Syracuse University

SURFACE

School of Information Studies - Faculty Scholarship

School of Information Studies (iSchool)

2003

# Translation Events in Cross-Language Information Retrieval: Lexical Ambiguity, Lexical Holes, Vocabulary Mismatch, and Correct Translations

Anne Roel Diekema
*Syracuse University*

Follow this and additional works at: https://surface.syr.edu/istpub

 Part of the Library and Information Science Commons

## Recommended Citation

Translation Events in Cross-Language Information Retrieval: Lexical Ambiguity, Lexical Holes, Vocabulary Mismatch, and Correct Translations (2003)

This Article is brought to you for free and open access by the School of Information Studies (iSchool) at SURFACE. It has been accepted for inclusion in School of Information Studies - Faculty Scholarship by an authorized administrator of SURFACE. For more information, please contact surface@syr.edu.

# TRANSLATION EVENTS IN CROSS-LANGUAGE INFORMATION RETRIEVAL: LEXICAL AMBIGUITY, LEXICAL HOLES, VOCABULARY MISMATCH, AND CORRECT TRANSLATIONS

by


## ANNE R. DIEKEMA

Bac., Haagse Hogeschool, 1993
M.L.S., Syracuse University, 1995


## DISSERTATION

School of Information Studies, Syracuse University

May 2003

Copyright 2003 Anne Roel Diekema

All rights reserved

## ABSTRACT

Cross-Language Information Retrieval (CLIR) systems enable users to formulate queries in their native language to retrieve documents in foreign languages. Because queries and documents in CLIR do not necessarily share the same language, translation is needed before matching can take place. This translation step tends to cause a reduction in the retrieval performance of CLIR as compared to monolingual information retrieval.

The prevailing CLIR approach and the focus of this study is query translation. The translation of queries is inherently difficult due to the lack of a one-to-one mapping of a lexical item and its meaning, which creates lexical ambiguity. This, and other translation problems, result in translation errors which impact CLIR performance.

To understand the events occurring in cross-language retrieval query translation and the relation of these events to retrieval performance, the study explored the following research questions:

1) *What kinds of translation events affect cross-language retrieval?*
2) *In what way does the presence of certain translation events in query translation affect retrieval performance?*

The study followed a two-phase multi-method approach. In phase one, a taxonomy of translation events was created through content analysis of queries and their translations in combination with an examination of the literature. In the second and final phase, a subset of the test queries was coded using the taxonomy resulting from phase one. These queries were then used in information retrieval experimentation to assess the impact of the translation events on retrieval performance.

## ACKNOWLEDGEMENTS

I would like to thank my committee members Jaklin Kornfilt, Barbara Kwasnik, Liz Liddy, Bob Oddy, and Jeff Stanton for spending time in the realms of cross-language information retrieval and statistics, and providing helpful comments and insights. It was a pleasure to work with them. I would also like to thank Jeffrey Katzer, one of my original committee members. Although Jeffrey passed away shortly before my proposal defense, his teachings kept inspiring me.

Information Retrieval (IR) is a big field and there are a number of people who have increased my understanding and influenced my thinking in this area. Liz Liddy, Arie Noordzij, and Bob Oddy were my original teachers in the field, while others like Jiangping Chen, Ted Diamond, Wen Hsiao, Wessel Kraaij, Farhad Oroumchian, Miguel Ruiz, and Arjen de Vries provided insight through discussions and work on research projects.

The IR system used in this dissertation was programmed in Perl. Many thanks go to Farhad Oroumchian and Arvind Srinivasan for helping me get my Perl legs in the early years when I did not know an array from a hash. Stéphane Dubon did great work setting up Linux boxes and networking my apartment. The readability of this dissertation was greatly enhanced by the work of Eileen Allen, Sarah Harwell, and Andrew Roginski who were all excellent editors.

I am especially grateful to Liz for providing me with a wonderful work environment for the last 7 years, where I gained experience as a researcher while working with a group of dedicated and inspiring colleagues on a wide variety of projects.

Naturally, there is more to life than working on a doctorate and I am blessed with a great set of friends who provided continuing support and welcome diversions during these many years. Thanks to Keith Berger and Bianca Flikweert for many wonderful meals, hikes, and (Dutch!) conversation. Thanks also to Marcus van Bers, Blake Rodgers, Arvind Srinivasan, and Kate Stewart for numerous social hours on the ice and the bike. Additional thanks go to Blake for simply being a joy to be around. I also enjoyed the letters and emails of Noor Evertsen and Françoise le Griep, friends who did not exactly live nearby but kept in touch nonetheless.

I could not have completed my studies without support from the home front. I am especially grateful for the company of Arie and Angus Diekema who always provided a listening ear and good company on many fabulous walks. And lastly, thanks to my (extended) family Jan Diekema, Marian Diekema-Hensums, Maurits Diekema, Myrthe Diekema, Simone Lücker, Aafke Stalman, and Lenie de Vries for taking an interest in my work and believing I could actually do this.

I dedicate this dissertation to my parents Jan and Marian, who have always stressed that getting educated is never a waste of time.

Financial support for this dissertation was provided through Beta Phi Mu in the form of a Eugene Garfield Doctoral Dissertation Fellowship and by the Institute of Scientific Information (ISI) in the form of a Doctoral Dissertation Proposal Award.

Anne Diekema
Syracuse, New York
June 30, 2003

## TABLE OF CONTENTS

## TABLES AND FIGURES

# 1 Introduction to the Study

*Die door de wereldt sal gheraken*
*Die moet connon huylen metten honden*
*Ende moet oock connen diverssche spraken*
*Die door de wereldt sal gheraken*[1]
*Anthonis De Roovere (ca. 1430-1482)*

## 1.1 Introduction

In our information-based society, trends towards globalization continue to diminish the significance of national borders in terms of trade and information exchange. International governmental, non-governmental and large corporate multinational organizations create information flows that span many nations. In addition, the rapid expansion of the World Wide Web, possibly the world's largest multilingual document repository, also contributes to this international information exchange. One of the major barriers to this global information exchange is raised by the multiplicity of languages in the flow of information.

Edwards (1994) estimates the existence of approximately 4,500 living languages, of which at least 30 are spoken by 30 million people or more (including non-native speakers).[2] It is clear that in order to function in this multinational, multilingual world, information exchange can no longer be restricted to a single language. Although the English language is spoken by a large number of people and English serves as the *lingua franca* for researchers, the exclusive use of English leaves publications in other languages inaccessible. At the same time, information in English is withheld from those millions who do not speak English.

Traditionally, the English language has been the main focus of information retrieval. Influential researchers were either from Britain (e.g. Cleverdon, Sparck Jones) or from the United States (e.g. Salton). Retrieval algorithms and heuristics originated almost without exception in English-speaking countries and are based on the English language. Over the years, these retrieval methods have been adopted by other language communities, creating a wide selection of language-specific monolingual retrieval systems. However, to ensure complete information exchange, information retrieval systems need to be multilingual or cross-lingual.

[1] Who gets ahead in this world, Has to be able to howl with the dogs, And also has to speak several languages, Who gets ahead in this world.
[2] Among the world's major languages are English (1400 million), Chinese (1000 million), Hindi (700 million), Spanish (280 million), Russian (270 million), and Portuguese (160 million).

## 1.2 Information Retrieval

Information Retrieval (IR) is the process in which users with an information need query a collection of documents to find those documents that satisfy their need.[3] For this to happen automatically in an electronic environment, the user types the query, which is then processed by the system to create an internal query representation. During processing, the system typically removes non-content bearing words or stop words such as articles, determiners, prepositions, and pronouns. The document collection is processed in a similar manner resulting in a list of document representations or an index. The query representation is then matched against the index to find documents that are similar to the query and thus are likely to be relevant to the query. Once the degree of similarity of documents to the query has been established, the documents with the highest similarity scores are presented to the user.

Typically, the similarity between documents and queries is determined by counting the occurrence of query terms in individual documents and the occurrence of these terms in the document collection as a whole.[4] These counts are based on the assumption that 1) the more often a term occurs in a document, the more likely it is to convey the subject of that document, and 2) terms that occur in only a few documents are often more valuable than terms that occur in many documents. Valuable terms have the ability to differentiate between documents and without them it would be quite difficult to make the distinction between relevant and non-relevant documents. Because some documents are long and are more likely to contain multiple occurrences of terms than short documents, document length is also taken into account. The assumption is that a term is a better subject indicator for a short document if it appears multiple times than it would be if it appeared an equal number of times in a long document. Also, terms have a greater chance of co-occurring in longer documents and might cause spurious correlations.

The document term counts (*term frequency*), the collection term counts (used in the *inverse document frequency*), and document length are combined in a so-called term weighting function. This weight is commonly referred to as a *tf·idf* weight because it is a multiplication of the term frequency (divided by the document length) and the inverse document frequency.[5] A weight can be calculated in this manner for each term in a document. The similarity score for a document can be calculated by summing all term weights. Document terms that do not occur in the query get the

---

[3] Appendix I contains a glossary for information retrieval and linguistics terminology.

[4] Once *words* are used in query or document representations they are typically referred to as *terms*.

[5] $w_{ij} = \frac{freq_{ij}}{length_j} * \log\left(\frac{N}{n_i}\right)$ where,

$w_{ij}$ = weight of term i in document j, $freq_{ij}$ = frequency of term i in document j, $length_j$ = length of document j, $N$ = the number of documents in the collection, $n_i$ = the number of documents with term i.

weight zero. Hence documents that do not contain any query terms will automatically get a score of zero. Retrieval systems differ in the formulas used for counting frequencies, and in the algorithms they use to calculate the similarity between documents and queries. Once a score has been calculated for each document, the documents are presented to the user in ranked order, with the most relevant document (the document with the highest score) in the first position. In monolingual IR (MIR) the queries and the documents are all written in the same language.

## 1.3 Cross-Language Information Retrieval

Cross-Language Information Retrieval (CLIR) is a special case of IR. In CLIR, retrieval is not restricted to the query language; rather queries in one language retrieve documents in multiple languages. Cross-language retrieval systems allow users to state their queries in their native language and retrieve documents in all the languages supported by the system. Thus CLIR techniques simplify searching by multilingual users and allow monolingual searchers to judge relevance based on machine translated results and/or to allocate expensive translation resources to the most promising foreign language documents (Oard and Diekema, 1998). For example, in the following figure a native English speaker queries a CLIR system in English to receive a set of results that spans two languages. The CLIR system in this example translates the English language query into Dutch to facilitate searching of the Dutch document database. The English query itself is used to query the English database. Because the user might not read Dutch, the results list from the Dutch database is automatically translated into English after retrieval has taken place. Next, the system combines results from both databases to present the user with a joint list.

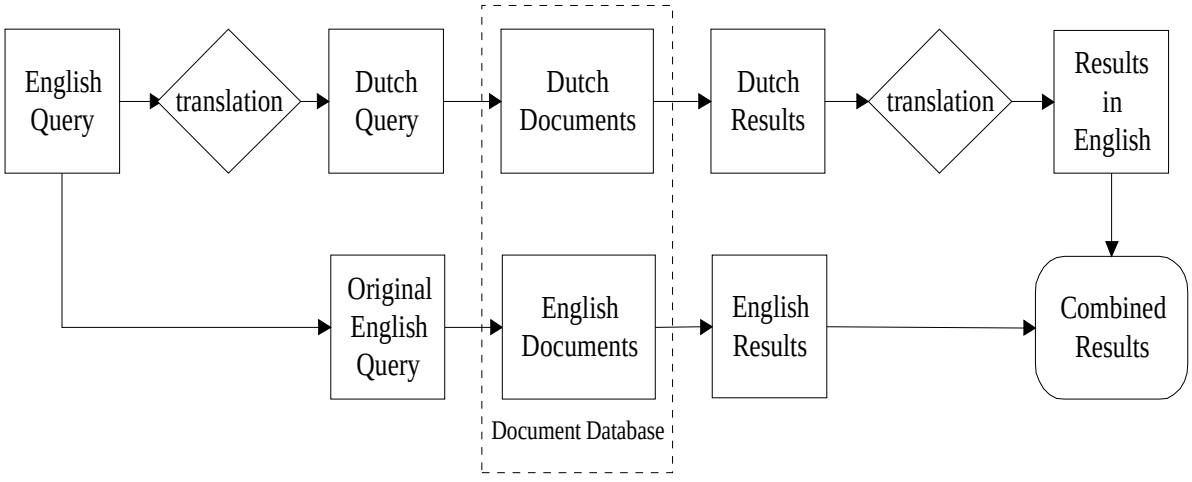


**Figure 1.1 CLIR using query translation.**

The field of CLIR has been in existence for almost as long as monolingual information retrieval although it went through a period of diminished research interest in the 1980s and early 1990s.

Older studies (e.g. Pigur 1964 (translated in 1979), Salton 1970, 1973) used controlled vocabularies (thesauri) to cross the language barrier. A controlled vocabulary specifies the complete set of terms that can be used to index and retrieve documents. By carefully controlling the term set, common retrieval problems, such as synonymy (different words with similar meanings) and homonymy or polysemy (identical words with different meanings), can be avoided.[6] For example, the vocabulary standardizes synonyms and near synonyms to one preferred form of a term (Canine, *see* Dog), and defines homonyms by providing definitions of terms (Turkey - country, Turkey - bird). A translated controlled vocabulary allows a multilingual collection to be indexed and searched in any of the vocabulary's languages while at the same time retaining all of the benefits of a controlled vocabulary. Maintaining and building controlled vocabularies is extremely time consuming and labor intensive, and therefore the majority of modern information retrieval uses free-text searching. In free-text searching none of the search and indexing vocabulary is previously specified or controlled. The documents are typically indexed by stems of the words that occur in the document. The information seeker now has to anticipate different vocabulary use when constructing the query, for example by naming all possible synonyms of a search term. Dealing with false hits caused by homonyms, is especially problematic. Current CLIR research, as well as this study, is typically based on free-text searching. As we will see later, the same vocabulary issues that cause the common retrieval problems in monolingual free-text searching are also problematic when translating the query terms or document terms into other languages for cross-language retrieval.

Conducting retrieval across languages alters the traditional view of information retrieval. In MIR, it is assumed that users are capable of determining the relevance of retrieved documents. Thus, the basic function of a monolingual retrieval system is to present the user with relevant retrieved documents. Since users of cross-language retrieval systems might not be able to read documents in foreign languages, merely presenting retrieved documents is no longer sufficient. Jones et al. (1998) view cross-language retrieval systems as part of a larger *information-access* environment in which documents and/or sections of documents such as document titles are translated into the user's native language using machine translation. Language translation is a unique feature of CLIR that sets it apart from MIR. Translation can appear in at least two places during the cross-language information seeking process, both before and after retrieval (Figure 1.1). This study addressed only the primary translation phase - before retrieval.

[6] Synonymy is problematic for retrieval because matching between synonyms cannot take place without additional processing. For example, a query with the term *canine* will not match relevant documents with the term *dog* even though both terms refer to the same concept. The terms are spelled differently and are thus considered to be different. Similarly, terms that are spelled the same are considered to be identical even though they might not share the same meaning, as is the case with homonymy or polysemy. The

Because queries and documents in CLIR do not always share the same language, translation is needed before matching can take place.[7] The literature explores four different translation options (see section 2.2): translating queries (e.g. Ballesteros and Croft 1996, 1997; Davis and Dunning 1995), translating documents (Oard and Hackett 1998; Kraaij 1997), translating both queries and documents (Dumais et al. 1997; Carbonnell et al. 1997; Diekema et al. 1999), and cognate matching [8] (Buckley et al. 1998). The general trend however, as is the case in this study, is to use query translation (Figure 1.1). As will become clear in the following sections, translation is the most problematic characteristic of CLIR.

Automatic translation of queries or documents requires lexical resources such as machine-readable dictionaries, ontologies, corpora, or machine translation systems.[9] One of the problems facing CLIR is that lexical resources are not widely available, especially for less common languages. Most resources require linguistic processing to create useable translation tools. Even ready-made resources such as machine-readable dictionaries need processing to filter out extraneous information, for example, the definition of the lexical item that is provided in dictionaries for human users but which tends to confuse computer systems (Hull and Grefenstette 1996). Multilingual ontologies with their rich internal knowledge structures are extremely time-consuming to create and are only used by a limited group of CLIR researchers (Gilarranz et al. 1997; Hamp and Feldweg 1997; Diekema et al. 1999). The largest resource for translation data is formed by multilingual corpora, which are used to mine translation data. Previous to the extraction of translation equivalents, alignment (e.g. at the sentence level) of a corpus is required (Carbonnell et al. 1997; Dumais et al. 1997; Nie et al. 1999; Sheridan et al. 1998). The most obvious solution to the translation problem is to apply machine translation (Gey et al. 1999; Oard and Hackett 1998). Machine translation performs reasonably well but is limited to only those language pairs that are available. In addition, machine translation is often not sufficient for short query translation because short queries lack context, which is used by machine translation systems to determine the translation.

---

query term *plant* retrieves documents about factories and documents about shrubs irrespective of the user's interest in gardening.

[7] A more detailed description of CLIR aspects can be found in Chapter 2 (section 2.2.1); the following paragraphs serve as a brief introduction to this topic.

[8] Cognate matching extends matching cognates (words that have identical spelling) across languages by allowing for minor spelling differences between the cognates.

[9] An ontology consists of terms, their definitions, and relationships between the terms; a machine-readable dictionary is a dictionary in machine (computer) readable form; a machine translation system automatically translates machine-readable text; and a corpus is a body of (machine-readable) text.

It is often thought that by acquiring the right translation resources the translation problem and thus CLIR is solved. Although a relationship between dictionary quality and retrieval performance has been observed (Kraaij 2001, McNamee and Mayfield 2002), natural language is too complex for this to be true. Even in the monolingual case, lexical ambiguity[10] and synonymy cause serious problems. When an additional language is added to the mix, problems grow worse. In the ideal situation, a word or phrase has only one sense and hence only one translation. However, many words have multiple senses, each of which may have single or multiple translations. Although phrases exhibit less translation ambiguity than single words, non-compositional phrases are problematic since they cannot be reconstructed from translations of the constituent terms.[11] In the worst situation, a word or phrase does not have a foreign language equivalent.[12] The CLIR literature recognizes three factors causing translation error in dictionary-based translations: lack of translations for technical terms and acronyms; the erroneous breaking up of non-compositional phrases in translation, and; the addition of multiple translation senses of a word to the translation (Ballesteros and Croft 1996, Hull and Grefenstette 1996). Similar and additional problems are described in the translation literature (see section 2.3.2).

CLIR performance is commonly expressed as a percentage of monolingual effectiveness. Reported values typically range from around 50% for unconstrained dictionary-based query translation to 98% or so for more sophisticated techniques.[13] On some test collections CLIR systems can even be more effective than monolingual systems due to careful exploitation of translation resources and combinations of query expansion techniques. The generally reduced performance of CLIR is thought to be caused by translation, which adds additional noise. A large part of the research is therefore devoted to finding reliable methodologies to reduce translation ambiguity. Solutions so far have included using part-of-speech taggers to restrict the translation options (Davis 1996), applying query expansion techniques (McNamee and Mayfield, 2002), using corpora for term translation disambiguation (Ballesteros and Croft, 1998), and using weighted Boolean models

[10] Lexical ambiguity occurs when words have multiple meanings (e.g. the word table could refer to a figure in a document or to furniture).
[11] The meaning of a non-compositional phrase is different than the sum of the meaning of its parts (e.g. “real estate”, “face value”). The literature suggests that phrase recognition strategies can substantially improve cross-language retrieval effectiveness (Ballesteros and Croft, 1997).
[12] This typically occurs with idiomatic expressions, specialized vocabulary, or culturally relative expressions (e.g. there simply is no English equivalent for the Frysian term *klûnen* = to walk on skates over boards or carpeted stretches of road around obstacles such as low bridges or patches of weak ice encountered while skating).
[13] In order to obtain the comparative performance figure, experiments typically use two sets of test queries (one in the same language as the documents and one in a foreign language). Running the same language queries constitutes a monolingual run and running the foreign language queries constitutes the cross-language run. Both sets are run using the same information retrieval system, but the language barrier is only crossed for the foreign language queries.

which tend to have a self-disambiguating quality (Hull 1997; Diekema et al. 1999; Hiemstra and Kraaij 1999).

## 1.4 Statement of the problem

As pointed out previously, on average, CLIR systems perform below their monolingual counterparts. Although both MIR and CLIR have to deal with lexical ambiguity as found in natural language, CLIR faces more extensive problems than does MIR. Researchers assume that the difference in performance can be explained by increased ambiguity and possible translation error created when crossing the language barrier (e.g. by using machine translation).

Although numerous studies in CLIR describe means to reduce the translation error and lexical ambiguity that hinders translation, no research so far has successfully identified the nature of the problems introduced by the translation other than by a comparison of a system's monolingual performance to its cross-language performance, expressed as a percentage of monolingual performance (see section 1.3). What is missing from the literature is a careful examination of the nature of translation events, such as lexical ambiguity and translation error, and the relation of these events to cross-language retrieval performance.

This study outlines the translation events facing the CLIR query translation process and shows how these translation events impact retrieval performance to provide a better understanding of the theoretical and practical implications of translation in CLIR.[14]

## 1.5 Research questions

To understand the problems facing cross-language retrieval and the relation of these problems to retrieval performance, the study explored the following research questions:

1) *What kinds of translation events affect cross-language retrieval?*
2) *In what way does the presence of certain translation events in query translation affect retrieval performance?*

### *1.5.1 What kinds of translation events affect cross-language retrieval?*

Until now, the literature has lacked a comprehensive taxonomy describing translation events that impact CLIR. Although translation successes are as important as translation problems when looking at query performance, the focus in the literature has been mainly on translation problems. Researchers, such as Ballesteros and Croft (1996) and Ruiz and Srinivasan (1998) do discuss sources of translation error, but the translation problems are not examined in detail. The translation errors in the literature often only form a subset of all possible translation errors and might be restricted to the linguistic resource used in the study. The main reason for the shallow analysis

might be that most CLIR research is empirically based and TREC[15] driven. Researchers analyze the performance of their systems by running the cross-language track TREC topics and, amongst other aspects of performance, summarily describe the sources of translation error encountered in the set of TREC queries.[16] Considering the wide diversity of natural language, it is highly unlikely that the translation problems encountered in the limited number of TREC cross-language queries[17] cover a wide range of possible translation events.

To thoroughly understand what kinds of translation events affect CLIR it is important to study a wide variety of queries until no new types of translation events are identified. In answering the first research question the study examined a large set of queries and their automatic translations. The criteria used to select taxonomy categories was guided by the translation literature and the CLIR literature. By comparing the original queries with their translations certain translation events became apparent (see Figure 1.2). Thus, the queries and their translations formed empirical entities from which a translation event taxonomy was created. The study examined query translations until a point of saturation occurred and no new translation event categories were identified. The study aimed to make the taxonomy as comprehensive as possible. Although only two languages were used in this study (see below), insights from the translation and CLIR literature were incorporated to ensure the taxonomy's applicability outside these two language areas.

The study restricted the languages under study to English and Dutch. To ensure a high level of understanding of issues encountered when crossing the language barrier, the study examined those two languages in which the bilingual researcher is skilled. The practice of using bilingual researchers is quite common in machine translation evaluation when correctness testing or accuracy scoring is involved (Arnold et al., 1994; Nyberg, Mitamura, and Carbonell, 1994). Translations from Dutch into English were compared to the original English queries as well as to their original version (see Figure 1.2). The first comparison showed whether the automatic English translation maintained the meaning of the Dutch original. A loss of meaning indicated a possible problematic translation event. To verify the translation event the automatic English translation entered a second comparison where it was compared to the English source query. A combination of the two comparisons formed the basis of the translation event taxonomy. The taxonomy was then extended with translation events gleaned from the literature.

---

[14] This study does not seek solutions to the translation problems facing query translation.
[15] The Text REtrieval Conference is a yearly event in which information retrieval systems run identical retrieval experiments for a grand scale comparative evaluation.
[16] For example, among the TREC cross-language track participants, the French non-compositional phrase *urse en peluche* (teddy bear) is well known to cause retrieval problems being translated as *fluffy bear*.
[17] 22 queries for TREC-6, 28 queries for TREC-7, and 28 queries for TREC-8.

### *1.5.2 In what way does the presence of certain translation events in query translation affect retrieval performance?*

After establishing a taxonomy of translation events, the second research question examined a possible relationship between translation problems and CLIR performance. It is important to realize, however, that all the problems facing MIR can also occur in CLIR, irrespective of whether a translation took place.

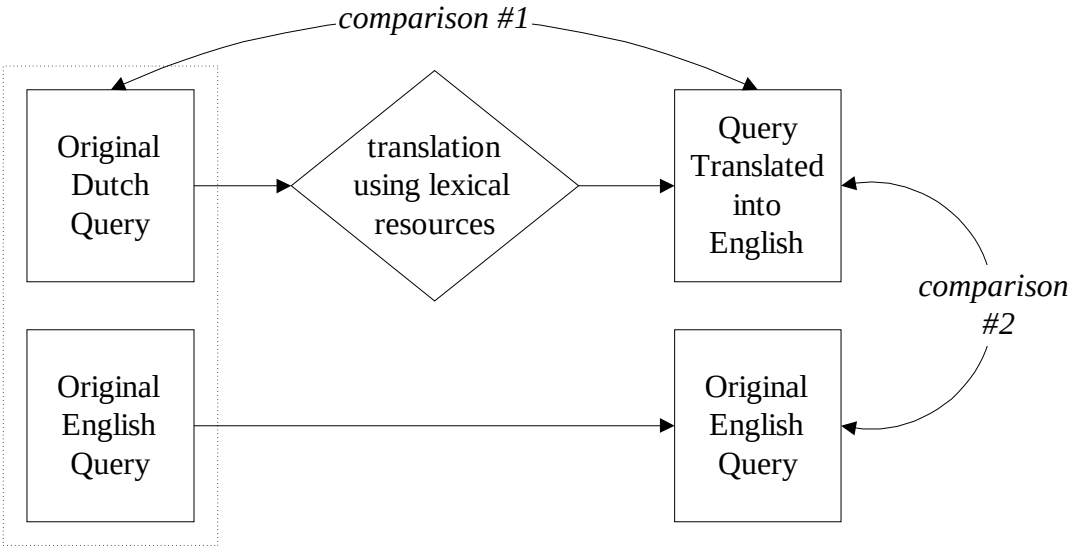


**Figure 1.2 Query translation comparison.**

Once the source language (original language) query has been translated into the target language query, retrieval is in effect taken back to a monolingual scenario. For example, after translation of the source language query we are left with a set of words that now form the target language query. This set of words faces the same problems that the source language terms would in a monolingual situation such as homonymy or polysemy (where the retrieval system doesn't distinguish between identical words even though they have different meanings), and synonymy (where similar documents are indexed by different words having similar meanings, leaving it up to the user to specify all possible synonyms for a term in the query). Thus, the problems facing the target language query at the point of retrieval can be divided into two categories: 1) problems unique to the cross-language situation caused by the translation step (e.g. mistranslation of phrases), 2) problems shared with MIR (e.g. synonymy). This study focused on the first category of problems although dealing with the second category of problems cannot be avoided because they are inherent to information retrieval itself.

To study the impact of the translation events on CLIR performance, this study carried out a set of information retrieval experiments with the queries used in the query translation analysis for the first research question.

In order to establish the impact of translation events on retrieval performance, it was essential to control for other factors possibly influencing retrieval performance such as the query, and the information retrieval system. Although neither Ruiz and Srinivasan (1998) nor McCarley (1999) found a relationship between statistical query features such as query length and CLIR performance, other literature suggests that these features might be important to information retrieval in general (Salton and McGill, 1984). Judging by the remarkably close performances of the different systems participating in TREC-7 (Voorhees and Harman, 1999), the actual information retrieval system used in the experiments is probably less of an issue as a possible source of variation.

## 1.6 Scope of the study

The study examined translation events as encountered in CLIR during query translation only. The study attempted to find a possible relationship between these translation events and CLIR performance. The query translations were created automatically from Dutch into English. An electronic free-text cross-language information retrieval system was used for the retrieval experiments with full-text electronic documents being the object of retrieval. The indexing of these (English) documents was derived and fully automatic. The retrieval algorithm used was OKAPI BM25.[18] Although translation occurs on two occasions in CLIR (see section 1.3), this study only focused on the translation step most closely related to matching. Linguistic processing issues, such as language recognition and character set conversions, were outside the scope of this study.

## 1.7 Limitations of the study

The study findings are dependent on the following: languages under study, lexical translation resources, cross-language retrieval system, test queries, and test documents.

As pointed out in the introduction, there are several thousand languages in this world and this study only covered two of them. The two languages, English and Dutch, not only both belong to the world's largest language group of Indo-European languages, but also share the same language sub-group of Germanic languages (Katzner, 1995). It could be argued that such a narrow selection limits the findings of this study. Naturally, there are likely to be translation events specific to the unidirectional Dutch-English language pair. Although using the translation literature and CLIR literature enabled the researcher to find additional translation problems that transcend this single

language pair, generalization to other language pairs might still be problematic. A possible source of bias that might have been introduced is enhanced retrieval performance due to the possibly larger number of cognates shared between the languages. However, the data showed that, in general, Dutch queries do not retrieve English documents very well before query translation. It is conceivable that the translation resource used, influenced the degree to which certain translation events were found. For example, a translation resource with only a few thousand entries is likely to result in more missing translation cases than might otherwise be the case. However, it is unlikely that a completely different set of translation events might be found with a different resource. As for the information retrieval system, after years of development, different types of retrieval systems have come quite close in their performance levels. Therefore, the choice of retrieval system is unlikely to be a major limitation to generalizability. However, inherent to information retrieval experimentation is the dependence on test collections. Although the TREC test collections are an answer to the traditional limitations connected to test collections, such as their small size (Sparck Jones, 1995), retrieval results do not necessarily carry over from one test collection to the next.

## 1.8 Significance of the study

The field of CLIR would benefit from a study of translation events because, despite being central to CLIR, they have not been investigated in great detail. The study broadens previous approaches to this problem by creating a taxonomy based on literature reviews in combination with a query analysis of a substantially larger number of queries than the 28-50 queries commonly found in CLIR studies. The study: 1) created a general taxonomy of translation events in CLIR, and 2) found a small correlation between some types of translation events and CLIR performance. Understanding in these areas increases knowledge in the field of CLIR about translation. This understanding aids in dealing with translation problems and benefits cross-language system design, which ultimately benefits global cross-lingual exchange of information. This study also informs monolingual retrieval and increases understanding of the relationship between the fields of translation and information retrieval.

## 1.9 Summary

There is increasing globalization of our information-based society. Only cross-language information retrieval systems, where users state their queries in their native language to retrieve documents in all the languages supported by the system, can support complete information exchange. Because queries and documents in CLIR do not always share the same language, translation is needed before matching can take place. This translation step causes a reduction in retrieval performance of CLIR as compared to MIR. Although numerous studies in CLIR describe

[18] The vector space and probabilistic algorithms are most commonly used and mathematically very close, especially in their basic form (without relevance feedback and other retrieval improvements).

means to reduce translation error and lexical ambiguity, no research has successfully identified the range of translation events and the extent of their impact on retrieval performance.

To study the translation events in cross-language retrieval and the relation of these events to retrieval performance, this study explored the following research questions:

1) *What kinds of translation events affect cross-language retrieval?*
2) *In what way does the presence of certain translation events in query translation affect retrieval performance?*

It answered these research questions by studying query translation from Dutch into English, and running retrieval experiments. This research resulted in a taxonomy of translation events, and a description of the impact of a subset of these events on CLIR performance.

# 2 Background

## 2.1 Introduction

This study examined translation events faced by CLIR query translation, and the impact of translation on retrieval performance. This chapter presents the context for the study. The chapter opens with an examination of matching in CLIR and the translation knowledge required to cross the language barrier. The next section examines translation, its problems as presented in translation literature and CLIR literature, as well as aspects of the two languages that form the focus of this study: English and Dutch. The chapter closes with a discussion on IR evaluation, CLIR evaluation, and methodological issues related to system evaluation in general and, more specifically, the use of test collections.

## 2.2 Matching and translation in Cross-Language Information Retrieval

As pointed out in section 1.3, Cross-Language Information Retrieval (CLIR) is a special case of information retrieval in which retrieval is not restricted to the query language itself but extends to all languages supported by the system (see Figure 1.1). In CLIR, matching takes place across languages, and queries in one language retrieve documents in another. What complicates CLIR is the fact that matching is taking place between quite distinct vocabularies. To facilitate matching across different languages, nearly all CLIR approaches apply some form of translation using various lexical resources. The following two sections discuss the different matching approaches and the translation knowledge needed in order for matching to take place.[19]

### *2.2.1 Matching approaches in CLIR*

Four general approaches to cross-language matching have emerged in CLIR: cognate matching, document translation, interlingual techniques, and query translation, the latter being the focus of this study.

Cognate matching essentially automates the process by which readers might try to guess the meaning of an unfamiliar term based on similarities in spelling or pronunciation. A simple version of cognate matching in which untranslatable terms are retained unchanged is often used in CLIR systems to match proper nouns and technical terminology (e.g. Ballesteros and Croft, 1997). Davis (1997) extended this technique using fuzzy matching to discover Spanish cognates for English words that did not appear in a bilingual dictionary. Buckley et al. (1998) applied a more sophisticated approach, creating equivalence classes for letter sequences with similar sounds (e.g., “c,” “k,” and “qu” share an equivalence class). Since the translation knowledge is embedded

[19] Section 2.2 is largely based on two sections from Oard and Diekema’s (1998) general introduction to CLIR.

directly in the matching scheme, cognate matching can be used in isolation. Most often, however, cognate matching is combined with other cross-language matching approaches.

Document translation is the opposite of query translation. In document translation, documents (or their representations) are automatically converted into each supported query language. Documents typically provide more context than do queries, so more effective strategies to limit the effect of translation ambiguity may be possible. Another potential advantage is that selected documents can be presented to the user for examination after retrieval without on-demand translation (Kraaij, 1997). On the other hand, the massive translation that is required can be an expensive undertaking, and the costs are even greater if several query languages must be supported. Erbach et al. (1997) suggest using document translation only for small collections in limited domains.

Interlingual techniques convert both the query and the documents into a unified language-independent representation (Diekema et al., 1999). Controlled vocabulary techniques based on multilingual thesauri are the most common examples of this approach. Concepts in the controlled vocabulary are represented by terms in multiple languages so that any of these languages may be used to index documents or to form queries. Controlled vocabularies can be time and labor intensive to create and are not flexible when it comes to adding new languages to the system. Some fully automated interlingual techniques have also been implemented. Latent semantic indexing (Landauer and Littman, 1990, 1991; Dumais, et al. 1997) and the generalized vector space model (Carbonnell et al., 1997) both use a document aligned training corpus to learn a mapping from one or more languages into a language-neutral representation. Document and query representations from either language can be mapped into this space, allowing similarity measures to be computed both within and across languages. Automatic interlingual techniques are highly computationally intensive and tend to work best on smaller document collections.

Query translation is a matching strategy in which the query (or some internal representation of the query) is automatically converted into every supported language. Query translation is relatively efficient and can be performed as needed. The principal limitation of query translation is that queries are often short and short queries provide little context for disambiguation. Homonymous words (those with more than one distinct meaning) produce undesirable matches even in monolingual retrieval (Krovetz and Croft, 1992). Translation ambiguity compounds this problem, potentially introducing additional terms that are themselves homonymous. For this reason, controlling translation ambiguity is a central issue in the design of effective query translation techniques.

Although query translation does face additional translation ambiguity it is the most commonly used matching approach in CLIR (Ballesteros and Croft, 1998). The main reason for this is that a query translation module can simply be added to an existing IR system to create a CLIR system. In addition, as pointed out earlier, query translation is efficient, can easily be extended to new languages, can be used for searching large document collections, and can be done while the user interacts with the system. Because query translation is the focus of most CLIR research and system implementations, it was chosen as the CLIR matching approach for this study.

### *2.2.2 Translation knowledge for query translation*

To study what kinds of translation events affect CLIR, this study examined a large number of CLIR query translations. CLIR query translation depends on some form of translation knowledge. That knowledge may be encoded manually by human lexicographers or extracted automatically from corpora. The literature typically refers to techniques using translation knowledge from manually encoded translation sources as knowledge-based approaches. Techniques using translation knowledge from corpora are referred to as corpus-based techniques. The four main sources of translation knowledge that have been applied to CLIR are ontologies, bilingual dictionaries, machine translation lexicons, and corpora. This section considers each of the four sources of translation knowledge in turn (see Figure 2.1).

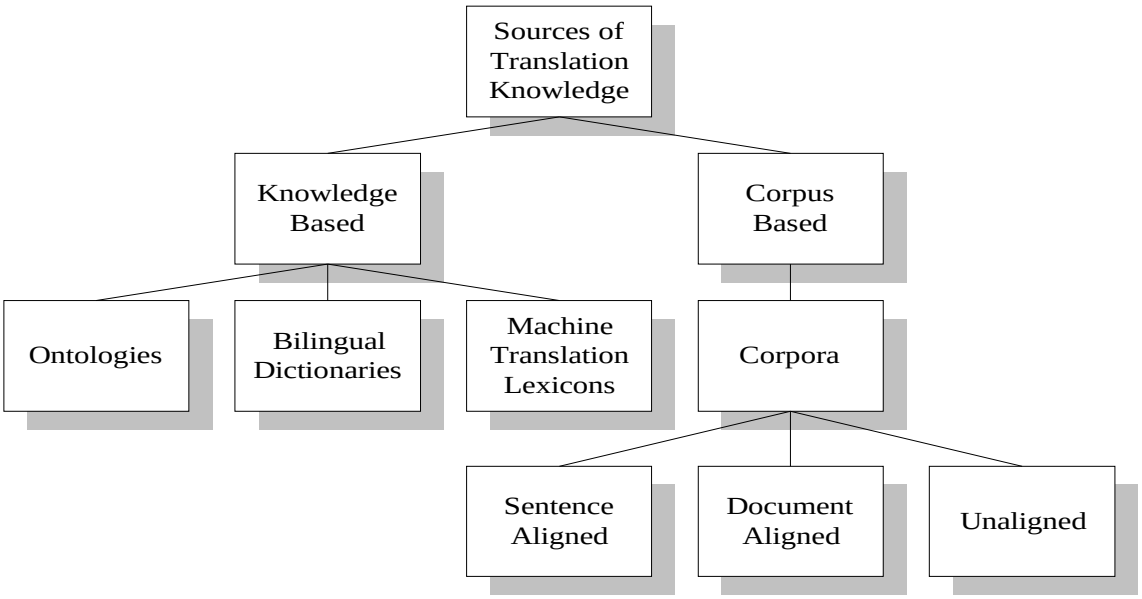


**Figure 2.1 Sources of translation knowledge.**

Ontologies[20] are structures that encode world or domain knowledge by specifying relationships between concepts. Thesauri are ontologies that are designed specifically to support information retrieval. At present multilingual thesauri are the dominant sources of translation knowledge in operational CLIR systems. Thesauri can support both controlled vocabulary and free-text retrieval, providing insight into hierarchical relationships (broader terms, narrower terms), synonymy, and more general associations (related terms). Such relationships can help experienced users better define queries by enhancing their understanding of the structure of knowledge for the topic being searched. The European Parliament's multilingual EUROVOC thesaurus is one example of a multilingual thesaurus. A common approach to create a multilingual thesaurus is to translate an existing monolingual thesaurus. General-purpose ontologies such as WordNet (Miller, 1990) are emerging as alternatives to traditional thesauri because their broader coverage permits use of sophisticated knowledge structures in broader domains than has heretofore been possible. By encoding additional relationships such as meronymy-holonymy (part-whole), WordNet explicitly captures a broader range of semantic knowledge than traditional thesauri. The EuroWordNet project is developing a multilingual ontology resembling WordNet with components in Dutch, English, Italian and Spanish that are linked by an "interlingual index." Gilarranz et al. (1997) have described how EuroWordNet might be used to support a query translation strategy.

Machine-readable bilingual dictionaries have been widely used to support query translation strategies (Ballesteros and Croft 1996, 1997, 1998; Gey and Chen 1998; Davis 1997; Fluhr et al. 1997; Hull and Grefenstette 1996; Kraaij and Hiemstra 1998; Kwok 1997; Nguyen et al. 1997; Yamabana et al. 1998) Bilingual dictionaries are typically designed for human use, so translations of individual terms are often augmented with examples showing how those terms could be used in context. It would be difficult to extract generalizations from those examples that could be used automatically, so machine-readable dictionaries are typically processed manually or automatically to reduce them to a bilingual term list, perhaps with additional information such as part-of-speech. In essence, dictionary-based translation consists of looking up each query term in the resulting bilingual term list and selecting the appropriate translation equivalents. The simplest way of using such a bilingual term list is to select every known translation for each term. That approach is often used as a baseline in dictionary-based CLIR evaluations. Both Radwan and Fluhr (1995) and Davis (1997) have shown that limiting the translations to those with the same part-of-speech (e.g., noun or verb) can improve retrieval effectiveness, and Kraaij and Hiemstra (1998) experimented with the use of preferred translations that were noted in their dictionary. Hull (1997) explored the ability of structured queries to further limit translation ambiguity, implementing a weighted Boolean matching strategy that exploited the observation that correct translations are more likely to co-occur than incorrect translations (see section 2.3.3.1 for more detail on sense disambiguation).

[20] As opposed to a branch of metaphysics.

Machine translation systems are becoming fairly widely available, although machine-readable dictionaries still cover a greater number of language pairs (Kraaij, 1997). Machine translation systems encode translation knowledge in a lexicon that contains the information needed for automatic analysis, translation and generation of natural language. One goal of natural language analysis is to disambiguate terms in ways that can limit translation ambiguity, and the lexicon is often designed to provide information that is useful for this purpose. The most straightforward way to apply a machine translation lexicon to CLIR is to simply use the machine translation system to translate the queries. However, queries are rarely provided as well formed sentences, so the effectiveness of this approach may be limited in query translation applications (Hull and Grefenstette, 1996; Kraaij, 1997). Machine translation systems necessarily choose a single preferred translation for each term, and Erbach et al. (1997) have observed that such a singular choice might adversely affect retrieval effectiveness. Examples of the use of machine translation for query and document translation can be found in Oard and Hackett (1998).

Corpora form an important source of translation knowledge for query translation in CLIR. There are three types of multilingual corpora that have been used in the literature: parallel corpora, comparable corpora and combined monolingual corpora. The corpora differ in their level of alignment: document, sentence and word, or no alignment. Parallel corpora are made up of translation equivalent sets, each containing a document and one or more translations. Comparable collections, on the other hand, are typically separately authored but related by topical content. Aligned document sets in comparable corpora may contain one or more documents in each language (Peters and Picchi 1997; Sheridan and Ballerini 1996). Combined monolingual corpora are made up of two or more monolingual corpora in different languages and are not related at all. Parallel corpora, since they contain exact translation equivalents, can be aligned at the sentence and term level whereas comparable corpora can only be aligned on the document level. Combined monolingual corpora cannot be aligned.

Document-aligned corpora are document collections in which useful relationships between sets of documents in different languages are known. The basic strategy for using document-aligned corpora is to represent each term using the pattern of aligned sets in which that term occurs and then to construct language-neutral representations of documents in any supported language using the resulting term representations. Techniques from linear algebra are typically used to compute and manipulate these term representations. When the language of each document is known, each term is typically tagged with a language marker in order to avoid undesired conflation of different concepts in other languages. Sheridan and Ballerini (1996) and Mateev et al. (1997) built a bilingual term list for query translation using term representations computed from a comparable

corpus of news stories that was aligned using classification codes, publication dates and cognates. They found the terms in each language that were most similar to each query term (using a vector similarity measure) and then used several of the most similar terms as the translated query.

Sentence and term aligned corpora are created by aligning individual sentences in parallel corpora using certain programming techniques. Davis (1997) used a sentence-aligned parallel corpus directly to augment dictionary based query translation without substantial improvement over a simpler dictionary-based technique. Research has shown that document and sentence aligned techniques may be most useful when the needed alignments are known within some portion of the same collection from which retrieval is desired (Carbonell et al., 1997). Although such a situation may exist in a few applications (e.g., if translations are being made routinely, but they are not available immediately), this factor is likely to somewhat circumscribe the utility of techniques based on document and sentence aligned corpora.

Unaligned corpora can be used in conjunction with a bilingual term list as an additional source of translation knowledge even if *a priori* document alignments are not known. Ballesteros and Croft (1997) applied fully automatic passage-level pseudo-relevance feedback using the query language portion of their unaligned corpus to refine the query representation. By augmenting the original query with terms appearing in top-ranked passages, monolingual pseudo-relevance feedback often improved recall without a significant adverse effect on precision. They then applied dictionary based query translation to produce a version of the query in the desired language, followed by fully automatic passage-level pseudo-relevance feedback using the portion of the unaligned corpus containing documents in that language. When applied individually, each pseudo-relevance feedback step improved CLIR effectiveness, and the combination outperformed either step alone.

## 2.3 Translation and its difficulties

Translation is the cause of the reduced performance of CLIR as compared to MIR and was the main focus of the study. By studying the translation events in CLIR query translation and their impact on retrieval performance, the study sought to establish a better understanding of the implications of translation in CLIR. The following section discusses translation and what makes it difficult. In addition, specific translation problems from the translation literature and the CLIR literature will be discussed. The section closes with a brief discussion of the two languages that are the object of translation in this study: English and Dutch.

### *2.3.1 Translation*

Translation is the process of transferring information from one language into an equivalent version in another language (Nida, 1964). What constitutes an equivalent version is a prominent topic in the rather prescriptive-oriented translation literature. Larson (1984) describes a translation

continuum ranging from very literal translation to unduly free translation (see Figure 2.2). Whereas unduly free translations change the facts of the original source text, literal translations replace source text words with literal equivalents. Literal translations appear unnatural because they lack the correct syntax and word choice of the target language. Idiomatic translations are desired as translations since they have both correct syntax and word choice, and are equivalent in content to the source text.

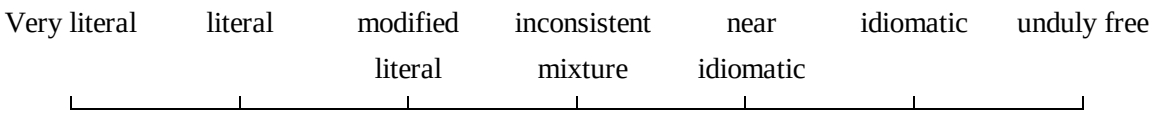


**Figure 2.2 Types of translation in a translation continuum.**

Creating an idiomatic translation is much more than a mere word replacement exercise to cross the language barrier. Besides an obvious linguistic component, translation also has anthropological, communication, cultural, historical, and psychological aspects (Schulte 1987; Hartmann, 1989). As we will see in section 2.3.2, some of these aspects might result in translation problems which, in turn, reduce cross-language retrieval performance.

The literature provides several reasons why translation is complex. The underlying reason for most of the translation problems is a natural language phenomenon called skewing. Skewing means that there is no one-to-one correspondence between a lexical item and its meaning. Due to the lack of this unique correspondence, a verbatim translation is not possible (Nida, 1958). Even within a single language it is difficult to find exact synonyms. Because lexicons (the collection of words used by a particular language community) across languages tend not to match, the translator will need to find solutions to the lack of equivalency. Still, it is not necessary to give up on translation because of these inherent limitations. Even everyday communications between native speakers involves interpretation, coding, and decoding of meaning and in a multilingual world translation is very much needed (Bell, 1991, Edwards, 1994, Steiner, 1975).

That skewing does not make translation entirely unworkable is pointed out by Snell-Hornby (1990) who introduces varying levels of interlingual relationships and recognizes 5 basic prototypes: 1) terminology or nomenclature (*zuurstof* = *oxygen*); 2) internationally known items and sets (*zaterdag* = *Saturday*); 3) concrete objects, basic activities, stative adjectives (*stoel* = *chair*); 4) words expressing perception and evaluation (*ongezellig* ≈ *not cosy*); 5) culture-bound elements (*hutspot* = *?*). Lexical equivalence between languages is typically found to be restricted to the first two prototypes and the majority of prototype 3. Lexical prototypes 4, and 5, contain the terms that make translation a formidable task.

A closer look at these two problematic prototypes suggests that there is a strong cultural aspect. Cultural differences are presented in the literature as another translation difficulty (Pause, 1997). For a translation to be completely equivalent to its source text, its target language readers should react to that text the same way as the source language readers would (Nida and Taber, 1969). Since the target language readers are likely to have a different cultural and historical background compared to the source language readers, complete equivalence as defined by Nida and Taber is next to impossible. Cultural differences between language groups are bound to create noise in a translation. Strangely enough, overlap between the lexicons can be the cause of misunderstanding. For example, unlike Italian, English and Dutch distinguish between naturally occurring *channels* and man-made *canals*. The English translation of the observations by the Italian astronomer Schiaparelli (1877) of *canali* (meaning *channels* but translated as *canals*) on Mars might be the origin of our fascination with the possibility of Martian civilization (Bell, 1991).

The lexical equivalency problems often originating in cultural differences between language communities make translation difficult. Pause (1997) refers to the conflict between on the one hand preserving the meaning of the source text while at the same time creating a natural target translation. It is tempting to create a literal translation that target language readers will be able to understand. However, literal translations sound clumsy and foreign. The meaning is not just in the words themselves but in the relationships between them. Selection of words depends on the context and the semantic structure. To a machine, the meaning of a sentence is often hidden behind its surface structure (Larson, 1984).

### *2.3.2 Specific problems in translation*

Prahl and Petzolt (1997) make a distinction between *Übersetzungsprobleme* (translation problems) and *Übersetzungsschwierigkeiten* (translation difficulties). Translation problems are known problems, such as a source term having multiple target translations. Translation problems are translator independent unlike translation difficulties that are issues facing the individual translator during the translation process. Gophinathan (1993) calls translation problems translinguistic problems and divides the problem spectrum into problems of meaning and problems of style. Problems of meaning result from words having 1) suggestive meaning as well as literal meaning; 2) socio-cultural meaning such as culturally specific lexical items, idioms, and folk images; and 3) false cognates. Problems of style result from mismatches between the structure of phonology, morphology, lexical level and syntax of the source and target languages. Problems of style mostly concern problems such as preserving sound effects (e.g. alliterations) or verse of the source text into the target text. The focus of the study was on translation problems rather than difficulties. It concentrated on problems of meaning rather than syntax since the creation of a naturally flowing translation that preserves verb tense and verse structure does not impact information retrieval

performance.[21] This following section will focus on specific translation problems from the translation literature: lexical ambiguity, lexical mismatches, lexical holes, figures of speech, multiword lexemes, specialized terminology, and false cognates (see Figure 2.3 below).

| **Translation problems from the translation literature** | |
|---|---|
| *lexical ambiguity* | words having multiple meanings |
| *lexical mismatches* | differing conceptual structures between language communities |
| *lexical holes* | unlexicalized concepts across languages |
| *figures of speech* | words that should not be taken literally or words that are used to create a certain literary effect |
| *multiword lexemes* | idioms, phrasal verbs, and collocations |
| *specialized terminology and proper names* | words used by certain discourse communities and names of people, places and organizations often do not appear in dictionaries |
| *false cognates* | words that seem to be the same across languages but are not |

**Figure 2.3 Specific translation problems from the translation literature.**

#### *2.3.2.1 Lexical ambiguity*

Lexical ambiguity occurs when words have multiple meanings. This lexical characteristic is also referred to as homonymy or polysemy.[22] What causes problems in translation are words with identical *written* form but different meanings, also called homographs. Homography can take place within the same part-of-speech (e.g. the noun *table* can refer to furniture of a figure in a document), or across categories (e.g. the noun and the verb *bark*) (Lehrberger and Bourbeau, 1988). For human beings, lexical ambiguity is usually easily resolved through the use of context. In the sentence "*she heard a loud bark from behind the fence*" it is clear *bark* means the sound of a dog and not the outside of a tree. We know this is the case because neither trees nor their skin can typically be heard. For translators who are non-native speakers, disambiguation might be more complex than for native speakers, especially when the sense distinctions are more complex and the translator is not familiar with the context clues needed to make the sense distinction. The sense of a word that comes to mind first when confronted with an ambiguous word often has a lexical equivalent in a foreign language that matches fairly closely. The secondary senses however, do not share this equivalency (Larson, 1984).[23]

[21] Naturally, meaning and syntax are intricately related and cannot be easily separated. Only syntactical issues that are directly related to meaning such as word order in phrases will be considered.
[22] Both phenomena pose identical problems to IR (see glossary for further detail).
[23] For example, the verb *to run* can be translated by *rennen*. The sentence "*the boy runs*" translates nicely into "*de jongen rent*". Using a different sense of *running* however does not share the same equivalency. The sentence "*the motor runs*" can not be translated into "**de motor rent*" but should be translated as "*de motor loopt*" which literally means **the motor walks*. An asterisk here indicates an nonsensical sentence.

The lack of a one-to-one mapping can take the following forms: monosemous words having more than one translation (1- *n*), polysemous words having only one translation (*n* –1), or polysemous words having multiple translations (*n* – *n*). In all cases translation choices will need to be made with the possibility of introducing noise to the translation.

#### *2.3.2.2 Lexical mismatches*

Lexical mismatches result from the fact that different language communities view the world in different ways. The worldview affects which concepts are lexicalized (see section 2.3.2.3), how concepts are related, and the lexical level of specificity of concepts (Arnold et al., 1994). For example, in Dutch there is a distinction, unknown in English, between the concept animal head (*kop*) and human head (*hoofd*). Different languages also have different orientations (e.g. hunting and fishing versus high technology) that are evident in their lexicon. Thus the Dutch language might have more terms related to hydraulic engineering than the Hopi language does. The problem caused by lexical mismatches is that there is no one-to-one mapping across languages - a problem similar to the problem caused by lexical ambiguity.

#### *2.3.2.3 Lexical holes*

A more extreme problem is formed by lexical holes where a lexical item does not have a lexical equivalent in the other language (Arnold et al., 1994). Lexical holes are often caused by culture-bound terms such as terms related to economics, food, politics, religion, and sports (Bugarski, 1985). Lexical holes can only be translated as phrases in the target language unless the translator can use a loan word, create a new term (e.g. the Dutch *klapschaats* becomes *clap skate* in English)[24], or find a cultural substitute (Beekman and Callow 1974; Saracevic 1989).

#### *2.3.2.4 Figures of speech*

Figures of speech do not translate literally and might cause erroneous translations if they are not recognized as such. The Dutch idiomatic expression *oude koeien uit de sloot halen* translates to *dragging old cows out of the canal*.[25] Dictionaries typically do not contain figures of speech, which further complicates their translation. Two well-known figures of speech are the euphemism, in which a term is used in place of the actual term to avoid saying something unpleasant or shocking, and hyperbole, which is an exaggeration. A special case of a figure of speech is found in words with connotative meanings. The word *dog* for example often has a positive meaning in the Western world as man's best friend. This positive connotation might not be shared by people from Islamic countries in which dogs can be regarded as pariahs. Translators have to be aware of connotative meanings because they are difficult to deal with in translation.

---

[24] *Washington Post* article: "From Bones to Clap Skates", Feb. 17, 1998, p. C4.

[25] The expression means *bringing up old histories* and is used to express the futility of dragging up the past.

#### *2.3.2.5 Multiword lexemes*

Idioms, phrasal verbs, and collocations are complicated in translation because of their non-standard compositionality (their meaning is not merely the sum of its parts, e.g. *real estate*), non-standard substitutability (the parts cannot be replaced by synonyms of these parts without invalidating the expression, e.g. *real = genuine *genuine estate*)[26], and non-standard morpho-syntactic properties (construction of terms follow idiosyncratic rules, e.g. plural = +s, **real estates*) (Storrer and Schwall, 1993). The meaning of multiword lexemes (MWLs) tends to get lost in translation when translating word-by-word (Arnold et al. 1994; Beekman and Callow 1974). As with figures of speech, MWLs create problems with failure to distinguish between a literal or idiomatic reading. Although some MWLs survive a literal translation, others result in utter nonsense. Phrasal verbs such as *to wake up* or *to lie down* often take different propositions or none at all in the translation. Not all MWLs have target translations.

#### *2.3.2.6 Specialized terminology and proper nouns*

Specialized terminology, such as scientific names, is often difficult to translate and is often only found in specialized dictionaries or term banks. Specialized terminology tends to be less ambiguous than regular vocabulary although regular vocabulary can have a specialized meaning when used in a certain subject area. Some translators report using literature in the target language to get a feel for possible term translations (Tomaszczyk, 1989). Specialized terminology is an area in which the translator might encounter lexical holes (see section 2.3.2.3). Proper nouns or proper names are names of people, places, and organizations. Proper nouns are important in translation but often do not appear in dictionaries (Wilks et al. 1996).

#### *2.3.2.7 False cognates*

False cognates, also known as false friends or *faux ami* are cognates that have different meanings across languages even though they are spelled the same. Left untranslated, false cognates create an erroneous translation. Newmark (1991) refers to this problem as lexical interference, and indicates that this can also be caused by homonyms when they are used in a different, less obvious, sense than is perceived by the translator. Examples of false cognates in Dutch and English are *bevel* and *drop*: *bevel* means *order* in Dutch but *incline* in English, drop means *licorice* in Dutch but *fall* in English.

### *2.3.3 Query translation problems in CLIR*

The CLIR literature identifies four general translation problems: 1) lexical ambiguity; 2) lack of translation coverage; 3) multiword lexemes; and 4) lexical resource creation error. Some of these problems, such as lexical ambiguity, also surface in the translation literature (see previous section) others are unique to CLIR.

[26] An asterisk indicates an nonsensical sentence.

#### *2.3.3.1 Lexical ambiguity*

As discussed previously, lexical ambiguity is caused by lexical items with identical spelling but with different meanings. This problem is compounded in translation due to the exponential increase of translation probabilities. For example, if a term has three source language senses and each of the three senses has two translations, there are suddenly 6 (3 times 2) translation possibilities. Lexical ambiguity is especially problematic in query translation since queries tend to be short and thus lack the context to aid sense selection. In its simplest form, query translation includes all possible translations in the target query, which dilutes the query's original meaning (Kwok, 1997).[27] However, in some cases the addition of alternate translations actually benefits recall because these translations include synonyms of the target term.

When the query translation process uses a lexical resource that provides multiple translations, additional disambiguation capabilities are needed. Yamabana et al. (1998) suggest utilizing the user to aid in sense selection. Obviously this would only work on the source language since the CLIR user cannot be assumed to be multilingual. Although this could cut down the number of possible translations, most researchers take an automatic approach to disambiguation. Davis (1996) uses part-of-speech information from the query processing stage during the dictionary lookup. By restricting the sense selection to the correct parts-of-speech, the number of translation options is reduced. A number of researchers have applied a form of automatic relevance feedback both before and after query translation to improve the query (Ballesteros and Croft, 1996, 1997; Bräschler et al. 1999). Automatic relevance feedback assumes that the top-ranked documents or passages are relevant and uses terms from these documents or passages to expand the query. Source query expansion was found to improve precision whereas target query expansion improved recall. Combining both query expansions improved retrieval even more, especially for short queries. Boolean models which have a self-disambiguating quality have also been used to deal with multiple translations (Hull 1997; Diekema et al. 1999; Hiemstra and Kraaij 1999). Using the OR operator to combine the multiple senses, and the AND operator to connect these sense sets. The assumption is that correct translations tend to appear in documents together unlike incorrect translations.

#### *2.3.3.2 Lack of translation coverage*

Lexical resources often lack information on how to translate abbreviations, acronyms, proper names, and technical terminology. Davis and Ogden (1998) found that out of all the terms their dictionary was not able to translate, 69% were proper names. Wechsler et al. (1997) consider proper names to be language neutral and try matching them across languages. Although a large

[27] A notable exception to this is machine translation (MT). Out of the four translation resources (ontologies, corpora, machine readable dictionaries, and MT), only the latter includes automatic disambiguation in the translation process, resulting in an exact (but not necessarily correct) translation.

number of proper names do not need to be translated between languages that are similar, special proper name dictionaries or omnastica are useful for other cases. A procedure similar to cognate matching (see section 2.2.1) can also be used to discover potential cognates for missing terminology (Davis, 1997). However, failure to recognize lexical items as being proper names or specialized terminology can also lead to erroneous translations (e.g. *Kurt Waldheim* is translated as *Kurt forest home* (Gaussier et al. 1998)).

#### *2.3.3.3 Multiword lexemes*

As described in section 2.3.2.5, the non-standard compositionality, non-standard substitutability, and non-standard morpho-syntactic properties of MWLs cause trouble in translation. Although the use of MWLs in MIR generally don't drastically improve retrieval performance (Fagan 1989; Lewis and Croft 1990; Croft et al. 1991), the use of phrases in CLIR is crucial to its performance (Gey et al. 1998; Ballesteros and Croft 1997; Hull 1997). Ballesteros and Croft (1998) show a 25% performance improvement with manual phrase translation over automatic word-by-word translation. Improvements due to the use of phrases have been reported by others as well (Hull and Grefenstette 1996; Radwan and Fluhr, 1995). The reason for the importance of phrases in CLIR is that the meaning of phrases often gets lost in translation and lexical resources have low coverage of domain specific MWLs (Kraaij 1997; Hull and Grefenstette, 1996). Also, single word compounds in languages such as German and Dutch often translate to phrases in English. Submitting the individual words that make up the compound with their dictionaries will not result in a translation with the same meaning (Gey et al. 1998). For example, a dictionary lookup of *speed* and *limits* is going to translated into (s*poed, snelheid, vaart, haast*) and (*grens, grenslijn, limiet, beperking*) even though one would be looking for the term *snelheidsbeperkingen*.

The literature shows several solutions to deal with MWLs. The most obvious solution is offered by Ballesteros and Croft (1998) who perform a dictionary lookup for phrases and revert to word-by-word translation in case the phrase is not found. Kraaij en Hiemstra (1998) use a translation chart to find the most probable phrase translation based on different combinations of multiple single word translations. Fluhr et al. (1997) employ a similar approach in which they translate phrases word by word but then recombine the single words based on the syntactic rules of the language pairs.

#### *2.3.3.4 Noise in lexical resource creation*

Before translation knowledge can be used by a machine, it needs to be processed. Since this processing is done automatically, errors can be introduced in the process. Wilks et al. (1996) introduce the term machine-tractability to distinguish between the processed lexical resource and a machine readable resource. A machine-tractable dictionary has been modified to remove all extraneous information that is only useful for human users of the dictionary and converted into a

form suitable for a given linguistic task. Hull and Grefenstette (1996) call a bilingual dictionary that has been processed to simply give a translation equivalent of a term a transfer dictionary. Errors are often caused by inconsistencies in the dictionary markup codes and lead to missing entries and the addition of irrelevant words (Hull 1997; Hull and Grefenstette 1996). Errors are also introduced when extracting translation knowledge from corpora due to the probabilistic nature of the alignment process (Carbonnel et al. 1997; Brown 1997).

### *2.3.4 Focus of translation*

The two languages used in the study are English and Dutch. The Dutch language and the English language are considered to be very close since they are both Indo-European, West-Germanic languages. The languages are different enough however, that monolinguals in either language cannot comprehend people from the other language group.

English is spoken by an estimated 350 million native speakers (Katzner, 1995). The language originated in England but spread out all over the world to parts of Africa and Asia, Australia, Canada, Ireland, New Zealand, Scotland, United States, Wales, Scotland (Finegan, 1987). English also spread to the scientific and technical professions where it is used as a *lingua franca*. The simple inflectional rules of English combined with the large *cosmopolitan* vocabulary are also reasons for the popularity of the English language (Finegan, 1987). For example, a number of English words originate in the Dutch language such as *cookie*, *waffle*, *maelstrom*, *yacht* (Katzner, 1995). English is closely related to German and Dutch and, like these two languages, also uses compounding to create new terms. Unlike Dutch where compounds are typically concatenated into a single word three different variations exist in English. Compounds can be written as one word (*iceage*), hyphenated (*ice-skate*), and two words (*ice floe*). No governing body regulates the English language and slight variations exist between the English language of the different countries.

Dutch (Nederlands) is spoken by over 20 million native speakers. Dutch is the official language of the Netherlands, Suriname (South America), the Netherlands Antilles (Caribbean), and Belgium where Dutch is one of the official languages along with French. The term *Dutch* stems from *Diets* or *Duuts* which refers to the (Low) German vernacular (Kooij, 1987). Dutch is close to both English and German. Although Dutch, like English, no longer shares the German noun morphology, it does share its inclination towards compounding. Unlike English, which typically creates noun-noun compounds, Dutch compounds can consist of noun-noun, adjective-noun, noun-

verb, adjective-verb, and particle-verb combinations.[28] Dutch compounds tend to be rather short as compared to German compounds. The Dutch language includes many English loan words such as computer, tennis, and internet. The spelling and grammar of the Dutch language is regulated by the Council for the Dutch Language (Raad voor de Nederlandse Taal) with members from the Netherlands and Belgium.

For this study, the main issue was that English and Dutch are sufficiently different that Dutch queries generally need to be translated before they can retrieve relevant English documents. The two languages are bound to share numerous cognates, and false cognates (see section 2.3.2.7), that have to be taken into account during the experiments, since they could facilitate or hinder CLIR without translation.

## 2.4 IR evaluation

This study used a system evaluation to compare the retrieval results of MIR and CLIR and in doing so studied the impact of translation on retrieval performance. The general purpose of an information retrieval evaluation is to test how well a retrieval system meets information needs. The literature recognizes two ways in which this can be carried out: system evaluation and user-based evaluation (Ellis, 1992).[29] Broadly speaking, a user-based evaluation studies user satisfaction with system interaction whereas system evaluation looks at the quality of the ranking of the document results (Buckley and Voorhees, 1999).

### *2.4.1 IR system evaluation*

A diagnostic evaluation as was used for this study requires a system evaluation which allows a high level of control over all four components of an evaluation: users, databases, searchers, and search constraints (Tague, 1992). Although controlling these variables necessarily makes the evaluation an abstraction of reality, system evaluations have aided the development of numerous retrieval techniques, such as query expansion, that have improved retrieval in operational settings (Buckley and Voorhees, 1999).

The information retrieval system evaluation tradition was started with the Cranfield experiments (Cleverdon and Mills 1963; Cleverdon et al. 1968). Modern day evaluations are carried out in the context of the Text REtrieval Conference (TREC). TREC is sponsored by the National Institute of Standards and Technology (NIST) and the Information Technology Office of the Defense

[28] In addition Dutch has derivational compounds that are not formed by simply adding two terms together but by creative use of suffixes, giving these compounds an almost phrasal quality in, for example, *langslaper* (someone who tends to sleep late) (Kooij, 1987).
[29] To illustrate the strong division between the two evaluations in the field of IR, Ellis (1992) speaks of the physical paradigm (system evaluations) and the cognitive paradigm (user-based evaluations).

Advanced Research Projects Agency (DARPA).[30] For the past seven years, retrieval systems from all over the world have participated in the yearly TREC evaluation. By running the same set of queries across all systems over a large document collection, retrieval performance comparisons are made. The combination of documents, test queries, and their relevance judgments is called a test collection. A test collection is a crucial component in system evaluations (see section 2.4.1.2). Unlike a regular retrieval collection, in a test collection, it is known in advance which documents are relevant and should be retrieved for each of the test queries.[31] For an experimental run, a set of test queries is used by the retrieval system to retrieve documents from the set of test documents. Based on the retrieved documents and the relevance judgments, evaluation measures concerning retrieval effectiveness can be calculated. The evaluation measures reflect how well a system does at finding relevant documents and ignoring irrelevant documents (Van Rijsbergen, 1981).

#### *2.4.1.1 Evaluation measures*

The most commonly used evaluation measures are recall and precision. These two measures, either individually or combined, form the basis for most system evaluation measures (see also section 3.4.1.3). Recall is the proportion of all the relevant documents that are retrieved. Precision is the proportion of retrieved documents that are relevant. For example, a certain document is either relevant to a query or it is not.[32] The same document can either be retrieved or not. By looking at the evaluation contingency table in Figure 2.4 we can determine recall and precision:

$$\text{recall} = \frac{\text{number of relevant documents retrieved}}{\text{total number of relevant documents}} = \frac{a}{(a+c)}$$

$$\text{precision} = \frac{\text{number of relevant documents retrieved}}{\text{number of retrieved documents}} = \frac{a}{(a+b)}$$

[30] The goals of TREC are: 1) to encourage research in text retrieval based on large test collections; 2) to increase communication among industry, academia and government by creating an open forum for the exchange of research ideas; 3) to speed the transfer of technology from research labs into commercial products by demonstrating substantial improvements in retrieval methodologies on real-world problems; and 4) to increase the availability of appropriate evaluation techniques for use by industry and academia, including development of new evaluation techniques more applicable to current systems (Voorhees and Harman, 1999).

[31] This is actually no longer true since the test collections have become too big to judge all documents against all queries. A technique called relevance pooling (see section 2.4.1.2) is used to solve this logistical problem and it discovers the most relevant documents for each query given certain assumptions.

[32] The unrealistic binary view of relevance is of course one of the major criticisms of system evaluations (see section 2.4.3).

It is misleading to report either only recall or only precision. Both measures have to be used together since system performance can vary widely depending on the number of documents retrieved (Hull, 1993).[33]

[33] If the cutoff point where one measures recall or precision is lower than the number of relevant documents it is impossible to obtain a recall of 1. If the cutoff is higher than the number of relevant documents and the number of relevant documents is low, it is impossible to obtain high precision, even though all relevant documents might have been retrieved in the top ranks.

| | Relevant | Non-Relevant |
|---|---|---|
| **Retrieved** | a | b |
| **Not-Retrieved** | c | d |

**Figure 2.4 Evaluation Contingency Table.**

A standard TREC retrieval evaluation results in 23 different performance measures (see table 2.5 below). However, the measure that is typically reported in the information retrieval literature is average precision.

| | Measure | Description |
|---|---|---|
| 1 | Recall at 1000 | After retrieving 1000 documents, how many of the known relevant documents were found? Set based measure over a group of 1000 retrieved documents, ignores rank. (Combines relevant and relevant retrieved measures). |
| 2-12 | Interpolated recall-precision averages at levels: 0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0. | Each recall precision average is computed by summing the interpolated precisions at the specified recall cutoff value (denoted by $\Sigma \boldsymbol{P}\lambda$, where $\boldsymbol{P}\lambda$ is the interpolated precision at recall level $\lambda$) and then dividing by the number of topics. $\frac{\sum_{i=1}^{NUM} P\lambda}{NUM} \quad \lambda = \{0.0, 0.1, 0.2, 0.3..., 1.0\}$ |
| 13 | Average precision | Non-interpolated for all relevant documents (averaged over queries). Average of the precision value obtained after each relevant document is retrieved. Reflects performance over all relevant documents and rewards systems that rank relevant documents high. |
| 14-22 | Precision at 5, 10, 15, 20, 30, 100, 200, 500, 1000 documents | Precision after *n* documents have been retrieved. Number of relevant documents retrieved *@rank n*/ *n*. |
| 23 | R-Precision | Precision after R documents have been retrieved, where R is the relevant documents for the topic. De-emphasizes exact ranking. Relevant documents retrieved / R. |

**Table 2.5 The 23 standard TREC measures.**

*2.4.1.2 TREC test collection and relevance pooling*

Information retrieval test collections preceding TREC were found to be too small because findings based on these collections did not apply to collections of a larger, more realistic size (Sparck Jones and Van Rijsbergen, 1975). With the TREC test collections, the size problem has been solved by providing large test collections. However, due to the large size of these collections, it is no longer feasible to find all relevant documents by comparing all documents against all test queries. For the TREC test collection, a new set of 50 topics is created each year. TREC distinguishes between topics which are the rather extensive statements of the information need provided by TREC (see Figure 3.2) and queries which are the statements submitted to the system.[34] The topics are created

[34] Queries can be based on the entire topic or on parts thereof.

by intelligence analysts and the documents are judged for relevance by the same analyst who generated the topic. As in this study, TREC considers relevance to be topical similarity. The judges are instructed to judge a document relevant if they would include it in a report on the topic independent of the amount of the information used. The 50 topics are used by TREC participants to retrieve sets of 1000 documents for each topic from the test documents.

To create a pool of relevant documents the top 100 documents for each topic from each participant are included in a so called relevance pool. The judges go through the relevance pools for each topic after the removal of duplicates. The assumption is that the most relevant documents are bound to be included in the pool and the rest of the collection does not need to be judged. All documents that are not included in the relevance pool are assumed to be not relevant. This process is called relevance pooling. For relevance pooling to work correctly one needs a large number and variety of search methods to the relevance pool and an adequate pool depth (TREC uses a depth of 100).

### *2.4.2 CLIR system evaluation*

In MIR experiments, researchers commonly vary the information retrieval system while keeping the test queries and documents constant. This allows for comparison between systems or comparison between different versions of the same system. The same practice is followed in CLIR experiments when comparing different systems. However, CLIR experiments vary the test queries rather than the system, to allow for comparison between the cross-language and monolingual capabilities of the same stem. The experiments in this research rely on varying the test queries.

By manually translating test queries into a foreign language and using these test queries as the cross-language equivalents, the cross-language performance can be compared directly to the monolingual performance of a system (see Figure 2.6).

Manual translation of queries is now a widely used evaluation strategy because it permits existing test collections to be inexpensively extended to any language pair for which translation resources are available. The disadvantage of this evaluation technique is that manual translation requires the application of human judgment, and evaluation collections constructed this way exhibit some variability based on the terminology chosen by a particular translator. Some insight into the contribution of alternative translation techniques can be obtained by comparing CLIR results with the effectiveness of a similar monolingual technique on the same collection. Typically expressed as a percentage of monolingual effectiveness, reported values range from around 50% for unconstrained dictionary-based query translation to 98% or so for more sophisticated techniques.

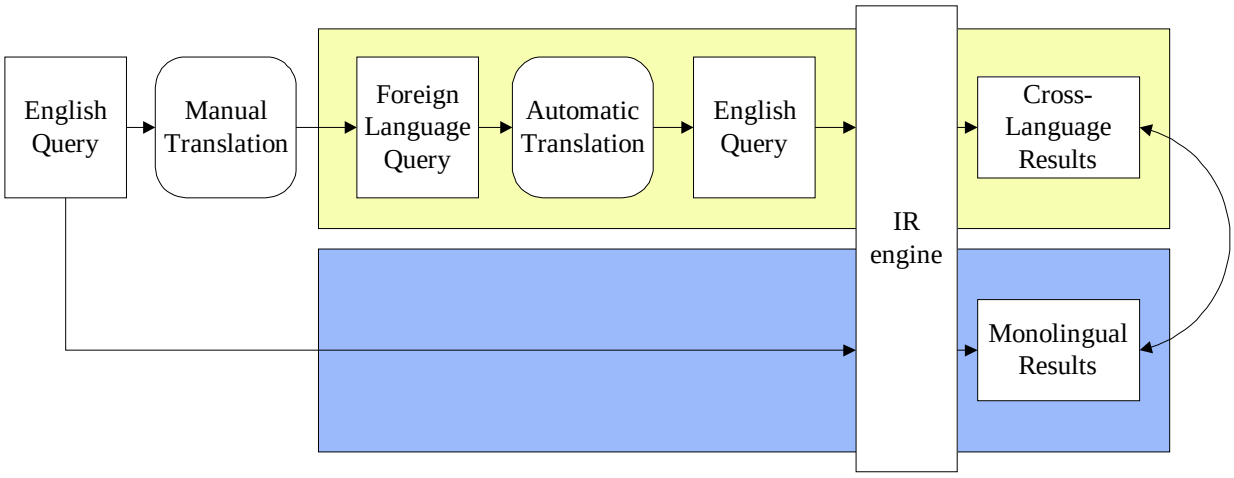


**Figure 2.6 CLIR system evaluation.**

Other CLIR evaluation techniques that are described in the literature are all to some degree based on mate finding in document aligned parallel corpora (Radwan and Fluhr 1995; Davis and Dunning 1995). In these evaluations, documents, or parts of documents, are used as foreign language queries in an attempt to find their parallel foreign language mate. Mate finding is a variation on known-item retrieval, a classic evaluation strategy in which the rank assigned to a unique item that is known to be relevant to the query is used as the measure of effectiveness (Landauer and Littman, 1991, 1992).

#### *2.4.2.1 CLIR test collection*

The Cross-Language Evaluation Forum (CLEF) test collection is the most comprehensive step in the direction of the creation of a large European language CLIR test collection (Braschler and Peters 2002). At the inception of this study, the CLEF test collection only had English, French, German, and Italian documents in its collection and the number of queries was relatively small. Recently Dutch and Spanish documents have been added. Modeled after the ad hoc TREC collection, the CLIR test collection is also based on the pooled relevance method (see section 2.4.1.2). The CLEF test collection had certain limitations caused by violations of the assumption behind the pooled relevance method, and did not have enough queries to be suitable for this study. For these reasons this study used the TREC ad hoc topics.

### *2.4.3 Methodological issues with IR system evaluations*

Information retrieval system evaluations have been criticized for some of the assumptions behind their evaluation. System evaluations assume that 1) relevance can be approximated by topical similarity; 2) relevance is binary; 3) relevance of a document is independent of other documents; 4)

relevance of a document is static; 5) user judgments are representative of all users; and 5) relevance judgments are available for all documents (Salton 1992; Hull 1993). These assumptions have been criticized as being rather unrealistic (Eisenberg and Barry 1988; Schamber et al. 1990; Saracevic 1970, 1975; Ellis 1996).

However, Voorhees (1998) has shown that even varying relevance judgments do not affect comparative system performance. Voorhees found that only some experimental runs swapped positions with changing sets of relevance judgments. Most queries have a core set of relevance judgments that seems to stabilize comparative rankings. As pointed out by Buckley and Voorhees (1999), the goal in a system evaluation is to compare systems and to obtain relative scores of evaluation measures not an absolute score. System evaluations do just that.

An additional criticism stems from the relevance pooling method (see section 2.4.1.2). The criticism is that the pools do not contain all relevant documents and that might disproportionally affect different systems. An incomplete pool would mean that the relevance judgments are incomplete and that the part of the collection outside the relevance pool and designated not-relevant may be relevant after all. Research has shown that the relevance pools are indeed incomplete (Harman, 1996). By extending the pool depth to 200 documents a median of 30 new documents were found per topic. It was also found that unique retrieval approaches tend to retrieve documents that other systems do not. This would mean that beyond the top 100 that the system contributed to the pool, the 900 additional documents, or a large portion thereof, would be considered not-relevant even though they might be relevant. Systems that take a more common approach do not have this problem because more documents (beyond their top 100) would have been retrieved by systems like them. However, Zobel (1998) found that even though results of similar systems did benefit somewhat from the pool, there was no correlation between the effectiveness score and the number of unjudged documents. It appears that relevance pooling is a sufficient approximation of judging every single document.

## 2.5 Summary

Term matching in CLIR is complicated because it requires matching between nearly distinct vocabularies. Nearly all CLIR approaches apply some form of translation to cross the language barrier and use a variety of lexical resources. The prevailing CLIR approach, and the focus of this study is query translation. By using ontologies, bilingual dictionaries, machine translation systems and corpora, the query is translated into all the languages supported by the system to retrieve documents in these languages. The translation of the query is inherently difficult due to the lack of a one-to-one mapping of a lexical item and its meaning, thus creating lexical ambiguity. In addition, query translation is complicated by the cultural differences between language communities and the way they lexicalize the world around them. These two translation issues

create many different translation problems such as lexical ambiguity, lexical mismatches, and lexical holes. In turn, these and other translation problems result in translation errors which impact CLIR performance. This study investigated the impact of these translation events in retrieval by carrying out an information retrieval experiment using the manual query translation evaluation method with the English TREC test collections.

# 3 Methodology

## 3.1 Introduction

The following chapter presents the methodological approach chosen to answer the study's research questions:

1) *What kinds of translation events affect cross-language retrieval?*
2) *In what way does the presence of certain translation events in query translation affect retrieval performance?*

The nature of the research questions required a two-phase multi-method approach (see Figure 3.1). The first phase was an exploratory phase in which a substantial number of queries and their translations were examined in combination with a review of the translation literature and CLIR literature. A translation event taxonomy was created through content analysis and an examination of the literature. In the second and final phase, 750 queries were coded using the taxonomy resulting from phase one. These queries were then used in information retrieval experimentation to assess the impact of the translation events on information retrieval performance. Since each phase and related research question required a different approach, this chapter is organized by research phase.

## 3.2 Phase one

Phase one of the study examined the first research question: *What kinds of translation events affect cross-language retrieval?* The answer resulted in a comprehensive taxonomy describing the sources of translation error in CLIR based on an extensive query analysis and review of the translation literature and CLIR literature.

### *3.2.1 Data collection phase one*

During phase one, two types of data were collected: queries and literature on translation problems.

A large number of queries are needed to discover which types of translation events face CLIR. At the same time, these queries should be suitable for the information retrieval experiment of phase two (see Figure 2.6). Phase two required these queries to be part of a test collection with available relevance judgments. The most commonly used modern-day test collection is the TREC ad hoc test collection which has 400 (query numbers 051-450) queries that were used in ad hoc retrieval experiments (Harman 1995; Voorhees and Harman 1999) (see also 2.4.1.2).

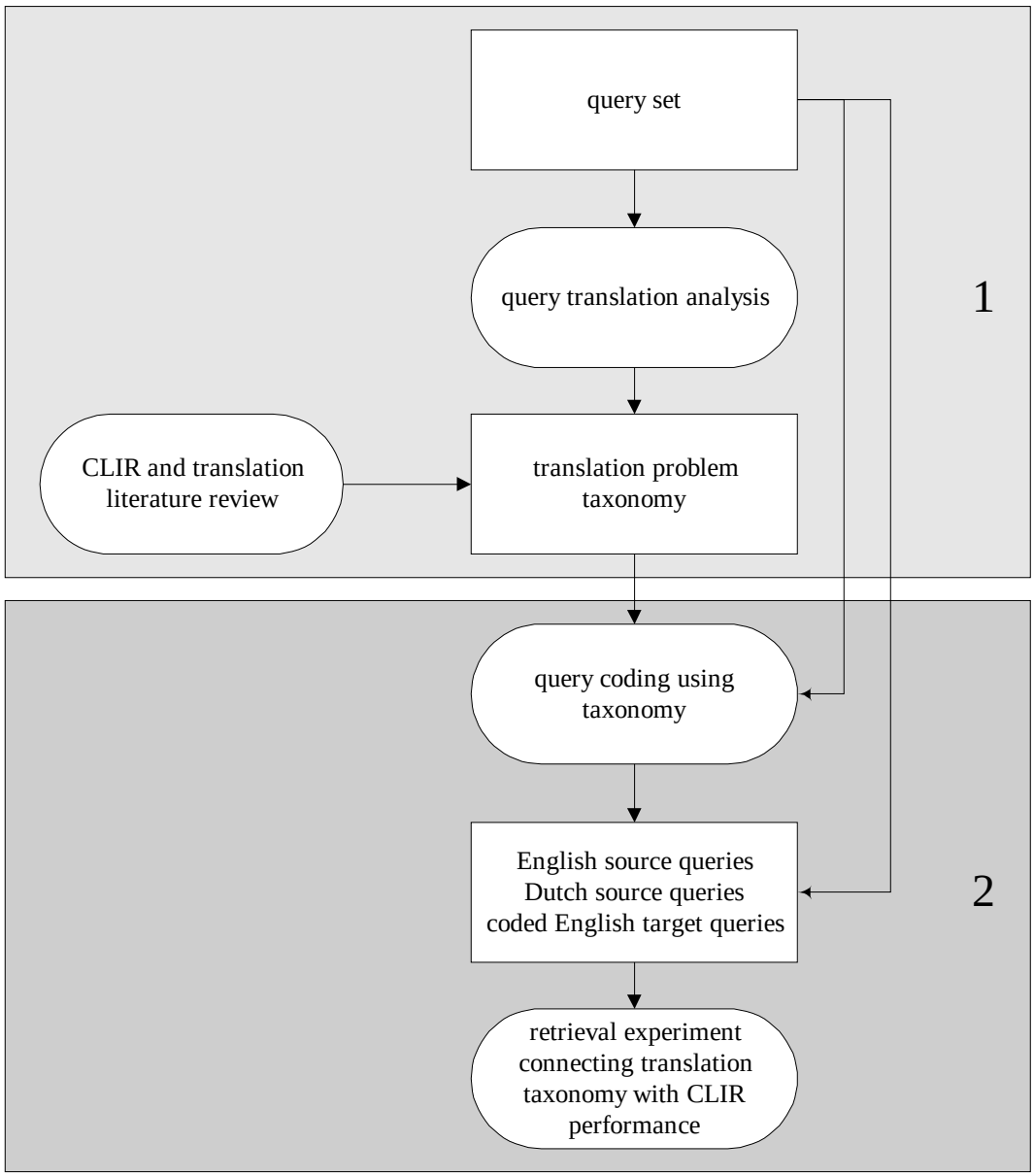


**Figure 3.1 The two research phases.**

The majority of TREC topics consist of 3 parts: a title, description, and a narrative (see Figure 3.2).[35] The title is the shortest version of a topic and describes it in a few key terms. The description is a longer version of a topic and describes the topic in a full sentence. The narrative lists explicitly what constitutes a relevant document and a non-relevant document. The narrative is intended to aid TREC judges in their relevance assessments but is often used as part of the topic for document retrieval. This study only used the title and description since these better resemble a

[35] This format is not consistent for all query sets. Notable exceptions are query numbers 051-100 and 201-250, the first having additional query fields, and the latter having only short description queries.

user query than the relatively long narrative (Cutting et al., 1997). Using both the title and the description provided two versions of the same query varying in query length.

```
<top>
<num> Number: 255
<title> Topic: Environmental
Protection

<desc> Description:
Name countries that do not practice
or ignore environmental protective
measures.

<narr> Narrative:
Nations that do not practice or
ignore environmental protective
controls degrade the progress other
nations have made in this vital area.
There are international efforts to
protect the environment. The actions
of some countries are of some
concern, however, because they may be
ignoring efforts to conserve and
protect the world's resources. The
objective of this topic is to
identify countries that do not have
environmental controls.

</top>
```

```
<top>
<num> Number: 426
<title> law enforcement, dogs

<desc> Description:
Provide information on the use
of dogs worldwide for law
enforcement purposes.

<narr> Narrative:
Relevant items include specific
information on the use of dogs
during an operation. Training of
dogs and their handlers are also
relevant.

</top>
```

**Figure 3.2 TREC Topics 255 and 426.**

For phase one, the study obtained manual Dutch translations for all of the 750 (400 titles plus 350 descriptions) queries. The manual Dutch translations were carried out by an independent Dutch translation bureau to eliminate the researcher from the translation process.[36] These manual Dutch translations can be considered Dutch equivalents of the English originals and both sets are viewed as source queries as is common in CLIR evaluations (see section 2.4.2).

The Dutch source queries were automatically translated back into English using a machine readable dictionary (Van Dale Groot Woordenboek Nederlands-Engels, 1997). To accomplish this automatic translation, a term list first needed to be extracted from the dictionary itself. The set of unique TREC query terms was used for dictionary lookup. All dictionary entries for each term were collected in a single file, which was processed to create a straight term translation list. This term list was used for the automatic translation.

[36] Translation bureau: *Drs. S.J. van der Ploeg B.V.*

All Dutch source queries were tokenized (based on white space) and each content term was looked up in the term list described above. If a term was on the list, it was replaced with all the translations listed in the entry. If the term was not on the list, it was kept untranslated. The original terms are kept in this case because these terms are often proper names that transcend the language barrier.

Data collection resulted in 750 source and target language query triples where each query has an English source query, a Dutch source query, and an English target query (see Figure 3.3).

| English source query | Dutch source query | English target query |
|---|---|---|
| The use of dogs worldwide for law enforcement purposes | *het wereldwijd gebruik van honden bij de ordehandhaving* | world-wide, throughout the world, all over the world, use, application, consumption, taking, custom, habit, practice, usage, dog, hound, cur, pooch, poochy |

**Figure 3.3 Sample source and target language triple.**

Aside from the collection of queries, phase one also required a review of the CLIR and translation literatures on translation problems. The collection of these data served two purposes: 1) verification of the translation problem categories resulting from the query translation analysis (see section 3.2.2), and 2) augmentation of the translation event taxonomy with additional translation event categories.

### *3.2.2 Data analysis phase one*

The data collection for phase one resulted in the following data: 750 source and target language query triples (English source query, Dutch source query, and English target query), and a listing of the translation problem categories as found in the literature. The first step in the data analysis was to perform content analysis (Babbie 1992) on the query triples. The analysis led to the discovery of different translation event categories.

By comparing the original English queries with the English translated queries and the English translated queries with their Dutch source queries (see Figure 3.4) it became clear what problems were encountered during the translation. For example, in Figure 3.3 the term *ordehandhaving* (law enforcement) got lost in the translation because it did not appear in the lexical resource. Another problem that emerged was the polysemous Dutch term *gebruik* (use) which resulted in the addition of all possible translations (*use*, *application*, *consumption*, *taking*, *custom*, *habit*, *practice*, *usage*) to the query although, for example, *custom* is clearly wrong in this case.

To carry out comparative analysis, a practice from descriptive translation studies (Toury, 1995), the queries were divided into textual segments (terms or phrases) to study translation equivalence. For example, the query "*the use of dogs worldwide for law enforcement purposes*" was divided into 6 segments (ignoring the articles and prepositions). Each segment was then aligned with its automatic translation to facilitate a comparison. This resulted in a number of source and target sub-triples for each query, that are each other's translations (see Figure 3.5). This sub-division of the query triples into translation equivalents is similar to the term alignment process in parallel corpus processing.

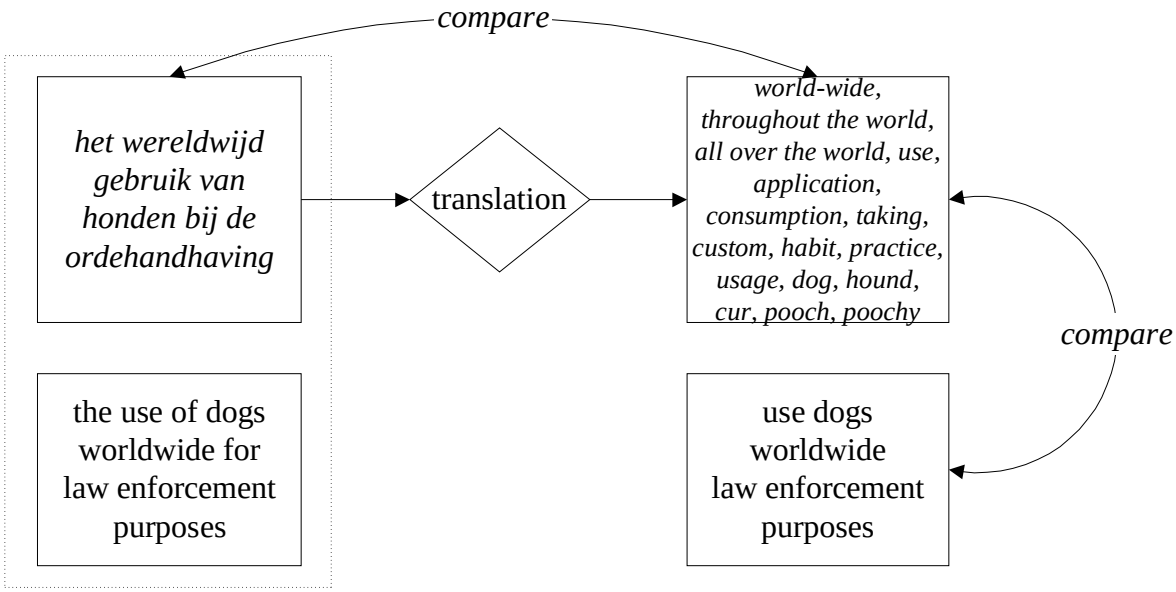


**Figure 3.4 Basic word-by-word translation after stop word removal.**

It should be pointed out that in descriptive translation studies the division into segments is normally limited to translation pairs, not triples[37] as is the case here (see also section 3.2.3). Another difference is in the level of analysis. Descriptive translation studies seem to focus on the origins of translations and the process of their creation rather than the lower level term equivalence comparison. Clearly, the lower level is more appropriate for this study since cross-language information retrieval is based on these lower level items (terms).

[37] For a comparative analysis in a descriptive translation study, the Dutch source query and its target translation would be studied as the translation pair.

| English source query | Dutch source query | English target query |
|---|---|---|
| Worldwide | *Wereldwijd* | world-wide, throughout the world, all over the world |
| Use | *Gebruik* | use, application, consumption, taking, custom, habit, practice, usage |
| Dogs | *Honden* | dog, hound, cur, pooch, poochy |
| Law | *ordehandhaving* | no translation |
| Enforcement | | |
| Purposes | | |

**Figure 3.5 Example of a source and target sub-triple.**

After establishing the source and target language sub-triples, content analysis took place. The unit of analysis in the study was the query, the unit of observation was the source and target language sub-triple. The sub-triples were examined for the presence of translation events (problems and successes), and these events were categorized into mutually exclusive and exhaustive translation event categories. These categories formed the basis of the translation event taxonomy (see section 4.2). For record keeping purposes the coding scheme used a numerical notation. Additional query features such as length, number of stop words, and number of sub-triples were recorded in the process.

The process to develop the translation event categories by processing the sub-triples in batches can be viewed as an inductive/deductive cycle (Shelly and Sibert, 1992). After processing a first batch of sub-triples a number of patterns emerged (induction), which were then tested against the next batch of sub-triples (deduction). For the study this means that the patterns were established in coding categories of translation events with certain attributes. With these categories the researcher revisited the data to see whether the categories still applied or needed to be changed and whether additional categories were needed. Each coding category represented a translation event and was identified by an event name, examples from the data of this type of event, and a definition of the event containing attributes of the event that distinguish it from the other event categories. Once the event categories were complete and no new categories were found, saturation of the event categories occurred. Once saturation occurred, a first version of the event taxonomy was created. This event taxonomy was then compared to the translation events that emerged from the literature review after which certain categories were adapted or added. This second event taxonomy was then tested for reliability.

To test the reliability of the translation event taxonomy, 16 random source and target language query triples were coded by the researcher and a Dutch native speaker with 20 years of English language experience. Cohen's Kappa Coefficient, a chance corrected inter-rater agreement measure for nominal data, was used to assess the reliability of the taxonomy (Cohen 1960; Banerjee 1999). The Kappa coefficient measures the proportion of inter-rater agreement after chance agreement has been removed.[38] For the suggested reliability test the coders were provided with: 1) a taxonomy description, 2) a taxonomy chart, 3) a coding sheet, and 4) a list of queries. Before starting the actual test the coders were provided with two test queries that were used to explain the process. After the process was clear the coders proceeded with the set of 16 queries. The queries were a mix of title and description queries – this feature was also selected at random.

The 16 queries in the reliability test had a total of 68 content words and together these words had 379 entries in the dictionary. Each of these dictionary entries was coded. For each content word there were four possible coding moments with codes representing: issues with the lexical resource, issues with the source word, issues with terms that remain untranslated, and issues with the translation building process itself (for a more detailed description of the coding process see section 3.3.1.2). The final number of coding instances for all the 68 content words was 1,116.[39]

After separately coding each query, the two coders went over the codes to discuss the agreements and disagreements before proceeding to the next query. The numbers reported below report the original number of disagreements even though some disagreements were resolved in discussions. However, the discussions did likely improve coding of the queries that followed. The coding proceeded until the same categories were seen over and over again and no new categories occurred. It is important to note that not all categories in the translation taxonomy appeared in this data.

Out of the total number of 1,116 coding instances, the coders agreed on 1006 instances and disagreed on 110 instances. The main reason for this high level of agreement was the occurrence of certain code sequences. For example, a 1.1 (lack of translation coverage) always led to a b1 (missing translation). If the coders agreed on 1.1 they automatically agreed on the b1 code as well. In addition, many of the categories were straightforward and quite discrete. Both of these reasons lead to an artificially high Kappa coefficient.

---

[38] $K = \frac{P_O - P_C}{1 - P_C}$, where $P_o$ is the observed portion of agreement and $P_c$ is the proportion of agreement expected by chance.

[39] Each dictionary entry can have up to 4 codes.

As mentioned above, the Kappa coefficient measures the proportion of inter-rater agreement after chance agreement is removed. To calculate this coefficient you measure observed proportion of agreement between coders and the proportion of agreement expected by chance. The more the observed proportion of agreement exceeds that of agreement expected by chance, the higher the coefficient. The reliability study reported in this paper resulted in a Kappa of 0.87.[40] A reliability score of this magnitude is considered to be excellent. Minor changes were made to the taxonomy to prevent disagreement in the most common areas.

### *3.2.3 Methodological and other issues*

The assumption behind CLIR evaluation of query equivalence might have affected the internal validity of the study. In CLIR research it is commonly assumed that the source pairs (Dutch and English in this case) are each other's exact equivalents. The equality assumption originates in the CLIR evaluation practice of capturing the relative performance of CLIR as compared to MIR. This is done by taking a retrieval system, its documents and queries, and translating the queries into a foreign language and retranslating the queries back into the original language (see section 2.4.2). By running both sets of queries, one can compare the performance of MIR and CLIR. The assumption that the two source queries (one for each language) are equivalent is not quite accurate since the foreign language source queries are a translation from the original queries and there are multiple ways they could have been translated. Hull and Grefenstette (1996) point out that the first (manual) translation to create the two source queries might already cause some translation error. Unfortunately this problem is inherent to the CLIR experimental method, as well as natural language in general, which exhibits vast variations in word choice, and cannot be resolved.

The use of a single unidirectional language pair (Dutch-English), the choice of lexical translation resource, and the use of TREC queries might restrict the external validity of the study. The use of translation and CLIR literatures however serves to soften the limitation of use of a single unidirectional language pair. It is conceivable that the translation resource used influenced the degree in which certain translation events were found. For example, a translation resource with only a few thousand entries is likely to result in more missing translation cases than might otherwise be the case. However, it is unlikely that a completely different set of translation events might be found with a different resource. The limitations brought on by the use of TREC topics,

[40] ( (0.8996 – 0.2134) / (1 – 0.2134) ) = 0.87, where 0.8996 is the observed portion of agreement, and 0.2134 is the proportion of agreement expected by chance.

although still a concern, is somewhat reduced by using a very large number of queries of varying length on various subjects.

## 3.3 Phase two

Phase two of the study answered the second research question *In what way does the presence of certain translation events in query translation affect retrieval performance?*

### *3.3.1 Data collection phase two*

The objective of the information retrieval experiments was to discover the extent of the impact of certain translation events on retrieval performance. To study the effect of the translation events, the English target queries were coded using the taxonomy developed in phase one, the codes indicating which events were present in each query. The experimental setup followed the CLIR practice of using source queries representing the monolingual situation and target queries representing the cross-lingual situation (see section 2.4.2). To measure the performance of CLIR as compared to MIR, the source and target retrieval runs can be compared (see Figure 3.6). This phase compared English source queries to English target queries (translated from Dutch into English).

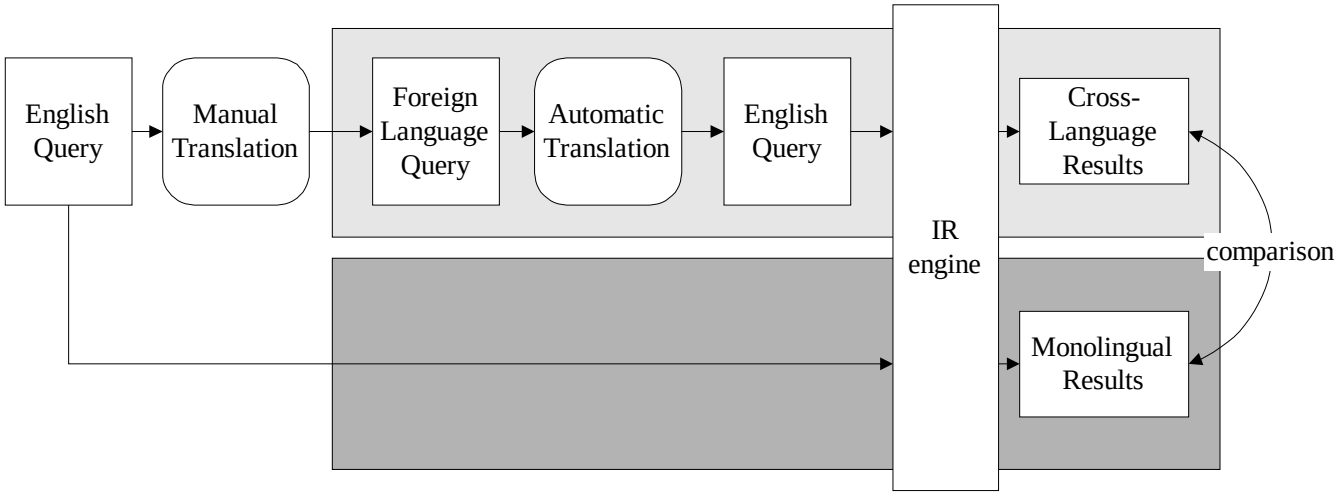


**Figure 3.6 Experimental design.**

#### *3.3.1.1 Experimental design*

The most intuitive approach to assessing the impact of the different translation events on retrieval performance is to compare retrieval performance of target queries with their source equivalents that did not undergo a translation step. However, since TREC topics tend to have anywhere from two to fifty relevant concepts (each liable to contain multiple translation events), partitioning queries into single event categories is unlikely (Hull, 1999). Restricting the study to one-word queries would solve this query categorization problem but would have made the study unrealistic. To address the problem of queries possibly having multiple translation events, the study used

multiple regression. Multiple regression allows two or more independent variables to predict scores on a dependent variable (Kerlinger and Pedhazur 1973; Roscoe 1975). Instead of using one independent variable (one of the many possible translation events) at a time, multiple regression allows the combination of numerous translation events to influence the dependent variable (retrieval performance).

In the study, the response variable represented the difference in retrieval performance between translation and no translation for each query (see Figure 3.7). The performance measures (see section 3.3.1.4) were based on the run results of the two different query sets (source and target queries). Translation was expected to have a negative impact on retrieval performance. The predictor variables represented the different translation events identified in phase one of the study plus additional factors that might have an impact on translation performance such as query length, and the number of senses (sense density).

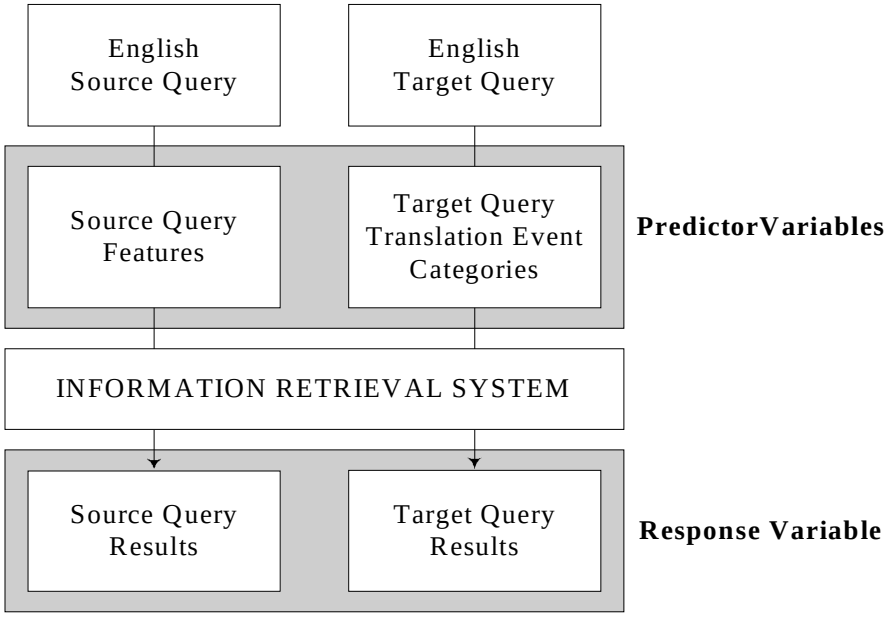


**Figure 3.7 Predictor and response variables.**

*3.3.1.2 Coding queries*

The translation events of all 750 English target queries were categorized using the translation event taxonomy. Using the taxonomy to categorize translation events in queries is perhaps best illustrated by using Figure 3.8 when coding a query. The Dutch source query for title query 190 *Gevallen van computerfraude (Instances of Fraud Involving the Use of a Computer)* has two terms to be translated: *gevallen* and *computerfraude* (*van* is a stopword and will be ignored). As Figure 3.8 indicates, there are four coding moments: Class I, translation builder, Class II, and Class III. Each

of these moments has a column in the coding chart (see Figure 3.9). For a detailed description of the taxonomy itself and its four components, see section 4.2.

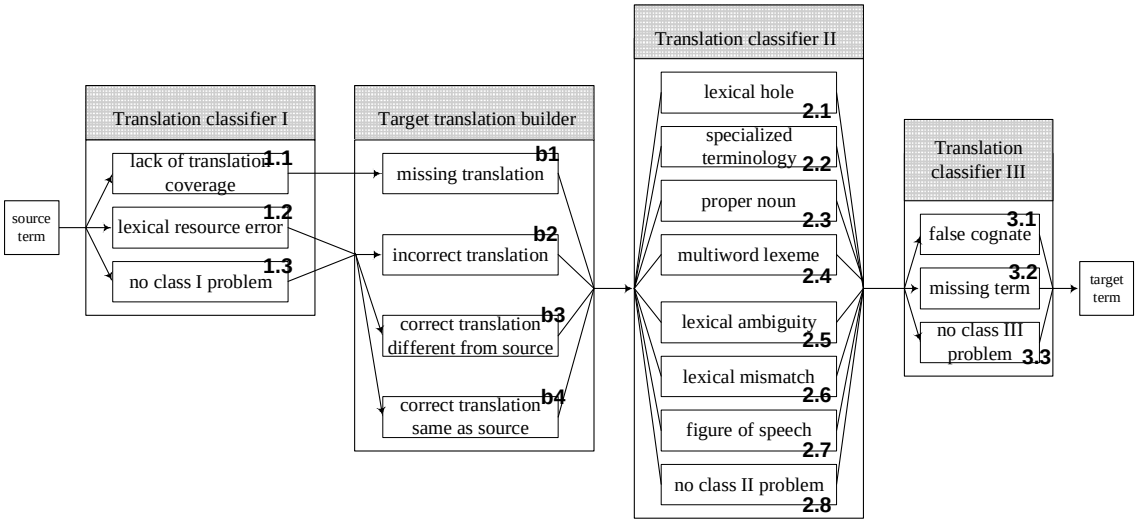


**Figure 3.8 Translation event taxonomy.**

In this example, the lexical resource did not contain any errors so all query term translations will have the initial code of 1.3 (coding Class I: 1.3 - no class I problem). A lookup of *gevallen* indicates that it has three entries, two of which are incorrect. The first entry (*gevallen1*) is incorrect (translation builder coding: b2 - incorrect) due to lexical ambiguity (coding Class II: 2.5 - lexical ambiguity). The third entry (*gevallen3*) has some correct translations – namely for the first noun sense – (coding translation builder: b3 - correct but different from source) but also some incorrect translations (coding translation builder: b2 - incorrect). Both are reflected in the coding chart. The second term (*computerfraude*) was translated correctly though different from the English source equivalent (coding translation builder: b3 - correct but different from source).

#### *3.3.1.3 Experimental process*

Although the study originally started out with 750 queries, query equivalency issues forced the researcher to remove 20 queries from the analysis. The premise of the retrieval experiments assumed equivalency between Dutch source query and English source query (see section 3.2.3). The 20 queries that violate this assumption have been removed from the analysis: 084D, 103D, 104T, 111D, 116T, 125D, 131D, 132D, 168D, 169D, 173D, 204D, 252D, 258D, 283D, 320D, 342D, 398D, 411D, 437D.

| Entry | POS | CLASS I | translation | builder | CLASS II | CLASS III |
|---|---|---|---|---|---|---|
| gevallen1 | \|ADJ | *1.3* | 1fallen | *b2* | *2.5* | *3.3* |
| gevallen2 | \|VB | *1.3* | 1come about | *b2* | *2.5* | *3.3* |
| gevallen2 | \|VB | *1.3* | 1come to pass | *b2* | *2.5* | *3.3* |
| gevallen2 | \|VB | *1.3* | 1happen | *b2* | *2.5* | *3.3* |
| gevallen3 | \|NN | *1.3* | 1case | *b3* | *2.8* | *3.3* |
| gevallen3 | \|NN | *1.3* | 1affair | *b3* | *2.8* | *3.3* |
| gevallen3 | \|NN | *1.3* | 2circumstances | *b2* | *2.5* | *3.3* |
| gevallen3 | \|NN | *1.3* | 2position | *b2* | *2.5* | *3.3* |
| gevallen3 | \|NN | *1.3* | 2situation | *b2* | *2.5* | *3.3* |
| gevallen3 | \|NN | *1.3* | 4thing | *b2* | *2.5* | *3.3* |
| gevallen3 | \|NN | *1.3* | 4contraption | *b2* | *2.5* | *3.3* |
| gevallen3 | \|NN | *1.3* | 4device | *b2* | *2.5* | *3.3* |
| gevallen3 | \|NN | *1.3* | 4contrivance | *b2* | *2.5* | *3.3* |
| gevallen3 | \|NN | *1.3* | 5chance | *b2* | *2.5* | *3.3* |
| gevallen3 | \|NN | *1.3* | 5luck | *b2* | *2.5* | *3.3* |
| computerfraude1 | \|NN | *1.3* | 1computer fraud | *b3* | *2.8* | *3.3* |

**Figure 3.9 Coding sheet for title query 190.**

Equivalence in translation is a very difficult issue and the topic has been quite extensively written about in the translation literature (Larson 1984). Absolute equivalency is almost impossible, especially for longer queries, so when do you consider the lack of equivalency problematic? Problems seem to arise when certain words and phrases are completely missing from the Dutch source query, or, when additional terms and phrases have been added to the Dutch source query. The Dutch query translations tend to be somewhat terse in their formulation, often leaving out information that is implied. Information that is already implied in other query terms is not included as a separate term. For example the term *aircraft* is absent from the Dutch query about McDonnel Douglas (131D) because the company is known to produce planes. Missing terms or phrases are problematic because the translation from Dutch back into English can never get back to the English source even if the translation is perfect. Queries with missing terms can be clearly identified. A few of the Dutch query translations add more specific terminology to the query in an effort to explain precisely what the user is interested in. Additional terms are problematic because they would give the Dutch source query an unfair advantage over the English source query. For example, in query 132D about "stealth" aircraft projects the Dutch query adds that these projects concern bombers and warplanes that are invisible to radar.

It should be made clear that the query equivalency issues, as described here, are different from the translation problems that are the focus of this study. An omission of a term from a source query is inherently different from an omission of a term in the target query. In the latter case the omission is caused by the lack of that term in the translation resource. The addition of a (more specific) term to the source query does not happen in an automatic translation where the only expansions are synonyms. Thus, removing queries with equivalency issues will not reduce the number of queries with certain translation events. In addition, since the omissions and additions are not caused by an inherent language difference, there is no danger in accidentally removing the translation events that are the object of this study.

The English target queries were coded using the taxonomy as described above. For each English source and target query pair, a retrieval run was carried out. Query 296 (both the title and the description) had an average precision of zero and was also removed from the analysis. The study was thus carried out with 728 queries: 347 title queries, and 381 description queries. The retrieval performance of both query versions was combined in a difference score (see section 3.3.1.4). All the information for each query (the translation event codes, the difference score, query length, and sense density) informed the data analysis (see chapter 4).

#### *3.3.1.4 Experimental measures*

The measures that were used to evaluate retrieval performance are all based on recall and precision (see section 2.4.1.1). A standard TREC retrieval evaluation results in 23 different performance measures. An analysis of the correlations between these measures for the monolingual run showed that nearly all measures were highly correlated. To avoid repeatedly testing hypotheses on highly correlated dependent variables, the choice was between a composite measure or selecting a single measure for testing purposes. Initially, a composite measure (*information retrieval quality*) was created using principal component analysis. However, the measure that is typically reported in the information retrieval literature is average precision. Average precision was also the measure that was most highly correlated with all the other 22 measures. For the sake of interpretability and dissemination of results, average precision is used as the main retrieval performance measure in this study.

Average precision over all relevant documents measures the precision at each relevant retrieved document which is then averaged per query and per query set. This non-interpolated measure favors systems that assign high ranks to relevant documents. Average precision over all relevant documents has good discrimination power, and is very suitable for system analysis and system tuning.

Since this study examined the difference in retrieval performance between monolingual and cross-lingual retrieval, a difference measure needed to be created as the independent variable. Two things are problematic about using straight difference scores to create the dependent variable: the issue of relative difference, and violation of the normality assumption which underlies parametric inference testing. Initially there is the problem of relative difference. That is, a small difference in scores, where the monolingual score is small to begin with, might be more important than a bigger difference with a large monolingual score. This problem can be solved by normalizing the score difference by dividing the difference score by the monolingual score for that query pair:

$D = \frac{P - P'}{P}$, where $P$ is the monolingual score and $P'$ is the cross-lingual score.

The second problem, lack of normality, occurs because the (relative) difference score *D* combines two different distributions, which often results in a distribution that does not lend itself to inference testing (i.e. multiple regression analysis). Since the distribution of difference score *D* is not suitable to be used for multiple regression testing we used unstandardized residual scores instead. Residual scores are created by regressing the monolingual average precision score on the cross-lingual average precision score. The residual score represents the variance in the cross-lingual score after the variance that can be explained by the monolingual score has been removed. Because the sign of the residual score does not necessarily indicate whether the difference in scores is positive or negative, it is difficult to interpret. Thus the difference score *D* is used in all non-inferential analyses and the unstandardized residual score is used for inferential purposes. The two dependent variables are highly correlated (Spearman's Rho -.781, $p < .001$).

#### *3.3.1.5 Test environment*

The study did not use the TREC CLIR test collection (see section 2.4.2.1) because at the time of this study's inception there were problems with the CLEF relevance pool used to create the relevance judgments.[41] In addition to these relevance pooling issues, the collection was relatively new and did not yet have a sufficient number of queries available for the study (see also section 3.2.1).[42] For these reasons, the study used the main TREC ad hoc retrieval test collection for the

[41] For the cross-language test collection there might not have been enough participants to guarantee a large enough relevance pool. Also, instead of using monolingual runs for each of the four languages to create the relevance pool, only the low quality merged run was used. Using the merged run has another negative implication. When using the pool depth of 100 of a merged run of four languages, one could conceivably find only 25 relevant documents for each language. This is problematic because 82% of the queries for TREC-7 had more than 25 known relevant documents.

[42] Since this study's inception, the Cross-Language Evaluation Forum (CLEF) has built up an impressive European language test collection. Although some of the same issues concerning the pooling method apply to the CLEF query sets, it is the current standard for cross-lingual research in European languages (Peters, 2002).

experiments in Phase two. Since the topic sets are spread out over multiple collections and all topics were used, the entire TREC collection had to be indexed (see Figure 3.10).

| TREC Collections | Topics |
|---|---|
| Text Research Collection Volume 1, Revised March 1994 Collection includes material from the Wall Street Journal (1987, 1988,1989), the Federal Register (1989), Associated Press (1989), Department of Energy abstracts, and Information from the Computer Select disks (1989, 1990) copyrighted by Ziff-Davis. | 051-100<br>101-150<br>151-200 |
| Text Research Collection Volume 2, Revised March 1994 Collection includes material from the Wall Street Journal (1990, 1991,1992), the Federal Register (1988), Associated Press (1988) and Information from the Computer Select disks (1989, 1990) copyrighted by Ziff-Davis. | 051-100<br>101-150<br>151-200<br>201-250<br>251-300 |
| Text Research Collection Volume 3, Revised March 1994 Collection includes material from the San Jose Mercury News (1991), the Associated Press (1990), U.S. Patents (1983-1991), and Information from the Computer Select disks (1991, 1992) copyrighted by Ziff-Davis. | 201-250 |
| Text Research Collection Volume 4, May 1996 Collection includes material from the Financial Times Limited (1991,1992, 1993, 1994), the Congressional Record of the 103rd Congress (1993), and the Federal Register (1994). | 251-300<br>301-350<br>351-400[43]<br>401-450 |
| Text Research Collection Volume 5, April 1997 Collection includes material from the Foreign Broadcast Information Service (1996) and the Los Angeles Times (1989, 1990). | 301-350<br>351-400<br>401-450 |

**Figure 3.10 TREC test collection.**

The retrieval system used in this study was implemented by the researcher and is based on City University's[44] Okapi (Online Keyword Access to Public Information) system (Robertson, 1997). The retrieval system used Okapi's BM25 ($k_1$=1, $k_3$=0.6, $b$=8) weighting function as described by Robertson et al.(1995).[45] Okapi is a probabilistic retrieval model and weighs each term based on a combination of four sources of information: collection frequency, term frequency, document length, and relevance (not used in this study). The sum of the term weights gives a document score and reflects a document's relevance to a query.

### *3.3.2 Data analysis phase two*

Data collection for phase two resulted in so called query vectors, each element in the vector representing information about a certain query. Each vector contained the query ID number, query length (for each of the different query versions), query sense density (for the source queries),

[43] The whole volume minus the Congressional Record.
[44] London, England.
[45] These tuning parameters can be adjusted to the test collection at hand but remained as listed here (TREC-4 settings) throughout all experiments in this study.

information about the number of each of the seventeen different translation events that occurred in that query, and retrieval scores for the different query versions.

BM25 $$\sum_{T \in Q} w \frac{(k_1+1)tf}{K+tf} \frac{(k_3+1)qtf}{k_3+qtf}$$

where,

$w$ Robertson–Sparck Jones weight (BM1):

$$w = \log \frac{(r+0.5)/(R-r+0.5)}{(n-r+0.5)/(N-n-R+r+0.5)}$$

| | |
|---|---|
| $R$ | number of documents known to be relevant = 0. |
| $r$ | number of these documents that contain the term= 0. |
| $N$ | number of documents in the collection |
| $n$ | collection frequency (number of documents containing the term) |
| $k_1$ | tuning parameter = 1 |
| $tf$ | term frequency (term occurrence in document) |
| $k_3$ | tuning parameter = 0.6 |
| $qtf$ | term frequency in query |
| $K$ | tuning parameter for document length: $K = k_1((1-b)+b\frac{dl}{ave.dl})$ |
| $b$ | tuning parameter = 8 |
| $dl$ | document length |
| $ave.dl$ | average document length |

**Figure 3.11 Best Match function 25.**

Exploratory data analysis was carried out to become acquainted with the data, identify outliers and other oddities, and to summarize data. Data transformations were carried out for those variables that were highly skewed (see section 4.4). A multiple regression analysis was carried out to discover whether any of the previously identified translation events caused a significant change in retrieval performance.

The regression analysis indicated that four of the independent variables contributed significantly to prediction of the residual difference in retrieval performance when comparing monolingual to cross-lingual information retrieval: missing specialized vocabulary, missing general terms, wrong translation due to ambiguity, and correct identical translation. Although the contribution of each of these variables is significant, their contribution to the total variance of the independent variable was small (see section 4.4).

### *3.3.3 Methodological and other issues*

In order to establish the impact of translation events on retrieval performance, it is essential to control for other factors possibly influencing retrieval performance such as the query, and the information retrieval system. Although neither Ruiz and Srinivasan (1998) nor McCarley (1999) found a relationship between statistical query features such as query length and CLIR performance, other literature suggests that these features might be important to information retrieval in general (Salton and McGill, 1984). Because query length correlated with all translation event variables, it turned out to be an interfering factor. Thus, the translation event variables were normalized by query length.

The study attempted to minimize the effect of query variability by using a large number of queries. This is important because IR experiments have shown that the difference in effectiveness between individual queries can be larger than the differences for the same query on different systems. The standard number of queries used by TREC to carry out system comparisons is 50. However, a power analysis by Diamond (1999) shows that the optimal test sample size is actually much larger (225).[46] Even though this study used a sizeable number of queries (728), the results still suffered from the effect of query variability (see section 5.3).

Judging by the remarkably close performances of the different systems participating in TREC-7 (Voorhees and Harman, 1999), the actual information retrieval system used in the experiments is less of an issue as a possible source of variation.

The validity of the experiment is determined by the assumptions behind both information retrieval evaluation and CLIR evaluation. It has long been argued that the relevance assumptions behind information retrieval evaluation are unrealistic (Saracevic 1975; Salton 1992; Ellis 1996; Blair 2002). Clearly, relevance is not binary, nor is it static or user independent, nor are all relevant documents of equal importance. However, the relevance judgments that are produced based on these assumptions are quite suitable for comparisons among systems.

The reliability of test collections, and thus of information retrieval experiments, has been called into question because different judges, and even the same judge over time, have been known to

---

[46] A power analysis by Diamond (with $\alpha$ = .05 (z=1.66), $\beta$ = .20 (z=.84), the minimum interesting difference between runs .05, and query precision standard deviation estimated at .30) estimates that the optimal test sample size is 225 and a sample of 50 might obscure differences between runs:

$$Sample \quad size = \left( \frac{\left( (1.66 + 0.84) * 0.30 \right)}{0.05} \right)^2 .$$

produce different relevance judgments. The problem of variable judgments, however, does not change comparative system performance and thus does not affect a test collection's reliability (Voorhees, 1998) (see also section 2.4.3).

As pointed out previously in section 3.2.3, the equivalency assumption behind CLIR evaluations that require a manual translation step is somewhat flawed. An unknown amount of translation error is created during the manual translation of English into Dutch and added to the translation error of the target translations from Dutch into English. The expected change in retrieval performance due to translation error might not be caused by translation from Dutch into English alone. Unfortunately this additional source of error is inherent to the experimental design and cannot be controlled.

### 3.4 Summary

The methodology to answer the study's two research questions required a two phase multi-method approach. The central element in all phases are the content equivalent query source and target language triples. The English source queries were created by taking title and description fields of 400 TREC topics. Manual translation of the English source queries into Dutch resulted in the Dutch source queries. Automatic translation back into English provided English target queries. Each query triple thus consists of an English source query, a Dutch source query, and an English target query. The first phase of the study used content analysis of the query triples, in combination with the CLIR and translation literature, to create a translation event taxonomy. During the second phase, the English target queries were coded using the translation event taxonomy. The English source and target queries featured in an information retrieval experiment resulting in performance scores. A multiple regression analysis was carried out to assess how the translation events impact retrieval performance.

# 4 Results

## 4.1 Introduction

This chapter presents the results of this study of translation events and their impact on information retrieval performance. As pointed out in chapter 3, the study followed a two-phase multi-method approach. In phase one, a taxonomy of translation events was created through content analysis of queries and their translations, in combination with an examination of the translation and cross-language information retrieval literature. In the second and final phase, the test queries were coded using the taxonomy resulting from phase one. These queries were then used in an information retrieval experiment to assess the impact of translation events on retrieval performance.

The first section of this chapter describes the translation event taxonomy resulting from phase 1 of this research. The next section describes the information retrieval experiment: the queries, the translation events, and retrieval results. The final section discusses an in-depth query analysis.

## 4.2 Taxonomy

The creation of the taxonomy was based on a review of the literature and analysis of the query triples. A query triple is a set of three equivalent queries: an English source query, a Dutch source query, and the English target query. For a more detailed description of query triples see section 3.1. The taxonomy was created in answer to the first research question: *What kinds of translation events affect cross-language retrieval?*

An analysis of a subset of the query triples was carried out to study the effect of translation on the queries. Content analysis showed that there are three kinds of translation events: 1) events concerning the lexical resource, 2) events concerning the source word, and 3) events concerning terms that remain untranslated. Events concerning the lexical resource are discovered at the initial term lookup. The term might not be listed, or its lexical entry may have errors as a result of the automatic conversion from machine readable dictionary to translation term list. Events concerning the source word capture possible linguistic issues that might hinder a translation. For example, the meaning of a non-compositional noun phrase might get lost in a word-by-word translation. Events concerning untranslated words affect terms that did not have an entry in the translation resource. Not only is the source term lost, but additional problems occur when an untranslated word exists in the target language but has a different meaning entirely (false cognate).

It is the result of a particular translation event that impacts information retrieval, not the actual event itself. It is important therefore that in categorizing the query using the translation taxonomy,

we also get an indication of what happened in the translation building stage. The translation building stage is the process of translating a source query term by automatic term lookup, and replacing that term with the foreign language equivalent term(s) listed in the resource. Or, in case the term is not listed, keeping the original source term. For example, lexical ambiguity sometimes results in an erroneous translation. Other times the ambiguity results in both an erroneous translation as well as a correct one.

Examination of the word-by-word translation building process revealed four basic events in a translation: 1) the term is translated correctly and is identical to the term used in the English source query, 2) the term is translated correctly but is not identical to the term used in the English source query translation, 3) the term is translated incorrectly, 4) the term could not be translated. In short, either the translation was correct, incorrect, or had no translation at all. This is of course a simplification of an actual translation where one will find several possible combinations of these four events (i.e. both correct and incorrect translations for a single entry), and variations in the magnitude of these events (i.e. depending on the number of translations found in a dictionary entry).

When we combine the three translation classifiers with the translation builder we get a query categorization schema (see Figure 4.1) that combines all aspects of the translation: the translation events, the result of the events, and the magnitude of the events.

For example, the translation codes for the query term *financiering* from title query 112 (EST[47]: *funding biotechnology*, DST: *financiering biotechnologie*, ETT: *financing, funding, bioengineering, biotechnology*) are as follows (see Figure 4.1):

> *financiering* = *financing*, 1.3|b3|2.8|3.3 – the term is listed in the dictionary (code 1.3), is translated correctly although the translation is different from the one used in the English source query (b3), there were no other problems with the term (code 2.8), (code 3.3).
> *financiering* = f*unding*, 1.3|b4|2.8|3.3 – same as the previous term but this time the translation is the same as the one used in the English source query (code 4). For a more detailed description about coding queries, see section 3.3.1.2.

The taxonomy was tested for reliability in a test where two judges coded 16 queries. The test resulted in a Kappa of 0.87. A reliability coefficient of this magnitude is considered to be excellent. The Kappa coefficient measures the proportion of inter-rater agreement after chance agreement has

[47] The query naming convention is as follows: EST=English source title, ESD=English source description, DST=Dutch source title, DSD=Dutch source description, ETT=English target title, ETD=English target description.

been removed. To calculate this coefficient you measure the observed proportion of agreement between coders, and the proportion of agreement expected by chance. The more the observed proportion of agreement exceeds that of agreement expected by chance, the higher the coefficient.

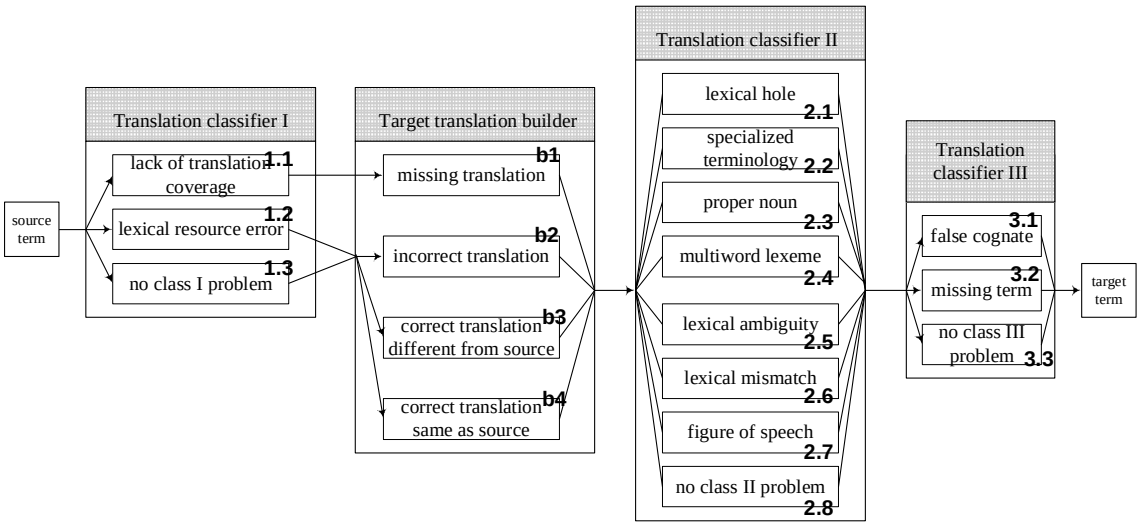


**Figure 4.1 Translation event taxonomy.**

## 4.3 Information retrieval experiment

### *4.3.1 Queries*

The English source and target queries were used in an information retrieval experiment, where the English source queries represented monolingual retrieval, and the English target queries represented cross-lingual retrieval.

The study originally started out with 750 queries: 350 title queries, and 400 description queries. Twenty queries were removed due to equivalency issues (see section 3.3.1.3) and two queries were removed because of average precision scores of zero for their monolingual runs. The study was thus carried out with 728 queries: 347 title queries, and 381 description queries. Each query had three versions: a Dutch source query, an English source query, and an English target query. The English source query was manually translated into Dutch to create an equivalent foreign language query. This Dutch source query was automatically translated back into English to create the English target query.

The following kinds of data were collected about the queries: query characteristics, translation events, and retrieval performance scores. Query characteristics are those query features that might

influence retrieval performance of that query, such as query length or the level of ambiguity of a query. Translation events were assigned to each query using the Taxonomy (see section 4.2) which specified what took place during the translation of the source query into the target query. Translation events indicate whether or not a term could be translated and whether it was translated correctly, etc. Retrieval performance scores were collected for the English source queries as well as the English target queries to compare performance between monolingual and cross-lingual information retrieval. The sections below examine the data collected about the queries in more detail.

#### *4.3.1.1 Query characteristics*

Query characteristics that were measured for this study were query length, source query sense density, and translation expansion effect. Query length is measured by the number of content words. The assumption is that longer queries will do better in cross-language retrieval. This is due to the automatic disambiguation effect between the different translations. Table 4.1 shows that the Dutch and English source queries have similar length characteristics. The English target query on the other hand is quite different. Since all possible translations are added to a translated query, its length can increase substantially. There is high variability in length among target queries, as is shown by the large standard deviation.

| *Query length in content words* | | | | | | |
|---|---|---|---|---|---|---|
| measure | mean | st. dev. | median | mode | min. | max |
| Dutch source | 5.69 | 4.14 | 4.50 | 2.00 | 1.00 | 24.00 |
| English source | 5.71 | 3.87 | 4.00 | 3.00 | 1.00 | 23.00 |
| English target | 34.92 | 32.99 | 25.00 | 1.00 | 1.00 | 206.00 |

**Table 4.1 Query length in content words.**

It is assumed that the greater the number of senses each word of the source query had, the more difficult the translation is going to be. Sense-density was measured for the English source query as well as for the Dutch source query. Sense density for the English source queries was measured by using the WordNet 6.0 sense index. For each query term the number of senses was counted and added to the total number of senses for that query. The total was then divided by the target query length to get the sense density for that particular query. Sense density for the Dutch source queries was calculated using the number of senses listed in the Van Dale Dutch-English dictionary. As is shown in Table 4.2, the sense-density for English source queries is somewhat higher than that of the Dutch source queries. The reason for this is likely to be twofold. The Dutch queries contain terms that do not occur in the Van Dale dictionary, and are thus only given one sense in the total sense count. In addition, the Dutch language uses more compounds. The longer the compounds, the fewer senses these terms tend to have.

| *Query sense density* | | | | | | |
|---|---|---|---|---|---|---|
| Measure | mean | st. dev. | median | mode | min. | Max |
| Dutch source | 2.58 | 1.39 | 2.33 | 1.00 | 1.00 | 10.00 |
| English source | 3.51 | 2.09 | 3.00 | 1.00 | 1.00 | 14.00 |
| *Query expansion effect* | | | | | | |
| Measure | Mean | st. dev. | median | Mode | min. | Max |
| English target | 4.43 | 2.93 | 4.15 | 0.00 | -0.50 | 19.67 |

**Table 4.2 Query sense density and expansion effect.**

The expansion effect measures the increased size of the English target query as compared to the Dutch source query from which it originated. The assumption is that the higher the expansion effect, the greater the number of erroneous translations (noise) have been added to the query. Expansion effect was calculated by subtracting the Dutch source query length from the English target query length, and dividing that number by the Dutch source query length. Table 4.2 shows that, on average, queries increased in size by a factor of 4.4.

### *4.3.2 Translation events codes*

All 728 queries were coded using the translation event taxonomy. The different translation events were the independent variables of this study. Each Dutch source query term – English target translation term pair was given an individual code vector. The vectors represent the query categorization schema. Although the translation event taxonomy presented 43 possible[48] translation events ranging from missing terms to terms without any problems, not all translation events were represented in the data. The 25,433 vectors traveled through the schema in 17 different ways (see Figure 4.2). Out of these 17 vectors, a number of vectors appear fewer than 10 times, and others well over a thousand times. The vector distribution is thus heavily skewed. Even though the translation event taxonomy is very expressive, not many different events seem to occur in this sizeable query set.

The four most common code vectors were: *wrong translation due to ambiguity*, *correct translation but different*, *wrong translation due to multi-word lexeme*, and *correct identical translation*. More than half the terms were translated incorrectly due to ambiguity (52.49%). In this case the Dutch source terms had multiple senses and all their translations were added to the query, some of them incorrect. Just under a quarter of all terms (22.77%) were translated correctly

[48] Although there are 288 unique ways to travel through the translation taxonomy chart (see Figure 4.1), only 43 paths are viable. For example, once a term is established as missing from the translation resource, the subsequent taxonomy path is restricted because the term can no longer be marked as incorrect or correct.

but used a synonym of the English source term. The distinction between *correct identical translation* and *correct translation but different* was made because this could potentially cause a retrieval performance difference between source and target queries. Just over 9% of the terms were translated correctly with the same terms used in the English source query. 11.86% of the terms were parts of phrases that should not be translated word-by-word but had been. The remaining 3.86% of translation events is spread out over 13 different code vectors.

To get an indication of how individual query terms fare in a translation, straight term counts were collected (in addition to the term-translation-pair counts presented earlier). A total of 4,360 query terms had to be translated. Out of these Dutch source terms, 2,280 (52.29%) had at least one translation that was correct and identical to that of the English source query. An even larger number of terms, 2,571 (58.97%) had at least one translation that was correct as a synonym of the English source query equivalent. A total of 3561 (81.67%) terms had at least one correct translation (either identical or a synonym to the Dutch source equivalent).

| code vector name | Vector | Frequency |
|---|---|---|
| **IV1: reverse lexical hole** | 1.1\|b1\|2.1\|3.3 | 2 |
| The Dutch language has no term for this phenomenon so it is forced to use the English terminology. (196ETT voucher in school vouchers) | | |
| **IV2: missing specialized terminology** | 1.1\|b1\|2.2\|3.2 | 108 |
| The source term is very specific and does not occur in a regular dictionary. The term cannot be translated and the source term is not a cognate. (170ETD borstimplantaten siliconengel – silicone gel breast implants) | | |
| **IV3: specialized terminology** | 1.1\|b1\|2.2\|3.3 | 35 |
| The source term is very specific but Dutch uses the English term. (066ETT natural language processing ) | | |
| **IV4: missing Proper Name** | 1.1\|b1\|2.3\|3.2 | 3 |
| The Proper Name in the source language does not appear in the dictionary. The term is lost in the translation. (163D Zuid-Vietnam – South Vietnam) | | |
| **IV5: Proper Name** | 1.1\|b1\|2.3\|3.3 | 154 |
| Again, the Proper Name in the source language does not appear in the dictionary but is a cognate. (109D Minnesota Mining and Manufacturing) | | |
| **IV6: missing Multi Word Lexeme** | 1.1\|b1\|2.4\|3.2 | 5 |
| The source term is part of a multi-word lexeme and is lost in the word-by-word translation. (091D met name genoemde ("name" not in dictionary) | | |
| **IV7: missing general term** | 1.1\|b1\|2.8\|3.2 | 77 |
| Term not in the dictionary (no obvious reason). The term is lost in the translation. (258D bevoegdheid) | | |
| **IV8: general term or number** | 1.1\|b1\|2.8\|3.3 | 20 |
| Term (number) not in the dictionary but is a cognate. (053D "200", or 167D sex) | | |
| **IV9: lexical creation error** | 1.2\|b2\|2.8\|3.3 | 87 |
| Error in creation of the lexicon, term translated incorrectly. (449D verklaren translated as "zich verklaren") | | |
| **IV10: wrong translation due to lexical hole** | 1.3\|b2\|2.1\|3.3 | 6 |
| Term cannot be translated back to the original English source term since the Dutch language does not have an equivalent term. Term translated incorrectly. (169D local, state, federal – Dutch source has regional instead of state). | | |
| **IV11: wrong translation of Proper Name** | 1.3\|b2\|2.3\|3.3 | 346 |
| The English Proper Name in the Dutch source query is translated (false cognate). Erroneous terms added to translation and possibly (parts of) the Proper Name get(s) lost in translation. (320 Fiber Optic Link around the Globe (FLAG) - link is translated as sly.) | | |
| **IV12: wrong translation due to Multi Word Lexeme** | 1.3\|b2\|2.4\|3.3 | 3017 |
| Dutch source query has a MWL that needs to be translated as a whole but gets lost in the word-by-word translation. Erroneous terms are added to the translation and the MWL might be lost in the translation as well. (416D stand van zaken – should be translated simply as status) | | |
| **IV13: wrong translation due to ambiguity** | 1.3\|b2\|2.5\|3.3 | 13350 |
| Dutch source term has multiple meanings. Erroneous terms are added to the translation. (130T betrekkingen – should be translated as relations but also gets job, and position.) | | |
| **IV14: wrong translation due to lexical mismatch** | 1.3\|b2\|2.6\|3.3 | 4 |
| The way of thinking between the two languages differs, term cannot be translated back to the English source term. (063 vertaalprogramma – translation programs but should be machine translation.) | | |
| **IV15: wrong translation due to figure of speech** | 1.3\|b2\|2.7\|3.3 | 134 |
| Figure of speech gets translated literally. Erroneous terms added to the translation and the meaning is lost. (069 nieuw leven inblazen - literally: blow new life into, but should be translated as revive) | | |
| **IV16: Correct translation but different (from English source)** | 1.3\|b3\|2.8\|3.3 | 5791 |
| Correct translation but translation is a synonym of the term used in the English source query. (195 fluctuaties – in English source as "shifts", translated as : fluctuation, drift, change, instability, swing.) | | |

| IV17: Correct identical translation | 1.3\|b4\|2.8\|3.3 | 2294 |
|---|---|---|
| Correct translation and identical term as the term used in the English source query. (405D heelal – translated as cosmos) | | |

**Figure 4.2 The seventeen translation code vectors.**

### *4.3.3 General retrieval performance*

As expected, the data shows that the majority (85.85%) of all cross-lingual queries performed below their monolingual counterparts. The remainder of the queries either showed no difference (3.57%), or outperformed the monolingual query (10.58%). A Wilcoxon Signed-Ranks Test revealed that this difference between monolingual and cross-lingual performance is statistically significant, Z(-20.195), $p < .001$.

To illustrate the dissimilarity between the two languages in this study, we compared the performance (in average precision) of the English source queries and the Dutch source queries in the identical task of retrieving English documents. Not surprisingly, there is significant difference between the two runs, Z(-21.884), $p < .001$. Mean average precision over all 728 queries is 0.2078 for English source queries and 0.0424 for Dutch source queries (see Table 4.3).

| Statistic | ES-run | ET-run | DS-run | ET-allcorr | ET-icorr | ES-ET/ES |
|---|---|---|---|---|---|---|
| Mean | 0.2078 | 0.0764 | 0.0424 | 0.0959 | 0.1160 | 0.3475 |
| SE of Mean | 0.0073 | 0.0050 | 0.0048 | 0.0053 | 0.0061 | 0.1180 |
| Median | 0.1400 | 0.0146 | 0.0000 | 0.0280 | 0.0352 | 0.8226 |
| Mode | 0.0005 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 1.0000 |
| Std. Dev. | 0.1982 | 0.1342 | 0.1305 | 0.1426 | 0.1656 | 3.1832 |
| Minimum | 0.0001 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | -58.2308 |
| Maximum | 0.9481 | 0.8100 | 0.8763 | 0.8100 | 0.8100 | 1.0000 |

**Table 4.3 Average precision data for all different runs.**

Two variations of the English target queries were created to study the effect of the erroneous terms that were added to the query. The ET-allcorr queries only contain query terms that are correct but different, and correct and identical. The ET-icorr queries contain only those translations that were correct and identical to those found in the English source queries. The effect of removing the error terms is revealed in the mean average precision of these two runs (see Table 4.3). Both runs do significantly better than the English target queries, -on Signed Ranks Test ($Z_{ET\text{-}allcorr}$ -10.205, $Z_{ET\text{-}icorr}$ -9.532), $p < .001$.

## 4.4 Statistical analysis

A statistical analysis and a query analysis (see section 4.5) was carried out to help answer the second research question: *In what way does the presence of certain translation events in query translation affect retrieval performance?*

### *4.4.1 The variables*

As is clear from the description in section 4.3.2, not all translation event independent variables display enough variability to be included in a statistical analysis. The decision about which independent variables to leave out was based on the following correlation analysis, with the dependent variable defined as the difference score *D* (see section 3.3.1.4), the normalized difference between monolingual and cross-lingual average precision. First, all the independent variables were normalized by query length to reduce the effect of length on the correlations. If the (Spearman's Rho) correlation between an independent variable and the dependent variable was significant ($p < .01$, 2-tailed), the variable was left in the analysis (see Table 4.4). This left 5 translation event independent variables; *missing specialized terminology* (IV2); *missing general terms* (IV7); *wrong translation due to ambiguity* (IV13); *correct translation but different* (IV16); and *correct identical translation* (IV17). Although correlations for these 5 variables were significant at the .01 level, the correlations were weak, ranging from (-).139 to .366. The two query characteristic independent variables were added for comparison. It is important to note that the latter two variables are highly correlated between themselves (rho = .834, $p < .001$).

The combination of dependent variable (*unstandardized residual differences*) and independent variables resulted in extended query vectors where elements of the vector represent the different variables. Each extended query vector contained nine values: query ID, 7 independent variables (which included the translation event variables), and one dependent variable. These extended query vectors were used in the multiple regression analysis.

Because the translation events are spread out over 17 different variables, some of them occurred so infrequently that they could not be included in the analysis. However, when we collapse the 17 translation events into three main translation events (*correct*, *missing*, and *wrong*), we can include all variables in the analysis. The *correct* category contains all those translation events where the translation was either correct and identical to the source query (IV17), correct but different from the source query (IV16), or the term could not be translated but was a cognate (IV1, IV3, IV5, and IV8). This *missing* category included variables that could not be translated for various reasons (IV2, IV4, IV6, IV7).

The *wrong* category included all variables where the translation was erroneous (IV9, IV10, IV11, IV12, IV13, IV14, IV15). Thus the 17 different translations were aggregated into 3 main translation events: *correct* (8,296), *missing* (193), and *wrong* (16,944). The extended query vectors for these aggregated variables contained seven values: query ID, 5 independent variables, and one dependent variable.

| IV | variable name | *f* | mean | SE mean | me-dian | mode | st. dev. | min. | max | $r_s$ | *p* |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | reverse lexical hole | 2 | 0.00 | 0.00 | 0 | 0 | 0.05 | 0 | 1 | 0.005 | 0.899 |
| 2 | **missing specialized terminology** | 108 | 0.15 | 0.02 | 0 | 0 | 0.41 | 0 | 3 | 0.290 | 0.000 |
| 3 | specialized terminology | 35 | 0.05 | 0.01 | 0 | 0 | 0.29 | 0 | 4 | 0.089 | 0.017 |
| 4 | missing proper name | 3 | 0.00 | 0.00 | 0 | 0 | 0.06 | 0 | 1 | 0.069 | 0.063 |
| 5 | proper name | 154 | 0.21 | 0.04 | 0 | 0 | 0.99 | 0 | 18 | -0.059 | 0.113 |
| 6 | missing multi-word lexeme (MWL) | 5 | 0.01 | 0.00 | 0 | 0 | 0.08 | 0 | 1 | 0.074 | 0.046 |
| 7 | **missing general terms** | 77 | 0.11 | 0.01 | 0 | 0 | 0.34 | 0 | 3 | 0.147 | 0.000 |
| 8 | general terms and numbers | 20 | 0.03 | 0.01 | 0 | 0 | 0.17 | 0 | 2 | 0.084 | 0.024 |
| 9 | lexical creation error | 87 | 0.12 | 0.01 | 0 | 0 | 0.38 | 0 | 3 | 0.062 | 0.096 |
| 10 | wrong translation due to lexical hole | 6 | 0.01 | 0.01 | 0 | 0 | 0.22 | 0 | 6 | -0.008 | 0.825 |
| 11 | wrong translation of proper name | 346 | 0.48 | 0.12 | 0 | 0 | 3.17 | 0 | 45 | 0.008 | 0.835 |
| 12 | wrong translation due to MWL | 3017 | 4.14 | 0.37 | 0 | 0 | 10.11 | 0 | 74 | 0.061 | 0.098 |
| 13 | **wrong translation due to ambiguity** | 13350 | 18.34 | 0.76 | 12 | 0 | 20.49 | 0 | 134 | 0.233 | 0.000 |
| 14 | wrong translation due to lexical mismatch | 4 | 0.01 | 0.00 | 0 | 0 | 0.10 | 0 | 2 | -0.016 | 0.674 |
| 15 | wrong translation due to figure of speech | 134 | 0.18 | 0.08 | 0 | 0 | 2.18 | 0 | 42 | 0.036 | 0.337 |
| 16 | **correct translation but different from source** | 5791 | 7.95 | 0.28 | 6 | 0 | 7.66 | 0 | 42 | -0.139 | 0.000 |
| 17 | **correct identical translation** | 2300 | 3.15 | 0.09 | 3 | 2 | 2.50 | 0 | 15 | -0.366 | 0.000 |
| 18 | **Dutch sense density** | n.a. | 5.69 | 0.15 | 4.5 | 2 | 4.14 | 1 | 24 | 0.206 | 0.000 |
| 19 | expansion effect | n.a. | 4.43 | 0.11 | 4.15 | 0 | 2.93 | -0.5 | 19.67 | 0.235 | 0.000 |

**Table 4.4 IV descriptive statistics, and correlation with difference score *D***

### *4.4.2 Multiple regression analysis*

A standard multiple regression analysis was carried out between unstandardized residual average precision as the dependent variable, and *missing specialized terminology* (IV2), *missing general*

*terms* (IV7), *wrong translation due to ambiguity* (IV13), *correct translation but different* (IV16), *correct identical translation* (IV17), and *Dutch sense density* (IV18) as independent variables. (*Expansion effect*, IV19, was left out of the analysis because of its high correlation with *Dutch sense density*).

The translation event independent variables were normalized by target query length, and underwent an arcsine transformation (recommended for proportional data such as these). *Dutch sense density* underwent a log transformation after a constant of 0.5 was added to each variable, and the dependent variable underwent a square root transformation (plus 0.5). There were no missing data (N=728).

An initial regression analysis was carried out to find extreme cases. Even after the transformations, the use of a $p < .001$ criterion for Mahalanobis distance[49] indicated 26 multivariate outliers. Analysis of these outliers showed that about one third of these outliers (9 cases) were caused by variable IV2 (*missing specialized terminology*), and 5 cases were caused by IV7 (*missing general terms*). The rest of the outliers (12 cases) had no variability among the translation event variables (all zero), which means that these queries vary on variables that were removed from the analysis. An additional 11 outliers were identified as cases where the prediction was more than three standard deviations from the regression model. The majority of these (6 cases) showed large differences in performance because the English target query failed to retrieve relevant documents. In three cases, none of the variables could predict the variability.
One case showed no variability on the translation event variables, and another exhibited problems with IV16 (*correct but different translation*).

Forty-four outliers that only appeared as univariate outliers were left in the analysis since query variability is the nature of information retrieval, and removing these outliers might not be warranted. All multivariate and model outliers were removed from the analysis to ensure they do not unduly influence the results (N= 691). However, although outliers were removed at the start of the regression analysis reported below, twenty-four new Mahalanobis outliers made their appearance, indicating the malformity of this dataset.[50]

[49] Statistic to identify multivariate outliers (see also glossary).
[50] A much stricter outlier removal (Mahalanobis criterion of $p > .01$) resulted in a significant $R$: $F(6, 652) = 15.33$, $p < 0.01$. $R = .353$, $R^2 = .125$, adjusted $R^2 = .116$.

| Variables | DV | DSD | IV2 | IV7 | IV13 | IV16 | IV17 | B | | β | $sr^2$ | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Intercept | | | | | | | | .715 | ** | | | |
| DSD | -.263 | | | | | | | -.018 | | -.108 | | |
| IV2 | -.141 | -.184 | | | | | | -.050 | ** | -.155 | .020 | |
| IV7 | -.136 | -.022 | -.063 | | | | | -.073 | ** | -.114 | .013 | |
| IV13 | -.272 | .505 | -.098 | .053 | | | | -.010 | * | -.103 | .006 | |
| IV16 | -.071 | -.277 | -.012 | .011 | -.301 | | | -.004 | | -.030 | | |
| IV17 | .435 | -.463 | -.076 | -.092 | -.430 | -.331 | | .040 | ** | .309 | .038 | |
| | | | | | | | | | | | .077 | |
| | | | | | | | | | | $R^2$ | .230 | a |
| Means | .704 | 1.056 | .065 | .033 | 1.33 | 1.035 | .762 | | | **Adj. $R^2$** | .223 | |
| St.dev. | .070 | .423 | .216 | .109 | .713 | .519 | .534 | | | $R$ | .480 | * |

**$p$ <.01 *$p$ <.05

a Unique variability = .077 Shared variability = .153

**Table 4.5 Multiple regression results.**

Correlations between the variables as well as unstandardized regression coefficients, semi-partial correlations, $R^2$, and (adjusted) $R^2$ are listed in Table 4.5. $R$ for regression was significantly different from zero, $F(6, 684) = 34.07$, $p < 0.01$. For the regression coefficients that differed significantly from zero, 95% confidence intervals were calculated. The confidence limits for (the arcsine of) *missing specialized vocabulary* (IV2) were -.073 to -.027, those for (the arcsine of) *missing general terms* (IV7) were -.116 to -.030, those for (the arcsine of) *wrong translation due to ambiguity* (IV13) were -.019 to -.002, and those for (the arcsine of) *correct identical translation* (IV17) were .027 to .054.

Four of the six independent variables contributed significantly to the prediction of retrieval performance differences when comparing monolingual to cross-lingual information retrieval: *missing specialized vocabulary* (IV2), *missing general terms* (IV7), *wrong translation due to ambiguity* (IV13), and *correct identical translation* (IV17). The largest unique contributions to this, admittedly small, $R^2$ difference come from (the arcsine of) *missing specialized vocabulary* and *correct identical translation* (.020 and .038 respectively). Half of the explained variability in the dependent variable is explained by the independent variables in combination. Together 23% (22.3% adjusted) of the variability in (the square root of) unstandardized residual average precision

was predicted by knowing the scores for these five independent variables. A logistic regression analysis provided similar results although the set of variables that contributed significantly was slightly different (IV13 and IV17 only - see section 4.4.3).

As is obvious from these results, a large part of the variability in these results cannot be explained by the independent variables and suggests a poor fit of the model. Information retrieval research has long been stymied by large query variability and this study is no exception. For a detailed discussion of these findings the reader is referred to chapter 5.

### *4.4.3 Additional regression analyses*

As explained in section 4.4.1, we aggregated the 17 translation event independent variables into three groups: *correct*, *missing*, and *wrong*. A standard multiple regression analysis was carried out between unstandardized residual average precision as the dependent variable, and *correct*, *missing*, *wrong*, and *Dutch sense density* as independent variables.
The new translation event independent variables were normalized by target query length, and underwent an arcsine transformation. *Dutch sense density* underwent a log transformation after a constant of 0.5 was added to each variable, and the dependent variable underwent a square root transformation (plus 0.5). Mahalanobis ($p > 0.01$) and model outliers were removed before the analysis leaving N=699.

$R$ for regression was significantly different from zero, $F(4, 694) = 33.66$, $p < 0.01$. The values for $R$, $R^2$, and (adjusted) $R^2$ were .403, .162, and .158 respectively. From these results it appears that aggregating the translation variables into the three groups did not result in a better understanding of the impact of translation on information retrieval performance than only using subset of the 17 translation variables.

Regression analyses were also carried out for title and description queries separately. As pointed out previously, these two query types are supposed to be different in nature, title queries typically being much shorter and less specific. As a starting point for both analyses we used the main analysis dataset of N=691 dataset (see section 4.4.2) after which it was divided into title queries and description queries: $N_{title}=312$, and $N_{description}=379$. $R_{title}$ for regression was significantly different from zero, $F(6, 305) = 19.75$, $p < 0.01$. The values for $R$, $R^2$, and (adjusted) $R^2$ were .529, .280, and .266 respectively. Interestingly, unlike in the main analysis, *missing general terms* (IV7) did not contribute significantly. $R_{description}$ for regression was significantly different from zero, $F(6, 372) = 7.184$, $p < 0.01$. The values for $R$, $R^2$, and (adjusted) $R^2$ were .322, .104, and .089 respectively. Also, unlike in the main analysis, *wrong translation due to ambiguity* (IV13) did not contribute significantly. Judging from the differences in $R^2$, the independent variables are better in predicting performance difference for title queries than for description queries.

A logistic regression analysis, which is less sensitive to distributional issues, was carried out with the same 6 variables and identical transformations as in the multiple regression analysis above. The dependent variable was dichotomized into queries with positive difference scores (monolingual better than cross-lingual), and those with zero or negative difference scores (identical performance or better cross-lingual than monolingual). The full model with all six predictors against a constant-only model was statistically reliable, $\chi^2$ (6, N=728) = 93.18, I < .001. The variance accounted for in retrieval performance difference is similar to that in the multiple regression analysis with Nagelkerke $R^2$ = .205. The Wald chi-square test indicated that only two independent variables showed a significant contribution: *wrong translation due to ambiguity* (IV13), and *correct identical translation* (IV17).

## 4.5 Query analysis

In addition to the statistical analysis, a traditional query analysis was carried out to better understand the queries and the query translation process. The analysis used the difference score *D* to divide the query set (N=728) into twelve groups (see frequency table 4.6), and analyze queries from each of the twelve groups. A minimum of 10 queries were analyzed for each of the 12 classes.

| Class | frequency | % | cum. freq. | cum. % |
|---|---|---|---|---|
| >=100 | 78 | 10.71 | 78 | 10.71 |
| >=90<100 | 226 | 31.04 | 304 | 41.76 |
| >=80<90 | 74 | 10.16 | 378 | 51.92 |
| >=70<80 | 52 | 7.14 | 430 | 59.07 |
| >=60<70 | 44 | 6.04 | 474 | 65.11 |
| >=50<60 | 47 | 6.46 | 521 | 71.57 |
| >=40<50 | 31 | 4.26 | 552 | 75.82 |
| >=30<40 | 25 | 3.43 | 577 | 79.26 |
| >=20<30 | 19 | 2.61 | 596 | 81.87 |
| >=10<20 | 15 | 2.06 | 611 | 83.93 |
| >=0<10[51] | 40 | 5.49 | 651 | 89.42 |
| <0 | 77 | 10.58 | 728 | 100 |
| | 728 | 100 | | |

**Table 4.6 Differences in average precision.**

### *4.5.1 Class query analysis*

The division of queries into twelve distinct performance difference classes might give the impression that queries can be easily classified, based on certain query characteristics. However, as pointed out by the statistical analysis in section 4.4.2, this is certainly not so. We cannot say that

[51] 26 queries in this class showed no difference in retrieval performance.

queries that lose an important term in the translation will do badly in retrieval and will end up in Class 1. Some queries still do well even though the majority of terms disappear in the translation. Conversely, some target queries that contain exactly the same terms as the source query plus one additional synonym do poorly in retrieval. While it is extremely difficult to predict retrieval performance based on the problems and successes achieved in the translation, the query analysis below provides some insight into the query translation process and the difference in monolingual and cross-lingual retrieval performance. See Table 4.7 for group results.

| | class1 | class2 | class 3 | class 4 | class 5 | class 6 | class 7 | class 8 | class 9 | class10 | class11 | class12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **ESlength** | 5.10 | 6.85 | 6.69 | 5.04 | 6.16 | 5.85 | 6.00 | 5.08 | 5.05 | 4.27 | 3.10 | 4.00 |
| **DSlength** | 4.72 | 6.96 | 6.80 | 4.85 | 6.32 | 6.04 | 5.84 | 5.28 | 4.53 | 4.33 | 2.78 | 4.04 |
| **ETlength** | 31.63 | 45.80 | 44.31 | 26.33 | 37.00 | 36.60 | 31.35 | 28.40 | 23.47 | 26.33 | 9.80 | 22.04 |
| **ESsnsdnsty** | 3.74 | 3.73 | 3.28 | 3.68 | 3.62 | 3.57 | 4.00 | 3.57 | 3.33 | 3.68 | 2.58 | 2.93 |
| **DSsnsdnsty** | 2.65 | 2.91 | 2.77 | 2.41 | 2.73 | 2.48 | 2.16 | 2.25 | 2.31 | 2.33 | 1.46 | 2.44 |
| **Expansion** | 4.47 | 5.33 | 5.01 | 3.92 | 4.61 | 4.47 | 3.54 | 3.77 | 3.70 | 4.62 | 1.08 | 3.87 |
| **Mono AVP** | 0.1245 | 0.2210 | 0.1970 | 0.2290 | 0.1615 | 0.2167 | 0.2986 | 0.3156 | 0.2497 | 0.2312 | 0.3475 | 0.1119 |
| **Cross AVP** | 0.0000 | 0.0051 | 0.0279 | 0.0573 | 0.0553 | 0.0962 | 0.1632 | 0.2004 | 0.1845 | 0.1937 | 0.3423 | 0.1599 |
| **Dutch AVP** | 0.0020 | 0.0476 | 0.0397 | 0.0481 | 0.0133 | 0.0360 | 0.0277 | 0.0954 | 0.0350 | 0.0760 | 0.1132 | 0.0346 |
| **Fair AVP** | 0.0092 | 0.0416 | 0.0720 | 0.0836 | 0.1111 | 0.1328 | 0.1933 | 0.1983 | 0.2254 | 0.1195 | 0.2377 | 0.1604 |
| **Good AVP** | 0.0169 | 0.0848 | 0.1359 | 0.1214 | 0.1266 | 0.1669 | 0.1787 | 0.1997 | 0.2465 | 0.1664 | 0.2104 | 0.1051 |
| **v1** | 0.00 | 0.00 | 0.00 | 0.02 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| **v2** | 0.45 | 0.22 | 0.04 | 0.10 | 0.09 | 0.06 | 0.00 | 0.00 | 0.11 | 0.00 | 0.05 | 0.06 |
| **v3** | 0.05 | 0.07 | 0.01 | 0.04 | 0.00 | 0.04 | 0.06 | 0.00 | 0.11 | 0.20 | 0.00 | 0.05 |
| **v4** | 0.03 | 0.00 | 0.01 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| **v5** | 0.14 | 0.18 | 0.14 | 0.15 | 0.64 | 0.21 | 0.10 | 0.60 | 0.11 | 0.13 | 0.50 | 0.06 |
| **v6** | 0.00 | 0.02 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| **v7** | 0.15 | 0.17 | 0.11 | 0.06 | 0.09 | 0.02 | 0.13 | 0.04 | 0.00 | 0.07 | 0.03 | 0.04 |
| **v8** | 0.06 | 0.04 | 0.00 | 0.06 | 0.00 | 0.02 | 0.06 | 0.00 | 0.00 | 0.00 | 0.00 | 0.01 |
| **v9** | 0.15 | 0.15 | 0.14 | 0.06 | 0.09 | 0.09 | 0.19 | 0.24 | 0.05 | 0.07 | 0.00 | 0.06 |
| **v10** | 0.00 | 0.00 | 0.00 | 0.12 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| **v11** | 0.35 | 0.61 | 1.24 | 0.19 | 0.34 | 0.49 | 0.26 | 0.32 | 0.21 | 0.13 | 0.05 | 0.22 |
| **v12** | 3.91 | 5.50 | 5.24 | 3.96 | 5.07 | 3.85 | 2.55 | 1.40 | 2.95 | 1.00 | 0.80 | 3.31 |
| **v13** | 17.71 | 24.96 | 23.68 | 12.69 | 18.07 | 18.81 | 16.19 | 16.60 | 11.11 | 15.80 | 4.10 | 9.18 |
| **v14** | 0.00 | 0.00 | 0.00 | 0.04 | 0.05 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| **v15** | 0.54 | 0.25 | 0.18 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.29 |
| **v16** | 6.12 | 10.04 | 9.45 | 5.90 | 8.77 | 9.55 | 8.13 | 5.96 | 6.00 | 6.20 | 2.65 | 6.35 |
| **v17** | 1.97 | 3.58 | 4.08 | 2.94 | 3.80 | 3.45 | 3.68 | 3.24 | 2.84 | 2.73 | 1.63 | 2.49 |
| **D** | 1.0000 | 0.9736 | 0.8537 | 0.7488 | 0.6534 | 0.5534 | 0.4531 | 0.3562 | 0.2634 | 0.1606 | 0.0153 | -3.0248 |

**Table 4.7 Class averages for dependent and independent variables.**

#### *4.5.1.1 Class I*

Class I contains queries in which monolingual performance is 100% better than cross-language performance. This group of queries is the second largest class of queries (10.7% of all queries) and all had a cross-lingual mean average precision score of zero. This means that the English target queries in general failed to retrieve a significant number of relevant documents.[52]

Analysis showed that for most short queries important keywords, or all keywords, are lost in the translation or that so many erroneous terms are added that the correct terms lose effectiveness. Of the 78 queries with a difference in average precision of 100%, 43 were title queries, which tend to be shorter in length, and 35 were description queries. Overall, most of these queries are short (60.25% of the Dutch source queries had 3 words or less). Shorter queries are more sensitive to term loss because when a 3-term query loses a term (i.e. term does not appear in dictionary) there is a content loss of 33%.

> Query 334T loses all terms in the translation:
> `EST: Export Controls Cryptography`
> `DST: Exportregels cryptografie`
> `ETT: exportregels, cryptografie`
>
> The terms *exportregels* and *cryptografie* did not appear in the dictionary. The English target query uses the Dutch source terms in that case. Since the Dutch terms are not cognates this query did not retrieve any relevant documents.
>
> Query 445T loses an important content term in the translation:
> `EST: women clergy`
> `DST: vrouwen in het ambt`
> `ETT: woman, female, lady, wife, spouse, queen, Mrs, Mistress, Madam, Madame, mistress, office, ministry`
>
> Since the term clergy is rather important in guiding this query, its loss proves to be momentous.

In a number of cases in this class, the English terms are represented in the Dutch source query as compounds. For example the 2-term phrase *arms export* is represented in the Dutch source query by *wapenexport*, and *food supplement* by *voedingssupplement*. We observe that the larger the number of terms that are collapsed into a compound, the more specific this compound becomes. The more specific a compound, the less likely it is to appear in a dictionary. When a compound

[52] A single relevant document retrieved at rank 1000 would not have a noticeable effect on average precision.

fails to get translated (compounds are unlikely to be cognates) a group of terms from the English source query, and much of it is lost.

#### *4.5.1.2 Class 2*

Class 2 contains queries in which the difference between monolingual and cross-lingual performance ranges from 90% to 100%. This is by far the largest class of queries (226 or 31.04%). The cross-lingual average precision scores in this class are all well below 0.1 (cross-lingual mean average precision score of 0.0051 compared to 0.2210 for monolingual average precision). Clearly, a lot goes wrong in this query class. The average query length for this class is much larger than for Class 1 (only 57 or 25.22% of the queries have three query terms or less). English source queries are on average 6.8 words long, and Dutch source queries 7.

The queries in Class 2 tend to expand with a large number of erroneous terms in the translation process. These terms create large target queries that retrieve non-relevant documents. On average, each query had about 25 erroneous terms added due to ambiguity alone (not counting other causes of error). For Class 1 the number of additional terms was 17. Another characteristic of Class 2 queries is that the query terms tend to be "weak" content-bearing terms. What this means is that these query words can appear in multiple contexts with many different, unrelated, terms. For example, although the topic area of query 288D survives the translation, 22 wrong translations are added and none of the correct terms can carry this query through.

> Query 288D has four "weak" content bearing terms:
> ESD: Weight control and diets in the U.S.
> DSD: Pogingen tot afvallen en diëten in de VS.
> ETD: attempt, try, bid, effort, endeavour, crack, go, shot, fall down, fall off, drop out, desert, abandon, defect, secede, lose weight, slim, waste, waste away, lose flesh, be left, be left over, be disappointing, not come up to one's expectations, not live up to one's expectations, be a disappointment, be a let-down, bear away, diet, regime, regimen, v, vs, US, USA

What appears to happen with queries in this class is that the additional erroneous terms cause the query's performance to drop near zero. Queries in this class exhibit loss of content bearing terms similar to the problems of Class1. Yet, Class 2 queries tend to be a little longer, perhaps retrieving a slightly larger number of relevant documents.

Overall, Class 2 queries would have performed better without any translation. The Dutch source query performed better than the English target query in retrieving English documents in 59 cases (44.09%). Using only the Dutch source query performs much better in an exceptional case (366D), outperforming the English source query.

Query 366D does better than its monolingual and cross-lingual versions:

```
ESD: What are the industrial or commercial uses of cyanide or its
derivatives?
DSD: Wat zijn de industriële of commerciële toepassingen van cyanide
of daarvan afgeleide producten?
ETD: commercial, use, employment, utilization, application,
adoption, practice, administration, implementation, enforcement,
cyanide, prussiate, de, product, production, commodity, exhibit
```

#### *4.5.1.3 Class 3*

Class 3 contains queries in which the difference between monolingual and cross-lingual performance ranges from 80% to 90%. Class 3 is the fourth largest class of queries with 74 (10.16%) queries. The cross-lingual average precision scores for this class are all still below 0.1 (except for two queries). The cross-lingual mean average precision score is 0.0279 (0.1970 monolingual).

The Class 3 queries are about the same length as the queries in the previous class but have, on average, fewer bad terms and a larger number of correct terms are added in the translation. Class 3 queries appear to be translated more or less correctly but are neutralized by large numbers of badly translated terms. When Class 3 target queries are rerun after removing all the bad terms, the impact of these noisy terms on performance becomes clear. With the clean target queries, mean average precision rises to 0.1359 which is very close to monolingual performance.

The translation of query 088T has the same terms as the English source query but also a large number of additional terms that lower cross-lingual performance:

```
EST: Crude Oil Price Trends
DST: Trends in de prijs van ruwe olie
ETT: trend, tendency, rage, fashion, price, fare, charge, price tag,
price ticket, prize, award, trophy, sports trophy, reward, prize,
rough diamond, rough, coarse, rugged, chunky, raw, crude, rough-
hewn, broad, slapdash, abrasive, harsh, rude, boisterous, unruly,
rough, wire-haired, rough-haired, oil, oil shares, oil stock
```

#### *4.5.1.4 Class 4*

Class 4 contains queries in which the difference between monolingual and cross-lingual performance ranges from 70% to 80%. Queries in Class 4 (52 queries, 7.14%) tend to be somewhat shorter, on average, than queries in previous classes. Half of the Dutch source queries contain three words or less. Cross-lingual mean average precision is 0.0573 and mean monolingual average precision is 0.2290. Compared to previous groups, cross-lingual performance is higher compared to previous classes, but very low.

Class 4 appears to have a large number of short queries with the addition of bad terms, which cause a drop in performance. This is true even when the number of additional bad terms is low. In many cases, the bad translations, though incorrect, were in the same subject area as the source query term. For example *scholen* (schools) is translated into *teach*, *instruct*, *train*, *drill*, and *education*. Judging by the performance difference, wrong translations in the same subject area might hinder performance even more than other bad translations. Even translations that might be considered acceptable can cause a performance drop as is illustrated by query 075T.

> Query 075T has the English source term in the translation but the term computerization causes a drop in performance:
> `EST: Automation`
> `DST: Automatisering`
> `ETT: automation, computerization`

#### *4.5.1.5 Class 5*

Class 5 contains queries in which the difference between monolingual and cross-lingual performance ranges from 60% to 70%. Class 5 contains 44 queries (6.04%). Compared to the previous class, monolingual mean average precision has dropped (0.1615), on average, while cross-lingual mean average precision remained about the same (0.0553).

Queries in Class 5 do not distinguish themselves clearly from other classes (beyond difference score *D*). Most of the terms of the analyzed queries survived the translation, but large numbers of erroneous terms were also added.

#### *4.5.1.6 Class 6*

Class 6 contains queries in which the difference between monolingual and cross-lingual performance ranges from 50% to 60%. Class 6 contains 47 queries (6.46%). The source and target queries of Class 6 are slightly longer than previous classes (similar to those in Class 2). Cross-lingual mean average precision is 0.0962 and monolingual mean average precision is 0.2167. Mean cross-lingual average precision for this class was a little bit higher than the previous classes.

Several queries in Class 6 were translated correctly but had some topically related translations added that caused a large drop in cross-lingual performance.

> Target query 431T looks reasonable but does not do well:
> `EST: robotic technology`
> `DST: robot technologie`
> `ETT: robot, automaton, technology, applied science`

This query is essentially the same because *robotic* gets stemmed to *robot* and the other query term (*technology*) is the same. Yet, the addition of *automaton* and *applied science*, fairly reasonable expansions, to the target query, caused a 54% drop in performance.

Again, a large number of erroneous terms were added to the queries in this class. However, crucial "stronger" query terms (e.g. *koolmonoxidevergiftiging* translated into *carbon monoxide poisoning*) often made it through the translation and probably were the main reason performance did not fall more.

#### *4.5.1.7 Class 7*

Class 7 contains queries in which the difference between monolingual and cross-lingual performance ranges from 40% to 50%. Class 7 contains 31 queries (4.26%). Mean average precision is improved compared to previous classes (cross-lingual=0.1632 and monolingual=0.2986). Cross-lingual mean average precision exhibited the greatest improvement - above 1.0 for the first time. The queries in Class 7 do not stand out from the other classes per se but have, on average, fewer erroneous terms added. Perhaps these queries are "easier" than the ones in previous classes.

#### *4.5.1.8 Class 8*

Class 8 contains queries in which monolingual performance is more than 30% (less than 40%) better than cross-lingual performance. Class 8 contains 25 queries (3.43%). Mean average precision is higher compared to previous classes (cross-lingual=0.2004 and monolingual=0.3156). Class 8 has the second highest mean average precision for both monolingual as well as cross-lingual performance of all classes. As with the queries in Class 7, there are no clear distinguishing features for this class.

#### *4.5.1.9 Class 9*

Class 9 contains queries in which the difference between monolingual and cross-lingual performance ranges from 20% to 30%. Class 9 contains 19 queries (2.6%). Mean average precision is about the same as the previous class for cross-lingual (0.1845) with a drop in monolingual performance (0.2497). A smaller number of erroneous terms were added to each query The average Dutch source query length was a little bit shorter. These queries appear to be more robust in regard to erroneous translations.

> Query 220D has strong terms that counterbalance bad terms:
> ESD: How do crossword puzzle makers go about making their puzzles?
> DSD: Hoe gaan de makers van kruiswoordpuzzels te werk?
> ETD: go, move travel, leave depart, be off, be going to, run, come,

```
be, get, walk, go around go about, be all right, work, fit, be in
charge, be in charge of, be about, be, go, happen, be about,
producer, architect, artist, author, Maker, Creator, (m, crossword,
crossword puzzle, work, job, employment, site, chore, duties, task,
action, deed, works, movement, oakum
```

#### *4.5.1.10 Class 10*

Class 10 contains queries in which the difference between monolingual and cross-lingual performance ranges from 10% to 20%. Class 10 is the smallest class of queries and contains only 15 queries (2.06%). Cross-lingual mean average precision is 0.1937 and monolingual mean average precision is 0.2312.

The majority of Class 10 queries are the so-called title queries (73.33%). Title queries tend to be shorter but also more to the point than description queries. Whereas description queries may contain a lot of additional terminology to elucidate the query, title queries only contain fundamental keywords. Interestingly, monolingual performance is lower compared to previous classes, perhaps explaining the better *D* scores in this class. Almost half the queries (46.67%) have a monolingual performance below 0.1.

Query 386T does badly in both cases (EST=0.0268, ETT=0.0233):

```
EST: teaching disabled children
DST: onderwijs aan gehandicapte kinderen
ETT: education, teaching, instruction, the field of education,
Education, about, around, away, on, against, on to, handicapped,
disabled, invalid, child, baby, infant, girl, thing, lass, dear
fellow, dear girl
```

#### *4.5.1.11 Class 11*

Class 11 contains queries in which the difference between monolingual and cross-lingual performance ranges from 0% to 10%, meaning that monolingual performance is slightly better or identical to cross-lingual performance. Class 11 contains 40 queries (5.49%).

As expected, mean average precision for cross-lingual retrieval (0.3423) is almost identical to that of monolingual retrieval (0.3475) for this class. Retrieval performance is also the highest of all previous classes. The majority (77.5%) of the queries have three words or less. Title queries (85%) outnumber description queries. However, not all queries did well. As in the previous class, there were a number of queries in this class (25%) that had poor monolingual performance (less than 0.1). These short title queries in this class tend to be less ambiguous than queries we have seen in previous classes. In most cases this facilitates good retrieval performance and translation.

The average expansion in this class is close to 1, which means that target queries were roughly twice the size of Dutch source queries. In no other class was average expansion this low.

Query 134T crosses the translation barrier with two additional terms:

```
EST: The Human Genome Project
DST: Het Human Genome Project
ETT: human, genome, project, scheme, plan
```

#### *4.5.1.12 Class 12*

Class 12 contains queries in which cross-lingual performance is better than monolingual performance, in varying degrees. Class 12 contains 77 queries (10.6%). Compared to other classes, monolingual performance is the lowest with a mean average precision of 0.1119 compared to that of 0.1599 cross-lingual. More than half the monolingual queries in this class (55.8%) had an average precision of below 0.1. The cross-lingual counterparts of these queries didn't fare much better, with a few exceptions where the vocabulary improved in translation as happened in query 352T.

Query 352T shows the different target query vocabulary for Chunnel:

```
EST: British Chunnel impact
DST: Effecten van de Britse Kanaaltunnel
ETT: effect, result, outcome, consequence, spin, stuff, twist,
curve, slice, side, power, stock, share, security, Briton, the
British, Brit, Britisher, Channel tunnel, tunnel under the Channel
```

When we look at queries that do well monolingually but do even better in translation, we see that the additional terms (correct translations) can help performance. For example, the additional terms of query 279T (*electromagnetic*, and *geographical pole*) must have boosted the English target query performance beyond that of the English source query. Note that the proper name *Poland*, as a translation of *polen* does not seem to influence query performance. The eastern European country is not likely to be associated with magnetic pole shifts (although you'd better check your maps of Poland for the correct declination if you plan to use your compass over there).

Query 297T additional terms boost the target query's performance:

```
EST: Earth magnetic pole shifting
DST: Verplaatsing magnetische polen
ETT: moving, movement, removal, transfer, transference, shifting,
transposition, displacement, move, relocation, permutation,
magnetic, electromagnetic, Poland, pool, pole, geographical pole,
pile
```

For some queries it is simply a matter of losing a disturbing term in the translation as happened in query 429D. Here the term *outbreaks* apparently hinders performance, since the target query does much better without it.

Query 429D is a case where a translation problem actually helped:

```
ESD: Outbreaks of Legionnaires' disease.
DSD: Epidemieën van de legionairsziekte.
ETD: Legionnaire's disease
```

*4.5.1.13 Query analysis reprise*

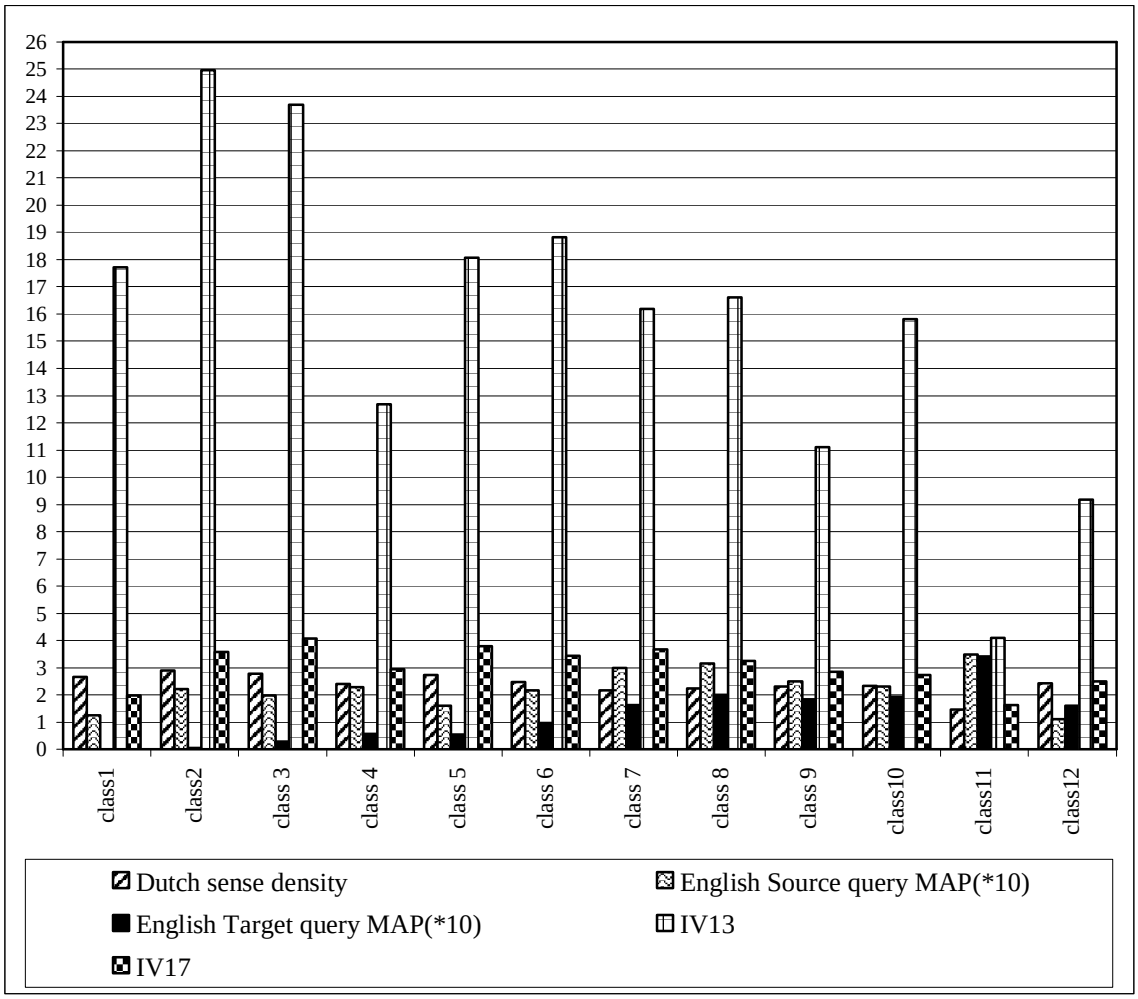


**Figure 4.7 Selective variables for query analysis classes. (Note that the Y-axis has three scales: 1) average number of translation events per query; 2) average Dutch sense density per query, and; 3) mean average precision multiplied by 10 (MAP * 10) for the English queries.)**

After looking at the different query classes it is evident that, although there are certain query characteristics that we intuitively link to query performance, none of these characteristics can be said to have a singular outcome. For example, when we look at the impact of erroneous

translations that are added to the query (IV13), Figure 4.7 shows that Class 4 has fewer of these terms than class 5, 6, 7, and 8. Yet, queries in Class 4 do worse when compared to their monolingual counterparts than in those other classes. The average number of correct terms in Class 4 is about one term less than in the subsequent four classes, which might explain the difference. As is also substantiated by the statistical analysis, there is no single translation event or other independent variable that can predict query performance but rather a combination of different variables, making the connection between the translation events in the queries and retrieval performance rather complex.

## 4.6 Summary

This chapter presented the results of the study of the impact of translation events on retrieval performance. The results were based on content analysis and literature review, a statistical analysis, and a query analysis.

In answer to research question 1, concerning the kinds of events affecting cross-language retrieval, the study presented the translation event taxonomy created through content analysis and a literature review. Although the taxonomy is very descriptive with regard to possible translation events (43), only a limited number of these events actually occurred in the study's queries (17).

In answer to research question 2 concerning the impact of translation on retrieval performance, the results of information retrieval experiments and a query analysis were presented. A statistical analysis indicated that only four of the 17 translation events contributed significantly to the performance differences between monolingual and cross-lingual performance: *missing specialized vocabulary* (IV2), *missing general terms* (IV7), *wrong translation due to ambiguity* (IV13), *correct identical translation* (IV17). However, the variance explained by these variables was modest (adjusted $R^2 = .223$). An additional analysis showed that translation event variables do better at predicting the performance difference for title queries as compared to description queries. A query analysis confirmed that the connection between translation events and retrieval performance is rather complex because the presence or absence of these events does not have a unidirectional effect. Chapter 5 provides a discussion of the results and implications of the findings of this study.

# 5 Discussion and Study Implications

## 5.1 Introduction

This study investigated translation events and their impact on retrieval performance. To this effect, a detailed taxonomy was developed which incorporated possible translation events from a query analysis and a literature review. A large number of queries (728) were coded with this translation event taxonomy to indicate what events took place during their translation from Dutch into English. The translated queries (English target queries) represent cross-lingual information retrieval, while the English source queries represent the monolingual case. The monolingual and cross-lingual versions of each query were used in a retrieval experiment to acquire retrieval performance scores for both. The impact of the translation events on retrieval performance was studied using a statistical analysis and a query analysis. As was stated in the previous chapter, the variation in performance between monolingual and cross-lingual retrieval can only be explained in small part by the events that occurred in the translation. In this chapter, these findings will be discussed both in terms of related information retrieval research, and their theoretical, methodological, and practical implications.

Section 5.2 of this chapter examines the two research questions and the main results of this study. The next two sections (5.3 and 5.4) discuss related research which contributes to explaining the study's findings. Section 5.5 presents additional findings. Section 5.6 discusses future research. The chapter ends with a conclusion (section 5.7) and study summary (section 5.8).

## 5.2 Answering the two research questions

This study sought to answer the following two research questions:

1) *What kinds of translation events affect cross-language retrieval?*
2) *In what way does the presence of certain translation events in query translation affect retrieval performance?*

The answer to the first research question lies in the translation event taxonomy (see section 4.2). The translation event taxonomy presented 43 possible translation events ranging from missing terms to terms without any problems. As pointed out in section 4.2, the translation event taxonomy included three main translation event categories: 1) events concerning the lexical resource, 2) events concerning the source word, and 3) events concerning terms that remain un-translated. Each main category was divided into more specific subcategories, each represented by a specific code. A query term translation received a code for each of the three main categories and a translation

correctness assessment. The combination of all four codes resulted in the final unique notation for a specific translation event.

In answer to the second research question we found that four of the six independent variables contributed significantly to prediction of retrieval performance differences when comparing monolingual to cross-lingual information retrieval: *missing specialized vocabulary* (IV2), *missing general terms* (IV7), *wrong translation due to ambiguity* (IV13), *and correct identical translation* (IV17). Although the contribution of each of these variables is significant, their contribution to the total variance of the independent variable is small (adjusted $R^2$ = .223). The independent variables were slightly more successful in predicting performance differences between title queries than description queries. Yet, most of the variability remains unexplained.

Given these results (presented in chapter 4), the question arises whether we can safely conclude that translation events have only a small impact on information retrieval performance ? Are these results expected? The next two sections examine possible explanations of the results.

## 5.3 The translation event taxonomy and query codes

### *5.3.1 Taxonomy*

The limited role of translation events in predicting retrieval performance might be due to certain shortcomings in the translation event taxonomy. However, it is not clear at this point what these shortcomings might be. The scope of the taxonomy is CLIR translation events. The taxonomy appears to be exhaustive enough to cover this domain. It incorporates the three instances where translation events occur: at the initial term lookup in the translation resource, at the term level itself (semantics), and at matching time. A value judgment about the term's translation correctness is also part of the taxonomy. The taxonomy has thus incorporated all known aspects about a term's translation. The level of granularity of the taxonomy is also quite fine: translation events in the query data only used about 40% of the possible event translation codes.

Might there be other aspects of the translation that need to be captured? This question is difficult to answer. Aggregating the translation event codes (see section 4.1) into three categories (*correct*, *missing* and *wrong*) was an attempt to capture the ultimate effect of translation on retrieval performance. After all, the reason for an erroneous translation (i.e. ambiguity in the source term) might not matter for retrieval. Just the fact that a wrong term has been added to the query is of importance. The adjusted $R^2$ dropped to .158, showing that even less of the variability in the difference score can be explained by using the aggregated variables.

### *5.3.2 Query codes*

The way the queries were coded might not have been expressive enough. By assigning each translated term a translation event code it becomes clear whether the term is correctly translated or not, and if not, whether the origin of the error lies in the translation resource, or in the semantic properties of the term itself. In case a term is missing, the target query uses the original source query term instead of a translation. In this case the translation event code expresses whether the source term provides a problem or whether it is a cognate. For example, in query 405T the term *Hussein* is not in the dictionary so it is added to the target query as-is. In this case this strategy works fine since *Hussein* is a cognate.

After a query is coded we know quite a number of things about the target query term. However, as became clear in the query analysis (section 4.5), not each query term is created equal. It is precisely this information that is not part of the translation event taxonomy and is therefore also missing from the query codes. As was noted in Class 2 for example, the target queries often had "weak" content bearing terms – terms that may appear in multiple subject areas. These terms might have been one of the reasons for the huge performance drop in that class. A certain translation event, e.g. a missing terminology, is likely to have a different effect on retrieval performance for a weak term than for a strong term. Conversely, the addition of a certain (erroneous) term might not matter to the retrieval outcome but other additions could seriously damage the query.

### *5.3.3 Relative term importance*

IR researchers have been trying to determine the relative importance of terms for many years, starting with the research on automatic indexing by Bookstein and Swanson (1974). In more recent research, Pirkola and Järvelin (2001) tried to identify the most important query terms and their impact on retrieval performance. While they were reasonably successful at predicting the best query term (term with most impact on retrieval performance), they could not distinguish between the good terms and the bad terms in advance. Pirkola and Järvelin also point out that while users may be able to recognize which search term is "logically" most important in a query, these terms are not necessarily the most valuable terms in regard to retrieval performance. This important observation is relevant for this study because it means that the connection between a translation event and retrieval performance might be more tenuous than originally assumed. A good term translation does not necessarily mean this term is going to contribute to successful retrieval, nor does a bad translation automatically result in a drop in retrieval, it might simply have little or no effect at all.

## 5.4 Query variability

A related explanation of the limited effect of translation events on retrieval performance difference is query variability. The phenomenon where the smallest possible change in a query (the addition or removal of a single query term) causes a discontinuous difference in retrieval performance is called query variability. We had hoped to reduce the effect of query variability by using a large sample size (see section 3.3.3), but it appears that in spite of the large sample size, query variability is responsible for most of the variability in the dependent variable.

Query variability has eluded the information retrieval community for quite some time, and is also evident in this study. System performance is usually represented by an evaluation average which hides the large variation among individual queries. However, good TREC systems hardly ever perform well on all queries, and even bad systems often do very well on a subset of queries (Buckley and Walz, 2000). As IR system developers know, retrieval improvements never work well across all queries. Different queries, or even different formulations of the same query can be highly variable in their retrieval performance.

### *5.4.1 Query variability in the current study*

The query analysis of this study (section 4.5) showed numerous examples of the high query variability that might be confounding the results. Some queries still performed fine (when compared to their monolingual counterparts) after most of their terms got lost in the translation, while other queries performed poorly after losing only a single term. As we speculated, some query terms are "stronger" than other terms and the loss of one term is not the same as the loss of another term. For example, when we compare the English source query 302T to the English target query (correct translations only), we see that even though the source terms *postpolio* and *trauma* are not present in the target translation, performance of the target query is slightly better than the source query itself.

```
EST: Poliomyelitis and postpolio trauma (average precision 0.1037)
DST: Poliomyelitis en postpoliomyelitissyndroom
ETT_icorr: poliomyelitis (average precision 0.1532)
```

Naturally, the presence or absence of other terms in the query affects how seriously the loss of a term impacts retrieval performance. For example, none of the English source terms of query 314T appear in the target query, resulting in a complete drop in retrieval performance.

```
EST: Three Gorges Project (average precision 0.3244)
DST: Drieklovenproject
ETT: drieklovenproject (average precision 0.0000)
```

When we study the addition of erroneous terms, we again see the effect of query variability. While some queries experience a large performance drop after the addition of a single erroneous term, others still do reasonably well or better after a large number of erroneous terms have been added.

```
EST: Impact of Government Regulated Grain Farming on International
Relations (average precision 0.1586)
DST: Invloed van overheidssteun aan graanboeren op de internationale
betrekkingen
ETT: influence, effect, domination, impact, weight, authority,
induction, government support, government assistance, grain farmer,
grain grower, corn farmer, corn grower, International,
international, post, job, position, office, situation, relation,
relationship, connection, reference, bearing, relative, kinsman,
kinswoman, purchase, obtaining, buying, derivation, recruitment,
moving in, occupation (average precision 0.2156)
```

### *5.4.2 Related research*

The TREC Query Track (1998-2000) was started in an attempt to understand query variability. By running many variations of the same query on different systems researchers hoped to discover, among other things, the source of performance differences. Buckley and Walz (2000) tentatively note that for most of their test queries, the variability caused by the query was typically larger than the variability caused by the type of system running the query.[53] The same authors refer to another analysis of their queries where long queries were found to be more variable than shorter queries, the latter performing much better, on average. Long query versions frequently performed best but also often did worst. The researchers caution however that the query-length factor may actually be a query "originator" factor because most of the longer queries were created by students rather than experts (evenly split for short queries). Query length did appear to play a role in this study as well. The class with the smallest differences between monolingual and cross-lingual performance (Class 11) had noticeably shorter queries than any other class. Similarly, classes with major performance differences had, on average, very long target queries. The TREC Query Track ended only two years after its inception, leaving behind a large dataset and the hope that future information retrieval researchers might provide insight into the issue of query variability.

Why can slight variations in queries cause large variations in query performance? One of the reasons is the considerable idiosyncrasy between a user's information need and the representation of that need in a user's query. Typically, an information need originates in an inadequate state of knowledge, resulting in poorly formulated queries. (Belkin, Oddy and Brooks 1982; Ingwersen 1992). Translating a complex information need into a (simplified) query representation causes a

[53] Note there is no system variability in this study since the system does not vary between runs.

further divergence between an information need and a query. Although users might not be able to state exactly what they are looking for yet, they know whether or not a retrieved document is relevant. System evaluation measures are based on relevance judgments. Thus, retrieval systems are evaluated on their ability to retrieve relevant documents based on imperfect information (queries). In this light it is not surprising that different variations of a query may cause fluctuations in the retrieval performance score.

Perhaps the biggest source of query variability however can be found in the term distribution of the document collection. Pirkola and Järvelin (2001) identify the best query term (from a retrieval perspective) based on the document collection and the statistical properties of the terms in that collection (see also section 5.3). In related work Cronen-Townsend et al. (2002), the document collection is utilized to measure query ambiguity and to predict query performance. By comparing the language usage of the query to that of the entire document collection these researchers create a clarity score for a query. Queries with low clarity scores tend to retrieve documents from several different topic areas. As a consequence, the retrieved document set likely contains more irrelevant documents in the higher ranks than might otherwise be the case. Cronen-Townsend et al. found that the clarity score correlates well with average precision measurements while it does not correlate with traditional term weighting measures such as *idf* (see section 1.2).

As is clear from the previous sections, query variability is the main reason why a large part of the variability in the difference between monolingual and cross-lingual retrieval cannot be explained by the independent variables in this study. Properties of the underlying document collection seem to dictate whether a good query term translation also means good retrieval performance or whether the addition of a erroneous translation harms the query or not.

## 5.5 Additional findings

Bearing query variability in mind, what are additional findings beyond the ones presented in section 5.2?

### *5.5.1 Retrieval performance differences*

The study found that there is a significant difference between monolingual and cross-lingual information retrieval performance. The majority of the cross-lingual queries performed below their monolingual counterparts.[54] More than half (51.92%) of the queries had a performance difference between monolingual and cross-lingual retrieval of more than 80%. The performance difference was negligible for only 40 queries (5.49%). For a small number of queries (77 or 10.58%) cross-lingual retrieval outperforms monolingual retrieval. These queries tend to be shorter and less

[54] This is to be expected since the study used a baseline query translation approach.

ambiguous than most queries in this study. By far the largest number of queries (226 or 31.04%) had a performance difference between 90-99%. These findings indicate that the cross-lingual retrieval baseline, where all possible translations are added to the target query without any means of word sense disambiguation, has a lot of room for improvement. Not surprising, most CLIR research revolves around reducing translation ambiguity or improving the lexical translation resource. When we compare the target query run with the ET-allcorr variation (target queries with correct translations only) however (see Table 4.3), it appears that more than disambiguation alone is needed. Even after only including correct translations in the ET-allcorr queries, their performance is very similar to the English target queries.

### *5.5.2 Query expansion and query length*

The fact that the target query run and the ET-allcorr run are very close, also has implications for query expansion. When we only add correct translations to the target query, we are in fact expanding the query because all translations (including synonyms) of a term are added. For example, in query 277T we are expanding *landmijnen* (land mines) with *land mine*, *claymore*, *claymore mine*, *fougasse*. In this study, the use of resource-driven expansion (i.e. synonym lookup in dictionary) does not guarantee an increase in retrieval performance.

There is a generally held belief in information retrieval that longer queries do better in information retrieval because of their self-disambiguating properties. The same does not appear to be true for target queries, especially when the large number of erroneous terms may steer a query in a different direction. The query class with the smallest performance differences between monolingual and cross-lingual retrieval performance contains queries that are much shorter than in any other performance difference class (see section 4.5.1.11). Also, recall that the correlation between performance difference and the independent variables is larger for title queries (typically shorter than description queries) than for description queries, indicating perhaps that description queries are more variable.

### *5.5.3 Term importance*

As pointed out previously, not all query terms are created equal. There appears to be a difference in query terms with regards to their importance to a query. It would help CLIR, as well as monolingual retrieval, enormously if it were possible to assess the importance of a query term (translation). This knowledge could guide the translation and creation of the target query by being more selective about what words are added to the target query.

### *5.5.4 Translation events*

Overall, performance loss appears to be caused by (a large number of) erroneous terms or the omission of important query terms. This finding is supported by the query analysis as well as the

logistic regression analysis (section 4.4.3). Conversely, the presence of correct translations has a positive effect on performance. However, the presence or absence of these conditions does not automatically mean trouble or success. Erroneous translations in the same subject area as that of the query appear to do more damage than other erroneous translations. The main goal of word-sense disambiguation should probably be to cut down on the number of translations that are added to a target query.

### *5.5.5 Phrases and compounds*

When examining the translation events in cross-language query translation in more detail, it appears that more than half of the erroneous term translations in a large query set are caused by ambiguity. The other main cause of translation error is the word-by-word translation of terms that are part of multiple-word expressions. About one third of the terms are translated correctly, some with terms identical to the English source queries, and others with synonyms. The compounds in the Dutch source queries represent multiple terms in the English source query. The combination of multiple terms in a single term reduces ambiguity (similar to pre-coordination in indexing), and compounds are often important query terms. The more specific the compound however, the less likely it is to appear in a translation lexicon. These observations about the translation results indicates the importance of a phrasal lexicon for CLIR.

## 5.6 Future research

### *5.6.1 Test collection and system changes*

Since the current study used an existing query set, it was unclear beforehand what translation events would occur and in what combinations they might manifest themselves. To answer some of the questions left unresolved, an artificial query set might be created. To allow a more thorough study of translation events, specific queries and query variations could be built. Each query version would have different events or event combinations. This study would study queries on a much smaller scale than the current study.

The current study was designed to get as clear a picture of the translation process as possible. To accomplish this, no commonly used retrieval or natural language processing techniques such as automatic relevance feedback, nor word-sense disambiguation were used. Also the word-by-word translation was rather simplistic. As a result, the approach to query translation can be viewed as a baseline where all term translations are added to the target query. An obvious extension to this research is to allow disambiguation and other techniques to arrive at the best query translation possible. In a more advanced approach, the set of query term translations should be treated as a single concept rather than a group of individual terms. This would require a structured query technique and would have implications for the way the translation event codes would be assigned.

Since the CLEF collection added a Dutch collection and now has a larger number of queries, it would be interesting to replicate this study using a different test collection, and to try multidirectional CLIR, not just Dutch-English but also English-Dutch.

### *5.6.2 Dealing with query variability*

These avenues of future research would all run into query variability issues as did this study. A different approach might be needed to tease out the effect of translation events on retrieval performance (see sections 5.3 and 5.4). Although this study's findings indicated that translation events had a significant effect on the difference between monolingual and cross-lingual retrieval, the presence of query variability may have clouded this effect. Future research should attempt to quantify this variability. This study showed that, as an effect of query variability, some query terms are more important from an information retrieval perspective than others. We called these terms "strong" query terms. It might be possible to quantify the strength of a term by using a measure akin to the clarity score of Cronen-Townsend et al. (2002), or the best term identification method of Pirkola and Järvelin (2001). This additional information about the query terms might be incorporated in an extended version of the translation event taxonomy.

## 5.7 Conclusion

The problems and successes in a translation are relatively independent of problems of information retrieval. Due to query variability, the query translation process and its impact on retrieval performance is chaotic. The smallest possible change you can make in a query (to add or remove a single query term) may cause a difference in results that is discontinuous. For example, take a 15 word target query and its 7 word source query. Imagine removing query terms from the target query, one at a time, until you are left with the same exact 7 terms as the source query. At this last point, the performance difference between source and target query will be zero because these queries are identical. However, because of the chaotic nature of retrieval, the performance difference between the previous eight query variations is going to be highly irregular and vary widely. The performance of the 12-term query may be much better than the 15-term version and the 8-term version may be much worse than the 15-term version. This chaotic behavior makes research in this area extremely challenging.

Even though query variability proved to be a confounding factor in the study, we still learned that query performance tends to suffer from missing specialized vocabulary, missing general terms, and wrong translations, whereas performance benefited from the correct identical translations. To improve the translation quality, CLIR system designers should work on word-sense disambiguation and the creation of specialized (term and phrase) lexicons. This will ensure that fewer erroneous terms are added to a target query while at the same time more correct terms are added. While this

is not a sure recipe for retrieval success, based on the results of this study, this is the best we can do at the moment.

## 5.8 Summary

The research presented in this study attempted to answer two research questions:

1) *What kinds of translation events affect cross-language retrieval?*
2) *In what way does the presence of certain translation events in query translation affect retrieval performance?*

In answer to the first research question a detailed taxonomy was developed which incorporated possible translation events from a query analysis and a literature review. The translation event taxonomy included three main translation event categories in addition to a translation correctness assessment.

To answer the second research question, a large number of English target queries (representing cross-lingual retrieval) was coded using the translation event taxonomy. In a retrieval experiment the performance of the English source queries (representing monolingual retrieval) was compared to that of the English target queries to asses the impact of the translation events on retrieval performance. A multiple regression analysis was carried out to see which translation events had an impact on retrieval performance. It was found that four of the six independent variables contributed significantly to prediction of the difference in retrieval performance when comparing monolingual to cross-lingual information retrieval: *missing specialized vocabulary* (IV2), *missing general terms* (IV7), *wrong translation due to ambiguity* (IV13), and *correct identical translation* (IV17). Although the contribution of each of these variables is significant, their contribution to the total variance of the independent variable is small (adjusted $R^2$ = .223). The small contribution of the independent variable to the total variance is likely to be caused by the information retrieval phenomenon called query variability. Here, unknown properties of the underlying document collection appear to diminish the correlation between a translation event and the retrieval performance of a term.

A traditional query analysis showed that there was a significant difference between monolingual and cross-lingual retrieval performance. More than half of the queries showed a performance difference between monolingual and cross-lingual retrieval of over 70%. The queries with less than 10% performance difference were shorter and less ambiguous than other queries in this study. The results also suggest that not all query terms are created equal and that some terms appear to be more important from a retrieval perspective than others. This also has implications for query

expansion because the results indicate that using a lexical resource for this purpose might not have the desired effect.

Future research should be aimed at trying to quantify term importance based on the unknown properties of the document collection in an attempt to explain a larger part of the variability in the retrieval difference score.

# 6 Bibliography


Al-Kasime, A.M. (1977). *Linguistics and Bilingual Dictionaries*. Leiden, The Netherlands: Brill. 131p.

Arnold, D., Balkan, L., Lee Humphreys, R.L., Meijer, S., and Sadler, L. (1994). *Machine Translation: An Introductory Guide*. Cambridge, Mass.: Blackwell. 240p.

Babbie, E. (1992). *The Practice of Social Science Research*. 6th edition. Belmont, Ca.: Wadsworth Publishing Company, 112p.

Baeza-Yates, R. and Ribeiro-Neto, B. (1999). *Modern Information Retrieval*. Reading, Mass.: Addison Wesley, 513p.

Ballesteros, L. and Croft, B. (1996). Dictionary Methods for Cross-Lingual Information Retrieval. In: *Proceedings of the 7th International DEXA Conference on Database and Expert Systems*, 1996 September 9-13; Zurich, Switzerland. New York, NY: Springer, 1996. 791-801

Ballesteros, L. and Croft, B. (1997). Phrasal Translation and Query Expansion Techniques for Cross-Language Information Retrieval. In: *Proceedings of the Association for Computing Machinery Special Interest Group on Information Retrieval (ACM/SIGIR) 20th International Conference on Research and Development in Information Retrieval*; 1997 July 25-31; Philadelphia, PA. New York, NY: ACM, 1997. 84-91.

Ballesteros, L. and Croft, B. (1998). Resolving Ambiguity for Cross-language Retrieval. In: *Proceedings of the Association for Computing Machinery Special Interest Group on Information Retrieval (ACM/SIGIR) 21st International Conference on Research and Development in Information Retrieval*; 1998 August 24-28; Melbourne, Australia. New York, NY: ACM, 1998. 64-71.

Banerjee, M.; Capozzoli, M.; McSweeney, L. and Sinha, D. (1999). Beyond Kappa: A Review of Interrater Agreement Measures. *Canadian Journal of Statistics*. Vol. 27, No. 1, 3-23.

Bart, P. van., Kerstens, J. and Sturm, A. (1998). *Grammatica van het Nederlands: Een Inleiding*. Amsterdam, The Netherlands: Amsterdam University Press. 262p.

Belkin, N.J., Oddy, R.N., and Brooks, H.M. (1982). ASK for Information Retrieval: Part I. Background and Theory. *The Journal of Documentation*. Vol. 38, 61-71.

Bell, Roger T. (1991). *Translation and Translating: Theory and Practice*. Essex, England: Longman. 298p.

Blair, D. C. (2002). Some Thoughts on the Reported Results of TREC. *Information Processing and Management*. Vol. 38, 445-451.

Blois, J., Buydens, J., Gougenheim, G., Hirschberg, L., Gougenheim, G. Michéa, and Rothenberg, M. *Problèmes de la Traduction Automatique*. Alabama, University of Alabama Press. 136p.

Bookstein, R., Swanson, D.R. (1974). Probabilistic Models for Automatic Indexing. *Journal of the American Society for Information Science*, Vol. 25, 312-318.

Boyce, B.R.; Meadow, C.T.; and Kraft, D.H. (1994). *Measurement in Information Science*. New York, NY: Academic Press, 283p.

Buckley, C., and Walz, J. (2000). The *TREC*-8 Query Track. . In: *Proceedings of the 8th Text REtrieval Conference (TREC-8)* E.M. Voorhees and D.K. Harman, (Eds.).; 1999 November 16-19; National Institute of Standards and Technology (NIST), Gaithersburg, MD. 65-76.

Braschler, M. et al. (1998). SPIDER Retrieval System at TREC7. In: *TREC-7 Working Notes*. Gaithersburg, MD: National Institute for Standards and Technology.

Braschler, M., Peters, C. (2002). The CLEF Campaigns: Evaluation of Cross-Language Information Retrieval Systems. In: *UPGRADE: The European Online Magazine for the IT Professional*, Vol. III, No. 3, 78-81.

Brown, R.D. (1997). Automated Dictionary Extraction for "Knowledge-Free" Example-Based Translation. In: *Proceedings of the 7th International Conference on Theoretical and Methodological Issues in Machine Translation*.

Buckley, C.; Mitra, M.; Walz, J.; and Cardie, C. (1998). Using Clustering and Super Concepts within SMART: TREC 6. In: *Proceedings of the 6th Text REtrieval Conference (TREC-6)*; 1997 November 19-21; National Institute of Standards and Technology (NIST), Gaithersburg, MD. 107-124.

Buckley, C. and Voorhees, E. (1999). *Theory and Practice in Text Retrieval System Evaluation.* A Tutorial Presented in Conjunction with the 22nd Annual International ACM/SIGIR Conference on Information Retrieval. Berkeley, CA. 1999 August 15.

Burgarski, R. (1985). Translation Across Cultures : Some Problems with Terminologies. In: *Scientific and Humanistic Dimensions of Language*, J.R. Jankowsky (Ed.). Festschrift for Robert Lado. Philadelphia, PA: Benjamins. 159-163.

Carbonnell, J.; Yang, Y.; Frederking, R.; Brown, R.; Geng, Y.; and Lee, D. (1997). Translingual Information Retrieval: A Comparative Evaluation. In*: Proceedings of the 15th International Joint Conference on Artificial Intelligence*; 1997 August 23-29, Nagaya, Japan. San Francisco, CA. Kaufmann, 1997. 706-715.

Carletta, J. (1996). Assessing Agreement on Classification Tasks: the Kappa Statistic. *Computational Linguistics*, Vol. 22, No. 2, 249-254.

Cleverdon, C.W. and Mills, J. (1963). The Testing of Indexing Language Devices. *Aslib Proceedings*, Vol. 15, No. 4, 106-130.

Cleverdon, C.W., Mills, J., and Keen, E.M. (1968). *Factors determining the performance of Indexing Systems*. Two Volumes. Cranfield, England.

Croft, W.B., Turtle, H., and Lewis, D. (1991). The Use of Phrases and Structured Queries in Information Retrieval. In: *Proceedings of the Association for Computing Machinery Special Interest Group on Information Retrieval (ACM/SIGIR) 14th International Conference on Research and Development in Information Retrieval*. 32-45.

Cronen-Townsend, S., Zhou, Y., Croft, W.B. (2002). Predicting Query Performance. *Proceedings of the Twenty-Fifth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. August 11-15; Tampere, Finland. 299-306.

Crystal, David. (1997). *A Dictionary of Linguistics and Phonetics*. 4th edition. Cambridge, Mass. : Blackwell. 426p.

Cutting, D., Pedersen, J. Noreault, T.; and Koll, M. (1997). *Real Life Information Retrieval: Commercial Search Engines*. Lesk, M. (Moderator). SIGIR Panel Session of at the 20th Annual ACM SIGIR Conference on Research and Development in Information Retrieval. Philadelphia, Pennsylvania, 1997, July 27-July 31.

Darlington, R.B. (1968). Multiple Regression in Psychological Research and Practice. *Psychological Bulletin*. Vol. 69, No. 3, 161-182.

Davis, M. (1997). New Experiments in Cross-Language Text Retrieval at NMSU's Computing Research Lab. In: *The Fifth Text Retrieval Conference (TREC-5)*. D.K. Harman, Ed. 1996, November. National Institute of Standards and Technology (NIST), Gaithersburg, MD.

Davis, M. and Dunning, T. (1995). A TREC Evaluation of Query Translation Methods for Multi-Lingual Text Retrieval. In: Proceedings of the 4th *Text REtrieval Conference (TREC-4)*; 1995 November 1-3; National Institute of Standards and Technology (NIST), Gaithersburg, MD.

Davis, M. and Ogden, (1997). Implementing Cross-language Text Retrieval Systems for Large-scale Text Collections and the World Wide Web. In: *Proceedings of the AAAI Symposium on Cross-Language Text and Speech Retrieval*. American Association for Artificial Intelligence. 2-10.

Diamond, T. (1999). Personal communication. October, 1999.

Diekema, A.; Oroumchian, F.; Sheridan, P.; Liddy, E. D. (1999). TREC-7 Evaluation of Conceptual Interlingua Document Retrieval (CINDOR) in English and French. In: *The Seventh Text REtrieval Conference (TREC-7)*. E.M. Voorhees and D.K. Harman (Eds.) 1998, November 9-11; National Institute of Standards and Technology (NIST), Gaithersburg, MD. 169-180.

Dumais, S. T.; Letsche, T. A.; Littman, M. L.; and Landauer, T. K. (1997). Automatic Cross-Language Retrieval Using Latent Semantic Indexing. In: *American Association for Artificial Intelligence (AAAI) Symposium on Cross-Language Text and Speech Retrieval*; 1997 March 24-26; Palo Alto, CA. 1997. 15-21.

Edwards, J. (1994). *Multilingualism*. London, England: Penguin. 256p.

Eisenberg, M.B. (1986). *Magnitude Estimation and the Measurement of Relevance*. Ph.D. dissertation, Syracuse University.

Eisenberg, M.B. and Barry, C. (1988). Order effects: A Study of the Possible Influence of Presentation Order on User Judgments of Document Relevance. *Journal of the American Society for Information Science*. Vol. 39, No. 5, 293-300.

Ellis, D. (1996). The Dilemma of Measurement in Information Retrieval Research. *Journal of the American Society for Information Science*. Vol. 47, No. 1, 23-36.

Ellis, D. (1992). The Physical and Cognitive Paradigms in Information Retrieval Research. *Journal of Documentation*, Vol. 48, No. 1, March. 45-64.

Erbach, G.; Neumann, G.; Uszkoreit, H. (1997). MULINEX: Multilingual Indexing Navigation and Editing Extensions for the World-Wide Web. In: *AAAI Spring Symposium on Cross-Language Text and Speech Retrieval*; 1997 March 24-26; Palo Alto, CA.

Fagan, J.L. (1989). The Effectiveness of a Nonsyntactic Approach to Automatic Phrase Indexing for Document Retrieval. *Journal of the American Society for Information Science*. Vol. 40, No. 2, 115-132.

Farwell, D., Gerber, L., and Hovy, E. (Eds.) (1998). *Machine Translation and the Information Soup: Proceedings of the Third Conference of the Association for Machine Translation in the Americas AMTA '98*. Langhorne, PA, USA, October 28-31, 1998. New York, NY: Springer. 532p.

Finegan, E. (1987). English. In: *The World's Major Languages*. B. Comrie (Ed.) New York, NY: Oxford University Press. 77-109.

Fluhr, C., Schmit, D., Elkateb, F., Ortet, P., & Gurtner, K. (1997). Multilingual Database and Cross-lingual Interrogation in a Real Internet Application. In: *AAAI Symposium on Cross-Language Text and Speech Retrieval : American Association for Artificial Intelligence*. 32-36.

Frakes, William B. and Baeza-Yates, R. (Eds.) (1992). *Information Retrieval : Data Structures & Algorithms*. Upper Saddle River, N.J. : Prentice Hall PTR. 504p.

Gaussier, E., Grefenstette, G., Hull, D.A., and Schulze, B.M. (1998). Xerox TREC-6 Site Report: Cross-Language Text Retrieval. In: *Proceedings of the 6th Text REtrieval Conference (TREC-6)*; 1997 November 19-21; National Institute of Standards and Technology (NIST), Gaithersburg, MD.

Gey, F., Jiang, H., and Chen, A. (1999). Manual Queries and Machine Translation in Cross-Language Retrieval and Interactive Retrieval with Cheshire II at TREC-7. In: *The Seventh Text REtrieval Conference (TREC-7)*. E.M. Voorhees and D.K. Harman (Eds.) 1998, November 9-11; National Institute of Standards and Technology (NIST), Gaithersburg, MD. 527-540.

Gey, F. and Chen, A. (1998). Phrase Discovery for Cross-Language Retrieval at TREC 6. In: *Proceedings of the 6th Text REtrieval Conference (TREC-6)*; 1997 November 19-21; National Institute of Standards and Technology (NIST), Gaithersburg, MD. 637-648.

Gilarranz, J.; Gonzalo, J. and Verdejo, F. (1997). An approach to Conceptual Text Retrieval Using the EuroWordNet Multilingual Semantic Database. In: *American Association for Artificial Intelligence (AAAI) Symposium on Cross-Language Text and Speech Retrieval*; 1997 March 24-26; Palo Alto, CA. 1997. 49-55.

Gopinathan, G. (1993). The Nature and Problems of Translation. In: *Problems of Translation*. G. Gopinathan and S. Kandaswamy (Eds.). Allahabad, India: Lokbharati Prakashan. 37-50.

Grefenstette, G. (1995). Comparing Two Language Identification Schemes. In*: Proceedings of the 3rd International Conference on the Statistical Analysis of Textual Data*; 1995 December; Rome, Italy.

Grossman, D. A. and Frider, O. (1998). *Information Retrieval: Algorithms and Heuristics*. Boston, Mass.: Kluwer Academic Publishers. 254 p.

Gruber, T.R. (1993). *Toward Principles for the Design of Ontologies Used for Knowledge Sharing*. KSL-93-04. Stanford, CA: Knowledge Systems Laboratory, Stanford University.

Gutt, E. (1991). *Translation and Relevance: Cognition and Context*. Cambridge, Mass.: Blackwell. 222p.

Hamp, B. and Feldweg, H. (1997). GermaNet – A Lexical-Semantic Net for German. Available at: http://www.sfs.nphil.uni-tuebingen.de/isd/english.html

Harman, D.K. (1995). The TREC Conferences. In: *Hypertext – Information Retrieval – Multimedia: Synergieeffekte Elektronischer Informationssysteme*, Proceedings of HIM '95. R. Kuhlen and M. Rittberger (Eds.) Kontanz, Germany: Universitätsverlag Konstanz. 9-28.

Harman, D.K. (1996). Overview of the Fourth Text Retrieval Conference (TREC-4). In: *Proceedings of the Fourth Text REtrieval Conference (TREC-4)*. National Institute of Standards and Technology (NIST), Gaithersburg, MD.

Hartmann,R. (1989). Lexicography, Translation and the So-Called Language Barrier. *Translation and Lexicography*. M. Snell-Hornby, E. Pöhl, and B. Bennani (Eds.) Amsterdam, The Netherlands: John Benjamins Publishing Company. 9-20.

Hiemstra, D. and Kraaij, W. (1999). Twenty-One at TREC-7: Ad-hoc and Cross-language Track. In: *The Seventh Text REtrieval Conference (TREC-7)*. E.M. Voorhees and D.K. Harman (Eds.) 1998, November 9-11; National Institute of Standards and Technology (NIST), Gaithersburg, MD. 227-238.

Hull, D.A. (1999). Personal communication. October and November, 1999.

Hull, D. A. (1993). Using Statistical Testing in the Evaluation of Retrieval Experiments. In: *Proceedings of the Association for Computing Machinery Special Interest Group on Information Retrieval (ACM/SIGIR)*. New York, NY: ACM. 329-338.

Hull, D. A. (1997). Using Structured Queries for Disambiguation in Cross-Language Information Retrieval. In: *American Association for Artificial Intelligence (AAAI) Symposium on Cross-Language Text and Speech Retrieval*; 1997 March 24-26; Palo Alto, CA. 1997. 84-98.

Hull, D. A. and Grefenstette, G. (1996). Querying across Languages: A Dictionary-Based Approach to Multilingual Information Retrieval. In: *Proceedings of the Association for Computing Machinery Special Interest Group on Information Retrieval (ACM/SIGIR) 19th International Conference on Research and Development in Information Retrieval*; 1996 August 18-22; Zurich, Switzerland. New York, NY: ACM, 1997. 49-57.

Ingwersen, P. (1992). *Information Retrieval Interaction*. London, UK: Taylor Graham.

Jones, G., Collier, N., Sakai, T., Sumita, K. and Kirakawa, H. (1998). CLIR Access a Case Study for English and Japanese. In: *Proceedings of the Joint Workshop of the IPSJ, SIG-FI and SIG-NL*. Tokyo, Japan, September 1998. 47-54.

Kalachkina, S.Y. (1987). Algorithmic Determination of Descriptor Equivalents in Different Natural Languages. *Automatic Documentation and Mathematical Linguistics*, Vol. 21, No. 4, 21-29. English translation from Russian.

Katzer, J. (1984). Inference and Simple Regression. Handout to IST770, Statistics in Behavioral Research. Syracuse, NY: Syracuse University, School of Information Studies.

Katzer, J., Cook, K. H., Crouch, W. W. (1991). *Evaluating information: A guide for users of social science research*. Third edition. New York, NY: , McGraw-Hill. 272p.

Katzner, K. (1995). *The Languages of the World*. New Edition. New York, NY: Routledge. 378p.

Karatzoglou, M. (1997). *Translib Edited Report*. (LIB/3-3038). Patras, Greece: Knowledge S.A.

Keen, E.M. (1991). Evaluation Parameters. In: *The SMART Retrieval System: Experiments in Automatic Document Processing*. G. Salton Ed. Prentice-Hall, Inc. Englewood Cliffs, NJ. 74-111.

Keen, E.M. (1981). Laboratory tests of Manual Systems. In: *Information Retrieval Experiment*. K. Sparck Jones (Ed.) London: Butterworths, 136-155.

Keen, E. M. (1992). Presenting Results of Experimental Retrieval Comparisons. *Information Processing and Management*, Vol. 28, 491-502.

Kerlinger, F.N. and Pedhazur, E. (1973). *Multiple Regression in Behavioral Research*. New York: Holt, Rinehart and Winston.

Kikui G. (1996). Identifying the Coding System and Language of On-line Documents on the Internet. In*: COLING-96: 16th International Conference on Computational Linguistics*: 1996 August 5-9; Copenhagen, Denmark. Copenhagen, Denmark: Center for Sprogteknologi, 1996. 652-657.

Kirk, R. E. (1995). *Experimental Design: Procedures for the Behavioral Sciences*. Third edition. Pacific Grove, CA: Brooks/Cole Publishing Company. 921p.

Kooij, J.G. (1987). Dutch. In: *The World's Major Languages*. B. Comrie (Ed.) New York, NY: Oxford University Press. 139-156.

Kowalski, G. (1997). *Information Retrieval Systems : Theory and Implementation*. Boston, Mass.: Kluwer Academic Publishers. 282p.

Kraaij, W. (1997). Multilingual Functionality in the Twenty-One Project. In: *American Association for Artificial Intelligence (AAAI) Symposium on Cross-Language Text and Speech Retrieval*; 1997 March 24-26; Palo Alto, CA 1997. 127-132

Kraaij, W. (2001). *TNO at CLEF-2001: Comparing Translation Resources*. In: "CLEF-2001 Working Notes". (http://www.clef-campaign.org).

Kraaij, W. and Hiemstra, D. (1998). Cross Language Retrieval with the Twenty-One System. In: *Proceedings of the 6th Text REtrieval Conference (TREC-6)*; 1997 November 19-21; National Institute of Standards and Technology (NIST), Gaithersburg, MD. 753-760.

Krovetz, R. and Croft, B. (1992). Lexical Ambiguity and Information Retrieval. *ACM Transactions on Information Systems*, Vol. 10, No. 2, 115-141.

Kwasnik, B. H. (1999). The Role of Classification in Knowledge Representation and Discovery. *Library Trends* Vol. 48, No. 1, 22-47.

Kwok, K. L. (1997). Evaluation of an English-Chinese Cross-Lingual Retrieval Experiment. In: *AAAI Symposium on Cross Language Text and Speech Retrieval : American Association for Artificial Intelligence*. 133-137.

Landauer, T.K. and Littman, M. L. (1990). Fully Automatic Cross-Language Document Retrieval Using Latent Semantic Indexing. In: *Proceedings of the Sixth Annual Conference of the UW Centre for the New Oxford English Dictionary and Text Research*, 1990, Waterloo, Ontario. 31-38.

Landauer, T.K. and Littman, M. L. (1991). A Statistical Method for Language-Independent Representation of the Topical Content of Text Segments. In: *Proceedings of the 11th International Conference: Expert Systems and Their Applications*: Volume 8. 1991 May 27-31; Avignon, France.

Landis, J.R. and Koch, G.G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics*, Vol. 33. 159-174.

Larson, M.L. (1984). *Meaning-Based Translation: A Guide to Cross-Language Equivalence*. New York, NY: University Press of America. 537p.

Lehrberger, J. and Bourbeau, L. (1988). *Machine Translation : Linguistic Characteristics of MT Systems and General Methodology of Evaluation*. Philadelphia, Penn.: John Benjamins Publishing Company. 240p.

Lewis, D. and Croft, W.B. (1990). Term Clustering of Syntactic Phrases. In: *Proceedings of the 13th International Conference on Research and Development in Information Retrieval*. 385-404.

Mateev, B., Munteanu, E., Sheridan, P., Wechsler, M. and Schäuble, P. (1998). ETH TREC-6: Routing, Chinese, Cross-Language and Spoken Document Retrieval. In: *Proceedings of the Sixth Text Retrieval Conference (TREC-6)*, National Institute of Standards and Technology (NIST), Gaithersburg, MD. 623-636.

McCarley, J.S. (1999). Should we Translate the Documents or the Queries in Cross-Language Information Retrieval? In: *37th Annual Meeting of the Association for Computational Linguistics*. College Park, MD.

McEnery, T. and Wilson, A. (1996). *Corpus Linguistics*. Edinburgh, Scotland: Edinburgh University Press. 209p.

McNamee, P. and Mayfield, J. (2002). Comparing Cross-Language Query Expansion Techniques by Degrading Translation Resources. In: *Proceedings of the Twenty-Fifth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*; 2002 August 11-15. Tampere, Finland. 159-166.

Meetham, A. R. and Hudson, R.A. (1969). *Encyclopaedia in Linguistics, Information and Control*. Oxford, England: Pergamon.

Miller, G. (1990). WordNet: An On-line Lexical Database. *International Journal of Lexicography*, Vol. 3, No. 4. Special Issue.

Neubert, A. (1985). Translation Across Language or Across Cultures? In: *Scientific and Humanistic Dimensions of Language*. J.R. Jankowsky (Ed.). Festschrift for Robert Lado. Philadelphia, PA: Benjamins.

Newmark, P. (1991). *About Translation*. Bristol, Penn.: Multilingual Matters Ltd. 184p.

Newton, H. (1997). *Newton's Telecom Dictionary*. 12th edition. New York, NY: Flatiron Publishing. 750p.

Nguyen, V.B.H., Wilkinson, R. and Zobel, J. (1997). Cross-Language Retrieval in English and Vietnamese. In: *AAAI Symposium on Cross-Language Text and Speech Retrieval : American Association of Artificial Intelligence*. 143-145.

Nida, E.A. (1958). Analysis of Meaning and Dictionary Making. *International Journal of American Linguistics*. Vol. 24. 279-292.

Nida, E.A. (1964). *Towards a Science of Translating: With Special Reference to Principles and Procedures Involved in Bible Translating*. Leiden, The Netherlands: Brill.

Nida, E.A. and Taber, C.R. (1969). *The Theory and Practice of Translation*. Leiden, The Netherlands: Brill.

Nie, J.; Simard, M.; Isabelle, P.; and Durand, R. (1999). Cross-Language Information Retrieval based on Parallel Texts and Automatic Mining of Parallel Texts from the Web. In: *Proceedings of the 22nd International Conference on Research and Development in Information Retrieval*. Berkeley, California: ACM/SIGIR. 74-81.

Nyberg, E. H. , Mitamura, T., and Carbonell, J. G. (1994). Evaluation Metrics for Knowledge-Based Machine Translation. In: *Proceedings of the 15th International Conference on Computational Linguistics (COLING – '94)*. Kyoto, Japan.

Oard, D. and Diekema, A. (1998). Cross-Language Information Retrieval. *Annual Review of Information Science (ARIST)*, Vol. 33, M. Williams (Ed.), Information Today Inc., Medford, NJ, 1998.

Oard, D. and Hackett, P. (1998). Document Translation for Cross-Language Text Retrieval at the University of Maryland. In: *Proceedings of the 6th Text REtrieval Conference (TREC-6)*; 1997 November 19-21; National Institute of Standards and Technology (NIST), Gaithersburg, MD. 687-696.

Oard, D.W., Dorr, B.J., Hackett, P.G. and Katsova, M. (1998). *Comparative Study of Knowledge-Based Approaches for Cross-Language Information Retrieval*. Institute for Advanced Computer Studies, University of Maryland. CS-TR-3897.

Office for Official Publication of the European Communities (1995). *Thesaurus EUROVOC Volume 3: Multilingual version*. Luxembourg.

Pause, P.E. (1997). Interlingual Strategies in Translation. In*: Machine Translation and Translation Theory*. C. Hauenschild, and S. Heizmann (Eds.) New York, NY: Mouton de Gruyter. 175-190.

Peters, C. (2002). Introduction to the Working Notes for the CLEF 2002 Workshop. In: *Working Notes for the CLEF 2002 Workshop*. 19-20 September, Rome, Italy.

Peters, C. and Picchi, E. (1997). Using Linguistic Tools and Resources in Cross-Language Retrieval. In: *Proceedings of the AAAI Symposium on Cross-Language Text and Speech Retrieval*. American Association for Artificial Intelligence. 1997. 179-188.

Pigur, V.A. (1979). Multilanguage Information-Retrieval Systems: Integration Levels and Language Support. *Automatic Documentation and Mathematical Linguistics*. Vol. 13, No. 1, 36-46. (English translation from Russian)

Pinker, S. (1994). *The Language Instinct: How the Mind Creates Language*. New York, NY: Harper Perennial. 494p.

Pirkola, A., Järvelin, K. (2001). Employing the Resolution Power of Search Keys. *Journal of the American Society for Information Science and Technology*, Vol. 52, No. 7, 575-583.

Prahl, B. and Petzolt, S. (1997). Translation Problems and Translation Strategies Involved in Human and Machine Translation: Empirical Studies. In*: Machine Translation and Translation Theory*. C. Hauenschild and S. Heizmann, (Eds.) New York, NY: Mouton de Gruyter. 121-144.

Radwan, K. and Fluhr, C. (1995). Textual Database Lexicon Used as a Filter to Resolve Semantic Ambiguity Applications on Multilingual Information Retrieval. In: *Proceedings of the 4th Annual Symposium on Document Analysis and Information Retrieval*. University of Nevada. 1995. 121-136.

Resnik, P. (1997). Evaluating Multilingual Gisting of Web Pages. In: *American Association for Artificial Intelligence (AAAI) Symposium on Cross-Language Text and Speech Retrieval*; 1997 March 24-26; Palo Alto, CA. 1997. 189-195.

Robertson, S. E. (1981). The Methodology of Information Retrieval Experiment. In: *Information Retrieval Experiment*. K. Sparck Jones (Ed.) London: Butterworths, 9-31.

Robertson, S.E. (1997). Overview of the Okapi Projects. *Journal of Documentation*. Vol. 53, No. 1, 3-7.

Robertson, S.E., S. Walker, S. Jones, M.M. Beaulieu, M. Gatford, and A. Payne. (1995). Okapi at TREC-4. In: *Proceedings of the 4th Text REtrieval Conference (TREC-4)*; 1995 November 1-3; National Institute of Standards and Technology (NIST), Gaithersburg, MD.

Roscoe, J. T. (1975). *Fundamental Research Statistics for the Behavioral Sciences*. 2nd edition. Philadelphia, PA.: Holt, Rinehart, and Winston. 483p.

Salton, G. (1970). Automatic Processing of Foreign Language Documents. *Journal of the American Society for Information Science*. Vol. 21, No. 3, 187-194.

Salton, G. (1981). The SMART Environment for Retrieval System Evaluation-advantages and Problem Areas. In: *Information Retrieval Experiment*. K. Sparck Jones (Ed.). London: Butterworths, 316-329.

Salton, G., (Ed.). (1971) *The SMART Retrieval System: Experiments in Automatic Document Processing*. Prentice-Hall, Inc. Englewood Cliffs, NJ. 556p.

Salton, G. (1973). Experiments in Multi-Lingual Information Retrieval. *Information Processing Letters*. Vol. 2, No. 1, 6-11.

Salton, G. (1992). The State of Information Retrieval System Evaluation. *Information Processing & Management*. Vol. 28, No. 4, 441-449.

Salton, G. and McGill, M.J. (1983). *Introduction to Modern Information Retrieval*. New York, NY: McGraw-Hill. 448p.

Sanderson, M. (1994). Word Sense Disambiguation and Information Retrieval. In: *Proceedings of the Association for Computing Machinery Special Interest Group on Information Retrieval (ACM/SIGIR) 17th International Conference on Research and Development in Information Retrieval*; 1994 July 3-6; Dublin, Ireland. 49-57.

Saracevic, T. (1970). The Concept of "Relevance" in Information Science: A Historical Review. In: *Introduction to Information Science*, T. Saracevic, (Ed.). New York, NY: Bowker, 111-151.

Saracevic, T. (1975) Relevance: a Review of and a Framework for Thinking on the Notion in Information Science. *Journal of the American Society for Information Science*, Vol. 26. 321-343.

Sarcevic, S. (1989). Lexicography and Translation Across Cultures. In: *Translation and Lexicography*. M. Snell-Hornby, E. Pöhl, and B. Bennani (Eds.) Amsterdam, The Netherlands: John Benjamins Publishing Company.

Schamber, L.; Eisenberg, M.B. and Nilan, M.S. (1990). A Re-examination of Relevance: Toward a Dynamic, Situational Definition. *Information Processing & Management*. Vol. 26, No. 6, 755-776.

Schulte, R. (1987). Translation Theory: A Challenge for the Future. *Translation Review*, Vol. 23. 1-2.

Sheridan, P., Ballerini, J.P. and Schäuble, P. (1998). Building a Large Multilingual Test Collection from Comparable News Documents. In: *Cross Language Information Retrieval*. G. Grefenstette, (Ed.). Boston, MA: Kluwer Academic.

Sheridan, P. and Schäuble, P. (1998). Cross-Language Information Retrieval (CLIR) Track Overview. In: *Proceedings of the 6th Text REtrieval Conference (TREC-6)*; 1997 November 19-21; National Institute of Standards and Technology (NIST), Gaithersburg, MD. 31-44.

Sibert, E. and Shelly, A. (1992). Qualitative Analysis: A Cyclical Process Assisted by Computer / Qualitative Analyse: Ein Computerunterstützter zyklischer Prozess. In: *Qualitative Analyse: Computereinsatz in der Sozialforschung*. G.L. Hüber, (Ed.). Münich, Vienna: R. Oldenbourg Verlag. 71-114.

Snell-Hornby, M. (1990). In: *Meaning and Lexicography*. Tomaszczyk, J. and Lewandowska-Tomaszczyk, B. (Eds.) Amsterdam, The Netherlands: John Benjamins Publishing Company. 209-226.

Snell-Hornby, M., Pöhl, E., and Bennani, B. (Eds.) (1989). *Translation and Lexicography*. Amsterdam, The Netherlands: John Benjamins Publishing Company.

Sparck Jones, K. (1981). *Information Retrieval Experiment*. London: Butterworths.

Sparck Jones, K. (1995). Reflections on TREC. *Information Processing & Management*. Vol. 31, No. 3, 291-314.

Sparck Jones, K. and Van Rijsbergen, C. *Report on the Need for and Provision of an "Ideal" Information Retrieval Test Collection*. British Library Research and Development Report 5266, Computer Laboratory, University of Cambridge, 1975.

Ruiz, M. E. and Srinivasan, P. (1998). Cross-lingual Information Retrieval with the UMLS: An Analysis of Errors. In: *Proceedings of the 61st Annual Meeting of the American Society for Information Science*. Pittsburgh, PA.: Oct. 24-29.

Steiner, G. (1975). *After Babel*. London: Oxford University Press.

Storrer, A. and Schwall, U. (1993). Description and Acquisition of Multiword Lexemes. In: *Machine Translation and the Lexicon: Proceedings of the Third International EAMT Workshop*. Petra Steffens, (Ed.). Heidelberg, Germany, April 1993. Heidelberg, Germany: Springer. 35-50.

Steffens, P. (Ed.) (1993). *Machine Translation and the Lexicon: Proceedings of the Third International EAMT Workshop*. Heidelberg, Germany, April 1993. Heidelberg, Germany: Springer. 251p.

Tabachnick, B. G. and Fidell, L. S. (2001). *Using Multivariate Statistics*. 4th edition. Boston, Mass: Allyn and Bacon. 966p.

Tague, J. M. (1981). The Pragmatics of Information Retrieval Experimentation. In: *Information Retrieval Experiment*. K. Sparck Jones (Ed.). London: Butterworths, 59-102.

Tague-Sutcliffe, J. (1992). The Pragmatics of Information Retrieval Experimentation, Revisited. *Information Processing and Management*. Vol. 28, 476-490.

Text REtrieval Conference Home Page: http://trec.nist.gov/

Tomaszczyk, J. (1989) L1-L2 Technical Translation Dictionaries. In: *Translation and Lexicography*. M. Snell-Hornby, E. Pöhl and B. Bennani (Eds.) Amsterdam, The Netherlands: John Benjamins Publishing Company.

Toury, G. (1995). *Descriptive Translation Studies and Beyond*. Amsterdam, The Netherlands: John Benjamins Publishing Co., 311p.

Voorhees, E.M. (1998). Variations in Relevance Judgments and the Measurement of Retrieval Effectiveness. *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 307-314.

Voorhees, E.M. and Harman, D. (1999). Overview of the Seventh Text REtrieval Conference (TREC-7). In: *NIST Special Publication 500-242: The Seventh Text REtrieval Conference (TREC-7)*. E.M. Voorhees and D.K. Harman, (Eds.). Gaithersburg, MD: Department of Commerce, National Institute for Standards and Technology, 1-24.

Van Rijsbergen, C. J. (1981). Retrieval Effectiveness. In: *Information Retrieval Experiment*. K. Sparck Jones (Ed.).London: Butterworths. 32-43.

*Van Dale Groot Woordenboek Nederlands-Engels*. (1997). Van Dale Lexicografie. Utrecht, The Netherlands. (based on the 2nd edition paper version published in 1991).

Wersig, G. and Neveling, U. (1976). *Terminology of Documentation : A selection of 1,200 basic Terms Published in English, French, German, Russian and Spanish*. Paris, France: The Unesco Press. 274p.

Wilks, Y.A., Slator, B. M. and Guthrie, L.M. (1996). *Electric Words: Dictionaries, Computers, and Meanings*. Cambridge, Mass.: MIT Press. 298p.

Yamabanda, K.; Muraki, K.; Doi, S; and Kamei, S. et al. (1998). Front-End for Cross-Linguistic Information Retrieval. In: *Cross Language Information Retrieval*. G. Grefenstette, (Ed.) Boston, MA: Kluwer Academic, 1998.

Zobel, J. (1998). How Reliable are the Results of Large-Scale Information Retrieval Experiments? In: *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 307-314.

# Appendix I: Glossary

| Term | Definition[55] |
|---|---|
| **ad hoc information retrieval** | *See* retrospective information retrieval |
| **alignment** | *See* corpus alignment |
| **article** | A term used in the grammatical classification of words, referring to a subclass of determiners which displays a primary role in differentiating the uses of nouns, e.g. *the/a* in English. (Crystal, 1997, p. 26) |
| **Boolean retrieval** | [In Boolean retrieval the query is stated using Boolean logic.] A Boolean query is given with the usual operators -AND, OR and NOT. The result set must contain all documents that satisfy the Boolean condition. (Grossman and Frieder, 1998, p. 176)<br>*See also* weighted Boolean |
| **CLIR** | *See* cross-language information retrieval |
| **coefficient of determination** | *See* R squared |
| **cognate matching** | Cognate matching essentially automates the process by which readers might try to guess the meaning of an unfamiliar term based on similarities in spelling or pronunciation. A simple version of cognate matching in which untranslatable terms are retained unchanged is often used in CLIR systems to match proper nouns and technical terminology. In more sophisticated approaches, equivalence classes are created for letter sequences with similar sounds (e.g., "c," "k," and "qu" share an equivalence class) increasing the number of cognates that can be matched across closely related languages. (Oard and Diekema, 1998) |
| **collection** | *See* document collection |
| **comparable corpus** | [Comparable corpora contain documents in different languages which are not direct translations of each other as is the case in parallel corpora. Documents in comparable corpora are created independently in each different language but share topicality. Alignment in comparable corpora usually only takes place on a document level.]<br>*See also* multilingual corpus ; parallel corpus ; alignment. |
| **content-bearing word** | [A word that has potential to identify the content of a document, a word with indexing value]<br>*See also* stop word |
| **controlled indexing** | Indexing using a controlled vocabulary.<br>See also controlled vocabulary ; free-text |
| **controlled vocabulary** | A controlled vocabulary is a finite set of index terms from which all index terms must be selected (the domain of the index). (Kowalski, 1997, p. 50) |
| **corpus** | [A] corpus in modern linguistics, in contrast to being simply any |

[55] For references in the definitions see *Bibliography* section.

| | |
|---|---|
| | body of text, might more accurately be described as a finite-sized body of machine-readable text, sampled in order to be maximally representative of the language variety under consideration. (McEnery and Wilson, 1996, p. 24)<br>*See also* multilingual corpus ; parallel corpus ; comparable corpus |
| **corpus alignment** | For the corpus to be of maximum utility, it is necessary to go further and identify which sentences in these sub-corpora are translations of each other, and below that level, which words are translations of each other. In order to bring the parallel sub-corpora into this more specific relationship with one another they need to be aligned. (McEnery and Wilson, 1996, p. 58)<br>*See also* multilingual corpus ; parallel corpus ; comparable corpus |
| **cross-language information retrieval** | Cross-Language Information Retrieval (CLIR) is a special case of Information Retrieval (IR) where retrieval is not restricted to the query language but queries in one language retrieve documents in multiple languages. (Oard and Diekema, 1998) |
| **cross-language track** | [A special track within TREC which focuses on evaluation of retrieval across languages.]<br>*See also* Text REtrieval Conference |
| **determiner** | A term used in some models of grammatical description, referring to a class of items whose main role is to co-occur with nouns to express a wide range of semantic contrasts, such as quantity or number. The articles when they occur in a language, are the main subset of determiners (e.g. *the/a* in English; other words which can have a determiner function in English include *each/every, this/that, some/any*...(Crystal, 1997, p. 112) |
| **document** | A document is a data object, usual textual, though it may also contain other types of data such as photographs, graphs, and so on. Often, the documents themselves are not stored directly in the IR system, but they are represented in the system by document surrogates. (Frakes and Baeza-Yates, 1992, p. 1)<br>*See also* document representation |
| **document collection** | [The database on which retrieval is carried out.] |
| **document frequency** | [The number of times a term appears in the entire document collection.]<br>*See also* term frequency |
| **document representation** | [The retrieval system's representation of the original document after document processing operations such as parsing, stemming, and stop word removal have taken place.] |
| **exhaustivity** | Exhaustivity of indexing. Number of index terms assigned to an indexed item. (Wersig and Neveling, 1976, p. 110)<br>*See also* specificity |
| **filtering** | [In information filtering or Selective Dissemination of Information] the user defines a profile (similar to a stored query) and as new information is added to the system it is automatically compared to the user's profile. [here the collection is changing and the query is static] (Kowalski, 1997, p. 162) |
| **free-text** | [As opposed to structured data. Free text has] minimal consistency in the vocabulary and styles of items discussing the exact same |

| | |
|---|---|
| | issue. The searcher has to be omniscient to specify all search term possibilities in the query. (Kowalski, 1997, p. 18)<br>*See also* structured data ; controlled indexing |
| **fuzzy matching** | Fuzzy matching allows matching between terms that differ slightly in spelling. Thus, terms with slight spelling variations can be matched which would be impossible in a regular exact match. This type of matching is especially useful in CLIR because some languages might have only slight spelling variations for certain terms (e.g. Netherlands - Nederland). |
| **homonymy** | A term used in semantic analysis to refer to lexical items which have the same form but differ in meaning. 'Homonymy' is illustrated from the various meanings of *bear* (=animal, carry) or *ear* (of body, of corn). (Crystal, 1997, p. 185). (*See also* polysemy) |
| **index** | An index is a file of the values of the searchable attributes sorted into a broadly accepted order like that of the Roman alphabet. Enumerated with each attribute value is a list of unique arbitrary identifiers for the records in question. These records are arranged by those unique identifiers in another file.<br>(Boyce, Meadow, and Kraft, 1994, p. 235) |
| **information filtering** | *See* filtering |
| **information retrieval** | Information Retrieval (IR) is concerned with the representation, storage, organization, and accessing of information items. (Salton & McGill, 1983, p. 1) |
| **inverse document frequency weight** | a weighting of word frequencies to reflect the difference between high rate of occurrence in all (most) documents and high in only in only a limited set of them.<br>(Boyce, Meadow, and Kraft, 1994, p. 103)<br>[The term *inverse* stems from the assumption behind this weight, which is that the discrimination power of a term is inversely proportional to the number of documents in which it occurs.]<br>*See also* term frequency |
| **IR** | *See* Information Retrieval |
| **lexical ambiguity** | Ambiguity which does not arise from the grammatical analysis of a sentence, but is due solely to the alternative meanings of an individual lexical item, is referred to as lexical ambiguity, e.g. *I found the table fascinating* (='object of furniture' or 'table of figures'). (Crystal, 1997, p. 17) (*See also* polysemy) |
| **logistic regression** | Logistic regression allows one to predict a discrete outcome such as group membership from a set of variables that may be continuous, discrete, dichotomous, or a mix. (Tabachnick, and Fidell, 2001, p. 517) |
| **machine-readable dictionary** | [Dictionaries in electronic, machine-readable form.]<br>*See also* transfer dictionary |
| **machine translation** | Automatic translation; mechanical translation. Translation performed by automatic means. (Wersig and Neveling, 1976, p. 77) |
| **Mahalanobis distance** | Mahalanobis distance is the distance of a case from the centroid of the remaining cases where the centroid is the point created at the intersection of the means of all the variables. (Tabachnick and Fidell, 2001, p. 68). Mahalanobis distance is distributed as a chi- |

square ($\chi^2$) variable, with degrees of freedom equal to the number of IVs. (p. 157).

**MRD** *See* machine readable dictionary

**MT** *See* machine translation

**multilingual corpus** Multilingual corpora contain texts in several [two or more] different languages. In practice, some multilingual corpora might more truly be described as small collections of individual monolingual corpora in the sense that they use the same or similar sampling procedures and categories for each language but contain completely different texts in those several languages. (McEnery and Wilson, 1996, p. 57)
*See also* parallel corpus ; comparable corpus

**multiple regression** A statistical procedure determining an equation which best predicts the dependent variable from two or more predictor variables. (Katzer, Cook, and Crouch, 1991, p. 255)

**natural disambiguation** *See* self-disambiguation

**noise** [random error]

**non-compositional phrase** A phrase in which meanings of the individual words can not be used to build up the meanings of larger units (e.g. real estate). (based on Crystal's definition of compositionality, 1997, p. 78)

**okapi** An African mammal (*Okapia johnstoni*) closely related to the giraffe but lacking the elongated neck. (Merriam Webster's Collegiate Dictionary)
Probabilistic information retrieval system developed by City University, London, England.

**ontology** Formally, an ontology consists of terms, their definitions, and axioms relating them. (Gruber, 1993)

**parallel corpus** Parallel corpora are corpora which, rather than simply employing the same sampling procedures, actually hold the same texts in more than one language. The basic notion of a parallel corpus pre-dates computer corpus linguistics by several centuries. From mediaeval times onwards, so-called 'polyglot' bibles were produced which contained the biblical texts side by side in Hebrew, Greek, Latin, and sometimes vernacular versions. In an almost identical way, a machine-readable parallel corpus presents the user with different translations of the same text. (McEnery and Wilson, 1996, p. 57)
*See also* multilingual corpus ; comparable corpus

**part-of-speech-tagger** [Software program which automatically assigns parts-of-speech)]
The task of part-of-speech assignment consists of assigning a word to its appropriate word class. In the systems developed to do that, the traditional basic part-of-speech distinctions, such as adjective, verb, noun, adverb and so on, have been supplemented with further relevant information , such as person and number. (McEnery and Wilson, 1996, p. 119)

**phrase** A term used in grammatical analysis to refer to a single element of structure typically containing more than one word, and lacking the subject-predicate structure typical of clauses. (Crystal, 1997, p. 292)

**polysemy** A term used in semantic analysis to refer to a lexical item which has

| | |
|---|---|
| | a range of different meanings, e.g. *plain* = 'clear', 'unadorned', 'obvious'...; opposed to monosemy (or univocality). A large proportion of a language's vocabulary is polysemic (or polysemous). (Crystal, 1997, p. 297) [The difference between homonymy and polysemy is a controversial issue in linguistics since it is often hard to tell whether one is dealing with a homonymous or polysemous word. Strictly speaking, polysemous senses of a word are related (e.g. *foot* - end of your leg) and *foot* - bottom of the mountain) as opposed to homonyms where words just happen to have the same form (*bark* - outside of a tree and *bark* -sound made by a dog) (Crystal, 1997). The difference is not found to be of consequence to information retrieval.<br>*See also* homonymy |
| **POS tagger** | *See* part-of-speech tagger |
| **precision** | The ratio of the number of relevant records retrieved to the total number retrieved. (Boyce, Meadow, and Kraft, 1994, p. 180)<br>*See also* recall |
| **preposition** | A term used in the grammatical classification of words, referring to the set of items which typically precede noun phrases (often single nouns or pronouns), to from a single constituent of structure. (Crystal, 1997, p. 305) e.g. *in*, *on*, *under*, *over*, etc. |
| **pronoun** | A term used in the grammatical classification of words, referring to the closed set of items which can be used to substitute for a noun phrase (or single noun). (Crystal, 1997, p. 312) e.g. *he*, *him*, *me*, *mine*, *she*, etc. |
| **pseudo relevance feedback** | Pseudo relevance feedback automatically reformulates a query. It assumes that the top *n* documents retrieved by a certain query are relevant. A selection of terms from these documents is then used to expand the original query. A new set of documents is retrieved with this expanded query and presented to the user.<br>*See also* relevance feedback |
| **query** | The statement a user makes to a retrieval system or service is what we call the *information need statement*, or INS. The statement made to a retrieval system is the *query*. (Boyce, Meadow, and Kraft, 1994, p. 178)<br>*See also* topic |
| **query representation** | [The retrieval system's representation of the user's query after document processing operations such as parsing, stemming, and stop word removal have taken place.] |
| **query translation** | [Cross-Language Retrieval approach in which the user's query statements are translated into the document languages.] |
| **R squared** | The multiple correlation coefficient, R square3d. It is interpreted as the percent of variation of the dependent variable accounted for by the independent variables in the multiple regression equation. (Katzer, Cook, and Crouch, 1991, p. 245) |
| **recall** | The ratio of the number of relevant records retrieved to the number of relevant records existing in the database. (Boyce, Meadow, and Kraft, 1994, p. 181)<br>*See also* precision |

| | |
|---|---|
| **relevance** | The principle measure in common use for assessing retrieval outcome is relevance, a term having many definitions which tend to cluster around two meanings: relevance as *relatedness* or *aboutness* and as *utility* or *value*. (Boyce, Meadow, and Kraft, 1994, p. 177) |
| **relevance feedback** | Relevance feedback reformulates a query. The user examines the set of retrieved documents and selects which documents are relevant. A selection of terms from these documents is then used to expand the original query. A new set of documents is retrieved with this expanded query and presented to the user.<br>*See also* pseudo relevance feedback |
| **retrospective information retrieval** | [In adhoc information retrieval the user carries out a retrospective search on a document collection. The collection is static and the queries are changing.]<br>*See also* information filtering ; information retrieval |
| **routing** | *See* filtering |
| **self-disambiguation** | [Most queries contain multiple terms and documents must contain all or most of these terms to be retrieved. The combination of all query terms acts as a form of disambiguation since only certain senses of words tend to co-occur together.] |
| **Spearman's rank order correlation coefficient** | Spearman's rho. A symmetrical measure of association based on ranks. (Katzer, Cook, and Crouch, 1991, p. 261) |
| **source language** | [The original language of a query or document.]<br>*See also* target language |
| **specificity** | Depth of indexing. The degree to which an assigned index term is co-extensive with the concept treated in the document. (Wersig and Neveling, 1976, p. 110)<br>*See also* exhaustivity |
| **stem** | A stem is the portion of a word which is left after the removal of its affixes (i.e. prefixes and suffixes). A typical example of a stem is the word *connect* which is the stem for the variants *connected*, *connecting*, *connection*, and *connections*. Stems are thought o be useful for improving retrieval performance because they reduce variants of the same root word to a common concept. (Baeza-Yates and Ribeiro-Neto, 1999, p. 168) |
| **stop list** | [A list of words that] are poor discriminators and cannot possibly be used by themselves to identify document content. In English, about 250 common words are involved, and it is easy to include them in a dictionary, sometimes called a negative dictionary, or *stop list*. (Salton & McGill, 1983, p. 71)<br>*See also* stop word |
| **stop word** | Stop words are terms deemed relatively meaningless in terms of document relevance and are not stored in the index. (Grossman and Frieder, 1998, p. 126)<br>*See also* stop list |
| **structure data** | Structured data is well defined data (facts) typically represented by tables. There s a semantic description associated with each attribute within a table that well defines that attribute. (Kowalski, 1997, p. 18)<br>*See also* free-text |

**synonym** *See* synonymy

**synonymy** A term used in semantics to refer to a major type of sense relation between lexical items; lexical items which have the same meanings are synonyms, and the relationship between them is one of synonymy. For two items to be synonyms, it does not mean that they should be identical in meaning, i.e. interchangeable in all contexts, and with identical connotations-this unlikely possibility is sometimes referred to as 'total synonymy'. Synonymy can be said to occur if items are close enough in their meaning to allow a choice to be made between them in *some* contexts, without there being any difference for the meaning of the sentence as a whole. (Crystal, 1997, p. 376)

**tagger** *See* part-of-speech tagger

**term** A word or phrase used to denote a concept. (Wersig and Neveling, 1976, p. 66) [Once words in query or documents have been processed by the retrieval system they are referred to as terms as opposed to words.]
*See also* word

**target language** [The language in which the query or document need to be translated.]
*See also* source language

**term frequency** $t_{i,k}$ The frequency of occurrence of term type *i* in record *k*. (Boyce, Meadow, and Kraft, 1994, p. 107)
*See also* document frequency ; inverse document frequency weight

**Text REtrieval Conference** In the early 1990's, the United States National Institute of Standards and Technology (NIST), using the text collection created by the United States Defense Advanced Research Project Agency (DARPA), initiated a conference to support the collaboration and technology transfer between academia, industry, and government in the area of text retrieval. The conference, named the Text Retrieval Conference (TREC) aims to improve evaluation methods and measures in the information retrieval domain by increasing the research in information retrieval using relatively large test collections on a variety of datasets. (Grossman and Frieder, 1998, p. 222)

**topic** TREC calls a natural language statement of information need a "topic" to distinguish it from a "query", which is the data structure actually presented to the retrieval system. The topics are formatted using a very simple SGML-style tagging. [TREC topics consist of a title, a description, and a narrative. All three, or a subset, can be used to build queries to pass on to the system.] (http://trec.nist.gov/data/testq_eng.html)
*See also* query

**transfer dictionary** [A transfer dictionary results from processing a machine readable dictionary so that all that is left is a word to word match across languages. A word can thus be transferred automatically to its foreign language equivalent(s).]
*See also* machine-readable dictionary

**translation** Translation is the replacement of a representation of a text in one

| | |
|---|---|
| | language by a representation of an equivalent text in a second language. (Meetham and Hudson, 1969) |
| **TREC** | *See* Text REtrieval Conference |
| **vector space model** | Both the query and each document are represented as vectors in the term space. A measure of the similarity between the two vectors is computed. (Grossman and Frieder, 1998, p. 11) |
| **weighted Boolean** | [Boolean queries return (large) unordered sets of documents. Relevance ranking remedies this problem. In weighted Boolean weights are assigned to facilitate relevance ranking.] A score is assigned such that an initial Boolean query results in a ranking. This is done by associating a weight with each query term so that this weight is sued to compute the similarity coefficient. (Grossman and Frieder, 1998, p. 12)<br>*See also* Boolean retrieval |
| **Wilcoxon signed-ranks test** | Wilcoxon matched-pairs signed-ranks test. A nonparametric significance test of the difference between two related groups based on ranks. (Katzer, Cook, and Crouch, 1991, p. 265) |
| **word** | The smallest entity in a language which can convey a specific meaning by itself. (Wersig and Neveling, 1976, p. 66)<br>*See also* term |
| **World Wide Web** | The World Wide Web is the universe of accessible information available on many computers spread through the world and attached to that gigantic computer network called the Internet. (Newton, 1997, p. 729) |
| **WWW** | *See* World Wide Web |

# Appendix II: Queries

Due to space limitations only a random selection (10%) of all 728 queries is included.[56]

052ESD: sanctions against South Africa.
052DSD: sancties tegen Zuid-Afrika.

052ETD: sanction, South Africa

053ESD: a leveraged buyout valued at or above 200 million dollars.
053DSD: een overname gefinancierd met vreemd vermogen (leveraged buy-out), waarmee 200 miljoen dollar of meer gemoeid is.

053ETD: take-over purchase, buying, taking-over, borrowing, underwriting agreement, finance, fund, defray the costs of, back, financier, strange, odd, unusual, weird, queer, surprised, foreign, exotic, strange, alien, imported, extraneous, introduced, unfamiliar, outside, other, fortune, riches, wealth, property, capital, means, power, capacity, ability, capability, be in a position to, have power to, be capable of, have great influence, be able to, leveraged, buy-out, , million, million, a million, one million, one million, dollar, buck, greenback, lake, loch, lough, more, -er, rather, further, anymore, no more, longer, any longer, more often, more

053EST: Leveraged Buyouts
053DST: Leveraged buy-outs
053ETT: leveraged, buy-outs

054EST: Satellite Launch Contracts
054DST: Overeenkomst lancering commerciële satelliet
054ETT: similarity, resemblance, likeness, correspondence, analogy, identity, match, sameness, equality, conformity, agreement, accord, accordance, concordance, unison, harmony, deal, bargain, launch, launching, blast-off, lift-off, commercial, satellite, moon, secondary planet

060ESD: either one or both sides of the controversy over the use of standards of performance to determine salary levels and incentive pay as contrasted with determining pay solely on the basis of seniority or longevity on the job.
060DSD: één of beide gezichtspunten in de controverse over het gebruik van prestatienormen om het salarisniveau te bepalen - het zogenaamde prestatieloon, in tegenstelling tot loonhoogte uitsluitend gekoppeld aan anciënniteit of aantal dienstjaren.

060ETD: one, both, either, either one, two, point of view, angle, aspect, perspective, viewpoint, controversy, polemic, dispute, argument, debate, use, application, consumption, taking, custom, habit, practice, usage, prestatienormen, salarisniveau, prescribe, lay down, determine, set, fix, stipulate, define, decide, ascertain, fix on, concentrate, concentrate on, qualify, modify, zich bepalen, confine, confine oneself to, restrict, restrict oneself to, so-called, supposed, alleged, so-

[56] The query naming convention is as follows: first three digits (051-450) = query number, EST=English source title, ESD=English source description, DST=Dutch source title, DSD=Dutch source description, ETT=English target title, ETD=English target description.

called, merit pay, antithesis, setoff, relief, opposite, foil, contrast, discrepancy, chasm, contradistinction, loonhoogte, exclusive, sole, pure, joined, linked, connected, coupled, seniority, number, year of service, seniority, working year, official year, financial year, fiscal year

069ESD: an attempt by the U.S. House of Representatives or a European country to revive the SALT II Treaty ceilings on weapons in order to limit President Reagan's military build up.
069DSD: pogingen door het Amerikaanse Huis van Afgevaardigden of een Europese regering om de bij SALT II vastgestelde limieten aan het wapenarsenaal te respecteren, om zo de militaire plannen van President Reagan te beteugelen.

069ETD: attempt, try, bid, effort, endeavour, crack, go, shot, American, American woman, American girl, American, American, house, residence, domicile, building, premises, theatre, home, establishment, firm, concern, case, casing, shell, House, Family, dynasty, court, household, delegate, representative, member, member of parliament, European, European woman, European girl, European, government, administration, rule, reign, regimen, SALT, ii, fix, determine, settle, arrange, appoint, assign, decide, decide on, decree, specify, lay down, provide, provide for, enact, set down, find, state, record, diagnose, assess, establish, ascertain, limit, verge, edge, reserve price, upset price, arsenal, ordnance, depot, arms depot, armoury, respect, revere, admire, appreciate, regard, defer to, observe, honour, soldier, serviceman, military man, militia man, troops, the troops, the military, military, army, war, warlike, armed, soldiers', in a military fashion, in a military way, plan, plan, scheme, project, strategy, blueprint, intention, design, plane, level, foreground, background, map, ground plan, floor plan, president, chairman, chairwoman, foreman, presiding judge, President, reagan, curb, check, suppress, rein in, control, restrain

074ESD: an instance in which the U.S. government propounds two conflicting or opposing policies.
074DSD: een kwestie waarin de Amerikaanse regering twee strijdige of tegengestelde beleidsdoelen hanteert.

074ETD: question, matter, issue, argument, dispute, American, American woman, American girl, American, American, government, administration, rule, reign, regimen, two, second, two, contrary, contratry to, adverse, adverse to, opposed, opposed to, inconsistent, inconsistent with, conflicting, incompatible, incompatible with, opposite, contrary, antithesis, opposite, antipodal, contrastive, incompatible, in the opposite direction, incompatibly, beleidsdoelen, handle, work, operate, employ, ply wield, manage, manipulate, manoeuvre maneuver

077EST: Poaching
077DST: Stroperij
077ETT: poaching, theft of crops, theft of growing crops, stealing crops, stealing growing crops

081ESD: a loss of revenue of a televangelist in the aftermath of the PTL scandal, or a financial crisis triggered by the scandal.
081DSD: inkomstenderving van een televisiedominee ten gevolge van het schandaal bij de PTL, of een financiële crisis veroorzaakt door dat schandaal.

081ETD: inkomstenderving, TV evangelist, consequence, result, effect, outcome, success, retinue, train, corollary, scandal, outrage, shame, disgrace, crime, ptl, financial, pecuniary, monetary, problem, critical stage, depression, slump, cause, bring about, bring on, occasion, create, produce, raise, scandal, outrage, shame, disgrace, crime

098ESD: individuals or organizations which produce fiber optics equipment.
098DSD: personen of organisaties die apparatuur voor de glasvezeltechnologie produceren.

098ETD: person, individual, people, role, character, figure, dramatis personae, organization, arrangement, system, set-up, society, association, outfit, apparatus, equipment, machinery, hardware, paraphernalia, glasvezeltechnologie, produce, make, manufacture, turn out, churn out, put out, generate

102EST: Laser Research Applicable to the U.S.'s Strategic Defense Initiative
102DST: Onderzoek naar laserstralen, dat bruikbaar kan zijn voor het Amerikaanse Strategic Defense Initiative
102ETT: investigation, examination, study, search, scrutiny, survey, inspection, inquiry, research, exploration, inquest, inquisition, check-up, test, check, nasty, grim, horrible, horrid, dismal, sick, ill, foul, unpleasant, as, laser beam, usable, useful, practicable, serviceable, employable, American, American woman, American girl, American, American, strategic, defense, initiative

107ESD: Japan's regulation of insider trading.
107DSD: de reglementering van handel met voorkennis in Japan.

107ETD: regulation, trade, trading, business, commerce, traffic, trafficking, transaction, deal, merchandise, goods, market, dealing, traders, dealers, commercial community, business community, shop, store, foreknowledge, prescience, Japan

112EST: Funding Biotechnology
112DST: Financiering van biotechnologie
112ETT: financing, funding, bioengineering, biotechnology

123ESD: studies into linkages between environmental factors or chemicals which might cause cancer, governmental actions to identify, control, or limit exposure to those factors or chemicals which have been shown to be carcinogenic.
123DSD: onderzoek naar een eventueel verband tussen omgevingsfactoren of chemische substanties en kanker, maatregelen genomen door de overheid om blootstelling aan dergelijke carcinogene factoren of stoffen op te sporen, te beheersen of te beperken.

123ETD: investigation, examination, study, search, scrutiny, survey, inspection, inquiry, research, exploration, inquest, inquisition, check-up, test, check, nasty, grim, horrible, horrid, dismal, sick, ill, foul, unpleasant, as, possibly, if necessary, if desired, if so desired, alternatively, any, any possible, such as, potential, bandage, dressing, sling, connection, bond, joint, correlation, link, context, coherence, cohesion, relation, relationship, association, level, contract, engagement, agreement, binding agreement, sanitary towel, sanitary napkin, security, surety, omgevingsfactoren, chemical, substance, material, matter, main point, main thing, essence, cancer, carcinoma, canker, cankerous growth, measure, step, move, enactment, proceeding, taken in, fooled, had, taken, done, take, put, have, get, take out, use, seize, capture, government, authorities, authority, council, corporation, the powers that be, blootstelling, similar, like, the like, such, such like, carcinogen, carcinogenic, factor, circumstance, influence, agent, cloth, fabric, textile, dust, brag, boast, show off, track, go by rail, go by train, travel by rail, travel by train, train it, spur, track, control, govern, rule, have control over have sway over sway, dominate, have a command of, to have a thorough command of, have mastered, limit, restrict, restrict to, limit to, confine, confine

to, keep, keep to, reduce, decrease, cut, cut down, cut down on, curtail, zich beperken, limit, limit oneself to, restrict, restrict oneself to, confine, confine oneself to

123EST: Research into & Control of Carcinogens
123DST: Onderzoek naar en beheersing van carcinogenen
123ETT: investigation, examination, study, search, scrutiny, survey, inspection, inquiry, research, exploration, inquest, inquisition, check-up, test, check, nasty, grim, horrible, horrid, dismal, sick, ill, foul, unpleasant, as, control, command, domination, check, carcinogen, carcinogenic

130ESD: the issue of Jewish emigration from the Soviet Union as it impacts on U.S.-Soviet relations.
130DSD: de emigratie van Joden uit de Sovjet-Unie, voorzover die van invloed is op de Amerikaans-Russische betrekkingen.

130ETD: emigration, Jew, jew, iodine, iodic, out, out, away, over, gone out, dead, Soviet Union, voorzover, influence, effect, domination, impact, weight, authority, induction, American, American, Russian, Russian, post, job, position, office, situation, relationship, relation, connection, reference, bearing, relative, kinsman, kinswoman, purchase, obtaining, buying, derivation, recruitment, moving in, occupation

133ESD: some design feature of the Hubble Space Telescope.
133DSD: een bepaald onderdeel van de Hubble ruimtetelescoop.

133ETD: particular, specific, fixed, set, specified, given, certain, definite, part, division, subdivision, branch, discipline, unit, spare, spare part, component, hubble, space telescope

148ESD: the Ethiopia-Somalia War, civil wars within those nations, the movement of refugees fleeing armed conflicts between or within Ethiopia and Somalia.
148DSD: de oorlog tussen Ethiopië en Somalië, de burgeroorlogen in beide landen, de vluchtelingenstromen veroorzaakt door de gewapende conflicten in of tussen Ethiopië en Somalië.

148ETD: war, Ethiopia, Abyssinia, Somalia, civil war, both, either, either one, two, land, touch down, splash down, land, disembark, stream of refugees, cause, bring about, bring on, occasion, create, produce, raise, armed, in arms, armoured, reinforced, assisted, protected, prepared, conflict, clash, Ethiopia, Abyssinia, ], Somalia

164EST: Generic Drugs - Illegal Activities by Manufacturers
164DST: Merkloze geneesmiddelen - Illegale activiteiten van producenten
164ETT: unbranded, no-brand, non-brand, medicine, drug, remedy, member of the resistance, member of the resistance movement, underground worker, resistance worker, illegal alien, illegal, unlawful, underground, activity, bustle, liveliness, action, active service, radioactivity, producer, maker

168EST: Financing AMTRAK
168DST: Financiering AMTRAK
168ETT: financing, funding, amtrak

172ESD: the effectiveness of the use of medical products and related psychological/psychiatric services in the cessation of smoking.

172DSD: de effectiviteit van medicamenten en daarmee samenhangende psychologische/psychiatrische hulpverlening bij het stoppen met roken.

172ETD: effectiveness, effectivity, efficiency, medicament, medicine, medication, remedy, drug, be connected, be linked, cohere, psychological, tactful, diplomatic, psychiatric, mental, assistance, aid, relief, fill, fill up, stuff, put, put in, put into, stick, stick in, stick into, stuff in, stuff into, pop, stop, halt, bring to a stop, bring to a halt, bring to a standstill, darn, mend, bind, bind the bowels, stop diarrhoea, cause constipation, stop, halt, come to a stop, come to a halt, come to a standstill, draw to a stop, draw to a halt, draw to a standstill, draw up, pull up, stopping, cease, stopper, plug, bung, fuse, cutout, darn, mend, stop, break, layover, freeze, stop, halt, ho, stop it, hold it, enough, avast, smoke, puff, puff at, smoke, steam, cure, smoke, cure, bloat

173EST: Smoking Bans
173DST: Rookverbod
173ETT: smoking ban, ban on smoking

176EST: Real-life private investigators
176DST: Privé-detectives bestaan echt
176ETT: privé-detectives, existence, living, livelihood, exist, be, be in existence, consist, consist in, consist of, include, be made up, be made up of, live, be possible, be kindred, dare, matrimony, wedlock, real, genuine, authentic, true, actual, regular, trueblue, trueborn, perfect, downright, thorough, veritable, legitimate, really, truly, genuinely, honestly, heartily, real, genuine

189ESD: a murderer's motive for killing a person or persons in a true case.
189DSD: het motief voor een moord op een of meer personen, zoals door de moordenaar zelf opgegeven.

189ETD: motive, cause, reason, ground, incentive, stimulus, motif, theme, design, pattern, figure, subject, murder, killing, assassination, homicide, gone, lake, loch, lough, more, -er, rather, further, anymore, no more, longer, any longer, more often, more, person, individual, people, role, character, figure, dramatis personae, murderer, murderess, killer, assassin, homicide, butcher slaughterer, cutthroat, give up, abandon, drop, quit, throw up, give, state, mention, report, return, set, assign, ask, propound, enter, bring up, spit, vomit, hawk up, cough up, expectorate, hand over, surrender, yield, yield up

190EST: Instances of Fraud Involving the Use of a Computer
190DST: Gevallen van computerfraude
190ETT: fallen, come about, come to pass, happen, case, affair, circumstances, position, situation, thing, contraption, device, contrivance, chance, luck, computer fraud

191ESD: an attempt or idea within the U.S. to improve student performance at any level (K through post-graduate) by any means (teacher pay, new equipment, new methods, new incentives, etc.).
191DSD: pogingen tot of ideeën omtrent het verbeteren van de schoolprestaties in Amerika, op alle niveaus (van K tot en met postuniversitair) op welke manier ook (betere salariëring van leerkrachten, nieuwe apparatuur, nieuwe methoden, nieuwe prikkels etc.).

191ETD: attempt, try, bid, effort, endeavour, crack, go, shot, idea, conception, notion, ideal, view, principle, notion, concept, conception, view, opinion, conception, plan, scheme, improve, get better,

improve, better, ameliorate, reform, reclaim, correct, amend, rectify, remedy, emend, revise, beat, improve on, schoolprestaties, America, gone, level, standard, water level, layer, stratum, postgraduate, gone, way, manner, mode, style, fashion, manners, breeding, habit, something better, anything better, better, recovered, well again, better class of, superior, better, payment, teacher, instructor, new things, new clothes, new, recent, unworn, unused, fresh, young, original, novel, fashionable, modern, up-to-date, up to the minute, apparatus, equipment, machinery, hardware, paraphernalia, new things, new clothes, new, recent, unworn, unused, fresh, young, original, novel, fashionable, modern, up-to-date, up to the minute, method, system, manual, primer, new things, new clothes, new, recent, unworn, unused, fresh, young, original, novel, fashionable, modern, up-to-date, up to the minute, tingling, stimulus, goad, incentive, stimulant, spur, prickle, thorn, & c

191EST: Efforts to Improve U.S. Schooling
191DST: Pogingen om het onderwijs in de VS te verbeteren
191ETT: attempt, try, bid, effort, endeavour, crack, go, shot, education, teaching, instruction, the field of education, Education, v, vs, US, USA, improve, get better, improve, better, ameliorate, reform, reclaim, correct, amend, rectify, remedy, emend, revise, beat, improve on

196EST: School Choice Voucher System and its effects upon the entire U.S. educational program
196DST: 'School Choice Voucher' systeem en de effecten daarvan op het gehele Amerikaanse onderwijssysteem
196ETT: school, education, schooling, mesh, choice, coupon, ticket, voucher, system, classification, method, effect, result, outcome, consequence, spin, stuff, twist, curve, slice, side, power, stock, share, security, gone, whole, entity, unit, unity, entirety, aggregate, entirely, fully, completely, totally, entire, whole, complete, full, American, American woman, American girl, American, American, onderwijssysteem

199ESD: a suicide that is effected through the assistance of a medically competent person -- doctor, nurse, medical technician etc. -- and the legality of such an assisted action.
199DSD: een geval van zelfdoding met behulp van een medisch deskundige -- arts, verpleegster, medisch specialist -- en de rechtmatigheid van zo'n handeling.

199ETD: case, affair, circumstances, position, situation, thing, contraption, device, contrivance, chance, luck, suicide, killing oneself, behulp, medical, medical, expert, expert in, expert at, professional, doctor, physician, nurse, male nurse, orderly, medical, medical, specialist, expert, authority, rightfulness, lawfulness, legitimacy, act, deed, dealings, doings, manoeuvre, operation, proceedings, transactions, report, minutes, discussion, deliberation, consultation, action, plot

208ESD: What are the latest developments in bioconversion -- the conversion of biological waste, garbage, and plant material into energy, fertilizer, and other useful products?
208DSD: Wat zijn de nieuwste ontwikkelingen op het gebied van bioconversie -- het omzetten van biologisch afval en plantaardig materiaal in energie, kunstmest en andere nuttige producten?

208ETD: new things, new clothes, new, recent, unworn, unused, fresh, young, original, novel, fashionable, modern, up-to-date, up to the minute, development, growth, maturation, generation, education, current, expansion, territory, domain, dominion, ground, home ground, area, district, region, zone, field, sphere, department, province, soil, land, command, sway, authority, rule, bioconversie, go round, go around, come round, come around, go running round, go running around, go racing round, go racing around, come running round, come running around, come racing round, come racing around, change, change position of, transpose, turn, turn over, reverse,

pull, pull over, move, shift, sell, convert, convert into, turn into, transform, transform into, transmute, realize, cash, transpose into, biological, organic, macrobiotic, defection, secession, apostasy, waste, waste matter, residue, refuse, rubbish, litter, trash, garbage, scraps, leavings, kitchen waste, spoil, vegetable, plant, material, materials, data, evidence, tools, energy, vigour, spirit, drive, go, power, fertilizer, artificial fertilizer, chemical manure, other, another, different, another, others, another matter, another thing, other matters, other things, next, other, useful, advantageous, profitable, efficient, product, production, commodity, exhibit

223ESD: What was responsible for the great emergence of "MICROSOFT" in the computer industry?
223DSD: Hoe valt de onstuitbare opkomst van Microsoft in de computerindustrie te verklaren?

223ETD: fall, drop, come down, go down, tumble, take a fall, take a spill, come a cropper, fall over, topple over, trip up, stumble, come, land, hang, fall in battle, be killed, be slain, be a failure, go, go for, take, take to, fancy, slope, slope down, subside, abate, die down, run, run into, join, meet, be laid, be played, unstoppable, irrepressible, rampant, rise, ascension, attendance, turnout, entrance, entry, emergence, appearance, development, initial development, rising, boom, ascent, origin, infancy, origination, call-up, enlistment, mobilisation, microsoft, computerindustrie, explain, make clear, account for, explicate, elucidate, declare, state, pronounce, certify, attest, zich verklaren, declare, declare oneself, explain oneself

230ESD: Is the automobile industry making an honest effort to develop and produce an electric-powered automobile?
230DSD: Houdt de autoindustrie zich serieus bezig met de ontwikkeling en productie van een elektrische auto?

230ETD: love, adore, idolize, like, be fond of, be partial to, care for, have a liking for, hold, stick, keep, retain, retain preserve, hold, keep on retain, maintain preserve, maintain, celebrate, observe, organize, give, hold to be, take to be, regard as, consider to be, consider as, take, stand, endure, zich houden, keep to, adhere to, abide by, comply with, observe, stick to, keep, pretend to be, sham, car industry, auto-industry, motor industry, serious, grave, earnest, straight, busy, busy with, busying, working, working on, occupied, occupied with, engaged, engaged in, industrious, development, growth, maturation, generation, education, current, expansion, production, manufacture, output, yield, produce, exhibit, electrically, electric, electrical, car, motor, motorcar, auto, automobile

234ESD: What progress has been made in fuel cell technology?
234DSD: Welke vorderingen in de brandstofceltechnologie?

234ETD: progress, advance, headway, demand, claim, account receivable, receivables, requisitioning, commandeering, brandstofceltechnologie

239ESD: Are there certain regions in the United States where specific cancers seem to be concentrated? What conditions exist that might cause this problem?
239DSD: Zijn er gebieden in de VS waar bepaalde soorten kanker blijkbaar veel vaker voorkomen dan in andere? Zijn er oorzaken aan te wijzen?

239ETD: territory, domain, dominion, ground, home ground, area, district, region, zone, field, sphere, department, province, soil, land, command, sway, authority, rule, v, vs, US, USA, goods,

ware, wares, articles, merchandise, produce, stuff, true, real, actual, veritable, genuine, correct, proper, right, whereas, where, what, that, which, wherever, everywhere, anywhere, since, as, particular, specific, fixed, set, specified, given, certain, definite, species, variety, sort, kind, type, variety, genre, class, quality, calibre, sort of, kind of, type of, cancer, carcinoma, canker, cankerous growth, evident, obvious, clear, manifest, apparently, evidently, much, a lot, a great deal, much, many, a lot, lots, plenty, many, a lot, a great deal, a great many, sleep, sleepiness, drowsiness, often, frequently, prevent, avert, preclude, obviate, anticipate, appearance, look, looks, aspect, air, presence, bearing, occurrence, incidence, get ahead, draw ahead, draw out in front, occur, happen, be found with, be met with, appear, be brought up, come before, come on, come up, seem, look to, come round, drive up, other, another, different, another, others, another matter, another thing, other matters, other things, next, other, cause, origin, source, root, about, around, away, on, against, on to, point, show, indicate, point, point to, zich wijzen, show, appear, be obvious, be apparent, be evident, become obvious, become apparent, become evident, show, point out, pronounce, give, pass

241ESD: Doctor and lawyer groups considering penalties against members of their professions for malfeasance and show the results of such investigations.
241DSD: Gevallen waarin verenigingen van artsen of advocaten sancties tegen vakgenoten overwegen in verband met ambtsmisdrijven. Welke resultaten levert dergelijk onderzoek op?

241ETD: fallen, come about, come to pass, happen, case, affair, circumstances, position, situation, thing, contraption, device, contrivance, chance, luck, club, association, society, union, guild, fellowship, combination, joining, junction, amalgamation, doctor, physician, lawyer, barrister, solicitor, attorney, advocate, counsel, supporter, advocaat, eggnog, eggflip, sanction, colleague, fellow craftsman, fellow worker, confrere, weigh again, reweigh, be overweight, give overweight, predominate, preponderate, prevail, consider, weigh, weigh up, think over, think out, ponder, contemplate, bandage, dressing, sling, connection, bond, joint, correlation, link, context, coherence, cohesion, relation, relationship, association, level, contract, engagement, agreement, binding agreement, sanitary towel, sanitary napkin, security, surety, ambtsmisdrijven, result, effect, issue, outcome, upshot, end, fruits, returns, supply, furnish, deliver, provide, give, produce, fix, do, bring off, do to, investigation, examination, study, search, scrutiny, survey, inspection, inquiry, research, exploration, inquest, inquisition, check-up, test, check

249ESD: How has the depletion or destruction of the rain forest effected the worlds weather?
249DSD: Welke invloed heeft de ontbossing of de vernietiging van het tropisch regenwoud op het weer in de wereld?

249ETD: influence, effect, domination, impact, weight, authority, induction, deforestation, disforestation, disafforestation, destruction, devastation, ruin, wrecking, dashing, annihilation, obliteration, nullification, annulment, reversal, quashing, rescission, cancellation, tropical, rain forest, wether, weather, weathering, callus, resistance, defence, defense, again, once more, once again, back, world, earth

251EST: Exportation of Industry
251DST: Verplaatsing van nijverheid
251ETT: moving, movement, removal, transfer, transference, shifting, transposition, displacement, move, relocation, permutation, industry

270ESD: Should the Food and Drug Administration (FDA) exercise more stringent control over the labelling and sale of food supplements.
270DSD: Moet de Food and Drug Administration (FDA) een striktere controle uitvoeren op etikettering en verkoop van voedingssupplementen?

270ETD: food, and, drug, narcotic, administration, fda, strict, stringent, rigorous, precise, check, check on, checking, control, supervision, control of, supervision of, control over, supervision over, surveillance, verification, inspection, examination, checkup, medical, monitoring, audit, auditing, screening, control, control point, checkpoint, turnstile, gate, ticket gate, ticket barrier, ticket box, export, do, execute, perform, carry out, implement, enforce, produce, finish, etikettering, sale, sales, voedingssupplementen

273ESD: There has been a significant, noticeable increase in volcanic and seismic (earthquake) activity.
273DSD: Is er Een duidelijke en opvallende toename van vulkanische en seismische activiteit (aardbevingen)

273ETD: clear, clear-cut, plain, obvious, evident, apparent, marked, broad, explicit, distinct, striking, conspicuous, marked, notable, noticeable, eye-catching, increase, growth, rise, volcanic, igneous, eruptive, violent, explosive, fiery, seismic, activity, bustle, liveliness, action, active service, radioactivity, ), earthquake, tremor, earth tremor, seism, quake

276ESD: The pros and cons of students wearing a school uniform or adhering to a dress code.
276DSD: Argumenten voor en tegen het dragen van een schooluniform of kledingvoorschriften op scholen.

276ETD: rest on, be supported, be carried, be born, be borne, carry, run, suppurate, discharge, be carrying, be pregnant, support, bear, carry, buoy up, sustain, wear, have on, be pregnant, yield, take, have, endure, schooluniform, kledingvoorschriften, flock, flock together, school, school, teach, instruct, train, drill, school, education, schooling, mesh

285EST: World submarine forces
285DST: Onderzeese strijdkrachten
285ETT: undersea, submarine, underwater, forces, services, armed forces, armed services

287EST: Electronic Surveillance
287DST: Elektronisch toezicht
287ETT: electronic, supervision, surveillance, inspection

288ESD: Weight control and diets in the U.S.
288DSD: Pogingen tot afvallen en diëten in de VS.

288ETD: attempt, try, bid, effort, endeavour, crack, go, shot, fall down, fall off, drop out, desert, abandon, defect, secede, lose weight, slim, waste, waste away, lose flesh, be left, be left over, be disappointing, not come up to one's expectations, not live up to one's expectations, be a disappointment, be a let-down, bear away, diet, regime, regimen, v, vs, US, USA

312EST: Hydroponics
312DST: Hydrocultuur

312ETT: hydroponics, aquiculture, water culture, sand culture

314ESD: Commercial harvesting of marine vegetation such as algae, seaweed and kelp for food and drug purposes.
314DSD: Op commerciële schaal winnen van planten uit zee, zoals algen, zeewier en kelp, om te dienen als voedsel of medicijn.

314ETD: commercial, scale, dish, plate, scales, balance, shell, win, gain, produce, obtain, mine, extract, enlist, secure, win, gain, make a profit, plant, plant out, grow, cultivate, plant, out, out, away, over, gone out, dead, after, in bloom, sea, ocean, the waves briny, the deep, the briny deep, the brine, flood, torrent, wave, alga, seaweed, wrack, sea wrack, kelp, seaware, eelgrass, kelp, varec, wrack, serve, serve as, serve for, be used as, used as, be, food, nourishment, fuel, medicine, medication, medicament, drug

322EST: International Art Crime
322DST: Misdaad in de internationale kunstwereld
322ETT: crime, criminal act, offence, criminality, criminal practices, outrage, moral offence, International, international, art world, world of art, art circles, artistic circles

328EST: Pope Beatifications
328DST: Zaligverklaringen Paus
328ETT: beatification, pope

329ESD: Mexico City has the worst air pollution in the world. The specific steps Mexican authorities have taken to combat this deplorable situation.
329DSD: Mexico Stad heeft de ergste luchtvervuiling van alle steden ter wereld. De concrete stappen die de Mexicaanse autoriteten hebben gezet om deze vreselijke toestand te bestrijden.

329ETD: Mexico, town, city, borough, town council, city council, misgiving, misgivings, notion, inkling, suspicion, intention, evil intention, evil intentions, erg, awful, terrible, dreadful, bad, lamentable, regrettable, deplorable, serious, critical, awfully, dreadfully, terribly, very, badly, air pollution, town, city, borough, town council, city council, world, earth, concrete, material, real, actual, tangible, definite, specific, particular, clearly, concretely, in a concrete manner, definitely, step, walk, stride, strut, march, tramp, step up, step down, pace, go out, go for a drink, spend the evening out, Mexican, Mexican woman, Mexican girl, Mexican, authority, set, definite, fixed, regular, stout, corpulent, portly, rotund, plump, thickset, regular, seat, set, put, place, move, fix, make, put on, brew, set to, start, settle, compose, set up, arrange, terribly, awfully, frightfully, enormously, shockingly, dreadfully, terrible, awful, frightful, enormous, dreadful, terrifying, horrible, fearful, ghastly, shocking, terrible, dreadful, shocking, awful, state, condition, situation, position, commotion, to-do, fuss, muddle, affair, dispute, challenge, contest, oppose, resist, combat, fight, suppress, counteract, control, contend, contend with

332ESD: Investigations that have targeted evaders of U.S. income tax.
332DSD: Welke onderzoeken zijn er gedaan om mensen op te sporen die de Amerikaanse inkomstenbelasting hebben ontdoken?

332ETD: investigation, examination, study, search, scrutiny, survey, inspection, inquiry, research, exploration, inquest, inquisition, check-up, test, check, done, finished, over, over, over with, fired, sacked, het, do, act, behave, be, deal, trade, do, make, take, put, go for, cost, clean, visit, human,

human being, man, mankind, humankind, people, person, folks, everybody, thing, creature, soul, track, go by rail, go by train, travel by rail, travel by train, train it, spur, track, track, print, footprint, trail, mark, spoor, trace, vestige, rails, railway, railway company, railroad, railroad company, rail, train, gauge, spur, gaff, calcar, spore, American, American girl, American woman, American, American, income tax, evade, elude, dodge, circumvent

339EST: Alzheimer's Drug Treatment
339DST: Medicijnen tegen Alzheimer
339ETT: medicine, medication, medicament, drug, Alzheimer's

345EST: Overseas Tobacco Sales
345DST: Verkoop van tabak in het buitenland
345ETT: sale, sales, tobacco, tobacco plant, foreign country, foreign countries, land outside of the dike, land outside of the dikes

350ESD: Is it hazardous to the health of individuals to work with computer terminals on a daily basis?
350DSD: Brengt dagelijks werken aan een computer terminal risico's voor de gezondheid met zich mee?

350ETD: bring, take, give, perform, present, send, put, drive, daily, diurnal, circadian, everyday, ordinary, day-to-day, common-or-garden, run-of-the-mill, workaday, daily, each day, every day, work, operate, function, run, act, take effect, warp, distort, settle, clean, char, work out, ferment, about, around, away, on, against, on to, computer, terminal, risk, hazard, chance, health, fitness, well-being, healthiness, wholesomeness, salubrity, bless you!, God bless you!

351EST: Falkland petroleum exploration
351DST: Oliewinning bij de Britse Falklands
351ETT: extraction of oil, extraction petroleum, recovery of oil, recovery of petroleum, falklands

363EST: transportation tunnel disasters
363DST: rampen in tunnels voor vervoer
363ETT: disaster, calamity, catastrophe, tunnel, underpass, fly-under, subway, transport, conveyance, haulage, transportation

366EST: commercial cyanide uses
366DST: commerciële toepassingen van cyanide
366ETT: commercial, use, employment, utilization, application, adoption, practice, administration, implementation, enforcement, cyanide, prussiate

368ESD: In vitro fertilization.
368DSD: In-vitrofertilisatie.

368ETD: in vitro fertilization

372EST: Native American casino
372DST: Indiaanse casino's
372ETT: Indian, American Indian, American Indian woman, American Indian girl, casino, kursaal, club, club-house, white tin-loaf

373EST: encryption equipment export
373DST: export van apparatuur voor encryptie
373ETT: export, apparatus, equipment, machinery, hardware, paraphernalia, encryptie

377ESD: The renewed popularity of cigar smoking.
377DSD: De hernieuwde populariteit van het roken van sigaren.

377ETD: renew, revive, resume, regenerate, renovate, reinvigorate, refresh, popularity, public favour, smoke, puff, puff at, smoke, steam, cure, smoke, cure, bloat, cigar, ticking off, telling off, scolding, dressing-down, reed mace, cat's tail, cattail spike

377EST: cigar smoking
377DST: roken van sigaren
377ETT: smoke, puff, puff at, smoke, steam, cure, smoke, cure, bloat, cigar, ticking off, telling off, scolding, dressing-down, reed mace, cat's tail, cattail spike

378ESD: Opposition to the introduction of the euro, the European currency.
378DSD: Het verzet tegen de invoering van de Euro, de nieuwe Europese munteenheid.

378ETD: resistance, protest, revolt, opposition, diversion, underground, gear ratio, introduction, institution, adoption, input, presentation, import, importation, Euro, new things, new clothes, new, recent, unworn, unused, fresh, young, original, novel, fashionable, modern, up-to-date, up to the minute, European, European woman, European girl, European, monetary unit, currency unit

388EST: organic soil enhancement
388DST: biologische meststoffen
388ETT: biological, organic, macrobiotic, fertilizer

394ESD: The education of children at home (home schooling).
394DSD: Het thuis lesgeven aan kinderen.

394ETD: teach, give lessons, give lessons in, about, around, away, on, against, on to, child, baby, infant, girl, thing, lass, dear fellow dear girl

396ESD: Sick building syndrome or building-related illnesses.
396DSD: Het sickbuildingsyndroom of andere met gebouwen samenhangende ziekten.

396ETD: sick building syndrome, other, another, different, another, others, another matter, another thing, other matters, other things, next, other, building, structure, construction, premises, hall, house, construction works, fabric, edifice, be connected, be linked, cohere, illness, sickness, disease, disorder

399ESD: The activities or equipment of oceanographic vessels.
399DSD: Activiteiten of apparatuur aan boord van oceanografische schepen.

399ETD: activity, bustle, liveliness, action, active service, radioactivity, apparatus, equipment, machinery, hardware, paraphernalia, about, around, away, on, against, on to, border, band, trim, collar, board, freeboard, edge, bank, shore, edge of a road, edge of the road, side of a road, side of

the road, shoulder of a road, shoulder of the road, roadside, verge, brim, leafblade, lamina, oceanographic, oceanographical, sheriff, alderman, ship, take on board

405ESD: What unexpected or unexplained cosmic events or celestial phenomena, such as radiation and supernova outbursts or new comets, have been detected?
405DSD: Welke onverwachte of onverklaarbare verschijnselen of fenomenen, zoals straling en ontploffingen van supernova's of nieuwe kometen zijn er aan de hemel of in de kosmos waargenomen?

405ETD: unexpected, unforeseen, unlooked-for, surprise, sudden, inexplicable, unaccountable, insolvable, phenomenon, symptom, sign, phenomenon, radiation, explosion, bang, supernova, new things, new clothes, new, recent, unworn, unused, fresh, young, original, novel, fashionable, modern, up-to-date, up to the minute, comet, about, around, away, on, against, on to, heaven, heavens, sky, Heaven, canopy, tester, baldachin, cosmos, universe, space, deep space, outer space, observe, perceive, discern, detect, notice, use, take, make the most of, fulfil, fulfill, do, perform, take hold of, catch hold of, replace, replace temporarily, fill in, take over, take over temporarily, deputize, deputize for, supply, act, hold, have, occupy, fill

409EST: legal, Pan Am, 103
409DST: Pan Am, vlucht 103, rechtszaken
409ETT: pan, tile, pantile, muddle, mess, tip, hollow, dip, a, m, flight, escape, flock, bevy, covey, gaggle, skein, span, wingspread, wingspan, reveal, lawsuit, legal business, legal matters, legal cases

411EST: salvaging, shipwreck, treasure
411DST: schatten in scheepswrakken
411ETT: value, rate, estimate, assess, appraise, put value on, set a value on, consider, deem, scheepswrakken

414EST: Cuba, sugar, exports
414DST: Cuba, suikerexport
414ETT: Cuba, suikerexport

420ESD: How widespread is carbon monoxide poisoning on a global scale?
420DSD: Hoe vaak komt koolmonoxidevergiftiging, over de hele wereld bekeken, voor?

420ETD: sleep, sleepiness, drowsiness, often, frequently, come, get, come round, come around, come over, call, touch, come about, happen, strike, come upon, come by, get hold of, move, be added, carbon monoxide poisoning, over, over, past, finished, across, over, left, remaining, spare, surplus, intact, whole, undamaged, decent, respectable, entire, unbroken, complete, full, quite a, quite some, some a hell of a, some one hell of a, all, very, very much, really most, completely, entirely, wholly, altogether, absolutely, at all, world, earth, settled, thrashed out, well-judged, deliberately, deftly

424EST: suicides
424DST: zelfdoding
424ETT: suicide, killing oneself

429EST: Legionnaires' disease

429DST: legionairsziekte
429ETT: Legionnaire's disease

433ESD: Is there contemporary interest in the Greek philosophy of stoicism?
433DSD: Bestaat er tegenwoordig nog belangstelling voor de Griekse leer van het stoïcisme?

433ETD: existence, living, livelihood, exist, be, be in existence, consist, consist in, consist of, include, be made up, be made up of, live, be possible, be kindred, dare, now, nowadays, these days, today, currently, latterly, interest, interest in, ; alleen in 1, Greek, Greek woman, Greek girl, Greek restaurant, Greek, Greek, Grecian, Hellenic, Byzantine Greek, Byzantine, Greek Orthodox, science, theory, principles, ism, doctrine, teachings, creed, faith, apprenticeship, lesson, ladder, stepladder, leather, football, pigskin, Stoicism, stoicism, resignation, impassiveness

439EST: inventions, scientific discoveries
439DST: uitvindingen, wetenschappelijke ontdekkingen
439ETT: invention, concoction, gadget, contraption, contrivance, scholarly, scientific, learned, discovery, find

442EST: heroic acts
442DST: heldhaftig optreden
442ETT: heroic, valiant, action, way of acting, behaviour, attitude, manner, bearing, demeanour, appearance, performance, show, appear, make one's appearance, enter, go on, come on, perform, act, act as, serve, serve as, pretend, pretend to be, occur, take action, proceed