Morphological Typology of Languages for IR

Ari Pirkola

University of Tampere, Department of Information Studies

Email: pirkola@cc.jyu.fi

Published in Journal of Documentation 57 (3), 330-348.

Abstract. This paper presents a morphological classification of languages from the IR perspective. Linguistic typology research has shown that the morphological complexity of each language of the world can be described by two variables, index of synthesis and index of fusion. These variables provide a theoretical basis for IR research handling morphological issues. A common theoretical framework is needed in particular due to the increasing significance of cross-language retrieval research and CLIR systems processing different languages. The paper elaborates the linguistic morphological typology for the purposes of IR research. It is studied how the indices of synthesis and fusion could be used as practical tools in mono- and cross-lingual IR research. The need for semantic and syntactic typologies is discussed. The paper also reviews studies done in different languages on the effects of morphology and stemming in IR.

1. Introduction

There are at least 4000 languages in the world [1, 2]. The precise figure depends on, for example, where to draw a line between a dialect and a distinct language.[1] Languages are classified on the basis of their supposed genetic relationships into language families on the one hand, and on linguistic grounds on the other. The language families include *Indo-European* (the largest family including the western languages), *Finno-Ugric* (including Finnish and Hungarian) and *Sino-Tibetan* (including Chinese). Some languages are difficult to include in the established families, and they are called isolates (e.g., Japanese). The traditional morphological typology distinguishes 4 language types. The syntactic typology by Greenberg divides languages into different types on the basis of the order of sentence elements [4].

This paper presents a morphological classification of languages from the standpoint of IR.
The paper considers morphology associated with *texts*, i.e., written form of languages. IR research is an international research area. Monolingual research is performed in different languages. Cross-language retrieval has become an important research area in a global scale [5, 6, 7]. It is difficult to follow and make research if one does not master the languages involved. This difficulty could be

relieved by a common linguistic framework applicable to IR. This study collects the results of morphological typology research done in linguistics and combines the results into a theoretical framework for IR research. It is shown in the present paper that the variation in morphological properties among world's languages is high. It is, however, also shown that the same morphological processes affect all world's languages and all languages can be described using the same morphological variables. This paper also discusses lexical-semantic variation in world's languages, but the theoretical framework only covers the structure of words.

The aim of the paper is also to provide practical tools for IR research, in particular for text retrieval research. *Text retrieval* refers to retrieving documents from text databases, i.e., electronic collections of documents, such as magazine, journal, and newspaper articles. Morphological typology research has shown that it is possible to describe the morphological complexity of each language using two variables, *index of synthesis* and *index of fusion* [8, 9, 10]. The former describes the amount of affixation in an individual language, and the latter the ease with which affixes can be segmented in words in a language. It is proposed in the present paper that, for each language, these variables could be utilized in IR within a language and across languages as practical tools in system development and evaluation.

The rest of this paper is organized as follows. Section 2 considers the central concepts of morphology. Section 3 considers the most important morphological phenomena related to information retrieval, i.e., inflection, derivation, and compound words, and reviews studies done on the effects of stemming in IR. Section 4 presents the traditional morphological typology as well as the recent one based on the variables of index of synthesis and index of fusion. In Section 5 the recent morphological typology is subcategorized for the purpose of IR. Section 6 considers how languages differ in inflection, derivation and the frequency of compound words. Section 7 discusses how the indices of synthesis and fusion could be utilized in empirical IR research and system development. In section 8 the need for semantic and syntactic typologies is discussed. Section 9 presents conclusions.

2. Core concepts of morphology

Morphology is the field of linguistics which studies word structure and formation. It is composed of *inflectional morphology* and *derivational morphology* [9, 11, 12]. *Inflection* is defined as the use of

[1] Saussure discusses the difference between a language and a dialect [3].

morphological methods to form inflectional word forms from a *lexeme*[2]. Inflectional word forms indicate grammatical relations between words. Derivational morphology is concerned with the *derivation* of new words from other words using derivational affixes. Compounding is another method to form new words. *A compound word* (or a *compound*) is defined as a word formed from two or more words written together. The component words are themselves independent words (free morphemes).

A *morpheme* is the smallest unit of a language which has a meaning [9, 15]. Morphemes are classified into (1) *free morphemes* and (2) *bound morphemes*. Free morphemes appear as independent words (in the form of their *allomorphs*, see below). Free morhemes are further divided into *lexical morphemes* and *grammatical morphemes*. The former are semantically significant words while the latter are function words. Bound morphemes do not constitute independent words, but are attached to other morphemes or words. Bound morphemes are also called *affixes*. Affixes are classified into *inflectional affixes* and *derivational affixes* on the one hand, and into *prefixes*, *suffixes*, and *infixes* on the other. Prefixes are attached to the beginning of words and suffixes to the end of words. Infixes, which are affixes attached within other morphemes, are used only in some languages, as in some *native American languages*.

The previous definitions can be illustrated with the following examples. In *English*, {*red*}[3], {*house*}, and {*when*} are all free morphemes. The first two are lexical morphemes whereas the morheme {*when*} is a grammatical morpheme (a function word). In speech and text morphemes are represented by *morphs*. *Allomorphs* are morph variants of a given morpheme. For example, in *Finnish* {*kalA*} (meaning *fish*) is a free morpheme, which has the allomorphs *kala* and *kalo*. An example of a *Finnish* bound morpheme is {*ssA*}. It has two allomorphs, *ssa* and *ssä*. These are suffixes which indicate the inessive case. They cannot stand as independent words but must be in combination with other morphemes or words. For example, the allomorph *ssa* can be attached into the allomorphs *kala* and *kalo*. This addition gives the words *kalassa* and *kalo(i)ssa*. In English the suffix *s* indicates a plural form. An example of a prefix and its use is the derivational prefix *un* in the word *unhappy*.

[2]A *lexeme* is a set of word forms which belong together [13], or a word considered as a lexical unit, in abstraction from the specific word forms it takes in specific constructions [14]. For example, the lexeme *sing* has the following *word forms* or *inflectional forms*: *sing*, *sang, sung, sings, singing*.

[3]The parentheses {} are used to denote morphemes.

Suffixes are more common than prefixes in world's languages [9]. There are many languages that almost entirely use suffixes in inflection and derivation, and they are also called *suffix languages*. For instance, in *Finnish* inflected word forms are formed only by means of suffixes. In derivation prefixes are also used but they are not common. The order of appearance of the derivational and inflectional suffixes is the same in most suffix languages, that is, a stem is followed by derivational suffixes and these are followed by inflectional suffixes. *Prefix languages* are not so common as suffix languages. *Thai* language and *Swahili* are examples of prefix languages. In prefix languages a stem is usually preceded by derivational prefixes, and these are preceded by inflectional prefixes.

3. Morphological phenomena in IR

The three main morphological phenomena, i.e., *inflection*, *derivation*, and *compound words*, all affect the effectiveness of text retrieval. 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 (or vice versa). For effective text retrieval, morphological processing is needed in most languages to handle inflected word forms. The morphological processing may be simple manual truncation or automatic *stemming* or *normalization* (*lemmatization*). In *stemming* affixes are removed from word forms [16]. The output is a common *root* or *stem* of different forms, which is not necessarily a real word. In lexicon-based *morphological analysis* word forms are normalized, i.e., word forms are turned into base forms which are real words. Morphological analysis also allows the splitting of compounds into their component words.

In text retrieval it has to be decided whether *derivatives* and their roots are conflated into the same form (or whether just inflected words are handled). The extent of derivation as well as morphological and semantic properties of derivatives vary between languages. In languages rich with *compound words* it must be decided whether compounds will be decomposed. If compounds are not decomposed, the component words of the compounds are not retrievable. However, in compositional compounds in particular the last component is often a valuable search key, as it is usually a hypernym of the full compound [17]. For instance, a (Finnish) request may concern sugars with *sokeri* (*sugar*) being one search key. If compounds are not split, the names of all sugar types should be listed: *hedelmäsokeri*, *ruokosokeri*, *rypälesokeri* (*fruit sugar*, *cane sugar, grape sugar*), etc. However, when compounds are split, one search key only, that is, *sokeri*, is enough. Compound

splitting is also important in dictionary-based cross-language retrieval. The translation of component words separately is often useful, because dictionaries may not include full compounds as such but only their components [18].

In *Japanese*, *Chinese*, and *Korean* texts there are no obvious word boundaries [19].[4]
Term segmentation is a process in which a string of characters is divided into words and other meaningful units [22]. The main problem with segmentation is that there are often several legitimate ways to segment a sentence due to various morphological, syntactic, and semantic factors [22, 23, 24, 25]. Segmentation is associated with *compound noun identification* which is the same kind of task as phrase identification in English [25].

As shown in this paper, for each language the decisions associated with morphological processing basically require three kinds of information, i.e., information on the degree of morphological synthesis and fusion as well as semantic fusion. It is possible to quantify this information using the measures of *index of synthesis* and *index of fusion* (Sections 4-5). It is proposed in this paper (Section 7) that the indices of synthesis and fusion could be used as guides for morphological processing decisions. The variables are computable allowing straightforward comparisons between many types of situations associated with IR morphology.

Due to stemming and normalization three kinds of benefits may be gained [26]. First, a user does not need to worry about morphology and truncation, because different forms of the key are automatically conflated into the same form. Particularly in the languages with complex morphology, such as Slovene and Finnish, it may be difficult to form a good query without morphological programs [17, 27]. Second, stemming and normalization may cause storage savings. This was shown by Alkula who used a Finnish test collection in her study and found that the number of index terms decreased substantially due to normalization [28]. This resulted in storage savings, though the number of addresses in the index was increased. A remarkable reduction in the number of index terms was also achieved when, besides normalization compounds were split, though compound splitting increases the number of index terms. Third, research has shown that stemming and normalization improve retrieval performance. *Recall* especially can be expected to improve as a larger number of potentially relevant documents are retrieved [29, 30].

[4] Large and Moukdad discuss the language barrier problem on the Web, including the issues related to different writing systems (scripts) [20]. Different writing systems are described in [21].

Research done in different languages has shown that stemming also improves *precision*. In his study Krovetz tested both an inflectional and a derivational stemmer in an *English* test collection [31]. Both stemming methods resulted in precision improvement compared with the situation where no stemming was performed. The performance improvements were significant in particular in the case of short documents. The derivational stemmer was more effective than the inflectional stemmer at high precision levels. Hull tested the effects of stemming in a large *English* test collection (180,000 documents) and found that stemming improved precision for short queries [29]. Savoy found that conflating plural nouns had positive effects on precision in *French* text retrieval [32]. Kalamboukis developed a stemming algorithm for modern *Greek* [33]. The algorithm was based on a suffix list, and quantitative (minimum stem lenght) and qualitative constraints. The researcher reported a clear improvement in precision due to stemming. Modern Greek has rich inflectional system, e.g., there are 41 inflectional suffixes for nouns. Abu-salem et al. tested Root, Stem, Word and Mixed indexing techniques in *Arabic* information retrieval [34]. The Root technique was reported to give the best precision. Arabic language is a root-based language with a *root* typically consisting of three consonants [9, 34]. *Stems* are longer forms which are formed according to fixed patterns. *Words* consist of stems and affixes.

A stemmer by Popovic and Willett for *Slovene* language contained a suffix list of over 5000 suffixes [27]. For Slovene, a sophisticated stemmer with a large suffix list is needed because of its rich morphology. For example, a noun referring to a person or an object has six features in a grammatical case and can appear in singular, plural and dual forms (see Section 6). The researchers found that stemming resulted in a significant increase in retrieval effectiveness. The effectiveness was measured as the number of relevant retrieved documents at document cut-off value 10. Ekmekcioglu and Willett used the same evaluation measure and showed that stemming increased retrieval effectiveness in *Turkish* retrieval [35].

The results of stemmming studies presented above are consistent, showing that in many languages stemming results in average performance improvements. Nonetheless, for single queries stemming and morphological analysis may be harmful, because longer word forms are more precise expressions than stems and base forms. For instance, in *Finnish* the inflectional forms of the lexeme *kuusi* in the sense of *spruce* and the inflectional forms of the lexeme *kuusi* in the sense of the numeral *six* are different. In normalization these are conflated into the same form (*kuusi*). Thus the unambiguous forms are turned into an ambiguous form. The Porter stemmer gives the same interpretation for the words *general*, *generous*, *generation*, and *generic* [29]. Normalization in the

case of *inflectional homonymy* where two (or more) lexemes share the same inflectional forms causes extraneous words (base forms) to be stored in a database index. In *Finnish*, the form *voin*, for example, gives the base forms *voida* (the base form of the verb *can*) and *voi* (meaning *butter*).

The conflation errors associated with stemming are caused either by *overstemming* or *understemming* [30, 36]. In overstemming the stem is too short, and words with different meanings are conflated to the same stem, e.g., *general* and *generation*. In understemming the stem is too long, and words with similar meanings are not conflated. If a stemmer is set towards overstemming, recall can be expected to increase, while choosing the policy of understemming enables users to do specific searches [30]. The concepts of overstemming and understemming do not apply to morphological analysis which gives base forms as its output. The effectiveness of morphological analysis is limited by the size of a lexicon [29].

4. Morphological typology

The traditional morphological typology dates back to the nineteenth century. It distinguishes three language types, i.e., isolating, agglutinative, and fusional languages [8, 9, 10]. This typology was later supplemented by the fourth language type, polysynthetic languages, in particular to explain the morphology of some *native American languages*. The four morphological types are ideal types rather than practical categories. There are languages that are close to some ideal type, e.g., *Chinese* and *Vietnamese* (isolating languages) and *Turkish* (an agglutinative language). Most languages, however, are mixed types sharing features of different ideal types.

Isolating languages have no morphology at all. The correspondence between words and morphemes is one-to-one. In *Vietnamese* words appear in the same invariable forms independent of their grammatical functions. This is shown in the following sentence [8]:

Khi toi den nha ban toi, chung toi bat dau lam bai.[5]
When I come house friend I 'plural' I begin do lesson (begin = bat dau)

'When I came to my friend's house, we began to do lessons.'

In *agglutinative languages*, the boundaries separating one morpheme from another in a word are

[5]Transcripted to Roman letters.

clear-cut, and morphemes are easily segmentable. In inflection affixes are added to invariable word stems. A classic example is *Turkish*. The *Turkish* word form *köpekleri* can be analyzed into the following morphemes: *köpek* (dog), *ler* (plural suffix), *i* (accusative suffix).

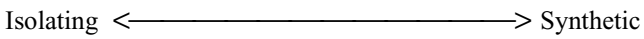
In *fusional languages*, there are no clear-cut boundaries between morphemes in a word. A monomorphemic word may consist of two or more meaning units. Typical examples of fusional words are the strong verbs of *Germanic languages*. For instance, the monomorphemic word *took* in *English* denotes two things, that is, the meanings 'to take' and to 'past tense'.

In *polysynthetic languages*, a word may consist of a large number of lexical and bound morphemes. A word consisting of several morphemes may form an entire sentence. Thus the difference between a word and a sentence is sometimes obscure in polysynthetic languages. The *Inuit (Eskimo)* language is often regarded as a typical polysynthetic language.

Most world's languages are mixed types. For instance, in *English* grammatical relations are shown mainly by means of prepositions. This resembles the pattern of isolating languages. The derivational and inflectional morphologies of English are in part agglutinative and in part fusional. For instance, the word *fortunate* (fortune + ate) is fusional. The form *fortunately* (fortunate + ly) is agglutinative.

Recent morphological typology is based on the traditional typology, but instead of distinguishing four distinct language types it operates with two independent variables, *index of synthesis* and *index of fusion* [8, 9, 10]. These variables seem to be useful also for IR as discussed below.

Index of synthesis (IS) refers to the amount of affixation in a language, i.e., it shows the average number of morphemes per word in a language. It can be illustrated by means of a scale, the end points of which are an isolating language and a (poly)synthetic language, as follows:



Each language falls on a given point on the scale. The languages in which synthesis dominates are on the right side and those with weak morphology on the left side on the scale.

Index of fusion (IF) refers to the ease with which morphemes can be separated from other

morphemes in a word. Agglutinative languages have low index of fusion, and in fusional languages it is high. In agglutinative words segmentation can be performed readily due to clear morpheme boundaries. In fusional words segmentation is difficult or impossible. Index of fusion also can be illustrated by means of a scale. The extremes are now agglutinative and fusional languages.

Agglutinative <——————————————> Fusional

All languages except for isolating languages fall between the two extremes. In isolating languages, by definition, there are no agglutinative or fusional morphological processes.

Table 1. Index of synthesis

Language	Index of synthesis
Vietnamese	1,06
Yoruba	1,09
English	1,68
Old English	2,12
Swahili	2,55
Turkish	2,86
Russian	3,33
Inuit (Eskimo)	3,72

Table 1 presents index of synthesis for eight languages [9]. For each case, the figures are calculated on the basis of 100 words of an unrestricted text sample. *Vietnamese* is close to an ideal isolating language and its index of synthesis is close to 1.0. *Inuit* is highly polysynthetic language with its index of synthesis being high. The other sample languages fall between Vietnamese and Inuit.

5. Morphological typology for IR

In this section the indices of synthesis and fusion are defined for the purpose of IR[6]. Index of synthesis can be divided into the following cases which are defined as follows:

- inflectional index of synthesis (IIS) - the number of inflectional morphemes per the total number of words (in a text sample)
- derivational index of synthesis (DIS) - the number of derivational morphemes per the total

[6]The classification is in part based on that of Greenberg's [37].

number of words

- compound index of synthesis (CIS) - the number of compound morphemes (components) per the total number of words

The following example sentences (*English*, *Finnish*) illustrate how IIS computed.

He was driving his car.

Hän ajoi autoansa.

The English sentence includes five words and one inflectional morpheme (*ing*); the IIS is 1/5. The corresponding Finnish sentence includes three words and three inflectional morphemes, i.e., the past tense suffix *i* in the word *ajoi*, and the suffixes *a* (accusative suffix) and *nsa* (genitive suffix) in the word *autoansa*. Thus, the IIS is 3/3. To get comparable figures for different languages (Section 7) the indices discussed in this section should be computed on the basis of parallel texts, as was done in this example (see parallel texts in Section 6).

Fusional changes can occur on morphological and semantic levels. Here *fusion* (both morphological and semantic) is defined as a process where the end product (a fused word) is something else than the sum of components. On a (sheer) *morphological level* the character set of the fused word is not exactly the same as the character sets of the component morphemes put together. *Strong verbs* of *Germanic languages* represent an extreme case. The *English* form *took* is monomorphemic, but denotes two things, that is, the meanings 'to take' and 'past tense'.

The morphological index of fusion can be divided into the following cases which are defined as follows:

- inflectional index of fusion (MorphIIF) - the number of fused inflected words per the total number of words
- derivational index of fusion (MorphDIF) - the number of fused derived words per the total number of words
- compound index of fusion (MorphCIF) - the number of fused compound words per the total number of words

Table 2 presents examples of agglutinative and fusional words (on a morphological level). The examples are from *English* (inflection and derivation) and *Swedish* (compounds). Swedish is a language of high frequency of compounds. The cases of *house* + *s* ---> *houses*, *read* + *er* ----> *reader*, and *järn* + *industri* ---> *järnindustri* represent agglutination. No structural changes occur when the affixes *s* and *er* are attached into the word stems *house* and *read*. The compound word *järnindustri* is formed in the same way without structural changes. The words *distributing* (*distribute* + *ing*), *cylindrical* (*cylinder* + *ical)*, and *gatubelysning* (*gata* + *belysning*) represent fused words. Now the product words of morphological processes differ from the cases where the components were put together as such.

Table 2. Examples of agglutinative and fusional words

Morphological process	Agglutination	Fusion
Inflection	house, houses	distribute, distributing
Derivation	read, reader	cylinder, cylindrical
Compounding	järnindustri (iron industry) järn (iron) + industri (industry)	gatubelysning (street lighting) gata (street) + belysning (lighting)

Compounding and derivation are often associated with meaning changes, and on a *semantic* level the index of fusion can be divided into the following types for IR:

- semantic index of fusion in compounding (SemCIF) - the number of fused compound words per the total number of compound words
- semantic index of fusion in derivation (SemDIF) - the number of fused derived words per the total number of derived words

On the semantic level, the meaning of a *compound expression* may be the same or different than the sum meanings of the component words. In the former case, compounds are called *transparent* or *compositional* [15, 38]. In the latter case, they are called *opaque* or *non-compositional*. The meaning of a transparent compound can be deduced on the basis of its component words (as far as the meanings of the component words are known). The meaning of an opaque compound cannot be deduced on the basis of its components. In the case of *derivatives*, *transparency* refers to the fact

that the meaning of a derivative is predictable on the basis of the meanings of its component morphemes. The meaning of an *opaque derivative* is unpredictable. In the cases of semantic fusion of compounds and derivatives, the character set of the fused word may or may not be the same as the character sets of the components put together.

Table 3 shows examples of transparent and opaque derivatives and compounds. Opaque derivatives and compounds may be originally created as opaque words or their meanings may change in the course of time. Sometimes the relationship of two forms can be established only through etymological research. The word *regard* is a derivative of the word *guard* [38]. Its meaning cannot be predicted on the basis of the meanings of the morphemes *re* and *gard*. In addition to semantic fusion morphological fusion has occurred in the word *regard*. In the same way the *French* compound *debonnaire* (*gentle*) has *lexicalized* into an independent lexeme. Etymological research has shown that it is a derivative of the phrase *de bonne aire* (meaning *of good stock*). The *Swedish* compound *jordgubbe* (*strawberry*) is an opaque compound - its meaning cannot be derived from the meanings of the components *jord* (*earth*) and *gubbe* (*old man*). The words *reader* and *kärnkraft* are transparent words. The addition of the affix *er* into the word *read* gives the word *reader* whose meaning is predictable ('read' and 'actor'). The same holds for the compound *kärnkraft* (*nuclear power*) whose meaning is a sum meaning of the meanings of the components *kärna* (*nucleus*) and *kraft* (*power*).

Table 3. Examples of transparent and opaque words

Morphological process	Transparent, Agglutination	Opaque, Fusion
Derivation	read, reader	guard, regard debonnaire, de bonne aire
Compounding	kärnkraft (nuclear power) kärna (nucleus) + kraft (power)	jordgubbe (strawberry) jord = earth, gubbe = old man

6. Differences in inflection, derivation, and compounding

In world's languages, the most usual *inflectional categories* of nouns are *number*, *a grammatical case*, and *a grammatical gender*. These are the main morphological phenomena that affect the

indices of inflectional synthesis and fusion.

In most languages there are two *morphosyntactic features* (*terms*) in the category of *number*, that is, *singular* and *plural*. Some languages have *singular*, *dual* and *plural*. In many languages singular is unmarked and plural is marked using a specific plural suffix. In *English* as in many other *Germanic languages* plural forms are normally marked using the suffix *s*. In the case of a language possessing several features in a grammatical case (see below) the situation is more complex since there may be several plural suffixes.

Grammatical relations can be shown using a word order, particles (such as prepositions), and *a grammatical case*. The morphological complexity of a language depends to a great extent on the method the language uses and on the number of morphosyntactic features in the category of case. In *English* grammatical relations are indicated by means of prepositions, only genitive case is marked (by a suffix). Because (for nouns) in addition to genitive forms only plural forms are marked, in English index of synthesis is relatively low (Table 1).

Table 4 shows the number of morphosyntactic features in the category of case for 8 languages [1]. *Hungarian* has 21 features. In *English* there are only 2 features (nominative and genitive; genitive is marked). *Finnish* represents a language of high index of synthesis (not shown in Table 1). This is in particular due to the high number of morphosyntactic features in the category of case (14 features). Because different affix types (number, affixes of different case features, and clitics) can be combined with one another in a single word, the number of word forms that a given Finnish lexeme may take is very high. It has been estimated that a Finnish noun has at least 2,200 word forms [13]. Even though many of these are only theoretical, the number of word forms used in everyday life is still high. The concept of grammatical case is not relevant to all languages (languages with weak inflectional morphology, e.g., many *Asian* languages).

Table 4. The number of morphosyntactic features in a grammatical case for 8 languages

Language	Number of features in case
English	2
Finnish	14
German	4
Hungarian	21
Lithuanian	7
Russian	6
Sanskrit	8
Serbo-Croat	7

Many languages possess *a grammatical gender*. *Germanic languages* typically have two or three genders. The definite form of a word depends on its gender. For instance, *Swedish* possesses two genders, *gender uter* and *gender neuter*. The definite suffixes for *gender uter* words are *en* and *n* and for *gender neuter* words *et* and *t* [39].

In some languages word inflection is associated with the inflection of word stems, e.g., *Welsh* [40] and *Finnish* [41]. This represents the case of inflectional fusion. The lexeme *käsi* (meaning *hand*) in *Finnish* has five *allomorphs* or *inflectional stems* [13]. These are listed below. As shown, different suffixes are attached into different stems.

käsi + kin (clitic; *also a hand*)
käte + nä (essive suffix; *as a hand*)
käde + n (genitive singular suffix; *hand's*)
kät + ten (genitive plural suffix; *hands'*)
käs + i + ssä (i = plural suffix, ssä = inessive suffix; *in the hands*)

World's languages differ remarkably from each other in the frequency of *derivatives* and *compounds* [38]. Compounds are common, for example, in *German*, *Dutch*, *Finnish*, and *Swedish*. *German* is also characterized by high frequency of derivatives. In German, compounds and derivatives are typically transparent [38]. In *English* and *French* derivatives and compounds are not so common. English and French are more opaque in their natures. A German compound is often translated by a phrase or a single word in English and French. The following sample words and

parallel texts (CLEF Topic 015)[7] in *German*, *English*, and *French* illustrate the situation . The same text is presented in *Finnish* in Appendix to illustrate different database index representations.

Bahnhof ('railway yard')	*Railway station*	*Gare*
Erdteil ('earth part')	*Continent*	*Continent*
Sprachwissenschaft ('language science')	*Linguistics*	*Linguistique*

In German transparent derivatives are common. German derivatives often correspond to phrases or single words in English and French, as shown below.

Ursache ('orignal matter')	*cause*	*cause*
Eintreten ('in come')	*enter*	*entrer*

Welche Faktoren *beeinträchtigen* die *Wettbewerbsfähigkeit* der europäischen Industrie auf den *Weltmärkten*?

What are the factors that damage the *competitiveness* of European industry on the *world's markets*?

Quels sont les facteurs qui nuisent à la *compétitivité* de l'industrie européenne sur les *marchés mondiaux?*

The German compound *Wettbewerbsfähigkeit* consists of the components *Wettbewerb(s)* (*competion*) and *Fähigkeit* (*potency*) and is translated by a single word in English *(competitiveness*) and French (*compétitivité*). The compound *Weltmärkten* is translated by a phrase in English (*world's markets)* and French (*marchés mondiaux*). The word *beeinträchtigen* (*damage*) is a derivative word containing the derivative prefixes *be* and *ein*.

7. The use of ISs and IFs in information retrieval

Because morphology is essential in IR, morphological phenomena have considerable effects on retrieval effectiveness. In languages of low inflectional IS (IIS) and inflectional IF (MorphIIF), inflection does not interfere matching to the same degree as in the languages of high IIS and

[7] CLEF- Cross-Language Evaluation Forum, http://www.iei.pi.cnr.it/DELOS/CLEF

MorphIIF. Retrieval can be expected to be more effective in these languages. Also, the costs of constructing effective stemmers/morphological analysers are lower for languages of low IIS and MorphIIF. In languages of high IIS and MorphIIF simple matching and indexing techniques are insufficient.

Whether derivationally related words and compounds should be handled depends in particular on the semantic DIF (SemDIF) and CIF (SemCIF) of a language. Low SemDIF (SemCIF) indicates high relative frequency of transparent derivatives (compounds) and suggests that handling of derivatives (compounds) would be useful. Low derivational and compound IS (DIS and CIS) as well as low derivational and compound IF (MorphDIF and MorphCIF) suggest that one may dispense with the morphological processing of derivates and compounds or that the costs of morphological processing are low.

The different morphological and semantic IS and IF variables could be used as practical tools in IR in different kinds of situations within one language and across languages. For instance, within one language they could be used:

- To predict the effectiveness of morphological processing
- To predict the effort required to construct effective stemmers/morphological analysers
- To show the problem areas of morphological processing in IR
- To predict the effects of morphological processing on the effectiveness of IR

The situations where the IS and IF variables could be used across languages involve the following:

- Comparing the results of monolingual IR experiments between different languages
- Comparing the results of different CLIR studies in which different languages are processed
- Designing global scale CLIR systems

Determining ISs and MorphIFs is a more straightforward task than determining the SemIFs of derivatives and compounds. This is because semantic transparency is a gradual phenomenon. Derivatives and compounds may be partially transparent. For instance, *blackbird* is a specific bird species. Thus *blackbird* is not the same as *black bird*. Before calculating semantic IFs, criteria have to be settled how to handle these kinds of intermediate cases. Nevertheless, what is required in the

determination of all types of ISs and IFs is the analysis of unrestricted text samples (according to given criteria).

The next two examples illustrate how SemDIF and SemCIF could be utilized in IR.

Derivatives. The category of derivative words is problematic in IR. Derivatives are often close to their root words in meaning, but sometimes there is only partial overlap in expected and real meanings, and sometimes it is difficult to establish the connection between a derivative and its root word (for instance, *regard* and *guard*). It is possible that the transparency of derivatives depends on derivative affixes. Thus for some affixes derivatives may be more transparent than for others. It is thus possible that some types of derivatives are useful and some types harmful in IR. Within a given language, SemDIFs can be calculated for derivatives of given affixes, and the effects of these derivatives on retrieval effectiveness can be tested. If the use of a given derivative type is useful, it is likely that the use of a different derivative type that has the same SemDIF is also useful. Thus it would not be necessary to make laborious IR experiments on all types of derivatives, but it would be possible to use SemDIFs as indicators of retrieval success. *Longman Dictionary of Contemporary English* lists over 300 derivational affixes [42]. Testing the effects of all the derivative types empirically would be an enourmous task. In many languages the number of derivative affixes is even higher.

Compounds. *Swedish* is rich in compound words. In IR it is possible to decompose compounds or leave them untouched. Compound splitting in Swedish can be done using a lexicon-based morphological analyser. The effects of compound splitting on IR in Swedish retrieval have not been tested. Neither it is clear to what extent Swedish compounds are transparent. Linguistic analysis of a sample text would show the proportion of transparent compounds. If the analysis indicates that SemCIF is low, showing that most compounds are transparent, this would suggest that compound splitting would be helpful in Swedish retrieval. The SemCIF for Swedish and Swedish IR results of compound splitting could be used as indicators of the effects of compound splitting on IR for other (compound) languages for which SemCIF is known.

8. The need for semantic and syntactic typologies

In addition to morphological properties, languages differ considerably from each other in semantic and syntactic features. Developing semantic and syntactic typologies for IR would be needed for the

same theoretical and practical reasons as in the case of morphology (Sections 1 and 7).

There seem to be significant differences between languages in the frequency of *lexical ambiguity*. Homonymy seems to be common in *Swedish* [9, 39]. In *English* the frequency of homonyms is higher than in *German* [38]. Chen and others reported that in English lexical ambiguity is more common than in *Chinese* [43]. The statistics showed that, on the average, an English word had 1.687 senses and a Chinese word 1.397 senses. For the 1000 top high frequency words, the number of senses for English and Chinese words were, respectively, 3.527 and 1.504.

Ullman identified different kinds of semantic tendencies in different languages, on the basis of which semantic typology of languages could developed [38]. The criteria for semantic language typology involve the following:

- The relative frequency of opaque and transparent words
- Synonymic patterns
- The relative frequency of particular and generic terms
- The relative frequency of polysemy
- The relative frequency of homonymy
- The relative independence of words, and the importance of context in determining their meanings

In the syntactic typology of Greenberg languages are divided into different types on the basis of the order of a subject (S), an object (O) and a verb (V) in a transitive sentence [4]. The most common types are SVO and SOV languages. In *Korean* and some other languages a word order is (to a large extent) free [44]. The syntactic type of a language is meaningful in syntactic parsing as well as determining collocations. For languages with a free word order, such as Korean, identifying collocations is more difficult than for languages with more stabile word order [44]. In addition to sentence structure the structure of syntactic phrases may vary between languages [9]. In *English* NPs are of the type AN (adjective, noun) while in *French* NPs are predominantly of the type NA.

9. Conclusions

With the increasing significance of global CLIR research [5, 7] and global scale CLIR systems [6] it

is important to know the universal linguistic features shared by different languages as well as differences among languages. Also, following IR research done in different languages requires a common linguistic framework. This paper presented a morphological typology for IR. The typology provides a theoretical framework for linguistically oriented IR reseach. To calculate different ISs and IFs for a given language is a relatively simple effort. When they have been established they could be used in several ways in IR research and system development and evaluation. Some applications were proposed here. At the University of Tampere we are experimenting with different languages in our CLIR research project and have the opportunity to study the utilization of ISs and IFs in cross-language text retrieval.

Linguistic research has shown that in addition to morphology languages differ considerably from each other in semantic and syntactic properties. We have planned to complement the morphological typology presented here by semantic and syntactic IR typologies. In this way a more complete picture of linguistic differences between languages can be achieved. Syntactic differences are of minor importance for IR, but indirectly syntax may be significant. For instance, the use of collocations is one method to recognize phrases. The more predictable syntactic structures a language possesses the more effectively collocations can be used.

Appendix

Table 5 illustrates the principal indexing methods used in Finnish text databases [28]: (1) no morphological analysis (inflectional index), (2) word form normalization (base form index), and (3) word form normalization and decomposition of compounds (base form index/compound splitting). Because one lexeme often has several inflectional forms, the inflectional index normally is largest. However, this cannot be demonstrated here.

Sample text (CLEF Topic 015*): Mitkä tekijät vahingoittavat Euroopan teollisuuden kilpailukykyä maailmanmarkkinoilla?* (What are the factors that damage the competitiveness of European industry on the world's markets?) The normalization and compound splitting of non-stop words occurring in the sample text go as follows:

tekijät →	tekijä (factor)
vahingoittavat →	vahingoittaa (damage)

Euroopan →	Eurooppa (Europe)
teollisuuden →	teollisuus (industry)
kilpailukykyä →	kilpailukyky (competitiveness)
kilpailukyky →	kilpailu (competitition), kyky (potency)
maailmanmarkkinoilla →	maailmanmarkkinat (world's markets)
maailmanmarkkinat →	maailman (world's), markkinat (market)
maailman →	maa (earth), ilman (without), ilma (air), maailma (world)

Table 5. Different index representations

Inflectional index	**Base form index**	**Base form index/Compound splitting**
euroopan	eurooppa	eurooppa
kilpailukykyä	kilpailukyky	ilma
maailmanmarkkinoilla	maailmanmarkkinat	ilman
tekijät	tekijä	kilpailu
teollisuuden	teollisuus	kilpailukyky
vahingoittavat	vahingoittaa	kyky
		maailma
		maailmanmarkkinat
		markkinat
		tekijä
		teollisuus
		vahingoittaa

Acknowledgements

I would like to thank prof. Kalervo Järvelin for his valuable comments and suggestions.

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 Project 44704).

References

1. Comrie, B. *The world's major languages*. London - Sydney: Croom Helm, 1987.

2. Katzner, K. The *languages of the world*. London: Routledge & Kegan Paul, 1977.

3. Saussure, F. de. *Course in general linguistics*. London: Duckworth, 1983.

4. Greenberg, J.H. Some universals of language with particular reference to the order of meaningful elements. In: Greenberg, J.H., ed. *Universals of language*. The MIT Press, 1966, 73-113.

5. Braschler, M., Krause, J., Peters, C. and Schäuble, P. *Cross-language information retrieval (CLIR) track overview*. http://trec.nist.gov/pubs/trec7/

6. Oard, D. Extending cross-language information retrieval to a global scale. In: *Workshop on Multilingual Information Management.* Granada, Spain, 1998. Also available at: http://www.glue.umd.edu/~oard/

7. Peters, C. *CLEF - Cross-Language Evaluation Forum*. 2000. http://www.iei.pi.cnr.it/DELOS/CLEF

8. Comrie, B. *Language universals and linguistic typology*. Chicago: The University of Chicago Press, 1989.

9. Karlsson, F. *Yleinen kielitiede*. [*General linguistics*]. Helsinki: Helsinki University Press, 1998. [In Finnish]

10. Whaley, L.J. *Introduction to typology: the unity and diversity of language*. Thousand Oaks - London - New Delhi: Sage Publications, 1997.

11. Bybee, J.L. *Morphology: a study of the relation between meaning and form*. Amsterdam - Philadelphia: John Benjamins Publishing Company, 1985.

12. Matthews, P.H. *Morphology*. Cambridge University Press, 1991.

13. Karlsson, F. *Suomen kielen äänne- ja muotorakenne.* [Phonological and morphological structures in Finnish]. Porvoo - Hki - Juva: WSOY, 1983. [In Finnish]

14. Matthews, P.H. *The concise Oxford dictionary of linguistics*. Oxford - New York: Oxford University Press, 1997.

15. Akmajian, A., Demers, R., Farmer, A. and Harnish, R. *Linguistics: an introduction to language and communication*. Cambridge, MA: The MIT Press, 1990.

16. Porter, M.F. An algorithm for suffix stripping. *Program*, 14, 1980, 130-137.

17. Pirkola, A. *Studies on linguistic problems and methods in text retrieval: the effects of anaphor and ellipsis resolution in proximity searching, and translation and query structuring methods in cross-language retrieval*. PhD Dissertation. University of Tampere, Department of Information Studies. Acta Universitatis Tamperensis 672, 1999.

18. Pirkola, A. The effects of query structure and dictionary setups in dictionary-based cross-language information retrieval. In: *Proceedings of the 21st Annual International ACM Sigir Conference on Research and Development in Information Retrieval*. Melbourne, Australia, 1998, 55-63.

19. Myaeng, S.H. Information retrieval with Asian languages: an introduction. *Information Processing and Management*, 35, 1999, 421-425.

20. Large, A. and Moukdad, H. Multilingual access to Web resources: an overview. *Program*, 34(1), 2000, 43-58.

21. Sampson, G. *Writing systems*. London: Hutchinson, 1987.

22. Lee, K.H., Ng, M.K.M. and Lu, Q. Text segmentation for Chinese spell checking. *Journal of the American Society for Information Science*, 50(9), 1999, 751-759.

23. Kando, N., Kageura, K., Yoshioka, M. and Oyama, K. Phrase processing methods for Japanese text retrieval. In: *ACM Sigir Workshop on Information Retrieval - Theory into Practice*. Melbourne, Australia, 1998, 13-19.

24. Oard, D. Effects of term segmentation on Chinese/English cross-language Information Retrieval. *Symposium on String Processing and Information Retrieval (SPIRE)*. Cancun, Mexico, 1999. Also available at: http://www.glue.umd.edu/~oard/

25. Wu, Z. and Tseng, G. Chinese text segmentation for text retrieval: achievements and problems. *Journal of the American Society for Information Science*, 44(9), 1993, 532-542.

26. Harman, D. How effective is suffixing? *Journal of the American Society for Information Science*, 42(1), 1991, 7-15.

27. Popovic, M. and Willett. P. The effectiveness of stemming for natural-language access to Slovene textual data. *Journal of the American Society for Information Science*, 43(5), 1992, 384-390.

28. Alkula, R. Merkkijonoista suomen kielen sanoiksi. PhD Dissertation. University of Tampere, Department of Information Studies. Acta Universitatis Tamperensis 763, 2000. [in Finnish]

29. Hull, D. Stemming algorithms: a case study for detailed evaluation. *Journal of the American Society for Information Science*, 47(1), 1996, 70-84.

30. Schinke, R., Greengrass, M., Robertson, A.M. and Willett, P. A stemming algorithm for Latin text databases. *Journal of Documentation*, 52(2), 1996, 172-187.

31. Krovetz, R. Viewing morphology as an inference process. In*: Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. Pittsburg, PA, 1993, 191-202.

32. Savoy, J. A stemming procedure and stopword list for general French corpora. *Journal of the American Society for Information Science*, 50(10), 1999, 944-952.

33. Kalamboukis, T.Z. Suffix stripping with modern Greek. *Program*, 29(3), 1995, 313-321.

34. Abu-Salem, H., Al-Omari, M. and Evens, M.W. Stemming methodologies over individual query words for an Arabic information retrieval system. *Journal of the American Society for Information Science*, 50(6), 1999, 524-529.

35. Ekmekcioglu, F.C. and Willett, P. Effectiveness of stemming for Turkish text retrieval. *Program*, 34(2), 2000, 195-200.

36. Frakes, W.B. Stemming algorithms. In: Frakes, W.B. and Baeza-Yates, R., ed. *Information retrieval: data structures & algorithms*. Englewood Cliffs, New Jersey, 1992

37. Greenberg, J.H. A quantitative approach to the morphological typology of language. In: Spencer, R.F., ed. *Method and Perspective in Anthropology*. Minneapolis: University of Minnesota Press, 1954, 192-220.

38. Ullman, S. *Semantics*: *an introduction to the science of meaning*. Oxford, 1967.

39. Hedlund, T., Pirkola, A. and Järvelin, K. Aspects of Swedish morphology and semantics from the perspective of mono- and cross-language information retrieval. Forthcoming in *Information Processing & Management*, 2000/2001.

40. Ashford, J.H. Free text retrieval in the Welsh language: problems, and proposed working practice. *Journal of Documentation*, 51(2), 1995, 118-125.

41. Karlsson, F. *A Finnish grammar*. Porvoo: WSOY, 1987.

42. *Longman Dictionary of Contemporary English*. Harlow: Longman Group, 1987.

43. Chen, H-H., Bian, G-W. and Lin, W-C. Resolving translation ambiguity and target polysemy in cross-language information retrieval. In: *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*. University of Maryland, MA, 1999, 215-222.

44. Kim, S., Yang, Z., Song, M. and Ahn, J-H. Retrieving collocations from Korean text. In: *Proceedings of the 1999 Joint Sigdat Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*. University of Maryland, MA, 1999, 71- 81.