Chapter 2

The process of succeeding can be seen as a series of trials in which your vision constantly guides you toward your target while in your actual performance you are regularly slightly off target. Success in any area requires constantly readjusting your behavior as the result of feedback from your experience.

Michael Gelb, Tony Buzan

2 Cross-Language Information Retrieval: an overview

2.1 Introduction

Due to the ongoing development of multilingual information retrieval systems, researchers within the information retrieval community are becoming increasingly more interested in the problem of Cross-Language Information Retrieval (CLIR) (Hull, 1997). The basic idea of CLIR is to cross the language boundary by providing access in one language (the source language) to documents written in another language (the target language) (Hull and Grefenstette, 1996; Oard and Diekema, 1998; Hull, 1997). As Internet resources (such as the world wide web) are increasingly available to more and more countries, researchers find it more appropriate to explore document collections that have not been written in their native language. Hull and Grefenstette (1996) also reason that "technological advances have overcome the network, interface and computer system differences that have obstructed information access in the past."

There are three main ways in which Cross-Language Information Retrieval approaches attempt to "cross the language barrier". Through

- query translation,
- document translation, or
- both (Oard, 1997).

Each of these methods will be described in more detail in Sections 2.3.1–2.3.3. This study (as described in Chapters 4 and 5) focuses on the use of query translation, and proposes various ideas to address some of the problems (such as ambiguities and vocabulary coverage) associated with dictionary-based translations (see Section 2.3.1).

2.2 Previous research – the origins of CLIR

Cross-Language Information Retrieval (CLIR) research started out with experiments using controlled vocabularies, associated dictionaries and thesauri (Pirkola, 2001). Currently, free text approaches dominate CLIR experiments. According to Pirkola et al. (2001) free text methods (see Section 2.3) can be further classified according to the resources used to cross the language boundary. This may be corpus-based resources (Section 2.3.1), or machine translation (Section 2.3.3). A significant

11

number of Cross- Language Information Retrieval approaches focus on existing linguistic resources and which are mainly machine-readable bilingual dictionaries (Section 2.3.2).

Previous research done by Hull and Grefenstette (1996) proved "that dictionary-based translation, where each term or phrase in the query is replaced by a list of all of its possible translations, represents an acceptable first pass at CLIR." In terms of the framework of controlled vocabulary systems, CLIR has been a subject of study for a long time (Hull, 1997). Earlier research by Salton (as cited in Hull, 1997) demonstrated that CLIR systems were equal in performance to monolingual performance when provided with a carefully, manually constructed bilingual thesaurus. However, Hull (1997) stated that "controlled vocabulary systems are less than equal for modern text retrieval for a number of reasons." For instance, the size and dynamic nature of modern information resources makes manual document indexing and thesaurus construction a difficult task. Oard's survey paper (as cited in Hull, 1997) extensively reviews the history of research on CLIR.

CLIR is considerably more complex than traditional Information Retrieval (IR) because some method for translating the document that needs to be queried or using document-ranking algorithms on must be developed. Several approaches have been proposed and tested, including using resources such as:

- Bilingual dictionaries (Hull and Grefenstette, 1996; Grefenstette, 1998; Ballesteros and Croft, 1996; Ballesteros and Croft, 1998b; Davis, 1998; Davis and Ogden, 1997),
- Thesauri (Mateev et al., 1996; Sheridan et al., 1997; Sheridan and Shäuble, 1997; Sheridan and Ballerini, 1996),
- Corpora (Rehder et al., 1998; Littman et al., 1998; Landauer and Littman, 1991; Carbonell et al., 1997; Peters and Picchi, 1997) and
- Machine translation systems (Yamabana et al., 1998; Gachot et al., 1998; Carbonell et al., 1997)

Some of the more recent theoretical research for using these strategies in terms of CLIR will be reviewed in this chapter.

2.3 The main approaches and strategies applied in CLIR

The basic approaches in CLIR involve

- *query translation* from the source language into the target language; and/or
- *document translation* from the target language (Pirkola, 2001).

There are several applications or scenarios in which the user of a retrieval system may be interested in finding information on (written in a language other than the user's native or preferred language). In some applications, a user may want to retrieve all possible relevant information in a multilingual text base, irrespective of the language in which the documents appear. For instance – this may occur when searching certain collections of traditional medicine information. In other cases a user may even have some language comprehension ability regarding the language the documents appear in, but the user may not have a sufficiently, rich active vocabulary in the document languages to confidently specify queries in those languages. In this instance, a cross-language search that permits the user to specify target language queries, but retrieve documents in their original language (the source language) is useful. Even if the user cannot read the retrieved documents, he/she at least has a relevant set of retrieved documents that can be manually translated.

In CLIR the main strategies for query translation are based on three different methods, which are:

- *Dictionary-based methods* with specific relevance to (bilingual) translation dictionaries (Oard and Diekema, 1998; Ballesteros and Croft, 1997 and Davis, 1996 and Hull and Grefenstette, 1996 and Ruiz and Srinivasan, 1998),
- *Corpus-based methods* (Davis, 1996 and Davis and Dunning, 1995 and Ruiz and Srinivasan, 1998), and
- *Machine translation*

A combination of these methods can also be applied (Hull, 1997). These methods will be discussed in more detail below, while also highlighting the problems with each specific method. The specific problems experienced in this particular study will further be analysed in Chapter 5, while solutions will also be proposed in context with the experiments done in this study.

13

2.3.1 Dictionary-based methods

In dictionary-based approaches, bilingual dictionaries are used for query translation. A very basic strategy for the query translation is to process the queries word-for-word and, for each source language word, look up the target language equivalents and place them into the target language query (see Figure 2.1). Retrieval results improve if the target language is structured by synonym groups (Pirkola, 1998), but these groups must each contain all the translations of a single source language word as given by the dictionary. This will be explained in more detail in Section 3.2.
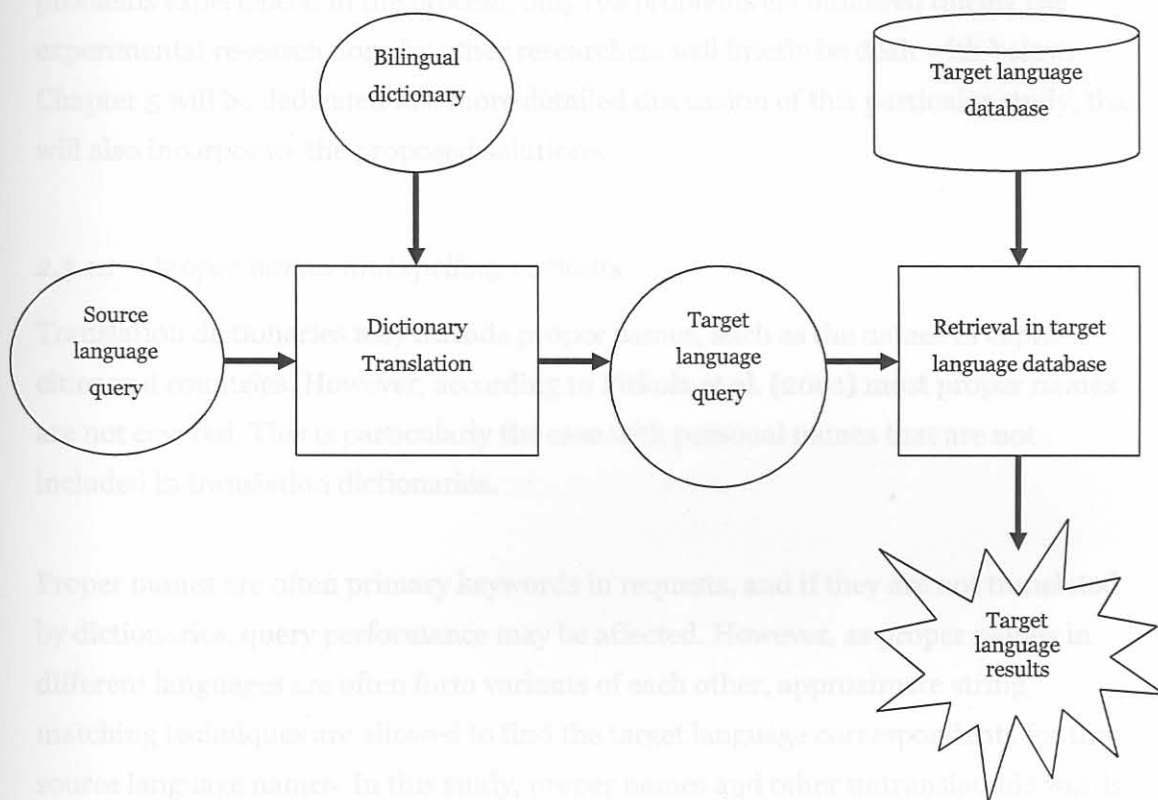
Figure 2.1 *Query translations via a bilingual dictionary* (Pirkola et al., 2001)

Several problems can be associated with the dictionary-based approach to CLIR, and Pirkola et. al (2001) identified the following as some of the most significant obstacles:

- untranslatable search keys (proper names and spelling variants) due to the limitations of translation dictionaries;
- processing of inflected words;
- phrase and compound word identification and translation; and

- lexical ambiguity in source and target languages (Ballesteros and Croft, 1997; Hull and Grefenstette, 1996).

Ballesteros and Croft's dictionary-based work tries to reduced translation ambiguity, by exploring the value of pre- and post-translation query expansion strategies. In doing so, their research supports the findings of Hull and Grefenstette (1996) that phrase translations are important to CLIR.

As the focus of this study is placed on the dictionary-based approach to CLIR and the problems experienced in the process, only the problems encountered during the experimental research done by other researchers will briefly be dealt with below. Chapter 5 will be dedicated to a more detailed discussion of this particular study, that will also incorporate the proposed solutions.

2.3.1.1   *Proper names and spelling variants*

Translation dictionaries may include proper names, such as the names of capital cities and countries. However, according to Pirkola et al. (2001) most proper names are not covered. This is particularly the case with personal names that are not included in translation dictionaries.

Proper names are often primary keywords in requests, and if they are not translated by dictionaries, query performance may be affected. However, as proper names in different languages are often form variants of each other, approximate string matching techniques are allowed to find the target language correspondents for the source language names. In this study, proper names and other untranslatable words are controlled by an advanced n-gram method (Pirkola, 2001). The n-gram method finds target language spelling variants for proper names in the source language. Proper name translation and matching in CLIR is complicated because proper names may be similarly inflected to common nouns (particularly in Finnish), and may have suffixes (representing different case and number features, as well as other grammatical categories). However, this is not the case with Zulu, because the proper name is not matched at all.

In morphologically complex languages (such as German, Dutch, Korean, Japanese, Arabic and Turkish), proper name searching in CLIR is further complicated by inflection (Pirkola et al., 2001). For example, the name *Gorbachev* is written as

Gorbatshov in Finnish, and it can also take on several inflectional forms, like *Gorbatshoville* (allative, *to Gorbachev*), *Gorbatshovin* (genitive, *Gorbachev's*).

2.3.1.2   *Special terms, technical terms and domain specific terms*

"General dictionaries often give many equivalents to a source language word, whereas special dictionaries typically give 1–2 equivalents only. The terms of special dictionaries are often unambiguous" (Pirkola et al., 2001). For these reasons, a special dictionary reduces the translation ambiguity problem. Two highly effective methods to address problems such as translation ambiguity and the lack of dictionary coverage, is <u>domain-specific</u> (for example medical, technological, scientific, social and cultural) and <u>general</u> (not limited to a single class or category) terminologies to translate the queries.

Pirkola et al. (2001) argues that "it is highly probable that the special dictionary/general dictionary approach to identifying and translating special terms in query formulation could be extended to other domains of study with great success." This is evident in the fact that many dictionaries can be used in CLIR translation systems. Although each of these dictionaries might have limited content, together they could cover different fields of study and interest.

2.3.1.3   *Word inflection*

If the source language appears in inflected form they cannot be readily translated, because they do not match dictionary headwords in base forms. A common method (called "stemming") for managing inflected search keys (and derivationally related keys) is to remove affixes from the word forms (Harman, 1991; Porter, 1980). The output is a common root or stem of different forms that may not necessarily be a real word. In lexicon-based morphological analysis, word forms are normalized into base forms that are real words. Morphological analysis also allows for the splitting of compounds into their component words (see section 2.3.1.5). This enables the matching of source language keywords with dictionary headwords (Pirkola et al., 2001). Alternatively, source language keywords and headwords can be conflated into the same form by a stemmer (Davis and Ogden, 1997). One problem related to stemming is where different headwords may be conflated into the same form. According to Hull (1997), the size of a morphological program's lexicon limits the effectiveness of morphological analysis. It is impossible to exhaustively list all the words of a language in the lexicon, as is the case with most translation dictionaries.

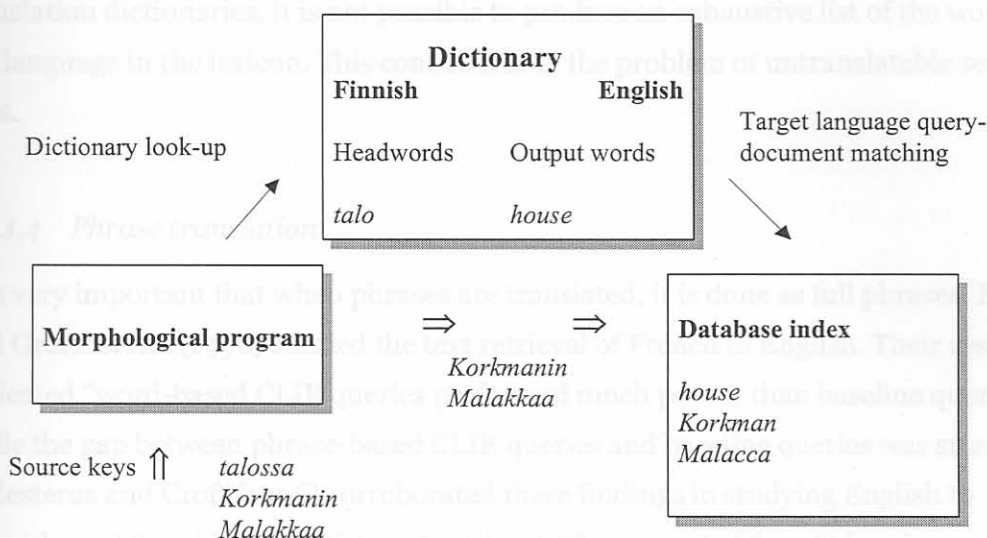This contributes to the problem of untranslatable search keys.

Figure 2.2 *Unrecognized words in morphological analysis* (Pirkola et al., 2001)

In Figure 2.2 Pirkola et al. (2001) portrays how different types of inflected words are managed from dictionary-based Finnish to English CLIR. "Most inflected keys are normalized and translated by a dictionary, like the form *talossa* (in the house) which is first normalized into *talo* and then translated into *house*. The unrecognized forms are sent unchanged into a CLIR query. The names *Korkmanin* (personal name in genitive) and *Malakkaa* (a spelling variant of the geographical name *Malacca* in accusative) represent typical cases of words not listed in the lexicon of a morphological program" (Pirkola et al. 2001). These unlisted words do not match the English index keys. Even if the word 'Malakkaa' were normalized, it would still not match the dictionary entry in the database. Pirkola et al. (2001) mentions "in dictonary-based CLIR, they could be handled similarly as untranslatable spelling variants in the case of a translation dictionary." (See Section 2.3.1.1 for more detail.) In investigating the inflection of words, it was found that CLIR effectiveness also depends on the monolingual component (i.e. the morphological processing of index keys). The effectiveness of stemming obviously depends on the language. Pirkola et al.'s research (2001) generally shows that "*recall* can be expected to improve due to stemming since a larger number of potentially relevant documents are retrieved." The research carried out in different languages (Spanish, Danish, German, French and Portuguese) has shown that stemming improved the *precision* of several of these languages.

Hull (1997) went a step further and proved that the effectiveness of morphological analysis is limited by the size of a morphological program's lexicon. As in the case of translation dictionaries, it is not possible to produce an exhaustive list of the words of the language in the lexicon. This contributes to the problem of untranslatable search keys.

2.3.1.4   *Phrase translation*

It is very important that when phrases are translated, it is done as full phrases. Hull and Grefenstette (1996) studied the text retrieval of French to English. Their research indicated "word-based CLIR queries performed much poorer than baseline queries, while the gap between phrase-based CLIR queries and baseline queries was small." Ballesteros and Croft (1996) corroborated these findings in studying English to Spanish, and Spanish to English text retrieval. They reported "a 55% loss in average precision for queries translated word-by-word compared with the original queries. A 30% loss in performance resulted from translation ambiguity and a 20% loss was due to inaccurate translation of phrases." In another study, Ballesteros and Croft (1997) showed that the correctness of the translations is just as important in automatic phrase translation. However, if phrase translation fails, phrase-based queries may perform slightly poorer than word-based queries.

"Phrases are not a major problem for languages in which multi-word expressions are compound words rather than phrases, such as German, Swedish, Finnish, and Dutch" (Pirkola 1999). However, if phrases are not identified and translated correctly, the effects on certain queries may prove to be decisive. Ballesteros and Croft (1998b) reported that retrieval performance improves when the phrases are translated by a dictionary. However, not all phrases are listed in dictionaries, which suggests the use of some additional or alternative translation method. Research by Ballesteros and Croft (1998b) and Fujii and Ishikawa (2001) showed that phrase translation, based on word collocation statistics in the target language to be useful. Fujii and Ishikawa (2001) further explored phrase translation in Japanese to English retrieval. In Japanese, as is the case with Zulu, technical terms are often phrases. New technical phrases are generated from existing words (in Zulu this is called 'Zululizing' the term), and the new phrases are not generally listed in dictionaries.

The importance of phrase translation cannot be emphasized enough in CLIR (Pirkola et al., 2001). It is not a problem for languages in which multi-word expressions are compound words rather than phrases, such as German, Swedish, Finnish and Dutch

(Pirkola, 1999). Nevertheless, if phrases are not identified and translated correctly, the effects on certain queries will be negatively impacted.
The results of Hull and Grefenstette's research (1996) done on French-English text retrieval are confirmed by Pirkola et al.'s research findings (2001). The word-based CLIR queries performed much poorer than the baseline queries, while the gap between phrase-based CLIR queries and baseline queries were small. For a review on other recent methods and results of phrase processing in CLIR, see Pirkola et al. (2001).

2.3.1.5   Compound words

Pirkola et al. (2001) defines a compound as "a word formed from two or more words that are written together." In CLIR, the distinction between compositional, non-compositional, semi-compositional phrases and compounds are important. Compositional compounds are "those compounds whose meaning can be derived from the meanings of component words" (Akmajian et al., 1990 as cited in Pirkola et al., 2001).

For instance, "the meaning of the Finnish word *kaupunginhallitus* (*city government*) comes from the meanings of the components *kaupungin* (*city*, in genitive) and *hallitus* (*government*)" Pirkola et al. (2001). As Pirkola phrased it: "in compositional compounds, a full compound typically is a hyponym (a narrower term) of its headword." This is very much the same for the Zulu language as illustrated in the following example: the meaning of the Zulu word *umakhalekhukhwini* (cellular phone) comes from the meanings of the components *u-ma-khala* (cry/ring) and *ekhukhwini* (in the pocket). Pirkola et al. (2001) refers to a compound whose meaning cannot be deduced based on its components, which is a non-compositional compound. The term semi-compositional compound refers to "a compound whose meaning is in part interpretable on the basis of the components" (Pirkola et. al. 2001).

Due to the productive nature of natural languages, words can be combined into any number of new compound words. Some languages such as German, Swedish, Finnish, Dutch and Afrikaans are characterized by a high frequency of compounds (Sheridan and Ballerini, 1996; Hedlund et al. 2001; Pirkola, 1999). For such languages, effective dictionary look-up and the searching for compound words in CLIR cannot only be solely based on full compounds, but must also be based on their component words. The decomposition of compounds and separate translation of

component words are often useful, because translation dictionaries may not include full compounds as such. According to Pirkola (1999), "the separate translation of compositional compounds will give correct senses."

2.3.1.6 Lexical ambiguity

It refers to words with the same spelling, but different meanings (homonymy) and words with many different meanings (polysemy). According to Pirkola et al. (2001) "the senses of homonyms are unconnected. A lexeme that has more than one sense is *polysemous*. The word *board*, for example, has several (sub)senses, e.g., (a) *a thin plank*, (b) *a tablet*, (c) *a table*, and (d) *food served at the table*." This example shows that the senses of a polysemous word are related to each other where one sub sense may be a metaphorical extension of another sub sense. Based on morphology, lexical ambiguity can be further divided into base form and inflectional ambiguity (Pirkola et al., 2001). First, *base form ambiguity* describes the condition where two (or more) lexemes–usually two separate headwords of a dictionary–have the same (base) form, as well as describing the condition in which one lexeme has two or more senses. *Inflectional ambiguity* again refers to a condition where two or more lexemes share at least one common inflectional form.

In CLIR, translation ambiguity is primarily caused by lexical ambiguity, which appears as an increase of irrelevant search key senses due to source and target lexical ambiguity. Ballesteros and Croft (1996), Grefenstette (1998), Hull and Grefenstette (1996) and Pirkola et al. (2001) attribute translation ambiguity and difficulty in managing phrases as the main reasons for the low effectiveness of plain dictionary-based CLIR queries.
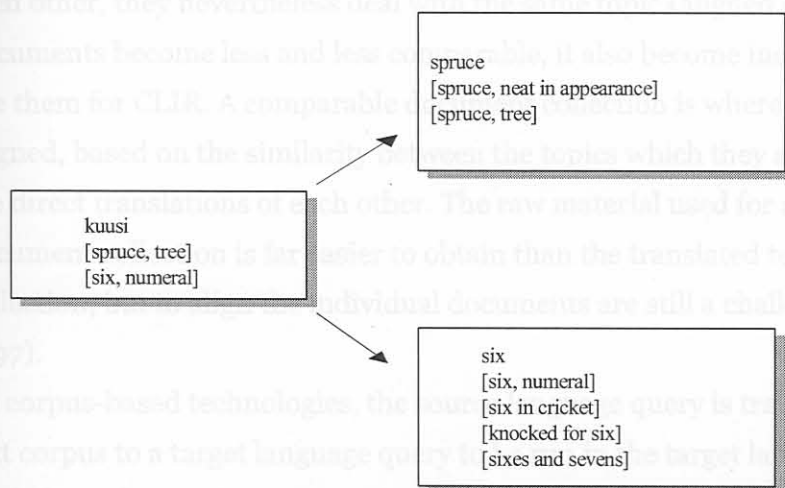
kuusi
[spruce, tree]
[six, numeral]

spruce
[spruce, neat in appearance]
[spruce, tree]

six
[six, numeral]
[six in cricket]
[knocked for six]
[sixes and sevens]

Figure 2.3 *Translation ambiguity* (Pirkola et al., 2001)

Figure 2.3 illustrates the increase of ambiguity in a translation process, as researched by Pirkola et al. (2001).

> The Finnish form *kuusi* is homonymous and has two senses: [spruce, tree], [six, numeral]. The English word *spruce* has two and the word *six* four senses (Collins dictionary, 1998): [spruce, neat in appearance], [spruce, tree], [six, numeral], [six in cricket], [knocked for six], [sixes and sevens]. Think that in monolingual Finnish and monolingual English searching the correct sense is [spruce, tree]. There is one extraneous sense in Finnish, [six, numeral], as well as English, [spruce, neat in appearance]. But in Finnish to English retrieval there are five extraneous senses, [spruce, neat in appearance], [six, numeral], [six in cricket], [knocked for six], [sixes and sevens].

From this example it can be deduced that lexical ambiguity (associated with CLIR queries) stems in part from a source language and in part from a target language.

2.3.2 Corpus-based methods

Corpus-based systems use parallel and/or comparable corpora for query translation (Hull, 1997). This method is also independent of dictionaries. A parallel corpus consists of pairs of documents, where the one document is in the source language of the user query and the other in the target language. Furthermore, the document pairs are translations of each other. Although the documents are not exact translations of each other, they nevertheless deal with the same topic (aligned corpus). Yet, as documents become less and less comparable, it also become increasingly difficult to use them for CLIR. A comparable document collection is where documents are aligned, based on the similarity between the topics which they address because they are direct translations of each other. The raw material used for a comparable document collection is far easier to obtain than the translated text used in parallel collection, but to align the individual documents are still a challenging task (Oard, 1997).

In corpus-based technologies, the source language query is translated in a parallel text corpus to a target language query to be run in the target language database (Pirkola et al., 2001).
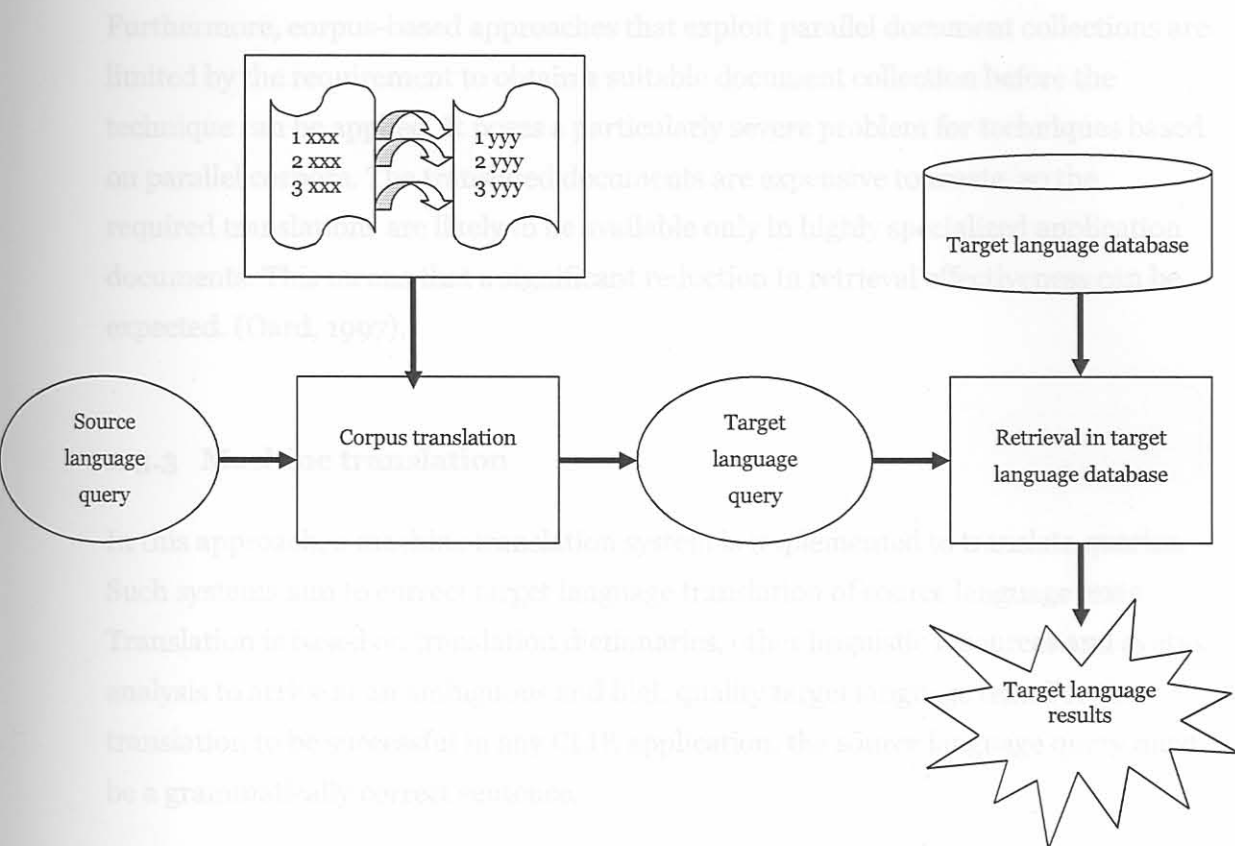
Figure 2.4 *Parallel corpora in CLIR* (Pirkola et al., 2001)

When a source language query is entered into the system, it is run against the source language documents of the parallel corpus. Best-matching documents are identified and their target language pairs retrieved. Statistical criteria and possible sentence-by-sentence alignment are used to identify best topic words to be used in the target language query. The target language query is a bag-of-words query, and run against the target language collection (Pirkola et al., 2001).

Although this approach delivers good results, it has several complications. First, the translation relationship established between parallel words in the text is usually domain dependent (for instance medical, technological, scientific, social, cultural or educational). This means "accuracy decreases outside the domain" (Davis, 1998). Second, because electronic parallel corpora is not readily available in different languages, this approach to CLIR is not practical in a South African context. It is simply too costly to acquire, because it is difficult to obtain existing translations of the correct document type. Furthermore, through the recall performance is high, the precision level is not acceptable. Because of this, the general application of this method should be demonstrated for it to become feasible. Currently, this method also remains very application dependent (Peters and Picchi, 1997).

Furthermore, corpus-based approaches that exploit parallel document collections are limited by the requirement to obtain a suitable document collection before the technique can be applied. It poses a particularly severe problem for techniques based on parallel corpora. The translated documents are expensive to create, so the required translations are likely to be available only in highly specialized application documents. This means that a significant reduction in retrieval effectiveness can be expected. (Oard, 1997).

2.3.3 Machine translation

In this approach, a machine translation system is implemented to translate queries. Such systems aim to correct target language translation of source language texts. Translation is based on translation dictionaries, other linguistic resources and syntax analysis to arrive at an ambiguous and high quality target language text. For the translation to be successful in any CLIR application, the source language query must be a grammatically correct sentence.

Arnold et al. (2001) identified three types of problems encountered in Machine Translation:

- *Ambiguity*: A word is lexically ambiguous when it has more than one meaning. An example of a lexically ambiguous word is "bank", which can refer to both a river bank and a financial institution.
- *Lexical and structural mismatches*: When two languages categorize the same word differently, a lexical or structural mismatch can occur. It is also the case when a concept expressed by multiple words in one language is expressed by one word in another language, or when a word in one language has no equivalent word or lexical unit in another language.
- *Multiword units (idioms)*: Idioms are generally phrases recognized as a unit. These phrases normally have a meaning that differs from the literal meaning of its parts taken together. An example in English would be the phrase "kick the bucket". It is common knowledge that this phrase has the same meaning as the word "die", but this meaning cannot be derived by taking the meaning of each word separately – the phrase must be considered as a whole.

Apart from the mentioned problems generally experienced with this approach, machine translation (MT) systems seem to be a straightforward choice for query translation. For each query *q*, an MT system will give a unique translation *q'* for it. In some instances, the translation is reasonable. But in other instances, the translation may depart from the original query. However, the machine translation system is not highly suitable for CLIR to use as a tool for query translation–as done in this study– for several reasons.
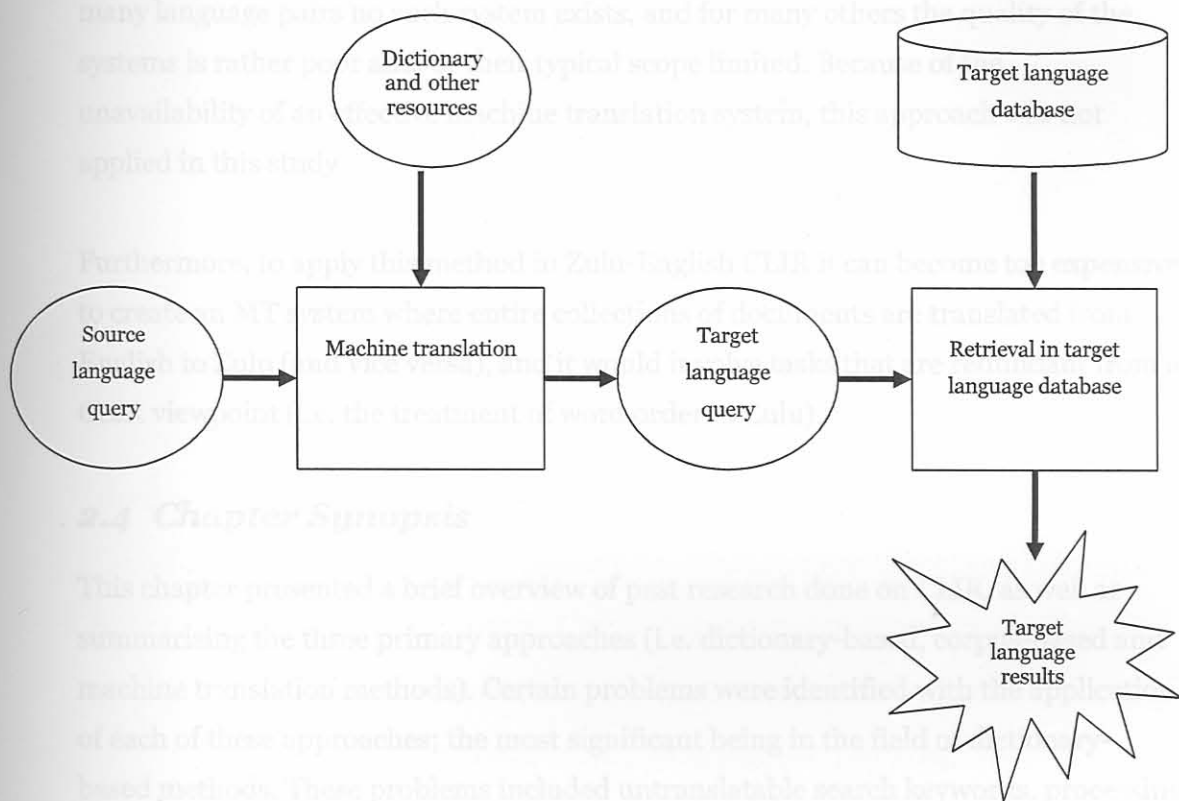

Figure 2.5 *CLIR based on machine translation* (Pirkola et al., 2001)

First, these systems put in a lot of effort trying to produce syntactically and semantically correct senses. Yet, this has no effect on current CLIR approaches, which operate on strings and not meanings (Hasnah and Evans, 1999).

Second, the machine translation system selects one translation of the word. This selection process is not only difficult, but also results in selecting the wrong target translation. In Hasnah and Evans' (1999) opinion, limiting the translation to a single specific word may result in losing several relevant documents that might contain other translations.

Also, typical queries in current IR systems tend to be very short. Therefore, the advantage of MT systems (which in principle can exploit syntactic and semantic aspects of context to improve translation) in respect to dictionary-based approaches is questionable.

Finally, another problem of this approach with regard to the South African context, is the unavailability of good machine translation systems for desired language pairs. For many language pairs no such system exists, and for many others the quality of the systems is rather poor and/or their typical scope limited. Because of the unavailability of an effective machine translation system, this approach was not applied in this study.

Furthermore, to apply this method in Zulu-English CLIR it can become too expensive to create an MT system where entire collections of documents are translated from English to Zulu (and vice versa), and it would involve tasks that are redundant from a CLIR viewpoint (i.e. the treatment of word order in Zulu).

2.4 Chapter Synopsis

This chapter presented a brief overview of past research done on CLIR, as well as summarising the three primary approaches (i.e. dictionary-based, corpus-based and machine translation methods). Certain problems were identified with the application of each of these approaches; the most significant being in the field of dictionary-based methods. These problems included untranslatable search keywords, processing inflected word-forms, identifying phrases and compound words (and translating them), and the lexical ambiguity that occurs in source and target languages. In explaining the different strategies offered by the dictionary-based approach to CLIR, it was concluded that this method would provide the best solution to the identified problems.

The next chapter will focus on Zulu as an indigenous language. The chapter will present to the reader a profile on the Zulu language, with a brief overview of the linguistic structure made up from the different noun classes and concords. The most important part of the chapter will focus on term creation in Zulu, which would help to form a better understanding of the empirical work and results obtained in Chapter 5.