titulus
On measuring semantic complexity: A contribution to a rapprochement of semantics and statistical linguistics
huius textus situs retis mundialis
http://www.christianlehmann.eu/publ/lehmann_sem_complexity.pdf
dies manuscripti postremum modificati
01.05.1977
occasio orationis habitae
–
volumen publicationem continens
Georgetown University Papers in Languages and Linguistics 14
annus publicationis
1978
paginae
83-120

On measuring semantic complexity

A contribution to a rapprochement of semantics and statistical linguistics

Christian Lehmann

University of Cologne

Abstract

In Part 1 it is suggested that a concept of semantic complexity that has empirical relevance can be formed if it is regarded as the specificity of meaning. This means that polysemy does not add to, but subtracts from the semantic complexity of a word. Markedness, on the other hand, adds to complexity; and semantic markedness theory can, in fact, be regarded as a part of the theory of semantic complexity. A measure of semantic complexity which depends on semantic representations in a predicate-calculus-type notation is formalized. The empirical correlate of the semantic complexity of a word thus calculated is its amount of information, as determined by checking it in a frequency dictionary.

In Part 2 Portuguese kinship semantics is treated as an example. First, a semantic analysis of all the terms is presented; second, the semantic complexity of each of the items is computed; third, the field is ordered according to five semantic relations which emerge as complexity relations; fourth, the frequency of the terms is determined by a Portuguese frequency dictionary, and the figures for the amount of information are calculated; and last, it is seen that there is a nearly 100 percent correlation between semantic complexity and amount of information. This is taken to corroborate the empirical nature of the complexity measure.

In Part 3 there is a discussion of some problems related to an improved application of the complexity measure. It is suggested that after some necessary refinements have been made, the check-up in a frequency dictionary will be able to serve as a partial test on the empirical adequacy of proposed semantic descriptions.

1 General questions about semantic complexity.

Several general questions can be asked about semantic complexity:
(1) What is semantic complexity?
(2) Why might one want to measure semantic complexity?
(3) How might one measure semantic complexity?
(4) What are the empirical correlates of semantic complexity?

1.1 What is semantic complexity?

Let us take this question to refer to a morpheme or a word. In order to determine the complexity of such a unit, we have to rely on its linguistic description. Since such descriptions will differ according to the theoretical framework chosen, the question of complexity can only be answered within a given framework.

In phonology, it has been possible to give an interesting answer to the corresponding question when generative phonology was combined with markedness theory. (See Chomsky and Halle 1968, Chapter 9 and an uncitable amount of subsequent discussion in the literature.) In this framework, the *significans* of a word is decomposed in bundles and series of phonological features, arranged in bidimensional matrices. Depending on its phonetic content and on the context in which it occurs, every feature has a certain expectation or naturalness, which is expressed in terms of marked and unmarked. That is to say: whenever a feature occurs such that it would be normal, most natural for it under the given circumstances, this occurrence is unmarked; whenever the occurrence of a feature is unexpected, less normal, it is marked. In lexical representations, features are specified only in terms of marked and unmarked; and since 'unmarked' is the same as 'not specified', the bidimensional matrices contain, in fact, only scattered *m*s, meaning 'marked'. Thus, the lexical representations contain only information which is not predictable by rule. Accordingly, there is a set of rules which fill in the predictable information and specify all feature values as + and -. The necessity of this latter step is disputable, and it has been abandoned in an alternative though largely equivalent proposal (Sanders 1974) which substitutes the *m*s by the indication of the features themselves. These are treated as one-valued, i.e. instead of having the values + or –, they are present or absent. The set of rules mentioned would in this case have the task of filling in the features whose presence is predictable. Now, the phonological complexity of a word is defined as the number of specifications present in its lexical representation, where specifications are *m*s in one model and features in the other.

Returning now to the question of complexity in semantics, we can say that the findings of markedness theory have to be incorporated here, too. For only positive information must be taken into account in deter-mining the complexity of representations, and it is markedness theory which gives us a principled way of distinguishing between positive information, present in the marked term, and absence of information, given in the unmarked term. On the other hand, it seems clear that an analogical transfer of the first of the two phonological models discussed above is not feasible. The arrangement of semantic representations in matrices whose file entries are features and cells contain their values is impossible the components of the *significatum* have to each other and to other *significata* which inexpressible by n'ary features. Instead, a lexical meaning has to be decomposed into elementary propositions, constructed and connected in a way very much like that of the predicate calculus. In this framework, the closest, analog to the former features are atomic predicates, which might also be called one-valued in the sense that they can only be present or absent from a given representation, but cannot hear values like + or – or *m*. Thus, when a certain feature is absent from a given *significatum*, this means in the first place that no information is present in that respect; and in the second place, when the *significatum* is opposed to another one containing the feature in question, the absence of the feature has to be taken to indicate its negation. Information of the latter kind, lacking in lexical representations, might be filled in by general rules just as in the second of the two phonological models. Therefore, the semantic complexity of a word would also be a function of the number of elementary propositions present in its lexical representation, where 'elementary' means 'containing only one predicate'.

This notion of semantic complexity, which is based on the number of components of a formalized semantic description, requires that we overturn certain inveterate preconceptions. We are accustomed to associate the property of significance with words like *life*, *mother*, *motion*, or *think*, rather than with words like *vegetate*, *grand-aunt*, *rush*, or *reflect*. We feel that the words of the first series are very much richer in shades of meaning than those of the second one; *mother* makes us think of much more than *grand-aunt*, *rush* is only one of the

numerous forms of motion, etc. If we check them in a common dictionary, we find that the entries of the words of the first series are much longer than those of the second. All this seems to show that they are ‘fuller of meaning’ and may make a linguist believe that their meanings are more complex.

On the other hand we know that the greater extension of a concept is accompanied by its lesser intension, and vice versa. If a word exhibits a broad range of applications, this implies that its meaning-core which remains invariant in all the uses has to be accordingly small. Conversely, in a word like *grand-aunt* whose meaning is almost independent of context we might say that everything belongs to the meaning-core. The reason for the fact that the words of the first series have so much less specific than the words of the second series. *Rush*, e. g. contains everything that constitutes the meaning-core of *motion* and additionally the specifics that distinguish it from other forms of motion. Conceiving of all these properties of meaning, which constitute the intension, the semantic specificity of a word, as elementary propositions, we are led to the opposite conclusion that the words of the second series are semantically more complex than those of the first.

But how do we account in this conception for the various nuances of meaning which do not make part of the invariant core? Will they not, too, show up in the form of elementary propositions and therefore contribute to complexity? This question has to be answered in two stages. In describing the meaning nuances of a word it has to be seen first which of them are due to the respective contexts. This is, in fact, the majority of the acceptations commonly indicated in the dictionaries. We find, for instance, that some dictionaries give as the thirty-third meaning of *come*: ‘close in’, e.g. in *the evening comes*. Actually, the dictionary has absolutely no business with this case. The semantic particularity of the example sentence lies in the fact that the subject of the process is a temporal expression. If it be the case that, from the point of view of the English language, time units come in a different sense than, say, grandaunts do – which I doubt – then it is the business of the rules of combinatory semantics to account for it. The different paraphrase required – or perhaps only permitted – by *come* in this context surely does not justify the establishment of a new acceptation in the dictionary. Most of what currently passes for polysemy will thus come to nothing when we put the combinatory semantics postulated by Katz and Fodor in 1963 into practice. What will remain constitutes the subject matter of the second stage of the answer to our question: the genuine polysemies which really call for a separate specification. Although they, too, are normally disambiguated by the context, a complete semantic description of the possible sentences requires in their case that the dictionary make available distinct alternatives among which the context selects. *Observe*, e.g. exhibits in all its uses a meaning-core which might be rendered by ‘direct one's attention to’; but when laws, norms, and the like constitute the object, there is an additional component which does not result by itself from this constellation and which says that one normally does not observe laws just as one observes rare birds, but that one obeys them.

The acceptations of a polysemous word are alternatives of which only one becomes operative in a given context. The elementary propositions which constitute them are, therefore, disjunctively ordered, whereas all those propositions which are operative simultaneously in a given context are conjunctively ordered. If we take semantic complexity to be a function of the number of elementary propositions present in a meaning, the question arises whether disjunctively ordered propositions contribute to complexity in the same way as conjunctively ordered do. We have to bear in mind that the unequivocal sense of a phrase which contains a polysemous word is arrived at by means of a reciprocal action of this word and the context: the polysemous word offers a number of alternative meanings, and it is the task of the context

to select one of them. The more alternatives the polysemous word leaves open, the more specific the meaning of the context has to be in order to resolve the ambiguity of the phrase, and the greater is therefore the contribution of the context to the integral meaning. Consequently, our conception of semantic complexity as meaning specificity requires that the semantic complexity of a word be, ceteris paribus, smaller the more polysemous the word is. This reasoning based on the syntagmatic axis of the semantic system can be paralleled by one related to the paradigmatic axis which leads to the same conclusion. Suppose *brother-in-law of x* had as its unique meaning 'brother of spouse of x'. This would be its intension; and any person x would have a certain range of relatives that fulfill the condition of being a brother of his spouse and constitute, thus, the extension of this meaning. Now, *brother-in-law of x* has an alternative meaning 'husband of sister of x' , thus requiring a disjunction of propositions in its *significatum*. In accordance with the alternative meaning there is, for any person x, an additional range of relatives whom he might call his brothers-in-law, so that the extension of the *significatum* is enlarged. Since greater extension corresponds to lesser intension, we have to conclude that the meaning of the ambiguous word is less complex than it would be if y *brother-in-law* had only one of its senses. In the determination of semantic complexity, the contribution of conjoined elementary propositions will be positive, that of disjoined propositions will be negative.

1.2 Why might one want to measure semantic complexity?

Complexity is the complement of simplicity and as such of theoretical interest to all linguists who regard the simplicity metric as the most promising candidate for an explicit evaluation criterion. Thus, if we were faced with two competing descriptions of the same semantic facts, we would, ceteris paribus, prefer the simpler, less complex one.

Second, there might be (and as will be seen later on, there is in fact) a range of interesting empirical facts which the notion of semantic complexity permits to appreciate in the first place. If semantics has quantitative empirical correlates, then an explicit criterion for the measurement of semantic complexity is necessary in order to describe and explain them. In fact, the greatest interest of the measurement of semantic complexity lies precisely in its being a method of an empirical semantics.

Third, there are various subdisciplines of applied linguistics which have a vital interest in linguistic complexity. The universal didactic principle which says that simpler matters have to be learnt and taught before more complex ones requires, in the case of vocabulary control, that we be capable of distinguishing semantically more or less complex vocables. In mentioning this matter in so superficial a fashion I do not, of course, want to imply that those words whose meaning is, according to my metric, relatively more complex will always prove to be the more difficult vocables in language learning. What I am saying is that semantic complexity will play a role in the design of a systematic vocabulary syllabus. Another such discipline is language pathology, which deals with phenomena whose description – and sometimes, whose cure – requires an understanding of semantic complexity.

There are, then, theoretical, empirical, and practical reasons which might motivate a linguist to try and measure the complexity of meaning.

1.3 How might one measure semantic complexity?

Let us suppose that the lexical representation of a *significatum* is formulated in a metalanguage very much like ordinary predicate calculus. The elementary propositions consist, then,

of predicates and various kinds of symbols gathered around them, and they are connected among each other by means of the logical junctors, i.e. conjunction, disjunction, implication, equivalence, etc. Molecular propositions are formed in this way, and they, in turn, may be joined to each other by junctors. We start from the hypothesis justified in Section 1.1 that conjuncts differ from disjuncts as to their weight in semantic complexity. Suppose that an independent elementary proposition is worth one point; it would then be natural for the valuation of conjunction to be additive so that a series of n conjoined elementary propositions that are inductively independent of each other would be worth n points. In the case of disjunction, we want to give lesser value to p ∨ q than to elementary p; and the more disjuncts we have, the smaller should the value of the whole disjunction be, up to the limiting case of an infinite series of disjuncts, which should be worth no point because it would represent the meaning of an infinitely polysemous word, that is, a word without meaning. The same is in order for the tautology p ∨ ¬p, which is logically equivalent to an infinite disjunction and which gives no positive information, either.

For a formal language such as we presuppose for our semantic representations, a concept which fulfills all of these requirements of linguistic common-sense has been developed in Carnap and Bar-Hillel (1952), although not with a view to approach the empirical problem of linguistic semantic complexity but to elaborate on Carnap's theory of inductive probability (Carnap 1950) and measure, e.g. the information that a certain hypothesis gives or the degree of confirmation that it receives by certain evidence. The authors define the amount of semantic information of a sentence (called 'inf', pp. 241-245) on the basis of a measure-function whose value is the inductive probability of the sentence. The inductive probability, in turn, is a function of the class of possible states of world, expressed by state-descriptions of the language system in question, which the sentence is compatible with (pp. 235-237; cf. the count in Lyons l977:47-51). Instead of this rather sophisticated procedure, I offer a simple formula which gives, as far as I can tell, the same results as Carnap's and Bar-Hillel's measure-function. It is based on the relative number of trues associated with the principal junctor of the molecular proposition in a truth-value table. Let n be the number of different elementary propositions in the molecular expression. Then 2^n is the number of lines in the truth-value table, i.e. the number of truth-values to be calculated for the principal junctor. Let $x/2^n$ be the relative number of trues in that column of truth-values. Now we define the semantic complexity c of a molecular proposition, in a way analogous to Carnap's and Bar-Hillel's amount of semantic information inf, as the negative dyadic logarithm of the relative number of trues:

$$c = ld\,\frac{1}{x/2^n} = -ld\,\frac{x}{2^n} = -\left(ld\,x - ld\,2^n\right) = n - ld\,x$$

This formula gives the desired results for our complexity coefficient c in the case of both conjunction and disjunction. In a series of n conjoined propositions, x is = l for any n. Since log 1 = 0, the formula gives c = n in this case, which is, of course, what we wanted. In a series of n disjoined propositions, x is = 2^n-1. With n = 2, say p ∨ q, we get $x/2^n = 3/2^2$ and c = 2 - ld 3 = 0.42. With p ∨ ¬p we get $x/2^n = 2/2$ and c = 0, which is, once more, what we wanted.

The general formula permits the derivation of specific formulas for not so simply structured cases. Conjunction of expressions with whatever internal structure does not offer problems because the *c*s may be calculated separately for each of the conjuncts and then be added. So let us take a glance at disjunction of molecular propositions. Call p_x and q_y the two disjuncts in which the proportions of trues have been determined as $x/2^m$ and $y/2^n$, respectively, and whose internal structure plays no role (because disjunction is commutative). The

proportion of trues beneath the disjunctor is then $x/2^m + y/2^n – xy/2^{m+n} = (2^m y + 2^n x - xy)/2^{m+n}$.[1] The formula for the complexity coefficient is, accordingly: $c = m + n - ld(2^m y + 2^n x - xy)$. In the case of more than two disjuncts, the formula has to be reapplied in a recursive fashion. Frequently the internal structure of the disjuncts is simply a series of conjunctions. This simplifies the formula because x and y are each = 1. Thus, the complexity coefficient for a disjunction $(p_1 \& p_2 \& ... p_m) \vee (q_1 \& q_2 \& ... q_n)$ is $c = m + n - ld(2^m + 2^n – 1)$. The elaboration of specific formulas for other cases, for instance implication or equivalence, is straightforward and need not concern us here. There might be semantic descriptions containing expressions of the form $(p \& q \& r) \vee (s \& q \& t)$. In such cases identical propositions have to be extracted from the disjuncts on the basis of distributivity;[2] otherwise, c would give erroneous results, because it presupposes that the components of the molecular expression be inductively independent.

There are numberless questions which remain open in this account. Some of them will be discussed later on when we treat an example in greater detail. I should like to conclude this section with a more general remark. The reader will have noted that the general formula for the complexity coefficient $c = -ld (x/2^n)$ bears a striking resemblance to the formula by which one calculates the amount of information (i) of a sign on the basis of its probability (p): $i = -ld\ p$; and he may well ask whether this is a coincidence. If we undertook an information-theoretical treatment of propositional logic (along similar lines as did Carnap and Bar-Hillel 1952), we should probably say that the amount of information conveyed by a logical expression is greater the more chances it has to be false, the limiting case being an expression which is never false, a tautology which conveys no information. Such reasoning would imply a concept ‘probability of an expression‘, conceived of as the relative number of trues beneath the principal junctor, just as the probability of a sign in information theory is its relative number of occurrences. In this way one might try to explain the fact that you come out with the general formula for the amount of information of a sign if you substitute the ‘relative number of trues‘ by ‘probability’ in a formula of semantic complexity that constitutes part of a model which considers semantic complexity as meaning specificity.[3]

1.4 What are the empirical correlates of semantic complexity?

The majority of the readers who have endured until here must find my ideas speculative and useless, if not worse. It is therefore time to point out that there are quite straightforward empirical correlates to my theoretical concept of semantic complexity. It leads, as the preceding paragraph already anticipated, almost necessarily to the expectation that there will be a correspondence between the semantic complexity of a word and its amount of information. The latter can be determined readily by consulting a frequency dictionary. It must be stressed that although the two formulas of c and of i are so similar, there is no circularity in this approach. The calculation of c is based on the semantic description of a word according to principles recognized in linguistic semantics, whereas the calculation of i is based on the frequency of a word, which is the result of frequency counts conducted according to principles recognized in statistical linguistics. The two coefficients are, therefore, entirely independent

[1] Thanks are due to Felicitas Lehmann and Holger van den Boom for help with this formula.

[2] Failure to recognize this vitiates the discussion of additivity for inf in Carnap/Bar-Hillel 1952: 249.

[3] It would be interesting to examine possible connections between semantic complexity as conceived here and as sketched in Thom (to appear). Although the two frameworks are entirely different, there are obvious points of contact, as e. g. when Thom has the species more complex than the genus, which would, in general, turn out to be the case in my framework, too.

of each other, and if there were a correspondence between them, this would confirm the empirical significance of the complexity measure and would, at the same time, constitute a fact in need of an explanation.

I will leave the demonstration that there is in fact a significant correlation between c and i to the next section and try here to explain why this should be so. It is relatively simple: semantic complexity has been conceived of as meaning specificity. The more specific the meaning of a word is, the smaller is its extension, the narrower its applicability. If a word has a narrow applicability, it will be rarely used. The less frequent or less probable words are, according to the formula, those with relatively higher amounts of information. Therefore, the greater the semantic complexity of a word, the greater its amount of information. Q.e.d. Let us repeat that while there exists this theoretical connection between the two concepts of semantic complexity and amount of information, there is no circularity involved in the method of measurement of semantic complexity which is based on this theory.

I am quite aware that in claiming a relation between information and meaning, I am contradicting nearly all of the better information theorists and others who have dealt with the subject, e.g. Bar-Hillel. In the case of the information theorists, as e.g. Shannon, their refusal to admit such a relation is certainly due to prudence, this being a consequence of linguistic incompetence. In the case of Bar-Hillel (1955:286f.), his conception of the meaning as the referent prevents him from seeing any relation between information and meaning. In general, it may be said that in those days nobody could see the relation because it presupposes semantic representations which have only been developed recently. It is, therefore, all the more remarkable that information theorists like Wiener have simply equated meaning and information, finding this 'entirely reasonable from the standpoint of commonsense' (Wiener l950:8).

The amount of information of a word is measured in bits. A bit is a minimum alternative, a binary decision between yes and no. If a word has fifteen bits of information, this means that speakers behave in their use of it as if they made fifteen binary decisions in selecting this word instead of any other in the lexicon. In information theory, these binary decisions are purely formal entities. In semantics, however we might be tempted to associate a content with them: since the atomic predicates or the elementary propositions containing them are either present or absent from a given *significatum*, why not regard them as the subject matter of the binary decisions? Such a decision would then consist of the choice or rejection of an elementary proposition. I said before that the speakers behave as if they made such decisions. Actually, neither information theory nor semantics claims that they really make them in performance. That a given word has fifteen bits of information means that in the code to which it belongs, fifteen yes-no questions would suffice to identify it. And in a parallel fashion, that a word has fifteen elementary propositions in its *significatum* means that by them it is distinct from all the other words in the same lexicon.

As the sober reader must have anticipated, this parallel is too close. A one-to-one correspondence between bits and elementary propositions is impossible for various reasons. First, bits are additive, elementary propositions are not. Second, why should the semantic unit that corresponds to a bit be just the proposition? The smaller units, predicates, arguments, and quantifiers, also have to be selected; for two elementary propositions may be distinct only by, say, one argument. (This same question has, of course, to be posed against our complexity metric, which also takes the proposition as the unit; cf. Section 3.) Third, as will be seen in the next section, the elementary propositions contribute with different weights to the overall complexity (this, too, has not been taken into account in the complexity measure), while bits are all equal. These provisos have to be taken very seriously and have an important consequence: we cannot expect to find a correspondence between the semantic complexity and the informa-

tion quantity of two arbitrarily selected words, at least not in the present near-zero state of sophistication in the theory of semantic complexity. We will have to compare words which are related in such a way that they have a certain number of semantic properties in common, but one of them has more than the other. In such a case the surplus in semantic complexity must have a clear parallel in a corresponding surplus in information quantity.

2 Semantic complexity in Portuguese kinship terminology

2.1 Principles of the analysis

For reasons explained in the preceding paragraph I base the exemplification on a semantic field. Since the application of the complexity measure presupposes a high degree of reliability and formalization of the semantic analysis involved, it seems preferable to choose a semantic field which has often been treated and consequently been the focus of a certain agreement between many semanticists. Besides all this, kinship semantics is one of the clearest cases to show that an explicit semantic analysis has to be formalized in a metalanguage very much like predicate calculus. This has also been noted by Kay (1974), whose analysis is not essentially different from mine. It is certainly an advantage for my argumentation that the principles of the analysis adopted here have already been published in an independent source; so I avoid the suspicion of making up an analysis to fit my complexity measure.

All the kin relations which are lexicalized in Portuguese can be decomposed into the two primitive relations ‘parent of’ and ‘married to’. We use two binary predicates, P(x, y) and MAR (x,y), reading ‘x is parent of y’ and ‘x is married to y’, respectively. Besides, we will need the relation $x \neq y$ which reads ‘x is not identical to y’, because whenever we want to specify two children of the same parent, we have to exclude the possibility of their being identical. Besides being a formal device, this has its empirical justification because Portuguese *irmão* ‘brother’ is not a reflexive relation; and it has to be explicitly stated in the semantic analysis because it cannot, as far as I can tell, be deduced by general semantic rules. Finally. we will use the property F(x), reading ‘x is female’ , and occasionally M(x), ‘x is male’. Quantifiers, though necessary in a complete predicate calculus notation. will be dispensed with here because they would change nothing in the analysis except to complicate it.

There are two complexity relations in the Portuguese kinship system which can be described as common markedness relations and will not, therefore, emerge from the analysis as a result but enter into it from the start. These are the relation between the male and female sexes in otherwise identical relatives, e.g. *tio* : *tia*, and the converse relation between the senior and the junior, e.g.: *tio* : *sobrinho*.

The sex case is too well known to need much exemplification. If you want to refer to uncles and aunts indistinctly, you say *tios* ‘uncles’, not *tias* ‘aunts’, and the same if you refer to one couple of an uncle and aunt. This is true for all of the kinship terms, even for the parents, who are *os pais* ‘the fathers’. The only exception are the spouses. While it is true that there is a generic term *esposos* which is the plural of *esposo* ‘male spouse’, as opposed to *esposa* ‘female spouse’, no one of these forms – especially not *esposo* – belongs to the same stylistic level as all the rest of Portuguese kinship terminology. The equivalent of *husband* is *marido*, and for *wife* one says *mulher* ‘woman’ on a colloquial and esposa on a formal level. *Marido* is not the unmarked term in either of the resulting pairs.

For our semantic analysis this has the consequence that while the female terms will be specified by F(x), the male terms, with the exception of *marido*, cannot be specified by M(x) because they are not necessarily male. The maleness of the neutral terms will be introduced

by a general postlexical rule whenever the term occurs either in the singular in a non-generic use or in the plural in contrast with the corresponding female term. This is a very common case of markedness in the lexicon, and it is neatly expressed by the fact that the marked and unmarked terms of a pair are distinct only by one elementary proposition (F(x) in this case) which is present in the marked and absent from the unmarked term.

As regards the explanation of the phenomenon – which is common in the entire Portuguese lexicon,[4] not only in kinship terminology – there is little of interest to be said. On the one hand, it is a fact that in all Portuguese-speaking societies the status of man is markedly superior to that of woman; male humans are regarded as humans par excellence. On the other hand, the Portuguese language with its well-developed twofold gender system offers ideal conditions to represent this sociocultural fact linguistically.

The case of the converse pairs[5] is not quite so straightforward. Kay (l974: 133f.) gives evidence from English and Tagalog to show that in such pairs the senior is unmarked and the junior marked.[6] The English examples are valid for Portuguese, too:

João tem um sobrinho. – João é tio.
Joã̃o tem um tio. – *Joã̃o é sobrinho.

The example shows that one of the terms can appear in certain syntactic contexts in which the other cannot – a phenomenon rather similar to the neutralization in favor of the unmarked term that we typically find in markedness pairs. Another piece of evidence is that we normally say *pai e filho*, *tio e sobrinho*, *avô e neto*, and not *filho e pai*, *sobrinho e tio*, *neto e avô*.[7] We therefore conclude that there must be a mark in the lexical representations of *filho*, *sobrinho*, and *neto* which is absent from *pai*, *tio*, and *avô*. Let us look at the descriptions of *pai* and *filho*. *X é pai de y* would have the representation P (x, y), nothing more. But would this not be the representation of y a *filho*. If we followed this argument, we would forget the relation between semantics and syntax. In any grammar of the kind presupposed here there must exist a regular mapping of places of arguments of relational predicates into syntactic functions like subject, object, or genitive complement.[8] Following generally accepted conventions of predicate calculus notation, we might provisionally stipulate that it is the first argument of a relational predicate which gets into subject position if one of the arguments is to be represented by the syntactic subject. Then from P (x, y) we get *x é pai de y*, but not *y é filho de x*. We might, of course, use a predicate Q, writing Q (x,y) and reading 'x is child of y' (cf. Kay 1974). Then the situation would be the reverse. In either case we would be able to associate

[4] And there is, of course, a parallel phenomenon concerning the masculine and feminine genders in the grammar; see Martin 1975.

[5] Actually, for pairs like *tia : sobrinha* to be perfect converses it would be necessary that *x é sobrinha de y* be synonymous to *y é tia de x*. This is not the case because in the first phrase x has to be female and in the second not, and vice versa with y. But this is a terminological issue which does not affect the argument.

[6] More evidence, though not directly relevant to Portuguese, may be found in Greenberg (1965: 100-111). Cahuilla offers interesting new data (Seiler 1977: 6): the converse terms are identical for corresponding seniors and juniors, except that the juniors have an additional suffix, which points to their markedness.

[7] Concerning the nature of this kind of irreversibility in such pairs see Malkiel 1959.

[8] This argument is not affected by the otherwise important problem of the eventual necessity to substitute the simple order relations implied in the predicate calculus notation by an explicit indication of the valencies or semantic roles played by the arguments in a Fillmorean sense. Cf. also Lyons (1977: 481 ff.).

one of the converse terms directly with the lexical representation, and for the other we would need a semantic rule which moves the second argument of the relational predicate in the first position so that it can be related to the syntactic subject. In the light of the markedness evidence just presented we opt for the predicate P, putting the burden of the semantic rule on *filho* and, in general, on the juniors of the converse pairs. This additional semantic rule might be regarded as the semantic surplus which marks the junior against the senior terms. It is also possible to represent this mark lexically, using P', the converse relation of P, in the semantic description of the junior terms, thus representing *x é filho de y* by P' (x, y). This is what we shall do in the analysis, and in order to be consistent we shall conventionalize the following: whenever the individual variable which appears as the subject of the kinship expression to be analyzed (by convention, this will always be x) constitutes part of a proposition in the semantic description, it has to occupy the place of the first argument; and whenever the variable which appears as the genitive complement of kinship expression (it will always be y) constitutes part of a semantic proposition, it has to occupy the place of the second argument. Of the two relational predicates we are dealing with, MAR will remain unaffected by this convention because it is a symmetrical relation; for P, however, it will sometimes have the consequence of changing it to P'. Since the apostrophe is the concretization of the semantic rule which constitutes the mark of the marked terms, we will in our complexity measuring count one point for it. Though this may seem somewhat arbitrary, it is at least consistent.

2.2 Semantic representations and complexity measuring of Portuguese kinship terms

The kinship terms to be analyzed are, in alphabetical order: *avô* ‘grandfather’, *avô* ‘grandmother’, *concunhada* ‘sister or wife (according to type) of *cunhado*’, *concunhado*, ‘brother or husband (according to type) of *cunhada*’, *cunhada* ‘sister-in-law’, *cunhado* ‘brother-in-law’, *esposa* ‘wife’, *filha* ‘daughter’, *filho* ‘son’, *genro* ‘son-in-law’, *irmã* ‘sister’, *irmão* ‘brother’, *mãe* ‘mother’, *marido* ‘husband’, *neta* ‘grand-daughter’, *neto* ‘grand-son’, *nora* ‘daughter-in-law’, *pai* ‘father’, *prima* ‘female cousin’, *primo* ‘male cousin’, *sobrinha* ‘niece’, *sobrinho* ‘nephew’, *sogra* ‘mother-in-law’, *sogro* ‘father-in-law’ , *tia* ‘aunt’, *tio* ‘uncle’. The terms not to be analyzed include the compounds of the type *tio politico*, *tio-avô*, the step- and half-relatives (the latter being compounds, too), and the ancestors and descendants beyond *avôs* and *netos*, respectively (the remoter of them being compounds, too). The order of presentation is systematic in a self-explanatory manner. See Table 1.

Table 1. Formal semantic representations for Portuguese kinship terms

x pai y:		P (x, y)	x mãe y:		P (x, y)
(1)			(2)	&	F (x)
x filho y:		P' (x, y)	x filha y:		P' (x, y)
(2)			(3)	&	F (x)
x irmão y:		P' (x, z)	x irmã y:		P' (x, z)
(4)	&	P (z, y)	(5)	&	P (z, y)
	&	x ≠ y		&	x ≠ y
				&	F (x)
x tio y:		P (z_1, y)	x tia y:		P (z_1, y)
(4.1)	&	P (z_2, z_1)	(5.1)	&	P (z_2, z_1)
	&	((P' (x, z_2)		&	((P' (x, z_2)
	&	x ≠ z_1)		&	x ≠ z_1)

	∨	(P (z_2, z_3)		∨	(P (z_2, z_3)
	&	MAR (x, z_3)		&	MAR (x, z_3)
	&	$z_1 \neq z_3$))		&	$z_1 \neq z_3$))
				&	F (x)
x sobrinho y:		P' (x, z_1)	x sobrinha y:		P' (x, z_1)
(4.5)	&	(z_2, z_1)	(5.5)	&	P (z_2, z_1)
	&	((P (z_2, y)		&	((P (z_2, y)
	&	$z_1 \neq$ y)		&	$z_1 \neq$ y)
	∨	(P (z_2, z_3)		∨	(P (z_2, z_3)
	&	MAR (z_3, y)		&	MAR (z_3, y)
	&	$z_1 \neq z_3$))		&	$z_1 \neq z_3$))
				&	F (x)
x primo y:		P' (x, z_1)	x prima y:		P' (x, z_1)
(6)	&	P (z_2, z_1)	(7)	&	P (z_2, z_1)
	&	P (z_2, z_3)		&	P (z_2, z_3)
	&	P (z_3, y)		&	P (z_3, y)
	&	$z_1 \neq z_3$		&	$z_1 \neq z_3$
				&	F (x)
x avô y:		P (x, z)	x avó y:		P (x, z)
(2)	&	P(z, y)	(3)	&	P (z, y)
				&	F (x)
x neto y:		P' (x, z)	x neta y:		P' (x, z)
(4)	&	P' (z, y)	(5)	&	P' (z, y)
				&	F (x)
x marido y:		MAR (x, y)	x esposa y:		MAR (x, y)
(2)	&	M (x)	(2)	&	M (x)
x sogro y:		P (x, z)	x sogra y:		P (x, z)
(2)	&	MAR (z, y)	(3)	&	MAR (z, y)
				&	F (x)
x primo y:		P' (x, z_1)	x prima y:		P' (x, z_1)
(6)	&	P (z_2, z_1)	(7)	&	P (z_2, z_1)
	&	P (z_2, z_3)		&	P (z_2, z_3)
	&	P (z_3, y)		&	P (z_3, y)
	&	$z_1 \neq z_3$		&	$z_1 \neq z_3$
				&	F (x)
x genro y:		MAR (x, z)	x nora y:		MAR (x, z)
(3)	&	P' (z, y)	(4)	&	P' (z, y)
				&	F (x)
x cunhado y:		P (z_2, z_1)	x cunhada y:		P (z_2, z_1)
(3. 5)	&	((MAR (x, z_1)	(4.5)	&	((MAR (x, z_1)
	&	P (z_2, y)		&	P (z_2, y)
	&	$z_1 \neq$ y)		&	$z_1 \neq$ y)
	∨	(P' (x, z_2)		∨	(P' (x, z_2)
	&	$z_1 \neq$ x		&	$z_1 \neq$ x
	&	MAR (z_1, y)))		&	MAR (z_1, y)))

				&	F (x)
x concunhado y:		P (z_3, z_1)	x concunhada y:		P (z_3, z_1)
(4.8)	&	((P' (x, z_3)	(5.8)	&	((P' (x, z_3)
	&	$z_1 \neq x$		&	$z_1 \neq x$
	&	MAR (z_1, z_2)		&	MAR (z_1, z_2)
	&	P (z_4, z_2)		&	P (z_4, z_2)
	&	P (z_4, y)		&	P (z_4, y)
	&	$z_2 \neq y$)		&	$z_2 \neq y$)
	v	(MAR (x, z_1)		v	(MAR (x, z_1)
	&	P (z_3, z_2)		&	P (z_3, z_2)
	&	$z_1 \neq z_2$		&	$z_1 \neq z_2$
	&	MAR (z_2, y)))		&	MAR (z_2, y)))
				&	F (x)

(Those not trained in reading this type of kinship analysis may find it helpful to be reminded that there are two types of *cunhados* and, accordingly, of *concunhados*. In the two disjoined series of conjunctions which make up the *significata* of the four terms, this does not become as clear as it might because one proposition – it is the first in each *significatum* – has been extracted from the disjunction by the law of distributivity; see Section 1.3.)

The number which appears in parentheses under each term is its complexity coefficient, which has been calculated according to the formulas for c given in Section 1.3, minding what has been said in Section 2.1, last paragraph. Let us recalculate one example: The representation of *sobrinho* consists of two conjoined propositions, followed by a disjunction whose first disjunct consists of two and whose second disjunct consists of three conjoined propositions. The formula gives us c = 2 for the first two conjuncts, plus one for the apostrophe appearing in the first. The specific formula for disjunction of conjunctions gives $c = 2 + 3 - \text{ld}\,(2^2 + 2^3 - 1) = 5 - \text{ld}\,11 = 5 - 3.5 = 1.5$ for the rest. Result: 2 + 1 + 1.5 = 4.5, as indicated in the table.

2.3 Semantic complexity relations

We are now in a position to order our semantic field in various ways according to the semantic complexity of the terms. Let us briefly return to the two markedness relations discussed in Section 2.1. As we put the sex specification only in the *significata* of the female terms (with the exception of *marido*), these are infallibly by one point more complex than the corresponding male terms. This is, though rather uninterestingly, the first semantic parameter by which the field can be arranged in ordered pairs and which is, simultaneously, a complexity relation. The same is true for the converse pairs, though the complexity surplus of the junior terms does not always equal exactly one point. (This is a – perhaps inadequate – consequence of the stipulation made in Section 2.1, last paragraph.) Thus, the senior-to-junior converseness is the second semantic relation which is at the same time a complexity relation.

We now come to complexity relations which have not been programmed in the semantic analysis but result from it. In the three central generations, there are, within consanguinity, minimal pairs whose members are distinct by the fact that the relationship of the second is mediated by one intervening sibling, whereas the relationship of the first, more direct, is not so mediated. The pairs, with their complexity coefficients indicated, are:

pai	:	tio	mãe	:	tia

(1)		(4.1)	(2)		(5.1)
filho	:	sobrinho	filha	:	sobrinha
(2)		(4.5)	(3)		(5.5)
irmão	:	primo	irmã	:	prima
(4)		(6)	(5)		(7)

There are four terms in the relationship by marriage which may be ordered in the same way:

marido	:	cunhado	esposa	:	cunhada
(2)		(3.5)	(2)		(4.5)

We see that the complexity coefficients of the terms in the second and fourth columns are constantly higher than those of the corresponding terms in the first and third columns. This is, then, the third complexity relation by which the field can be ordered.

In a parallel fashion we can form ordered pairs in which the relationship of the second term is mediated by a marriage not present in the first term:

pai	:	sogro	mãe	:	sogra
(1)		(2)	(2)		(3)
filho	:	genro	filha	:	nora
(2)		(3)	(3)		(4)
irmão	:	cunhado	irmã	:	cunhada
(4)		(3.5)	(5)		(4.5)

Cunhado/-a, which appears a second time due to its ambiguity, is the one failure in the picture: it ought to be more complex than *irmã*(*o*) as are the other terms in the second and fourth columns as opposed to those in the first and third. There are two more pairs which might, according to the ambiguity of *concunhado*, figure either in this type of relation or in the preceding:

cunhado	:	concunhado	cunhada	:	concunhada
(3.5)		(4.8)	(4.5)		(5.8)

They fit both of the respective complexity relations.

The last semantic relation we will examine is that of more distant generation to proximate generation. The pairs are:

pai	:	avô	mãe	:	avó
(1)		(2)	(2)		(3)
filho	:	neto	filha	:	neta
(2)		(4)	(3)		(5)

It turns out that this semantic relation is definable as a complexity relation, too.

Now that the third, fourth, and fifth semantic relations have emerged from the analysis as complexity relations, certain linguistic facts come to mind which resemble very much the

markedness phenomena discussed in Section 2.1 (cf. once more Greenberg l966: l00-111). As regards the relation of lineal to collateral relatives, there is the paradigmatic defectivity of the collateral terms which points to their markedness: whereas the lineal consanguineous relatives *pai*, *filho*, and *irmão* have corresponding affinal relatives opposed to them, viz. *sogro*, *genro*, and *cunhado*, which are monomorphemic, in the case of the collateral consanguineous relatives *tio*, *sobrinho*, and *primo* such corresponding affinal terms are either compounds, viz. *tio político*, *sobrinho político*, or altogether missing, as there is no **primo político*.

Neutralization may be observed in the case of the affinal vs. consanguineous relationship: In certain colloquial styles, especially in the lower social classes, one may refer to what is properly *sogros* as *pais*, and to *genros* as *filhos*. In certain contexts this substitution is obligatory for all speakers; for instance, a couple may talk about their joint *pais* but not about their *sogros*. And as regards the case of the generation distance, it is common to refer to the descendants in general as *nossos filhos* and to the ancestors as *nossos pais*. It is true that *nossos netos* and *nossos avós* may be used in the same sense, but these exclude the *filhos* and *pais*, respectively, so that it remains true that the former terms are more inclusive. Thus we see that on closer analysis, all of our five complexity relations may be described as markedness relations. In a certain sense, the theory of semantic complexity is an extended semantic markedness theory although the former cannot be reduced to the latter since this cannot incorporate several features of the theory of semantic complexity, such as the special treatment of polysemous *significata*.

The result of this section is that there are several linguistically significant ways in which we may arrange our terms in series of ordered pairs, and with the single exception of the pair *irmão* : *cunhado* (*irmã* : *cunhada*), there is consistently a corresponding complexity relation between the members of the pairs. I am conscious of the fact that given the complexity relations between males and females and those between all the first and second columns analyzed in this section, the correspondences in the third and fourth columns are a logical consequence and not additional evidence. Their enumeration is, however, a presupposition to a better understanding of what follows.

2.4 Semantic complexity and amount of information.

The empirical correlate to the semantic complexity of a term is its amount of information. Therefore, in order to test the empirical significance of our analysis, we check the kinship terms in a Portuguese kinship dictionary, calculate, on the basis of the frequencies found there, the amount of information for each term, arrange them once more according to the five semantic relations discussed in the preceding section, and look to see whether there is a correspondence between the differences in complexity and those in information quantity. It is clear that the transformation of frequencies in information quantities changes nothing in the statistical relations between the terms, except for inverting them. I do this in consistency with my argument according to which the intrinsic, direct relation exists between complexity and amount of information and not between complexity and frequency.

Duncan's (1971) frequency dictionary is based on a count of 500,000 words. Words with less than five occurrences were not included in it. That is why I give <4 as the frequency of some of the terms. In the calculation of the information quantity of these it is consequently only possible to indicate a lower limit, which is the same for all of them.[9]

[9] It is probable that some of the missing terms occurred in Duncan's corpus with lesser frequency. A word with only one occurrence would have i = 18.9.

One adjustment had to be made in the data: Since Duncan's dictionary – that is, Duncan's computer – neglects diacritic marks, it creates new homonyms. There appears an entry *avo*, which is a rare word and presumably did not occur in the corpus. In the thirty occurrences displayed, it mixes up the joint occurrences of *avô* and *avó*. Instead of removing this from the data, I distributed the thirty occurrences between the two words, extrapolating the two figures by means of a rule-of-three on the basis of the quantitative relations between males and females in the whole field.

In the alphabetical list which follows I give for each term first the absolute frequency according to Duncan – he does not indicate relative frequencies – and second the amount of information calculated by the formula i = – ld p.[10]

Avô 20, 14. 6; *avó* 10, 15.6; *concunhada* <4, >16.9; *concunhado* <4, *>16.9; cunhada* <4, >16.9; *cunhado* 6, 16.3; *esposa* 12, 15.3; *filha* 109, 12.2; *filho* 240, 11.0; *genro* <4, >16.9; *irmã* 37, 13.8; *irmão* 62, 13.0; *mãe* 181, 11.4; *marido* 88, 12.5; *neta* 6, 16.3; *neto* 19, 14.7; *nora* <4, >16.9; *pai* 388, 10.3; *prima* 8, 15.9; *primo* 11, 15.5; *sobrinha* 6, 16.3; *sobrinho* 14, 15.0; *sogra* 5, 16.6 *sogro* 13, 15.2; *tia* 21, 14.5; *tio* 25, 14.3.

We now repeat the arrangement of the field according to our five semantic complexity relations, giving this time each term accompanied by its i (cf. the frequency tables displayed in Greenberg 1966: 106 ff). See Table 2.

Table 2. Sex relation

pai (10.3)	:	mãe (11.4)	filho (11.0)	:	filha (12.2)
irmão (13.0)	:	irmã (13.8)	tio (14.3)	:	tia (14.5)
sobrinho (15.0)	:	sobrinha (16.3)	primo (15.5)	:	prima (15.9)
avô (14.6)	:	avó (15.6)	neto (14.7)	:	neta (16.3)
marido (12.5)	:	esposa (15.3)	sogro (15.2)	:	sogra (16.6)
genro (>16.9)	:	nora (>16.9)	cunhado (16.3)	:	cunhada (>16.9)
concunhado (>16.9)	:	concunhada (>16.9)			

Exempting the case of the *avós*, which necessarily conforms to the others, and that of the *genros* and *concunhados*, which cannot serve to confirm or disconfirm anything, there is not a single exception to the rule that the female terms have a higher amount of information than the corresponding male terms. Second, this corresponds exactly to our observation about the complexity relation between males and females. The correlation is so close that there is even

[10] The probability is taken to be simply p = x/n, where x is the absolute frequency and n the size of the corpus, 500,000 in this case. We can circumvent the calculation of p by substituting it by x/n in the formula for i: i = -ld (x/n) = -(ldx – ldn) = ldn – ldx.

an average difference of one bit in the information quantities which corresponds to the difference of one proposition in the semantic complexities.

The single case that behaves exceptionally, i.e. the pair *marido* : *esposa*, has an explanation. *Esposa*, which is exactly as complex as *marido*, has so much higher an i because it shares with *mulher* the occurrences of the Portuguese equivalent to 'wife'. This hypothesis could be proved if Duncan's dictionary distinguished between the relational and non-relational uses of *mulher*, which it does not.[11] There is, however, an indirect proof: If we presuppose a general frequency relation of 2: 1 between corresponding males and females, as it emerges from our semantic field, we should expect a similar relation in the pair *homem* : *mulher*. *Homem* 'man' has 614 occurrences in Duncan; but instead of the expected ca. 307 occurrences of *mulher* there are 377. If we separate the 70 excess occurrences, in the assumption that they represent the meaning 'wife' (which has, of course, no analog in *homem*), and add them to the 12 occurrences of *esposa*, we get 82 occurrences for the counterpart of *marido*. The amount of information would be of 12.6 bits, almost identical to the 12.5 bits of *marido*. Viewed from this angle, the pair *marido* : *esposa* can be taken to confirm the general rule. See Table 3.

Table 3. Senior-to-junior converseness

pai	:	filho	mãe	:	filha
(10.3)		(11.0)	(11.4)		(16.3)
tio	:	sobrinho	tia	:	sobrinha
(14.3)		(15.0)	(14.5)		(16.3)
avô	:	neto	avó	:	neta
(14.6)		(14.7)	(15.6)		(16.3)
sogro	:	genro	sogra	:	nora
(15.2)		(>16.9)	(16.6)		(>16.9)

This time there is not a single exception to the rule that the juniors have a higher i than their respective seniors. So here we have a 100 percent correlation between semantic complexity and amount of information. See Table 4.

Table 4. Terms distinct by an intervening sibling

pai	:	tio	mãe	:	tia
(10.3)		(14.3)	(11.4)		(14.5)
filho	:	sobrinho	filha	:	sobrinha
(11.0)		(15.0)	(12.2)		(16.3)
irmão	:	primo	irmã	:	prima
(13.0)		(15.5)	(13.8)		(15.9)
marido	:	cunhado	esposa	:	cunhada
(12.5)		(16.3)	(15.3)		(>16.9)

[11] It may be remarked here that so-called 'semantic counts' of the kind that has been done in 1938 by Lorge and Thorndike for English are an urgent desideratum for the kind of research exemplified by this article.

Once more, there is an exceptionless rule that the terms of the second and fourth columns have a higher amount of information than the corresponding terms in the first and third columns, even if we regard *esposa* instead of the artificial **marida* calculated in the discussion of the first relation. Thus, this is one more 100 percent correspondence between c and i. It is also worth mentioning that the information surpluses, which in the first two relations centered around an average of one bit, this time are markedly higher; this has also a parallel in the complexity surpluses. See Table 5.

Table 5. Terms distinct by an intervening spouse

pai (10.3)	:	sogro (15.2)		mãe (11.4)	:	sogra (16.6)
filho (11.0)	:	genro (>16.9)		filha (12.2)	:	nora (>16.9)
irmão (13.0)	:	cunhado (16.3)		irmã (13.8)	:	cunhada (>16.9)
cunhado (16.3)	:	concunhado (>16.9)		cunhada (>16.9)	:	concunhada (>16.9)

The difference in the amount of information is constant here, too. Even the pairs *irmão* : *cunhado*, *irmã* : *cunhada*, which had been exceptional in their complexity relations, here fit into the picture. This makes us suppose that the fault is with the complexity measure, which should have attributed a higher complexity to the *cunhados*. We shall return to this point. Apart from this case, there is once more a correspondence between semantic complexity and amount of information. See Table 6.

Table 6. Terms distinct by an intervening generation

pai (10.3)	:	avô (14.6)		mãe (11.4)	:	avó (15.6)
filho (11.0)	:	neto (14.7)		filha (12.2)	:	neta (16.3)

There is a constant difference in amount of information which has its exact parallel in the difference in complexity.

The result of this section is twofold. First, there are five semantic relations in our field which define series of ordered pairs of terms; and exactly the same ordering, without exception, can be imposed on the pairs using the criterion of information quantity. It must also be mentioned that, in distinction to the case of the preceding section, the relations between columns three and four are not a logical consequence of the relations obtaining between columns one and two and can therefore be regarded as independent evidence. Second, this same ordering can, with one apparent and one real exception, be effected by means of a third criterion which is the complexity measure. I take this to be confirmatory evidence of its empirical nature.

3 Some open questions

The results are no doubt impressive as far as they go, but various uncomfortable points remain. First, the question comes up naturally: would the same results have been achieved in any other semantic field? The answer is: yes and no. The fact is that kinship terminology is the one semantic field which is on the one hand so clearly structured and has on the other hand received so much attention by linguistics that we are in a position to give it formalized treatment which is the necessary condition for the application of the complexity measure. Its successful application to other data presupposes an equal degree of explicitness in their analysis. Only on this condition will it be possible to reply to the suspicion that the type of semantic bring out the inner complexity of kinship terms and might fail to do this job in other semantic fields.

Second, there is a series of questions which point to interrelated problems and cannot, therefore, be treated separately: If there is such a close relation between semantic complexity and amount of information, why did the analysis have to be done in terms of ordered pairs? Would it not have been simpler to order all the terms of the field in one line according to their complexity, then order them once more according to their amount of information, and finally compare the two orders? Do they not have to coincide, too? Furthermore, why is there no numeric correspondence between the c and the i of the same word? How do we account for the gross discrepancies, e.g. *pai*, c = 1, i = 10.3, as opposed to *prima*, c = 7, i = 15.9? Finally, why did the complexity measure fail in the pair *irmão* : *cunhado*?

To give a partial answer to the first of these questions: I have tested the correlation between the two ordered series, and although there is a promising correlation coefficient, it is far from overwhelmingly impressive. The reasons for this and all the other problems are, to my mind, the following: Recall that we extracted identical propositions from disjunctions of conjoined propositions. If we had left them in the disjunctions, they would have contributed at most 0.4 points to complexity (this varies according to the number of conjoined propositions involved in the disjunction). Now that they were extracted, they contributed one point each. But the non-identical propositions frequently have much in common. In *cunhado*, e. g. we have the disjunctive specification of either the spouse of a sibling or a sibling of the spouse. In either case, we have the two predicates MAR and P (or P') and the statement of the non-identity of the siblings; only the arguments are different. This is the natural consequence of the fact that the two senses of *cunhado* are not wholly disparate but are, in fact, converses: barring sex differences, *cunhado* is a symmetric term. Now *irmão* (as well as *primo*) is also a symmetric term; but since its *significatum* reads, so to say, the same from left to right and from right to left, thus being the same for both of the brothers, it needed to be stated only once, whereas the *significatum* of *cunhado*, which is the reverse for the second of the two *cunhados*, had to be stated in a disjunction. So there is fault either with our semantic analysis, which does not make the symmetry explicit, or with the complexity measure, which does not account for the fact that the two disjuncts are almost identical. If, in measuring the complexity of *cunhado*, we were permitted to regard the two senses as identical, we would come up with c = 4.5, which does seem more adequate.

There is a different way of approaching the same problem: Note that, intuitively, the higher complexity of *cunhado* as opposed to *irmão* (and, in fact, the different complexity in all the pairs in the semantic relations 3, 4, and 5 of Section 2.4) is based on the fact that there are more different persons involved in the relationship. This is expressed in the semantic representations, since that of *cunhado* operates with four distinct variables, while in *irmão* we have only three. So this should be taken account of by the complexity measure.

The solution to the problem appears to be: Count symbols instead of propositions and extract identical symbols from disjunctions by a rule like distributivity. But then we immediately face a whole mess of new problems: Would it not make a difference whether two identical symbols appeared in totally different or in similar propositional frames, whether, say, x appeared twice as the first argument of P or once as the first and once as the second? The construction of elementary propositions could not simply be regarded as a sort of conjunction of symbols; otherwise, each symbol would count only once in a given *significatum*, since (p & p) ↔ p. As I did not know the solution to such problems, I contented myself with regarding propositions as unanalyzable, establishing thus the proposition as the unit of measure.

There is a related set of problems. *Esposa* has the same complexity as *mãe*, but their information quantities differ markedly: *mãe* has 11.4 bits, *esposa* has 15.3 or, if we admit the figure for **marida*, 12.6 bits. *Genro* has c = 3 and i >16.9, whereas *irmão* has c = 4 and i = 13.0. If we look at the propositions that make up for these c's we see that *esposa* and *genro* have the predicate MAR where *mãe* and *irmão* have P or P' (*irmão* has even a surplus non-identity statement). The observed disproportions seem to be explicable if we admit that not all predicates contribute to complexity with equal weight: MAR ought to contribute more than P. This is linguistically justifiable because the proposition MAR (x, y) implies many more other propositions than does P (x, y). While the latter requires only that x be adult, the former requires this for both x and y. While for P the sexes of x and y are unimportant, MAR requires that they be distinct. While P is a relation of two to indefinitely many, MAR is a relation of one to one. And so on. On the other hand, the statement x ≠ y has no implications at all, and that is why it ought not to raise the complexity of *irmão* as opposed to *genro* essentially.

This different implicative potential (cf. Lehmann l975: l08-110) of the primitive predicates should be taken into account in the complexity measure. But this, again, raises problems. When we argued about the markedness relation between the male and female terms, we saw that the complexity measure has to be applied to the lexical representations because derived representations may contain specifications which must not contribute to complexity. If we want to include the implicative potential of a predicate into the complexity, how do we distinguish between implied propositions that have to be counted and others that do not? And if this problem is resolvable: Does it suffice to determine the implicative potential for every predicate separately and once for all, so that the implicative potentials of the predicates of conjoined propositions would be additive? Or would we have to consider the predicates present in a given *significatum* jointly and derive from them the set of distinct propositions that it is possible to derive? Exemplifying: Each of the two propositions making up the representation of *sogro* implies that z is human and animate. Does this count twice or once? And if we opt for the latter solution: would we not necessarily also have to make all the mediate inferences which are possible only if two or more propositions are combined? Where would this end? Finally: do the logical properties of the relations contribute to complexity? And if so: would transitivity or intransitivity be marked? As I did not know the solution to such problems, I contented myself with regarding predicates, too, as unanalyzable, counting only those propositions which are necessary to an unambiguous specification of the meaning of the terms. But I am sure that if these problems were solved, the questions posed at the beginning of this discussion would receive a satisfactory answer.

Doubtless I have raised more problems than I have solved. But if I am right in principle that there is a necessary correlation between the amount of information transmitted by a word and its semantic complexity as here conceived, and if this latter has any linguistic significance, this means that in the future, after the necessary refinements have been made, the check-up in a frequency dictionary can be used as a partial test of the adequacy of semantic

descriptions.[12] This is where I see the possibilities of a rapprochement between semantics and statistical linguistics.

References

Bar-Hillel, Yehoshua. 1955. An examination of information theory. *Philosophy of Science* 22.86- 105. Cited according to reimpr. Bar-Hillel 1964: 275- 297.

Bar-Hillel, Yehoshua. 1964. *Language and information. Selected essays on their theory and application.* Reading, Mass.: Addison-Wesley.

Carnap, Rudolf. 1950. *Logical foundations of probability.* Chicago: University of Chicago Press.

Carnap, Rudolf and Yehoshua Bar-Hillel. 1952. An outline of a theory of semantic information. *Technical Report* No. 247. Cambridge, Mass.: MIT Research Laboratory of Electronics. Cited according to reimpr. Bar-Hillel 1964: 221-274.

Chomsky, Noam and Morris Halle. 1968. *The sound pattern of English.* New York: Harper and Row.

Duncan, John C., Jr. 1971. *A frequency dictionary of Portuguese words*. 2 vols. Stanford University, Ph.D. Dissertation, University Microfilms No. 71-19, 676.

Greenberg, Joseph H. 1966. Language universals. In: *Current trends in linguistics 3*. Edited by Thomas A. Sebeok. The Hague/Paris: Mouton. 61-112.

Katz, Jerrold J. and Jerry A. Fodor. 1963. The structure of a semantic theory. *Language* 39. 19-29.

Kay, Paul. 1974. On the form of dictionary entries: English kinship semantics. In: *Towards tomorrow's linguistics*. Edited by R. W. Shuy, Ch.-J. N. Bailey. Washington, D. C.: Georgetown University Press. 120-138.

Lehmann, Christian. 1974. Isomorphismus in sprachlichen Zeichen. In: *Linguistic workshop II*. Edited by H. Seiler. München: Fink. 98-123.

Lyons, John. 1977. *Semantics*. Cambridge: Cambridge University Press.

Malkiel, Yakov. 1959. Studies in irreversible binomials. *Lingua* 8. 113-160.

Martin, John W. 1975. Gênero? *Revista Brasileira de Lingüística* 2. 3-8.

Sanders, Gerald A. 1974. The simplex feature hypothesis. *Glossa* 8. 141-192.

Seiler, Hansjakob. 1977. Two systems of Cahuilla kinship expressions: Labeling and descriptive. *Akup* 27. Köln: Inst. f. Sprachwissenschaft der Univ.

Thom, René. To appear. La double dimension de la grammaire universelle. In: *Language universals. Papers from the conference held at Gummersbach/Cologne, Germany, October 3-8, 1976*. Edited by H. Seiler.

Wiener, Norbert. 1950. *The human use of human beings*. Boston: Houghton Mifflin; London: Eyre and Spottiswoode.

[12] There is no space here to discuss the obvious problems related to differences in the corpora which frequency dictionaries are based upon.