# RELATIVE SEMANTIC COMPLEXITY IN LEXICAL UNITS

Bo Ralph

Department of Computational Linguistics
University of Göteborg
Göteborg, Sweden

## Summary

The lexical component of a human language is typically heterogeneous and extremely complex. Before we can come to grips with the underlying lexical organization, we must reduce the bewildering complexity. Methods must be elaborated by which the interrelations between the units of the lexicon can be elucidated. This paper describes how a Swedish lexical material stored in a computer has been semantically stratified as a stage in the semantic analysis of the items included in the data base. In particular, a minor subset of the lexical items, consisting of current words in the language, has been selected as metalanguage in the definitions. It is argued that, in this way, a means of describing the relative semantic complexity in lexical units is provided.

## Introduction

The semantic and syntactic interrelations between the lexical units of a human language are notoriously complex and intricate, whether considered from the individual language-user's point of view or from the perspective of the collective competence of a language community. Indeed, they are so complex that, when it comes to thorough semantic analysis, scholars have only been able to handle small portions of the lexicon at a time. The typical lexico-semantic study has therefore concerned single lexical items or small groups of semantically interrelated items, in particular so-called word-fields or semantic domains.

On the other hand, there seems to be a growing sentiment among linguists that the lexical component is very basic to the functioning of language. The crucial role of the lexicon cannot, however, be adequately understood unless the scope is widened. Detailed knowledge is, admittedly, quite indispensable in constructing an overall model of the lexicon; but large-scale lexical investigations are just as necessary in order to reveal the underlying principles of lexical organization. Consequently, computer-based lexicology should rank high as a branch of computational and theoretical linguistics.

## The Heterogeneity of Lexicons

Lexical inventories that have developed spontaneously do not usually constitute neat and clear-cut systems. They are typically skewed in the sense that many phenomena which may seem quite marginal have nonetheless given rise to a rich vocabulary, in contrast to the lexical sparsity characterizing several domains that are logically more fundamental to man. To take just one example, there are, in Swedish, rather few expressions for eating while there is a great variety of verbs for making all sorts of noises displaying only minor acoustical (and perceptual) differentiation. Our creative capacity simply seems to be more nourished by our imagination with regard to sounds than by our imagination with regard to food consumption. That the asymmetry is quite arbitrary is emphasized by the fact that other essential human activities may produce a rich vocabulary. For instance, very fine distinctions can, in Swedish, be expressed monomorphemically in the field of walking.

Such disproportions as those just mentioned are basically due to historical accidents, i.e. pure chance, more or less. Consequently, they are language-specific rather than universal and cannot be ascribed to any general tendencies in the human mind. The same holds for all culture-dependent expressions. Thus, if the lexicons of many languages tend to contain words for buildings and vehicles, it is primarily because human beings tend to develop such things and, secondarily, need to name them. It can be concluded that the reason for the recurrence of such terms in various languages all over the world is not essentially (psycho)linguistic but, rather, a corollary of comparable extra-linguistic circumstances.

Cultural conditions may also give rise to other types of lexical heterogeneity. The lexicon of a language may be viewed as comprising different strata, some of which contain common words used by everyone, others containing words used exclusively by specialists. Technical language – where "technical" should be taken in a broad sense – in various fields, such as medicine, law, economy, technology, etc.; some forms of language used in certain professions or by certain socially defined groups, like traders, priests, or outlaws – these are examples of vocabulary strata that are likely to be fully mastered only by relatively few individuals. It is to be deplored when the language of professional debaters, for instance in politics and esthetics, also develops in this direction, as is often the case.

Other strata of language may be quite familiar to a majority of the language-users although they are less frequently employed, being tied up with different styles, registers, or contextual settings. This may apply to the vocabulary of honorific language, religious language, etc. Such differentiation in vocabularies as has been exemplified here is manifested in a language-specific way, but the very existence of differentiation is a universal trait. It has been suggested that lexical inventories can be subdivided into various domains obeying different sets of rules that govern the relations between language and reality. In other words, there may well be various kinds of word meanings (cf. Fillmore 1978).[5] Information about a many-splendoured world is to be conveyed by means of language. The phenomena referred to are quite different in nature, and so the semantic content of lexical items may vary accordingly.

In most authentic vocabularies there is a gradient ranging from more or less purely grammatical operators and structure-dependent items (such as the copula, connectives, quantifiers, etc.), over items that are partly system-oriented, partly more semantically weighted (e.g. pronouns, deictic expressions, prepositions), all the way to items simply indexing "encyclopedic" phenomena. There is much fluctuation from language to language in this regard, since the division of labour between vocabulary and grammar proper may vary. Thus the proportion of words with primarily grammatical functions may differ to a high degree between languages. However, the grammar-oriented part of the vocabulary tends to be shared by most speakers, more differentiation being found at the other extreme.

Fillmore has mentioned a number of ways in which languages may differ with respect to word semantics. There are such features as relative analyticity, i.e. the degree of semantic transparency characterizing the total lexical system, taxonomic depth, by which is meant the dosage of particular as compared to generic terms, patterns of meaning extension, areas of synonymy elaboration, collocational patterns, etc. (Fillmore 1978, p. 155-157).[5] In fact, different domains within the vocabulary of a single language may vary a great deal in these respects. For instance, terminology is often, although not always, harder to analyse than are common words. In particular, terminology tends to invite heavy borrowing of foreign lexical material; in this way the portion of arbitrary lexical units increases.

It cannot be doubted that somewhere behind the confusing complexity of the lexicon there is a clue as to what human beings find imperative to recognize as delimited concepts. The categorization reflected by lexical inventories is considerably disguised through the heterogeneity which is a basic characteristic of the lexical component, as has been emphasized repeatedly. As a first step, then, methods must be elaborated by which the complexity can be duly handled. In particular, the semantic redundancy of the authentic lexicon must be reduced.

## Reducing Redundancy

It is very natural in lexico-semantic analysis to take word definitions as a point of departure. It can be argued that a defined word is semantically more complex than each word used in the definition of that word. Also, it is a well-known fact that circularity very easily creeps into definitions. Although circularity in definitions has occasionally been the target of investigation and has served successfully as a basis for determining semantic relatedness (e.g. Calzonari 1977),[2] it should, ideally, be controlled.

One way of achieving maximal reduction of semantic redundancy in the lexicon is, of course, to define all lexical entries by means of an effective metalanguage, e.g. a minimal defining vocabulary. Our interest can then be focused

on this minimal word-list on the assumption that it covers the same semantic range as the complete vocabulary defined by it. In practical lexicography, defining vocabularies have been utilized in, for instance, The General Basic English Dictionary (1942);[8] Michael West, An International Reader's Dictionary (1965);[10] and, in a project having much wider scope and, therefore, holding greater theoretical interest, in Longman's Dictionary of Contemporary English (1978).[3]

Defining vocabularies are intuitively attractive. They seem to capture the notion of basic vocabulary, the general lexical subset included in everybody's vocabulary. In some exceptional cases it is very easy to isolate this subset. In Dyirbal, for instance, a Queensland Australian language, there is a special vocabulary used in certain social contexts; hence it is referred to as "mother-in-law language" (Dixon 1971).[4] In this subsystem, Dyangul, the same grammatical rules apply, but the vocabulary is very restricted so that, for instance, each Dyangul verb corresponds to several in the common language. Therefore, the Dyangul vocabulary can be taken as a model for a semantic classification of words in Dyirbal.

A slight disadvantage in using defining vocabularies is the levelling of depth in the linguistic analysis. The lexicon is considered on two fixed levels alone: that of the lexical entries and that of the basic defining words. As is well known, however, lexical units play very different roles in the language they are part of. Not infrequently, the semantic interrelations within given sets can only be represented in a multi-layered fashion. I do not wish to claim that the human lexicon is, in any strict sense, hierarchically organized, but various subdivisions of it may well be.

For instance, to catch something means roughly 'to get hold of something', to fish means 'to try to catch fish', and to angle means 'to fish with a hook and line'. Consistent use of a minimal defining vocabulary would yield definitions like 'to try to get hold of fish with a hook and line' for to angle. This is by no means a totally inadequate definition. To angle is clearly related to verbal expressions like to get hold of; the semantic relatedness becomes apparent in a comparison with other verbs, such as to interrupt, to sneeze, or to twinkle. The verbal acts designated by to catch, to fish, and to angle are, however, not absolutely on a par with each other. Both to fish and to angle "contain" an element of catching. It can be argued that they differ from each other, and from to catch, in the way the catching is specified. To fish explicates the object caught, viz. 'fish'. That fish is caught is presupposed by to angle as well, but with the additional specification of the fishing method employed. However, the two types of specification are not equal with respect to the verbal act 'to catch'. While 'to catch' is presupposed as an element in to fish, the whole meaning 'to try to catch fish' is incorporated in to angle. The relations can be expressed by bracketing in the following manner:

to catch – '(to try to get hold of)'
to fish – '(to catch [= to try to get hold of] (fish))'
to angle – '(to fish [= to catch (= to try to get hold of) (fish)] (with a hook and line))'

The closer relationship between to fish and to angle may be indicated by making use of to fish in the definition of to angle. Parallel treatment of pairs or groups of verbs to the effect that one verb may contain not only the general semantic properties of another verb but actually the other verb itself has been suggested by, among others, Binnick (1971)[1] and Fillmore (1978).[5]

In fact, this relative semantic stratification of the lexicon is rather similar to Weinreich's strategy for investigating the semantic content of the lexical inventory. Weinreich gives the following presentation:

Stratum 0: terms definable only circularly and by ostensive definition
Stratum 1: terms whose definitions contain only stratum-0 terms, but without circularity
Stratum 2: terms whose definitions contain only stratum-0 and stratum-1 terms, without circularity
Stratum n: terms whose definitions contain only terms of strata 0, 1, 2, ... n - 1.

He concludes that the metalanguage will be made up of the complete ordinary language except for stratum n (Weinreich 1962).[9]

A similar line of reasoning is at the bottom of the organization of the Swedish lexical material analysed in the

project Lexical Data Base, carried out at the Department of Computational Linguistics, University of Göteborg. A minimal defining vocabulary is, in principle, utilized in definitions. In addition, however, words not included in the defining vocabulary proper are occasionally allowed in definitions, with the requirement that they should be ultimately reducible to strict defining vocabulary units. The minimal defining vocabulary comprises words denoting very fundamental concepts pertaining to physical elements and forces, geometrical notions, topographical properties, state and movement, location, time, causation, basic organisms, physical and mental functions of organisms, etc., as well as more culture-sensitive and conventionalized concepts, such as colours, artefacts, social conditions, and the like.

A larger subset than the defining vocabulary is the so-called fully defined vocabulary. This part of the vocabulary is provided with elaborated definitions. Together with the defining vocabulary it makes up the semantic hard core of the lexicon taken as a whole. We are not likely to find more candidates for this part of the vocabulary no matter how much material is included in the data base. Instead, new material tends to be of a more specific kind, e.g. terminology known by only a few people, almost obsolete words, non-permanent compounds that have barely passed the threshold of lexicalization, but which are easily analysable in terms of the well-defined part of the vocabulary; in short, words which do not add anything further to the basic semantic system of the lexicon. These latter items are not assigned any proper definitions but are semantically specified more summarily.

Thus the data base is, in principle, divided into three strata:

(1) the defining vocabulary, whose units are axiomatic in a logical sense and highly restricted in number;
(2) the fully defined vocabulary, whose units have carefully formulated definitions based on the defining vocabulary;
(3) the paraphrased vocabulary, whose units are semantically described by approximation.

In line with the above reasoning as regards relative semantic complexity, we allow entities from the fully defined vocabulary to enter into definitions. They are ultimately reducible to defining-vocabulary units. The definitions are more elegantly formulated in this manner, but, in particular, the interrelations between lexical items are more revealingly stated. Such an approach, building on extensive lexical cross-referencing, implies several theoretical commitments. Therefore, it should be emphasized that the data base described here is aimed at contributing to revealing the semantic interrelations between lexical items, in the first place. This, however, should not be taken to mean that our goal has been an ideal ultimate representation of the semantic structure in the lexicon.

## Investigating Relative Semantic Complexity

Careful selection of defining units and adequate definitional formats are a prerequisite for an acceptable result of the empirical work under way. It is true that lexicographers involved in practical undertakings naturally seek to attain consistent and adequate formulations in definitions. The requirement is even stronger if semantic structure is the main object of analysis.

Monolingual dictionaries usually take the reader's knowledge of the language in question for granted. As a consequence, the definitions may not be explicit enough. For instance, the Swedish causative verbs fylla, glödga, runda, släta, svärta all agree in focusing on the result of the respective activities. In a standard dictionary, ISO,[6] they are defined by verbal phrases very similar to each other in structure:

| Verb | Definition |
|---|---|
| fylla 'to fill' | göra full 'to make full' |
| glödga 'to make glowing' | göra glödande 'to make glowing' |
| runda 'to round' | göra rund(are) 'to make round(er)' |
| släta 'to smooth' | göra slät 'to make smooth' |
| svärta 'to blacken' | göra svart 'to make black' |

However, the verb of the paraphrases, göra 'to make', implies quite different activities in the respective cases, perhaps something like 'to regulate', 'to treat', 'to shape', 'to grind', and 'to colour'. By aiming at this higher degree of exactitude, we both acquire a better knowledge of the basic semantic properties of the lexical entries paraphrased and obtain good candidates for

the eventual defining vocabulary.

Although the material is stored in and manipulated by the computer, intuition and Sprachgefühl play dominant roles in this work. Therefore, it is urgent to employ methods which may guide our intuitions in a favourable direction. Since, in Swedish, there is no "mother-in-law language", the units of the defining vocabulary have to be determined by a number of methods with the joint goal of finding the minimal workable set of defining words. One important method implies large-scale paraphrasing of verbs. In a first round, we concentrate on such verbs as have equivalent paraphrases involving the base morpheme of the original verb, retained in the verb complement in the paraphrase. Such verbs are, for instance, the following:

| Verb | Verb paraphrase |
|---|---|
| bind-a 'to bind' | fästa med band 'to fix with a band' |
| cykl-a 'to cycle' | åka (med/på) cykel 'to go by bicycle' |
| fisk-a 'to fish' | fånga fisk 'to catch fish' |
| gul-na 'to turn yellow' | bli gul(are) 'to become (more) yellow' |
| hamr-a 'to hammer' | slå med hammare 'to hit with a hammer' |
| kant-a 'to edge' | förse med kant(er) 'to provide with an edge (or: edges)' |
| kup-a 'to make convex' | forma kup-ig 'to shape convex' |
| såg-a 'to saw' | kapa med såg 'to cut with a saw' |
| tor-ka 'to dry' | göra torr(are) 'to make dry (or: drier)' |
| tvivl-a 'to doubt' | känna tvivel 'to feel doubt' |

The Swedish paraphrases are just as natural as the simple verbs in the examples given. There are hundreds of analogous cases. In a host of other examples, there are quite conceivable paraphrases of basically the same kind, although less conventionalized as collocations. The following verbs are of this type:

| Verb | Verb paraphrase |
|---|---|
| blöd-a 'to bleed' | utsöndra blod 'to give off blood' |
| dörj-a 'to whiff' | fiska med dörj 'to fish with a whiffing-line' |
| fuml-a 'to fumble' | bete sig fuml-igt 'to act fumblingly' |
| köp-a 'to buy' | förvärva genom köp 'to acquire through purchase' |
| pensl-a 'to paint' | bestryka med pensel 'to paint with a brush' |
| plåg-a 'to pain' | åsamka plåg-a (or: plåg-or) 'to cause pain' |
| skrik-a 'to cry' | utstöta skrik 'to ejaculate a cry' |
| stapl-a 'to pile' | upplägga i stapel 'to arrange in a pile' |
| sörj-a 'to grieve' | känna sorg 'to feel grief' |
| träl-a 'to toil' | arbeta som en träl 'to work like a slave' |

The verbs of the paraphrases are usually deprived of much of the specific content characterizing the original, simple verbs. They emphasize the purely verbal element in the respective events. Most of the specific meaning lies, instead, in the verb complement in the paraphrases. It is easily seen that the paraphrase verbs represent different degrees of abstractness, i.e. they are semantically complex to a varying extent. They are always, however, less complex than the corresponding simple verbs they derive from in the analysis. Considering the resemblance of these verbs to pronouns and other pro-forms, "pro-verb" would be a fitting term.

Once the set of verbs to be used in paraphrases is established, it may also be employed for verbs with morphologically dissimilar paraphrases. For instance, älska 'to love' : hysa kärlek 'to feel love', may be classified together with other verbs of emotion. A similar mode of analysis may also be applied to such verbs as cannot be associated with paraphrases in any apparent way: drabba 'to afflict', hämta 'to fetch', märka 'to notice', etc.

As to the formats of the definitions, it is obvious that the paraphrases signal some fundamental properties of the paraphrased verbs, besides

the nature of the respective reduced verb. Some verbs incorporate an instrument, even morphologically recognizable (cykla, dörja, såga, etc.), others an object implied in the event (blöda, kanta); still others focus on the result of the event (gulna, skrika, stapla, torka), or on the phenomenon or state perceived by an experiencer (plåga, störja). A class which is potentially very large integrates an adverbial specification of the event itself rather than the actants involved (e.g. linka 'to limp', porla 'to purl', tindra 'to twinkle').

Relations similar to those obtaining between verbs and verbal phrases within a language may be found if corresponding verbal expressions are compared across languages. This is a widespread and natural method for reinforcing observations on patterns of language structure. In certain lexical domains one language may have developed single (i.e. relatively arbitrary) verbs, while another language may express the same content by phrases. For instance, there is a large family of motion verbs in both English and Swedish. In Japanese, the same meanings are usually rendered by various forms of the basic verb for 'to walk' (aruku) augmented by one of a number of mimetic adverbial elements. Interestingly, some adverbs of the phrasal collocations thus arising in Japanese are, themselves, limited to a very restricted context. This amounts to saying that the phrases are just as lexicalized as the single verbs in English and Swedish (cf. Fillmore 1978).[5]

Transferring this interlingual comparison to one language only, we may note that verbal paraphrases lend themselves to classificatory work in an analogous way. Verbs may be more or less productive as pro-verbs in paraphrases, they may establish more or less natural paraphrases, they may occur in phrases which have corresponding single verbs or not, they may be more or less synonymous or antonymous to verbs established as pro-verbs, etc. By comparing different pro-verbs and their respective paraphrases with each other we may also find that a pro-verb may occur in the paraphrase of another pro-verb, thus producing semantic links of the type discussed above. In such cases, the relative semantic complexity is clearly recognizable.

Verbs, in particular, are highly rewarding in such work as has been described here. Other word classes are, however, accessible to basically the same type of analysis. Of course, we are aware of many problems connected with this approach, e.g. the question of syntactic compatibility between original items and their paraphrases, the relative arbitrariness in selecting defining units, etc. Furthermore, there are many features in the approach resembling generative semantic theory of the early 1970's (and, incidentally, the work outlined in Mel'čuk and Žolkovskij 1969[7] and elsewhere); consequently, the same type of criticism as has been raised against that theory applies to the present work.

We do not find this too embarrassing. Our work is chiefly empirical, starting with observable facts, i.e. the words themselves, gradually eating our way down into deeper semantic structure. Thus, in a way, we are working in the opposite direction compared with the generative semanticists. We have no wish to reduce all lexical items to a single underlying category of units, and we are not prepared to press all lexical items into one basic semantic schema. Rather, we hope to be able to shed some light on the richness of the semantic system of Swedish, by elaborating a semantically based convertibility system. The method we have used seems to us to provide a versatile means to such an end.

## References


1. Binnick, R.I. 1971. Bring and Come. Linguistics Inquiry 2. 260-265.
2. Calzolari, N. 1977. An Empirical Approach to Circularity in Dictionary Definitions. Cahiers de lexicologie 31. 118-128.
3. Dictionary of Contemporary English. 1978. Harlow & London: Longman.
4. Dixon, R.M.W. 1971. A Method of Semantic Description. Semantics, ed. by D.D. Steinberg and L.A. Jakobovits. Cambridge: University Press.
5. Fillmore, C.J. 1978. On the Organization of Semantic Information in the Lexicon. Papers from the Parasession on the Lexicon, ed. by D. Farkas et al. Chicago: Chicago Linguistic Society. 148-173.
6. ISO = Illustrerad svensk ordbok. 1977. 3rd ed., 3rd pr. Stockholm: Natur och Kultur.
7. Mel'čuk, I.A. and A.K. Žolkovskij. 1969. Towards a Functioning 'Meaning-Text' Model of Language. Essays on Lexical Semantics, Vol. II, ed. by

V.Ju. Rozencvejg. 1974. Stockholm: Skriptor. 1-52.

8. The General Basic English Dictionary, ed. by C.K. Ogden. 1942. New York: W.W. Norton & Co.

9. Weinreich, U. 1962. Lexicographic Definition in Descriptive Semantics. Problems in Lexicography, 2nd ed. by F.W. Householder and S. Saporta 1967. Bloomington: Indiana University. 25-44.

10. West, M. 1965. An International Reader's Dictionary. London: Longman.