NICHOLAS D. ANDREYEV

ALGORITHMISATION OF LINGUISTIC RESEARCH USING THE STRUCTURAL-PROBABILISTIC PROPERTIES OF LANGUAGE UNITS

First of all, I must confess that the algorithmisation in question is only a by-product of the structural-probabilistic analysis, whose main and openly immodest purpose is to re-write anew the whole book of language theory, and especially of its development.

Nevertheless, the algorithmisation of linguistic research, both partial and in full, is feasible, and has already been done in dozens of experiments during the last 15 years. This algorithmisation is essentially based upon the assumption that quantitative characteristics are an inherent part of the language structure.

In order to prove this hypothesis, let us consider the so-called "intuitive statistics": first, its facts, then their theoretical interpretation.

TABLE 1.

ХЛОП ТОП, ОХРИП СТОЛП, ...
ПРИЁМ, ПОЙТИ, ПЫЛЬ, ПЕРЕХОД, ...

If you ask a Russian man in the street, what series of words would be easier to prolong, the upper or the lower one, nine times out of ten (or, better to say, 99 times out of 100), he gives a quick and correct answer.

And how to interprete this well-verified and therefore undeniable fact? Does this knowledge, belonging to the man in the street, constitute a part of the language structure inside his brain, or is it something beyond that structure?

TABLE 2.
I. *PALEONTOLOGY, SUN, SOUP, ...*
II. *EVENT(S), CLOUD(S), VEGETABLE(S),....*

Now let us turn to such a simple case as plural form of nouns. Of course, it is possible to say: *the planet of two suns, these two soups are incompatible, the distinguished professors do have different paleontologies,* but all these cases are rarities of the kind, and we know the fact *without* any count.

On the other hand, the plurals of the second group are quite common, and occur as often as (or even more frequently) than their respective singulars. Consequently, every native speaker of English *feels* the difference between class I and class II, once more *not* having counted anything at all.

TABLE 3.
1. *NEVER*
2. *RARELY*
3. *NOT CLEAR*
4. *FREQUENTLY*
5. *ALWAYS*

Psycholinguistic experiments prove that a man, sufficiently mature, has a probabilistic ladder in his mind, on the five steps of which not only the life situations, but also language phenomena are disposed.

TABLE 4.
AVERAGE: Usually about 30 % of the whole set of noun occurences are plural.
A. Singularia tantum B. Predominantly singular class (1-10 %) C. Average class (20 %-40 %) D. Predominantly plural class (50 %-90 %) E. Pluralia tantum

Returning to the singulars and plurals, we may observe that there exists a probabilistic scale of nouns which divides all of them into five groups, according to their respective proportions between plural and singular forms.

Thus we have come closely to the conceptions of categorial measure and its classifying role. The figures in the table 4 express the categorial measure of plurality; of course it could be done in terms of singularity.

The short time, which is at my disposal, does not permit me to supply you with strict mathematical definitions, the latter being too complicated to be understandable at first glance.

TABLE 5.

Any strong governing is a case of high categorial measure (*CM*), e.g.:

Verb	*CM*	*Class*
LOVE	0.98	4. (High transitivity)
KNOW	0.51	3. (Middle t.)
WORK	0.02	2. (Low t.)
LAUGH	0	1. (Intransitivity)

Average factual occurrence of a direct object is 35% of factual verb occurrences in English texts.

Traditional structural linguistics has already elaborated a conception which has some kinship to the notion of categorial measure; I mean the so-called " strong governing ". But the limits of the phenomenon have never been established, - at least, unanimously.

Of course, the boundaries between classes 4,3,2,1, are not thin lines, they have some probabilistic thickness, but such is the very nature of the system of classes under consideration here.

TABLE 6.

The Correlational Functional (*CF*).

Verb	*CM*	*CF*	*Class*
AVERAGE	0.35	$\frac{0.35}{0.35} = 1$	—
LOVE	0.98	$\frac{0.98}{0.35} = 2.8$	4
KNOW	0.51	1.5	3
WORK	0.02	0.1	2
LAUGH	0	0	1

These four classes are a clear case of a probabilistic distinctive feature (*PDF*). Another one is represented by the five classes based on the *CM* of plurality/singularity.

Here we proceed to a more sophisticated notion of the correlational function which is defined as a ratio of the individual categorial measure of a processed linguistic unit to the average categorial measure.

When the correlational function is equal or near to 1, we have a typical representative of the set, whose properties are standard (or quasi-standard) from the considered point of view.

If the *CF* substantially exceeds 1 (the boundary of substantiality is determined by factor analysis), then we have met a representative of an upper class (whatever its name).

Last, if the *CF* is substantially less than 1, then before us there is an item from a lower class. Sometimes, it is very important to make a distinction between the lower class and the *zero* class, as well as between the higher class and the *absolute* class. This particular situation may be found in the case of plurality/singularity.

The categorial measure being the basis for the classification and for establishing oppositions, we come to the notion of probabilistic distinctive feature, which may be defined (without mathematical details and therefore not too strictly) as a feature generating quantitative differences,

and the latter ones, when strong and stable enough, may be defined as creating qualitative oppositions at the level of mental perception.

TABLE 7.

Interaction between two *PDF*'s: between transitivity and commentability.

I know him -	Transitivity
I know that he wouldn't do it -	Commentability
I hope that he (&) -	both times

Verb	*CF* (Tr-ty)	*CF* (Com-ty)	Σ*CF*
LOVE	2.8±0.6	0	[2.8]
KNOW	1.5±0.5	3.1±1.1	[4.6]
HOPE	0	5.5±1.6	[5.5]

A case of probabilistic complementary distribution (*PCD*).

Probabilistic distinctive features are not independent of each other. Of course, it is not officially forbidden to say:

I know him, and his habits, and his wish to do it exactly as usual, and that he never would do it another way,

but such examples are very rare. For all practical purposes, we may consider transitivity and commentabilitity as nearly absolutely excluding each other. By the way, it is wrong to interpret the latter as a particular case of the former; English *to hope*, French *espérer*, German *hoffen*, and Russian НАДЕЯТЬСЯ, - they all are intransitive, and at the same time commentable, which proves the point.

So, when the possibility of simultaneous realisation of two (or more) *PDF*'s is near to zero (or equal to), then we have the right to speak about a probabilistic complementary distribution.

TABLE 8.

A heuristic task: to investigate the (English) verb and to discover *PCD*'s, as many cases as possible.

A Routine For Its Solution

1. Fix all the factual syntactic links of verbs in a given corpus of texts.
2. Find the average *CM* for each type of link.
3. Sort out a prescribed quantity of the most frequent verbs in each type of link.
4. Measure the *CF*'s, i.e. the individual values of them for each verb sorted out, thus establishing the *PDF*'s.
5. Correlate the *PDF* values considering them as components of a multi-dimensional vector.
6. Dine well if you have found a new case of a *PCD*: it's rarer than a wife who neglects a new vogue more often than once a year.

It may be easily seen, that all the procedures included in the set of routines admit algorithmisation and subsequent computerisation (except number 6, for obvious reasons).