*COLING 82, J. Horecký (ed.)*
*North-Holland Publishing Company*


# COMPUTATIONAL DATA ANALYSIS FOR SYNTAX

Ludmila Uhlířová - Iva Nebeská - Jan Králík
Czech Language Institute
Czechoslovak Academy of Sciences
Prague
Č.S.S.R.

Methodology and results of complex computational analysis of present-day standard Czech are presented. According to computer programmes various linguistic observations were achieved, concerning especially dependency syntax.

## INTRODUCTION

The aim of this paper is to present the methodology and results of a detailed analysis of syntactic structures in Czech, performed with the aid of computer. This work is a part of a large and long-term project, known as Quantitative analysis of present-day standard Czech, which is carried out in the Department of mathematical linguistics at the Czech language Institute under the leadership of M. Těšitelová [1], [2]. Some hitherto achieved results have already been published in special monographs /e.g. the frequency dictionaries of publicist and administrative styles [3], [4], a volume of quantitative characteristics of publicist style [5]- these prints were issued by the Czech language Institute, Prague/ or as articles and papers e.g. in PBML 31 and 32 [6] and PSML 7 and 8 [7]; other articles are to be published within a short time /as e.g. the frequency dictionary of scientific Czech/ or they have been already prepared for publication.

## PROPOSITIONS AND SYNTACTIC CODE

The target of the syntactic analysis as well as of the whole project is to obtain a coherent set of mutually related quantitative parameters concerning the Czech language system, its functioning in different communicative spheres, as well as its stylistic values. To guarantee the reliability and representativeness of re-

sults, the research is based on the corpus of 540 000 word-forms /occurrences/ chosen from non-narrative /i.e. newspaper, administrative and scientific/ texts. During the preparatory works all word-forms in 180 samples - each sample consisting of 3000 running word-forms - were supplied with a special code carrying both morphological and syntactic information about parts of speech, all main morphological categories, relevant in Czech /such as case, nominal gender and number, person, number, tense, mood, voice of verbs etc. with a very detailed subcategorization/ and all main syntactic categories /such as subject, object, predicate, attribute, types of adverbials etc., again with the necessary subcategories/. The syntactic code was basically digital, with three exceptions: the characters "+" and "-" expressed whether the dependent sentence element follows or precedes the governing word /sentence member/, and the character "!" was used for special defective sentence constructions. For purposes of frequency lists, each word-form was given a basic lexical information /lemma/.

The coding and lemmatization was done by linguists in such a way, that after marking boundaries between sentences and clauses each word-form was given information about its syntactic function /"membership in sentence"/ and about its governing word in the dependency tree. Each clause was then given information about its structural position in the complex or compound sentence; then coordination between words or between clauses was marked; finally, information about the linear arrangement of words in sentences, of clauses in complex and compound sentences, and of sentence wholes in text was added, so as to enable the complete reconstruction of running texts by computer, if necessary, at any time.

The whole corpus together with the encoded information came by means of punched cards /80 columns/ over a current input programme into external computer memory on magnetic tapes /each tape library can carry max. 30 texts, the whole corpus is contained in 7 tapes/. Each record got a special translation zone to guarantee the possibility of obligatory /non-standard/ sorting with respect to different alphabetic ordering of some Czech letters /record size on tape is 130, block size BS=6502/. The automatic processing has been executed at the computer TESLA 200 in the computer-centers ÚTZCHT and ÚFPL, Czechoslovak Academy of Sciences, all programmes and

their modifications were written in internal programming language "APS".
The survey of the main results, given below, is ordered from the most simple to the more complicated ones, with respect to the computer programmes and with respect to the linguistic information obtained.

## SIMPLE COMPUTATIONAL CHARACTERISTICS

The first set of programmes gives by its simple structure as to the programming technique the essential totals of frequencies /occurrences/ of encoded syntactic categories, individual syntactic features and various items under investigation. Number of results of this kind was obtained by means of reading or repeated reading of the magnetic tape library and through simple adding of different code items.

The received data offer us a basic survey about the frequencies of sentence elements and simple, complex and compound sentences, about types of syntagms /both determinative and coordinative/, about the frequencies of two-element and one-element sentences and their patterns, about the frequencies of types of subordinate clauses, some word-order and clause-order characteristics and the frequency ratio of simple and complex/compound sentences. In addition, there have been collected some data concerning how often various sentence elements are expressed by a nominal or an adverbial phrase and how often they are expressed by a dependent clause, concerning the most frequent types of complex sentences, including frequencies and functions of various syntactic connectors. Using the cycles within counting programmes the distributions of syntactic units were obtained in a similarly simple and prompt way. This part of the computer work yielded the length distributions of clauses and sentences expressed in number of words, or in number of clauses.

## COMPOUND COMPUTATIONAL CHARACTERISTICS

By doubling or chaining of testing subprogrammes and cycles another set of programmes was constructed for more complicated searching and output of syntactic characteristics. Specially commented tab-

les, as well as larger sets of numeric data supplying rich material for further steps of analysis were obtained. Whereas the results of the programme set mentioned above referred to frequencies of individual syntactic categories, the programmes reported about in this paragraph were concentrated on their relationships. Attention was paid especially to the relationships between syntactic elements and their part-of-speech appurtenance, to the syntactic relevance of some morphological categories /e.g. of noun cases/, to the correlation between sentence length and complexity of its structure, to the relationship between types of subordinate clauses and their linear position in complex sentences etc. Some of the statistical data obtained have confirmed our intuitive expectations /e.g. concerning syntactic functions of parts of speech and syntactic functions of cases of nouns/, others lead us to a deeper insight into interrelations between linguistic levels, esp. about the connection between the lexical and the syntactic levels.

## TYPES OF VERBAL CONTEXTS SEARCHED BY COMPUTER

Using the computer operation memory we overcame the technical impossibility of a reverse magnetic tape reading. This enabled us to prepare the third set of programmes with many variations. As an output we received whole sentences, sentence types or their components with required code combinations or with immediate verbal contexts. Thus we could study not only abstract syntactic categories as such, reported above, but we also could take into account the concrete lexical manifestations of various syntactic elements, units and categories.

Some interesting tendencies were found, concerning the insertion of certain lexical types into different syntactic positions, e.g. types of adjectives typical for predicative positions and other typical for attributive positions. The relationship between the semantics of the co-ordinated syntactic elements and their functions in topic-comment structure was studied, the correlation between the frequencies of subordinate clauses and lexical semantics of the governing predicate was proved /with predicates expressing the attitude of the speaker to the content of communication/, the correlation between morphological category of infinitive and semantic category of modality was found. A special attention was

paid to the syntactic structures with verbs to be and to have.

## MAIN RESULTS AND PERSPECTIVES

The hitherto made experiences have shown that even a very extensive statistico-linguistic project can be successfully carried through, if there is the aid of computer. Results obtained up to now offer a very detailed picture about the functional load of syntactic elements and units in texts from various styles and from various communicative spheres of the present-day Czech. However, the importance of the project, which has not yet been completely finished, consists also in the recognition and understanding of quantitative linguistic principles, relationships, tendencies and general laws.

Statistics of syntax contrasts by some of its features with the statistics of other linguistic levels. If compared with some hierarchically lower levels, such as with phonology or morphology, and their units, the sentence as the basic syntactic unit is structurally much more complex /not representing the mere summ of elements and forms of the lower levels/ and, larger, too. For these reasons it disposes with a considerably higher combination possibilities, and consequently of richer posibilities of individual usage of linguistic means of its creation and usage during the communicative process. On the other hand, if compared with concrete lexical items, most properties of sentence are of abstract, categorial nature; the inventory of sentence patterns is strictly limited in number, and therefore they are repeated very often in texts, which of course contributes to the neutralization of their stylistic value.

The computer aided quantitative analysis of syntax proved to be a valuable counterpart of qualitative structural research in describing and evaluating the functioning of language means in communication.

REFERENCES:

[1] Těšitelová, M., Otázky lexikální statistiky /Academia, Praha, 1974/.

[2] Těšitelová, M., Využití statistických metod v gramatice /Academia, Praha, 1980/.

[3] Frekvenční slovník současné české publicistiky, Těšitelová, M. /ed./, /Ústav pro jazyk český, Praha, 1980/.

[4] Frekvenční slovník současné administrativy, Těšitelová, M. /ed./, /Ústav pro jazyk český, Praha, 1980/.

[5] Kvantitativní charakteristiky současné publicistiky, Linguistica II, Těšitelová, M. /ed./, /Ústav pro jazyk český, Praha, 1982/.

[6] Prague Bulletin of Mathematical Linguistics vol.31 and 32 /Universita Karlova, Praha, 1979/.

[7] Prague Studies in Mathematical Linguistics vol.7 /Academia, Praha, 1981/ and vol.8 /Academia, Praha, in print/.