COLING 82, J. Horecký (ed.)
North-Holland Publishing Company
© Academia, 1982

MARKEDNESS AND FREQUENCY: A COMPUTATIONAL ANALYSIS

Henry Kučera

Brown University
Providence, Rhode Island

U.S.A.

When the markedness analysis is extended to the lexical and grammatical levels, the question arises whether an analogue of the markedness/frequency correlation, observed in phonology, also exists on these higher linguistic levels. This article presents evidence that in some interesting cases, such as tense and aspect forms in English, the correlation does not hold and that this failure is due not simply to superficial stylistic factors but rather to the inability of the markedness hypothesis to provide an adequate framework for the analysis of the English verbal system.

Computational linguistics provides important potential tools for the testing of theoretical linguistic constructs and of their power to predict actual language use. The present paper falls into this general area of inquiry. The specific issue, which I shall deal with, is the relation of the so-called marked and unmarked categories which have been posited for different levels of linguistic description. Greenberg (1966) considers this concept to be one of the important notions in the study of language universals and devotes a long essay to this subject. He argues that the concept of the marked and unmarked can be shown "to possess a high degree of generality in that it is applicable to the phonological, the grammatical, and the semantic aspects of language." It is in particular this generality claim which I plan to address in my report. A consideration of these questions seems especially appropriate at this COLING conference since the theory of markedness had its home in Prague.

The concept of markedness originated in phonology and is most systematically presented in the work of Trubetzkoy (1939). On this level, one can observe a clear and convincing correlation between the markedness property and the statistical facts of language use. The unmarked phonemes, e.g. short vowels, occur with considerably higher frequencies than the marked ones, e.g. long vowels, in all languages where this opposition has been investigated. Greenberg (1966:67-69) gives data from Icelandic, Sanskrit, Czech, Hungarian, Finnish, Karok and Chirichua to demonstrate this point. Similar statistical relations can be found with reference to other unmarked vs. marked oppositions, such as voiceless vs. voiced obstruents, non-palatalized vs. palatalized consonants, non-nasal vs. nasal vowels, etc. (For some statistical data in Czech, German, and Russian, cf. Kučera and Monroe 1968.) This correlation between structure and language statistics is important since it supports the basic notion that the markedness relation, in reflecting an informational economy in language coding, has the expected statistical effects.

In the early 1930's, the markedness relation was extended, again in the work of the Prague school (primarily by Jakobson 1932), to higher levels of language, specifically to the semantic analysis of the lexicon and the meaning of grammatical categories, such as case, number, person, tense and aspect. Significantly, Jakobson's initial example of this non-phonological markedness relation was a lexical one, the Russian words *osël* 'donkey' and *oslica* 'female donkey'. "Wenn ich *osël* sage, bestimme ich nicht, ob es sich um ein Männchen oder ein Weibchen handelt, aber fragt man mich '*èto oslica*?' und ich antworte '*net*, *osël*', so wird hier das männliche Geschlecht angekündigt - das Wort is in verengter Bedeutung angewandt." As in the English contrast between *man* and *woman*, or *lion* vs. *lioness*, the crux of the issue is that the first term, the unmarked one, has two possible usages, the general and the specific one, while the second term, the marked one, carries only a single value with reference to the relevant semantic concept which, in all the above lexical examples, is the feature [feminine]. It is because of this dual function of the unmarked term that such sentences as *Is this lion a lion or a lioness?* -- although perhaps somewhat unusual -- are certainly not semantically incongruous. In his later work, Jakobson (1971) provided a more formal definition of the markedness relation on higher language levels: "The general meaning of a marked category states the presence of a certain (whether positive or negative) property A; the general meaning of the corresponding unmarked category states nothing about the presence of A, and is used chiefly, but not exclusively, to indicate the absence of A. The unmarked term is always the negative of the marked term, but on the level of general meaning the opposition of the contradictories may be interpreted as 'statement of A' vs. 'no statement of A', whereas on the level of 'narrowed', nuclear meanings, we encounter the opposition 'statement of A' vs. 'statement of non-A'."

It is important to realize that the markedness relation, as given in this definition, is a more complex one than Trubetzkoy's privative phonological opposition. This has certain interesting consequences to which I will turn later in my paper. At this juncture, however, I want to consider the natural question of whether the correlation between markedness and frequency also holds in non-phonological cases as, it might be reasonable to assume, the theory would predict.

On the lexical level, the markedness/frequency correlation seems to hold rather well. In representative English texts, such as the one-million-word Brown Corpus, the frequencies of the unmarked members of the opposition in such pairs as *man* vs. *woman*, *lion* vs. *lioness*, or *he* vs. *she*, are from 3.5 to 5.5 times greater than the frequencies of the marked members of each opposition. Since the unmarked member of such lexical oppositions does indeed have a broader semantic scope--capable of both the generic and the specific usage--its higher frequency is hardly surprising. Even so, the correlation is not without interest since it points out that certain relations (such as homonymy, hyponymy, as well as markedness proper), which are present in the design of a language system, make a more economical use of the available formal distinctions possible.

What is of particular interest to us, however, is the claim of the adherents of the markedness analysis that the oppositions within grammatical categories, such as number, person, tense or aspect, also exhibit the same logical relation and thus should be analyzed within this framework. The opposition of the singular number vs. the plural number usually provides a 'classical' illustration of the various properties that are supposed to support this analysis (for details, cf. Greenberg 1966:73). The singular, being "unmarked", is said to have two potential functions, a general one (as in such generic usages as *The beaver builds dams*) as well as a specific one, indicating individuality (as in *The beaver ran under the house*). Plurals, on the other hand, are said to represent the marked member of the opposition and signal this specific plural function only. Arguments referring to form are often given to support this kind of markedness analysis: in many languages, singulars have no overt mark while the plural is marked by an affix, so that the markedness relation appears to be reflected formally in a semiologically diagrammatic manner. Several other arguments as to the correlation between formal and functional marking have been given; Greenberg (1966) provides a good summary.

The question of whether an analogue to the frequency/markedness correlation exists for grammatical categories was raised already in Greenberg's essay. Greenberg found the problem difficult, largely because there were very few studies available to him which gave information about the frequency of grammatical categories. Some examples which Greenberg collected showed the expected frequency distributions: singulars more frequent than plurals, or a greater frequency of so-called direct cases (nominative, accusative and vocative) than of oblique cases (genitive, dative, ablative, locative, instrumental, etc.) in a number of languages. In other instances, however, the evidence was either inconclusive or contrary to the putative markedness analysis. In Josselson's Russian word count (Josselson 1953), for example, the imperfective aspect of verbs (which is considered the unmarked member of the opposition) is slightly less frequent (46.9%) than the marked perfective (53.1%).

The availability of new frequency data of grammatical categories now makes it possible to reopen the question and to see whether the results can shed light on the adequacy of the markedness hypothesis in this area. The main source for my analysis is the one-million-word Corpus of Present-day American English, assembled at Brown University in the 1960's and recently analyzed grammatically (cf. Francis and Kučera 1982). This data base does constitute a representative sample of printed American English, consisting of 500 samples of texts, each about 2,000 words long, drawn from a variety of sources ranging from newspapers to learned articles and fiction. The grammatical analysis of the corpus is based on a taxonomy of 87 grammatical classes or "tags" (including six syntactically useful punctuation tags), representing an expanded and refined system of word-classes, supplemented by major morphological information and some syntactic information.

In some relatively straightforward cases, our English data clearly support the expected frequency correlation that is consistent with the markedness analysis. The prevalence of singular over plural forms in common nouns shows, in spite of some interesting stylistic differences, an overall frequency ratio similar to that between unmarked and marked lexical forms, namely about 3:1. The same is true of those case forms which are still formally marked in English: roughly the same 3:1 ratio holds for the nominative of personal pronouns (i.e. forms such as 'I, he, she, we, they') vs. objective forms (i.e. 'me, him, her, us, them').

With regard to the more interesting cases, such as tense, the statistical evidence from the Brown Corpus offers both greater problems and greater insight. The adherents of the markedness hypothesis have often attempted to fit the notion of tense into the markedness framework. Tense in some Slavic languages, for example, has been generally analyzed into a binary opposition of the marked past vs. the unmarked nonpast, an analysis substantiated on the grounds that the present forms, too, serve a more general function than simply the localization of the activity as overlapping with the speech moment--such as that of the gnomic present, historical present, or "programmed" future. Greenberg (1966) assumes the same kind of analysis for Latin and Sanskrit and presents figures which show that the "unmarked" present in both of these languages is more frequent than the "marked" past and the future, although the differences between the present and past figures, in the Sanskrit sample, are really quite small.

Let us assume, for a moment, that the past--nonpast relation has some appeal as a linguistic universal and assume that in English, too, the simple present is unmarked and the simple past is the marked member of this opposition. If one takes the frequency figures of the entire Brown Corpus into account, then the past tense predominates above the present; there are altogether 21391 occurrences of the simple present vs. 26172 occurrences of the simple past. The frequency data is thus the reverse of what one might have assumed under the markedness analysis. But the fact of real interest is not this discrepancy alone but rather the fact that, in looking at the individual genres of writing represented in the data base, and counting the present vs. past tense occurrences separately for each genre category, a quite idiosyncratic pattern emerges, as the following table indicates.

DISTRIBUTION OF SIMPLE PRESENT AND PAST FORMS

Genre	Number of	Percent of	Number of	Percent of
	Present Tense	Occurrences	Past Tense	Occurrences
A. Press: Reportage	1271	16.08	2526	31.97
B. Press: Editorial	1405	26.77	700	13.34
C. Press: Reviews	910	28.83	504	15.97
D. Religion	934	27.41	511	14.99
E. Skills and Hobbies	2262	34.15	617	9.31
F. Popular Lore	1973	21.19	2272	24.40
G. Belles Lettres	3302	21.96	3501	23.29
H. Miscellaneous	1034	24.18	405	9.47
J. Learned	3319	24.87	1481	11.10
Subtotal: Informative	16410	24.03	12517	18.33
K. General Fiction	995	13.17	3032	40.13
L. Mystery/Detective	1052	15.47	2643	38.87
M. Science Fiction	253	15.74	531	33.04
N. Adventure/Western	1065	13.17	3702	45.77
P. Romance	1206	14.65	3048	37.02
R. Humor	410	18.89	699	32.21
Subtotal: Imaginative	4981	14.46	13655	39.63
CORPUS	21391	20.82	26172	25.47

In six of the fifteen genre categories, the simple present is more frequent than the past; in nine, the opposite is true. The stylistic reasons for this tense distribution are fairly clear from the characteristics of the genres involved: the present tense prevails in the descriptive genres, while the past predominates greatly in what might be called the narrative genres. Interestingly enough, this tense distribution groups all imaginative prose (Genres K through R) with A. Press: Reportage, in this narrative category. The other two newspaper genres (B. Press: Editorial and C. Press: Reviews), on the other hand, are grouped together with the descriptive genres.

The genre dependency of the present and past tense forms makes a meaningful statement about their possible markedness relation and their frequency a rather hopeless enterprise. The same difficulty, as it turns out, is not limited to the present/past opposition but encompasses all grammatical categories, including other tense and aspect forms, such as the perfect and the progressive. The chi-square statistical test returns a highly significant value when calculated for the distribution of these forms over the fifteen genres of the corpus. The chi-square for the perfect is 634.00, for the progressive aspect 806.70, for the simple present 840.35 and for the simple past 12391.56. All these figures are highly significant, even at the 1% level of significance (at P = 0:01 for 14 degrees of freedom, the critical value of chi-square = 29.1). Consequently, the null hypothesis that the uneven distribution of grammatical forms among the genres of the corpus is due to chance has to be rejected.

The apparent impossibility of determining a stable frequency of such grammatical forms as tense and aspect in English dooms the attempts to find a correlation between markedness and frequency, at least in these cases. The question arises, however, whether such problems bear only on this statistical correlation or whether the entire difficulty is due to more fundamental causes.

There are essentially two basic difficulties that the markedness hypothesis faces in its analysis of grammatical categories: the first is that it needs to assume, by virtue of the theory, the possibility of identifying some invariant element in the meaning of the grammatical form, similar to the invariance notion that prevails in structuralist phonology. This, in essence, amounts to the advocacy of isomorphy between the set of grammatical forms and the set of corresponding meanings (often quite abstractly perceived), a claim which, in many instances, is realistically untenable. If we look at the English tense system, the problem becomes quite pronounced. In English, the function of the individual tense forms depends crucially on the lexical character of the verb or the structure of the verb phrase. Taking Vendler's influential classification of English verbal constructions as our point of departure (Vendler 1957), it can be shown that the simple present exists, in English, in natural usage only for states. *Jack loves Mary*, *Peter hates to shave*, *Charles knows German*, *The Charles Bridge spans the Vltava river*, are all descriptions of states. So are simple present tenses of verbs whose basic semantic function is to denote activities: such a verb as speak, for example, requires the progressive for a present activity reading: *Jane is speaking French*; its simple present can denote only a habit or an attribute (forms of a state): *Jane speaks French*, i.e., has the skill of speaking the language. This is also the reason, of course, for the clear activity-state difference between such sentences as *Jack is smoking* and *Jack smokes*.

Verbs and verb phrases that Vendler labeled accomplishments and achievements have no natural use of the simple present at all. Accomplishments are telic verbs which require a process to reach the intended goal, while achievements denote an instantaneous "leap" into a new state. Both, in their simple English past tense forms, correspond to the Slavic perfectives, denoting complete events: *Peter built a house*, *Jane bought a new coat*, *Čapek wrote many novels* (accomplishments), *The tyrant died*, *Jack spotted the plane*, *His wife found the key* (achievements). Notice that neither of these two classes of verbs has a simple present that would have the capacity of denoting an activity overlapping with the moment of speech: *Peter builds a house*, *Jane buys a new coat*, *The tyrant dies*, *His wife finds her key*, are possible only in reportive style as descriptions of complete past events. They can be used in a newspaper headline, a chapter heading, or "historical present" constructions but not as descriptions of ongoing activities. Note also, however, that some of the same verbs with a plural object again denote states: *Peter builds houses* designates Peter as a house builder.

Even such a small set of examples, which could be greatly extended to other English verbal categories, such as the progressive, clearly demonstrates that the notion of form--meaning isomorphy, which seems essential to markedness analysis of these forms, is not tenable. The large differences in frequency that we have noticed in the occurrences of these tense forms in the Brown Corpus are thus not only "stylistic" in a superficial sense, but stylo-semantic: they occur because of the differing needs for the expression of states, activities or events, each of which requires some verbal tense forms and blocks others.

The logical relation between marked and unmarked forms, predicted by the markedness theory, presents another difficulty for the analysis. As I have already mentioned, the definition of markedness on the lexical and the grammatical level is not logically the same as the so-called privative opposition in phonology. Rather, as Lyons (1977:305) has pointed out, the concept of markedness in the lexicon and grammar is a special case of the relation of hyponymy. The term hyponymy can be used as a more suitable designation for what, in logic, has been often discussed in terms of class inclusion. The hyponymy relation can be best illustrated on examples involving the relation of simple lexical items: so the word *rose* is a hyponym of *flower*, for example, with *flower* being the superordinate term of the relation. Hyponymy is definable in terms of unilateral implication. So, for example, the verb *waltz* can be established to be a hyponym of *dance* by virtue of the implication: *She waltzed all night* → *She danced all night* (but, of course, not the converse).

As Lyons also suggests, the Praguian markedness relation is, essentially, a special case of hyponymy. The principal difference is that the unmarked term has two meanings, the general (which gives it the usual status of a superordinate term) and the narrow or nuclear, which has a more specific sense, depending on context, and puts it in opposition to the marked term. Lyons suggests that the markedness relation may differ from the simple hyponymy relation by its potential of being reflexive: *Is that dog a dog or a bitch?* is meaningful, though rather odd (Lyons, 1977:308). The logical basis of the morphological markedness relation thus clearly requires that the unmarked term must have the potential of expressing the "general" meaning. There must therefore be at least some contexts in which the unilateral implication, required by the theory, holds. However, with reference to the English tense system, the implications from one member of a tense opposition to another again vary in dependence on the lexical character of the verb. Markedness theorists have proposed, for example, that the contrast between the progressive and non-progressive verbal forms in English be viewed as constituting a marked vs. unmarked opposition (cf., for example, Kopečný 1948). Under this analysis, entailments from the "marked" progressives to the "unmarked" simple forms should hold. In fact, however, we do get entailments from the past progressive to the simple past (or the present perfect) for activity verbs, but not for accomplishment and achievement verbs: *Jack was walking in the park* → *Jack walked (has walked) in the park*, but *Jack was walking to the park* does not entail *Jack walked (has walked) to the park*, and *Martha was writing a dissertation* does not entail *Martha wrote (has written) a dissertation*. Thus we see again that a markedness analysis which tries to relate the simple and the progressive forms of the English system in a single relation of potentially reflexive hyponymy is bound to fail.

In conclusion, let me mention an example which illustrates how the frequency data may offer useful clues for an evaluation of theoretical constructs. Czech -- like other Slavic languages -- has two sets of verbs of motion, the so-called *determinate*, e.g. *jít* 'to go (on foot)', *jet* 'to go (by vehicle)', and the *indeterminate*, e.g. *chodit* 'to go (on foot)', *jezdit* 'to go (by vehicle)', etc. In the conventional markedness analysis, the determinate verbs are considered marked (for integrity or directionality of the action), the indeterminate unmarked. Language statistics.indicate, however, that every single one of the determinate verbs is more frequent than the corresponding indeterminate. Overall, the allegedly marked verbs are 4.38 times as frequent as the supposedly unmarked ones. (Cf. Jelínek et al., 1961). A closer analysis reveals that this ratio, which of course is the exact reverse of what one would expect from the markedness relation, ceases to be a mystery if one considers the semantics of these verbs more carefully (cf. Kučera 1980). While I cannot give the complex detail here, let me simply point out that the entailment relations, which represent the essence of the logical relations of markedness, never hold between these verbs. Sentences with the "marked" verbs do not entail the sentences with the "unmarked" ones under any circumstances: *Bratr jde po ulici* 'Brother is walking down the street' does not entail *Bratr chodí po ulici* 'Brother is walking up and down the street'; Sestra jela do Prahy 'Sister went (drove) to Prague' does not entail *Sestra jezdila do Prahy* 'Sister used to go (drive) to Prague'. The essence of the problem, as far as all the verbs of motion are concerned, lies in the fact that the relation between the determinate and indeterminate sets is not one of markedness, i.e., of logical inclusion, but rather one of logical exclusion. The determinate verbs denote an activity in a single direction, the indeterminate a complex action.

The evidence provided by computational and statistical techniques, while not by itself decisive, suggests the weak points of the application of the markedness framework to the analysis of grammatical categories, such as tense and aspect. A semantic reanalysis of the categories themselves then indicates that the relations of these sets of forms is more complex than the invariance-oriented markedness hypothesis is able to accommodate and that the vagaries of the performance data are the net result of such complexities.

REFERENCES:

[1] Francis, W.N. and Kučera, H., Frequency Analysis of English Usage: Lexicon and Grammar (Houghton Mifflin Co., Boston, 1982).

[2] Greenberg, J.H., Language universals, in: Sebeok, T.A. (ed.)., Current Trends in Linguistics 3 (Mouton, The Hague, 1966).

[3] Jakobson, R., Zur Struktur des russischen Verbums, in: Charisteria Gvilelmo Mathesio (Prague, 1932).

[4] Jakobson, R., Shifters, Verbal Categories and the Russian Verb, in: Selected Writings 2 (Mouton The Hague, 1971).

[5] Jelínek, J., Bečka, J.V., and Těšitelova, M., Frekvence slov, slovnich druhu a tvarů v českém jazyce (Státní pedagogické nakladatelství, Prague, 1961).

[6] Josselson, H.H., The Russian Word Count (Wayne University Press, Detroit, 1953).

[7] Kopecný, F., Dva příspěvky k vidu a času v češtině, Slovo a slovesnost, 10 (1948) 151-158.

[8] Kučera, H., Markedness in Motion, in: Chvany, C.V. and Brecht, R.D. (eds.), Morphosyntax in Slavic (Slavica Publishers, Columbus, 1980).

[9] Kučera, H. and Monroe, G.K., A Comparative Quantitative Phonology of Russian, Czech, and German (American Elsevier, New York, 1968).

[10] Lyons, J., Semantics 1 (Cambridge University Press, Cambridge, 1977).

[11] Trubetzkoy, N.S., Grundzüge der Phonologie. Travaux du Cercle Linguistique de Prague 7 (Prague, 1939).

[12] Vendler, Z., Linguistics in Philosophy (Cornell University Press, Ithaca, 1967).