Designing English-Kambaatissata Bilingual Electronic Dictionary: Using Parallel Corpora

Temesgen Heliso Woymo

A Thesis Submitted to:

The Department of Linguistics

in Partial Fulfillment for the Degree of Master of Science in Computational Linguistics

Addis Ababa University

Addis Ababa, Ethiopia

Jun 2013

ADDIS ABABA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

This is to certify that the thesis submitted by Temesgen Heliso Woymo, entitled: Designing English-Kambaatissata Bilingual Electronic Dictionary: Using Parallel Corpora and submitted in partial fulfillment of the requirements for the degree of master of science complies with the regulations of the university and meets the accepted standards with respect to originality and quality.

Signed by the Examining Committee:

Examiner Wondwosen Tesfaye Signature _____ Date July 12, 2013

Examiner GRMIAS ABEBT Signature _____ Date July 12, 2013

Advisor Sebsibe H/M (PhD) Signature _____ Date 12 July 2013

Advisor Dereb Ado Signature _____ Date 12 July 2013

Chair of Department of Graduate Program Coordinator

Abstract

The main aim of this study is to design English-Kambaatissata bilingual electronic dictionary using English-Kambaatissata parallel corpora. Literature review has been made on Kambaatissata phonology and morphosyntax. Then, based on the knowledge on Kambaatissata morphology, the study adopted statistical machine translation approach. To do so, IBMModel1that is a word alignment model and widely used in working with parallel bilingual corpora and which implements expect maximization algorithm has used.

In general, 1194 Kambaatissata and English sentences were used from parallel raw text. The raw texts were collected from the English-Kambaatissata Constitution of the Federal Democratic Ethiopia (1995), a training material for primary school mother tongue teachers and a Bible story book of children. A database of word alignment probabilities were developed from the aligned sentences. These probabilities were used to select the translation of English-Kambaatissata word and Kambaatissata-English. After having the translation equivalent of English and Kambaatissata, parts of speech and gloss of the English terms from the English WordNet were extracted.

The accuracy of the designed prototype was tested using 273 or 20% of the retrieved dictionary terms. Based on the manual evaluation, the result shows that 61.5% of the translation was correct.

Key Words: Bilingual electronic dictionary, parallel corpora, design, alignment and enhancement.

Acknowledgement

I am very grateful to all who helped me generously by giving their time, energy, knowledge and other resources while undertaking this research. Without the contribution of these people, the study could not have come to completion.

First and foremost, I would like to express my deepest thank to Sebsibe H/Mariam (PhD) and Derib Ado (PhD) my thesis advisors, for their invaluable and unreserved assistance in providing me with relevant advice, critical comments and constructive suggestions throughout my thesis work.

My gratitude is also extended to Philipos Paulos (MA), Temesgen Senbato (MA), Tadesse Bunaro and Meles Samuel for their technical support concerning the corpus data in general and in evaluating the generated dictionary entries manually.

My deepest appreciation also goes to my lovely wife Meseret Getachew for her all-rounded encouragement and support not to be listed. I also appreciate my staff head Adisse Ersado for his support by giving me the laptop computer to do this research.

At the end but not the least thanks go to my classmates in the Computational Linguistics MSc Program of 2011 entries for their encouragement.

Table of Contents

List of Figures

List of Algorithms

List of Acronyms

CLIR- Cross-lingual information retrieval

MRD- Machine readable dictionary

NLP- Natural Language Processing

NLTK- Natural Language Toolkit a comprehensive python library for natural language and text analytic

WA- Word alignment

POS- Part of Speech

WordNet- a dictionary resource/ lexical database with glosses, synonymy sets (SynSets), example sentences and the like

1. CHAPTER ONE

1.1. Introduction

Kambaatissata is a language which is spoken in the Kambaata-Tambaaro Zone of the Southern Nations Nationalities and Peoples Region (SNNRPR) of Ethiopia. According to (Freeman and Pankhurst, 2001 as cited in Treis, 2008), the Kambaatissata speakers live in the high land areas around the Hambarichcho massif of south west of the Ethiopian capital of Addis Ababa, between the Omo River to the west and the Billate River to the east. It is bounded by Hadiyya to the north of west, Wolayita to the south and Halaaba to the east (Treis, 2008).

The total population Kambaata Tambaaro Zone is 771,698. Among them 510,000 are mother tongue speakers of Kambaatissata (Kambaata Tambaaro Zone Culture Tourism and Government Communication Head Office, 2011).

Information can exist in various forms, but accessing the right information at the right time is a pre-requisite for functional efficiency (Birungi, 1995). Now accessing information is becoming critical to the success of all users within a given society. This is because the economic development of any country depends upon effective utilization of the stored information, especially for developing countries (Birungi, 1995). Today a large amount of information is available in the form of written text. Developing countries like Ethiopia can facilitate their vision of development by making this store of knowledge accessible to their citizens. This can be done with the help of different techniques such as natural language processing, speech recognition, machine translation and information retrieval (Alemayehu & Willett, 2002).

There are different linguistic resources which are used in natural language processing. These resources are different in their type and role. These include corpus, WordNet, thesaurus, dictionary and the like.

Corpus (corpora are) is a linguistic resource that is a collection of texts assumed to be representative of a given language, or other subset of a language, to be used for linguistic analysis. Its applications include language teaching, lexicography and linguistic research (Francis & Kucera, 1964).

The usage of parallel corpora opens up the possibilities of creation of new resources. Extraction of lexical data by means of word alignment (WA) from bi-text (parallel corpora) is broadly used in many fields of computational linguistics, such as machine translation, cross-language information retrieval, and also bilingual thesaurus or dictionary creation (Pietrzak, 2009).

WordNet is a lexical database which was established in 1985 by a group of psychologists and linguists at Princeton University (Miller, 1985). The initial idea was to provide an aid to use in searching dictionaries conceptually, rather than merely alphabetically.

Thesaurus is a kind of dictionary which deals on words relation especially, as synonyms and antonyms. Its plural form is thesauri. It provides structured vocabularies for describing art. The first requirement for a thesaurus to be useful is that it provides a hierarchical structure that has an unambiguous interpretation. Some hierarchically organized thesauri mix the sub/super class relation with a part-of relation (Bechofer & Goble, 2001).

Another type of linguistic resource is a dictionary. It is a lexicon or a reference book containing an alphabetical list of words with information about them. Dictionary can be of different types based on different criteria. These can include bilingual, monolingual, multilingual, etymological, rhyming, mini, pocket, thesaurus, glossary, crossword, machine readable (electronic) and the like.

A bilingual dictionary gives words in two languages. One of the languages is grouped alphabetically in separate halves of the book with translations into the other language.

In contrast to traditional dictionaries, which are designed to be used by human beings, dictionaries for natural language processing (NLP) are built to be used by computer programs. The final user is a human being but the direct user is a program. Such a dictionary does not need to be able to be printed on paper. The structure of the content is not linear, ordered entry by entry but has the form of a complex graph. Because most of these dictionaries are used to control machine translation or cross-lingual information retrieval (CLIR) the content is usually multilingual and usually of huge size.

It is widely recognized that machine readable dictionaries are important sources of lexical data which is suitably analyzed, extracted and formalized. It could be used effectively in the construction of lexical components for NLP applications.

Lexical databases play a central role in all natural language processing applications, ranging from simple spellcheckers to more complex machine translation systems (Briscoe, 1991). In most cases, they constitute the sole parameter information for the corresponding software, and apart from some very basic methods such as stemming relying on pure string processing principles and low linguistic requirements, hardly

any language technology application can avoid relying on a minimal lexical resource (Lovins, 1968).

Machine-readable dictionaries (MRDs) have been seen as a likely source of information for use in NLP. This is because they contain an enormous amount of lexical and semantic knowledge collected together over years of effort by lexicographers.

Manual construction of large lexical-semantic databases demands enormous human resources, and there is a growing body of research into the possibility of automatically extracting at least a part of the required lexical and semantic information from everyday dictionaries (Amsler, 1980). Everyday dictionaries are obviously not structured in a way that enables their immediate use in NLP systems.

Electronic dictionaries offer great potential benefits for users: they are quick and easy to use, they can provide access to large amounts of data, and they are interactive (Nesi, 1999). Using electronic bilingual dictionaries have to do with allowing learners to search for terms they wish to communicate in the target language. Electronic dictionaries serve as much more than just mere word translators. Apart, from being portable, they are convenient in terms of providing multi search paths and speeding up the search process.

The expediency and speed of electronic dictionaries give access to a wide variety of searches through the presentation of entries and other composite elements, which supply updated information from different available sources.

The empirical methods are gaining ground in the field of linguistics. Especially, the automatic analysis of big corpora is used in constructing of models of language

based on empirical data. Soaring popularity of automatic construction of linguistic resources, collecting of corpuses from the web, constructing of grammars, lexicons and dictionaries complement the work performed till now manually by linguists (McCracken, 2003).

The usage of parallel corpora opens up the possibilities of creation of new resources. Extraction of lexical data by means of word alignment from bi-text (parallel corpora) is broadly used in many fields of computational linguistics, such as machine translation, cross-language information retrieval, and also bilingual thesaurus or dictionary creation, which is of interest in this research.

Most bilingual dictionaries do not contain the detailed information that may find necessary for translation (Deshmukh, 2011). The same thing happened to the designed English-Kambaatissata bilingual electronic dictionary. In order to solve the problem, the enhancement is done using the English WordNet.

1.2. Classification of the Language

Kambaatissata belongs to the Cushitic branch of the Afro-Asiatic language phylum, more precisely to the Highland East Cushitic (HEC) language group. The hitherto little documented language is spoken by more than 600,000 speakers in an area approximately 300 km south-west of the Ethiopian capital Addis Ababa (Treis, 2012).

The Kambaatissata language has been spelled in different ways. The common ones are Kambaata, Kambata, Kambatta, Kembata, Cambata and Cambatta. The Kambaata people refer to their language by the term Kambaatissata, which includes the derivational formative *-issata* for names of or as Kambaati afoo "the mouth of

Kambaata" (Treis, 2008). The term Kambaata can be used both in reference to the ethnic group and to the Kambaata country. In the linguistic literature, it is also used to denote the language. But in this research, the term *Kambaata* is used to refer to the people and the term *Kambaatissata* is used to refer to the language to be consistent.

1.3. Statement of the Problem

Machine readable electronic dictionary is very vital resources for natural language applications. Nevertheless, dictionaries and other lexical resources are not yet widely available in electronic form (Mayfield & McNamee, nd).

A key resource for many approaches to cross-language information retrieval (CLIR) is a bilingual dictionary. Unfortunately, like other cross-language resources, machine-readable bilingual dictionaries that are suitable for use in CLIR are scarce (Mayfield & McNamee, nd).

Translated lexicons play a vital role in several applications related to machine translation (MT). Such lexicons are used for cross-language information search (Etzioni, Reiter, Soderland, & Sammer, 2007).

Lexical coverage is critical but resources (Machine-readable bilingual dictionaries, parallel corpora, and MT systems) are scarce. There is a growing consensus in the CLIR community that lexical coverage is the most important factor in CLIR performance (Mayfield & McNamee, nd).

There is no Kambaatissata-English bilingual electronic dictionary which is one of the necessary lexical resources to develop NLP applications. This inconvenience

There is no Kambaatissata-English bilingual electronic dictionary which is one of the necessary lexical resources to develop NLP applications. This inconvenience made the researcher to design Kambaatissata-English bilingual electronic dictionary system using statistical machine translation approach.

Tools that increase lexical coverage are desirable. This research solves one of the problems of the lexical resource scarceness by designing Kambaatissata-English bilingual electronic dictionary.

In general, by designing English-Kambaatissata bilingual electronic dictionary, the following research questions were expected to be answered.

- How bilingual electronic dictionary of English-Kambaatissata could be designed using statistical machine translation approach?
- In which way an English WordNet could be used to enhance the dictionary entries of the language?

1.4. Objective of the Study

1.4.1. General objective

The general objective of this research is to design English-Kambaatissata bilingual electronic dictionary using parallel corpora and enhancing its performance using the English WordNet.

1.4.2. Specific objectives

The specific objectives of this research are:

I. to collect the parallel corpora of English and Kambaatissata languages;

II. to handle the preprocessing of the raw texts of both English and Kambaatissata;

III. to select and use appropriate tool for building English-Kambaatissata bilingual electronic dictionary;

IV. to use the parallel corpora of English and Kambaatissata for the creation of English-Kambaatissata bilingual electronic dictionary;

V. to measure the performance of the English-Kambaatissata bilingual electronic dictionary;

VI. to enhance the English-Kambaatissata bilingual electronic dictionary using the English WordNet.

1.5. Significance of the Study

The significance of this study is presented as follows.

1. As it is stated by Teresa (2011) (cited in Meyer, 1988), the priority for scientists is to spread progress among societies that do not share the same language, and the way to achieve this is through translation and the first sources they look to when they do not know an equivalent are bilingual electronic dictionaries (Meyer, 1988). The English-Kambaatissata bilingual electronic dictionary is very important source for spreading the technological advancement of the world to the Kambaatissata speaking society by providing the equivalent translation of English terms into Kambaatissata.
2. Translation lexicons play a vital role in several applications related to machine translation (MT). Such lexicons are used for cross-language information search (Etzioni et al., 2007). Thus, the people who like to make a study on Kambaatissata in general as well as in specific areas like machine translation, and cross-language information retrieval are beneficiaries.

3. Researchers in the languages which are highly dialectically related like Halaaba, Sidaamo, Qabeena, and Hadiyya can be beneficiary by using the output of this research as a reference as well as motivation.
4. Designers of bilingual electronic dictionaries for the other languages by using the algorithm as the initial items.
5. It can be an input for English-Kambaatissata machine translation, bilingual information retrieval system, and thesaurus construction.

1.6. Scope of the Study

The corpus data used in this research was not as such large in size. It had only 1,194 pair of sentences. It was also collected from three different sources; from the Federal Democratic Republic of Ethiopia Constitution (1995), the Bible story book for children and the training material for primary school mother tongue teachers.

The output of this work is the probabilistic structure of the dictionary, aligning a list of Kambaatissata potential headwords with their English translation equivalent(s), where the degree of correspondence is expressed in terms of alignment probabilities. This English-Kambaatissata probabilistic dictionary is considered to be half-finished product, the basis for the full-fledged dictionary, revised and enhanced by linguists. Its enhancement is only on English WordNet side. Since there is no Kambaatissata WordNet, there is no enrichment when retrieving from English to Kambaatissata.

1.7. Methodology

The focus of this research was designing English-Kambaatissata bilingual electronic dictionary based on statistical machine translation approach. To do so, the following

procedures, tools and techniques were used to achieve the stated objective of the study.

1.7.1. Data Sources

The corpus data was collected from the Federal Democratic Republic of Ethiopian Constitution (1995), training material for primary schools mother tongue teachers prepared by Kadida Gamela Woreda Education Office and the children Bible story book.

The text corpora used here were by having the available one. That was because there were shortage of parallel corpora of English and Kambaatissata.

1.7. 2. Methods and Procedures

All the tasks in the word alignment (WA) field are highly language-dependent. Questions like word order, inflectional paradigms, morphological complexity, alphabets used, etc. are of vital importance for the final success (Pietrzak, 2009).

This complex nature of WA field leads to text preprocessing activity. This was included normalization (using all letters in small letters case, removing punctuation marks, symbols and stop words), text tokenization and lemmatization and creation of list of aligned sentences.

The Text pre-processing process was the important part of the research. This was because texts in their raw form may be in different cases (capital or small). In order for the machine to read properly, it needs to get them in the same form and WA usually works on lists of aligned sentences. Why the researcher concentrated on pre-processing was because the precision of WA depends highly on the preprocessing.

1.7. 3. Development Tools

Python programming language is a simple yet powerful programming language with excellent functionality for processing linguistic data. Due to that it was preferred with English WordNet and other modules like NLTK, IBMModel1 and tk-inter to develop interactive user friendly interface and all the programs to manipulate the files and develop the prototype. Thus, this research also used python to develop the bilingual electronic dictionary.

1.7. 4. Evaluation Techniques

There are two widely used ways of evaluation. These are automatic comparison with existing bilingual MRD and manual evaluation of a subset of the dictionary entries. For this research using manual evaluation method was chosen. This was because there is no bilingual MRD to compare in Kambaatissata-English or vice versa.

1.8. Organization of the Paper

The paper is structured as follows:

As the effectiveness of the research depends mainly on the accurate word alignment, the analysis of related works carried out in this field is presented in section two. In section three, the literature about Kambaatissata is presented. In section four, the methodology of the research was introduced step by step. These included some information about the input data or provided by parallel corpus used and described the whole process of dictionary creation step by step. The fifth section dealt with the design and experimentation of the prototype as well as measuring the accuracy of the created dictionary. The final part dealt with the conclusion and recommendation of the future research directions.

1.9. Definitions of Terms (Operational Definitions)

In this part of the research the terms used to describe special meaning or which contain out of the usual meaning are described as the following:

Under-resourced languages- refer to the languages which do not have much literacy works compared to other developed languages.

Electronic dictionary is a dictionary whose data is in digital form.

Python is a kind of programming language.

Tokenization is splitting texts into different smaller parts.

IBM model 1 is the model which is used to translate words of a source language into target language using the lexical translation probabilities.

White space is the space between the two words in the written text.

Source language is the language of the user information request.

Target language is the language of the document collection or to be retrieved as translation.

Translation does not mean a full translation of the user request incorporating correct grammar and word order as if it had been carried out by a human translator. Rather, it is to perform a type of concept mapping.

Lemma is normally the head words in a dictionary. They represent word forms of the same lexeme. For example in: "go", "goes", "going", "went" and "gone", "go" is the lemma form.

Inflection is variation of the form in a word, often affix (the ending of the word) showing different grammatical information such as tense, number, gender, case, etc. Example: walk, walked.

Derivation is the morphological process which derives or creates a new word by varying the meaning or POS or both.

Morpheme is meaningful grammatical unit consisting of a word, such as fan, or word element, such as *–ed* in worked, which cannot be divided into smaller meaning full or grammatical parts.

Parallel corpora are pairs of texts which contain data in a main language and a translation thereof.

Accuracy is the amount of word forms that are correctly translated in the texts.

Alignment is the process of matching texts in both languages sentences or words correspond to each other on the same level.

Corpus (plural = corpora) is a body of stored texts in written and/or a transcription of recorded spoken language in the electronic form.

Transnumeral is the term that refers to inherently both plural and singular nouns.

CHAPTER TWO

2. Literature Review

2.1. Introduction

This chapter is divided into four different parts. The first part provides a review of literature about linguistic resources; corpus, dictionary (bilingual dictionary and electronic dictionary), thesaurus and WordNet. The second part presents about text preprocessing including concept of stop words removal, tokenization, and lemmatization. The third part discusses about the development tool; python programming language. The fourth part focused on machine translation approach and the last part discussed about related works in the Kambaatissata language as well as electronic dictionary.

2.1.1. Linguistic Resources

2.1.1.1. Corpus

A corpus is a collection of text documents, and corpora are the plural of corpus, which is assumed to be representative of a given language, or other subset of a language, to be used for linguistic analysis (Francis & Kucera, 1964).

Corpus has been classified in a finer scheme of classification characterized by its inherent features:

- Loosely, a corpus refers to anybody of text;
- Most commonly, it refers to a body of machine-readable text and;

- More strictly, it refers to a finite collection of machine-readable texts sampled to be maximally representative of a language or a variety of it (McEnery & Wilson, 1996, P. 215).

A corpus may contain texts in a single language (monolingual corpus) or text data in multiple languages (multilingual corpus). Multilingual corpora that have been specially formatted for side-by-side comparison are called aligned parallel corpora.

Corpus is a large collection of linguistic data, either written text or a transcription of a recorded speech, which can be used as starting point of linguistic description or as a means of verifying a hypothesis about a language (Crystal, 1995). Thus, it refers to a large collection of written and spoken text sample, available in machine-readable form, accumulated in scientific manner to represent a particular variety or use of language.

In principle, a corpus is actually designed for accurate study of linguistic properties, features and phenomena observed in a language. According to (Dash, 2005) a systematically compiled corpus however, small in size, should adhere to the following criteria:

- A corpus should faithfully represent both a common and special linguistic features.
- The corpus should be large and wide to encompass texts from various disciplines.
- A corpus should be a true replica of physical texts available in printed form.

- A corpus should be available in the electronic form for easy access by the end users in order to enable the common users as well as the language researchers to use the database in multiple tasks related to language description and analysis, statistical analysis, language processing, translation, etc. (Dash, 2005, P. 12).

Corpus may be defined acrostically from the letters used to compose the term in the following way:

C: Compatible to both man and computer,

O: Operational in research and application,

R: Representative of a language or variety,

P: Process-able both in man and machine,

U: Unlimited in amount of data and samples, and,

S: Systematic both in formation and representation (Dash, 2005, P. 4).

The scarceness of corpora is not necessarily the insurmountable obstacle since human decoders can align a small number of sentences (about 1000 sentence) (Al-Onaizan et al., 2000).

Dictionary publishers are increasingly recognizing the value of electronic versions of dictionaries and are putting more information in these versions than the ones that appear in the print versions. It needs several efforts to enhance a dictionary database as a resource for computational applications. These efforts include much greater use of corpus evidence in creating definitions and associated information for an entry.

Corpus-based methods have also been used in the construction of a thesaurus (McCracken, 2003).

2.1.1.2. Dictionary

A convenient and valuable source of linguistic information is a standard dictionary. It is easy to use, and if used intelligently, very informative. Far from being a dull dry reference book, the dictionary is a vast storehouse of interesting information about an infinite number of useful word tools (Deshmukh, 2011).

2.1.1.3. Bilingual Dictionary

The bilingual dictionary is useful to have exact equivalence of a word. A dictionary contains important facts far beyond simple definitions and guides to pronunciation and spelling (Deshmukh, 2011).

A bilingual dictionary is a dictionary which registers the equivalences of meanings in two languages. Bilingual dictionaries are seldom diachronic and usually alphabetic in arrangement. The difference between a monolingual dictionary and a bilingual one is made not only in the number of languages in which they are written but also in their essential purpose.

A bilingual dictionary consists of an alphabetical list of words or expressions in one language (the 'source language') for which, ideally, exact equivalents are given in another language (the 'target language'). The purpose is to provide help to someone who understands one language but not the other.

2.1.1.4. Electronic Dictionary

The term electronic dictionary can be used to refer to any reference material stored in electronic form that gives information about spelling, meaning, or use of words.

Computational lexicology was coined to refer to the study of machine-readable dictionaries (MRDs), and emerged in the mid-1960s and received considerable attention until the early 1990s. 'Machine-readable' does not mean that the computer reads the dictionary, but only that it is in electronic form and can be processed and manipulated computationally (Amsler, 1982).

Conventional dictionaries contain a lemma with various descriptions. A machine-readable dictionary may have additional capabilities. As (Deshpande, 2012), machine-readable versions of everyday dictionaries have been seen as a likely source of information for use in natural language processing because they contain an enormous amount of lexical and semantic knowledge (Deshpande, 2012).

A machine-readable dictionary is used for a number of different purposes in information systems, like word sense disambiguation, information retrieval, automatic text classification and automatic text summarization (Aas & Eikvil, 1999).

Dictionary entry consists of several fields of information for different applications. Among the standard fields, translation for the bilingual dictionaries is one of them. Each of these fields has proven useful for different applications, such as for building semantic taxonomies and machine translation (Deshpande, 2012).

An electronic dictionary is a dictionary whose data exists in digital form and can be accessed through a number of different media. Lexical databases play a central role

in all natural language processing applications ranging from simple spellcheckers to more complex machine translation systems (Briscoe, 1991).

In most cases, they constitute the sole parameter information for the corresponding software, and apart from some very basic methods such as stemming relying on pure string processing principles and low linguistic requirements, hardly any language technology application can avoid relying on a minimal lexical resource (Lovins, 1968).

According to (Ahmed, 2008) the print version of the dictionary has the following drawbacks:

- It is time consuming- locating the required word takes a lot of time and the user has to pass different stages i.e. first he/she has to locate the first letter of the word and then to locate the second, the third till the last letter of the word.
- Users should have the dictionary at hand; the users should have to carry the dictionary with them. Many people do not feel comfortable time and place to have the dictionary with them.
- Users should have a means to get the dictionary, buy or lend.
- The dictionary is not available anytime, anywhere; availability of the dictionary is limited.
- Some of the pages may not be available because of the different reasons (Ahmed, 2008, PP. 9-10).

Electronic dictionaries offer great potential benefits for users: they are quick and easy to use, they can provide access to large amounts of data, and they are

interactive. It is also important to take into consideration that most web based dictionaries are currently available (Nesi, 1999).

2.1.1.5. Thesaurus

Thesaurus is a kind of dictionary which deals with words relation especially, as synonyms and antonyms. Its plural form is thesauri. It provides structured vocabularies for describing art. The first requirement for a thesaurus to be useful is that it provides a hierarchical structure that has an unambiguous interpretation. Some hierarchically organized thesauri mix the sub/super class relation with a part-of relation (Bechofer & Goble, 2001).

2.1.1.6. WordNet

WordNet is a dictionary designed for programmatic access by natural language processing systems. NLTK includes a WordNet corpus reader, which is used to access and explore WordNet (Perkins, 2010, P. 8).

WordNet is a hierarchically organized large lexical database of English. Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (SynSets), each expressing a distinct concept (Miller, G. A, Beckwith, Fellbaum, Gross, & Miller, 1990).

SynSets are interlinked by means of conceptual-semantic and lexical relations. The resulting network of meaningfully related words and concepts can be navigated with the browser. Its structure makes it a useful tool for computational linguistics and natural language processing (Miller et al., 1990).

WordNet was designed to capture several types of associative links, although the number of such links was limited by practical considerations. WordNet was not designed as a lexical resource, so that, its entries do not contain the full range of information that is found in an ordinary dictionary (Fellbaum, 1998). WordNet has found widespread use as a lexical resource, both in research and in NLP applications. WordNet is a prime example of a lexical resource that is converted and mapped into other lexical databases.

The most obvious difference between WordNet and a standard dictionary is that WordNet divides the lexicon into five categories: nouns, verbs, adjectives, adverbs, and function words. Actually, WordNet contains only nouns, verbs, adjectives, and adverbs. The relatively small set of English function words is omitted on the assumption that they are probably stored separately as part of the syntactic component of language. The realization that syntactic categories differ in subjective organization emerged first from studies of word associations (Miller et al., 1990).

WordNet is organized by semantic relations. Since a semantic relation is a relation between meanings, and since meanings can be represented by SynSets, it is natural to think of semantic relations as pointers between SynSets.

2.1.2. Text Preprocessing

2.1.2.1. Stop Words

Stop words are common words that generally do not contribute to the meaning of a sentence, at least for the purpose of information retrieval and natural language processing. Most search engines filter stop words out of search queries and documents in order to save space in their index (Perkins, 2010).

In information retrieval, a document is traditionally indexed by frequency of words in the documents. Statistical analysis through documents showed that some words have quite low frequency, while some others act just the opposite (Baez-Yath & Ribeiro-Neto, 1999). The common characteristic of these words is that they carry no significant information to the document. Instead, they are used just because of grammar. We usually refer to this set of words as stop words (Raghavan & Wong, 1986).

In processing English texts, the relatively small set of English function words is omitted on the assumption that they are probably stored separately as part of the syntactic component of language. Stop words are words which found in both written and spoken languages. Although they found in the natural language, they cannot carry the main content of the text.

The application of stop words has already been explored in many fields. In digital libraries, for instance, elimination of stop words could significantly reduce the size of the indexing structure and obtain a compression ratio of more than 40% (Baez-Yath & Ribeiro-Neto, 1999). On the other hand, a stop word list provides a good resource for information retrieval. It can speed up the calculation and increase the accuracy at the same time (Salton & Buckley, 1988).

Up to now, a lot of stop word lists have been developed for English language. These stop word lists are traditionally extracted by frequency analysis of all the words in a large corpus. Results from different corpora are usually quite similar to each other and they are commonly used as standards (Yang, 1995).

2.1.2.2. Tokenization

Tokenization is the process of separating parts of the text into different parts. This can take place using different criteria. For instance, the tokenization may follow the white space as the delimiting or it may use other symbols or punctuation marks.

Tokenization is a method of breaking up a piece of text into many pieces, and is an essential first step for any process. Tokenization is the process of splitting a string into a list of pieces, or tokens. We can split a paragraph into a list of sentences, sentences into tokens and the like (Perkins, 2010, P. 8).

Tokenization is a method of breaking up a piece of text into many pieces. Word tokenizer works by separating words using spaces and punctuation, allowing you to decide what to do with it. It is possible to use the NLTK default tokenizers to tokenize text corpora. If it is not suitable for the NLTK default tokenizer, you can provide your own tokenizer instances (Perkins, 2010, P. 9).

2.1.2.3. Lemmatization

Lemmatization is similar to stemming, but is more akin to synonym replacement. Lemma is a root word, as opposed to the root stem. Unlike stemming, it left with a valid word which means the same thing. But the word you end up with can be completely different (Perkins, 2010, P. 28).

For English, we can use the WordNet Lemmatizer to find lemmas. This can be done by calling from the NLTK as importing module.

The WordNet Lemmatizer is a thin wrapper around the WordNet corpus, and uses the morphy() function of the WordNet Corpus Reader to find a lemma. If no lemma

is found, the word is returned as it is. Unlike with stemming, knowing the part of speech of the word is important. The default part of speech is a noun. Instead of just chopping off the affixes like the Porter Stemmer, the WordNet Lemmatizer finds a valid root word. Where a stemmer only looks at the form of the word, the lemmatizer looks at the meaning of the word. And by returning a lemma, one will always get a valid word (Perkins, 2010, P. 28).

2.1.3. Development Tools

2.1.3.1. Python Programming Language

Python is a simple yet powerful programming language with excellent functionality for processing linguistic data (Bird et al., 2009, P. Xii).

Python programming language is a dynamically-typed, object-oriented interpreted language. Although, its primary strength lies in the ease with which it allows the programmer to rapidly prototype a project, its powerful and mature set of standard libraries make it a great fit for large-scale production-level software engineering projects as well. Python has a very shallow learning curve and an excellent online learning resource (Madnani, nd, P. 3).

Python has a comprehensive library for natural language and text analytic called NLTK. Originally it is designed for teaching; it has been adopted in the industry for the research and development due to its usefulness and breadth of coverage (Perkins, 2010, P. 8).

2.1.4. Machine Translation

Language is an effective medium of communication. It represents the ideas and expressions of human mind. According to Tripathi and Krishna (2010), there are more than 5000 of languages in the world which reflect the linguistic diversity. It is difficult to know and understand for an individual every language. Hence, the methodology of translation was adopted to communicate the message from one language to another (Tripathi & Krishna, 2010). But as (Lewis, Gary, & Charles, 2013) states the number of languages are 7,105 identified as living languages in the world.

Research efforts have been on exploring the possibilities of automatic translation of one language (source text) to another language (target text). There have been major intuitive from various research organizations and government agencies to develop tools for automatic translation of text in order to achieve wider outreach and bridge the gap of language diversity.

As it is known, machine translation is one of the research areas in computational linguistics. According to (Tripathi & Krishna, 2010), the objective of machine translation is to restore the meaning of the original text in the translated verse. In general, the process of translation has two levels; meta-phrase (word-to-word translation) and paraphrase or gist equivalence (Tripathi & Krishna, 2010).

2.1.4.1. Types of Machine Translation Approaches

There are different methods of machine translation presented by different scholars. These may include dictionary based approach, rule based approach (direct approach, transfer based approach), knowledge based approach and corpus based (statistical

based and context based) approach. Corpus based machine translation approach has dominated over other approaches because of high level of accuracy achieved during translation (Tripathi & Krishna, 2010).

2.1.4.2. Statistical Machine Translation Approach

Tripathi and Krishna introduced the idea of statistical machine translation. These statistical methods are applied to generate translated version using bilingual corpora. Most of the research in machine translation has focused on using statistical methods on very large corpora to learn translations of words and phrases. However, more and more researchers are starting to incorporate syntax into such methods (Tripathi & Krishna, 2010).

Statistical machine translation approach can be statistical word-based model, statistical phrase-based model, statistical syntax-based model and example-based model (Tripathi & Krishna, 2010).

There are numerous applications for word alignments in natural language processing. These applications crucially depend on the quality of the word alignment (Yarowsky, David, & Wicentowski, 2000). An obvious application for word alignment methods is the automatic extraction of bilingual lexica and terminology from corpora (Smadja, Frank, Kathleen, McKeown & Hatzivssiloglou, 1996).

Statistical alignment models are often the basis of single-word-based statistical machine translation systems (Berger et al., 1994). In addition, these models are the starting point for refined phrase-based statistical or example-based translation systems (Franz & Hans, 1998). In such systems, the quality of the machine

translation output directly depends on the quality of the initial word alignment (Franz & Ney, 2000).

2.1.5. Related Works

This part focused on what was done on Kambaatissata which was reviewed as resource for this research work, on bilingual electronic dictionary and on parallel text alignment. As much as the researcher's knowledge goes, the following works were done.

Kambaatissata-Amharic dictionary was developed by the Kambaata Tambaaro Zone Information and Culture Main Department (2005). The dictionary is in the printed copy; not in electronic form. It gives the translation from Amharic to Kambaatissata. It has 13,080 entries.

Even though it is not published, the grammar and structure of Kambaatissata was prepared by Philipos entitled: *SIR: Kambaatissa Afee Seerahaa Rogahaa* (Philipos, 2012). In his work, he used Latin script for the Kambaatissata language as other Cushitic languages like Sidaama and Affan Oromo. He presented vowels of Kambaatissata as five in number and as they can be doubled to indicate the length. Concerning the consonants, he put as there are 27 in number. And also he extended his work to syntax level.

English-German Bilingual machine readable dictionary was designed by taking two corresponding texts (English and German) and developed algorithm to determine lexical alignments by using statistical methods over texts combined

with the optional support of an MRD (Robert, Russell, & Warwick, 1989). This work is related with this research work on its development algorithm.

A Manipuri-English bilingual electronic dictionary was designed and implemented. For the work, the data was collected from many sources including daily newspapers, weekly and monthly journals, and Manipuri Text Corpus which was created in Manipuri University (Poireiton, Ningombam, Mamata, & Syam, 2012). This work is related with this research on data collection methodology.

Hansard bilingual corpus was created for the purpose of building machine translation system using purely statistical techniques. Such a system is made using exclusively statistical non-linguistic methods to induce translations (Brown, 1993). Though this current research used statistical method as they did, there is a difference of the output; bilingual electronic dictionary for this research and parallel corpora for their work.

English-Japanese parallel corpora were constructed and tagged. They prepared Japanese translations of sample English abstracts to make it easier for users to search for a good model of both their target abstract and their target component sentence. Japanese equivalents were voluntarily constructed on a sentence-to-sentence basis by Ricoh's software engineers; thereby aligning English-Japanese sentence pairs of sample abstracts was done manually. They used both manual alignments as well as automatic one (Narita, nd.).

The English-Afrikaans parallel corpora were created by Draghoender and Kanhov (2010). They prepared with the purpose of bilingual dictionary creation. They created English-Afrikaans corpora manually by gathering parallel texts and compiling them into one raw text file. They used Uplug in combination with the parallel corpus to generate an English-Afrikaans dictionary. In their work, the single characters which are not words were removed; numbers and punctuation marks. They used Google Translate as an evaluator (Draghoender & Kanhov, 2010). Their work is somehow related with this research on the data preprocessing steps.

Polish-Basque bilingual dictionary was constructed semi-automatically based on WA of parallel corpora by (Pietrzak, 2009). They used three free word alignment tools to perform word alignment. These are PLUG Word Aligner (PWA) which comprises two word alignment systems, NATools which are designed to create bilingual dictionaries using statistical methods and GIZA++.

Amharic-English, English-Amharic Multimedia Dictionary was developed by Ahmed for master's degree at Addis Ababa University. The dictionary system developed was to facilitate and collect requests from different dictionary users. The dictionary system was put in the database of the SQL Server 2005. To display Amharic text online they have used WEFT tool. They used ASP.Net for the system development (Ahmed, 2008). The relationship between the project work done by Ahmed and this research work is only on its output being bilingual electronic dictionary.

English-Macedonian machine readable dictionary created by using parallel corpora by Saveski and Trajkovski in the Faculty of Computing, Engineering and

Technology, Staffordshire University, and Faculty of Electrical Engineering and Information Technologies, Ss. Cyril and Methodius University. The bilingual dictionary was generated based on parallel corpora of English and Macedonian. To do so, they used statistical machine translation approach and manual evaluation method with volunteer fluent speakers of both English and Macedonian. They have got the accuracy of 79.8% (Saveski & Trajkovski, 2010). Their work is related with this study much more than others. This is because their bilingual dictionary generation method is based on parallel corpora using statistical machine translation approach and the evaluation method they used is related with this study; manually by domain experts.

Yang and Li have constructed automatically English-Chinese Parallel Corpora. In their work, the alignment between English and Chinese words starts from simple English words. Each simple word of English is translated into a set of Chinese words through dictionary lookup. Each of these possible translations will then match with the characters in the Chinese (Christopher & Wing, 2003). Their work is related with this study only by being alignment. For this research, the alignment was on the word level but for their work it was in the sentence level.

CHAPTER THREE

3. KAMBAATISSATA

3.1. Introduction

The Kambaata live in the Southern Nation Nationality of Ethiopia in Kambaata Tambaaro Zone. Their language is referred to as Kambaatissata. Kambaatissata is a Highland East Cushitic language, part of the larger Afro-Asiatic family and spoken by the Kambaata. Dialects are Donga, Kambaata and Tambaaro (Korhonen, Mirja, & Ronald, 1986).

Kambaata is located in the southern region of Ethiopia, bordered by Wolayta, Hadiyya, Tambaaro and Halaaba. Scientists categorize the Kambaata as Highland East Cushitic ethnic group. The area where they live is about 1,040.39 km² (Treis, 2008).

In spite of its importance as a vernacular that is widely spoken (nearly one million) it lacked a well developed literature. Before the Revolution of 1974, publishing and broadcasting in Kambaatissata was very limited. But after 1974 Revolution, the government undertook a literacy campaign in several languages, one of them being Kambaatissata.

Now a day, Kambaatissata is the instructional medium for primary schools throughout the Kambaata Tambaaro Zone except the Tambaaro Woreda. Moreover, some literary works, newspapers, magazines, education resources, official documents and religious writings are written and published in Kambaatissata.

3.2. Orthography

Before Kambaatissata was introduced as medium of instruction in primary school in 1986 E.C. (1993), its orthography in the Latin script was developed, which largely follows the orthography of Oromo (Griefenow-Mewis, 2001) as cited in (Treis, 2008). The Kambaatissata orthography is very similar to the Sidaama orthography, (Yri, 2004) as cited in (Treis, 2008). The description of the orthographic conventions is predominantly based on (Matewoos, 1992) as cited in (Treis, 2008) and supplemented by information from primary school books (Treis, 2008).

The Kambaatissata data in this paper were written in the official orthography produced by (Maatewoos, 1992) as cited in (Treis, 2008). The following graphemes are not in accordance with the IPA conventions: <ph> = /p'/, <x> = /t'/, <q> = /k'/, <j> = /dʒ/, <c> = /tS'/, <ch> = / ʧ /, <sh> = / ʃ /, <y> = /j/ and <'> = /ʔ/. Length is indicated by double letters, e.g. <aa> = /a:/, <bb> = /b:/, and <shsh> = /S:/ (Treis, 2008).

Due to an idiosyncratic convention, the second consonant of a glottal stop-sonorant cluster is generally written as double, although the cluster only consists of two phonemes, e.g. <'mm> = /ʔm/. Nasalization is marked by a macron, e.g. <ā> = /ã/. Word-final unstressed i does not occur orthographically, irrespective of its phonological status (Treis, 2008).

As it is mentioned in Kambaatissata-Amharic dictionary, it has thirty two letters to represent sounds in the orthography and out of them six are double letters */ph, ch, sh, ts, ny and zh/* and the rest twenty six are single letters. These are:

A B C CH D E F G H I G K L M N NY O P PH Q R S SH T TS U V W X Y Z ZH.

From those twenty six single letters, five are vowel letters. All capital letters have their equivalent small letters. The long vowel is represented by double vowels. (Kambaata Tambaaro Zone Culture Tourism and Government Communication, 1995). But there is a glottal consonant /ʔ/ represented by [‘] and used.

3.3. Morphosyntax

Morphosyntax has to do with how sounds combine to form words and sentences. The term "morphosyntax" is a hybrid word that comes from two other words – morphology and syntax.

Kambaatissata has SOV (subject–object–verb) order. The phonemes of Kambaatissata include five vowels (which are distinctively long or short), a set of ejectives, a retroflexed implosive, and glottal stop (Korhonen et al., 1986).

Kambaatissata has four open word classes; nouns, verbs, attributes (with the sub-word classes like adjectives, numerals and demonstratives) and ideophones/ interjections (Treis, 2008).

3.3.1. Noun Morphology

Nouns in Kambaatissata can be inflected to indicate agreement (number and gender), empathy, person (first, second and third), honorific and the like.

Nouns in Kambaatissata can be derived from nouns, adjectives and verbs. The de-verbal agent nominals are much more common. These can be to form agentive, action, manner and gerundive nouns.

3.3.2. Verb Morphology

Verbs in Kambaatissata can be both inflected and derived. Verbs can be inflected to indicate agreement (person, number, and gender), aspect (imperative, imperfective and negation), honorificity, etc. Verb derivation can take place to show passive, causative and the like.

3.3.3. Adjectives Morphology

Adjective is a word which is used as a modifier of a noun to denote quality, quantity, extent, or to specify the noun as it is distinct from something else. Adjectives can be inflected as well as derived for different purposes. They can be inflected for gender and number in Kambaatissata. They also derived from other parts of speech.

CHAPTER FOUR

4. Designing and Experimentation

4.1.Introduction

The previous chapter discussed corpora data collection methodology, corpora preprocessing methodology, model selection methodology, dictionary creation methodology, dictionary enhancement methodology and evaluation methodology with justification. This chapter focuses on designing a prototype for further implementation of the approach. To do so, parallel texts have been used.

Machine readable dictionary can be done manually or using automatic translation system. Manual translations are very expensive both in terms of time and manpower. Because of this reason, the researcher was interested to design electronic dictionary using machine translation approach.

This research started with collecting parallel texts and followed with preprocessing activities. These activities included data normalization (including changing letters into small letter case, removing punctuation marks and other symbols inside the text, tokenization, stop word removal and lemmatization), making sentence level alignment or creating aligned sentences and implementing IBMModel1 to do word level alignment, enhancing the created English-Kambaatissata bilingual dictionary with English WordNet and evaluating the system performance with domain experts.

4.2.Architecture of the System

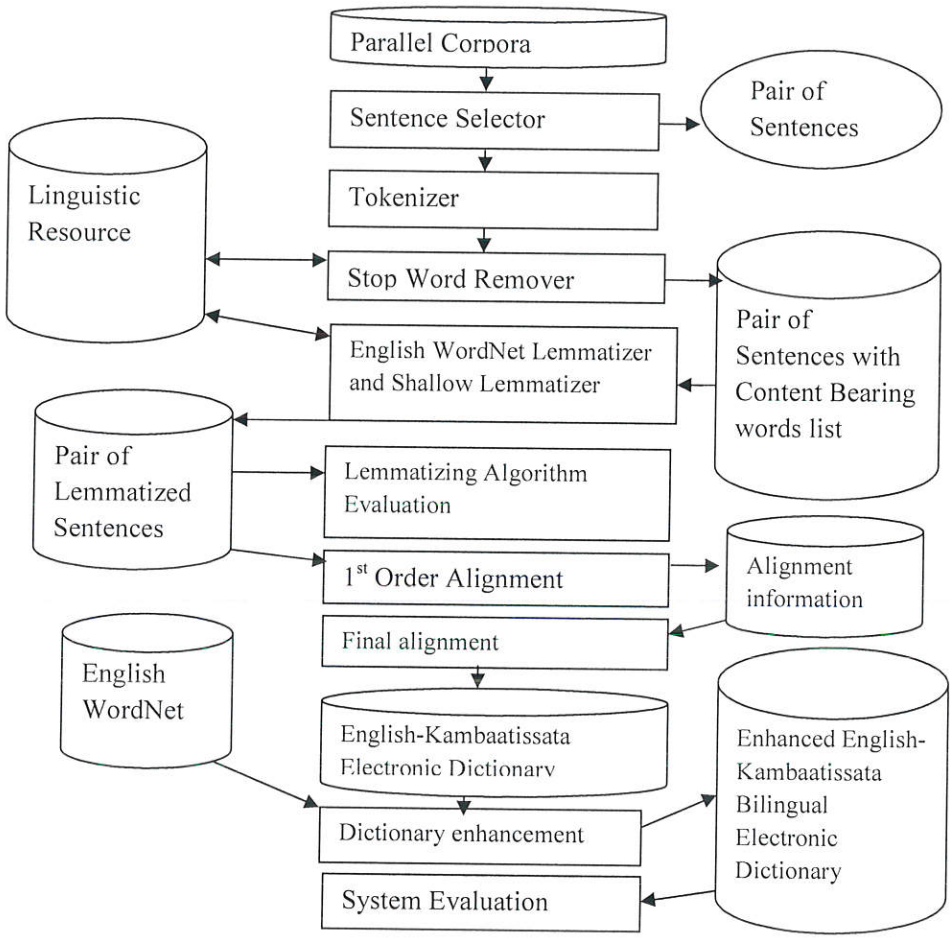

Figure 1: The Architecture of the System

4.3.Text Preprocessing

The data used for the experimentation of this research is the parallel text corpora of English and Kambaatissata. There was scarceness of parallel corpora of English and Kambaatissata. But there are some texts available as the parallel texts for both English and Kambaatissata. These texts were the Ethiopian Federal Republic Constitution of (1995), training material for teachers of mother tongue teaching in the primary schools of SNNPR Kambaata Tambaaro Zone Kadida Gamela Woreda and the Bible story book for children. The total number of data of parallel texts was 1,194 sentences.

The corpora were made to pass through a series of preprocessing steps. These were presented below step by step.

4.3.1. Changing Capital Letters into Small Cases

All letters of both language texts were changed into the small letter cases. This was done in assumption of the same word could be found in the same letter case. This was done by using lower() function of the python. As a result, the distinction between the word *The* and *the* is ignored. Changing all into small letter case was preferred since majority of letters in the written texts are in small letter case. This was done on the python programming language by having the following algorithm.

Algorithm for Changing Capital Letter Cases into Small

Input: Kambaatissata/ English/ raw texts

Output: Corpora in lower letter case

Initialize list

While not end of file of Kambaatissata/English/ text do

    if other case

    change into small letter case and put the changed token into list

    end if

end while

Algorithm1: Algorithm for Changing Capital Letter Cases into Small

4.3.2. Removing Punctuation Marks

There are different punctuation marks and symbols in written texts which are put intentionally or unknowingly. Even though they have different grammatical functions in the written texts, they cannot carry the main content of the text which is used in bilingual electronic dictionary creation. Due to that, they were removed from both Kambaatissata and English texts by having the following algorithm.

Algorithm for Removing Punctuation Marks

Input: Kambaatissata/ English/ small letter case raw texts

Output: Cleaned corpora (without punctuation marks)

Initialize list

while not end of file of Kambaatissata/English/ text do

    if punctuation marks

    remove and put the token into list

    end if

end while

Algorithm 2: Algorithm for Removing Punctuation Marks

4.3.3. Tokenization

Tokenization is the process which takes place before any real process which can be applied on the input text. Concerning this research, tokenization is done using the NLTK default (white-space) as the word delimiter. This was done using the following algorithm.

Algorithm for Tokenization:

Input: Kambaatissata/ English/ raw texts

Output: Tokenized corpora

Initialize list

while not end of file of Kambaatissata/English/ text do

    if white space

    split and put the splited tokens into list

    end if

end while

Algorithm 3: Algorithm for Tokenization

4.3.4. Removing Stop Words

There are stop words list of English words in the NLTK corpora but there is no stop words list of Kambaatissata readily available. Thus, the researcher consulted the domain experts of the language and prepared them by identifying the English equivalent from the NLTK corpora.

Kambaatissata stop words can be inflected for different purposes. Because of that, they can be found in different forms. In addition to inflecting, some of them can be found as different words with the same meaning like '***te***' and '***indo***' to represent the English '***or***' and also attached as suffix like *'martan-indo waltan' will you come or go*. These all forms of stop words were identified and removed from the raw text of

Kambaatissata. Independent stop words of Kambaatissata identified for this research are 156 in number.

In general, the identified Kambaatissata stop words include personal pronouns, demonstrative pronouns, relative pronouns, reflexive pronouns, conjunctions, prepositions, modal verbs and quantifiers.

Personal pronouns include those nine independent personal pronouns which are listed in (Treis, 2012). These include an, at, a'nnu, is, ise, issa, na'oot, a'nno'oot, and isso'oot.

In addition to personal pronouns, demonstrative pronouns were listed as stop words of Kambaatissata. These have included kaan, taan, kara, tara, kuun, tiin, kuru, tiru, kanni, tanne, karri, tarri, kannii(ha), tannee(ha), karrii(ha), tarrii(ha), kanniichch, tanneechch, karriichch, tarriichch, kanniin, tanneen, karriin, tarriin, kanneen, tanneen, karraan, tarraan, kanne, hikuuphiru, kun, tin, kuru, hika'e and hikan.

Relative pronouns used in this research as stop words include hakanisse, hakkusii, hakkada, hattita, hattiguta, hakku, ma, hakannee, aye and ayebi.

Possessive pronouns of Kambaatissata which were used as stop words in this research include kibha, ki'ineb, nibbi, kiisibiihu ki'nneeesibiihu, issebi issob isseb, issib, issiha, hikuuphiru, esaa, kun, tin, kuru, gagisib, gagisab, hika'e and hikan niibb, nibbihu, and isseha.

Reflexive pronouns of Kambaatissata which were used as stop words in this research include gaggutisse, gaggunkusse, gaggu'nkuk, gaggunku'ne, gaggunkussa, gagunkus, gagunkui, gagunnkuune and gagunkus.

Another parts of speech which were listed as stop words of Kambaatissata in this research include the conjunctions which are found independently. These include te, indo, amo, amonichch, amonsi, hittajata, hikkanin, birssi, hikkajata, hittajata, tessu, ikkodaa, barggammi, ikkodaan, anannagin, te, illanqaxe, illanqaxechch, hittiki hikkanniichch awwanni, teesuu, bargginini, zakkiin, zakkishshi, hikkusini, hikkus, hikkanitannee, bargginini hikkanniga hittigunta ikkoda,

Prepositions of Kambaatissata were listed as stop words in this research. This was because they were listed as stop words of English in the NLTK. These include alin, worron, azin, biren, etarin, bizzi, ale, mererroo, haddan, aaze aazeeni, qunxa, qunxaan and woroon.

Modal verbs and be verbs were used as stop words in this research. These include yos, yose, yosa, atto'u, danddano dandditaau danddeeno, a'aa and aa.

Quantifiers of Kambaatissata were used as stop words in this research. These include horru, horunku, xali, xalla, anuku-anuku, amant-amant, qakichchu kakichchu qahu, batinnaashsha and hollama.

The stop words of English were retrieved from the NLTK corpus stop words list. These function words were omitted on the assumption that they cannot carry the main content of the text and they were not used as the dictionary terms in this research. The retrieved English stop words were 128 in number. From those retrieved 128 stop words of the NLTK resources, '***don***, ***s***, and ***t***' were not removed from the texts of English. This was because they are not convincing stop words. But, the rest 125 stop words were used as stop words of English and removed from the English text.

In general, 156 stop words of Kambaatissata and 125 English stop words were used in this research in order to not be used as the content bearing words. To remove the stop words of both English and Kambaatissata texts, the following algorithm were used in the python programming language.

Algorithm for Stop Words Removal

Input: List of tokenized Kambaatissata/English/ raw text

Input: List of tokenized Kambaatissata/English/ stop words

Output: Corpora of content bearing words

Initialize integer n=zero

For all lists of stop words:

For all lists of Kambaatissata/English raw text:

If word/token in list of stop word:

Assign index of token in the lists of Kambaatissata/ English raw text to n

Delete token with index n from list of Kambaatissata/ English raw text

Algorithm 4: Algorithm for Stop Words Removal

4.3.5. Lemmatization

Lemmatization is the process of changing different variants of a word into the same word root form. There is WordNet Lemmatizer to find lemmas of English.

For Kambaatissata the lemmatizing algorithm was done by using the following suffixes to be removed. These include */–ssa, -kkaata, -nta, -nka, -hu, -aannu, -*

a'yyoo, -amayyoou, -aagu, -ooiichch, -ammoochch, -oo'iichch, -ammeehaa, -umbua, -umba'a, -chuta, -ichchuta, -chchta, -ndo, -s, se, -chu, -chut, -ichchuta, -akat, -ooiichch, -íichch, -am, -aada, -teen, -siis, -aan, -ichchu, -chchu, -nta, -chuta, and -íin. The lemmatization for Kambaatissata was applied using the following algorithm.

Lemmatizing Algorithm for Kambaatissata Raw Text:

Input: List of tokenized Kambaatissata content bearing terms

Output: List of lemmatized Kambaatissata content bearing terms

Initialize suf= 'ssa' or 'kkaata' or 'nta' or 'nka' or 'hu' or 'ammeehaa' or 'umbua' or 'chuta' or 'ichchuta' or 'ammoochch' or 'a'yyoo' or 'ooiichch' or 'aannu' or 'íin' or 'aagu' or 'amayyoou' or 'oo'iichch' or 'chchta' or 'ndo' or '*s*' or '*se*' or 'chu' or 'chut' or 'akat' or 'ooiichch' or 'íichch'

For all lists of Kambaatissata words:

If a word has suffix suf:

Strip a suf from a word.

End if

Algorithm 5: Algorithm for Lemmatizing Kambaatissata Raw Text

4.4.Evaluating Lemmatizing Algorithm

The output of the preprocessed parallel corpora was tested manually. For the testing purpose 10% (60 sentences) or 539 terms for English and 430 entries of Kambaatissata texts were used by following systematic random sampling technique. The number of terms to be evaluated were not the set or unique vocabularies but with

repetition. Based on the experiment 72.5% for English and 70.1% for Kambaatissata terms were lemmatized correctly. The 24.1% for English and 22.1% for Kambaatissata terms were retrieved as they were (not passed through lemmatization process) and the rest 4.4% of English and 7.8% of Kambaatissata terms were lemmatized wrongly. These results were satisfying and encouraged further experiments of the word alignment.

4.5.Sentence Level Alignment

Much work has been reported in sentence alignment using different techniques. These techniques include sentence length or word correspondences. Length-based approaches use the longer sentences in one language to be translated into longer sentences in the other language and shorter sentences to be translated into shorter sentences. In the word correspondences approach, probabilistic score is assigned to each proposed correspondence of sentences, based on the scaled difference of lengths of the two sentences (in characters) and the variance of this difference (William & Kenneth, 1993).

Sentence-length-based methods are relatively fast and fairly accurate. Word-correspondence-based methods are generally more accurate but much slower, and usually depend on bilingual lexicon.

In Kambaatissata, a single word may be translated into a phrase or a sentence. For example,

English: '*His brothers and sisters*' can be translated as:

Kambaatissata: '*Hizaakas*'.

Short sentence in Kambaatissata may be translated into long sentence in English and vice versa. For example,

English: '*He called his sisters, brothers, sons and daughters for his brother's wedding ceremony.*' This can be translated into:

Kambaatissata: '*Hizes bolochchi osusi hizakas ga'e'u.*'

The above examples show that the number of tokens or length of sentences is not equal. This indicates that it is not possible to apply sentence length algorithm for word alignment. There are word alignment tools like NATools and GIZA++. But they need aligned sentences with equal tokens each. As a result, applying word-correspondence technique will be effective than length based. Therefore, word-correspondence approach was preferred for this research. This is because it does not need knowledge of the languages or the corpus beyond division into words and sentences.

The raw text of both languages have already tokenized and normalized. What left was constructing aligned sentences which were ready to be trained on IBMModel1. This was done by writing the following algorithm on python programming language.

Algorithm for Sentence Alignment

Input: List of sentence of parallel Text of Kambaatissata and English language

Output: List of aligned Kambaatissata and English sentences

Opening file where aligned sentence will be stored

Writing string [AlignedSent(to file

Initializing counting integer j=1

```
for all token in list of lines
while j is less than<3 do
    if j is equal to 1 do
    write sentence tokenized by white space into file
     write comma /,/ to file
     make the value of j=2
     break
     if j is equal to 2 do
   write Kambaatissata sentence tokenized by white space into file
   write comma as string , to file
   write string AlignedSent ( to file
    make the value of j=1
    break
end while loop
write string ] to file
end for loop
close file
```

Algorithm 6: Algorithm for Sentence Alignment

4.6. Word Alignment

Since the main objective of this research is to design bilingual machine readable dictionary of English-Kambaatissata from bilingual corpora, it is vital to think of machine translation approaches for a lexical translation. Here, there is no need of grammatical information to implement word alignment. For this research word-translation model, the well-known IBM Translation Model1was preferred.

4.7. IBMModel 1 Implementation

As it is known that in each language the number of words in a sentence and the order of words is governed by the language. Kambaatissata has SOV (subject–object–verb) order. But the English language follows SVO (subject-verb-object) order. When we align words from both languages, we have gotten the challenge of not having equal number of words and tokens in both sentences of both languages. In order to solve the problem, the IBMModel 1 was used.

In IBM Model 1 all alignments have the same probability by using a uniform distribution. Hence, the word order does not affect the alignment probability. It describes the essence of statistical alignment as trying to model the probabilistic relationship between the source language string S and the target language string T.

There are many ways we could define P(S/T) when P = probability, S = source language word, and T = target language word. A very simple but natural model is one based on lexical translation that is word-to-word translation.

IBM Model 1 assumes that each word in a given language is a translation of exactly zero or one word of the target language, (Koehn, 2010). That is why this model was selected and implemented in this research.

In a parallel text (or when we translate), we align words in one language with the words in the other. There is formalizing alignment with an alignment function here. Mapping an English target word at position i to a Kambaatissata source word at position j with a function a: i→ j. We would like to estimate the lexical translation probabilities t(e|k) from a parallel corpus but we do not have the alignments. If we had the alignments, we could estimate the parameters of our generative model. If we had the parameters, we could estimate the alignments. The data is not complete by itself. If we had complete data, we could estimate a model. If we had a model, we could fill in the gaps in the data. This problem leads to the expectation maximization of the IBM Model 1.

Expectation maximization (EM) in its initial stage assigns equal probability for all words uniformly. The algorithm runs upon a sentence-aligned parallel corpus and generates word alignments in aligned sentence pairs. The process is divided into 2 main stages.

Stage 1: Studies word-to-word translation probabilities by collecting evidence of an English word were the translation of a Kambaatissata word from the parallel corpus.

Stage 2: Based on the translation probabilities from Stage 1, generates word alignments for aligned sentence pairs. For example:

- initialize model parameters (uniform);
- assign probabilities to the missing data;
- estimate model parameters from completed data;
- iterate steps 2-3 until convergence;

Example: The English and Kambaatissata pairs of sentences are aligned below:

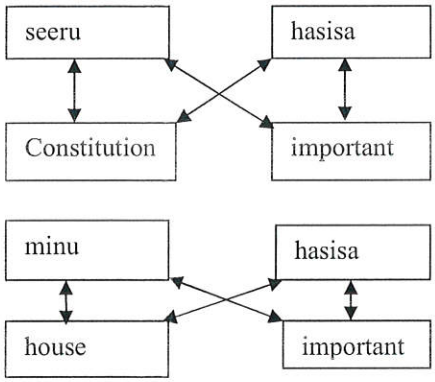

Figure 2: The English and Kambaatissata Pairs of Sentences Alignment Example

In the initial step all alignments equally likely. Then, the model learns that ***seeru*** is often aligned with ***Constitution, minu*** is with ***house***, and ***hasisa*** is with ***important*** and do in the same way for the whole words throughout the whole corpora and finally creates a pigeon hole principle or a model. This reveals the inherent hidden structure. This is done by counting the probability for the words alignment and then creating the translation model.

EM Algorithm consists of two steps. The first one is expectation step which applies model to the data. Parts of the model are hidden; here they are alignments. In this step, the algorithm assigns probabilities to possible values using the model. The second step is the maximization step. It estimates the model from data. To do so, it takes assigned values as fact by collecting counts (weighted by probabilities) then estimates the model from counts. It iterates these steps until convergence done finally.

In general, it needs to be able to compute probability of alignments in the expectation step and count collection in the maximization step.

4.8.English-Kambaatissata Bilingual Electronic Dictionary Creation

After the implementation of the IBM model 1, the next step was making a Kambaatissata-English or English-Kambaatissata bilingual dictionary. This was done by having an alignment of English word with maximum probability of the Kambaatissata equivalent word. This was done by having the following algorithms which indicates the maximum probability word as the translation of English-Kambaatissata

Algorithm for Constructing Dictionary from Results of IBMModel1 Training

Input: List of aligned sentences

Output: English-Kambaatissata dictionary

Opening file where dictionary will be stored

for all list of English words:

initialize m=0

for all list of Kambaatissata words:

if alignment probability of English word to Kambaatissata word > m:

m = alignment probability of English word to Kambaatissata word

write newline to the file

write English word to file

write single space to file

write Kambaatissata word to file

end if

end for

end for

close file

Algorithm 7: Algorithm for Constructing Dictionary from Results of IBMModel1 Training

4.9.Enhancement of Kambaatissata-English Electronic Dictionary with English WordNet

Most bilingual dictionaries do not contain the detailed information that is necessary for translation, (Deshmukh, D. L. 2011). In order to solve such problem, the enhancement is done using the English WordNet.

In the previous stage of the dictionary making what is done is mapping one word of Kambaatissata with its equivalent in English. This is not as such satisfactory since it loses many other dictionary features. In order to resolve this problem, mapping of the created dictionary with the English WordNet was done. This enabled the created dictionary to have additional dictionary features. These extracted features were glosses and POS. This extracted feature of the dictionary from the English WordNet has enhanced the created bilingual dictionary of Kambaatissata-English.

4.9.1. Extracting Gloss from WordNet

WordNet is a large lexical database of English and it is publicly available. In it, nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (SynSets), each expressing a distinct concept. A SynSet contains a brief definition ("gloss").

From the objective of this research, extracting gloss from WordNet for each automatically aligned English term to enable the English-Kambaatissata machine readable user to understand the meaning of English word with different context. So, the gloss of each English term is extracted using the algorithm presented below.

Algorithms for Extracting Glosses from English WordNet

Input: list of English-Kambaatissata dictionary words

Output: English-Kambaatissata dictionary with parts of speech

Initialize dict = {}

for all list of English-Kambaatissata dictionary:

    put English word as a key and Kambaatissata word as value into dict

for all key and value in dict:

    for all sense of key in WordNet:

    extract definition of sense from WordNet

Algorithm 8: Algorithm for Extracting Gloss from English WordNet

4.9.2. Extracting POS from English WordNet

The majority of the WordNet's relations connect words from the same part of speech (POS). That is because WordNet consists of four POS; nouns, verbs, adjectives and adverbs. Here, these four POS were extracted for each dictionary term from WordNet. The algorithm for extracting POS for each English dictionary term from WordNet is presented below.

Input: list of English-Kambaatissata dictionary words

Output: English-Kambaatissata dictionary with English gloss

Initialize dict = {}

for all list of English-Kambaatissata dictionary:

    put English word as a key and Kambaatissata word as value into dict

for all key and value in dict:

    for all sense of key in WordNet:

        extract parts of speech of sense from WordNet

Algorithm 9: Algorithm for Extracting POS for each Dictionary Term from WordNet

4.10. System Evaluation

There are two widely used ways of evaluation for automatically extracted dictionaries. These are automatic comparison with existing electronic dictionary and manual evaluating a subset of the dictionary terms. Since, there is no English-Kambaatissata machine readable bilingual dictionary, the second technique was selected for the purpose of evaluation of the English-Kambaatissata bilingual dictionary extracted during this study. This was done by four language experts who can speak and write both languages.

From the 1365 total vocabularies 20 percent (273) dictionary entries were evaluated. The selection of the entries were done randomly form the dictionary by picking

every 5th entry. Each selected entry from the subset was assigned '**correct (C)**' or '**somewhat correct (SC)**' or '**wrong (W)**'.

There is a result of the evaluation below in the graphical representation.

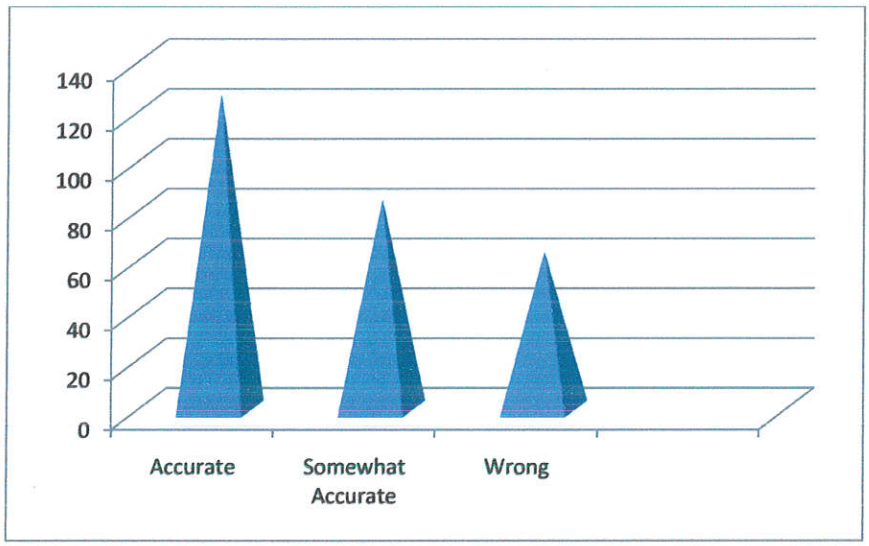

Figure 3: Results of the Manual Evaluation of the Extracted Dictionary

With the assumption of giving a single measure of the accuracy of the dictionary generated, it was combined the results by using the following formula (Charitakis, 2007).

$$\text{Accuracy} = \frac{\text{Correct Translations} + 0.5 * \text{Somewhat Correct Translations}}{\text{Number of Translation Evaluated}}$$

For example, if there are six translations, two are accurate, two are somewhat correct, and the last two are wrong, then the accuracy will be (2+0.5*2)/6=3/6= 50%.

$$\text{Accuracy} = \frac{126 + (0.5 * 84)}{273} = 0.6153 = 61.5\%$$

By using this formula, it was concluded that the accuracy of the extracted dictionary was 61.5%.

CHAPTER FIVE

5. CONCLUSIONS AND RECOMMENDATIONS

5.1. Conclusions

Before any experimentation stage, the gathering of parallel corpora of English-Kambaatissata was done. The collected corpora were not ready for bilingual electronic dictionary creation and in general for NLP application. Because of this reason, it passed through series of preprocessing. Then, alignment of the source language texts into its equivalent parts in the target language was done. This was done in two levels; sentence level and word level. Sentence level alignment was done by using the python programming language module called *align*. On the other hand, word alignment or dictionary creation was done by using purely statistical machine translation tool the IBMModel 1 of one of the python models. The tool was selected because of its application that it is not dependent on linguistic information like word order and number of words in the source language sentence and in the target language sentence. It simply makes word alignment based on probability and for this research it was assigned to use maximum probability to be the alignment. After dictionary generation, its enhancement took place.

The generated bilingual electronic dictionary had no other dictionary features like POS, glosses etc. rather it was simply mappings of word equivalents. In order to get the missed dictionary features, enhancing the dictionary entries with English WordNet was done. After doing so, performance measuring of the bilingual electronic dictionary performed by looking the accuracy of the system output or generated dictionary entries with domain experts who can speak and write both

languages. As the evaluation result indicates, the system translated 61.5% accurately.

This accuracy measured value indicated that the system missed 38.5% of the data to translate correctly. This can tell us that there were limitations for the research. This can lead us to see each steps of designed English-Kambaatissata bilingual electronic dictionary. There were limitations identified as the following:

- There was lack of standard parallel corpora. This means there was no parallel corpora of large in number and aligned in the sentences level. If so it would be easier to generate bilingual electronic dictionary with better performance and following simple steps.
- The English WordNet lemmatizer which was used to lemmatize the English text could not handle the whole text properly. It failed to handle 27.5% of the data correctly. There were the same shortcomings of the shallow lemmatizer designed by the researcher to lemmatize Kambaatissata text. It lost 29.9%.
- Every language can be governed by its own. Due to that there were places which were not answered by IBMModel 1. Since it does not consider any linguistic information.

In general, bilingual electronic dictionary is one of the very important linguistic resources. It is used for cross-lingual information retrieval, machine learning and the like applications in addition to its indirect applications of overall development of the nations. In addition to this importance, it is very useful for the people who like to make a study on the area of bilingual electronic dictionary as well as related

like applications in addition to its indirect applications of overall development of the nations. In addition to this importance, it is very useful for the people who like to make a study on the area of bilingual electronic dictionary as well as related linguistic resources like WordNet and thesaurus by having this study as motivation as well as by using the algorithms in some ways.

6.2. Recommendations

Research is a means to give solutions for the world's everyday problems. Due to that, all stake holders should take part directly or indirectly in the research work. The current study hopefully, has demonstrated the possibility of developing and it is a promising field for future improvements. The area is just at its initial stage. Developing this area requires the collaboration of researchers and funding organizations. Based on the result of the current study, it is important to recommend the following points for future investigation and collaborative work.

- Generating automatic electronic dictionary needs standard parallel corpora for testing and making experimentation. But there is no standard parallel corpora developed yet. This is an area of future research. Thus, researchers, government and funding organizations should work together to develop standard parallel corpora for English-Kambaatissata.
- The shallow lemmatizing algorithm used in this study doesn't work well for every words in the language. Additionally, there are conditions which never addressed by the shallow lemmatizing algorithm. Hence, there should be further study to come up with better Kambaatissata lemmatizer.
- The English WordNet lemmatizer used in this research cannot work for all English words. It is a good research area to improve English WordNet lemmatizer.

- The study was conducted for the first time by using statistical machine translation approach of English-Kambaatissata bilingual electronic dictionary. It wasn't compared to any other bilingual dictionary designed by this approach. Further study is needed to figure out the better approach that works for automatically designing English–Kambaatissata bilingual electronic dictionary.
- To enhance Kambaatissata word senses, it is important to have Kambaatissata WordNet. Therefore, future projects could be investigated to fill the gap.
- The accuracy of the designed prototype was 61.5%. To improve its accuracy, the questions like word order, inflectional paradigms, morphological complexity, etc. which are vital importance for the final success should be answered. Therefore, designing the English-Kambaatissata bilingual electronic dictionary considering word order and morphological complexity is very important future research area to come up with.

In general, government, researchers, non-governmental funding organizations and the society in general should involve in solving the problem by contributing what we can.

References

Aas, K. & Eikvil, L. (1999). Text categorization: A survey. *Technical Report 941.* Norwegian Computing Center.

Ahmed Aragaw (2008). Online Amharic-English English-Amharic Multimedia Dictionary (Master Thesis). Addis Ababa University. Addis Ababa, Ethiopia.

Alemayehu, N. & Willett, P. (2002). Stemming of Amharic words for information Retrieval. *Literary and Linguistics Computing. 17 (1),* 1-17.

Al-Onaizan, Y. U., Germann, U., Hermjakob, K., Knight, P., Koehn, D., Marcu, D. & Yamada, K. (2000). Translating with scarce resources. *In Proceedings of the 17th National Conference on Artificial Intelligence*, 672–678.

Amsler, R. A. (1980). The structure of the merriam-webster pocket dictionary (Ph. D. Dissertation). Texas University. Texas at Austin.

Amsler, R. A. (1982). Computational lexicology: A research program. *In American Federated Information Processing Societies Conference Proceedings*. National Computer Conference.

Baeza-Yates, R. & Ribeiro-Neto, B. (1999). *Modern information retrieval*. Addison Wesley Longman, Boston.

Bechofer, S. & Goble, C. (2001). Thesaurus construction through knowledge representation. *Data and knowledge engineering, 37,* 25–45

Belachew G. T., (2002). *The impacts of the socio-culturalstructures of the Kambata on their economic development*. OEFSE: Vienna. 127

Berger, A. L., Peter, F., Brown, S. A., Pietra, D., Vincent, J., John, R., Gillett, J. D.,

Lafferty, H. P., & Lubos, U. (1994). The candidate system for machine translation. *In proceedings of the ARPA workshop on human language technology plainsboro, New Jersey*, 157–162.

Bird, S., Klein, E. & Loper, E. (2009). *Natural language processing with python.* O'Reilly Media. Gravenstein Highway North, Sebastopol.

Birungi, P. (1995, Ed.). *Improved strategies for employment and human resource utilization: In information and documentation sector.* Strategies for Human Resource Development for Information Management in Africa, Addis Ababa: UNECA, PADIS. 49-57

Briscoe, T. (1991). *Lexical issues in natural language processing.* University of Cambridge Computer Laboratory New Museums Site. Pembroke Street, Cambridge, UK

Brown, P. F., Pietra, D. S., & Mercer, R. L. (1993). *The mathematics of statistical machine translation: Parameter estimation.* Computational Linguistics. 19(2), 263-311.

Carstairs-McCarthy, A. (2002). *An introduction to English morphology: Words and their structure.* Edinburgh University Press LTD 22 George Square, Edinburgh

Charitakis, K. (2007). *Using parallel corpora to create a Greek-English dictionary with Uplug.* Department of Computer and Systems Sciences (DSV). KTH-Stockholm University 164 40 Kista. Stockholm, Sweden.

Christopher, C. Y. & Wing, L. K. (2003). Automatic construction of English/Chinese parallel corpora. *Journal of the American Society for Information Science and Technology.* Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong, Shatin, New Territories, Hong Kong, Shatin, New Territories, Hong Kong.54(8):730–742. Retrieved from{yang, kwli}@se.cuhk.edu.hk.

Claire, A. F. & Puskas, G. (1986). *Phonetics and phonology: Reader for first year English linguistics.* University of Geneva. Updated by Cornelia Hamann and Carmen Schmitz (March 2005). University of Oldenburg.

Crystal, D. (1995). *The Cambridge encyclopedia of the English language.* Cambridge: Cambridge University Press.

Dagan, I. & Church, W. (1994). Termight: Identifying and translating technical terminology. In: *Conference on Applied Natural Language Processing.* 34-40.

Dash, N. S. (2005) *Corpus linguistics and language technology.* Mittal. New Delhi.

Deschryver, G. (2003). Lexicographers' dreams in the electronic dictionary age. *International Journal of Lexicography, 16 No. 2.* Oxford University Press.

Deshmukh, M. D. L. (2011). The pivotal role of dictionary in translation. *The Criteria: An International Journal in English.* ISSN 0976-8165 *II.* Issue, I. Retrieved from http://www.the-criterion.com

Deshpande, M. A. (2012). A survey: Structure of machine readable dictionary. *International journal of engineering and innovative technology (IJEIT).* 1, 4.

Draghoender, A. & Kanhov, M. (2010). *Creating a Reusable English-Afrikaans*

Parallel Corpora for Bilingual Dictionary Construction. Department of Computer and System Science (DSV). Stockholm University.

Etzioni, O., Reiter, K., Soderland, S., & Sammer, M. (2007). Lexical translation with application to image search on the web. In: The Proceedings of Machine Translation Summit.

Fellbaum, C. (ed.). (1998). *WordNet: An electronic lexical database.* MIT Press. Cambridge, Massachusetts:

Francis, W. N., & Kucera, H. (1964). *Brown corpus manual: Manual of information to company.* Department of Linguistics, Brown University, Rhode, Island

Franz, O. J. & Hans, W. (1998). Improving statistical natural language translation with categories and rules. *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics.* Montreal, Canada. 985–989.

Franz, O. J. & Ney, H. (2000). A comparison of alignment models for statistical machine translation. *The 18th International Conference on Computational Linguistics.* Saarbr ¨ ucken, Germany. 1086–1090.

Gangemi, A., Navigli, R., & Velardi, P. (nd.). The ontoWordNet project: Extension and axiomatization of conceptual relations in WordNet. *Laboratory for Applied Ontology*, ISTC-CNR, viale Marx *15* (00137) Roma, Italy.

Koehn, P. (2010). Statistical machine translation. Cambridge University Press. ISBN: 0521874157.

Korhonen, E., Mirja, S. & Ronald, J. S. (1986A). *A dialect study of Kambaata-Hadiyya (Ethiopia)*. Afrikanistische Arbeitspapiere. 6-41

Korhonen, E., Mirja, S. & Ronald, J. S. (1986B). *A dialect study of Kambaata-Hadiyya (Ethiopia)*. Afrikanistische Arbeitspapiere. 71-121

Lewis, M. P., Gary, F. S., & Charles, D. F. (eds.). (2013). *Ethnologue: Languages of the world (17th ed.)*. Dall as, Texas: SIL International. Retrieved from: http://www.ethnologue.com

Lovins, J. B. (1968). *Development of stemming algorithm: Mechanical translation and computational linguistics*. 22-30.

Madnani, N. (nd). *Getting started on natural language processing with python.* Retrieved from nmadnani@ets.org. McCracken, J. (2003). *Oxford dictionary of English: Current developments*. European Association for Computational Linguistics. Budapest, Hungary.

Mayfield, J. & McNamee, P. (nd). *Converting on-line bilingual dictionaries from human-readable to machine-readable form*. The Johns Hopkins University Applied Physics Laboratory 11100 Johns Hopkins Road. Retrieved from http:// www.mcnamee.com.

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

Meyer, I. (1988). *The general bilingual dictionary as a working tool in" thème"*. 368-376.

Miller, G. A. (1985). WordNet: A dictionary browser in information in data. *Proceedings of the first conference of the UW centre for the new oxford dictionary*. University of Waterloo. Waterloo, Canada.

Miller, G. A., Beckwith, R., Fellbaum, C. D., Gross, D., & Miller, K. (1990). *WordNet: An online lexical database.* 235–244.

Narita, M. (nd). *Constructing a tagged E-J parallel corpus for assisting Japanese software engineers in writing English abstracts*. Software Research Centre Ricoh Koishikawa, Bunkyo-ku. Tokyo, Japan. Retrieved from narita@src.ricoh.co.jp.Mercer, & Roossin, P. (1988). A statistical approach to language translation. *4th conference on computational linguistics*. Coling, Budapest, Hungary.

Nesi, H. (1999). Dictionaries on computer: How different markets have created different products. In: *Symposium on Language Learning and Computers Held at Chemnitz University of Technology.*

Nielsen & Sandro (2008). *The effect of lexicographical information costs on dictionary making and use. Lexikos.* 170–189.

Perkins, J. (2010). *Python text processing with NLTK 2.0 cookbook.* Packt.

Philipos P. (2012). SIR: *Kambaatissa afee seerahaa rogahaa.* Unpublished

manuscript.

Pietrzak, J. (2009). *An extraction of a Polish-Basque dictionary from parallel corpus*. Euskal Herriko Unibertsitatea/Universidad del País Vasco. The University of the Basque Country. Retrieved from Justina.o.pietzak@gmail.com

Poireiton, M. S., Ningombam, S., Mamata, D. H. & Syam, B. P (July, 2012). A Manipuri-English bilingual electronic dictionary: Design and implementation. *International journal of engineering and innovative technology* (IJEIT), *2*. Issue 1

Raghavan, V., & Wong, S. (1986). A critical analysis of vector space model for information retrieval. *Journal of the American Society for Information Science, 37, No. 5.* 279-287.

Random House Webster's Unabridged Dictionary (2nd ed.), (1987). Random House. New York, Toronto, London, Sydny and Auckland.

Ritchey, T. (1998). General morphological analysis: A general method for non-quantified modeling. *16th EURO conference on operational analysis*. Brussels. 3-4.

Robert, C., Russell, G. & Warwick, S. (1989). *Deriving translation data from bilingual tex.* Unpublished manuscript.

Russo-Lassner, G., Lin, J. & Resnik, P. (2005). A paraphrase based approach to

machine translation evaluation. *Technical Report LAMP-TR-125*, CS-TR-4754, UMIACS-TR-57, University of Maryland, College Park

Salton G. & Buckley C. (1988). Term weighting approaches in automatic text retrieval. *Information Processing and Management.* 24. 513-523

Saveski, M. & Trajkovski, I. (2010). *Development of an English-Macedonian machine readable dictionary by using parallel corpora.* Skopje, Macedonia.

Sidney, L. I. (2001). *Dictionaries: The art and craft of lexicography* (2nd ed.). Cambridge University Press. Cambridge.

Smadja, Frank, McKeown, Kathleen, R., & Hatzivssiloglou, H. (1996). Translating collocations for bilingual lexicons: A statistical approach. *Computational linguistics, 22(1).* 1–38.

Bersenholtz, H. & Trap, S. (eds). (1995). *Manual of specialized lexicography: The preparation of specialized dictionaries*. John Benjamins, Amsterdam.

Tesfaye, H. & Haile, D. (1992). *Kambata study journal.* Addis Ababa University Press. Addis Ababa.

Treis, Y. (2008). *A grammar of Kambaata (Ethiopia)*. Part 1: Phonology, Nominal Morphology, and Non-verbal Predication. Cologne: Köppe.

Treis, Y. (2011). Polysemous agent nominals in Kambaata (Cushetic). *Special issues*

of sprachtypologi and universalienforschung. 64,(4). 369-381.

Treis, Y. (2012A). *Categorical hybrids of Kambaata. 00719313*, (version 1-19).

Treis, Y. (2012B). *Number in Kambaata: A category between inflection and derivation. 00278439*, (version 1).

Tripathi, S. & Krishna, S. J. (2010). Approaches to machine Translation. *Annals of Library and Information Studies. 57.* 388-393.

William, A. G. & Kenneth, W. C., (1993). *A program for aligning sentences in bilingual corpora.*

Wu, D. & David, C. (2007). Syntax and structure in statistical translation. *Workshop at HLT-NAACL.*

Yang, Y.M. (1995). Noise reduction in a statistical approach to text categorization. In: *Proceedings of the 18th ACM International Conference on Research and Development in Information Retrieval.*

Yarowsky, David & Wicentowski, R. (2000, PP. 207–216). Minimally supervised morphological analysis by multimodal alignment. In: *Proceedings of the 38th annual meeting of the association for computational linguistics.* Hong Kong.

Annexes

Annex 1. List of English Stop Words

All	Were	Being	For	Himself
Through	Such	Had	Herself	From
Was	Yourselves	About	Ours	Do
Does	Them	Only	Because	Should
Both	At	They	Those	Nor
Themselves	Ourselves	Once	Me	Below
Until	Its	Itself	Further	Whom
Am	Your	When	Him	Above
Having	Will	Or	Then	Where
His	Again	These	What	To
With	Own	She	Most	Myself
Hers	Up	Did	Which	For
After	Against	Each	Of	Any
While	How	Some	Between	Very
So	Not	Doing	On	Who
That	Have	He	Are	Into
Be	Few	Down	This	I
Too	Here	If	Out	There
Than	Is	Her	Been	An

The	During	Over	A	Now
More	In	Both	Yourself	Where
Do	But	Yes	We	Under
Through	Can	As	You	Off
By	Yours	Their	Theirs	It
Had	Are	My	They	Before
Same	No	Him	Why	And
Other	Just	Our	All	

Annex 2. List of Kambaatissata Stop Words

kanní tanne	at	a'nnu	is	ise
karríichch	na'oot	a'nno'oot	isso'oot	kaan
taan	kara	tara	kuun	tíin
kuru	tiru	an	karri	tarrí
kanníi	kanníiha	tannee	tanneeha	karríi
karríiha	tarríi	tarríiha	kanníichch	issa
tanneechch	tarríichch	kanníin	tanneen	karríin
tarríin	kanneen	tanneen	karraan	tarraan
kanne hikuuphiru	kun	tin	kuru	hika'e
hikan hakanisse	hakkusii	hakkada	hattita	hattiguta hakku
ki'nneeesibiihu	ma	hakannee	aye	ayebi
batinnaashsha	ki'ineb nibbi	kiisibiihu	issebi	issob
hikuuphiru	issiha	esaa	kun	tin
kakichchu	kuru	gagisib	gagisab	hika'e
gaggunkussa	hikan	niibb	nibbihu	isseha
gaggutisse	gagunkus	te	indo	amo
gaggunkusse	birssi	gaggu'nkuk	gaggunku'ne	gagunkus
gagunkui	amonsi	hittajata	hikkajata	hikkanin

gagunnkuune	amonichch	tessu	hittajata	ikkodaa
barggammi	ikkodaan	anannagin	teesuu	illanqaxe
illanqaxechch	hittiki	worron	te	awwanni
bargginini	zakkiin	zakkishshi	hikkusini	hikkus
ikkoda	alin	qakichchu	mererroo	haddan
hikkanitannee	bargginini	hikkanniga	hittigunta	aaze
hikkanniichch	azin	biren	etarin	yosa
amant-amant	qunxa	qunxaan	woroon	atto'u
danddano	a'aa aa	kibha	horru	horunku
dandditaau	yos	xalla	xali	qahu
danddeeno	yose	hollama	issib	isseb
anuku-anuku	ale	aazeeni	bizzi	ku

Annex 3. Sample Aligned Sentences

[AlignedSent(['aim', 'language', 'teach', 'enable', 'student', 'four', 'language', 'skill', 'listen', 'speak', 'read', 'write'],['afoo', 'ros', 'qome', 'quuxu', 'ros', 'sholo', 'afoo', 'dand', 'gonu', 'xawaaqq', 'anabbab', 'xaaf', 'mooshsh', 'dag', 'mesheeshsh']),AlignedSent(['kambaata', 'language', 'true', 'language'],['Kambaatissata', 'afoo', 'afoo', 'garita']),AlignedSent(['student', 'able', 'use', 'four', 'language', 'skill', 'possible', 'use', 'language', 'without', 'problem'],['ros', 'sholo', 'afoo', 'dand', 'mooshsh', 'dag', 'afoo', 'hawwu', 'ta\x92mminui', 'dand', "amma'n"]),AlignedSent(['problem', 'speak', 'liste', 'first', 'language', 'challege'],['afoo', 'bac', 'afoo', 'xawaaqq', 'gonsu', 'hawwi']),AlignedSent(['challenge', 'write', 'read', 'write', 'language'],['kee\x92mm', 'afoo', 'afoo', 'xawaqq', 'afoo', 'anabbab']),AlignedSent(['make', 'first', 'language', 'skill', 'full', 'develop', 'language', 'important', 'advance', 'four', 'skill', 'language'],['afoo', 'dand', 'afoo', "le'ga", 'sholo', 'afoo', 'dand', 'mesheeshsh', 'has'])]

Annex 4. Sample Result of Aligned English-Kambaatissata Words Using IBMModel1 with their Alignment Probability

Represent	tuk	0.398768096352
Code	sera	0.499999999151
Customary	ken	0.337030021651
consider	xud	0.159295724202
chinese	le'ii	0.507720646451
declaration	chayina	0.174225526639
bear	bazi	0.302219565116
yellow	bula	0.110984449503
month	aggana	0.192692437448
four	shollo	0.511053170506
cooperative	xaaxxita	0.165282025811
skin	gogga	0.499932128537
follow	awwant	0.999999766528
loyalty	mooltooii	0.0938217159168
settlement	faqaadi	0.142857118941
logographic	lagaamu	0.957574719429
row	hoya	0.101899650352

othrwise	daqqam	0.25
pardon	faunyita	0.250034773307
solution	ke'iaa	0.500000613136

DECLARATION

THIS THESIS IS MY ORIGINAL WORK AND HAS NOT BEEN SUBMITTED

FOR A DEGREE IN ANY OTHER UNIVERSITY.

Temesgen Heliso Woymo

THE THESIS HAS BEEN SUBMITTED FOR EXAMINATION WITH OUR

APPROVAL AS UNIVERSITY ADVISORS

Sebsibe Haile Mariam (PhD)

Derib Ado (PhD)