PREFACE

The papers in this volume were read at the Conference on Quantitative and Computational Linguistics arranged by The Linguistic Association of Finland at Seili, August 29-30, 1979. A simple format was chosen in order to avoid delays in the publication of the papers.

Sincere thanks are due to the Council of Humanities (Valtion humanistinen toimikunta) for a grant that facilitated publication of the volume.

Copies of the volume can be ordered from Suomen Kielitieteellinen Yhdistys, c/o Kaisa Hääkkinen, Fennicum, Turun yliopisto, SF-20500 TURKU 50, Finland.

Kaisa Hääkkinen  Fred Karlsson

ISSN 0357-0622

Turun yliopiston offsetpaine 1979
Jouko Lindstedt

SOME BASIC PROBLEMS OF QUANTITATIVE MORPHOLOGICAL TYPOLOGY

0. Morphology has always occupied a central position in linguistic typology. Consequently, in the domain of morphology the abuse of typological statements and an erroneous conceptualization of linguistic types as absolute classes have manifested themselves most plainly. A new, relative, and multidimensional way of conceiving morphological typology was first put forward by Edward Sapir, whose ideas were the necessary basis for Joseph H. Greenberg's (1960) later quantitative model, consisting of special morphological indices calculated for samples of text.

Greenberg's original model has been improved in some respects, particularly as concerns the mathematics. The indices have been revised so that they all share the same interval of variation from 0 to 1 (Krupa 1965), and their mutual relations have been subjected to correlation and regression analysis, which, in turn, has led to new ideas of the concepts "type" and "typological class" (Krupa—Altmann 1966; Altmann—Lehfeldt 1973: 39-48). There have also been some thoughts concerning the relationship between quantitative typology and the modern concept of the fuzzy set (Pierce 1977).

In spite of these developments Greenberg's typology cannot yet be considered adequate. So far the indices have only been calculated for some thirty languages, and the samples used have not been very extensive. There is still something rather tentative in the very foundations of the whole system, so that its user can never feel quite safe. Altmann and Leh—
feldt (39) do not want to call it "typology" without quotation marks, but their scepticism is the result of their programmatic attempt to reshape all typological thinking. Irrespective of whether such attempts are necessary (and what their direction should be), quantitative morphological typology now requires clarification of the possible sources of its inaccuracies—on its own terms.

1. Typological statements are traditionally statements about the whole (subsystem of a) language; quantitative indices are, however, calculated on the basis of text samples. What is the relation of the text to the language? It is of course the complex old *langue*/*parole* problem that we encounter in this question: as *la langue* is a social norm, and *la parole* is a product of specific human actions, they do not have the same ontological status and, consequently, we cannot identify language with an infinitely large utterance population the parameters of which can be estimated by means of sampling. On the other hand, sociolinguistics has shown that when a rule of a language is not absolute—and typological "rules" seldom are—the only way to reveal new regularities is to study speech. It is necessarily true that a text represents the language *somehow*: we never base typological research on texts which contain violations of the grammatical rules of the language in question. Hence, the sample text has already passed through the filter of language before reaching our analysis.

So, a text can tell us something about the language behind it; but the exact nature of this something remains a little unclear. According to Altmann and Lehfeldt (64), "[es ist] angemessener, statt mit dem Sprachsystem mit Texten zu arbeiten, da der Text eine natürliche Gewichtung der Bestandteile des Systems darstellt." But in what sense "natürliche"? We can assume that when the random and systematic (i.e., stylistic) variations depending on the choice of particular texts have been eliminated, the estimated parameters of the whole hypothetical infinite utterance population are direct reflections of certain properties of the grammatical rules; they are thus...
(theoretical) properties of the language in question, although they are not among the (social) rules that primarily constitute it.

As Charles Muller (1968:91) has pointed out, this all means that the transition from text to language is always possible irrespective of the length of the sample. A sample that cannot be representative of the whole language is, in any case, a product of a certain style, a certain idiolect—"un certain état de langue". Therefore, large samples are not better approximations of language than shorter ones; they are simply more representative. Quantitative typology is thus perfectly reconcilable with the notion of language and speech as different kinds of entities.

2. To make a qualitative typology quantitative, simple transition from grammar to text is often not sufficient, since the parameters also require a great deal of explication and readjustment in order to become measurable. A distinction involving several characteristics has to be simplified or divided into separate parameters.

The traditional isolative/inflective/agglutinative division, the essence of morphological typology, is a good example of quantification problems. Theoretically it is easy to describe this trichotomy by means of the two unidimensional distinctions agglutination/fusion and synthesis/analysis, but for different linguists "analytic construction", for instance, does not mean the same thing. According to Sapir and Greenberg, it refers to the low morphological complexity of words, so that a language with a low M/W (morpheme-to-word) ratio—and, consequently, with a high value of W/M—would be called "analytical". But in the European linguistic tradition "analyticism" means use of prepositions and other auxiliary words (instead of case endings) to connect the words in a sentence; i.e., the European notion of analysis is morphosyntactic, not narrowly morphological as is Greenberg's.

It is important to note that these two criteria of analysis may give totally contradictory results. During all its history, Bulgarian has manifested a strong tendency towards
increasing analysis, in the traditional sense, and yet the M/W index clearly shows that modern Bulgarian is a more "synthetic" language than was Old Bulgarian (i.e., Old Church Slavonic; see Lindstedt, forthcoming). To some extent this problem is, however, merely terminological since another Greenbergian index, that of isolation (O/N), closely corresponds to the morphosyntactic concept of analysis, being also fit for the measurement of the Bulgarian development (op. cit.). The discrepancy between W/M and O/N is probably typical of many stem-isolating (as opposed to root-isolating) languages.

As for the agglutination/fusion dimension, the problems are here even greater. For Greenberg (1960:185), "what is involved is the degree of morphophonemic alternation. --- If both morphs in a construction belong to morphemes which are automatic, the construction is called agglutinative." This definition leads to intuitively unsatisfactory results. First, it turns our attention from the properties of the actual construction (in praesentia) to the paradigmatic relations (in absentia): the English plural cows would not be agglutinative, since the plural morpheme has non-automatic allomorphs in some words, such as oxen. Hence, the ratio of agglutinative constructions per morpheme juncture (A/J) for English would be as low as 0.30 (Greenberg 1960:193)! Second, Greenberg's definition attributes to the juncture a characteristic that is defined as a property of the morphemes on its two sides: but why should ox-en not be an agglutinating construction, with a clear-cut morpheme boundary?

Mikko Korhonen (1969:207-208) has pointed out that A/J cannot tackle the problem of "symbolism", such as in goose/geese—Greenberg (:187) has to resort here to "infixes". Of course, his index system, first published in 1954, was necessarily saturated with the mechanistic linguistic theory prevailing at that time. Agglutination is traditionally associated with ease of segmentation, but A/J can only be calculated when the segmentation has already been carried out because the number of junctures should be known.
Since agglutination can also be defined as a one-to-one relation between the morphs and morphemes in a syntagm, it might be worth trying to replace A/J by a morph-to-morpheme index, whose highest value would be 1, indicating a perfectly agglutinative language (cf. Lyons 1968:188-191 and Wells 1978:29). Though in appearance similar to Greenberg's original indices, this new index would presuppose a functional definition of the morpheme as the minimal unit of grammatical analysis (i.e., "moneme"), not as a class of morphs as in the distributionalist tradition.

3. Besides the more theoretical questions dealt with above, there are some crucial problems related to the reliability of the calculation procedure of the quantitative indices. These problems are basically of two kinds—statistical and operational.

The need for operational definitions, used for the identification and segmentation of linguistic elements, arises from the requirements of objectivity: a calculation procedure must be defined with such a precision that different linguists will obtain identical or at least similar results by using it (Altmann—Lehfeldt 1973:52). Several concepts used in Greenberg's typology are too vague in this respect. Cowgill 1966 and Korhonen 1969 contain useful discussion about the problems of positing zero morphemes, distinguishing between clitics and affixes, segmenting inflexional endings into one or more morphemes and deciding whether or not a given morpheme is automatic. Yet, it may turn out that these kinds of questions are not crucial for some at least theoretically well-defined indices, such as M/W. I have calculated M/W for the Old Church Slavonic text analysed by Cowgill. I found 223 morphemes where he had found 229, a fact which points to a high degree of unambiguity in the definition of this index. However, had Cowgill not explicitly stated his way of treating some problematic points, the difference between our results might well have been far greater.

As far as I know, nobody has noticed the considerably greater
ambiguity in the isolation index O/N. In its formula, N is
the number of nexuses, a nexus being defined as an "instance
of the use of a principle to indicate relations between words
in the sentence" (Greenberg 1960: 187). The principles of
indicating relations between words are concord, pure in-
flexion and order; O is the number of the "order" cases,
which are defined as "absence of an inflectional morpheme in
a word" (ibid.). Moreover, it is explicitly stated that a
morpheme such as the -um of the Latin masculine accusative
singular of adjectives has to be counted "a number of times,
one for each distinct feature" (in the case of -um, for two
concordial and one inflexional feature). For Greenberg, a
nexus is thus a means of connection and there may be more
than one nexus in a given connection between two words.

If we now turn to the important calculations of Warren Cow-
gill (1966), we cannot help suspecting that he has misunder-
stood the whole notion. For him, a nexus seems to be a
relation between two words (:124), i.e. the connection it-
self, not the means of expressing it. In consequence, the
instances of simultaneous concord and pure inflexion (pos-
sibly with multiple concord, as in -um) now count as only
one nexus each and the O/N index is bound to grow higher as
the total number of nexuses decreases. All this of course
makes comparison of Coghill's indices with those calculated
by Greenberg meaningless, though we may consider Coghill's
calculation procedure as such more appropriate for typo-
logical purposes.

For evidence, I recalculated the O/N index, as well as its
complement indices of pure inflexion and concord, for Cog-
gill's Old Church Slavonic sample, defining "nexus" as a
"connection" in the sense used in dependence grammar. The
results were practically identical:

<table>
<thead>
<tr>
<th></th>
<th>O/N</th>
<th>Pi/N</th>
<th>Co/N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coghill:</td>
<td>.41</td>
<td>.33</td>
<td>.26</td>
</tr>
<tr>
<td>recalculation:</td>
<td>.42</td>
<td>.33</td>
<td>.25</td>
</tr>
</tbody>
</table>

It goes without saying that application of Greenberg's
original criteria would have given a significantly lower
value of O/N. The following comparison of some indices calculated by Greenberg and Cowgill is also most revealing:

<table>
<thead>
<tr>
<th>Language</th>
<th>O/N</th>
<th>I/W</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classical Sanskrit (Greenberg)</td>
<td>.16</td>
<td>.84</td>
</tr>
<tr>
<td>Vedic Sanskrit (Cowgill)</td>
<td>.26</td>
<td>.97</td>
</tr>
<tr>
<td>Old English (Greenberg)</td>
<td>.15</td>
<td>.90</td>
</tr>
<tr>
<td>Gothic (Cowgill)</td>
<td>.37</td>
<td>.98</td>
</tr>
</tbody>
</table>

It seems improbable for a language having more inflexional morphemes (I) per word to have a higher index of isolation as well. According to Cowgill (:132), "the index of Isolation --- is directly proportional to the number of uninflected words in the text," a statement which, taken literally, is an additional proof of his unconscious redefinition of the concept "nexus".

4. Although some users of the Greenbergian typology are only too aware of the need for statistical evaluation of their results, as a rule they totally neglect even the most elementary statistics. The results of the calculations are presented in terms of mean values and no attention is paid to the other parameters of the distribution of the index values. But such a value of M/W as, for instance, 2.17 means only that in a certain sample of 100 words there were 217 morpheme segments, so that we would certainly not be surprised to find 216 or 225 segments in some other sample of the same language or even the same idiolect.

The statistical reliability of index values is dealt with by Pierce 1966 (for phonological indices, see Pierce 1962) and, most notably, by Korhonen 1969 (:213-216). The former article shows the necessity of drawing samples from different styles so as to characterize the whole language under study. Expanding each separate sample from 100 to 200 words is not considered by Pierce to increase the statistical reliability—but one might ask what a 20,000-word sample would do. What can further be criticized is his apparent assumption that the index values are always subject to the normal distribution. At least M/W is not likely to be so since it is its inverse.
number $W/M$ that is likely to approximate the normal distribution through the binominal distribution. The $\chi^2$-test might be of some value here.

The central problem is the relation between the size of the sample and the accuracy of the estimations based upon the sample. Korhonen (216) makes it a necessary condition that two digits on the right of the decimal point should be significant. The required size of the sample can only be calculated for those indices whose distribution is known; assuming that $A/J$ is subject to the binominal distribution, Korhonen arrives at the result that, in the most unfavourable case ($A/J=0.50$), as many as 10,000 junctures are needed.

It is easy to show that if statistical methods had been applied, certain conclusions in current literature, based on the comparison of the index values of different languages, would have been formulated with more caution. Cowgill (125), for instance, points out that in several Indo-European branches $A/J$ has been increasing up to our days. He is of course right, on the whole; yet, he is unable to establish where a given difference in index values is really significant. Using his figures for illustration, may we consider New Testament Greek ($A/J=0.12$) to be more agglutinative than Homeric Greek ($A/J=0.10$)? The difference between the two values is 0.02 and it has the standard deviation of

$$\sigma = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}} = 0.040,$$

where $p_1$ and $p_2$ are the two values of $A/J$ and $n_1$ and $n_2$ are the respective sizes of the samples, i.e. the number of junctures ($= J = 100 \cdot M/W - 100$ in a 100-word sample). Since the difference 0.02 is only 0.5 times the standard deviation, with the probability of 62 per cent it is an error to assert that this difference is not caused by mere chance. But if we compare Old English (0.11) and New English (0.30), we find that this difference is significant: the difference (=0.19) is 3.02 times its standard deviation (=0.063) and, consequently, the tables show only a 0.3% risk of error.
We are still left with the problem of how to eliminate stylistic variation within different languages. It would hardly be feasible to choose translations of the same text, as suggested by Altmann and Lehfeldt (:64). It is an additional problem that an integrate sample text is not a random sample, because only the first word can be chosen randomly. Perhaps we should develop the indices according to the principles put forward, albeit rather implicitly, in Pierce 1962 and 1966: instead of a random variable $X$ whose mean value $\bar{X}$ is calculated on the basis of a sample of $n$ words, we take a new variable $\bar{X}$ whose values are the $\bar{X}$'s of different samples; for each language, we choose a random sample of $N$ samples of $n$ words and consider $\bar{X}$ as the best estimation of the value of the index in question. Graphically:

$$
\begin{align*}
&\text{n words} \\
&\downarrow \\
&\bar{X} \\
&\downarrow \\
&\bar{X}
\end{align*}
$$

The advantages of this kind of conversion are as follows:

i. $\bar{X}$ can account for various styles of the language;

ii. the $N$ texts can be chosen more randomly than the $n$ words;

iii. $\bar{X}$ has a more continuous distribution than $x$, especially in the indices where the latter can only have the values 0 and 1, as in $A/J$ and $O/N$;

iv. the standard deviation of $\bar{X}$ is $\sqrt{n-1}$ times smaller than that of $x$;

v. the standard deviation of $\bar{X}$ is $\sqrt{N-1}$ times smaller than that of $\bar{X}$. However, the confidence intervals are not restricted proportionally unless $N$ is big enough (above thirty).
5. In short, the difficulties connected with quantitative morphological typology are serious and largely unresolved, partly because so few researchers are aware of them. On the other hand, these difficulties are often dependent upon the problems of grammatical description in general, so that they cannot be used to show any fundamental defects in Greenberg’s original conception. Duly amended, the indices will certainly find interesting applications, but before that considerable efforts are needed to improve the theoretical and operational definitions and to develop appropriate statistical routines.

Notes:
1. That the corpus is in this sense secondary to the grammar has often been emphasized by Esa Itkonen (see, e.g., Itkonen 1978).
2. Actually, I only found 99 words in his 100-word sample. The obvious source of this difference is the word nėsme 'I am not', whose constituents are clearly visible (cf. nė 'not', jesme 'I am') so that Cowgill has evidently counted it as two words.
3. I would like to thank Kimmo Koskenniemi for making this point clear to me.
4. Or n junctures, n nexuses, n morphemes, depending on the denominator of each index.

Bibliography:


Greenberg, Joseph H. 1960: "A quantitative approach to the

Itkonen, Esa 1978: *Grammatical Theory and Metascience.* (= Current Issues in Linguistic Theory, 5.) Amsterdam: John Benjamin.


Lindstedt, Jouko (forthcoming): "Edin cjelosten podchod kām analitizma na bǎlgarskiija ezik." (To appear in the proceedings of the First Bulgaro-Scandinavian Symposium on the Bulgarian Language and Literature.)


