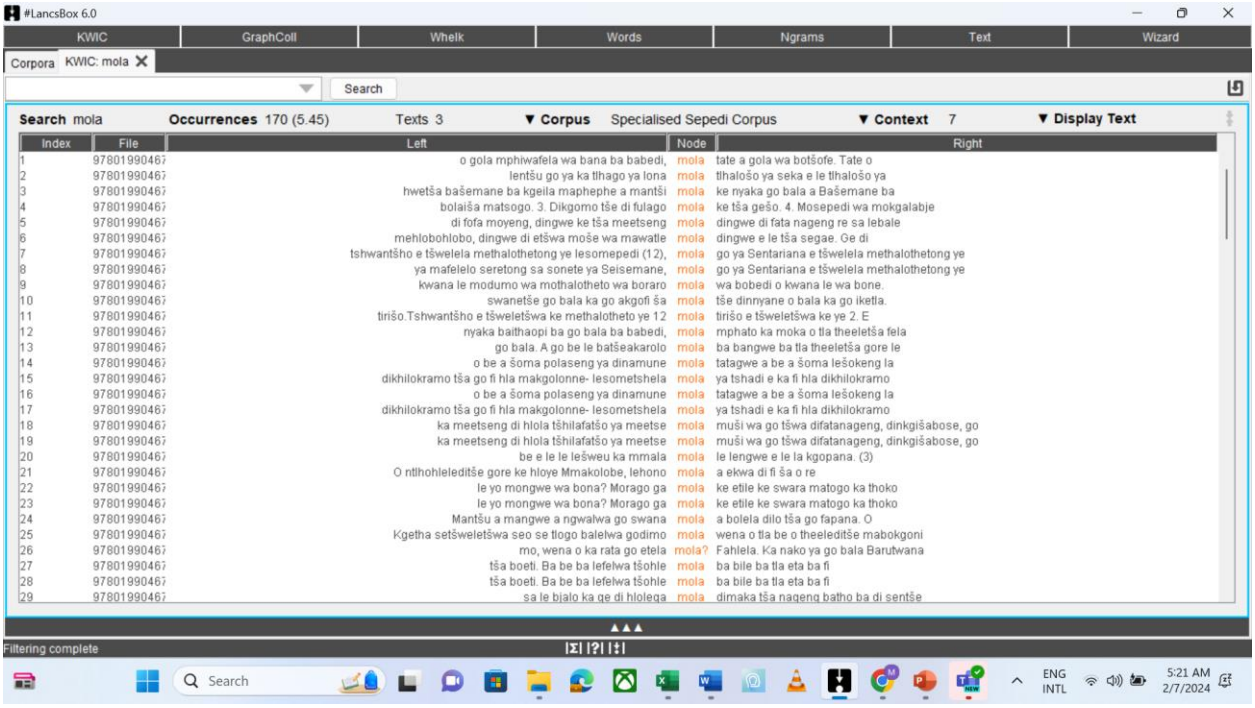


***Figure 4. 26**: KWIC tool displaying mola as a search word in the SSC*

The sampled 100 KWIC lines show that *mola* appears 79 times positioned between clauses with no comma usage (see **Figure 4.27**),16 times positioned between clauses with comma usage before (see **Figure 4.28**) and five times positioned at the beginning of a sentence (see **Figure 4.29**).

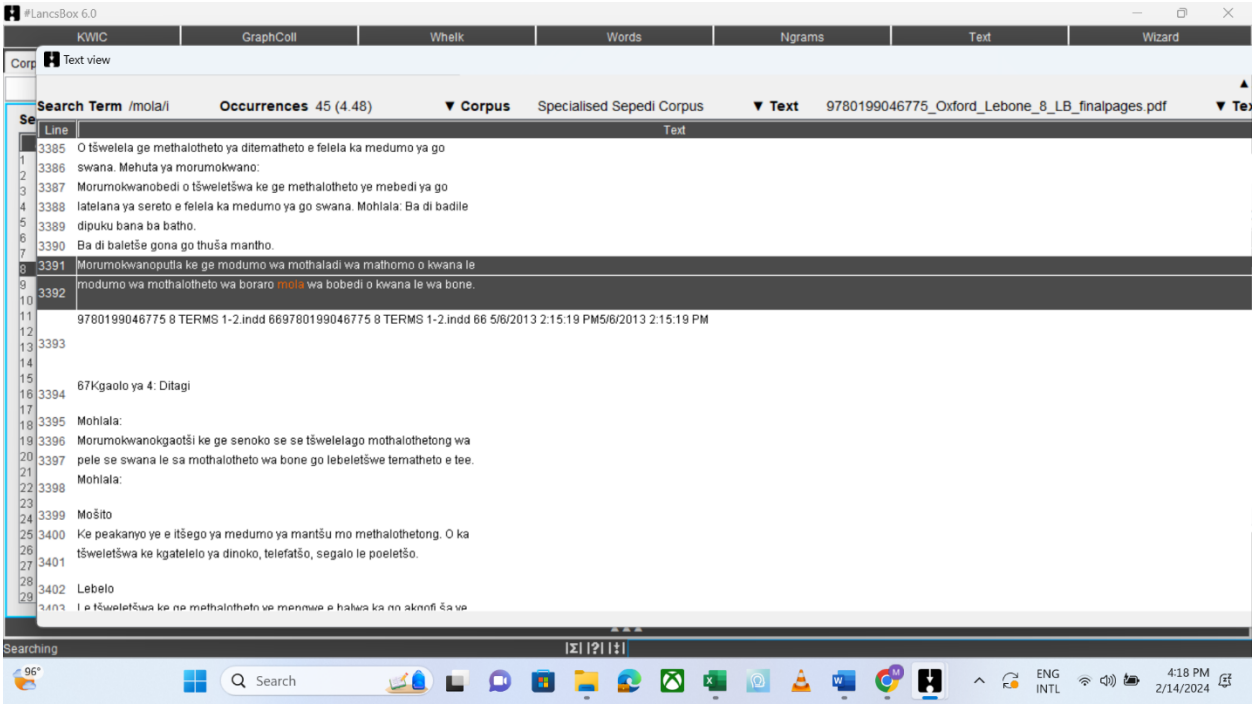


***Figure 4. 27**: The conjunction mola positioned between clauses with no comma usage in the SSC*

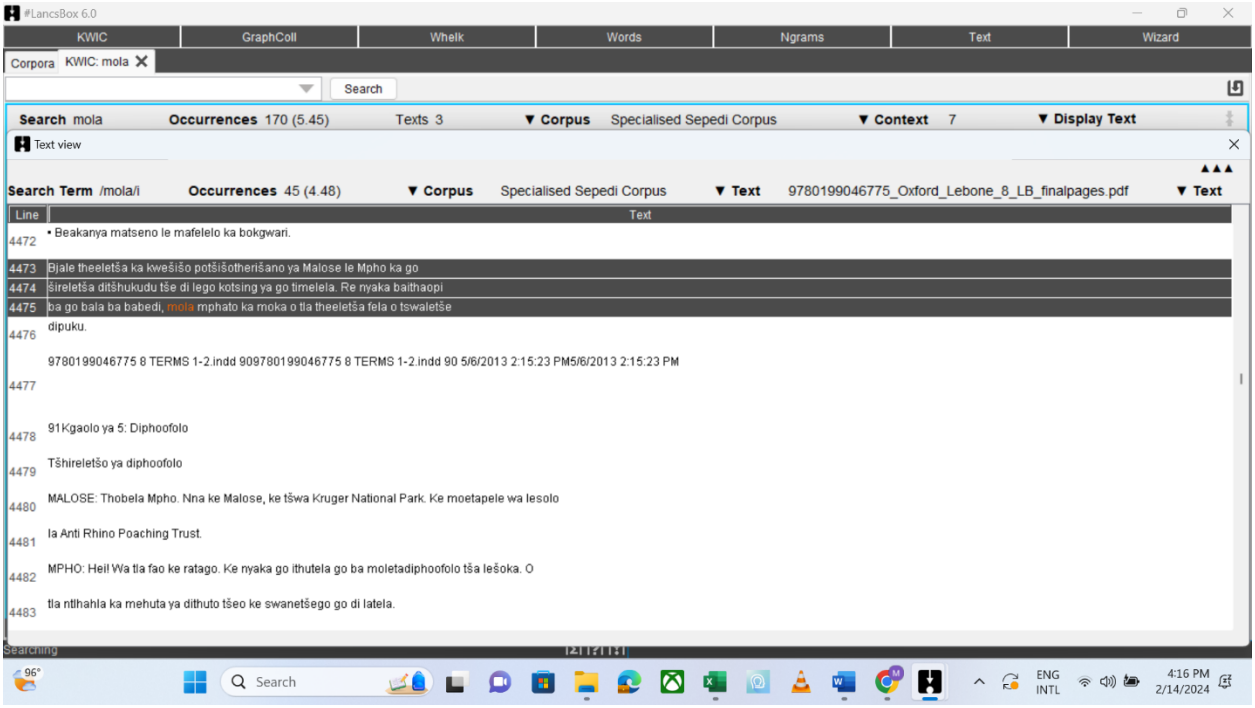


***Figure 4. 28**: The conjunction mola positioned between clauses with comma usage before in the SSC*

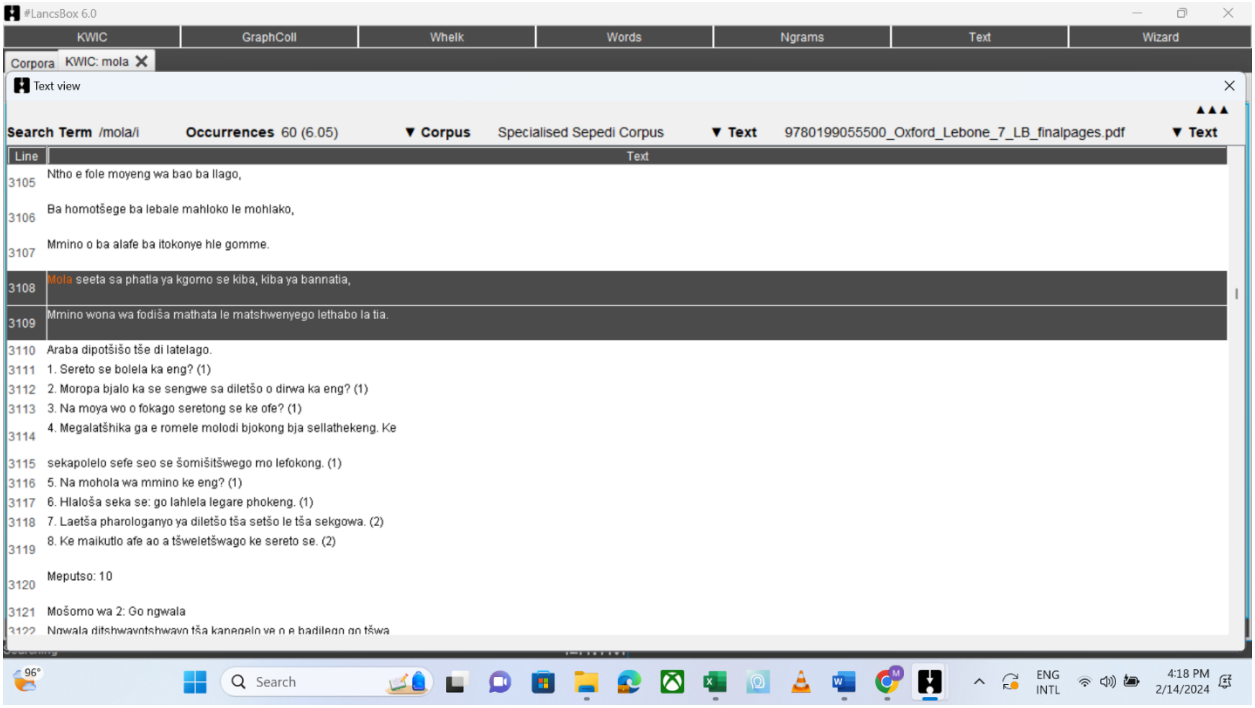


***Figure 4. 29**: The conjunction mola positioned at the beginning of a sentence in the SSC*

The results above clearly show that the use of *mola* positioned between clauses with no comma usage is more predominant than its usage with a comma before and at the beginning of a sentence. In terms of Generative Grammar theory, these findings entail that the conditions of the usage of the conjunction *mola* positioned between clauses with no comma usage, its usage with a comma before, as well as its usage at the beginning of a sentence form part of the mental faculty housing laws and rules on the grammatical usage of this conjunction.

When the conjunction *mola* is used in the SSC, it expresses:

- a contrast/change and can be translated as ‘whereas’, and
- an alternative situation or contrast which can be translated as ‘while’ in English.

Consider the following examples gleaned from the SSC:

- *Sekolo se ile sa tšea sephetho sa gore ba tla tšea Lesedi a tle sekolong gomme a rute bana ba sekolo go betla le go dira tše dingwe gomme bana ba sekolo le bona ba tla mo thuša ka tša sekolo **mola** barutiši ba tla reka ditšweletšwa tša gagwe*. 'The school took a decision that they will take Lesedi to come to school and teach learners carving and doing others but in return, learners will help him/her with school work **while** teachers will buy his/her products'.

- *Mma o gola mphiwafela wa bana babedi, **mola** tate a gola wa batšofe.* 'Mother is receiving social grant for two kids, **whereas** father receives pensioners social grant'.

- *Ba šegofaditšwe ka barwa ba babedi le dikgarebe tše tharo. **Mola** rakgolo a sa phela, go kwagala gore ba be ba fela ba etela Malawi.* 'They are blessed with two sons and three daughters. **While** grandfather was still alive, it is said that they used to visit Malawi'.

The above concordance lines clearly confirm that the Sepedi conjunction *mola* can be translated as 'whereas' and 'while' in English.

The insights from the SSC on the usage and meaning of Sepedi conjunctions have been exhaustively discussed. In the following section, the authentic occurrences of the same conjunctions in the GSC is investigated.

### 4.6 Syntactic and semantic features of Sepedi conjunctions in the GSC

The aim of this section is to discuss the syntactic and semantic features of the same conjunctions discussed above, but now in the GSC. The section will also attempt to compare the usage and meaning of these conjunctions as observed in the SSC and GSC. Let us commence with the same conjunction which opened our discussion above, namely *ebile*.

#### *4.6.1 The conjunction: ebile*

Using *ebile* as a search word in the GSC, the results produced 2,426 KWIC lines, as displayed in **Figure 4.30** below.

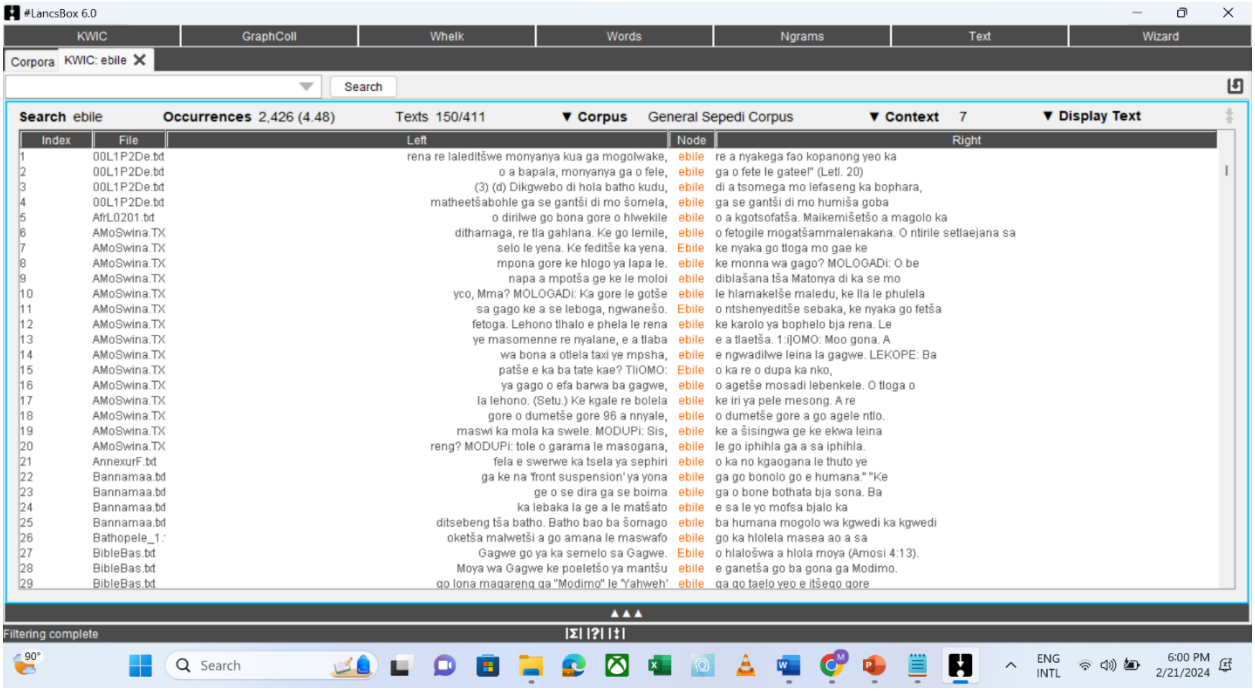


***Figure 4: 30**: KWIC tool displaying ebile as a search word in the GSC*

The 100 KWIC lines from GSC show that *ebile* appears 49 times positioned between clauses with no comma usage (see **Figure 4.31**), 26 times positioned between clauses with comma usage before (see **Figure 4.32**) and 25 times positioned at the beginning of a sentence (see **Figure 4.33**).

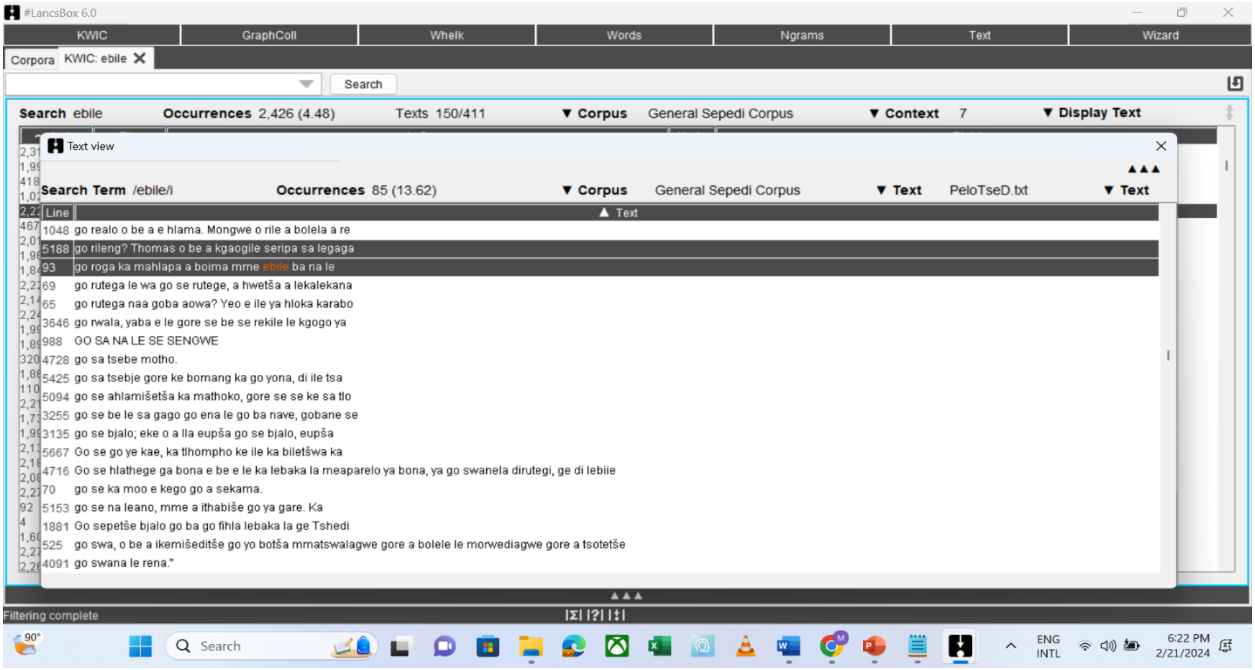


***Figure 4: 31**: Conjunction ebile positioned between clauses with no comma usage in the GSC*

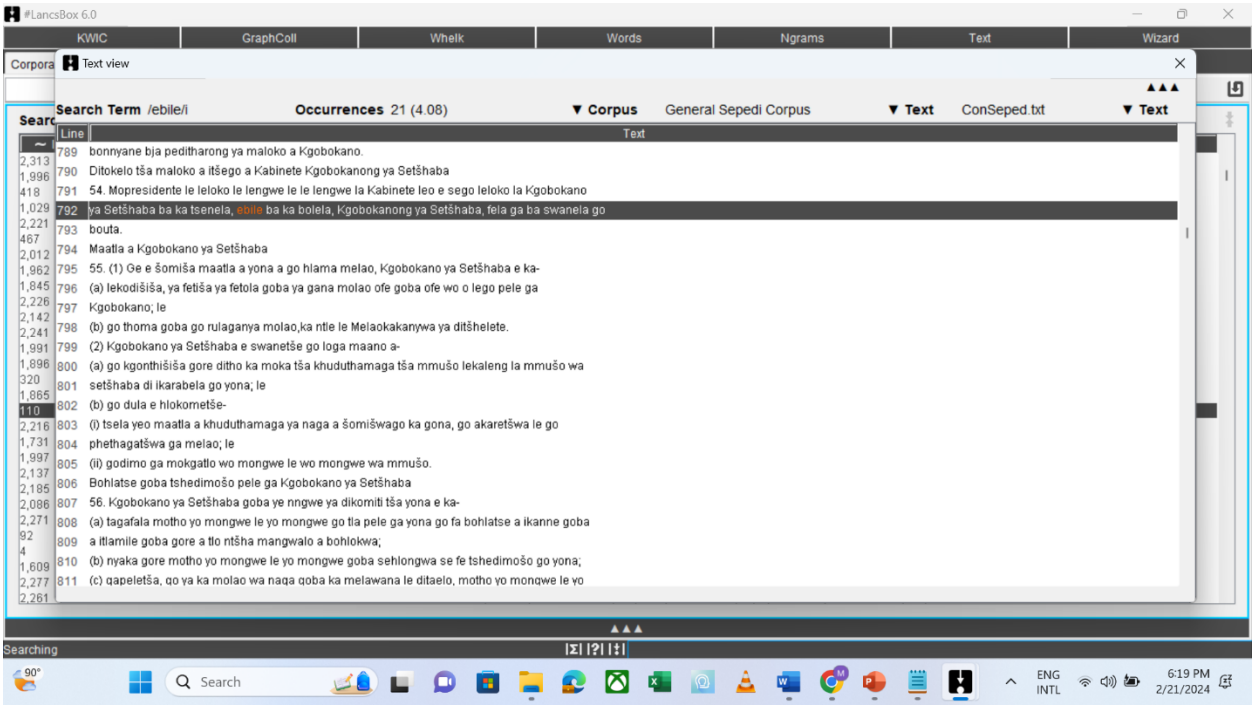


***Figure 4. 32**: Conjunction ebile positioned between clauses with comma usage before in the GSC*

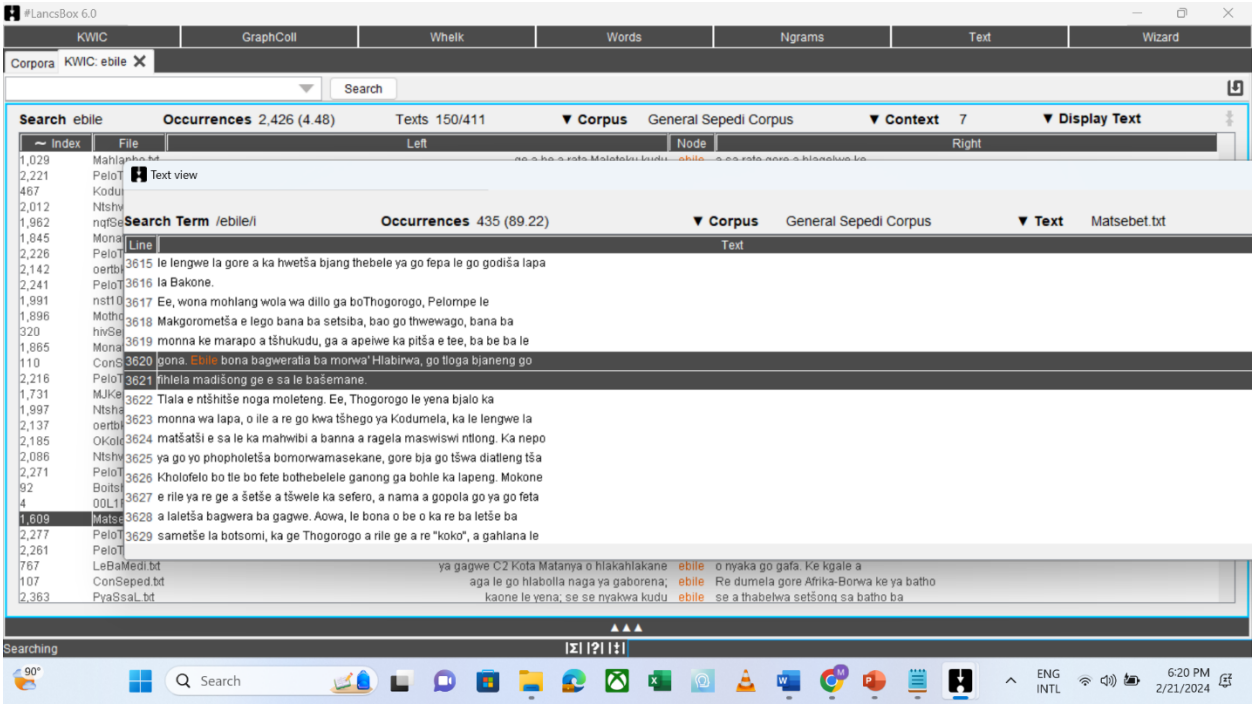


***Figure 4. 33**: Conjunction ebile positioned at the beginning of a sentence in the GSC*

The results show that the use of *ebile* positioned between the clauses with no comma usage is more frequent than its use with comma before and at the beginning of a sentence. The results simply demonstrate that the usage of the conjunction positioned between clauses with no comma usage is predominant as compared to its usage with comma before and positioned at the beginning of a sentence.

These results coincide with those found concerning the SSC. The results from both corpora show predominance of the usage of the conjunction positioned between clauses with no comma usage as compared to its usage positioned between clauses with comma usage before as well as positioned at the beginning of a sentence.

The semantic features of the conjunction *ebile* were again checked by going through the concordance lines. When the conjunction is used in the GSC, it expresses the following senses:

- expresses comparison which can be translated as ‘also’, and
- expresses addition which can be translated as ‘and’.

Let us consider the following examples gleaned from the GSC concordance lines:

- "*O thoma maaka, Sello. "Mpule o mo tsena ganong a kgohlotše mahlo ka makalo. "Ke neng moo o ilego wa nngwalela lengwalo ka se go fetole? A ke re le maloba mo* ***ebile*** *ke go rometše seswantšho seo o se kgopetšego? Bjale o bolela ka eng?"* 'He starts lying, Sello. "Mpule interrupts him speaking while gawking. "When did I not respond your letter to me? Isn't it that a day before yesterday I ***also*** sent you the picture you requested? Then what are you talking about?'

- *"Aowa, melaetša ke tlo e fihliša Mokone." Mopostola o ntšhitše leo,* ***ebile*** *a etšwa ka sefero*. 'No, I will deliver the message Mokone." The Apostle said so, ***and*** came out from the front entrance'.

- *Thokgorogo, bjalo nna ke romilwe go go tsebiša gore o tsebe gore fa ga ba rate bašomi ba matepe.* ***Ebile*** *godimo ga fao ba re o tsebe gore ge o ka hlaba kolobe gape o feditše, gobane o tlo ba o le thakadu ge e tennwe ke molete.* 'Thokgorokgo. I am sent here to inform you that you must know that moody workers are unwanted here. **And** on top of that, being absent again from work will mean you are resigning'.

From the above example, it is clear that the Sepedi conjunction *ebile* can be translated as 'also' and 'and'. In the case of the SSC, the findings indicated that the conjunction is used to express: 1) an addition which can be translated as 'furthermore', and 2) a contrast or change which can be translated as 'even'. Although 'also', 'and' and 'furthermore' are different terms, it cannot be denied that they are closely related terms. Therefore, it can be inferred that the conjunction conveys more or less similar meanings in the case of these terms. The only sense which shows huge discrepancy in the meaning of this conjunction in the two corpora is 'contrast or change', which is found in the SSC and not in its general counterpart.

It is perhaps worthwhile to now move on to the second conjunction that was sampled, namely *ge*.

### *4.6.2 The conjunction: ge*

Using *ge*. as a search word in the GSC produced 74,694 KWIC lines (see **Figure 4.34**).

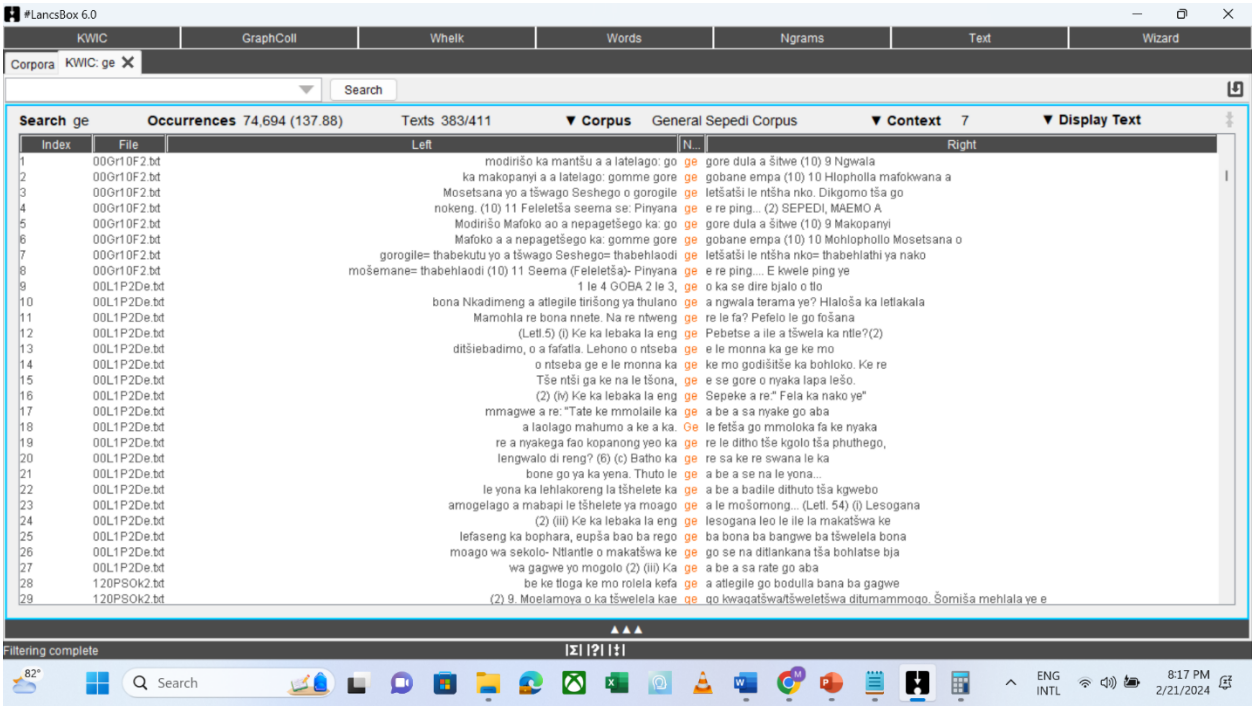


***Figure 4. 34**: KWIC tool displaying ge as a search word in the GSC*

In the 100 KWIC lines sampled, the results showed that ge appears 60 times positioned between clauses with no comma usage (see **Figure 4.35**), 13 times positioned between clauses with comma usage before (see **Figure 4.36**) and 27 times positioned at the beginning of a sentence (see **Figure 4.37**).

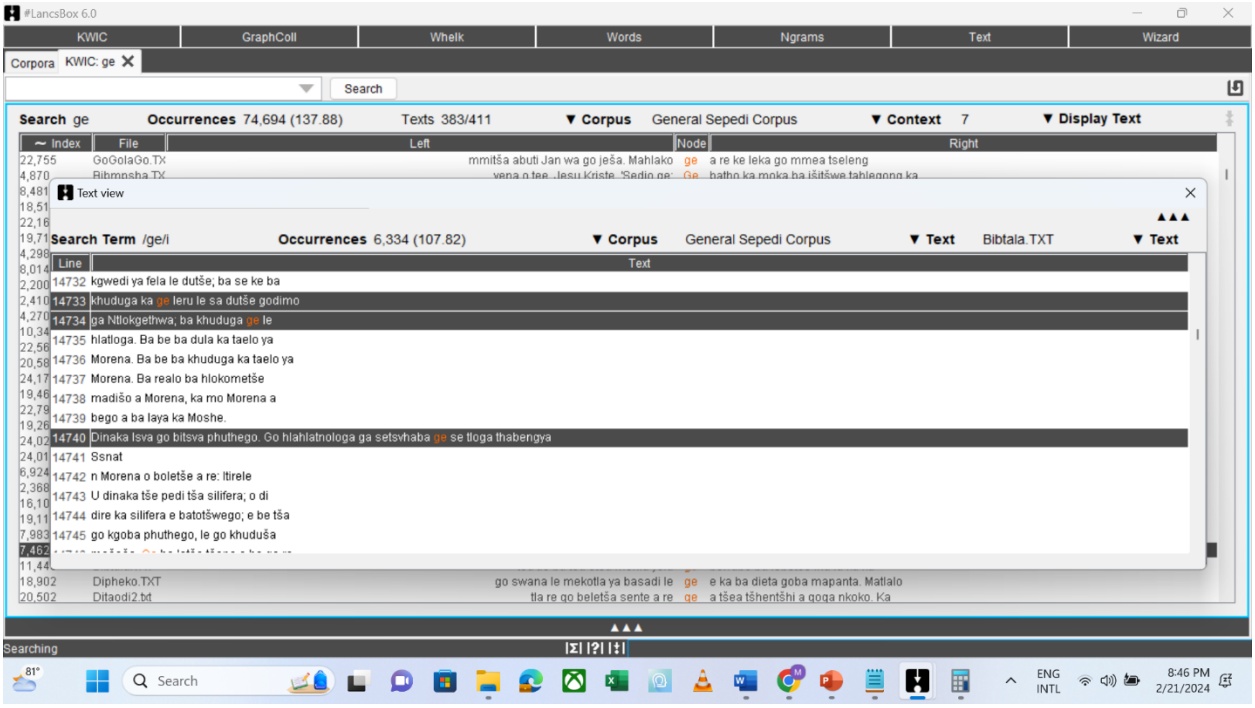


***Figure 4. 35**: Conjunction ge positioned between clauses with no comma usage in the GSC*

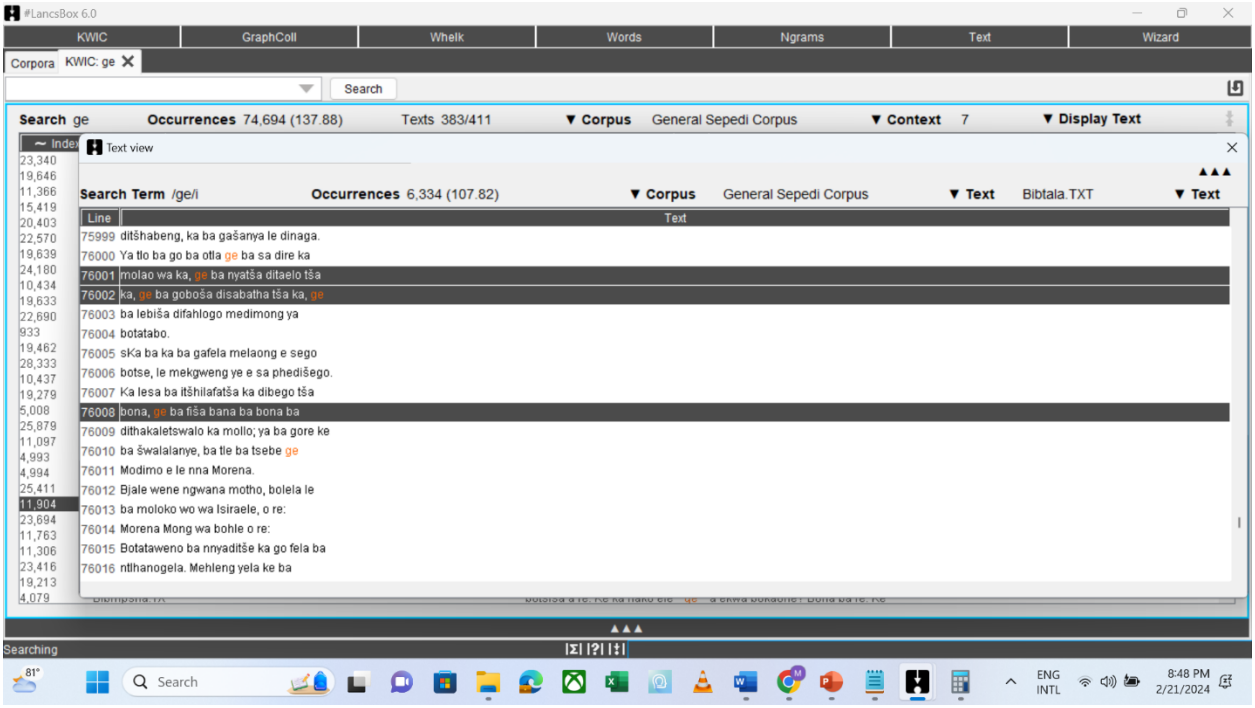


***Figure 4. 36**: Conjunction ge positioned between clauses with comma usage before in the GSC*

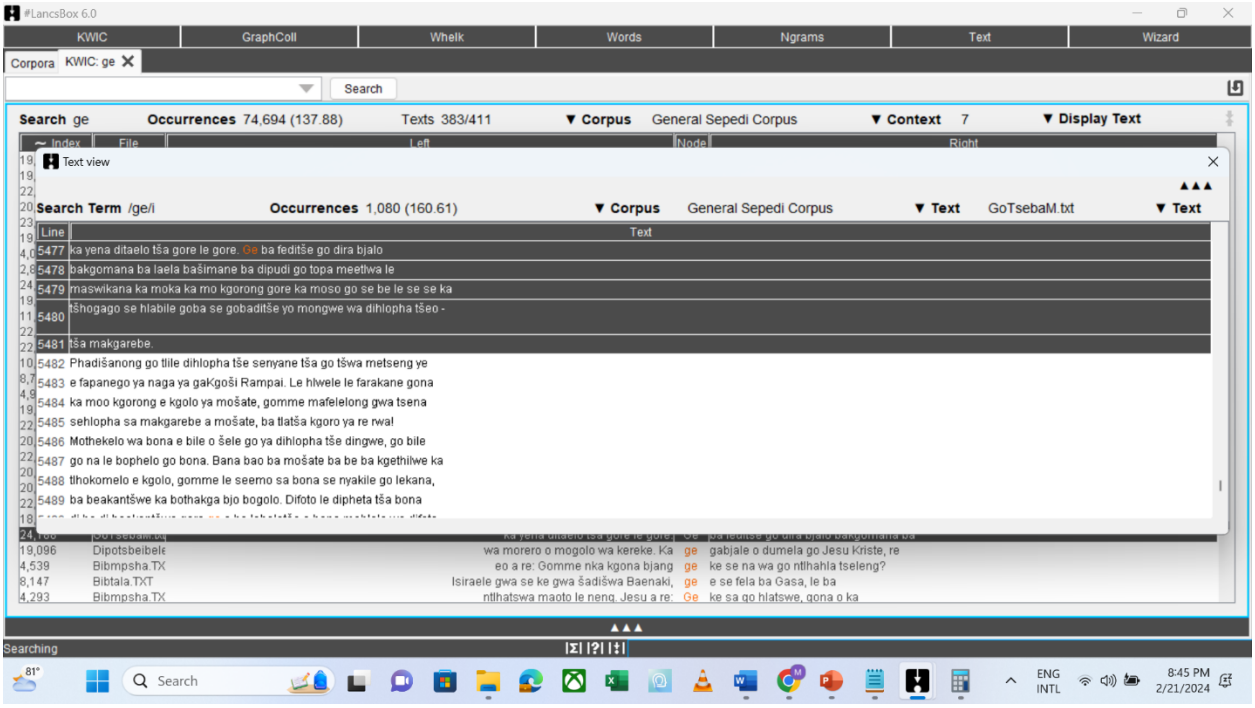


***Figure 4. 37**: Conjunction ge positioned at the beginning of a sentence in the GSC*

These results clearly demonstrate that the use of *ge* positioned between clauses with no comma usage is more frequent in the GSC than its use with comma before as well as at the beginning of a sentence.

These results from the GSC seem to suggest a different trend in terms of the usage of this conjunction. In the SSC, there was a preference of usage of the conjunction positioned at the beginning of a sentence, whereas in the GSC the dominant preference is the usage of the conjunction positioned between clauses with no comma usage. Furthermore, from the SSC, it was discovered that one of the syntactic features of the conjunction is that it is also positioned between the clauses with comma usage after. However, in the sample of concordance lines from GSC, there were no instances in which the comma was used after the conjunction.

Concerning semantic features, the conjunction has been found to express the following senses in the GSC:

- expresses a logical order or time which can be translated as ‘when’, and

- shows a condition which can be translated as ‘if’ and ‘even if’.

Consider the examples below from the sampled concordance lines:

- *Ka la ka moswane **ge** go esa, banna bao ba fiwa tsela, bona le dipokolo tša bona. **Ge** ba tšwile mo motseng, ba sešo ba ya kgole, Josefa a botša molaki wa gagwe a re: Nanoga o latele banna bao, mme **ge** o ba hweditše o ba botšiše o re: Le reng le lefetša botse ka bobe?* ‘Tomorrow **when** the morning comes out, men and their donkeys will be helped to locate a route. **When** they are out of the village, before going far, Joseph will say to his servant: quickly follow those men, and **when** he finds them he will ask: why are you repaying good with evil?’

- *‘Ba ba lokago pele ga Modimo ga se bao ba kwa molao ka ditse; ba go tlo thwe ba lokile ke badiri ba molao.’ Gobane bantle ba ba se nago molao wo, **ge** ka noši ba dira tše di bolellwago ke wona molao wo, ke go re mo go bona molao o gona, le **ge** ba se ba wona molao.* ‘The righteous before God are not the hearers of the law; the righteous are those who enforce the law. Because foreigners without this law, **if** personally they do as the law says, it means there is law amongst them, **even if** the law is not for them’.

From the above examples, it becomes apparent that the Sepedi conjunction *ge* can be translated as ‘when’, ‘if’ and ‘even if’ in English.

When comparing these semantic features to those obtained in the SSC, it appears that only the sense ‘when’ is common in the two corpora. To highlight again, the SSC results indicated that senses carried by the conjunction include ‘after’, ‘when’ and ‘then’. Therefore, the senses ‘if’ and ‘even if’ found in the GSC are absent in the SSC, and the meanings ‘after’ and ‘then’ observed in the SSC could not be traced in the GSC. Let us now proceed to the third conjunction that was sampled, namely *goba*.

### *4.6.3 The conjunction: goba*

Using *goba* as a search word in the GSC produced 18,919 KWIC lines, as displayed in **Figure 4.38**.

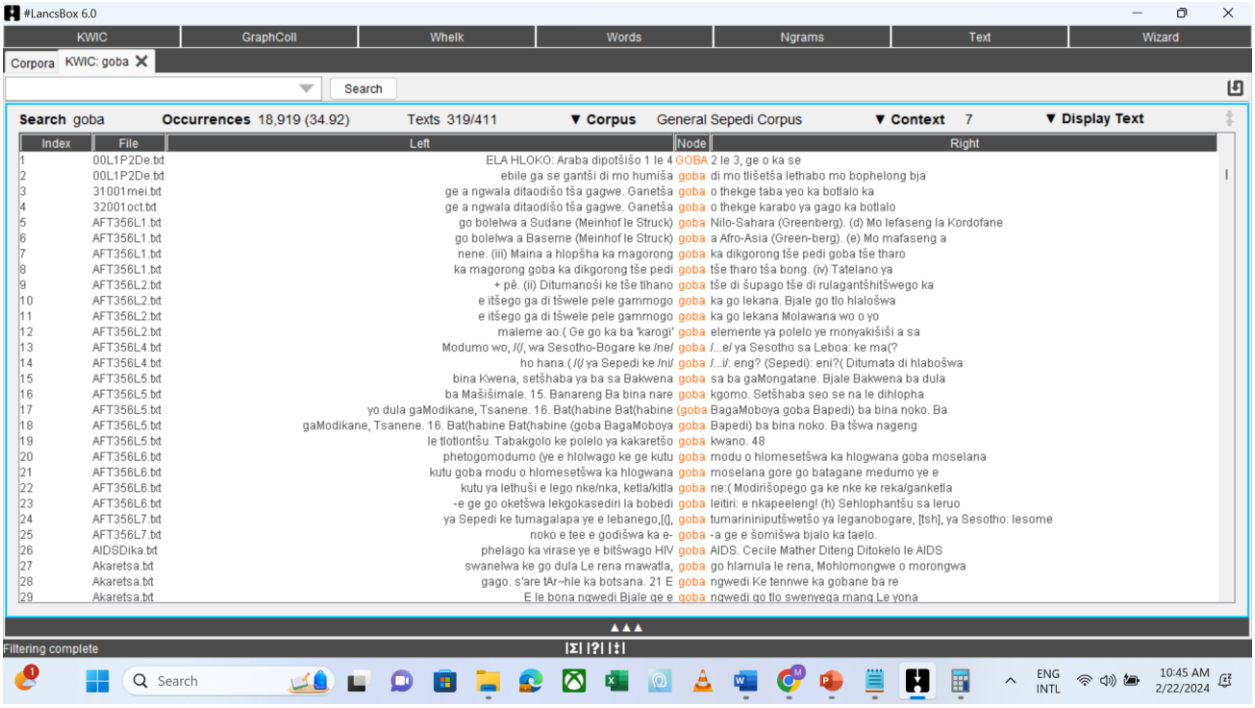


***Figure 4. 38**: KWIC tool displaying goba as a search word in the GSC*

The sampled 100 concordance lines demonstrate that *goba* appears 84 times positioned between clauses with no comma usage (see **Figure 4.39**), 11 times positioned between clauses with comma usage before (see **Figure 4.40**) and five times positioned at the beginning of a sentence (see **Figure 4.41**).

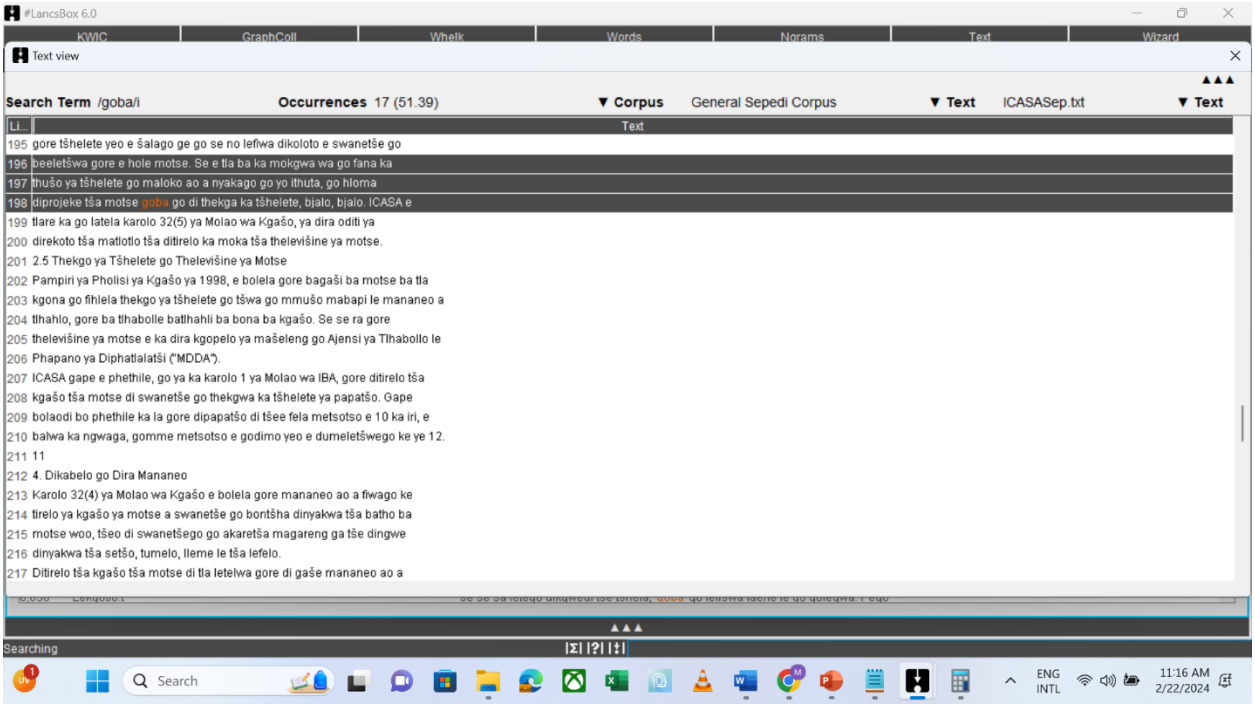


***Figure 4. 39**: The conjunction goba positioned between clauses with no comma usage in the GSC*

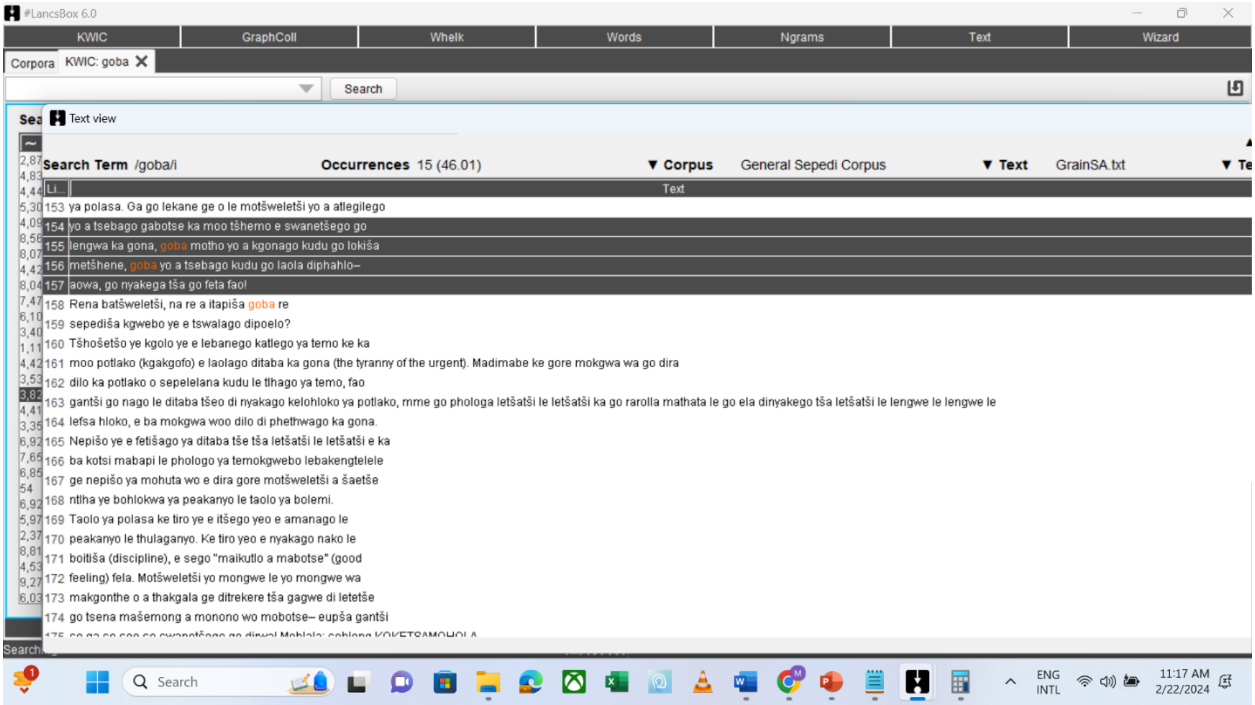


***Figure 4. 40**: The conjunction goba positioned between clauses with comma usage before in the GSC*

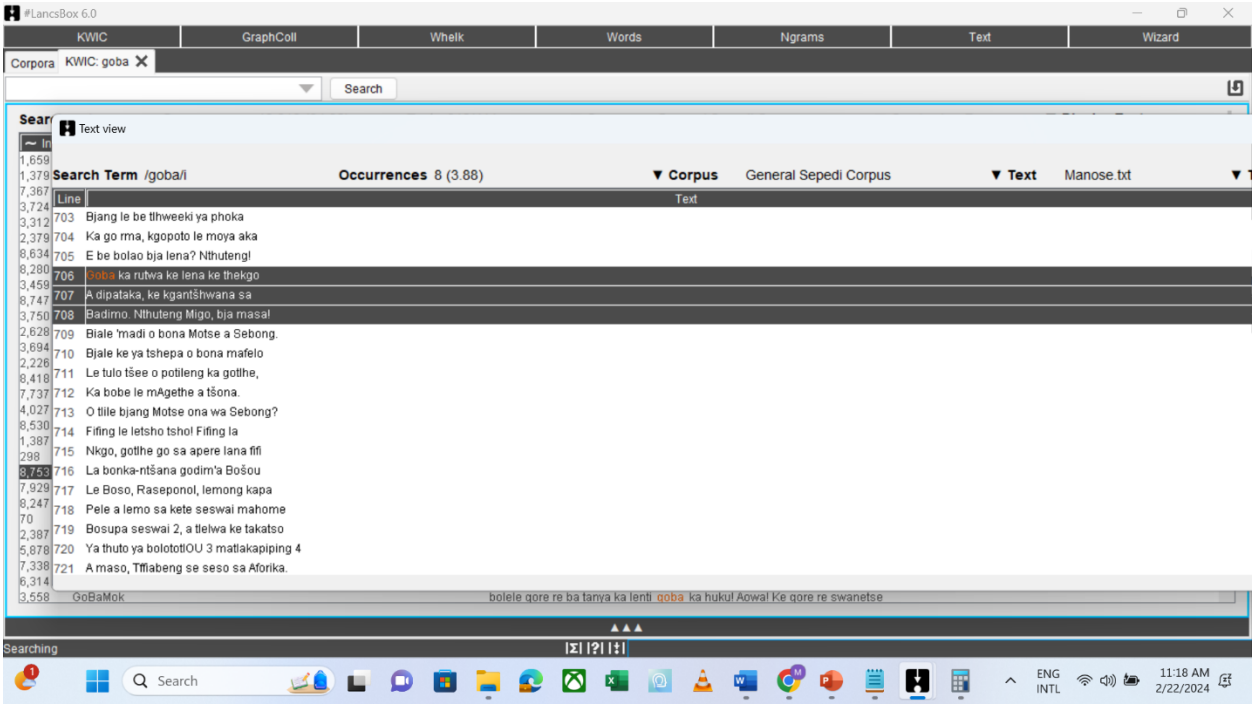


***Figure 4. 41**: The conjunction goba positioned at the beginning of a sentence in the GSC*

These results clearly indicate that the usage of *goba* positioned between clauses with no comma usage is more frequent in the GSC than its use with comma before and at the beginning of a sentence.

The findings reveal that the usage of the conjunction positioned between clauses with no comma usage is a common syntactic feature between the SSC and GSC. However, GSC results reveal an additional feature, namely the usage of the conjunction positioned at the beginning of a sentence. This feature was not present in the case of the SSC.

The results indicate that when the conjunction *goba* is used in the GSC, it expresses alternatives or choices which can be translated as 'or'. Consider the examples below from the excerpted concordance lines:

- *Go nyaka gore mohlankedi wa tshedimošo **goba** mohlankedi wa maleba wa setho sa setšhaba goba hlogo ya setho sa poraebete, go tšea magato a bjalo, **goba** go emiša go tšea magato a bjalo, go ya ka moo lekgotla le bonang go hlokega mo nakong yeo e laeditšwego ka taelo ya lekgotla.* 'It requires the information officer

**or** a relevant officer who is a member of the public or a private member leader, taking such steps, **or** to stop taking such steps, according to how the council sees a need as specified by the order of the council'.

- *A le re go tlo letla mohla a sekiša lena? **Goba** le re a ka forwa boka motho ge a forwa?* 'Are you saying it will be allowed the day he/she judges you? **Or** he/she can be deceived like when a person is deceived?'

From the above examples, it is apparent that the Sepedi conjunction *goba* can be translated as 'or' in English. This sense is common in both the SSC and GSC. Therefore, the findings have some convergence in the case of this conjunction, when considering its semantic features. Let us move on to the fourth conjunction that was sampled, namely *gomme*.

### *4.6.4 The conjunction: gomme*

The results of the KWIC tool reveal that the conjunction *gomme* has 13,891 KWIC lines in the General Sepedi Corpus, as shown in **Figure 4.42**.

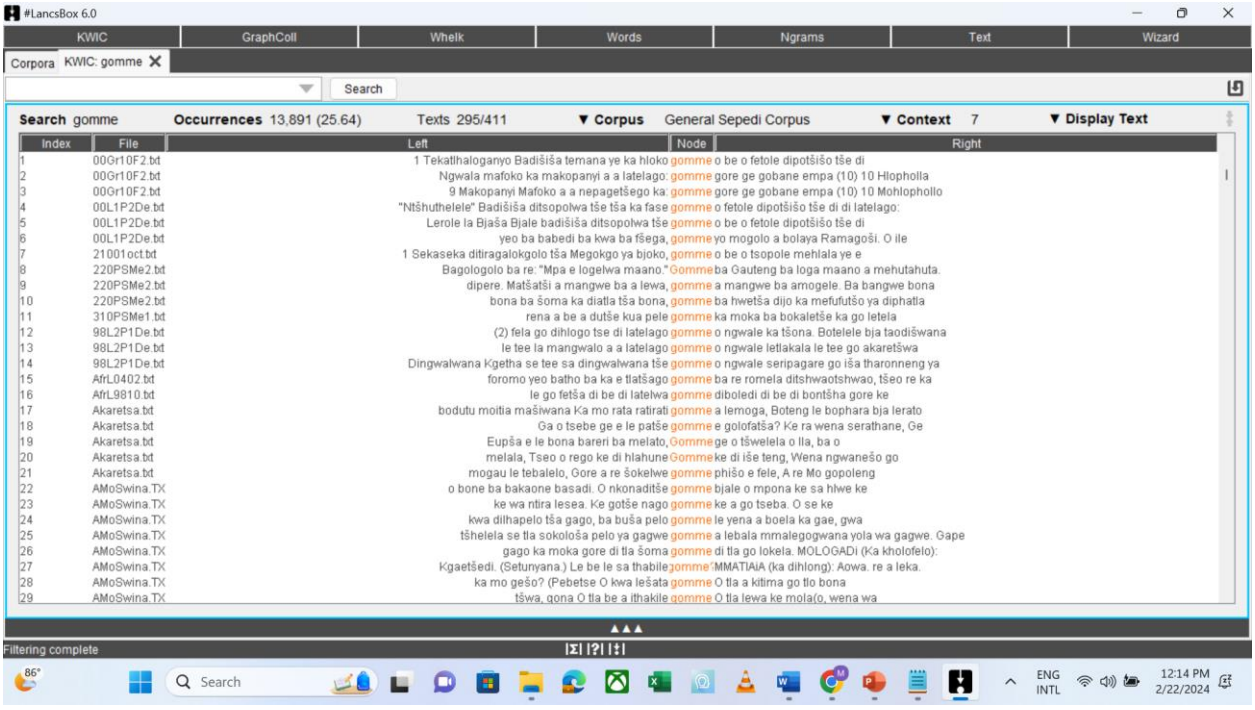


***Figure 4. 42**: KWIC tool displaying gomme as a search word in the GSC*

The sampled 100 KWIC lines show that *gomme* appears 43 times positioned between clauses with no comma usage (see **Figure 4.43**), 40 times positioned between clauses with comma usage before (see **Figure 4.44**) and 17 times positioned at the beginning of a sentence (see **Figure 4.45**).

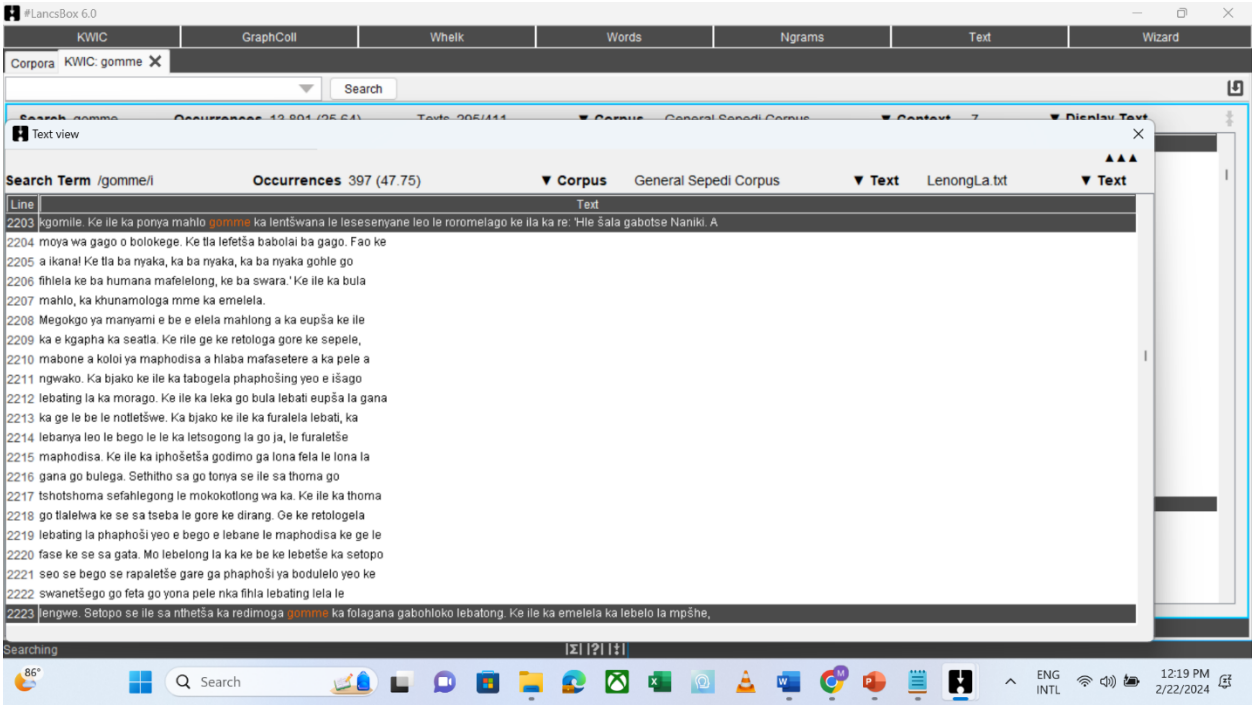


***Figure 4 43**: The conjunction gomme positioned between clauses with no comma usage in the GSC*

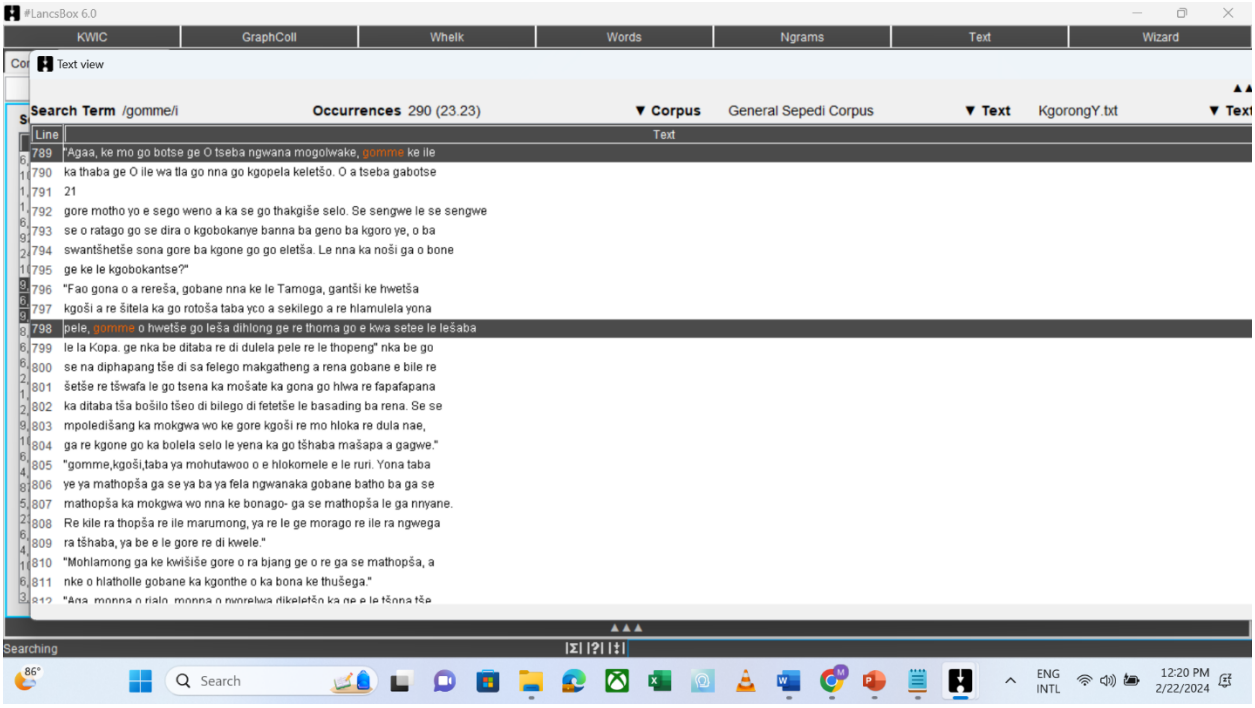


***Figure 4. 44**: The conjunction gomme positioned between clauses with comma usage before in the GSC*

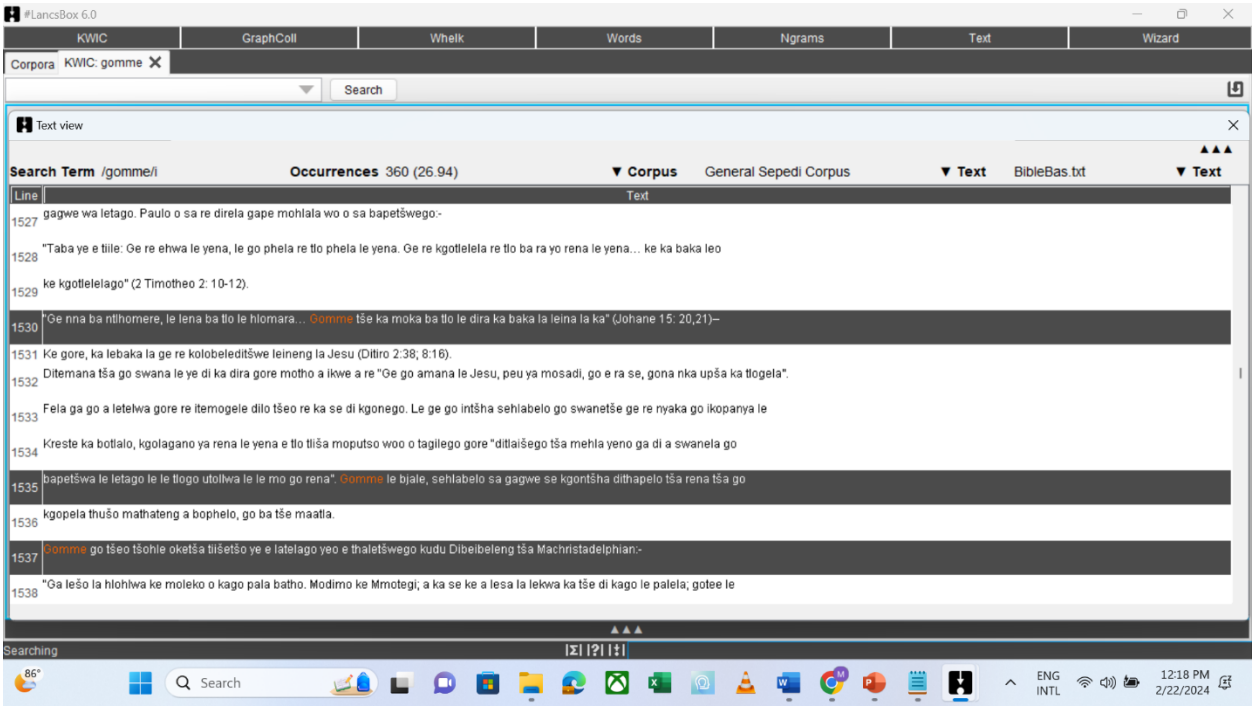


***Figure 4. 45**: The conjunction gomme positioned at the beginning of a sentence in the GSC*

These results show that the usage of *gomme* positioned between clauses with no comma (43 times) and with comma usage (40 times) are both more frequent, with only a difference of three usages in the GSC. The usage at the beginning of a sentence is somewhat infrequent.

The findings of the SSC also indicated predominance of the usage of the conjunction positioned between clauses with no comma. Therefore, this syntactic feature is common between the two corpora. However, in the GSC results, another syntactic feature was observed, that of usage of the conjunction positioned at the beginning of a sentence. This feature was not prevalent in the SSC and therefore marks the difference in the usage of the conjunction between the two corpora.

When the conjunction *gomme* is used in the GSC, it expresses addition, which can be translated into 'and' in English. Consider the following examples from the sampled concordance lines:

- *O na le mabjoko* ***gomme*** *o tseba taba yeo bjalo ka ge le nna ke e tseba.* 'You have brains **and** you know this matter as well as I do know it.'/ 'He has brains **and** he knows this matter as well as I do know it'.
- *Kgarebjana yeo, mmogo ba ilego ba šegofatšwa ka mošemane,* ***gomme*** *bjale lerato lapeng la Hlabirwa ya ba la go gahlanya dinoka le mawatle*. 'With the young lady, together they were blessed with a baby-boy, **and** now there was excessive love in Hlabirwa's home'.
- *Ba be ba no bonagala gore "ga se barutegi" (Ditiro 4:13) ge go tlišwa go tšeo.* ***Gomme*** *le go bareri ba bego ba rutegile (bjalo ka Paulo), bothata bja maleme bo be bo ba šitiša.* 'They just seemed to be "uneducated" (Acts 4:*13*) when those were brought. **And** even to preachers who were educated (like Paul), language barrier was a hindrance'.

From the above examples, it is clear that the Sepedi conjunction *gomme* can be translated as 'and' in English. These results coincide with those obtained in the SSC. Therefore, this is also another common semantic feature for this conjunction in the two corpora. Let us proceed to the fifth conjunction sampled, namely *gore*.

### *4.6.5 The conjunction: gore*

Using *gore* as a search word in the GSC, the results produced 63,444 KWIC lines, as shown in **Figure 4.46** below.


***Figure 4. 46**: KWIC tool displaying gore as a search word in the GSC*

The 100 KWIC lines sampled indicated that *gore* appears 90 times positioned between clauses with no comma usage (see **Figure 4.47**) and 10 times positioned between clauses with comma usage before (see **Figure 4.48**).

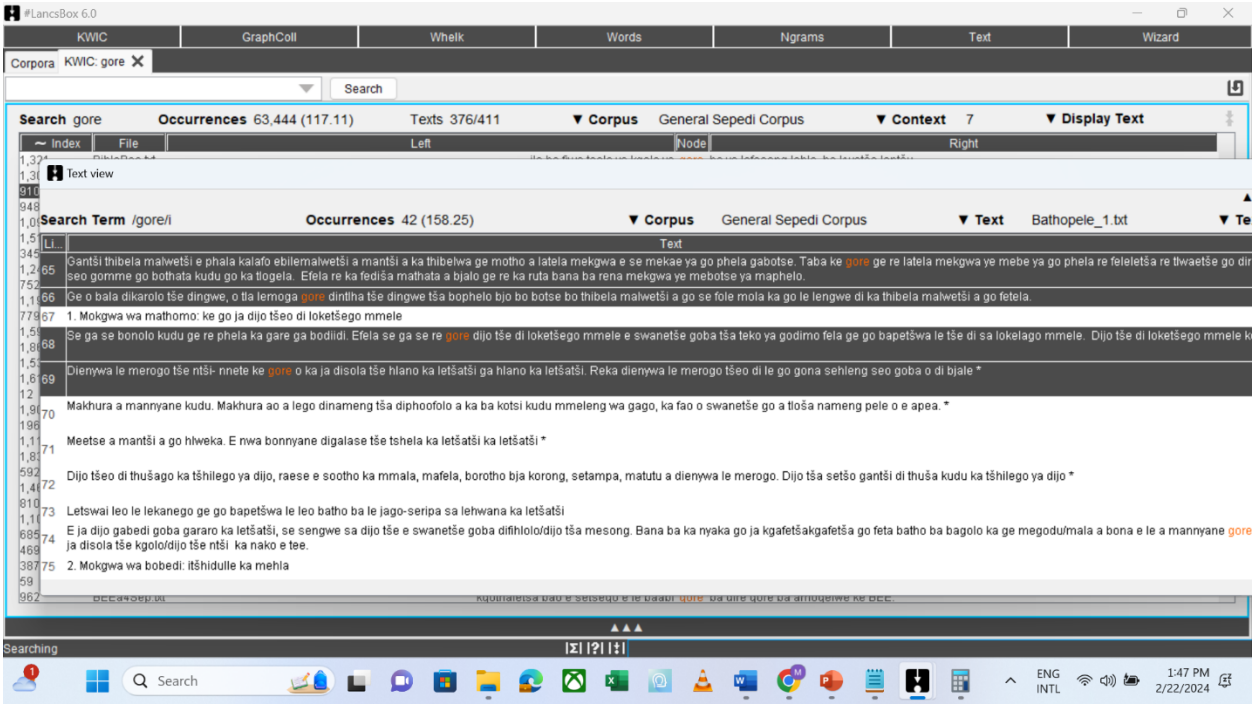


***Figure 4. 47**: The conjunction gore positioned between clauses with no comma usage in the GSC*

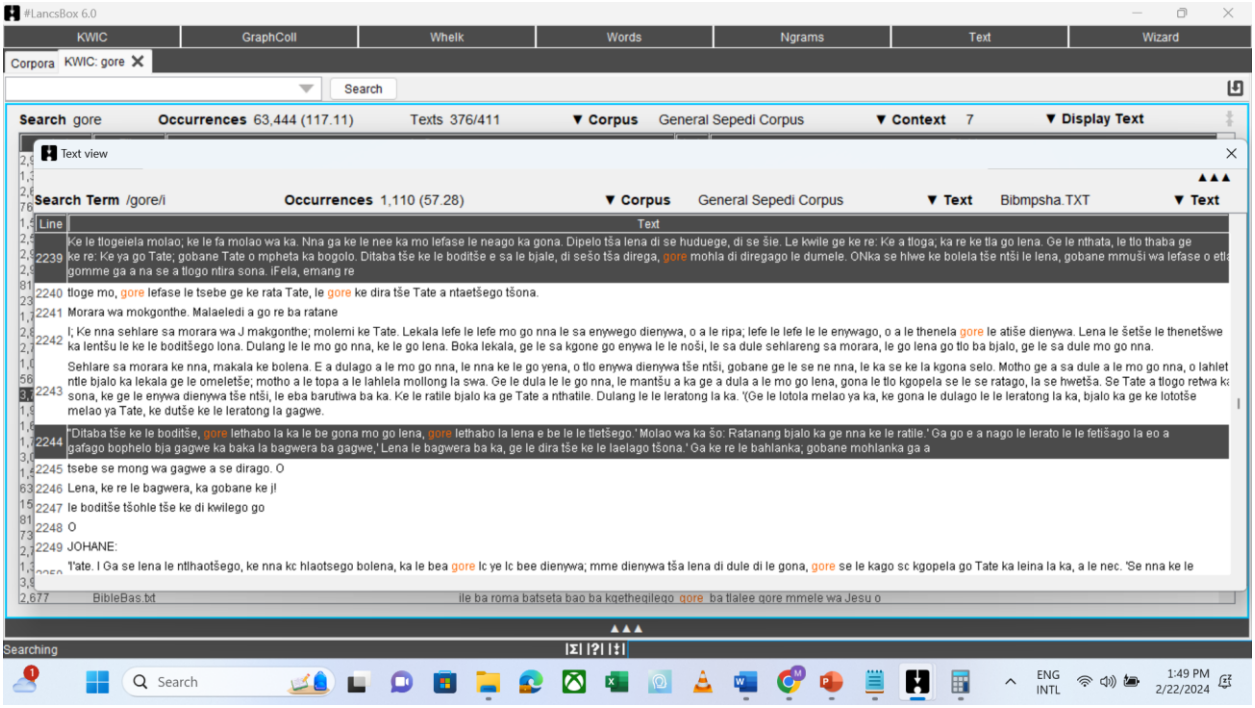


***Figure 4. 48**: The conjunction gore positioned between clauses with comma usage before in the GSC*

These results clearly show that the use of *gore* positioned between clauses with no comma usage is more frequent in the GSC than its usage with comma before.

Therefore, it is apparent that the feature of the usage of conjunction *gore* positioned between clauses with no comma usage is the preferred one in both the SSC and GSC. However, the usage of the conjunction differs between the two corpora in that the feature of usage of conjunction *gore* positioned at the beginning of a sentence, which had two occurrences in the SSC was not observed in the case of the GSC.

When the conjunction *gore* is used in the GSC, it expresses cause and effect, which can be translated as 'because', 'so that', and 'that' in English. Consider the examples below from the sampled concordance lines:

- *A a di kgone ka nnoši ka **gore** ga a motho yo a ka dumelago go dikišana motse le motho wa go swana le yena.* 'Let him do it alone **because** there is no one who can agree to share a home with someone like him'.
- *Ke gona le ka upša la mo lebalela, la mo homotša, **gore** motho eo a se tlo metšwa ke nyamo ge e mo golela. Ke ka baka leo le kgopelwago, **gore** le mo dire ka lerato.* 'Therefore you better forgive him, and warn him, **so that** he does not get swallowed up in sorrow when it grows for him. That is why you are requested, **that** you may serve him with love'.

From the above examples, it is clear that the Sepedi conjunction *gore* can be translated as 'because', 'so that', and 'that' in English. The senses 'that' and 'so that' could also be found in the SSC. However, the meaning 'in order', present in the SSC, could not be traced in the GSC. Furthermore, the sense 'because', prevalent in the GSC, could not be observed in the SSC. These two senses conveyed by the conjunction mark distinction in its usage in the two corpora.

Let us now consider the sixth and final conjunction that was sampled, namely *mola*.

### *4.6.5 The conjunction: mola*

Using *mola* as a search word in the GSC produced 6,604 KWIC lines (see **Figure 4.49**).

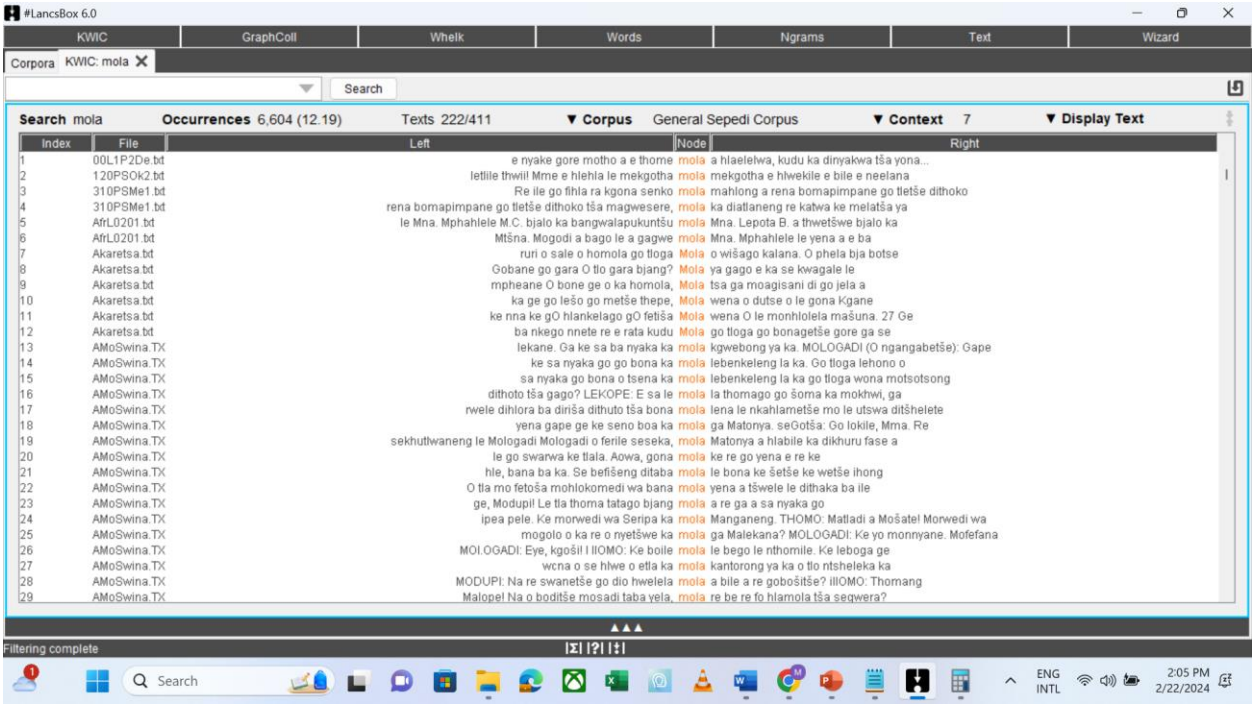


***Figure 4. 49**: KWIC tool displaying mola as a search word in GSC*

The 100 KWIC lines excerpted reveal that *mola* appears 72 times positioned between clauses with no comma usage (see **Figure 4.50**), 16 times positioned between clauses with comma usage before (see **Figure 4.51**) and 12 times positioned at the beginning of a sentence (see **Figure 4.52**).

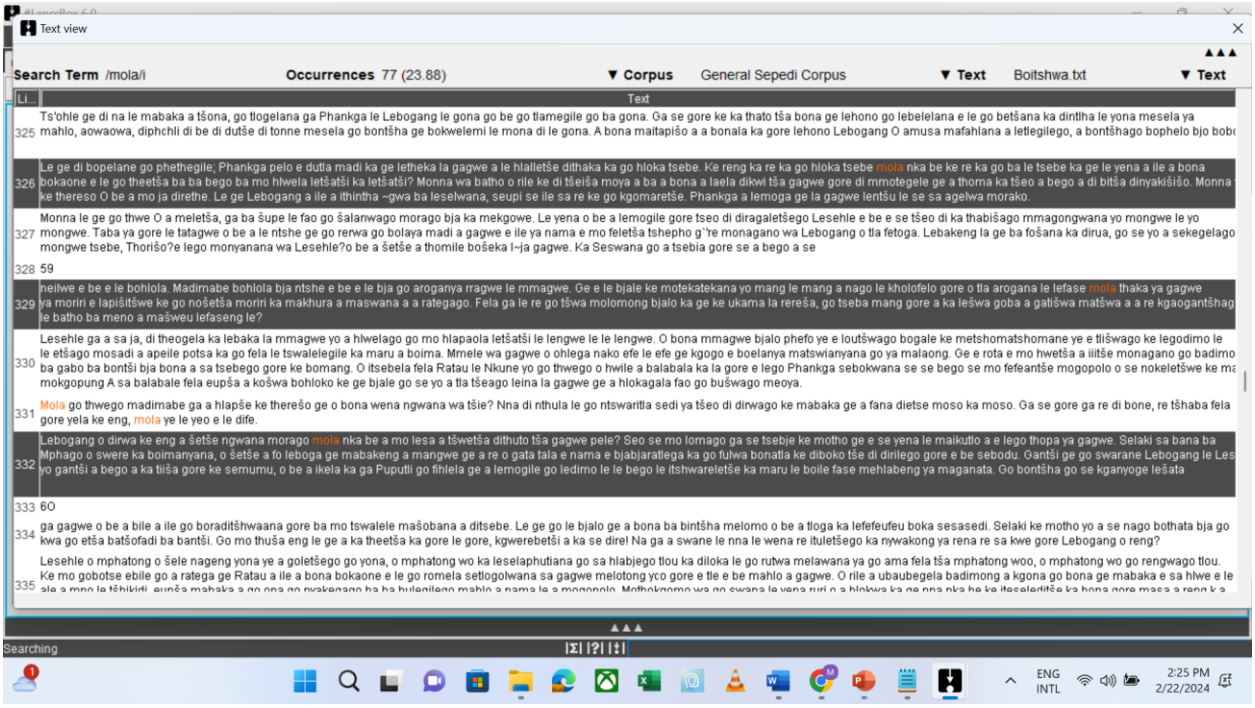


***Figure 4. 50**: The conjunction mola positioned between clauses with no comma usage in the GSC*

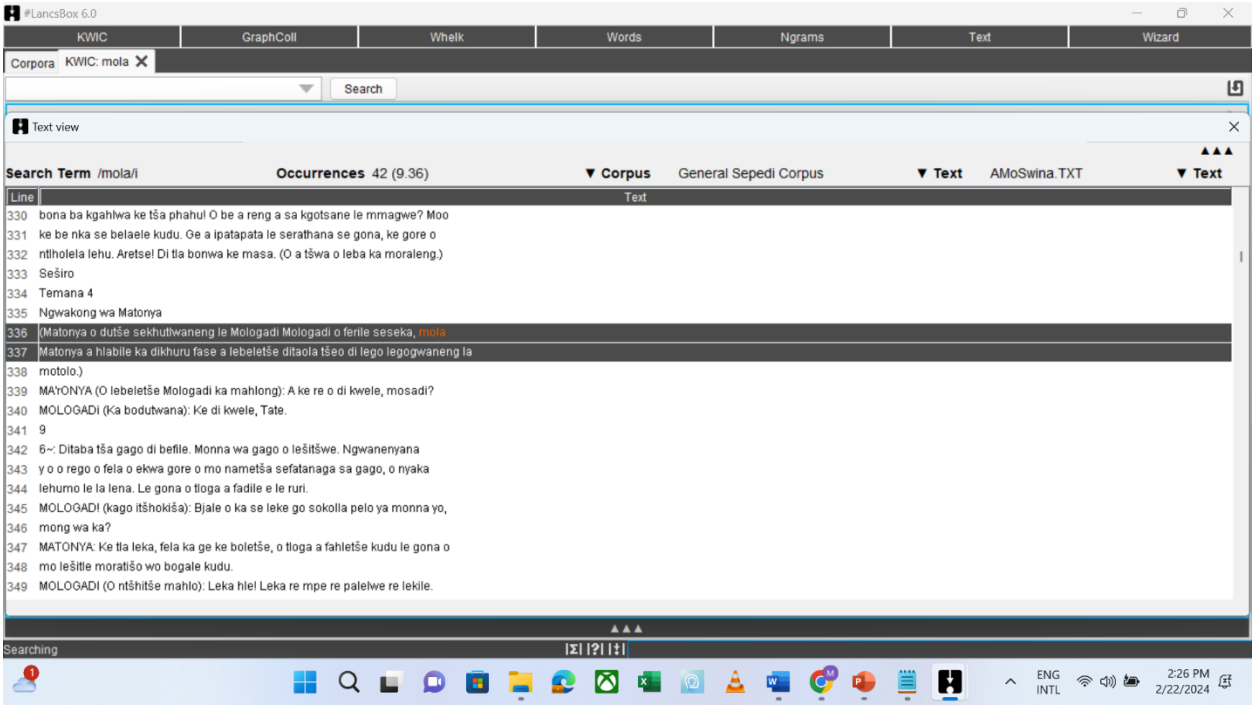


***Figure 4. 51**: The conjunction mola positioned between clauses with comma usage before in the GSC*

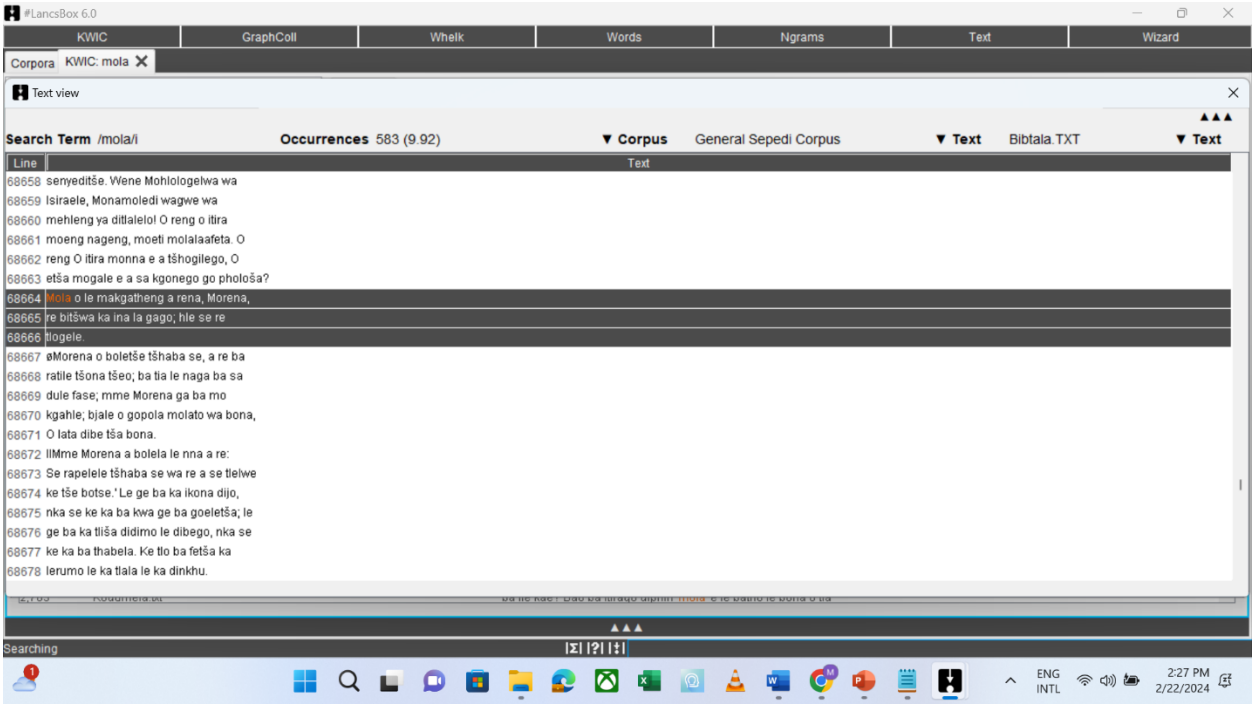


***Figure 4. 52**: The conjunction mola positioned at the beginning of a sentence in the GSC*

These results demonstrate that the use of *mola* positioned between clauses with no comma usage is predominant in the GSC than its usage with comma before and at the beginning of a sentence.

These findings are in agreement with those of the SSC. Both corpora show dominance of the usage of the conjunction positioned between clauses with no comma usage.

The results further indicate that when the conjunction *mola* is used in the GSC, it expresses the following senses:

- contrast or change, which can be translated as ‘whereas’,
- logical order or time, which can be translated as ‘during’ and ‘subsequently’, and
- cause and effect, which can be translated as ‘since’ in English.

Consider the following examples sampled from the concordance lines:

- *Ba bantši ba sepela ka dipese go ya mošomong **mola** ba bangwe ba sepela dithekisi goba dinamelwa tša bona.* 'Others travel by buses to work, **whereas** others travel by taxis or their private transports'.

- *Bakgomana ba dilete ba be ba agile Jerusalema, **mola** ba agile le kua metseng ya Juda; e mongwe le e mongwe o be a dula lapeng la gagwe motseng wa gabo wa Baisiraele.* 'County chiefs lived in Jerusalem, **during** their dwelling in the cities of Jude; each one of them used to live in his home, the hometown of the Israelites'.

- ***Mola** a se hloramišago madulong a sona ka mehla, a bowa le nao tša gago. **Mola** a felellago ka sefero, ye kgolo ya retoga e farafarilwe ke bakgomana. Go fihleng a dula fase, setšhaba se emeletše. **Mola** a šurušwago wa go thekga ditho, a kebamiša hlogo e le gona go ba dudiša fase.* '**Since** the chief persecuted it from where it lives, every day, following your footsteps. **Since** he left, he came back surrounded by chiefs. He arrived and sat down, the people stood up. **Subsequently** to being escorted in a relaxed way, he then raised his head to make them sit down.

From the above examples, it is clear that the Sepedi conjunction *mola* can be translated as 'whereas', 'during', 'subsequently' and 'since' in English.

The results show that the semantic feature of whereas/while is common between the two corpora. However, the senses of 'during', 'subsequently' and 'since' observed in the GSC, could not be found in the SSC. These three meanings, therefore, indicate that there is a distinction in the sense conveyed by the conjunction in the two corpora.

The preceding section has highlighted the commonalities and differences in the syntactic and semantic features of Sepedi conjunctions in the SSC and GSC. In the section that follows, the syntactic and semantic features of Sepedi conjunctions outside the two corpora are provided. A comparison of the findings obtained from the two corpora and the general usage of the conjunctions outside the corpora is then performed.

## 4.7 Discussion

The findings obtained from the SSC and GSC indicate that similarities in the syntactic and semantic features of conjunctions surpass the differences. In terms of syntactic features, findings from both corpora showed that the usage of Sepedi conjunctions positioned between clauses with no comma usage is the predominant feature. This was observed in the analysis of five out of the six conjunctions that were investigated, namely *ebile*, *goba*, *gomme*, *gore* and *mola*. Generative Grammarians would, therefore, argue that the usage of Sepedi conjunctions positioned between clauses with no comma usage forms part of the innate rules governing production of grammatical sentences or utterances in Sepedi.

In terms of the semantic features, the results also indicated several commonalities in the sense or meaning carried by Sepedi conjunctions in the SSC and the GSC. For instance, the conjunction *ebile* conveys more or less similar meaning, which is 'also', 'and' and 'furthermore' in both corpora, and the conjunction *ge* has the sense 'when' common in both corpora. Furthermore, the findings also show similarities in the meaning conveyed by the conjunction *goba,* which is 'or' in both corpora. Moreover, the results from the SSC concerning the conjunction *gomme* coincide with those obtained in the GSC. The conjunction has the sense 'and' common among the two corpora. The conjunction *gore* also has the senses 'that', 'so that' that are common between the two corpora. Lastly, results also showed that the meaning(s) 'whereas/while' carried by the conjunction *mola* (or has only one common meaning) coincide (or coincides has only one meaning) in the two corpora.

It is perhaps vital to point out that the SSC represents LSP, since it represents language belonging to the field of Linguistics. The GSC, on the other hand, is representative of LGP, since it encompasses general language resources that were gathered from the internet and supplemented by plain texts received from the Department of African Languages at the University of Pretoria. Therefore, it is safe to infer that the findings of the study indicate that Sepedi conjunctions are used in a similar manner and often carry the same meaning in the LSP and LGP. However, it is also necessary to find out if Sepedi conjunctions are used and defined in a similar way as reflected in the two corpora in their everyday usage outside the corpora, as well as in scholarly sources.

### *4.7.1 The conjunction: ebile*

From Sepedi scholarly material, the examples below show that the conjunction *ebile* is positioned between clauses with a comma usage before and without a comma usage. Consider the following examples provided by Poulos and Louwrens (1994) and Nokaneng and Louwrens (1991).

Example provided by Poulos and Louwrens (1994:372):

- *O a lwala,* ***ebile*** *ga se a ya mošomong lehono*. 'He is ill, he didn't **even** go to work today'.

Example provided by Nokaneng and Louwrens (1991: 138):

- *O na le lehufa* ***ebile*** *o hloya batho.* 'He/She is jealous **and** hates people'.

It was alluded in the foregoing discussion that in terms of usage, the conjunction *ebile* was found positioned between clauses with a comma before and without a comma in both corpora. Therefore, this observation from Sepedi scholarly books coincides with that from both corpora. Furthermore, it was clear from corpus analysis findings that the predominant feature between the two corpora was that the conjunction *ebile* is positioned between clauses with no comma usage. Therefore, the example provided by Nokaneng and Louwrens (1991) confirms this predominance.

From the above examples, it is clear that the conjunction *ebile* is used to express contrast or change which can be translated as 'even' and also addition which can be translated as 'and'. The sense 'and' was observed in both corpora. However, it was alluded that this sense 'even' was found in the SSC and not in its GSC counterpart. Therefore, the examples from scholarly sources indicate that the senses of the conjunction that were found in the corpora are also seen in its usage outside the corpora. Let us proceed to the second conjunction, namely *ge*.

### *4.7.2 The conjunction: ge*

Concerning the usage of this conjunction, Poulos and Louwrens (1994: 370-371) provide the following examples:

- *Ke tlo mmotša **ge** ke boa*. ‘I will tell him **provided that** I will come back’.
- ***Ge** a aga ntlo ye botse, ke tla e reka*. ‘**If** he builds a nice house, I will buy it’.

It is clear from these examples that the conjunction *ge* is used positioned between clauses with no comma usage and also positioned at the beginning of a sentence. In the SSC, there was preference of usage of the conjunction positioned at the beginning of a sentence, whereas in the GSC the dominant preference was the usage of the conjunction positioned between clauses with no comma usage. Therefore, it is apparent that the two usages of this conjunction are common in the corpora and the material outside the corpora.

Concerning the semantic features, it is apparent from the above examples that the conjunction *ge* shows a condition which can be translated as ‘provided that’ and ‘if’. The senses could be traced only in the GSC and were absent in the SSC. Therefore, it is only the meaning from the GSC in this case that coincides with that found outside the corpora. Let us now consider the third conjunction, namely *goba*.

### *4.7.3 The conjunction: goba*

One syntactic feature of this conjunction is observed from Sepedi scholarly material. In this feature, the conjunction *goba* is positioned between clauses with no comma usage. Consider the example below provided by Nokaneng and Louwrens (1991: 138):

- *O tseba go ngwala **goba** o tseba go bala*? ‘You know how to write **or** you know how to read’.

The above example confirms the common syntactical feature between the two corpora, as was indicated previously when the conjunction was analysed. In terms of semantic features, the example above expresses an alternative or choice, which can be translated as ‘or’. This meaning also confirms what was observed in both corpora. However, another sense of the conjunction was indicated by Ziervogel and Mokgokong (1975: 310) in the *Groot Noord-Sotho Woordeboek* ‘Comprehensive Northern Sotho Dictionary’, i.e., the conjunction *goba* can be translated as ‘either or’. This feature could not be found in the two corpora. Let us move on to the fourth conjunction, namely *gomme*.

### *4.7.4 The conjunction: gomme*

Three syntactic features of this conjunction are observed in the Sepedi reference scholarly materials. Consider the following examples provided by Poulos and Louwrens (1994), Lombard, van Wyk and Mokgokong (1985) and Ziervogel *et al*. (1969).

Example provided by Poulos and Louwrens (1994:372):

- *O hlwele sepetleleng sebaka,* ***gomma*** *ga se a fola*. 'He stayed in hospital for a long time, **and** did not get well'.

Example provided by Lombard *et al*. (1985:177)

- *Ke mmoditše* ***gomme*** *o ganne nang*. 'I asked him **but** he refused bluntly'.

Example provided by Ziervogel *et al*. (1969:85)

- *A mmotšiša a re:* ***Gomme*** *ke ka lebaka la eng ge o kgahlwa ke seema seo?* 'He asked him: **And** for what reason is it that that proverb affects you so much?'

From the above examples, it is clear that the conjunction *gomme* is positioned between clauses with comma usage before, positioned between clauses without comma usage and positioned at the beginning of the sentence. Since the common feature between the two corpora was that the conjunction is positioned between the clauses with no comma, it is apparent that this usage is common between the corpora and the general usage of the conjunction outside the corpus. Furthermore, the usage of the conjunction positioned at the beginning of the sentence could only be traced in the GSC. Since the scholarly materials are treated as representing general language usage, it can be inferred that the findings confirm the usage of the conjunction in the LGP.

From above examples, it is apparent that the conjunction *gomme* can be translated as 'and' and 'but', as indicated in the English translations. Therefore, the sense 'and' is common between the general usage of the conjunction outside the corpus and the two corpora. However, the meaning 'but' was not present in both corpora. Let us proceed to the fifth conjunction, namely *gore*.

### *4.7.5 The conjunction: gore*

One syntactic feature is observed from the *Groot Noord-Sotho Woordeboek* 'Comprehensive Northern Sotho Dictionary'. Consider the examples provided by Ziervogel and Mokgokong (1975: 343-344) in this dictionary:

- *O a bona **gore** re sa ja.* 'You see **that** we are still eating'.
- *O tlo tla **gore** a re thuše.* 'He'll come **in order** to help us'.
- *Enwa dihlare tše **gore** o se babje.* 'Drink this medication **lest** you become ill'.

From the above examples, it is clear that the usage of this conjunction positioned between clauses with no comma usage is a common feature between the corpora and scholarly sources. The conjunction can be translated as 'that', 'in order' and 'lest'. The sense 'that' was also found in both corpora. Furthermore, the sense 'in order' was present in the SSC, but absent in the GSC. Therefore, the sense 'in order' marks similarities between the meaning of the conjunction in the scholarly materials and SSC. However, the sense 'lest' indicates distinction in the meaning carried by the conjunction between the scholarly sources and the corpora. Let us now consider the sixth and final conjunction, namely '*mola'.*

### *4.7.6 The conjunction: mola*

One syntactical pattern can be observed in the Sepedi scholarly material. Consider the following example provided by Poulos and Louwrens (1994: 371):

- *Ke tlo mmotša **mola** a fihla*. 'I will tell him **when** he arrives'.

From the above example, it becomes apparent that the conjunction is used positioned between the clauses with no comma usage. The observation above is in agreement with findings from both corpora. In terms of the meaning conveyed by the conjunction, the sense 'when' was not found in both corpora. Therefore, semantic features of the conjunction indicate difference in the meaning carried by the conjunction between the corpora and scholarly materials.

From the above discussion, it is apparent that the similarities in the usage and meaning of Sepedi conjunctions between the scholarly sources or everyday language and corpora surpass the differences. This, therefore, indicates that the corpus-based approach to investigating linguistic phenomena provide valid, credible and invaluable information that can benefit linguists, language educators, researchers, as well as students. The findings further reveal that material developers can also draw useful insights from corpus data, that can guide them provide accurate information on parts of speech and also provide usage examples that are a true reflection of the function and usage of parts of speech.

## 4.8 Conclusion

This chapter has presented discussion of the usage of Sepedi conjunctions in the SSC and GSC. It was highlighted that BONSE dictionary was used as a source from which to gather conjunctions to be considered for inclusion in the study. A total of 20 Sepedi conjunctions from the dictionary were selected and their frequency of occurrence in both corpora was proffered.

The Whelk Tool was employed to note the distribution of the identified conjunctions in the individual files of the SSC and their comparison was carried out. The frequency of occurrence of the conjunctions in the two corpora was used as the criteria in the selection of conjunctions to be considered for inclusion in the comparative analysis. Six conjunctions with high frequency of occurrence were selected and analysed. The KWIC tool offered by LancsBox X was used to retrieve concordance lines from the two corpora, which showed the authentic and contextual usage of the conjunctions. The analysis was firstly conducted for the SSC, focusing on the syntactic and semantic features of these conjunctions. The analysis then moved to the GSC, in which the same procedure was followed. The findings obtained from the SSC were then compared to those found in the GSC. Lastly, the findings obtained from the corpora were compared to the general usage and meaning of conjunctions outside the corpora.

This chapter presented analysis and interpretation of data for the present study. The following chapter presents an overview of how the aim and objectives of the study were met, the contribution of this research to existing literature, acknowledgement of observed

limitations, implications for future research, as well as recommendations based on the findings obtained.

# CHAPTER 5: CONCLUSION

## 5.1 Introduction

This chapter presents an overview of how the aim and objectives of the present study were met. Furthermore, it presents the contribution of this research to corpus-based studies in teaching and learning material development and also provides acknowledgement of observed limitations with the present research. Moreover, the chapter ends with recommendations based on the results and offers implications for future research. The aim of the study has been to compare the syntactic and semantic features of Sepedi conjunctions between the SSC and the GSC. The investigation also considered frequency of occurrence of the conjunctions in the SSC and the GSC. The present study sought to address the following objectives, deriving from the aim of the study:

- To determine the frequency of Sepedi conjunctions in the SSC and the GSC.
- To establish whether syntactic and semantic features of Sepedi conjunctions are similar or different between the SSC and the GSC.
- To determine if there are any similarities and differences in the usage and meaning of Sepedi conjunctions between the scholarly sources and the corpora.

Prior to summarising the findings and demonstrating how the above-mentioned objectives were achieved in this present study, it is essential to provide a brief overview of the chapters that make up the present study.

## 5.2 Summary of chapters

In **Chapter 1**, the primary goal was to offer a contextual background of the research problem that the present study aimed to address. It was necessary to explore and compare how conjunctions are used and defined between the LSP and LGP. Furthermore, the chapter discussed the rationale behind conducting the present study and addressing the identified research problem. Although many studies have been conducted to investigate parts of speech in the South African Indigenous languages, very few studies have been done to investigate conjunctions, especially within the context of the Sepedi language. This study, therefore, endeavoured to bridge this gap in existing

literature and contribute to our understanding of usage and meaning of Sepedi conjunctions.

Also, in order to address the previously mentioned research problem, the present study employed the corpus-based approach as its method. It was alluded that it was only in 2002 that a fully-fledged corpus-based study was conducted on South African Indigenous languages, and consequently, the study contributes to the body of ongoing corpus-based research in the South African setting. Lastly, the study emphasised the research questions it set out to address, in addition to the aim and objectives already described and the study's delineation.

The aim of **Chapter 2** was to provide a comprehensive literature review of studies relevant to the present study. It presented various scholars' perspectives on the subject matter. It began with studies conducted on an international scale. These comprised research by Simpson and Mendis (2003), Gabrielatos *et al*. (2007), and Roslim *et al*. (2021), among others. The discussion of studies on the African continent then followed, which included studies conducted by Kawalya *et al*. (2009), Toscano and Sewangi (2005) and Okeke and Okeke (2022). Studies from the South African setting were then covered, including those by Gauton and de Schryver (2002), Van Olmen *et al*. (2019), Gauton *et al*. (2004), and Taljard and de Schryver (2016). The chapter also encompassed studies done within the context of education, specifically on teaching and learning material development. Lastly, the theoretical framework underpinning this study was discussed. For the present study, generative grammar theory was used as a theoretical framework. This theory put forward several views. Firstly, descriptive grammar is preferable to prescriptive grammar. Secondly, grammars should characterise competence, not performance. Thirdly, grammars should be fully explicit. Fourthly, analyses of language should be as broad as possible. Fifthly, there should be generalisations in grammar theory. And lastly, grammars should be psychologically relevant (Wasow 2003). As the sole corpus-based study focused on the authentic usage and meaning of the Sepedi conjunction conducted within the South African setting, it was emphasised that the current study will make a substantial contribution to the field.

**Chapter 3** provided the research methodology employed in the present research. The study used both the qualitative and quantitative method. Qualitative research emphasises linguistic (words) data over numerical data and employs meaning-based analysis methods rather than statistical approaches (Nieuwenhuis 2016). On the other hand, quantitative research depends on numerical data from a specific subgroup of a population to draw generalisable conclusions about the broader population under study (Maree & Pietersen 2016). Furthermore, there was discussion of the differences between the intuition-based and corpus-based approaches to linguistic analysis. Moreover, the corpus-based approach versus the corpus-driven approach as types of approaches that employ corpora as the basis for analysis were explored.

The use of the corpus-based approach for data analysis and interpretation in the study was also emphasised. It was indicated that two corpora, namely the SSC and the GSC, are used in the present study. The SSC consisted of texts from three Sepedi textbooks and the GSC encompassed texts from the internet which were supplemented with general texts received from the University of Pretoria, Department of African Languages. A definition of a corpus was given, along with a description of the design process. Descriptions of several corpora, according to the purpose they mean to serve were given, including general corpus, specialised corpus, written corpus, spoken corpus, synchronic corpus, diachronic corpus, learner corpus, and monitor corpus. Corpora that are classified in terms of design criteria were discussed, including the monolingual corpus, bilingual or multilingual corpus, parallel corpus and comparable corpus. Aspects that relate to corpus design were highlighted, including balance, representativeness and size.

The chapter then went on to detail the several kinds of corpus-query software that can be used to manipulate a corpus. These software types include Sketch engine, ParaConc, WordSmith Tool, MonoConc, AntConc and LancsBox X. It was said that the program to be used in the present study was LancBox X. Additionally, the entire process of developing the SSC for the current study and constructing the GSC utilising the BootCat toolbox was detailed. Lastly, the different tools (KWIC, Whelk, GraphColl, Words, Ngrams and Text Tools) offered by LancsBox X, some of which were used in the present study,

were discussed and screenshots were provided displaying how to access and search words in each tool.

**Chapter 4** presented the results of the analysis of the syntactic and semantic features of Sepedi conjunctions. Subsequently, the study's results were correlated with its aim and objectives in order to make it evident how they were met.

It may be imperative to include a separate section specifically for summarising the results of this study and also indicating how each of the objectives of the research were fulfilled.

### 5.3 Summary of findings

The study's first objective was to determine the frequency of Sepedi conjunctions in the SSC and GSC. In the case of comparing frequency of occurrence, there was some contrast in the frequency of the conjunctions between the two corpora. Notably, conjunctions ***eitše*** and *erile* exhibited zero occurrence in the SSC, contrasting sharply with their high prevalence in the GSC. Similarly, *kapa* and *nkane* manifested infrequent appearance in the SSC, with only one occurrence for each, whereas their occurrence was significantly high in the GSC.

The second objective was to compare the syntactic and semantic features of Sepedi conjunctions as observed in the SSC and the GSC. The results indicated that similarities in the syntactic and semantic features of conjunctions outweigh the differences. Concerning the syntactic features, the results from both corpora indicated that the usage of Sepedi conjunctions positioned between clauses with no comma usage is the dominant feature. This was seen in the investigation of five of the six conjunctions that were analysed, i.e., *ebile, goba*, *gomme*, *gore* and *mola*. In terms of the Generative Grammar approach, which was used as the theoretical framework for the study, the condition of the usage of the conjunctions positioned between clauses with no comma usage would be taken as forming part of the general and innate laws governing the use and function of conjunctions, particularly in Sepedi. Of course, other conditions that were found, e.g., the usage of the conjunctions without a comma before or after, the usage of the conjunctions at the beginning of a sentence, etc., would also be considered as forming part of the universal laws and rules governing formation of grammatical sentences or utterances.

In terms of the semantic features, the findings also revealed several commonalities in the meaning conveyed by Sepedi conjunctions in the SSC and the GSC. For example, the conjunction *ebile* conveys similar meaning, which is 'also', 'and' and 'furthermore' in both corpora, and the conjunction *ge* has the sense 'when' common in both corpora. Moreover, the results also revealed similarities in the meaning carried by the conjunction *goba,* which is 'or' in both corpora. Furthermore, the results from the GSC regarding the conjunction *gomme* coincide with those obtained in the SSC. The conjunction possesses the sense 'and', which is common in the two corpora. The conjunction *gore* also carries the senses 'that', 'so that', and these could be observed in both corpora. Lastly, the findings also indicated that the meaning 'whereas/while' conveyed by the conjunction *mola* is common among the two corpora.

The above findings clearly indicate that there are commonalities in the manner in which conjunctions in Sepedi are used and defined between the LSP and the LGP.

The third objective has been to determine the commonalities and differences in the usage and meaning of Sepedi conjunctions between the scholarly sources or everyday language and the corpora. The results indicated that the similarities in the usage and meaning of Sepedi conjunctions between the scholarly sources or everyday language and corpora outnumber the differences. Therefore, this was an indication that the findings obtained from the corpora coincides with the usage and meaning of conjunctions outside the corpora. This in turn signals that a corpus can be trusted and be seen as representing the language as a whole. Therefore, the corpus-based approach is able to provide linguists, researchers, language educators and students with valid, credible and invaluable insights. This further shows that material developers can benefit immensely from corpus data when writing about parts of speech in general and conjunctions in particular.

In the preceding section, a summary of the results has been provided. Now it is time to highlight the potential contribution of the present research to the body of knowledge concerning the present topic.

## 5.4 Contribution of present research

Within the South African Indigenous languages, there are several corpus-based studies on teaching and learning material development. However, it is evident that the area of conjunction usage or their function within the Sepedi language, has remained largely unexplored. This study, therefore, endeavours to bridge this gap in existing literature and contribute to our understanding of usage and meaning of conjunctions in Sepedi.

Furthermore, the corpus-based approach was used in this study's data analysis and interpretation. Since the corpus-based method is still very much in its infancy in the South African context, as was mentioned in the discussion that came before it, the current study significantly contributes to the ongoing corpus-based research in South Africa.

Moreover, the majority of research that utilised this approach in the South African setting, if not all of them, did concentrate on other aspects of part of speech, with none of them particularly addressing the use and meaning of Sepedi conjunctions. Consequently, this study is the first to use the methodology to look into the use and meaning of Sepedi conjunctions. Thus, it is clear that, in the context of South Africa, the current study significantly advances the field of corpus-based investigations. It will also benefit material developers who wish to write on the usage and meaning of Sepedi conjunctions. The present study can assist material developers and linguists to use conjunctions that are more frequent instead of focusing on those that are hardly used by the speakers of the language. In addition, language educators can also draw from the findings of the study when teaching conjunctions and put more emphasis on those conjunctions that have been found to be more frequent in the corpora. The corpus-based approach, when combined with appropriate corpus-query software, allows electronic texts to be studied and manipulated in their original form. Since the method is still not so popular within South African, the present study also contributes by raising awareness of the benefits the method brings for linguists, researchers, students, etc. who wish to investigate various linguistic phenomena.

## 5.5 Limitations

Only one tittle of school textbooks was used for the purposes of the present study and only three learners' textbooks within the title that are used in Senior Phase were purposefully selected. Therefore, this can be seen as a limitation to the study, as other tittles of learners' textbooks might have presented different findings concerning the usage and meaning or function of Sepedi conjunctions.

Furthermore, there may be other Sepedi conjunctions outside the corpora with high frequency of usage that were not necessarily included in the analysis. However, the conjunctions that were analysed were sampled from a dictionary that was compiled using a corpus. Therefore, the general assumption is that its compilers conducted research on conjunctions with high frequency of occurrence before commencing with the dictionary compilation process. Furthermore, the fact that only six conjunctions formed the focus of the study could also be a limiting factor on the findings obtained. An analysis of all frequent Sepedi conjunctions may yield sufficient and credible findings concerning the usage and meaning of conjunctions. However, the limited scope of the research could not allow for more conjunctions to be analysed.

## 5.6 Future research implications

The relationship between the corpus-based method and pedagogical materials within South African has not yet been adequately explored. Therefore, this area of research still needs to be investigated to assist material developers (i.e., developing) use corpus-based results to supplement their intuition and experiences when teaching and learning material. Furthermore, research that can investigate different titles from the ones explored in the present study is needed. That will strengthen further strengthen the findings of the present study on the usage and meaning or function of Sepedi conjunctions. Moreover, the corpus-based method is a useful tool that can be used to investigate other parts of speech in future. Lastly, only six conjunctions formed the focus of the present research. Therefore, future research is needed that can include more conjunctions or focus on different conjunctions from the ones investigated in the present study. That can further strengthen or refute the findings of the present study.

## 5.7 Recommendations

Using corpus results to inform material developers and linguists on inclusion, exclusion, and treatment of parts of speech is a good move in this day and age. Therefore, researchers seeking to investigate various linguistic phenomena should consider employing the corpus-based method, if they wish to arrive at valid, reliable and transferable findings. Moreover, LancsBox X is an extremely helpful software package when trying to query a corpus. The user-friendliness of the software is commendable. As mentioned in Chapter 3, the program can be downloaded and installed with ease and is available without charge. For researchers seeking to virtually see linguistic phenomena, such as collocation graphs, distributions of terms, frequency of occurrence, etc., LancsBox X is a valuable corpus-query tool that can enable them to achieve exactly that.

## 6. LIST OF REFERENCES

Anthony, L. 2004. AntConc: a learner and classroom friendly, multi-platform corpus analysis toolkit. *Proceedings of IWLeL 2004: An Interactive Workshop on Language e-Learning*. Waseda University, Tokyo, 10 December, 2004:1–13.

Anthony, L. (2013). A critical look at software tools in corpus linguistics. *Linguistic Research 30(2)*, 141 – 161.

Atkins, S., Clear, J., & Ostler, N. (1992). Corpus design criteria. *Literary and linguistic computing*, 1-16.

Babbie, E. (2010). *The practice of social research. 12th Edition.* Belmont: Wadsworth.

Babbie, E., & Mouton, J. (2001). *The practice of social research. South African edition.* Cape Town: Oxford University Press.

Baker, M. (1995). Corpora in translation studies: An overview and some suggestions for future research. *Target, 7(2)*, 223 – 243.

Bapela, M., Mphela, T., & Ratshivhambela, M. (2013). *Oxform Lebone Grade 7 LB.* Cape Town: Oxford University Press.

Barlow, M. (2001, January 02). *Analysing parallel texts with ParaConc.* Retrieved from https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.553.3644&rep=rep1&type=pdf.

Barlow, M. (2003). *Concordancing and corpus analysis using MP 2.2.* Houston: Athelstan.

Barlow, M. (2008). *ParaConc and parallel corpora in Contrastive and Translation Studies.* Houston: Athelstan.

Bowker, L., & Pearson, J. (2002). *Working with Specialized Language: A Practical Guide to Using Corpora.* London & New York: Routledge.

Brezina, V., Timperley, M. & McEnery, T. 2018. #LancsBox X. v. 4.x [Software]. Available at: http://corpora.lancs.ac.uk/LancsBox X.

Brown, E.K, & Miller, J.E. (1991). *Syntax: a linguistic introduction to sentence structure*. 2nd ed. London: Routledge.

Chapel, L., & Clause, C. (2021). Chomsky's Theory of Language Acquisition | Stages & Examples. Retrieved from Humanities Courses / Introduction to Humanities: Help and Review: https://study.com/academy/lesson/noam-chomsky-on-language-theories-lesson-quiz.html

Chomsky,N. (1986). *Knowledge of language: Its nature, origin and use*. New York: Praeger.

Conrad, S. (2010). What can a corpus tell us about grammar? In A. O'Keeffe, & M. McCarthy, *The Routledge Handbook of Corpus Linguistics* (pp. 227-240). London & New York: Routledge.

Creswell, J. W., & Creswell, J. D. (2018). *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches.* Los Angeles: SAGE.

Davies, M. (2008). Retrieved from The Corpus of Contemporary American English (COCA): https://www.english-corpora.org/coca/.

de Schryver, G. (2007). *Bilingual Oxford Northern Sotho-English.* Cape Town: Oxford

de Schryver, G. M., & Gauton, R. (2002). The Zulu locative prefix ku-revisited: A corpus-based approach. *Southern African Linguistics and Applied Language Studies*, 20(4), 201-220.

de Schryver, G. M., & Nabirye, M. (2010). A quantitative analysis of the morphology, morphophonology and semantic import of the Lusoga noun. *Africana Linguistica*, 16, 97-153.

Department of Basic Education. (2012). *Setatamente Sa Pholisi Sa Lenaneothuto Le Kelo Sepedi Leleme la Gae: Mephato Ya 7-9.* Pretoria: Government Printers.

Dlamini, P. (2021). *Avoiding potholes in translation: A practical perspective on translation between English and IsiZulu.* Pietermaritzburg: University of KwaZulu-Natal Press.

Froehlich, H. (2015). *Corpus Analysis with Antconc.* Programming Historian. Retrieved from https://programminghistorian.org/en/lessons/corpus-analysis-with-antconc.

Gabrielatos, C. (2005). Corpora and Language Teaching: Just a fling or wedding. *The Electronic Journal for English as a Second Language*, 1-32.

Gabrielatos, C., Davies, M., Rayson, P., Hunston, S., & Danielsson, P. (2007). If-conditionals as modal colligations: A corpus-based investigation. *In 4th Corpus Linguistics Conference.*

Gauton, R., de Schryver, G. M., & Mohlala, L. (2003). A corpus-based investigation of the Zulu nominal suffix-kazi: A preliminary study. *In Proceedings of the 4th World Congress of African Linguistics, New Brunswick*, 373-380.

Gauton, R., de Schryver, G., & Mohlala, L. (2004). A corpus-based investigation of the Zulu nominal suffix -kazi: A preliminary study. *Proceedings: 4th World Congress of African Linguistics. New Brunswick, 2003*, 373 – 380.

Gouws, R., & Prinsloo, D. (2005). *Principles and practice of South African lexicography.* Cape Town: Sun Press.

Granger, S. (2003). The international corpus of learner English: a new resource for foreign language learning and teaching and second language acquisition research. *TESOL Quarterly*, 538-546.

Greene, J. (2005). Combining qualitative and quantitative methods in social inquiry. In B. Somekh, & C. Lewin, *Research methods in the social sciences.* (pp. 274 – 281.). London: SAGE.

Halle, M. (1962). Phonology in generative grammar. Word, 18(1-3):54–72.

Hausser, R. (2011). *Corpus linguistics, generative grammar, and database semantics.* Berlin ; New York: De Gruyter Mouton.

Holmes, J. (1988). Doubt and certainty in ESL textbooks. *Applied Linguistics. 9*, 21-44.

Kawalya, D., De Schryver, G. M., & Bostoen, K. (2019). A corpus-driven study of the expression of necessity in Luganda (Bantu, JE15). *Southern African Linguistics and Applied Language Studies*, 37(4), 361-381.

Kennedy, G. (1987). Quantification and the use of English; a case study of one aspect of the learners' task. *Applied Lingustics. 8*, 264-286.

Kennedy, G. D. (1998). *An Introduction to Corpus Linguistics .* London and New York: Longman.

Kenny, D. (2009a). Corpora in translation studies. In M. e. Baker, *Routledge encyclopedia of translation studies.* (pp. 50 – 53). London: Routledge.

Kilgariff, A. (2008). The Sketch Engine as a Common Platform for Showcasing Language Resources. *Proceedings of the Sixth International Conference: Formal Approaches to South Slavic and Balkan languages.* (pp. 15 – 20.). Dubrovnik,.

Kilgariff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., & Rychlý, P. &. (2014). The Sketch Engine: Ten Years On. *Lexicography, 1(1)*, 7 – 36.

Koshane, S., Mpe, M., & Mphela, T. (2013). *Oxford Lebone Grade 8 LB.* Cape Town: Oxford University Press.

Kruger, A. (2002). Corpus-based translation research: Its development and implications for general, literary and Bible translation. *Acta Theologica Supplementum, 2*, 70 – 106.

Kunilovskaya, M., & Koviazina, M. (2017). Sketch Engine: A toolbox for linguistic discovery. *Journal of Linguistics, 68(3)*, 503 – 507.

Lee, D. Y. (2012). What corpora are available? In A. O'Keeffe, & M. McCarthy, *The Routledge Handbook of Corpus Lingusitics* (pp. 107-121). London & New York: Routledge.

Lewin, C. (2005). Elementary quantitative methods. In B. Somekh, & C. Lewin, *Research methods in the social science* (pp. 215 – 225). London: SAGE Publications.

Lombard, D. P. (1993). *Northern Sotho (special) :* Study guide for SNS100-V. Bd. 1 (Rev. ed). Pretoria: Univ. of South Africa.

Lombard, D., van Wyk, E., & Mokgokong, P. (1985). *Introduction to the Grammar of Northern Sotho.* Pretoria: J.L van Schaik.

Louwrens, L., Kosch, I., & Kotzé, A. (1995). *Northern Sotho.* Germany: LINCOM EUROPA.

Makhalemele, S., Mpe, M., & Mphela, T. (2013). *Oxford Lebone Grade 9 LB .* Cape Town: Oxford University Press.

Maree. (2016). First Steps in research 2. In K. Maree, & J. Pietersen, *The quantitative research process* (pp. 162-172). Pretoria: Van Schaik .

Maree, K. (2016). First Steps in Research 2. In J. Nieuwenhuis, *Introducing qualitative research* (pp. 50-70). Pretoria: Van Shaik.

McEnery, T., & Wilson, A. (1996). *Corpus Lingusitcs.* Edinburgh: Edinburgh University Press.

McEnery, T., Xiao, R., & Tono, Y. (2006). *Corpus-Based Language Studies. An afvanced resource book.* USA & Canada: Routledge.

Mindt, D. (1995). *An Empirical Grammar of the English Verb: Modal verbs.* Berlin: Cornelsen.

Mncwango, L. (2017). *Corpus-based critical discourse analysis of the portrayal of body parts in selected Zulu novels. Unpublished Master's dissertation.* Pretoria: University of Pretoria.

Moemi, M., Ramusi, M., Kgatla, M., Mothiba, K., & Ngulube, M. (2013). *Platinum A re šogeng thari.* Cape towm: Maskew Miller Longman (Pty) Ltd.

Nokaneng, M., & Louwrens, L. (1991). *Segagešo. Mphato 7.* Pretoria: Via Afrika Limited.

Ojo, T. A., & Mathabathe, R. (2021). An investigation into the effectiveness of the Curriculum and Assessment Policy Statement (CAPS) in south african schools. *International Journal on Integrating Technology in Education (IJITE) Vol.10, No.2*, 23-38.

Okeke, G. T., & Okeke, C. O. (2022). On the semantic-pragmatic interface of Igbo verbs of perception. *Cogent Arts ; Humanities*, 9(1), 2025991.

Poulos, G., & Louwrens, L. (1994). *A Lingustic Analysis of Northern Sotho.* Pretoria: Via Afrika Limited .

Prinsloo, D. (2015). Corpus-based Lexicography for Lesser-resourced Languages - Maximizing the Limited Corpus. *Lexikos, 25*, 285 – 300.

Recski, L.J. 2006. Corpus linguistics at the service of English teachers. *Literatura y lingüística*, 17:303–324.

Roslim, N., Aziz, A., Abdullah, M. H., & Nimehchisalem, V. (2021). Corpus-informed Prepositions: What to Learn and What is Suitable for Learning. *ournal of Asia TEFL*, 18(3), 1003.

Saldanah, & O'Brien, S. (2014). *Research methodologies in translation studies.* London & New York: Routledge.

Schiffer, S. (2015). Meaning and Formal Semantics in Generative. *Erkenntnis 80 (1)*, 61-87.

Scott, M. (1998). *WordSmith Tools Manual. Version 3.0.* Oxford: Oxford University Press.

Simpson, R., & Mendis, D. (2003). A corpus-based study of idioms in academic speech. *TESOL quarterly*, 37(3), 419-441.

Simpson, S., Briggs, J., & Swales, J. M. (1999). *The Michigan Corpus of Academic Spoken English.* Ann Arbor, MI: The Michigan Regents of the University of Michigan.

Stanborough, R. (2019). Born This Way: Chomsky's Theory Explains Why We're So Good at Acquiring Language. Retrieved from PARENTHOOD: https://www.healthline.com/health/childrens-health/chomsky-theory

Taljard, E. (2012). Corpus-based language teaching: An African language perspective. *Southern African Linguistics and Applied Language Studies*, 30(3), 377-393.

Taljard, E., & de Schryver, G. M. (2016). A corpus-driven account of the noun classes and genders in Northern Sotho. *Southern African Linguistics and Applied Language Studies*, 34(2), 169-185.

Tognini-Bonelli, E. (2001). *Corpus Linguistics at Work.* Amsterdam; Philadelphia: John Benjamins.

Toscano, M., & Sewangi, S. (2005). Disovering usage patterns for the Swahili amba- relative forms Cl. 16, 17, 18: Using corpus data to support autonomous learning of Kiswahili by Italian speakers. . *Nordic Journal of African Studies*, 14(3), 44-44.

Van Olmen, D., Breed, A., & Verhoeven, B. (2019). A corpus-based study of the human impersonal pronoun ('n) mens in Afrikaans: Compared to men and een mens in Dutch. *Languages in Contrast*, 19(1), 79-105.

Wasow, T. (2003). Generative Grammar. In M. Aronoff, & J. Rees-Miller, *The Handbook of linguistics* (pp. 296-318). UK ; USA: Blackwell.

Wynne, M., & Berglund, Y. (2012). *Corpus Linguistics course.* Retrieved from [online]: https://weblearn.ox.ac.uk/access/content/group/3a217dfd-a8cd-4034-8564-

Xiao, R. (2009). Theory-driven corpus research: Using corpora to inform aspect. In A. Lødeling, & M. (. Kytö, Corpus Linguistics: An International Handbook (pp. 987-1008). Berlin: de Gruyter.

Yule, G. 2010. The study of language. New York: Cambridge University Press.

Ziervogel, D., & Mokgokong, P. (1975). *Groot Noord-Sotho Woorde-book.* Pretoria: J.L. van Schaik and The University of South Africa.

Ziervogel, D., Lombard, D., & Mokgokong, P. (1969). *A Hand book of the Northern Sotho Language.* Pretoria: J.L. van Schaik.

***Addendum A**: A formal letter addressed to the publication company*

Make today matter

FACULTY OF EDUCATION
Department of Humanities Education

**Publishing company**

**Oxford University press (OUP)**

**THE PUBLISHING COMPANY**

**RE: REQUEST FOR USE OF SEPEDI TEXTBOOKS FOR RESEARCH PURPOSES**

**RESEARCH TITLE: THE REPRESENTATION AND DISTRIBUTION OF CONJUNCTIONS IN SELECTED SEPEDI HOME LANGUAGE TEXTBOOKS: A CORPUS-BASED INVESTIGATION**

Dear Ms Barbara Strydom,

I am writing to request permission to use selected Sepedi textbooks for research purposes for my Master's thesis. I am registered for the MEd in African Languages education at the University of Pretoria (u13413199). Specifically, I am interested in investigating the representation and distribution of conjunctions in selected Sepedi home language textbooks. My study will be corpus-based and will rely on data collected from grade 7-9 Sepedi textbooks. This is an important research question as it has the potential to inform the development of future language curricula in South Africa. Since my study is a corpus-based one, I would need to have the textbooks in electronic format in order to be able to query them by means of computer software.

If possible, I would like to request your assistance in obtaining electronic copies of the textbooks listed below in order to conduct my research. Alternatively, I would request permission to scan and convert these books to electronic (.txt format) in order to enable me to query the contents computationally. I understand the importance of adhering to the copyright act, and I assure you that I will comply with all relevant

regulations in my research. I am therefore willing to sign a confidentiality agreement to ensure that the data is not misused.

The books that I would need the permission for are the following:

Oxford Lebone Kreiti ya 7

Oxford Lebone Kreiti ya 8

Oxford Lebone Kreiti ya 9

Thank you for your consideration of my request. I look forward to hearing from you soon.

Yours faithfully,

Mpho Mahlobogoane

Mpho Mahlobogoane (MEd Student)

Signature: 

U13413199@tuks.co.za

Cell: 078 273 7434

Dr Connie Makgabo (Supervisor)

Signature: 

connie.makgabo@up.ac.za

Cell: 072 923 8838

Dr Erick Nzimande

(Co-supervisor)

Signature:

nzimaen@unisa.ac.za

Cell: 076 423 4985

***Addendum B**: A letter from the publication company, recognising the academic intent and the importance of fostering research*

**Oxford University Press**
**Southern Africa (Pty) Ltd**

Vasco Boulevard
N1 City, Goodwood
Cape Town | 7460
South Africa
(**PO Box** 12119 | 7463)

**Telephone** | 021 596 2300
**Email** | oxford.za@oup.com

www.oxford.co.za

09 October | 2023

To Mpho Mahlobogoane,
Faculty of education
Department of humanities education
University of Pretoria

Dear Mr Mahlobogoane

This letter serves as permission for you to make use of selected OUPSA titles in your research titled:

THE REPRESENTATION AND DISTRIBUTION OF CONJUNCTIONS IN SELECTED SEPEDI HOME LANGUAGE TEXTBOOKS: A CORPUS-BASED INVESTIGATION

The titles in question are:

- Oxford Lebone Kreiti ya 7
- Oxford Lebone Kreiti ya 8
- Oxford Lebone Kreiti ya 9

As per your email dated 19 September 2023 in which you agreed, this approval is granted on the basis that you will share with OUP any judgement you make of the content before publication of your research.

Best wishes,

Barbara Strydom

Publishing Manager: Schools Languages

Oxford University Press