VOCABULARIES

Åke Viberg

Introduction

This paper will treat the organization of vocabularies in spoken languages. To most people, 'vocabulary' conjures up the image of a dictionary. Dictionaries contain a lot of information about words, information about their form, such as spelling and pronunciation, about grammar and to some extent also about meaning. How much and what type of information is given about each word depends on the type of user the dictionary maker had in mind. Native speakers may be primarily interested in spelling or the meaning of uncommon words, while foreign language learners may be interested in knowing the meaning even of common words. Dictionaries always serve various practical purposes and should be judged accordingly.

From a theoretical point of view, the most interesting question is how words are organized in the mind or brain of ordinary language users. What type of information is needed about words when they are used in conversations or when we are reading our morning paper? Presently, we can only give incomplete answers to such simple questions. It is obvious that speakers use a lot more information about individual words than is found even in a large dictionary. Another problem is how we find or access all the information we need. A simple alphabetical listing of words, as in a printed dictionary, will not do. We must be able to access words in our mental store via many different pathways according to their form or to the many different aspects of meaning. To describe what is in the mind of a single speaker of a language is thus a formidable task. On top of this, we are intrigued by the great variety that can be observed between the 5.000 or so spoken languages that still exist today. The form a word has in one language is obviously not related to the form it has in another language, unless the languages are historically related or have borrowed words. With respect to the meaning, however, there must be some correspondence, since languages

can be translated into one another without distorting too much of the original meaning. This is an example of the crosslinguistic perspective, which is concerned with patterns of diversity and similarity across languages. In this paper, I will try to provide some partial answers to questions like these. In particular, I will attempt to show that it is possible to compare the vocabularies of spoken languages in a systematic way with respect to their global organization.

Categorization

If you see something oblong and black moving in a zigzag over the ground, you might say 'Watch out for the snake!' and if you see something grey and red moving through the air and landing on your windowsill, you might say 'What a beautiful bird!' Using a word such as 'snake' or 'bird' involves categorizing an individual specimen of something as an instance of a general concept and naming it. Such a concept (or category) may cover an extensive body of knowledge. We know that, e.g., a bird has feathers, wings and a beak and that typical birds build a nest, lay eggs and can sing. Only some of that knowledge is necessary to identify something as a bird and exactly what information we use varies with the situation. Categorization is important for language, as it allows us to use a single name for a number of individual phenomena. It is also important for thinking, because categorization allows us to make a number of predictions. Once we are told that something is a snake we can expect it to crawl and to hiss and to try to bite us, if we come too close.

Concepts are organized into conceptual fields, where concepts are related to one another in systematic ways. One central type of relation is the hierarchical relation. The concept 'bird' covers a number of more specific concepts such as 'robin' and 'lark'. At the same time, 'bird' is an instance of the more general concept 'animal'. In this way, we get the hierarchy:

animal ⟶ bird ⟶ robin, lark.

One way of looking at concepts is to regard them as structured sets of attributes that can take a number of values (see Barsalou 1992). An animal, for example, has attributes such as Sex (values: male/female), Age (child/adult) and Species (with a large number of values: human, horse, donkey, chimpanzee...). The combination of various values of the attributes allows us to form a large number of more specific concepts such as female + horse + adult (named: 'mare') and child + horse ('foal'). Naming or, to use

a more general term, lexicalization, involves relating a concept to a word-form. Note, however, that you can easily form a great number of natural concepts that are easy to think of but lack a name (i.e. that are not lexicalized) in a given language. One example would be female + child + chimpanzee. Actually, conceptual fields usually contain a great number of lexical gaps (unnamed concepts) of this type.

Lexicalization

Words can be characterized as the combination of a concept and a word-form. We have already seen that concepts outnumber words and that many concepts do not have a corresponding word. The concepts that are realized as words differ greatly with respect to how easily they can be expressed. Words have various types of *composition*. A concept can be named by a simple word (a lexical root) e.g. *knife*, a derived word, e.g *cutter* or a compound such as *carving knife* or *wire cutter*. From a formal point of view a *knife* is much simpler than *carving knife*, which not only is longer but also has an internal structure (*carv* + *-ing* + *knife*). In general, simple words have a tendency to name concepts which feel more basic or familiar. As will be demonstrated below, there are certain concepts that tend to be lexicalized in a simple way in most (or possibly all) languages. Another aspect of lexicalization is *syntactization*. A word has a number of distinct grammatical properties, which a concept does not have when it is not tied to language in some way. The most important of these are related to the word class, for example, whether the word is a noun such as *knife*, which in English inflects for definiteness and plural (*the knives*), or a verb such as *cut*, which takes tense markers (*is cutting/has cut*).

Words tend to have more than one meaning, a characteristic referred to as *polysemy*. The word *glass*, for example, refers to a physical object in *He held a glass in his hand* and to a substance in *a vase (made) of glass*. A word like *spoon* can refer to a physical object or to a measure: *two spoons of sugar*. Conceptually, there is a great difference between an object you can hold in your hand and a substance or a measure. Automatically, we realize that it can meaningfully be said *He took a spoon in his hand and bent it* or *He took a spoon of sugar in his hand and tasted it*, but not the opposite way around. The physical object 'spoon' and the measure 'spoon' represent two completely different concepts, even if there may be a natural link between the two meanings of the corresponding word. A single word can have a number of meanings, which often form quite clear patterns.

Glass, for example, follows the pattern: name of substance > object made of substance. Words which follow this pattern can be found in many languages, but only a minority of the object names are related to the name of the corresponding substance and exactly which words allow this pattern vary from language to language. In English, for example, *wood* (substance) is completely unrelated to *tree*, but in Swedish, two related word forms are used (*trä* 'wood' and *träd* 'tree'), and in Danish, the same word-form is used in both meanings: *træ* 'wood'/'tree'. Even if there may be a natural association between the different concepts a polysemous word refers to, this is far from always the case. In many cases, not even native speakers can find a natural conceptual link between two given meanings of a polysemous word-form which is still felt to be a single word. The extensive patterns of polysemy found in the lexicon only partly correspond to relations in the conceptual structure. This is one of the strongest motivations for regarding conceptual structure and lexical semantic structure as two separate systems (closely related but in a complex way).

Polysemy is an important driving force behind *historical lexical change*. The basic meaning of a word is all the time adapted to new contexts and extended to new meanings, often in accordance with patterns of polysemy which are already established in the language. In some cases, the original meaning has been lost and the new meaning subsequently felt as the basic one. The comparison of the meanings of words with similar forms in two related languages can often serve as an illustration of such change. In Swedish, *rita*, which historically is the same word as English *write*, means 'draw = make a drawing'. In older Germanic languages, the verb also had a more concrete meaning 'scratch, cut', which describes how a text or drawing is produced. The extension of the meaning in both cases is based on the relation between method of production and the result. In the historical process, different results have been lexicalized in English and Swedish. The Swedish word for 'write', *skriva*, is related to Latin *scribere*, which had a similar original meaning 'scratch, cut, carve'. The sense 'make a drawing' of English *draw*, also follows from the principle method > result, with the difference that the original meaning 'pull' is still extant. The historically related word *dra(ga)* in present-day Swedish only has this more concrete meaning and is the closest semantic equivalent of English *pull*.

Polysemy can also give rise to new grammatical markers. This process is known as *grammaticalization*. This usually proceeds in several steps. The verb *have* historically is derived from a verb with a more con-

crete meaning 'grasp, seize'. In present-day English, the basic lexical meaning is Possession. In some uses, concrete physical possession is still involved: *Peter has a book in his hand*, but usually the possession is of a more abstract type: *I have a camera but I don't have it with me*, or even more abstract: *I have an idea, I have a terrible headache.* In these abstract examples, the object is still a noun (phrase), which motivates regarding *have* as a lexical verb in these uses. The meaning is further extended in examples such as *You have to leave now* (Obligation), *Peter has left* (Temporal auxiliary). In the last two examples, the grammatical construction has also shifted and *have* functions as a grammatical auxiliary. In the last example, it can even be inverted like a grammatical verb: *Has Peter left?* In some other languages, the verb has gone full circle and become an ending. In French for example, the future endings are historically derived from a 'have' verb. In some forms of the future paradigm verbal forms like *je sortirai*, 'I will leave'; *tu sortiras* 'you (sing.) will leave', the formal similarity can still be observed: *j'ai* 'I have', *tu as* 'you (sing.) have', etc. Grammaticalization in its most developed stage involves a combination of semantic shifts with grammatical and phonological shifts (as when its status as an independent word is lost and an ending has arisen). Grammaticalization of this type is a very common process across languages (see Hopper/ Closs-Traugott 1993).

Lexical organization for rapid access

One basic characteristic of vocabularies is their *size*. In spite of the fact that we can form many more concepts than we have words for, we still know literally tens of thousands of words in our native language. According to some estimations, we commmand as many as 100.000 words – at least passively. The exact figure is hard to estimate for a number of reasons. It depends, for example, on how we define a word. Should, to take just one example, *Christmas tree* be regarded as a separate lexical item or just a combination of *Christmas* and *tree*? The combination is not self-evident. In other languages such as Swedish, the same phenomenon is called 'Christmas fir' (*julgran*) and regarded as one word. It is also difficult to say when a person knows a word. Is it enough to know that a *thrush* is a kind of bird or should one also be able to recognize it in a reliable way?

A common assumption is that there are actually two independent but closely connected stores in the mind or brain (Levelt 1989). One contains related *forms* such as *hat-cat-mat* without necessarily taking the meaning

into consideration and the other contains so-called *lemmas*, which represent the meaning and grammatical properties of words. Lemmas are thus related according to meaning such as *hat-cap* or *cat-dog* and according to grammatical properties, which tend to cluster and form word classes (or parts of speech) such as *Noun, Adjective* and *Verb*. In the following, I will concentrate on lemmas and the semantic and grammatical properties of words. For the sake of simplicity, lemmas will usually be referred to as 'words'. When a distinction is needed, 'word' in this sense will be contrasted to '(word-)form'.

An important fact about our command of the vocabulary is the *speed* with which we can find or access words. When a person speaks, the speed easily reaches 150-200 words a minute or two to three words a second. This means that the person who is speaking must find more than a hundred words each minute among tens of thousands of words. Imagine finding 100 books among ten thousand books in a library! In order to achieve that, the books must be stored in a systematic way, as they usually are in a public library. The same is true of our mental store of words, the dictionary in the brain. Traditionally, grammar has been regarded as the systematic part of the description of a language, while the vocabulary has often been regarded as a loosely organized list of words. But even the vocabulary is highly structured. To a considerable extent this structure is achieved with a restricted number of very general *lexical relations*. There are two major types of such relations: *hierarchical* and *contrastive*. Vast hierarchies can be formed between words with a superordinate and subordinate or *hyponymic* meaning, as it is usually called. An example would be *thing – artifact – tool – screwdriver – Philips screwdriver* or *organism – plant – tree – beech – copper beech*, where the words are ordered from the semantically most general to the most specific. At each step, the more specific word is a *hyponym* to the words with a more general meaning which appear above it in the hierarchy. Some of the more general words are found only in more formal or technical registers, but there are usually several levels even in everyday language. Another type of hierachy is represented by the part – whole relation or *partonomy* such as *house* (Whole) and *roof, wall, door, window* (Parts). These parts may have parts in their turn: window–pane, *window–sill*. Verbs form hierarchies based on manner (*troponomy*): *move – walk – stroll/waddle/wobble/stride/tiptoe/plod/trudge*. *Contrastive patterns* is used as a cover term for the many types of *contrast* (or opposition) and for *synonymy* ('no contrast').

What has been said so far about lexical structure and its relation to conceptual structure is summed up in Figure 1. (Some of the terms will not be introduced until later.)

Conceptual structure	
Categorization	(Comprises both linguistic and non-linguistic categories or concepts)
Lexical structure	
Lexicalization	
Composition	Root > Derived word > Compound > Phrase
Syntactization	Word class properties, Syntactic frame
Patterns of polysemy > Grammaticalization	
Lexical relations	
Hierarchic structure	Hyponomy – Partonymy – Troponymy
Contrastive patterns	Contrast/Antonomy – Synonomy

Figure 1
A characterization of lexical structure and its relation to conceptual structure.

Meaning and grammar in the lexicon

Individual words (i.e. lemmas) have a number of semantic and grammatical properties. Semantically, the word *run* signifies an event involving motion, which relates it to words such as *walk*, *swim* and *jump*, and grammatically, it has properties such as taking tense inflection and taking directional complements such as *to school*. In spite of the fact that there are a great number of such semantic and grammatical properties which can be combined with one another in many different ways, the most basic and frequently used properties tend to occur together and form clusters. The most fundamental clusters of grammatical properties have traditionally been called word classes and have been identified with names such as noun, verb, adjective and adverb. On strictly semantic grounds, words tend to form clusters, too, known as semantic fields or groups of words with a related meaning. In the next two sections, word classes and semantic fields will be briefly characterized.

Word classes

Traditionally, word classes have been characterized either in notional (semantic) or formal terms (or some type of combination). Thus, in notional

terms, nouns typically are names of persons, places and things, verbs refer to actions and events and adjectives describe properties. There are, however, many problems with this type of definition. It is not exhaustive and does not account for a number of obvious counterexamples such as *accident* and *theft*, which are regarded as nouns in spite of the fact that they refer to events. The exact delimitation of nouns and verbs in English is rather based on formal criteria. Nouns can be combined with grammatical markers such as the definite article, while verbs can be inflected for tense. The problem with formal criteria of this type is that they are language-particular and not cross-linguistically valid. The majority of the world's languages lack articles and even morphological tense is far from universally present. In an article treating word class systems from the perspective of the field linguist, Schachter (1985) suggested that word classes should first be established in individual languages on the basis of language-particular formal criteria. As a second step, notional definitions could be used to name classes and to identify word classes across languages. According to Schachter, nouns and verbs are found in most (or all) languages, while many languages lack a formal class of adjectives or have a very small such class. In such languages, property concepts are either lexicalized as nouns or as verbs. Instead of phrases with an adjective such as a *happy man*, phrases of the following types based on nouns or verbs are found: *a man with happiness*, *a rejoicing man*. A distinct class of adverbs also seems to be lacking in many languages.

In the last ten years, a number of specific proposals have been made in order to explain why there are word classes or, in more theoretical terms, why semantic and grammatical properties have such a strong tendency to cluster across languages. One such proposal is the *time-stability scale* presented by Givón (1984), according to which nouns tend to lexicalize the most time-stable concepts while verbs tend to designate rapid changes and adjectives have an intermediate position:

Most Time-Stable Nouns ◄► Adjectives ◄► Verbs *Least Time-Stable*

The scale explains why property concepts tend to be realized as either nouns or verbs in languages with no or a very restricted class of adjectives.

Another proposal treats the word classes from a discourse perspective. Hopper and Thompson (1984) regard nouns and verbs as universal lexicalizations of prototypical discourse functions. The prototypical function of

a verb is to report an event, while the prototypical function of a noun is to introduce a participant which can be further 'manipulated' and referred to in the following discourse. The prototypical function is more or less clearly realized in different contexts where a word is used. Thus, in a sentence such as *Peter kicked Harry* the verb *kick* asserts the occurrence of an event, while this is no longer the case in sentences such as *To kick Harry was cruel* or *The man kicking Harry broke his leg*. Similarly in *Suddenly, a bear appeared*, the noun *bear* is used to introduce a new discourse participant, while no participant is introduced in a sentence such as *Bear-hunting may be dangerous*. The discourse function of a word also has consequences for the morphological and syntactic properties characteristic of a certain word class. In a prototypical discourse environment, a noun such as *bear* can take the full set of modifiers and inflections, which is not possible in an environment such as *bear-hunting*. The same applies to *kick*, which has an overt subject and is marked for tense only in the first example above, where it asserts the occurrence of an event.

Semantic field structure

One way of looking at the semantic structure of the vocabulary of a language is to regard it as a set of semantic fields. In very general terms, a *semantic field* is a group of words which are closely related in meaning. In order to delimit a field, reference has often been made to a superordinate term covering all and only the words belonging to the same field, for example animal (field: Animals): *cat – dog – pig* or *move* (field: Motion): *go – run – put – pull*. It is, however, not always possible to find a superordinate term for words which belong together semantically. A more powerful definition is to say that a semantic field is organized around a core concept shared by all the members of the field (Miller and Johnson-Laird 1976). From a strictly semantic point of view, words could be classified into fields without considering the grammatical properties. There are, however, strong reasons to believe that words belonging to the same word class are related in the mental lexicon. Studies of slips of the tongue have shown that the incorrect and the intended correct word belong to the same word class in the majority of cases. There are also strong regularities across languages which lexical core concepts are realized as a certain word class. In particular, this applies to the most unmarked verbs and nouns and to the unmarked adjectives in languages that distinguish these as a word class. In the next section, English verbs will be used as an example to illustrate how a classification

into semantic fields can be used to show the basic semantic structure of a representative part of the lexicon, in a way that can be extended to cover the lexicon in general.

As an example, Figure 2 shows how the 50 most frequent verbs in English can be classified into semantic fields. (The numbers show the rank when the verbs are ordered according to descending frequency, based on Francis/Kucera 1982.)

"Concrete verbs"

Posture	*Motion*		*Possession*	*Existence & Production*
	Reflexive	*Objective*		
43 stand	***9 go***	36 put	***2 have***	***7 make***
	11 come	38 bring	***10 take***	
	26 leave	48 set	***14 give***	
Manipula-	31 follow		***15 get***	*Organic life*
tion	32 turn		35 keep	
	42 move		44 provide	
33 hold	45 run		46 need	39 live

Mental verbs

Meta-linguistic	*Verbal Commun.*	*Perception*	*Cognition*	*Desire*
25 call	***6 say***	***12 see***	***13 know***	29 want
50 mean	23 tell	***17 find***	***20 think***	
	30 ask	21 seem		
	34 write	24 show		
		28 feel		
		47 hear		

Grammatical verbs

Dynamic system	*Aspectual*	*Causal*	*Modal*	*Modality*
1 be	27 begin	40 let	***3 will***	41 try
4 do	49 start		***5 can***	
19 use			***8 may***	
22 become			***16 shall***	
			18 must	

Other fields: 37 work

Figure 2
The 50 most frequent verbs in English classified into semantic fields