Linguistic research has shown that languages differ considerably from each other in morphological, semantic and syntactic properties. With the increasing significance of global research and global scale, it is useful to know the universal linguistic features shared by different languages as well as differences between languages. This knowledge is useful in system development and evaluation.

There are at least 4000 languages in the world. Linguistic typology can be used to categorize these languages on linguistic grounds. Based on linguistic typologies it is possible to consider the variation in morphological, semantic, and syntactic properties among the world's languages and the implications of differences for mono-lingual and cross-lingual information retrieval (IR). First, the writing systems will be discussed followed by other related topics.

Writing Systems

There are differences among languages in terms of characters and rules. The Chinese and Roman writing systems can be used to demonstrate the concept. The most obvious difference between Chinese and English texts is the use of different characters. Writing system (script) refers to a set of characters and rules how to apply the characters in writing. We may differentiate e.g. Chinese, Roman, and Arabic writing systems. Depending on the writing system a character may be associated with a word sense, a syllable, or a phoneme.

Term Segmentation

Different writing systems require different text processing techniques in mono-lingual and cross-lingual information systems. A good example of text processing needed in some languages only is term segmentation in many Asian languages. In Japanese, Chinese, and Korean texts there are no obvious word boundaries. Term segmentation is a process in which a string of characters is divided into words and other meaningful units. 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. Segmentation is associated with compound noun identification, which is the same kind of task as phrase identification in English.

Basic Concepts of Morphology

As mentioned earlier**,** morphology is an essential element in differentiating between languages. Therefore, basic principles of morphology are discussed below as follows:

- Inflectional morphology
- Derivational morphology
- Compounding
- Morpheme
- Free morphemes
    - lexical morphemes
    - grammatical morphemes

- Bound morphemes, affixes
    - prefixes
    - suffixes
    - infixes

Morphology is the field of linguistics which studies word structure and formation. In another way it is a way of classifying the languages of the world. It is composed of inflectional morphology and derivational morphology. Inflection refers to the use of morphological methods to form inflectional word forms. Derivational morphology is concerned with the derivation of new words from other words using derivational affixes. Compounding (compound word formation) is another method to form new words.

A morpheme is the smallest meaningful unit of a language. Morphemes are classified into (1) free morphemes and (2) bound morphemes. Free morphemes appear as independent words. E.g.: book, car. Free morphemes 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. E.g.: re-, -ly, the plural-s. 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 ends of words. Infixes, which are affixes attached within other morphemes, are used only in some languages, as in certain native American languages.

Recent Morphological Typology

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 (Comrie, 1989; Whaley 1997). 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. Word order is less important for these languages than it is for analytic languages, since individual words express the grammatical relations that would otherwise be indicated by syntax. 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:

Isolating □ <==========>Synthetic

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 a low index of fusion, while 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 2 presents the index of synthesis for eight languages (Karlsson, 1998). 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 a highly polysynthetic language, its index of synthesis being high. The other sample languages fall between Vietnamese and Inuit.

Table 2. 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

Differences in Inflection

In the world's languages, the most usual inflectional categories of nouns are number, grammatical case, and 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 x 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 the genitive case is marked (by a suffix). Because (for nouns) in addition to genitive forms only plural forms are marked, in English the index of synthesis is relatively low (Table 2).

Table 3 shows the number of morphosyntactic features in the category of case for 8 languages (Comrie, 1987). Hungarian has 21 features. In English there are only 2 features (nominative and genitive; genitive is marked). Finnish represents a language

with a high index of synthesis. This is due in particular 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. The concept of grammatical case is not relevant to all languages (languages with weak inflectional morphology, e.g., many Asian languages).

Table 3. 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 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.

In some languages word inflection is associated with the inflection of word stems, e.g., Welsh and Finnish. This represents the case of inflectional fusion. For example, the Finnish lexeme "käsi" ('hand') has five allomorphs or inflectional stems.

Differences in Compounding and Derivation

The world's languages differ remarkably from each other in the frequency of derivatives and compounds. 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. In English and French derivatives and compounds are not so common. English and French are also more opaque. A German compound is often translated by a phrase or a single word in English and French. The following sample words and parallel texts in German, English, and French illustrate the situation.

German	**English**	**French**
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.

German	**English**	**French**
Ursache ('orignal matter')	cause	cause

Eintreten (’in come’)	enter	entrer

German: "Welche Faktoren beeinträchtigen die Wettbewerbsfähigkeit der europäischen Industrie auf den Weltmärkten?"

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

French: "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".

The Implications of Different Morphological Features for IR and CLIR

Because morphology is essential in IR, morphological phenomena have considerable effects on retrieval effectiveness. Regarding inflection, in languages with a low inflectional index of synthesis and index of fusion, inflection does not interfere with matching to the same degree as in the languages with high inflectional index of synthesis and index of fusion. Retrieval can be expected to be more effective in these

languages. Also, the costs of constructing effective stemmers/morphological analyzers are lower for languages with low indices of synthesis and fusion. In languages with high indices of synthesis and fusion simple matching and indexing techniques are insufficient.

Whether derivationally related words and compounds should be handled in IR and CLIR depends in particular on the transparency of derivatives (compounds) in a language. High frequency of transparent derivatives (compounds) suggests that handling of derivatives (compounds) would be useful. On the other hand low frequency of transparent derivatives (compounds) suggests one may dispense with the morphological processing of derivates and compounds or that the costs of morphological processing are low (Pirkola, 2001).

Semantic Differences in Languages

In addition to morphological properties, languages differ from each other in semantic features.

There seem to be significant differences between languages in the frequency of lexical ambiguity. Homonymy seems to be common in Swedish (Hedlund, 2001). In English the frequency of homonyms is higher than in German (Ullman, 1967). Chen and others reported that in English lexical ambiguity is more common than in Chinese (Chen et al., 1999). 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 respective number of senses for English and Chinese words were 3.527 and 1.504.

Different kinds of semantic characteristics in different languages, all of which may have effects in mono-lingual and cross-lingual IR. The characteristics are as follows:

- Frequency of opaque and transparent words
- Synonymic patterns
- Frequency of polysemy
- Frequency of homonymy
- Frequency of particular and generic terms
- Independence of words, and the importance of context in determining their meanings

Syntactic Differences

In addition to morphological and semantic properties, languages differ from each other in syntactic features. The following features will be discussed in the coming section:

- S (subject), V (verb), O (object)
- The most common SOV and SVO
- Languages with free word order
- The structure of syntactic phrases may vary
- Word order meaningful in syntactic parsing and determining collocations.

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 (Greenberg, 1966). The most common types are SVO and SOV languages. In Korean and some other languages word order is (to a large extent) free. In addition to sentence structure, the structure of syntactic phrases may vary between languages. In English NPs (noun phrases) are of the type AN (adjective, noun) while in French NPs are predominantly of the type NA. The syntactic type of a language is meaningful in syntactic parsing as well as in determining collocations. For languages with free word order, such as Korean, identifying collocations is more difficult than for languages with more stable word order.

Conclusion

To conclude, Linguistic typology studies and classifies languages according to their structural features. Its aim is to describe and explain the structural diversity of the world's languages. In other word, it compares languages and studies the structural similarity and differences between languages in order to reach some general calcification or typology.

Despite the fact that the languages of the world show an enormous variety, there are basic principles that govern the structure of all languages. Therefore, generalizations can be made across unrelated and geographically non-adjacent languages according to the occurrence and co-occurrence of specific structures with respect to phonology, morphology, syntax and semantics.

Referances

Chen, H-H., Bian, G-W. and Lin, W-C. (1999). 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, pp. 215-222.

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

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

Hedlund, T., Pirkola, A. & Järvelin, K. (2001). Aspects of Swedish morphology and semantics from the perspective of mono- and cross-language retrieval. Information Processing & Management, 37(1), 147-161.

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

Pirkola, A. & Hedlund, T. & Keskustalo, H. & Järvelin, K (2001). Dictionary-based cross-language information retrieval: problems, methods, and research findings. Information Retrieval, 4(3/4): 209-230.

Ullman, S. (1967). Semantics: an introduction to the science of meaning.

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