Aspects of Swedish Morphology and Semantics from the Perspective of Mono- and Cross-language Information Retrieval.

Turid Hedlund*, Ari Pirkola and Kalervo Järvelin
Dept. of Information Studies
University of Tampere
P.O. Box 607
FIN-33101 Tampere, Finland

*and Swedish School of Economics and Business Administration
Library
P.O. Box 479
FIN-00101 Helsinki, Finland

Corresponding author:
Turid Hedlund
Swedish School of Economics and Business Administration
Library
P.O. Box 479
FIN-00101 Helsinki, Finland

phone: +358-9-43133378
fax: +358-9-43133425
email: turid.hedlund@shh.fi
Abstract

This paper analyzes the features of Swedish language from the viewpoint of mono- and cross-language information retrieval (CLIR). The study was motivated by the fact that Swedish is known poorly from IR perspective. This paper shows that Swedish has unique features, in particular gender features, the use of fogemorphemes in the formation of compound words, and a high frequency of homographic words. Especially in dictionary-based CLIR, correct word normalization and compound splitting are essential. It was shown in this study, however, that publicly available morphological analysis tools used for normalization and compound splitting have pitfalls that might decrease the effectiveness of IR and CLIR. A comparative study was performed to test the degree of lexical ambiguity in Swedish, Finnish and English. The results suggest that part-of-speech tagging might be useful in Swedish IR due to the high frequency of homographic words.

Keywords: Text retrieval, Cross-language information retrieval, Swedish language, Natural language processing
1 Introduction

Our society is depending on written communication, which today often is created and
stored in digital form. We have an increasing amount of full text material in different
languages available through the Internet and other information suppliers. Information
retrieval (IR) from operational full text databases or text retrieval is characterized by
the lack of controlled vocabularies. A relatively new research area is cross-language
information retrieval (CLIR). It is a process of selecting and ranking documents in a
language different from the query language.

Text retrieval, involves the use of natural language in the form of textual descriptions
of a topic. The variation in word forms and homographic and polysemous expressions
causes disambiguation problems in text retrieval, and the effect increases in CLIR.
We can express the same information content in many different ways depending on
who is the writer and for whom or for what purpose a text is written.

Full text retrieval as a response to a query in natural language involves methods
developed in computational linguistics (Grishman 1986) and natural language
processing (NLP) (Strzalkowski 1994; 1995). Increasing recall and precision depends
on these applications to match non-identical expressions belonging to the same
concept. Computational linguistics offers research in language analysis and
knowledge representation (syntactic and semantic knowledge), and NLP applications
rely strongly on statistical text analysis and machine readable texts and dictionaries
tools relevant for IR purposes involve morphological analysis programs, part of

Because English has been the main language for IR system development, much research on IR and NLP involves English. However, IR systems for small languages like Finnish and Swedish and other Nordic languages cannot be developed properly without studying their special features. Although Spanish and Chinese have rendered special tracks in the TREC Conferences¹ and even Finnish (Alkula & Honkela 1992) has been explored the results cannot be applied on linguistically quite different languages.

This paper analyzes Swedish as document and query language for IR. Swedish is spoken as a native language by 8 - 9 million people, mainly in Sweden but also by a minority population in Finland. However, due to close relationships between the Nordic countries and the other Scandinavian languages the number of people who speak Swedish and can understand it is much larger (Teleman, Hellberg & Andersson 1999). Approximately, about 20 million people have a basic knowledge of Swedish.

Since the Scandinavian languages Swedish, Danish and Norwegian are close, results of scientific research on one language can help to create a deeper understanding of the same phenomena in the other Scandinavian languages (Teleman et al. 1999). Thus, a careful generalization of the results in this study can be made to all three languages.

So far we can report very few analyses of the Scandinavian languages concerning IR, only a small Norwegian study (Fjeldvig & Golden 1988). Swedish and the

Scandinavian languages have unique and novel features which affect IR. We will identify and present these features in the Swedish language relevant to IR on morphologic, syntactic as well as on semantic level. The problem areas studied are a) morphological features such as inflection, derivation, gender and compound words, b) semantical features as homonymy, polysemy and hyponymy. The viewpoints for the analysis are full text IR, database indexing, query formulation and CLIR. Publicly available tools to overcome some of the problems have some pitfalls, which we analyze. We believe that these questions deserve serious research efforts.

The rest of this paper is organized as follows. Section 2 summarizes the linguistic problems of IR. Section 3 considers IR and natural language processing. In Section 4 the main linguistic features of Swedish are analyzed from the perspective of IR. In Section 5 the use of NLP tools in Swedish IR and CLIR is discussed. Section 6 presents concluding remarks.

2 Linguistic problems in IR

An ideal case in searching would be that a search would give all the relevant documents of a database, ranked in an order of descending relevance, and none of the irrelevant documents. In a typical case, however, search results often include many irrelevant documents, and many of the relevant documents are not retrieved. This is primarily caused by linguistic phenomena. The basic linguistic problems of IR can be classified as follows: (1) the selection of alternative concepts and search keys, (2) the morphological variation of search keys, (3) referred and omitted search keys, (4) search key ambiguity, and (5) multilinguality.
The first problem, *the selection of alternative concepts and search keys*, is related to the fact that different documents which consider the same topic may use, and often do use, different concepts and expressions. Typically only part of the relevant documents are retrieved, as users are not able to produce exhaustively alternative concepts for the concepts of requests and are not able to expand the queries using synonyms. The problem of query expansion has been investigated in several studies (Efthimiadis 1996; Kekäläinen 1999). In IR terms, the search concepts which describe a given aspect of a request and which thus are alternative concepts are either in hierarchical or associative relations to each other. In linguistic terms, these relationships include hyponymy, a hierarchical subordinate relation (*animal - horse*) meronymy, a relationship between the whole and its parts (*hand – finger*), and other sense relations, such as antonymy (*hot – cold*) (for associative relations). In IR terms, the search keys which represent a given concept are equivalents. In linguistic terms, equivalence covers synonymy (*Information Retrieval – IR*).

The second problem, *the morphological variation of search keys* is associated with the matching process. In morphologically simple languages, such as English, word form variation can be handled by quite simple means, e.g. for nouns, by recognizing the plural forms (*-s*) (Harman 1991; Porter 1980). However several European languages are morphologically much more complex (e.g., German, French, Italian and the Scandinavian languages). Documents are not retrieved if the search key and its occurrence in a database index (the index term) are not identical in form. Thus a search key given in a base form does not match with the inflected forms of the key. The same problem also concerns derivationally related keys. Moreover, in a database
index, search key occurrences may be embedded in compound words. All these forms may, however, reflect the information need behind the query. In this study a compound is defined as a word formed by two or more components that are spelled together. The term phrase is used for the case where the constituents are spelled separately.

The third problem, referred and omitted search keys covers anaphoric and elliptical keys. In the sentences: “Where is Mary? She is in the garden.”, the pronoun she and the noun Mary are coreferential, and the pronoun is an anaphor. An example of omitted search key or ellipsis is found in the sentences: “I have a dog. You have one too.” In certain cases in proximity searching a considerable improvement in retrieval performance is achieved if anaphors and ellipses are resolved (Pirkola & Järvelin 1996a; 1996b). The fourth problem, search key ambiguity, covers polysemy, a word with more than one sense or subsenses (star – a star in the sky or star - a famous person) and homonymy, different words with the same form (bank – a financial institution or bank – the side of a river). Due to the ambiguity of search keys the intended senses of search keys are not always present in retrieved documents. In other words, despite successful matching irrelevant documents are retrieved.

The fifth problem, multilinguality, concerns multilingual and monolingual document collections from where documents can be retrieved using more than one language. Cross-language retrieval is concerned in particular with the issues associated with the cross-lingual equivalence of words (Ballesteros & Croft 1997; Davis 1998; Hull 1998; Oard & Dorr 1996; Pirkola 1998).

3 IR and natural language processing
Natural language processing involves linguistic methods and the analysis can take place on different levels of a language, i.e., **morphological, syntactic and semantic** levels. On a morphological level the structure of words is analyzed. Recognizing different word forms as variants of the same basic word affects both indexing (word weights) and retrieval (matching). Syntactic analysis determines the structure of phrases and sentences, while semantic analysis investigates the meaning or sense of words and sentences.

Commonly used methods in document indexing are word form normalization (Koskenniemi 1983; Pirkola 1999) and stemming (Harman 1991). A **stemmer** removes affixes from the word forms and the output is a common root, not necessarily a real word. A similar type of process is **normalization**, but in this case the output is the base form, a real word. Due to stemming and normalization three kinds of benefits may be gained (Harman 1991; Järvelin 1995; Alkula & Honkela 1992). 1) A user does not need to worry about truncation and inflection, because different forms of the key are automatically conflated into the same form. 2) Stemming and normalization result in storage savings. 3) Stemming and normalization may improve retrieval performance, especially recall since a larger number of potentially relevant documents are retrieved. However, no significant improvement in performance was found by Harman (1991) in her experiment with simple stemmers for the English language. For inflectionally more complex languages the results are not necessarily the same.

Swedish is rich in compound words, and is thus confronted with the problem of embedded search keys. **Splitting the compounds** into their components allows the use
of the component words as separate search keys. For instance, the decomposition of the compound hustak (roof of the house) gives the expansion keys hus (house) and tak (roof). If the compound hustak was truncated and used as a key, documents including the word tak would not be found.

In IR part-of-speech (POS) tagging may be used to identify central words (word classes, especially nouns) and phrases of a sentence. In CLIR part-of-speech tagging is useful in matching the source language keys with translation dictionary entry words.

A syntactic parser (program) determines the structure of a sentence according to a particular grammar (Grishman 1986). The parsing procedure may involve the assignment of a tree structure to the input sentence. Linguistic transformation, such as transforming active sentences into passive can be of potential value for IR. Syntactic analysis can be used as a basis for further analysis, e.g., anaphor resolution.

**Word sense disambiguation** is an NLP method which aims at finding correct senses for word occurrences. It has been studied intensively in IR and other fields. The methods used in the studies include dictionaries (Dagan et al. 1991; Guthrie et al. 1991; Krovetz and Croft 1989), knowledge bases (Hirst 1987), statistical methods (Brown et al. 1991; Schütze & Pedersen 1995), multiple knowledge sources (McRoy 1992), thesauri (Voorhees 1993) and pseudo-words (Sanderson 1994; Sanderson 1997). Most IR studies have reported no or only slight improvements in retrieval performance due to word sense disambiguation (Krovetz & Croft 1992; Sanderson 1994; Sanderson 1997; Voorhees 1993).
4 Linguistic features of Swedish from IR perspective

In Swedish we can identify the following features that are likely to affect information retrieval: 1) a fairly rich morphology with 2) gender features (gender ‘uter’ and gender ‘neuter’) and 3) high frequency of compound and derivative word forms, 4) common noun phrases are less frequent in Swedish than for example in English 5) a high frequency of homographic word forms.

Morphology. The inflectional and derivational morphology is more complex than that of English. Nouns can be divided into five declination types according to the plural suffixes they take, i.e. -or, -ar, -er(r), -n, and ø (no suffix). Genitive forms are formed by the suffix -s. Definite forms of nouns are expressed using the articles den, det (sing.) de (pl.), and the definite suffixes -en(-n) (gender ‘uter’) and -et (-t) (gender ‘neuter’). Due to the fairly rich inflectional morphology, nouns have several inflectional forms and even the stems change (umlaut) (Table 1). Therefore simple indexing and matching methods are insufficient for Swedish IR. Inflection interferes with the weighting of index terms, and in matching documents are lost due to inflected words.

2 Gender is an inalienable property of a Swedish noun. A noun is defined as gender “uter” or gender “neuter” according to the inflectional suffix it takes in definite form singular (uter: -en, -n, neuter: -et, -t). Homonyms often have different gender, for example: plan –en (plan), plan –et (plane).
<table>
<thead>
<tr>
<th>Base form</th>
<th>Definite form sing.</th>
<th>Plural</th>
<th>Definite form pl.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flicka (girl)</td>
<td>Flicka-n</td>
<td>Flick-or</td>
<td>Flickor-na</td>
</tr>
<tr>
<td>Bro (bridge)</td>
<td>Bro-n</td>
<td>Bro-ar</td>
<td>Broar-na</td>
</tr>
<tr>
<td>Bok (book)</td>
<td>Bok-en</td>
<td>Böck-er</td>
<td>Böcker-na</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Flicka-s</td>
<td>Flickan-s</td>
<td>Flickor-s</td>
<td>Flickorna-s</td>
</tr>
<tr>
<td>Bro-s</td>
<td>Bron-s</td>
<td>Broar-s</td>
<td>Broarna-s</td>
</tr>
<tr>
<td>Bok-s</td>
<td>Boken-s</td>
<td>Böcker-s</td>
<td>Böckerna-s</td>
</tr>
</tbody>
</table>

**Table 1** Examples of inflected noun forms

Adjectives are congruent with their headwords in gender and number (Table 2).

Irregularities to the general inflectional rules exist in nouns as well as in adjectives and verbs. In Swedish strong verbs have vowel change. For a comprehensive description of inflection we refer the reader to books on Swedish morphology, e.g., (Hellberg 1978; Kiefer 1970; Malmgren 1994; Teleman et al. 1999).

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>en vit hare (uter)</td>
<td>den vita haren (the white rabbit)</td>
<td>vita harar (white rabbits)</td>
<td>de vita hararna (the white rabbits)</td>
</tr>
<tr>
<td>ett vit hus(neuter)</td>
<td>det vita huset (the white house)</td>
<td>vita hus (white houses)</td>
<td>de vita husen (the white houses)</td>
</tr>
</tbody>
</table>

**Table 2.** Examples of adjective inflection

For the problem of inflectional forms, morphological analysis programs should be used for word normalization at the indexing and searching stages. We shall consider morphological analyzers and word form normalization in Section 5.

**Gender.** Swedish nouns have two grammatical genders, gender _uter_ and gender _neuter_ (Teleman et al. 1999) The separation of gender in Swedish has positive effect
on the effectiveness of anaphor resolution algorithms, as in the algorithm for pronoun resolution in Swedish, constructed by Fraurud (1988). The gender based resolution is based on the fact that the pronouns den (uter), det (neuter) (meaning it) should be in agreement with their correlates. Homographic nouns often possess different genders (Teleman, Hellberg & Andersson 1999), thus the use of gender is helpful also in the resolution of homographs.

Derivatives. New words are formed by derivation and compounding. Derivatives and their stem words may belong to different part-of-speech categories. The possibility to change part-of-speech class of a word by derivation enables for example the derivation of a noun (an actor or an action) from a verb. Such words are in many cases lexicalized. Words belonging to different part-of-speech categories have different derivational suffixes, and the inflection of derived and underived words is similar. Examples of noun suffixes are -are –het and -ning, e.g., målare (painter), skönhet (beauty) and förkortning (abbreviation). Examples of adjective suffixes are –bar and -aktig, e.g., mätbar (measurable) and livaktig (vivid). The suffixes –era and -na, e.g., spionera (spy), vakna (wake up) are examples of verb suffixes (Malmgren 1994).

Derivation is a common method to form antonyms. They are often formed by prefixes, i.e., o-, a-, in-, il-, im-, ir-, e.g., lycklig / olycklig (happy / unhappy), symmetrisk / asymmetrisk, tolerant / intolerant, legal / illegal, produktiv / improduktiv, rationell / irrationell. Derivational prefixes are, as in the examples above in many cases of foreign origin (Malmgren 1994).
Derivatives can enter into compounds as a stem and the rules for derivation must precede those of compounding, for example:

- *lära* - *lärare* - *lärarförbundet* (teach - teacher - teacher association)
- *fängsla* - *fängelse* - *fängelsestraff* (arrest - prison - penalty of imprisonment)
- *antik* - *antikvitet* - *antikvitetsaffär* (antique - antiquity - antique shop)

In a normalization process for IR, derivatives may be related back to the original word/stem. This increases ambiguity and sometimes makes it impossible to find the right translation alternative for CLIR, because in many cases the derived forms are lexicalized and have separate entries in a translation dictionary.

**Compounds and Phrases.** Swedish is characterized by high frequency of compound words. Phrases are not so common as in English. Compounds are easier to deal with than phrases, because there is no need for identification. On the other hand for effective IR and CLIR, compounds need to be decomposed. Especially in CLIR this is often necessary, because for many compounds dictionaries include only the components of compounds.

A typical feature of Swedish is also the use of *fogemorphemes* in compound word formation. Fogemorphemes are elements used to join the constituents of compounds. The fogemorpheme types are as follows:

- (omission) *flicknamn* (maiden name)
- *s* *rättsfall* (legal case)
- *e* *flickebarn* (female child)
- *a* *gästabud* (feast, banquet)
- *u* *gatubelysning* (street lighting)
Sometimes the component preceding the fogemorpheme is a stem, e.g., *gat(a)*
sometimes a base form, e.g., *rätt*. Proper weighting of index terms and effective
matching require that fogemorphemes could be handled and correct base forms of
component words could be identified.

Many nominal compounds are lexicalized so that it is not always possible to derive
their meaning from the meanings of component words, e.g., *jordgubbe* (strawberry).
In many compound words, however, the last component is a hyperonym of the full
compound being thus a valuable key (Pirkola 1999). For exemple, compound words
denoting to different types of schools have the component *skola* (school) as their last
component. In IR the decomposition of compounds would often be helpful in cases
like this. The last component may also be used as an ellipsis later in the text.

Particle verbs (compound verbs where one component is a particle) are special forms
of compounds. There are four types of particle verbs: always tightly compounded,
e.g., *påminna* (to remind), always loosely compounded (except in past participle),
e.g., *tycka om / omtyckt* (to like), both loosely and tightly compounded, but with no
change in sense, e.g., *omtala / tala om* (speak about), and with a change in sense, e.g.,
*avbryta / bryta av* (interrupt / break) (Malmgren 1994). From IR point of view, some
particle verbs are similar to nominal compounds and some similar to nominal phrases.
Tightly compounded particle verbs are mostly lexicalized, but with the other forms
we can see the benefits of identifying phrases and the benefits of splitting compounds
into components.
**Homonymy and polysemy.** Homonymy and polysemy are two different forms of lexical ambiguity. Homonyms are different lexemes spelled similarly. A polysemous word is a word that has several subsenses which are related with one another (Teleman et al. 1999). For example the word *stjärna* (star) may denote to a famous person or a star in the sky. Due to lexical ambiguity the intended senses of keys are not always implemented in the retrieval results, and irrelevant documents are retrieved. In CLIR the negative effects of lexical ambiguity on retrieval performance are often more severe than in monolingual retrieval, because in CLIR both source and target language keys may be ambiguous.

Homographs are very common in Swedish. According to Karlsson (1994) around 65% of Swedish words in running text are homographs. In Finnish the percentage is only 15% and in English around 50%. For example, the Swedish form *för* can be used as a preposition, verb, noun, and adverb. Since Swedish is a language rich in homographic word forms it is possible that the disambiguation of homonyms or the analysis of the the meaning of a word from its context would have a positive effect on retrieval results, contrary to the findings for English, e.g., (Sanderson 1994).

5 Natural language analysis programs for Swedish

**Normalization.** *SWETWOL* is a morphological analyzer developed to analyze Swedish words (Karlsson 1992). The program incorporates a full description of Swedish morphology, and morphological descriptions are assigned. Normalization
returns all the derivational and inflectional forms of a word to the base form by recognizing / deleting inflectional and derivational affixes.

Let us consider some examples of Swedish word normalization using SWETWOL (web-version, used in October 1999). Table 4 shows that most sample words are correctly normalized. The noun *behandling* (treatment) is normalized to the verb *behandla* (to treat). In CLIR this may be a problem, because dictionaries typically give nouns and verbs as separate entries. It would benefit the translation if the normalization output would give both forms. In monolingual IR this problem is less severe since both document and query words are treated similarly.

<table>
<thead>
<tr>
<th>Input word</th>
<th>Normalization to base form</th>
<th>SWETWOL analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>avgöras</td>
<td>avgöra (decide)</td>
<td>V PASS INF (verb, passive, infinitive)</td>
</tr>
<tr>
<td>utsläpp</td>
<td>utsläppa (dumping)</td>
<td>V ACT IMP (verb, active, imperative)</td>
</tr>
<tr>
<td></td>
<td>utsläpp (outlet)</td>
<td>N NEU INDEF SG/PL NOM (noun, neuter, indefinite, sg./pl. nominative)</td>
</tr>
<tr>
<td><strong>behandling</strong> (treatment, usage, management)</td>
<td><strong>behandla</strong> (treat, deal with)</td>
<td><strong>VDER/-ning N UTR INDEF SG NOM (verb derivation /-ning, noun, uter, indefinite, sg. nominative)</strong></td>
</tr>
<tr>
<td>väsentligt (essentially, mainly)</td>
<td>väsentlig (essential, fundamental, main)</td>
<td>A NEU INDEF SG NOM (adjective, neuter, sg. nominative)</td>
</tr>
<tr>
<td><strong>operation</strong> (operation)</td>
<td>operation (operation)</td>
<td>N UTR INDEF SG NOM (noun, uter, indefinite, sg. nominative)</td>
</tr>
</tbody>
</table>

Table 4. Normalization for Swedish

**Compound splitting.** Table 5 presents some examples of compound splitting by SWETWOL. In the first example “skogsindustrin” (forest industry) the components *skog* (forest) and *industrin* (the industry) are joined with the fogemorpheme “s”. The
latter component *industrin* is normalized to the base form *industri*, which is a hyperonym of *skogsindustrin* and therefore often a valuable search key. The former component *skogs* has retained the “s”, and is not normalized to the base form *skog*. In monolingual IR this has the effect that the matching process only concerns the genitive form *skogs* and compounds with the first component *skogs*. The compounds with *skog-* (the fogemorpheme “s” not used) like *skogbevuxen* (wooded), *skoglös* (treeless) are not found.

For CLIR this is still more puzzling since the translation alternatives are several for almost every word. If the compound word *kärnkraftverk* (nuclear power plant) (example 2 in table 5) is found in the dictionary, we will get an unambiguous translation. If the compound is split, we will have a situation where the last component is normalized to base form but the other components are retained in the form they exist in the compound word. In this case we have an example of omission (See Section 4 on fogemorphemes), where *kärna* is transformed to *kärn*.

For the word *toppmötet* (meeting, summit) there are two outputs *tupp* and *möte* and *tupp, mö* and *te* (example 3 in table 5). The latter case provides an example of a semantically false coordination.
Table 5. Examples of compound splitting in Swedish

<table>
<thead>
<tr>
<th>Input word</th>
<th>Compound splitting</th>
<th>SWETWOL analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>skogsindustrin</em> (the forest industry)</td>
<td>skogs#industri (forest # industry)</td>
<td>&lt;N&gt; # N UTR DEF SG NOM (noun # noun, utter, definite, sg. nominative)</td>
</tr>
<tr>
<td><em>kärnkraftverken</em> (the nuclear power plant)</td>
<td>&quot;kärn#kraft#verk&quot; (kärn-a = kernel, seed, grain, nuclear etc. kraft = force, strength, etc. verk = work(s), labor, plant, mill etc.)</td>
<td>&lt;N&gt; # &lt;N&gt; # N NEU DEF PL NOM (noun # noun # noun, neuter, definite, pl. nominative)</td>
</tr>
<tr>
<td><em>toppmöte</em> (summit, (top-level) meeting)</td>
<td>&quot;topp#möte&quot; (topp= done!, agreed!, it's a bargain! top crest, summit, pinnacle, peak, apex möte= meeting, encounter, appointment, date, gathering, assembly, conference)</td>
<td>&lt;N&gt; # N NEU DEF SG NOM (noun#noun, neuter, def. sg. nominative)</td>
</tr>
</tbody>
</table>

**Lexical ambiguity.** Homograph disambiguation might be useful in Swedish IR due to the high frequency of homographic expressions in Swedish. We performed a test in which 35 Finnish test requests used in IR and CLIR research were translated into English and Swedish. We picked out 60 Finnish, English and Swedish equivalent words representing different part-of-speech classes, and studied the degree of ambiguity in the three languages. We used off-the-shelf dictionaries\(^3\) to count the

Ruotsi-suomi-suursanakirja [Swedish-Finnish comprehensive dictionary] / Lea Lampén, Porvo : WSOY, 1973 (70.000 words)
number of homographic word forms and translation alternatives. By translation alternatives is meant alternative translations (including polysemy) and translation examples to a word in similar dictionaries that are used for CLIR-research. The test with homographic base form words in Table 6 show an average of 0,22 homographic expressions for Swedish, 0,07 for English and no homographic expressions in the Finnish test sample. The number of translation alternatives for the same test sample is shown in Table 7. The number of translation alternatives is depending on the dictionary used, in this case standard dictionaries of approximately the same size. Similar translation tools would have to be used for CLIR.

<table>
<thead>
<tr>
<th>Average number of homographic base form words</th>
<th>Finnish</th>
<th>English</th>
<th>Swedish</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nouns</td>
<td>0</td>
<td>0,05</td>
<td>0,25</td>
</tr>
<tr>
<td>Verbs</td>
<td>0</td>
<td>0,15</td>
<td>0,4</td>
</tr>
<tr>
<td>Adjectives+Adverbs</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Average all POS</td>
<td>0</td>
<td>0,07</td>
<td>0,22</td>
</tr>
</tbody>
</table>

**Table 6.** Average number of homographic base form source words in dictionaries

<table>
<thead>
<tr>
<th>Average number of transl. alternatives for base form words</th>
<th>Fin-Swe</th>
<th>Eng-Swe</th>
<th>Swe-Fin</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nouns</td>
<td>20,8</td>
<td>8,1</td>
<td>8,7</td>
</tr>
<tr>
<td>Verbs</td>
<td>43,9</td>
<td>36,4</td>
<td>19,3</td>
</tr>
<tr>
<td>Adjectives + Adverbs</td>
<td>5,7</td>
<td>7</td>
<td>5</td>
</tr>
<tr>
<td>Average all POS</td>
<td>23,4</td>
<td>17</td>
<td>11,0</td>
</tr>
</tbody>
</table>

**Table 7.** Average number of translation alternatives for base form words
To test the effect of normalization we used the test sample, this time using the inflected word form of the original test requests. The (TWOL)$^4$ morphological analyzers were used in the experiments. In this case (Table 8), we can see a clear difference in the languages. For Finnish nouns we can see a small increase in translation alternatives, and a higher for verbs and adjectives + adverbs. English shows practically no difference, but in Swedish we have a considerable increase in translation alternatives for all part-of-speech classes.

<table>
<thead>
<tr>
<th>Average number of translation alternatives of inflected word forms after morphological analysis, the increase in percent to words in base form (Table 7)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Nouns</strong></td>
</tr>
<tr>
<td>-----------------------------------</td>
</tr>
<tr>
<td>Nouns</td>
</tr>
<tr>
<td>Verbs</td>
</tr>
<tr>
<td>Adjectives + Adverbs</td>
</tr>
<tr>
<td>Average all POS</td>
</tr>
</tbody>
</table>

Table 8. Average number of translation alternatives after normalization

A general cautiousnes is needed about the prospects of word sense disambiguation strategies and the hope that they should significantly improve retrieval performance (Lewis & Sparck Jones 1996). But also in this case we have no research result for Swedish.

---

One method to disambiguate word senses is the use of a concordance, where the context of the word is listed along with the word. Homonymous words can be disambiguated by establishing their co-occurrence or collocability. For example, as in English in Swedish the word “bank” means both financial institution and reef or sand bank. Bank in the sense of financial institution is probably occurring with words like “lån” (credit), “pengar” (money), “ekonomi” (economy), while the word “bank” (reef) occurs with “jord” (earth), “flod” (river) etc.

Consider one of our test requests:

“Skuldkrisen i Sydamerika. Hur har skuldsättningsproblemet utvecklats?”
“South American debt crisis. How has the debt problem developed?”

From a dictionary we can establish the following meanings for the Swedish words skuld, kris and skuldsättning:

<table>
<thead>
<tr>
<th>Word</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>skuld</td>
<td>debt, amount due, liabilities, fault, blame, guilt</td>
</tr>
<tr>
<td>kris</td>
<td>crisis, depression</td>
</tr>
<tr>
<td>skuldsättning</td>
<td>getting into debt</td>
</tr>
</tbody>
</table>

If we can establish that the words skuld and skuldsättning are usually co-occurring, we can establish that the right translation would be “debt crisis”, not “guilt crisis” or “guilt depression”. However, the method with collocating words for disambiguation purposes is not easy to implement in operational IR-systems, and specially not if we only have a relatively short query to work with. This is the case for dictionary-based methods in CLIR.
Part-of-speech tagging might be useful in Swedish IR because of the high frequency of homographic word forms. The words in the query and in the documents are part-of-speech marked and in the matching process the part-of-speech marks for the search keys and the document index terms are acknowledged. With this method, for example, the preposition för (because) could be separated from the verb för(a) (bring). In CLIR part-of-speech tagging of source language words has been used successfully for finding correct translation equivalents in the dictionary (Ballesteros & Croft 1998).

With CLIR-research in mind we tested part-of speech-marking, using the same test sample of 60 words in Finnish, English and Swedish. The effect on the number of translation alternatives is shown in Table 9. We can see a small decrease in the number of translation alternatives for English nouns, no decrease for Finnish nouns and a larger for Swedish nouns. The results agree with earlier research by Leppänen (1995) (Finnish) and Sanderson (1994, 1997) (English). From the test we come to the conclusion that part-of-speech-marking might be effective in CLIR for Swedish.

<table>
<thead>
<tr>
<th>Average number of translation alternatives after POS-marking</th>
<th>the decrease in percent to normalization without POS-marking (Table 8)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Fin-Swe % decrease</td>
</tr>
<tr>
<td>Nouns</td>
<td>22,9</td>
</tr>
<tr>
<td>Verbs</td>
<td>45,5</td>
</tr>
<tr>
<td>Adjectives+Adverbs</td>
<td>5,0</td>
</tr>
<tr>
<td>Average all POS</td>
<td>24,5</td>
</tr>
</tbody>
</table>

Table 9. Average number of translation alternatives after POS-marking
6 Conclusions

The present situation with Swedish natural language IR is that it is lacking research. Neither has there been any CLIR research in Swedish. A Norwegian study (Fjeldvig & Golden 1988) done ten years ago with a fairly small test database is the only research result that might give some leads on how the Swedish language behaves in similar research situations. The Norwegian study shows the effectiveness of automatic normalization and truncation but an improvement in the results for compound splitting was not shown.

We have described the properties of Swedish language and pointed out a number of research problems for Swedish IR. We also argued that they deserve serious research attention, because 1) the Nordic countries have a fairly large population, 2) they are technically advanced countries with lots of IR-related activity, and because 3) the features of Swedish differ from those of many other languages, such as English and Finnish. The research results for other languages do not necessarily apply to Swedish due to its unique features, e.g., the frequency of homographs, the use of fogemorphemes in compound words, and particle verbs.

We have also shown that the publicly available morphological analysis tools for word normalization and compound splitting have pitfalls that might decrease the effectiveness of IR and CLIR. The identification of base forms and the analysis of
compounds, in particular may present problems for IR. Therefore the morphological analysis programs cannot be utilized directly but need to be tuned for IR applications.

References


