Information Research, Vol. 7 No. 2, January 2002,

Compounds in dictionary-based cross-language information retrieval

Turid Hedlund
Department of Information Studies
University of Tampere
Finland

Abstract

Compound words form an important part of natural language. From the cross-lingual information retrieval (CLIR) point of view it is important that many natural languages are highly productive with compounds, and translation resources cannot include entries for all compounds. Also, compounds are often content bearing words in a sentence. In Swedish, German and Finnish roughly one tenth of the words in a text prepared for information retrieval purposes are compounds. Important research questions concerning compound handling in dictionary-based cross-language information retrieval are 1) compound splitting into components, 2) normalisation of components, 3) translation of components and 4) query structuring for compounds and their components in the target language. The impact of compound processing on the performance of the cross-language information retrieval process is evaluated in this study and the results indicate that the effect is clearly positive.

Introduction

Cross-language information retrieval (CLIR) deals with the problem of presenting a query in one language and retrieving documents in one or several other languages. The use of natural language in topics and documents makes it obvious that many of the research questions deal with aspects of word morphology and semantics (Strzalkowski *et al.* 1999; Smeaton 1999). Compound words form an important part of natural language, since compounding is a major way of forming new words. From the information retrieval (IR) point of view, compounds may be content bearing words in natural language sentences and therefore important for the retrieval result. Compound words have been discussed in the main forums for research and evaluation on cross-language information retrieval systems, the TREC conferences and the CLEF evaluation forum (1). The problem with compound handling for cross-language information retrieval is acknowledged for many languages, but we still lack research results on the importance and effects to the retrieval result of compound handling in natural language queries. In this study compound words are analysed from the cross-language information retrieval perspective. Methods of handling compounds in an automated process as well as the effects of compound handling on retrieval results are presented.

Since natural language cross-language information retrieval is faced with the task of identifying, normalising, translating and matching query words to the database index, linguistic tools and linguistic analysis are in use. Natural language analysis tools are also extended in information retrieval and cross-language information retrieval applications also to a sub-word level, e.g., morphological analysers are used for the decomposing of compounds and normalisation of words (Sparck Jones 1999). Normalization of words is a way to improve recall also in monolingual information retrieval. By reducing the number of word forms, more successful matches can be made (Strzalkowski 1995). Stemming algorithms, able to produce word stems, are used for the same purpose (Porter 1980).

A general introduction to compounds and their relevance from an information retrieval perspective is presented in Section 2. In dictionary-based cross-language information retrieval, stemming or normalisation of words to base forms using morphological analysis programs is necessary to be able to match the right dictionary entry. Typically in Germanic languages (for example, Swedish, Norwegian, Danish, German and Dutch) the compounds are joined by special elements, "fogemorphemes" (Hedlund *et al.* 2001a). However, splitting compounds into constituents using morphological analysis programs does not necessarily produce lexical base form words due to the use of "fogemorphemes", inflection, and word stems to join compounds. Section 3 provides a description of compound formation in three compound languages, Swedish, German and Finnish, the first two being Germanic languages while Finnish is a Fenno-Ugrian language.

If one cannot translate a compound as a whole, one does not normally get a one-to-one translation for each component in a compound, but rather several alternative translations. Also, a component might remain untranslated due to limitations in the lexicon of the morphological analyser or a translation dictionary. It may also be a proper name or a spelling variant of a proper name, or a spelling error. Components might also remain untranslated as described above due to fogemorphemes and inflection.

For compounds containing more than two components we can in many cases identify compounds within the original compound. For example the German compound "*Universitätshochhaus"* contains the compound "*hochhaus*" (Warren 1978).

Compound based structuring and the use of proximity operators have been tested for several language pairs at the University of Tampere by Pirkola (1998), and Hedlund *et al.* (2001b; 2001c). Methods for component processing, handling fogemorphemes and the use of proximity operators are presented in Section 4, followed by a discussion in Section 5 on the relevance of the findings and their importance to cross-language information retrieval research.

Compound definitions, categories and relevance from an information retrieval perspective

Compounds can be defined according to orthography (Akmajian *et al.* 1997), namely compounds where the components are written together and compounds in the form of phrases. The way to form compounds, that is, joining words together into one word or having them separated as a phrase is a language dependent feature. Swedish, German, Dutch and many other Germanic languages belong to the former type, likewise Finnish (a Fenno-Ugric language), while English and French are examples of the latter one. In this study, the term *compound* refers to a case where the components are written together. The term *phrase* refers to a case where components are written separately.

What is important for purposes of information retrieval is the categorisation according to the part of speech of the components. The combinations are numerous, we can identify noun - noun (raindrop), adjective - noun (bluebell), adjective - adjective (light green), noun - verb (heart-broken), preposition - verb (over-ride) as common combinations but also many others. The part of speech of the compound is normally defined on the basis of the last component (Malmgren 1994; Akmajian *et al.* 1997). From the information retrieval point of view nouns and combinations containing nouns are often content-bearing words and therefore mo! st! important. The syntactic structure of compounds, components within components, and the left- or right-branching structure are used for compound handling strategies in this study. Compounds also have a paradigmatic structure, e.g., berry is a hyperonym of several types of berries (blueberry, strawberry).

Compounds are used to form new words in a language and are therefore productive. The word formation process is complex and different stages can be identified. We have **occasional compounds** that are used to express a specific thing or process, e.g., (*skolbokshylla*, a Swedish word for a shelf containing school books). When forming a new concept as a compound, the semantic structure is often **transparent** or **compositional** (Malmgren 1994). The meaning of the compositional compounds can be derived from the meanings of their components, e.g., *Handelsvertrag*, a German word for trade agreement where the components are *Handel* (trade) and *Vertrag* (agreement).

When a compound is frequently used in a language the denotation of the original transparent compound is sometimes transformed into a special concept. In Finnish the word refrigerator is *jääkaappi* (*jää* = ice and *kaappi* = cupboard), which is no longer a cupboard containing ice, but has the special meaning of refrigerator. The compound is stable and is no longer an occasional compound.

Compounds where the meaning of the compound cannot be derived directly from the meanings of the components are **opaque** or **non-compositional**, e.g., the English word *strawberry* and the Swedish word *jordgubbe* (meaning strawberry). The word for towel in German *Handtuch* (hand cloth or towel) is opaque to a certain degree but we can identify a semantic relationship to the components forming the compound. This is quite a usual pattern for compounds and completely opaque compounds are not very common (Fleischer & Barz 1992). Opaque compounds denote in many cases common concepts and are therefore **lexicalised** and can often be found in dictionaries.

From the information retrieval point of view and especially for the translation process in cross-language information retrieval using dictionaries the categorisation presented above is important. Occasional compounds rarely appear as entries in translation dictionaries, transparent compounds may appear as entries in comprehensive dictionaries or dictionaries created for a special domain. The word *Handelsvertrag* (trade agreement) is most certainly appears in a dictionary of economics, whereas frequent opaque and lexicalised compounds often appear as entries even in small general translation dictionaries.

For the above-mentioned reasons we need ways to handle compounds in dictionary-based cross-language information retrieval when they cannot be directly translated. A solution is to split the compounds into components and translate them separately since the components can be seen as contents bearers from the information retrieval point of view. This holds for many transparent and occasional compounds. In many cases the splitting of non-compositional compounds and translating the components often increases ambiguity, and adds no value to the retrieval performance. The automatic translation process (below) is designed to work in a way similar to human understanding of language. If a compound word is found in the dictionary it will be translated, and only if not will the process of splitting it into components and translating them take place.

Compound formation in Swedish, German and Finnish

The simplest form of compound formation is to join two word stems, base forms or inflected forms, e.g., the Swedish word, *livbåt*, (lifeboat) of *liv* (life) and *båt* (boat). This new word can in turn be joined to form other words, e.g., *livbåtsbesättning* (lifeboat crew). The German word *Welthandel* (world trade) is formed from *Welt* (world) and *Handel* (trade) and is a part of a compound *Welthandelsorganisation*(world trade organization). Similarly the Finnish word *Lentokoneonnettomuus* (air plane accident) is formed of the three components *lento* (flight) *kone*(machine) and *onnettomuus* (accident). Note that in Swedish and German the fogemorpheme "s" is used to join the constituents (see Figure 1). Several components in a compound usually indicate that the compound is occasional. In English compounds in the form of phrases are common, as can be seen in the translations of the examples above.

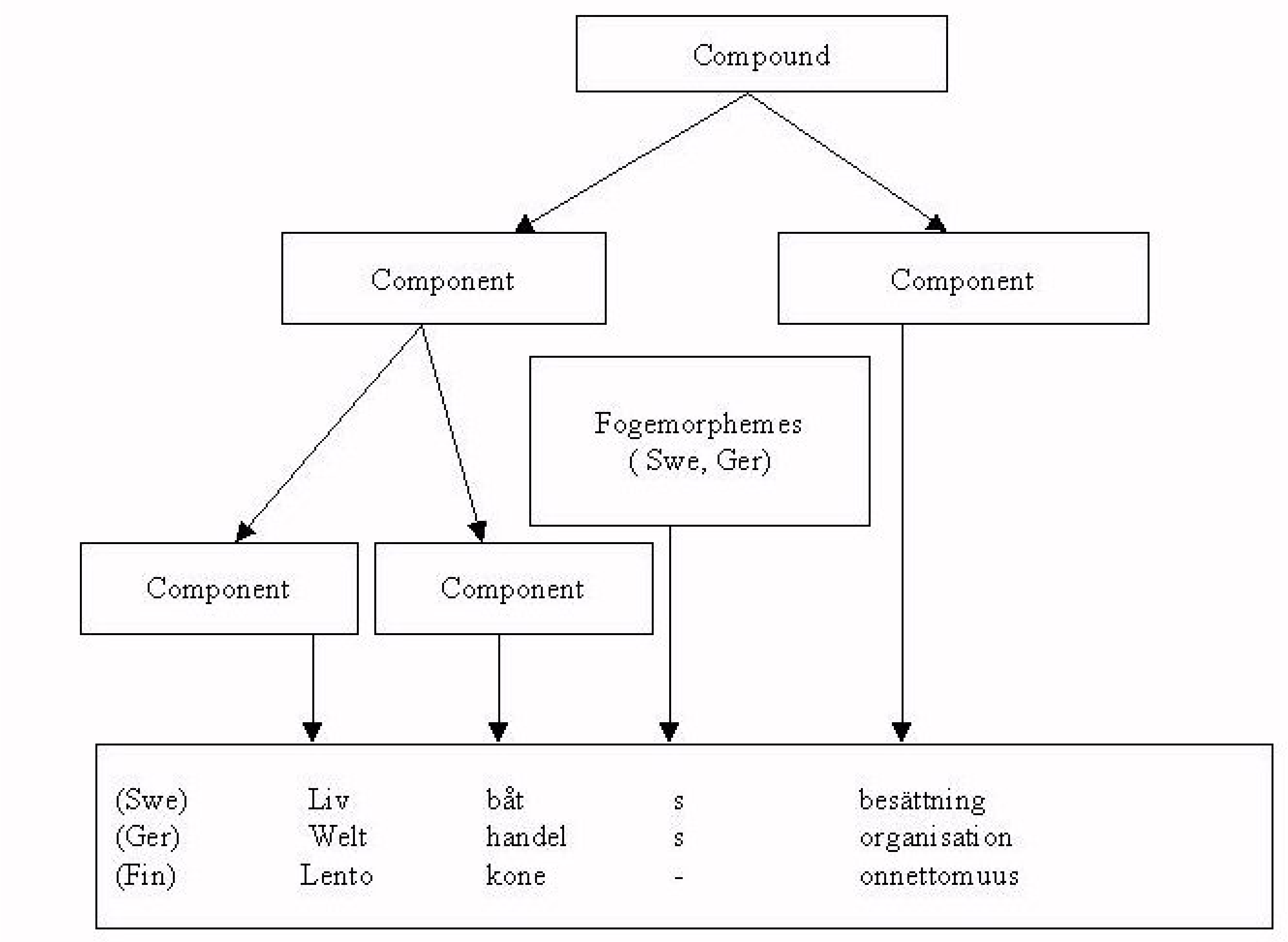
Compound
Component
Component
Fogemorphemes
(Swe, Ger)
Component
Component
(Swe) Liv båt s besättning
(Ger) Welt handel s organisation
(Fin) Lento kone - onnettomuus

Figure 1. Examples of compound noun formation in Swedish, German and Finnish

The use of fogemorphemes to join compound is typical of the Germanic languages. In Swedish the fogemorpheme types are as follows:

- *(omission)*__ *flick(a)namn* (maiden name)
*-s*______*rätt**s**fall* (legal case)
*-e*_____ *flick**e**barn* (female child)
*-a*_____ *gäst**a**bud* (feast, banquet)
*-u*_____ *gat**u**belysning* (street lighting)
*-o*_____ *människ**o**kärlek* (love of mankind)

Sometimes the component preceding the fogemorpheme is a stem, e.g., *gat(a)* sometimes a base form, e.g., *rätt.* From the information retrieval point of view, proper weighting of index terms and effective matching require that fogemorphemes be handled and correct base forms of component words identified.

In German the most common fogemorpheme types are:

*-s*_____ *Handel**s**vertrag* (trade agreement)
*-n*_____ *Affe**n**haus* (monkey house)
*-e*_____ *Gäst**e**bett* (guest bed)
*-en*____ *Fotograph**en**ausbildung* (training of photographers)
*-er*____ *Gespenst**er**haus* (ghost house or haunted house)
*-es*____ *Freund**es**kreis* (circle of friends)
*-ens*___ *Herz**ens**brecher* (hearthbreaker)
- *(omission)*__ *Sprach(e)wissenschaft* (linguistics)

The morpheme boundaries of the constituents in a compound are called fogemorphemes in Swedish, Fuge-elements in German. They mark the elements of one constituent and the following constituent in a compound and are mostly connected with nouns and sometimes verbs as the first element of a compound. For the use of fogemorphemes some rules can be established but there is no complete scheme for how the components are connected (Fleischer & Barz 1992). They are historically often derived from plurals, from oblique case suffixes and combinations of the two.

Although there are no fogemorphemes in the Finnish language, compound splitting can still produce components in inflected forms or in the form of derivatives. For example the compound noun *tupakastavieroituskurssi* (course to break the habit of smoking) includes the components *tupakasta* (an inflected form of *tupakka* = tobacco), *vieroitus* (wean, a derived form of *vieroittaa* = wean, to break a habit) and *kurssi* (a base form for course).

To give an indication of the frequency of compounds in text for different languages a sample of 100,000 words of newspaper text was collected for each language (Swedish, Finnish and German) in this study. The text corpus was processed in a way similar to building up a document database index. Stop words, that is, frequent words, e.g. articles and pronouns in the text were eliminated. Then the text was processed by a morphological analyser, and the criteria for a compound was that it could be separated into constituents by a morphological analyser. The test indicated that for Finnish 8.7%, for Swedish 9.8% and for German 10.2% of the words are compounds.

We can clearly see common patterns in compound formation in the three different languages. The Germanic languages, Swedish and German, naturally have more in common, but for information retrieval or cross-language information retrieval purposes solutions for how to handle compounds may to some extent be language independent. In information retrieval and cross-language information retrieval we need to identify compound components for translation, and that means considering how to handle fogemorphemes, word stems and inflected word forms. In the following section compounds are processed for a bilingual translation process.

Compound processing for bilingual cross-language information retrieval

The problems with compounds and compound splitting in Germanic languages (German, Dutch, Swedish) were mentioned as important in several papers presented at the Cross-language Evaluation Forum (2001). Generally compound splitting and the translation of components is found to improve retrieval results (Savoy 2001; Kraaij 2001; Riplinger 2001; Hedlund *et al.*2001c).

In this section problems and solutions to three important questions concerning compounds in dictionary based cross-language information retrieval are discussed. They are 1) normalisation, 2) translation of components and 3) query structuring of components in the target language. Since the impact of compound processing on the actual query process is of interest, evaluation results for this way of processing compounds will be presented.

The test settings in this study are:

- CLEF- 2000 topic set in Swedish, Finnish and German (33 queries), http://www.iei.pi.cnr.it/DELOS/CLEF/
-
- CLEF- 2001 topic set in Swedish, Finnish and German (47 queries), http://www.iei.pi.cnr.it/DELOS/CLEF/
- LA Times English document collection provided by CLEF

Other resources:

- Motcom Swedish - English translation dictionary (60,000 entries) by Kielikone plc. Finland
- Motcom Finnish - English translation dictionary (110,000 entries) by Kielikone plc. Finland
- Oxford Duden German - English translation dictionary (260,000 entries)
- Morphological analysers: SWETWOL, FINTWOL, GERTWOL and ENGTWOL by Lingsoft plc. Finland
- InQuery retrieval system, Center for Intelligent Information Retrieval at the University of Massachusetts.

Normalisation of components

Morphological analysers used to decompose compounds into constituents are effective, but do not necessarily present the constituents in base forms. This is especially the case with analysers for Germanic languages with fogemorphemes, but also words containing components that are stems or inflected word forms are not automatically normalised and presented in base forms.

Fogemorpheme algorithms to transform the output of morphological analyzers to base form words have been tested for Swedish - English translations using CLEF 2000 Swedish test queries (Hedlund et al. 2001b). The algorithms are able to handle only the most common fogemorphemes. Table 1, presents the translation results of a sample of 10 compounds including fogemorphemes is presented.

Swedish compoundn	Normalised components	English translation
medlems|land	medlem land	member country member land
befolknings|konferens	befolkning konferens	population conference
världs|marknaden	värld marknad	world market
brand|bekämpnings|olyckor	brand bekämpning olycka	fire bekämpning accident
samarbets|område	samarbet område	samarbet area
flyg|plans|olyckor	flyg plan olycka	aviation plane accident
världs|handels|organisation	värld handel organisation	world trade organisation
undervisnings|metod	undervisning metod	education method educational method
varu|hus|tak	vara hus tak	goods building roof article building roof

Table 1: Sample normalisations of compound components

The algorithm fails to normalise the component *samarbets* (samarbet is a word stem) to the right base form *samarbete* (co-operation) and for the word *bekämpning* the dictionary did not provide any translation. In some examples, the sense of the Swedish compound does not transform into a correct English phrase although components are found in the dictionary. We obtain nonsense combinations like "aviation plane accident","article building roof", "goods building roof". The reason for this is that the Swedish compounds contain non-compositional compounds as components *flygplan* (aeroplane) and *varuhus* (department store). This phenomenon is discussed later in Section 4.2. However, the algorithm developed and used for handling fogemorphemes, obviously reduces the number of non-translated components.

Two different automatic test runs, (4) one using the fogemorpheme algorithm and one without it were performed with the CLEF 2000 test set of Swedish topics and the English document database provided by CLEF. The average precision rates over 11 recall levels from 0 to 100 for the eleven queries containing fogemorphemes indicate a slight effect on the result (Table 2).

Data set	Average precision over recall levels	
Queries containing fogemorphemes	*Fogemorpheme algorithm used*	*Fogemorphemealgorithm not used*
Swedish queries containing fogemorphemes (N=11)	0,2898	0,2711

Table 2: The effect of the fogemorpheme algorithm

For individual queries the effect depends on whether the translated compound components can express the sense of the original word in the translation. For non-compositional compounds the effect of compound splitting and fogemorpheme handling does not have any effect, or the effect is negative, since additional translations add noise to the query. However, since we are dealing with constituents of compounds the actual effect on the search result also depends on other factors, such as to what extent the constituents are important topic words. A similar algorithm for German has been developed and used in dictionary-based cross-language information retrieval tests (Hedlund *et al.* 2001c).

The inflected word forms appearing as constituents, e.g., the Finnish word *ympäristön|suojelu* (environment protection), where the first component is the genitive form of *ympäristö* (environment), also need to be normalised to base form in order to match the entries of a translation dictionary. A morphological analyser can be helpful in this case, by normalising the compound components to base forms in a two-stage process. The original compound is first decomposed into components and then the single components are normalised.

By applying these two features to a system for compound processing the majority of the compounds can be handled for translation. The unsolved problem with component normalisation is word stems and other irregularities in compound formation; they cannot be handled easily in an automated process.

Component translation and structuring of queries

A problem with splitting compounds in cross-language information retrieval is to find an optimal solution as to how to combine the translation alternatives in the final query structure. This is especially true for multi-compositional compounds (three or more components). Multi-compositional compounds are quite frequently used in text. For example, the Swedish word *metangasfyndigheter* (deposits for methane gas) is composed of three components, *metan gas fyndigheter* (deposits). A general Swedish - English dictionary gives the following translations (Table 3).

Swedish	English
metan	methane
gas	gas gauze
fyndighet	deposit

Table 3: Sample translations from Swedish to English

A solution using structuring of the translation alternatives for Finnish to English CLIR was originally developed by Pirkola (1998). It includes the translation of every component and combining each translation equivalent to each other using a proximity operator. The proximity statements are joined using the syn-operator of the InQuery retrieval system. Different grouping strategies of components in combination with a proximity structure have been used in later CLIR studies by Hedlund *et al.* (2001b; 2001c).

For the example key above, the initial query structure used in Hedlund *et al.* (2001b) combines all the component translations in a synonym structure using a strict word order operator "OD" and a proximity operator 5 allowing the window size to span enveloping 6 words (2). See Figure 2. For proximity operators see Section 4.3.

```
#syn(#od5(methane gas deposit) #od5(methane gauze deposit)).
```

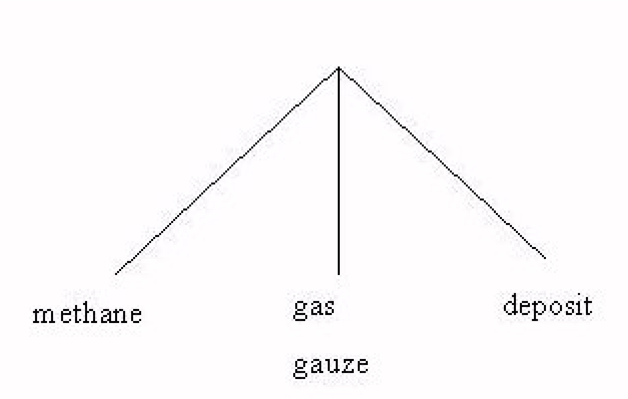

Figure 2: Query structure for the three component compound

This construction when matched to a document database index requires that all words in the construction in this exact order match the document index. It would thus not be able to retrieve a paraphrased version of the concept, e.g., "deposit of methane gas" or the word pair "methane gas". In the worst case this might affect the retrieval result negatively due to no matches for this concept.

The syntactic structure of a compound can almost always be divided into two main constituents. These may in turn be made up of compounds or other complex units (Malmgren 1994; Warren 1978). The structure is either left branching or right branching (see Figure 3). An alternative, according to the above mentioned structure is to group the components into pairs of consecutive components in order to form meaningful pairs. This has been and tested by Hedlund *et al.*(2001c).

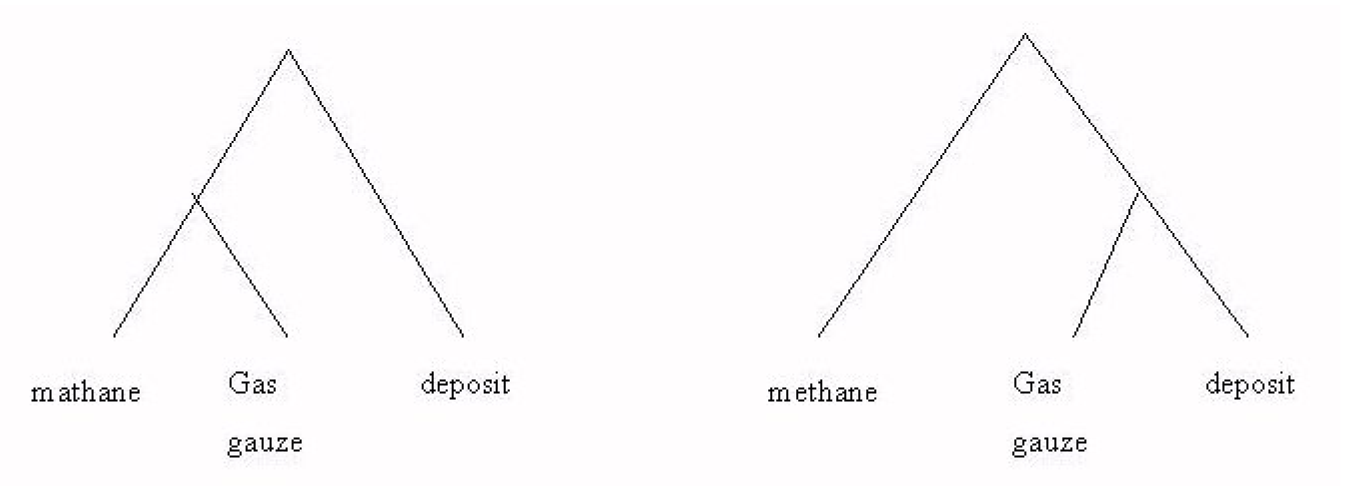

Figure 3. Left branched and right branched compound structure.

The previous example grouped into consecutive pairs would be as follows. Note that also the word order operator is changed to unordered window (UW):

#syn(#uw5(methane gas) #uw5(methane gauze) #uw5(gas deposit) #uw5(gauze deposit).

More formally this is expressed as follows:

Compound word = *CW* [$comp_1$+$comp_2$+$comp_3$+...$comp_n$]

Consecutive pairs, *(i,i+1)* are formed, *i*=1...*n*-1

where *n* is the number of components.

The other aspect, also discussed in Section 3, is that compounds in many cases also contain components that themselves are compounds. These are naturally consecutive. In the above example both the components *metan* and *gas* form the compound *metangas* and the components *gas* and *fyndighet* form the compound *gasfyndighet* (gas deposit). In an earlier study by Hedlund *et al.* (2001c) our attempt was made to translate all component pairs directly by looking them up in a translation dictionary. For example, if the words *metangas* and *gasfyndighet* can be translated to methane gas and gas deposit, the query formulation would be like this:

#syn(#uw5(methane gas) #uw5(gas deposit)).

Another problem with compounds is that not all components are included in a dictionary, for example in Section 4.1, Table 1, the Swedish compound *brandbekämpningsolyckor* (accidents in fire fighting). The components are *brand* (fire) *bekämpning* (fighting) *olyckor* (accidents, conflagrations). In this case the word *bekämpning* is not translated by a sample dictionary. Combinations including a word in a foreign language do not usually retrieve any documents.

#syn (#uw5(fire bekämpning) #uw5(bekämpning accident) #uw5(bekämpning conflagration))

Moving the untranslatable component outside the proximity construction can be helpful for the matching process.

#syn(bekämpning #uw5(fire accident) #uw5(fire conflagration)).

The idea of still having it as part of the SYN construction is that if it is a proper name or a loan word the matching with the document database could still take place.

The phenomenon of untranslatable constituents in compounds is of immediate interest for several source languages used in dictionary-based bilingual cross-language information retrieval. In a test with Finnish queries in an English document database (47 test queries from the CLEF 2001 collection) the effect is hardly noticeable in the average precision over recall values for all the queries. However, for some individual queries containing untranslated compound components the effect is greater (see Table 3).

	Average precision over 11 recall points	
	Untranslated components added inside the phrase construction	**Untranslated components moved outside the phrase construction**
Single query (CLEF 2001 topic no 49)	0,0355	0,1977
Single query (CLEF 2001 topic no 65)	0,0196	0,1419
All Finnish-English queries	0,3894	0,3962

Table 4: Test with Finnish-English CLIR, for untranslated components in compounds

Proximity operators and window size

Compounds in German, Swedish and Finnish are usually translated into English using a phrase construction, e.g., the German compound *Fussballweltmeisterschaft* is translated as "world soccer championship", the Swedish *Eutanasifall* as "incidents of euthanasia", and the Finnish *Tietokonevirus* as "computer virus" (3). The importance of using phrases in and problems with phrase identification for CLIR is discussed in Ballesteros and Croft (1997).

Windowing techniques and the span for the text window or window size have been discussed for use in information retrieval by Jacquemin (1996) and Haas and Losee (1994). In the study by Jacquemin a linguistic approach is taken and the author stresses that the words included in the window should be lexically and syntactically related or constitute of a correct partial noun phrase. Haas and Losee focus on an optimal window size and confirm that limiting the span of the words to between three and five is a good value. Zhou (1999) has documented phrasal terms in English documents. The findings show a clear predominance of two-word phrases (75.9%), while three-word phrases were not very common (14.4%). A window size of five words and a free word order would according to the findings of Zhou cover most (roughly 95%) of the phrases.

These results give some background knowledge on the processing of compounds for bilingual CLIR when the source language is a compound rich language and the target language in most cases uses a phrase instead of a joint compound. A direct translation by looking up a compound in a bilingual dictionary would give a phrase as translation equivalent, e.g., *Welthandelsorganisation* (world trade organisation). The translation can be marked as a phrase #2(world trade organisation). This meaning that the exact sequence of words, in this word order, has to match the document database, in order to retrieve relevant documents on this subject.

Enlarging the window size and allowing for any word order #uw5(world trade organisation) means that a document containing the phrase "organisation for world trade" would be retrieved. But we would not get a match for the phrases "trade organisation" or "world trade".

On the other hand, if the source compound word is split into the constituents *Welt*, *Handel* and *Organisation*, and these are grouped according to the rules in Section 4.2., e.g., *Welt+handel* and *Handel+s+Organisation* we could get direct translations of the component pairs, forming the phrases "world trade" and "trade organisation".

Proximity operators and window size can also be used in cases where no direct translations for the whole compound nor for constituent pairs can be obtained. If in the above case the split components: *Welt* (world, universe), *Handel* (trade, business), *Organisation* (organisation), and the combinations that can be made, are marked as phrases we would get the following combinations, using a synonym structure:

#syn(#uw5(world trade organisation) #uw5(world business organisation) #uw5(universe trade organisation) #uw5(universe business organisation))

Both the proximity operators #od and #uw have been used in cross-language information retrieval tests in the studies by Hedlund *et al.* (2001b; 2001c) for the CLEF campaigns 2000 and 2001. The actual effect of changing the word order operator and the window size is small. A test with Finnish - English bilingual CLIR (47 Finnish topics, English document database) showed no actual effect in the average precision values over all recall levels 0 to 100 for all the 47 queries (see Table 5).

	Average precision over 11 recall points	
Data set	**#od n (exact phrase length)**	**#uw 5 + n**
Individual query (CLEF 2001 no 45)	0,6417	0,6564
Individual query (CLEF 2001 no 48)	0,1280	0,1253
Individual query (CLEF 2001 no 63)	0,2671	0,3985
Individual query (CLEF 2001 no 86)	0,3113	0,3709
All Finnish - English queries (47)	0,3906	0,3962

Table 5: The effect of proximity operators with Finnish-English CLIR

The positive effect is noticeable only for three individual queries (queries 45, 63 and 86 in Table 5). The test, however, shows that the free word order operator (#uw) might be a better choice in an automated system since for 44 queries the result was better or equal and the largest negative result was an average precision drop of 0,0015 (query 48 in Table 5).

Effects of compound processing in bilingual cross-language information retrieval

To evaluate the effect of compound handling in an automated process for bilingual dictionary based CLIR several tests with German CLEF 2001 topics and the English document collection were performed. The test runs were performed by using different translation resources for the same 47 German topics (see also the study by Hedlund *et al.* (2001c). The resources were 1) a comprehensive dictionary, containing translations for many compounds, 2) the same dictionary, except that every compound word was first removed from it. This limited dictionary thus contains no direct translations for compound words. A direct translation through a comprehensive dictionary has a higher average effectiveness (see runs 2 and 5 in Table 6). However the test with a limited dictionary also indicates that the compound! h! andling process works fairly well.

In order to test the compound handling process further, additional tests were made (Table 6.). A baseline for German - English bilingual CLIR was established.

Run 1 in Table 6 is a baseline for German - English CLIR - no compound splitting or translation of components was performed. The compounds that could be translated directly using the translation dictionary were used in a similar way as for any other single word, that is, a synonym structure for translation alternatives was used.

Run 2 in Table 6 is a test run including the compound processing features described in Section 4. That is, splitting of compounds into constituents, normalisation of constituents to base forms, translation using a comprehensive dictionary and a structured query construction using an unordered proximity operator and a window size of 5 + n (n = spaces between words in the construction).

Test run 3 in Table 6 includes the best possible translation alternatives for the compounds in the query. This means that ambiguities in the translations provided by the dictionary were manually eliminated. It is, however, disambiguation based solely on the translation of individual compound words in the topic, not on phrases and multi-word concepts or semantic structures in the topic sentences.

Test run	Average precision over 11 recall values
1. Comprehensive dictionary no compound process	0,3520
2. Comprehensive dictionary + compound process	0,3830
3. Comprehensive dictionary (manual disambiguation of compounds)	0,3737
4. Limited dictionary no compound process	0,3057
5. Limited dictionary + compound process	0,3547

Table 6: Evaluation results for the effects of compound processing and dictionaries. The topic sets are the German-English CLEF 2001 queries (N=47).

Test runs 4 and 5 in Table 6, were performed with the limited dictionary. Run 4 does not include the compound process and since the dictionary does not contain any translations for compound words, no compounds were translated in this run. They are included in the query as such. Test run 5 was performed using the compound process as described in the case of test run 2.

As expected, the manual disambiguation of compounds, test run 3, performed well. Likewise, run 4, where compound translation was eliminated by not including compounds in the dictionary and by not including any compound handling process, got the lowest score. But surprisingly test run 2, where the compound process was tested using a comprehensive dictionary was the most effective, although the difference from test run 3 is very small. An explanation for this is that the compound process for some queries adds more weight to the compounds in a query using many alternative translations and combinations of components. Test runs 1 and 2 allow a comparison for how well the compound process works as a complement to the comprehensive dictionary. The result for the run with the process included is clearly better. This is also true for test runs 4 and 5 where the effect of the process is compared using a limited dictionary not containing translations for compounds. The conclusion is! t! hat the compound process is effective for handling compounds.

Discussion and conclusion

The importance of compound processing in dictionary-based cross-language information retrieval was investigated in this study. We found that around ten percent of remaining content words in running text are compounds for the three languages studied. This can be reformulated even more impressively: it means that more than twenty percent of morphemes in running text are in compounds! Compounding is a way to form new words and concepts. Compounds are often occasional. Studying compound processing in cross-language information retrieval is important, since we know that translation dictionaries, even comprehensive ones, are limited and cannot hold entries for all compounds. Compounds and their features are described in this study from the information retrieval and cross-language information retrieval point of view, and therefore not all linguistic features for compound formation are considered in this study. The linguistic reality is much more complex than described here.

The compound handling process is implemented as a component into an automated cross-language information retrieval system. This makes it possible to evaluate the final process as well as individual steps in the process using conventional evaluation measures, the average precision over different recall levels. The topics and the document database were provided by the international CLEF evaluation forum, which ensures the availability of the test material.

The solutions proposed in this study rely on the morphological and syntactic structure of compounds. For normalisation of compound components the features of fogemorphemes in Germanic languages, as well as the inflected components or components in the form of word stems have to be taken into account. The left- or right-branching structure for the compound components form the linguistic basis for the novel *grouping strategy* with pairwise combination and translation of consecutive components.

The query construction phase has to take into account the features in the target language. This involves the use of proximity operators and phrase construction, in this study due to the fact that the target language in the document database language is English.

The evaluation results of this study indicate that compound processing as a whole has a clearly positive effect on retrieval results. Each step in the process has a relatively small impact on the result. However, summing up each of them adds to the positive effect of the system. One can naturally discuss whether general conclusions should be drawn on the basis of this relatively small test sample. Test situations do not completely cover the complexity of natural language information retrieval, and additional test sets with different topics and different document databases could yield other results. In this study the document database contains newspaper text, which implies that the topics created are adapted to this content. Scientific articles or other types of texts would use different types of topics and the concepts and vocabulary would be different.

The relative importance of a system processing compounds for dictionary based cross-language information retrieval naturally depends on the number of compounds present in queries as well as the available translation resources. A direct translation of compounds is clearly more precise and therefore effective. Using comprehensive dictionaries we are able to directly translate more compounds than with limited and small translation dictionaries. But on the other hand, no dictionary can hold entries for all occasional compounds in a language, so there will always be a need for systems handling compounds even if their function in some cases is to work as a complement to a direct translation.

The analysis of compounds and their features from the cross-language information retrieval point of view in this study could be extended to semantic analysis to determine, for example, paradigmatic relations and to syntactic analysis to determine syntactic structures of phrases. This kind of information could be in use in information retrieval and cross-language information retrieval applications to determine valuable keys. The descriptions provided by the morphological analysis program like part of speech, gender, etc. could also be valuable for disambiguation of translation alternatives.

Notes

(1) TREC, Text REtrieval Conference, http://trec.nist.gov/
CLEF is a forum for evaluating cross-language information retrieval solutions on European languages, http://www.iei.pi.cnr.it/DELOS/CLEF/

(2) The operators referred to are used in InQuery retrieval system. OD refers to "Ordered window" or strict word order, UW refers to "unordered window" or free word order.

(3) The examples are from the year 2001 CLEF topics in German, Swedish and Finnish

(4) The query formulation files for the runs in this study can be obtained from the author, e-mail: turid.hedlund@shh.fi

Acknowledgements

The *InQuery* search engine was provided by the Center for Intelligent Information Retrieval at the University of Massachusetts.
ENGTWOL (Morphological Transducer Lexicon Description of English): Copyright (c) 1989-1992 Arto Voutilainen and Juha Heikkilä.
FINTWOL (Morphological Description of Finnish): Copyright (c) Kimmo Koskenniemi and Lingsoft plc. 1983-1993.
GERTWOL (Morphological Transducer Lexicon Description of German): Copyright (c) 1997 Kimmo Koskenniemi and Lingsoft plc.
SWETWOL (Morphological Transducer Lexicon Description of Swedish): Copyright (c) 1998 Fred Karlsson and Lingsoft plc.
TWOL-R (Run-time Two-Level Program): Copyright (c) Kimmo Koskenniemi and Lingsoft plc. 1983-1992.
MOT Dictionary Software was used for automatic word-by-word translations. Copyright (c) 1998 Kielikone plc, Finland.

This research is part of the research project *Query structures and dictionaries as tools in concept-based and cross-lingual information retrieval* funded by the Academy of Finland (Research Projects 44703; 49157).

A special expression of gratitude to the FIRE research group for many comments and to Mr. Heikki Keskustalo and Ms. Eija Airio for valuable programming assistance.

References

- Akmajian, A., Demers, R., Farmer, A. and Harnish, R. (1997) *Linguistics: An introduction to language and communication.* Cambridge, MA: MIT Press.
- Ballesteros, L. and Croft, B. (1997) "Phrasal translation and query expansion techniques for cross-language information retrieval". *Working Notes of the AAAI Spring Symposium on Cross-Language Text and Speech Retrieval,* Stanford, CA: Stanford University. Available at http://www.ece.umd.edu/medlab/filter/sss/papers/ballesteros.ps [Accessed 20 January 2002]
- Fleischer, W. and Barz, I. (1992) *Wortbildung der deutschen Gegenwartssprache*. Tübingen: Max Niemeyer Verlag.

- Haas, S. W. and Losee, R. M. (1994) Looking in text windows: Their size and composition *Information Processing & Management*, **30**(5), 619-629.
- Hedlund,T., Pirkola, A. and Järvelin, K. (2001a) Aspects of Swedish morphology and semantics from the perspective of mono- and cross-language information retrieval. *Information Processing & Management*, **37**(1), 147-161.
- Hedlund, T., Keskustalo, H., Pirkola, A., Sepponen, M. and Järvelin, K. (2001b) Bilingual tests with Swedish, Finnish and German queries: Dealing with morphology, compound words and query structuring. In Carol Peters (Ed.) *Cross-Language Information Retrieval and Evaluation: Proceedings of the CLEF 2000 Workshop, Lecture Notes in Computer Science*, 2069, pp. 211-225. Heidelberg: Springer.
- Hedlund, T., Keskustalo, H., Pirkola, A., Airio, E. and Järvelin, K. (2001c) "UTACLIR @ CLEF 2001". *Working Notes for CLEF 2001 Workshop*. Sophia Antipolis: European Research Consortium for Informatics and Mathematics.Available at http://www.ercim.org/publication/ws-proceedings/CLEF2/hedlund.pdf [Accessed 20 January 2002]
- Jacquemin, C. (1996) What is the three that we see through the window: A linguistic approach to windowing and term variation. *Information Processing & Management*, **32**(4), 445-458.
- Kraaij, W. (2001) "TNO at CLEF 2001: Comparing translation resources". *Working Notes for CLEF 2001 Workshop*. Sophia Antipolis: European Research Consortium for Informatics and Mathematics. Available at http://www.ercim.org/publication/ws-proceedings/CLEF2/kraaij.pdf [Accessed 20 January 2002]
- Malmgren, S. G. (1994) *Svensk lexikologi. Ord, ordbildning, ordböcker och orddatabaser. [Swedish lexicology. Words, word formation, dictionaries and word databases.] Lund: Studentlitteratur.*
- *Pirkola, A. (1998) The effects of query structure and dictionary setups in dictionary-based cross-language information retrieval. Proceedings of the 21st ACM/SIGIR Conference, pp. 55-63 New York, NY: Association for Computing Machinery*
- *Porter, M.F. (1980) An algorithm for suffix stripping. Program, 14, 130-137.*
- *Riplinger, B. (2001) "Mpro-IR in CLEF 2001." Working Notes for CLEF 2001 Workshop. Sophia Antipolis: European Research Consortium for Informatics and Mathematics. Available at http://www.ercim.org/publication/ws-proceedings/CLEF2/ripplinger.pdf [Accessed 20 January 2002]*
- *Savoy, J. (2001) "Report on CLEF-2001 experiments". Working Notes for CLEF 2001 Workshop. Sophia Antipolis: European Research Consortium for Informatics and Mathematics. Available at http://www.ercim.org/publication/ws-proceedings/CLEF2/savoy.pdf [Accessed 20 January 2002]*
- *Smeaton, A. F. (1999) Using NLP or NLP resources for information retrieval tasks. In Tomek Strzalkowski (Ed.) Natural language information retrieval. Dordrecht: Kluwer Academic Publishers.*
- *Sparck Jones, K. (1999) What is the role of NLP in text retrieval? In Tomek Strzalkowski (Ed.) Natural language information retrieval. Dordrecht: Kluwer Academic Publishers.*
- *Strzalkowski, T. (1995) Natural language information retrieval. Information Processing & Management,* ***31****(3), 397-417.*
- *Strzalkowski, T., Lin, F., Wang, J. and Perez-Carballo, J. (1999). In T. Strzalkowski (Ed.) Natural language information retrieval. Dordrecht: Kluwer Academic Publishers.*
- *Warren, B. (1978) Semantic petterns of noun-noun compounds. Göteborg: Acta Universitatis Gothoburgensis. (Gothenburg studies in English 41)*

How to cite this paper:

Hedlund, T (2002) "Compounds in dictionary-based cross-language information retrieval" *Information Research*, **7**(2) [Available at http://InformationR.net/ir/7-2/paper128.html]

Last updated: 20th January, 2001

Check for citations, using Google Scholar

Contents 10518 **Web Counter** **Home**