Adverbials in Czech: Models for their frequency distribution

Radek Čech, Ludmila Uhlířová

„Text is a virtual transformation of a set of words from lexical units to semantic elements“
(Hřebíček 2007: 74)

1. Introduction

Luděk Hřebíček, whose 80th birthday we celebrate this year, is an internationally respected author of general text theory, which he formulated and elaborated in the course of his scientific path through life. His text theory (cf. Hřebíček 1992, 1995, 1997, 2000, 2007), in which he brilliantly presented his original ideas and also significantly developed some discoveries made by Altmann in the eighties (partly as Altmann´s co-author and co-editor), is based on the idea that fundamental principles of text structures can be explained by the means of Menzerath-Altmann´s law (Altmann 1980) as its core, and with the numerous interrelations between two kinds of fundamental linguistic units, *constructs* and *constituents*. The present contribution, inspired by Hřebíček´s philosophy of language, deals with one concrete class of constituents and a class of their constructs.

Let us say that *words/lexical units* are constituents of constructs on the immediate higher level, i.e. of *phrase* or *clause*. If we omit a level and jump to sentence, the relation may become more complex. There are complex interrelations between them which can be defined and described in various terms. In the present contribution we deal with just one. We ask the following question: Do syntactic categories of a certain type, as will be specified below, show similar distributional features as other types of constituents and do they abide by the same statistical law(s)?

2. Development of syntactic analysis – from a description to an explication

As Köhler (2012) pointed out recently, the field of syntax remained – for a long time – less affected by quantitative methods than the lower levels of language. It was concentrated mainly on descriptive and applied aspects rather than on a theoretical (explanatory) analysis. Only today syntax became one of the fields which come more and more in the centre of interest of quantitative studies. The

interests of quantitative researchers are expanding rapidly now; effective and sophisticated mathematical methods and procedures of study are being developed.

There are many reasons why traditional syntactic statistics dealt mainly with the counting of particular syntactic phenomena, and not with general hypotheses and laws, which would "explain why languages are as they are" (Köhler 2012:7). The crucial ones should be sought in the very nature of syntax, namely in the manifold and complicated interrelations both between its constituents and constructs and between interrelations to other levels of language. Let us compare: In the Czech language, according to Ludvíková (1987), there are 36 speech sounds: 11 vowels and 26 consonants, and from the sum of all theoretically possible bigrams only about 60 per cent are attested, with different functional loads in the phonological system and with different frequencies in texts. The limited inventory of phonemes made it possible to perform phonological statistics already in the first half of the 20th century, such as those done by linguists of the pre-war Prague school generation, e.g. Trnka (1935), Mathesius (1947), Vachek (1940), and later on Ludvíková (1968) and others. On the other hand, there does not exist any "inventory", or any "list" of Czech sentences. What we have is a finite set of abstract sentence (or clause) patterns (schemes, types) and rules for their application. They are stored in our minds and described in grammars. However, due to our linguistic competence we are able to create (or "generate"), theoretically, an infinite number of utterances. This fact is considered to be one of the most important (and the most astonishing) aspects of human language. But it poses both theoretical and methodological problems. First, a huge variability of utterances leads to relatively big differences among particular syntactic descriptions, even within the same or similar methodological framework. Further, for an adequate statistical syntactic analysis, appropriate data must be prepared which is not a trivial task. Fortunately, because of the rapid development of computational linguistics in the last few decades, syntactically annotated corpora of Czech are available now (e.g., Králík, Uhlířová 2007; Vidová Hladká et al. 2008; Hajič et al. 2006).

To sum up, after a period of syntactic research that has struggled with many difficult problems (in comparison to phonology or morphology), it is time to take steps to a deeper understanding of syntactic functioning. So, we are trying to follow this research direction. Our analysis may be taken as an attempt to show how a descriptive analysis (Uhlířová 1975) can be reinterpreted in the light of current quantitative linguistic knowledge.

3. Specification of the task

In the following, we deal with one kind of sentence constituent – the adverbial. A plausible assumption to be tested, namely that the frequency distribution of adverbials abides by the same/similar law(s) to those that are already known for

syntactic and/or other constituents, can be based on the hitherto achieved experience: 1. There exist already a number of empirical frequency distributions as well as their probability models in the field of syntax; see Köhler (2012) for the present state of the art. 2. There exist typical frequency distributions of constituents of other levels of language which have been theoretically derived and attested on hundreds of languages; see Altmann (1980, 1988, 1993, 2001), Best (2005), Altmann, Köhler (1996), Grzybek (2006), Köhler (2005), Köhler, Altmann (2000), Köhler, Naumann (2009), Wimmer (2005), Wimmer, Altmann (1999, 2006) and many other authors. 3. No uniform frequency distributions of any language phenomenon/unit (not only in the field of syntax) have been found so far; it seems that nothing is uniformly distributed in language. 4. All frequency distributions commonly known are shifted in some way or other, but none display symmetry. 5. Generally: Human language is a phenomenon which abides by laws of a probabilistic nature – the idea which was once proclaimed by Zipf (1935, 1949), Mathesius (1947), Halliday (1993), and other great personalities in the course of the whole 20th century and which is being successfully formulated in a strict mathematical way now by Hřebíček, Altmann, Köhler and those mentioned already above. 6. The existence of underlying regularities in frequency distributions can be interpreted as a result of a diversification process (Altmann 2005). This process – alongside the opposite process of unification (cf. Zipf 1949) – has a decisive impact on language form and it follows from general principles which control human language behavior, such as the least effort principle (Zipf 1949) or a self-regulation in the synergetic model of language (Köhler 1986, 2005).

4. Syntactic framework and language material

We start from the "classical" dependency grammar formalism, according to which the finite verb is the centre of the sentence; the arguments headed by the finite verb are subject(s), object(s), or adverbial(s). We accept the definition and classification of adverbials given in the well-known grammar by Šmilauer (1966), inspired, basically, by the well-known views of Tesnière. Šmilauer´s syntactic framework was applied to the oldest Czech treebank, the Czech Academic Treebank corpus, compiled (and tagged) as early in the seventies by a team of linguists at the Czech Language Institute; in the seventies, it was the best structural syntactic description available for the given purpose (cf. Králík, Uhlířová 2007). Due to the consistency of tagging, the Czech Academic Treebank could be technically modernized later (see Czech Academic Corpus 2.0 Guide 2008; Vidová Hladká et al. 2008), and it is still used and still respected, even though today we work with a number of other syntactic frameworks as well as with huge corpora and treebanks. The data used in the analysis are taken from four non-fiction text samples from the Czech Academic Treebank: history of architecture, psychology, sociology and communication engineering, 250 ad-

verbials from the beginning of each text. We took only adverbials expressed by a noun, adverb or adverbial clause, not by pronouns. (For more details, see Uhlířová 1975). Each adverbial was labeled with one of the thirteen possible labels: place, time, manner, degree, means, aspect, cause, purpose, condition, concession, origin, originator, and result.

5. Statistical procedures and interpretation of the data

Frequencies of the adverbial classes are given in Table 1. The data in the columns reads as follows (from left to right): adverbial class r = rank, f = absolute frequency, f_r = relative frequency in per cent.

Table 1
Frequencies of adverbials.

Adverbial	***r***	***f***	f_r
Place	1	273	27.3
Time	2	204	20.4
Manner	3	172	17.2
Means	4	68	6.8
Aspect	5	61	6.1
Condition	6	59	5.9
Measure	7	52	5.2
Cause	8	30	3.0
Result	9	18	1.8
Origin	10	18	1.8
Purpose	11	17	1.7
Concession	12	16	1.6
Originator	13	12	1.2
∑		1 000	100

The data can be interpreted in two steps.

First step: On the primary level of interpretation, we can simply count the occurrences of adverbials attested in the samples, as is done in the second and third columns. According to their frequencies, we may divide all adverbials into three groups. (a) Adverbials with frequencies higher than 10 per cent (in the four samples taken together). Here go the adverbials of place, adverbials of time and adverbials of manner, which, in the total, make roughly two thirds of all adverbials. (b) Adverbials with frequencies within the interval from 5 up to 10 per cent; they express means, aspect, condition and measure. (c) The lowest frequencies are attested with the adverbials of cause, origin, result, purpose, concession and originator. It can hardly be denied that even such an elementary

result of counting is of a certain descriptive value. The absolute as well as the relative frequencies (given in per cent in the third column) show different quantitative weights of adverbials in non-fiction texts: With the increasing rank the frequency of the adverbial class decreases, and so does, implicitly, its relevance in the semantic structure of the text (in the respective non-fiction field). Absolute and/or relative frequency may serve as a quantitative indicator of the content of the text.

Moreover, in each of the four samples, a certain diversification in frequencies may be seen, due to different thematic and stylistic factors. For example, in the text on the history of architecture (dealing with the development of building styles) a higher frequency of the adverbials of time is found, whereas in the text on communication engineering, in which processes in electronic circuits are described and explained, there is a higher frequency of the adverbials of condition. Unfortunately, the data from our four texts are too small to allow considering "boundary" conditions in more detail.

However, such a "traditional" interpretation could be considered as sufficient and useful, let us say, only forty years ago.

Second step: On the advanced level we make a step from the empirical frequencies of adverbials to their probabilities. This step is theoretically decisive and reflects the present demands on quantitative linguistics as "empirical science" in the sense of Altmann´s, Hřebíček´s and Köhler´s theoretical approach to linguistics. The questions sound: What are the numerical representations of word class frequencies and do they abide by a probabilistic law? What model should be used? What are the (dis)advantages and limits of particular models? etc. In the following sections three approaches to a modeling of the distribution of adverbials will be discussed.

5.1. Models for the distribution of adverbials

It has been shown by Hammerl (1990), Liu (2009), and Köhler (2012) that the Zipf-Alexeev approach seems to be a good model for a representation of different word classes (parts of speech, dependency type, motifs). Here we start from the assumption that the frequency in class r takes on values proportional to the preceding class, $r - 1$. This is based simply on the a priori diversification (which decreases) and the a posteriori ranking which captures this process. Hence the relative rate of change of frequency y'/y is proportional to the relative rate of change of the rank (r) in the following way:

$$\frac{dy}{y} = \frac{a + b \ln r}{r} dr .$$

The solution of this differential equation yields

$$y = Kr^{a+b\ln r},$$

where K is simply a constant, or, if taken as a probability distribution (discrete or continuous), it may be considered a normalization constant.

Therefore, we expected that the distribution of adverbial classes (see Table 1) should follow this model. We have found out that the Zipf–Alekseev distribution fits to our data with a very good result: We used it simply as a continuous function. The goodness-of-fit, tested with the help of the determination coefficient, yields $R^2 = 0.97$ (with parameters $K = 273.1124$, $a = -0.0383$, $b = -0.49745$), see Fig. 1.

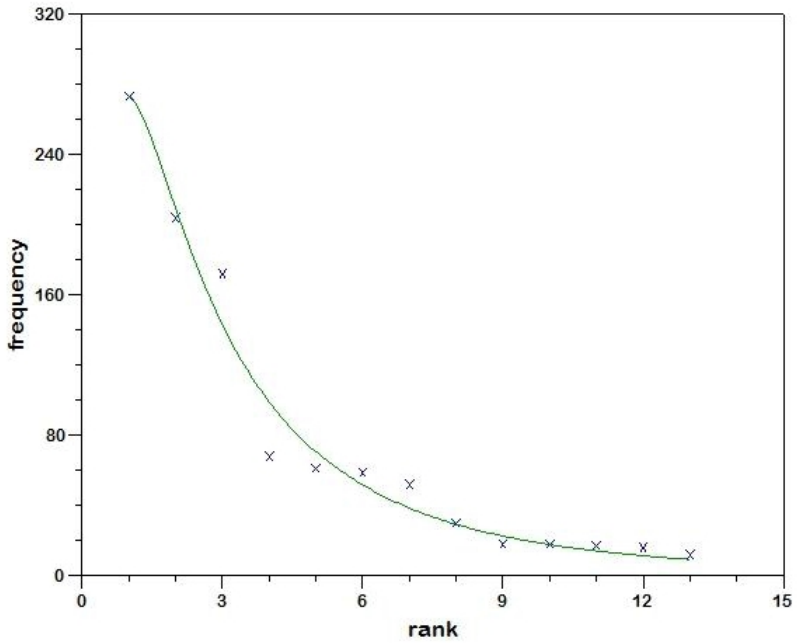

Figure 1. The distribution of all adverbials and the result of the fitting of the Zipf-Alekseev function to the data.

Now let us have a look at our thirteen classes of adverbials with regard to their part-of-speech appurtenances, and, once more, let us ask about their frequency within the parts-of-speech classes. An adverbial expressing place, may belong to a noun, adverb or be a complete clause, etc. Let us test the numbers of adverbial frequencies separately within each of their parts of speech. Though some of the differences in absolute frequencies seem to be quite large, the Zipf–Alekseev function can be fitted to the data with very good results again: the goodness-of-fit is tested with the help of the determination coefficient R^2. The values of R^2 are the following: $R^2 = 0.98$ for nouns (with parameters $K = 260.4290$, $a = -1.1856$, $b = -0.0078$), $R^2 = 1$ for adverbs (with parameters $K = 104.1095$, $a = 0.5208$,

b =-1.4542), R^2 = 0.96 for dependent clauses (with parameters K = 28.3545, a = 0.0356, b = -0.5672), see Figures 2 to 4.

At first sight, we may repeat our conclusion already achieved in the previous point: The assumption that it is possible to find a model of the frequency distribution of adverbials is valid also with regard to their parts of speech; the model is fully compatible with the data. However, the course of the function, as is presented in Figure 3 and 4, does not seem to be plausible for the modeling of observed distributions – there is first an increase and then a monotonically decreasing part. It means that the "mere" good fitting does not mean automatically the best choice of a model. Consequently, we have looked for other models.

Table 2
Part-of-speech frequencies of adverbial classes.
R^2 expresses results of the goodness-of-fit of the Zipf–Alekseev function.

Adverbial	**Noun**	**Adverb**	**Clause**
Place	263	9	1
Time	96	104	4
Manner	79	75	18
Means	68	-	-
Aspect	46	13	2
Condition	30	-	29
Measure	21	30	1
Cause	11	-	19
Result	18	-	-
Origin	18	-	-
Purpose	10	-	7
Concession	4	-	12
Originator	12	-	-
$\sum$	676	231	93
R^2	0.98	1	0.96

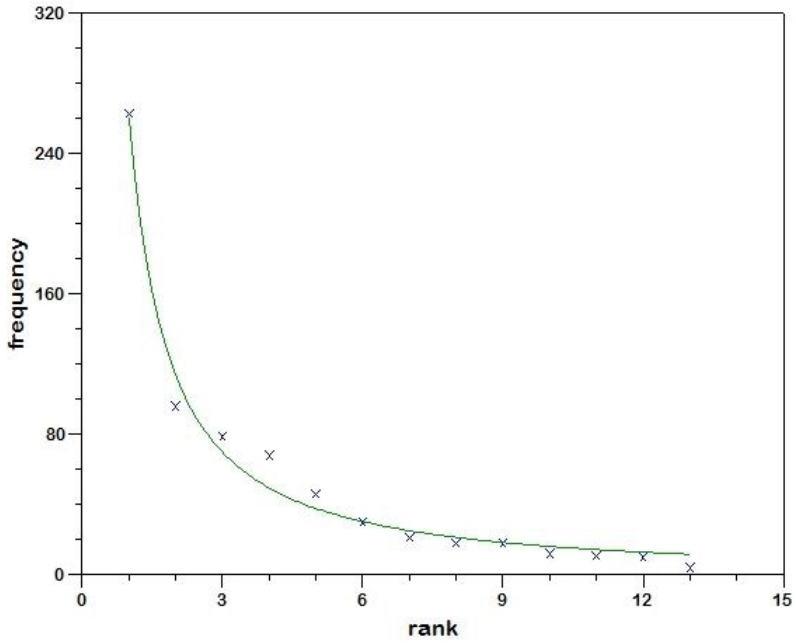

Figure 2. The distribution of adverbials expressed by nouns and the result of the fitting of the Zipf-Alekseev function to the data.

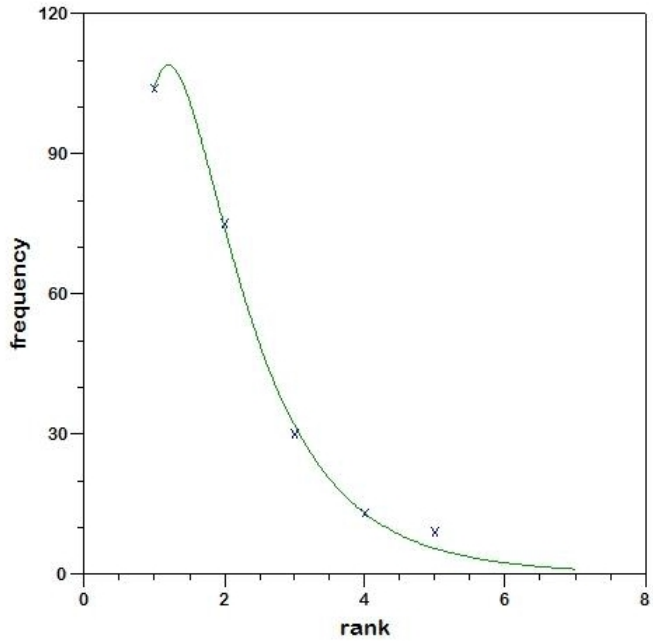

Figure 3. The distribution of adverbials expressed by adverbs and the result of the fitting of the Zipf-Alekseev function to the data.

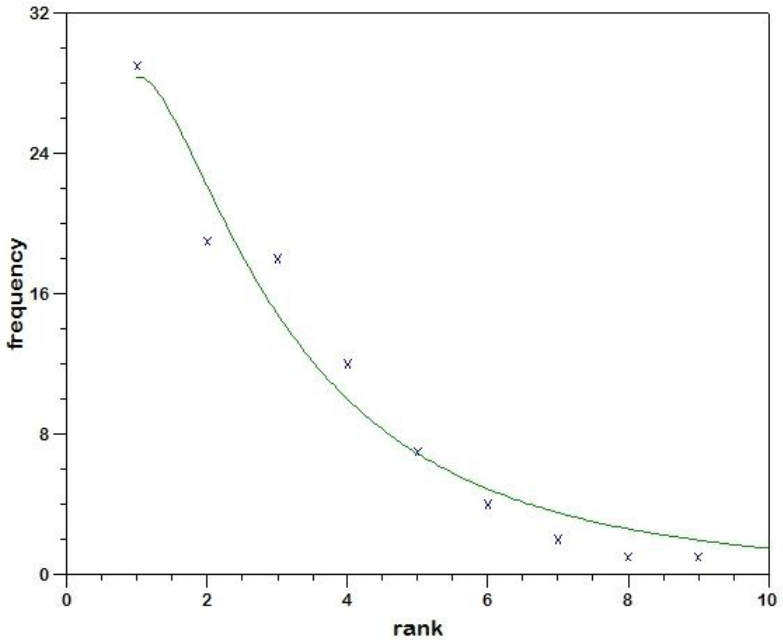

Figure 4. The distribution of adverbials expressed by clauses
and the result of the fitting of the Zipf-Alekseev function to the data.

Since Zipf, it has been well-known that a simple power-law function can be used as an acceptable model for different kinds of distributions; it is based on the assumption that the relative rate change of frequency *y'/y* is proportional to the relative rate of change of the rank (*r*) in the following way:

$$\frac{dy}{y} = \frac{b}{r} dr ,$$

where *b* is the constant. The solution of the equation yields the power-law function

$$y = Kr^b ,$$

where *K* is a well interpreted constant – it usually corresponds approximately to the highest frequency.

The results of fitting the function to the data are presented in Table 3. Except for adverbs, the fits bring worse results than the Zipf-Alekseev function. This result seems to corroborate Köhler's statement according to which this model is more appropriate for data with a bigger inventory size (e.g., word forms or lemmas) (cf. Köhler 2012: 75). Even though the results can be considered acceptable, we are striving for better results.

Table 3
The results of the fitting of the power-law function to the data; the fitting of all adverbials is visualized in Figure 5.

	K	b	R^2
All	298.7741	-0.9023	0.90
Adverb	109.5577	-1.0807	0.98
Noun	260.6419	-1.1985	0.88
Clause	31.0882	-0.8874	0.87

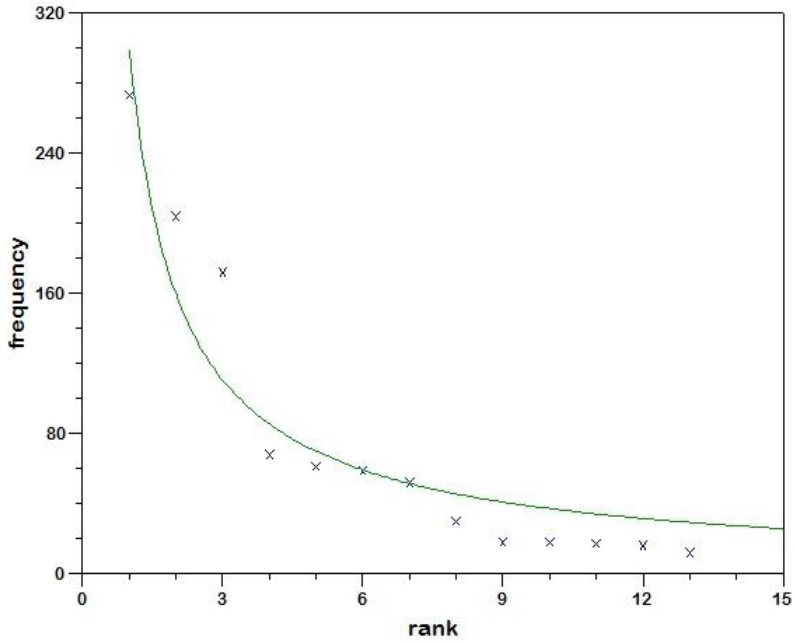

Figure 5. The distribution of all adverbials and the result of the fitting of the power-law function to the data.

Based on a unified theory (Wimmer, Altmann 2005), we assume that the relative change of frequency y could be proportional to the change of rank in a degree which is represented by a function expressing mutual relations between speaker's and hearer's impacts on a process of communication. Specifically,

$$\frac{dy}{y} = \frac{a+br}{cr} dr,$$

where a is a constant (it differs with regard to a specific class of language units, e.g., parts of speech, clauses, morphs), br is the impact of the speaker (he

changes *r* constantly according to *b*) and *cr* is the impact of language community (it restricts speaker's tendency to perform too much change in his speech). The solution of the equation is a function

$$y = Ce^{\frac{\frac{br}{c}+a\ln r}{c}},$$

after a simplification

$$y = Ce^{ar}r^{b},$$

where *C* is a constant of integration. Despite the fact that this function is identical with a "long" version of the well-known Menzerath-Altmann law, the identity is purely formal because there is no identity in theoretical derivation (cf. Köhler 2012: 75). Fitting the function to the data yields excellent results, see Table 4. Consequently, we consider this model to be the best one for an analysis of the frequency distribution of adverbials.

Table 4
The results of the fitting of the function derived from the unified theory to the data; the fitting of all adverbials is displayed graphically in Figure 6.

	C	*a*	*b*	R^2
All	385.5941	-0.3337	0.0189	0.97
Adverb	440.5010	-1.4413	1.5788	0.99
Noun	267.4394	-0.0304	-1.1157	0.98
Clause	45.1149	-0.4662	0.3009	0.97

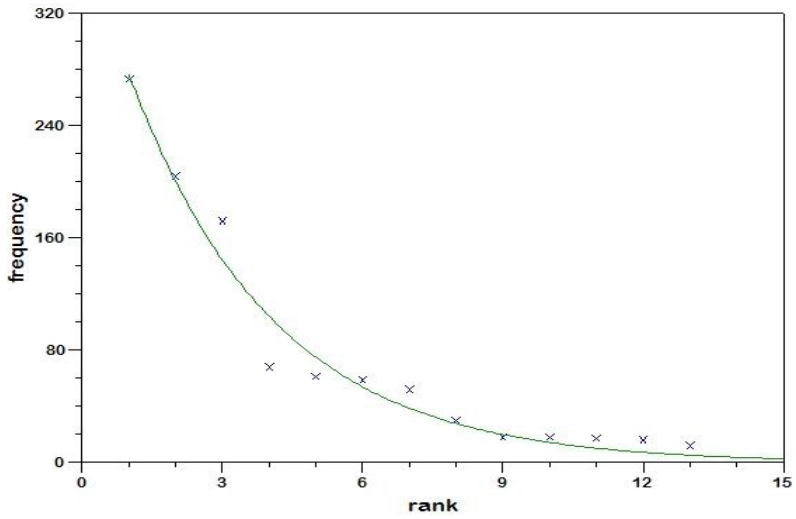

Figure 6. The distribution of all adverbials and the result of the fitting of function derived from the unified theory to the data.

6. Discussion

No new model has been "discovered" (by the way, it was not our aim). Anyway, we did not find anything contrary to what was intuitively expected. We may conclude that we have - most modestly - contributed to the general quantitative understanding of the syntactic constituents and their constructs: Adverbials manifest the same distributional universals as other types of constituents, and they abide by the same statistical law(s). As for models, we tried to demonstrate that the choice of the model is not a trivial task. Obviously, we are still at the beginning of this kind of syntactic research and only further research will reveal which models are more useful than others.

Our results can be interesting also for "traditional" (i.e. mostly descriptive) grammarians: The meaningfulness and logic of the classical dependency framework, especially of the classification of adverbials into the thirteen semantic classes, which we chose for our empirical counting, was fully supported by the results.

But at the same time, we have shown the beginnings of a particular research strategy which is well known from physics: In the first step, one describes a class of entities and shows that they follow some regularity. In the next step, one analyzes an individual class and shows that it is not unique but forms again a hierarchically lower stratum; e.g. one has the set of place adverbials whose elements can be classified as nouns, adverbs or clauses, and their distribution is again an expression of some regularity. In the third step one looks at the nouns, classifies them and states that the same conjecture (in best cases a law) holds again, but the parameters are different. This step will be repeated, so to say, ad infinitum, just as done in physics where from time to time a new, smaller particle will be discovered. The problem in linguistics is that this way is not possible without mathematics, but if it is done by means of mathematics, the time will come in which we shall be able to forecast the result at the next lower level.

Acknowledgments
We are deeply indebted to Gabriel Altmann for his substantial help and advice in mathematical matters, generously given to us during the writing of this text. We would also like to thank to Reinhard Köhler for inspiring ideas and comments.

References

Altmann, G. (1980). Prolegomena to Menzerath's Law. In: Grotjahn, R. (ed.), *Glottometrika 2: 1–10.* Bochum: Brockmeyer.

Altmann, G. (1988). *Wiederholungen in Texten.* Bochum: Brockmeyer.

Altmann, G. (1993). Science and linguistics. In: Köhler, R., Rieger, B.B. (eds.), *Contributions to quantitative linguistics: 3–10.* Dordrecht: Kluwer.

Altmann, G. (2001). Theory Building in Text Science. In: Uhlířová, L., Wimmer, G., Altmann, G., Köhler, R. (eds.), *Text as a Linguistic Paradigm: Levels, Constituents, Constructs. Festschrift in honour of Luděk Hřebíček: 10–20*. Trier: WVT.

Altmann, G. (2005). Diversification processes. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.), *Quantitative Linguistik. Ein internationales Handbuch. Quantitative Linguistics. An International Handbook: 646–659*. Berlin, New York: de Gruyter.

Altmann, G., Köhler, R. (1996). "Language forces" and synergetic modelling of language phenomena. In: Schmidt, P. (ed.), *Glottometrika 15: 62–76.* Trier: WVT.

Best, K.-H. (2005). Satzlänge. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.), *Quantitative Linguistik. Ein internationales Handbuch. Quan--titative Linguistics. An International Handbook: 298–304.* Berlin, New York: de Gruyter.

Czech Academic Corpus 2.0 Guide (2008). [Available at: http://ufal.mff.cuni.cz/rest/CAC/doc-cac20/cac-guide/eng/html/index.html]

Grzybek, P. (2006). On the science of language in light of the language of science, In: Grzybek, P. (ed.), *Contributions to the Science of Text and Language: 1–14.* Dordrecht: Springer Verlag.

Hajič, J., Panevová, J., Hajičová, E., Sgall, P., Pajas, P., Štěpánek, J., Havelka, J., Mikulová, M., Žabokrtský, Z., Ševčíková-Razímová, M. (2006). *Prague Dependency Treebank 2.0*. Philadelphia: Linguistic Data Consortium.

Halliday, M. A. K. (1993). Quantitative studies and probabilities in grammar. In: Hoey, M. (ed.), *Data, Description, Discourse: 1–25.* London: Harper Collins Publishers.

Hammerl, R. (1990). Untersuchungen zur Verteilung der Wortarten im Text. In: Hřebíček, L. (ed.), *Glottometrika 11: 142–156.* Bochum: Brockmeyer.

Hřebíček, L. (1992). *Text in communication: supra-sentence structures*. Bochum: Brockmeyer.

Hřebíček, L. (1993). Text as a construct of aggregations. In: Köhler, R., Rieger, B. B. (eds.), *Contributions to Quantitative Linguistics: 33–39.* Dordrecht: Kluwer.

Hřebíček, L. (1995). *Text levels. language constructs, constituents and the Menzerath-Altmann law.* Trier: WVT.

Hřebíček, L. (1997). *Lectures on text theory*. Prague: Oriental Institute.

Hřebíček, L. (2000). *Variation in Sequences. Contributions to general text theory*. Prague: Oriental Institute.

Hřebíček, L. (2007). *Text in semantics. The principle of compositeness.* Prague: Oriental Institute.

Köhler, R. (1986). *Zur linguistischen Synergetik. Struktur und Dynamik der Lexik*. Bochum: Brockmeyer.

Köhler, R. (2005). Synergetic linguistics. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.). *Quantitative Linguistics. An International Handbook: 760–774.* Berlin, New York: de Gruyter.

Köhler, R. (2012). *Quantitative Syntax Analysis*. Berlin, Boston: Mouton de Gruyer.

Köhler, R., Altmann, G. (2000). Probability distributions of syntactic units and properties. *Journal of Quantitative Linguistics 7*, *189–200.*

Köhler, R., Naumann, S. (2009). A contribution to quantitative studies on the sentence level. In: Köhler, R. (ed.), *Issues in Quantitative Linguistics: 34–57*. Lüdenscheid: RAM-Verlag.

Králík, J., Uhlířová, L. (2007). The Czech Academic Corpus (CAC), its History and Presence. *Journal of Quantitative Linguistics 14*, *265–285.*

Liu, H. (2009). Probability distribution of dependencies based on Chinese dependency treebank. *Journal of Quantitative Linguistics 16*, *256–273*.

Ludvíková, M. (1968). Kombinatorika českých fonémů z kvantitativního hlediska. *Slovo a slovesnost 29*, *56–65*.

Ludvíková, M. (1987). Čísla o hláskách. In: Těšitelová et al., *O češtině v číslech: 91–108.* Praha: Academia.

Mathesius, V. (1947). Úvod do fonologického rozboru české zásoby slovní. In: *Čeština a obecný jazykozpyt:* 59–86. Praha: Melantrich.

Šmilauer, V. (1966). *Novočeská skladba*, Praha: Státní pedagogické nakla--datelství.

Trnka, B. (1935). *A phonological analysis of present-day Standard English.* Praha: Univerzita Karlova.

Uhlířová, L. (1975). O frekvenci příslovečného určení v souvislém textu. *Naše řeč 58*, *133–142.*

Vachek, J. (1940), Poznámky k fonologii českého lexika. *Listy filologické 67,* 395–402.

Vidová Hladká, B., Hajič, J., Hana, J., Hlaváčová, J., Mírovský, J., Raab, J. (2008). *Czech Academic Corpus 2.0*. Philadelphia: Linguistic Data Consortium.

Wimmer, G., Altmann, G. (1999). *Thesaurus of univariate discrete probability distributions*. Essen: Stamm.

Wimmer, G., Altmann, G (2005). Unified derivation of some linguistic laws. In: Köhler, R., Altmann, G., Piotrowski, R. G. (eds.), *Quantitative Linguistik. Ein internationales Handbuch. Quantitative Linguistics. An International Handbook: 760–775*. Berlin, New York: de Gruyter.

Wimmer, G., Altmann, G. (2006). Towards a unified derivation of some linguistic laws. In: Grzybek, P. (ed.), *Contributions to the Science of Text and Language: 329–337.* Dordrecht: Springer Verlag.

Zipf, G. K. (1935). *The psycho-biology of language. An Introduction to Dynamic Philology*. Boston: Houghton-Mifflin. Cambridge: M.I.T. Press (2nd edition, 1968).

Zipf, G. K. (1949). *Human behavior and the principle of least effort*. Cambridge: Addison–Wesley.