Journal of Quantitative Linguistics
2002, Vol. 9, No. 1, pp. 87–96

0929-6174/02/0901-087$16.00

Book Review

Karl-Heinz Best (Ed.), *Häufigkeitsverteilungen in Texten* [Frequency Distributions in Texts]. Göttingen: Peust & Gutschmidt Verlag 2001. 310 pp. [Göttinger Linguistische Abhandlungen; 4]

Reviewed by Peter Grzybek

The so-called "Göttingen Project on Quantitative Linguistics" has repeatedly been reported about in earlier issues of JQL (Best, 1998; Best & Altmann, 1996). This project is not a closed research institution involving a particular number of local scholars, but an open structure of research coordination to which anyone interested may contribute. The object of this project is, generally speaking, the study of frequency distributions (FD) of language units in texts; more specifically, it predominantly focuses on word length in individual texts, but also on sentence length, word classes, and related topics.

The theoretical background of this project is basically represented in two major articles by Wimmer, Köhler, Grotjahn, and Altmann (1994) and Wimmer and Altmann (1996); these need not be commented upon here in detail, since they should be familiar to readers of JQL. These authors' basic assumption is that the frequency distribution of linguistic units is not randomly (chaotically) organized, but follows specific rules (or laws). Specifically, it is assumed that the probability of a particular unit (e.g., the class of words with length x) is proportionally related to the probability of its neighbouring unit (e.g., the class of words with length $x - 1$). This proportional relation is not considered to be constant (which would result in an equation of the type $P_x = aP_{x-1}$, but is represented by a function obtaining the form $g(x)$. Thus we obtain the basic formula $P_x = g(x)P_{x-1}$, which, depending on the specific form of $g(x)$, results in various frequency distributions. The function $g(x) = ax^{-b}$ is regarded to be the basic organizational function; it results in the so-called Conway-Maxwell-Poisson distribution which, if $P_0 \neq 0$, has the form

$$P_x = \frac{a^x}{(x!)^b} P_0, \quad x = 0, 1, 2 \tag{1}$$

Other functions are seen as modifications of this basic function; further details need not be mentioned here, and can be found in the articles mentioned above.

Now, Karl-Heinz Best, who is one of the major promoters and coordinators of the Göttingen project, has edited a volume of articles on related questions: *Frequency Distributions in Texts* [Häufigkeitsverteilungen in Texten]. It is empirically, rather than theoretically, oriented, an orientation which is two-fold, however: on the one hand, the specific hypotheses mentioned above are tested with regard to particular texts; on the other hand, and in addition to this, the book's (and the project's) intention can be characterized by what may be called the "explorative formulation of hypotheses," that is, the intention to generalize theoretically the individual results achieved and to derive specific hypotheses from this. In this sense, the empirical approach characterizing the book under review can be regarded as the "bottom-up" face of the Göttingen project coin, whereas the theoretical foundations (and the theoretically postulated relevance for linguistic studies) would thus have to be considered its top-down face. Quite logically, then, the book under review intends to present the state of the art of the Göttingen project, in the first place; but it also wants to point to remaining and emerging questions, and to delineate perspectives for future research.

The whole volume is divided into four major parts:

I. Part one is represented by a synoptic "foreword" (v–xx) by Karl-Heinz Best;
II. Part two involves "Studies on Frequency Distributions in Texts" (pp. 1–280); without a doubt, this part represents the center of the whole volume: in it, no less than 18 articles by various authors can be found;
III. Part three (pp. 281–283) offers a short "Biographical Note on S.G. Čebanov (1897–1966)"; and
IV. Part four, which concludes the volume, is an "Annotated Bibliography on Works of the Göttingen Project" (part IV, pp. 284–310).

Let us start with part I which, following a short general introduction (I.1.), presents the theoretical foundations of the Göttingen project (I.2.), as outlined above. For most readers of JQL, the state-of-the-art report on word length studies (I.3.) is likely to be more relevant. Three categories of analyses are distinguished: text analyses on various historical levels of German texts (Old High German, Middle High German, Early New High German, New High German) are summarized as "German" (I.3a), whereas "Lower German" as a German dialect is represented in a separate subchapter (I.3b), just as are studies on languages other than German, summarized as "Foreign Languages" (I.3c).

I.3a As to "German" texts, there seems to be a tendency for the hyperpoisson distribution to represent a good model for all historical levels, eventually surpassed by Poisson distribution in Old High German. In fact, Best (p. ix) claims the hyperpoisson distribution to be a basic model for German texts in general, be they literary texts, letters, fables, songs, juridical, or other text types. If this assumption should turn out to be justified on a wider level of analyses, factors such as "text type (genre)," "text length," "authorship," or others would not have to be considered as factors influencing the type of frequency distribution within a given language, but only, if so, its parameters. It goes without saying that this is a rather daring, but also tempting, hypothesis, and it will be interesting to see its future fate, when more specific studies will be available.

I.3b Interestingly enough, Low German texts (colloquial language, short stories, poems) seem to follow a different type of distribution: 106 of the 110 texts analyzed may well be fitted to the positive negative binomial distribution.

I.3c Quite characteristically, languages other than those mentioned thus far, are called "foreign languages." Meanwhile, data are available from about 40 different languages (for some of which there are data from only one single text, however). The central conclusion of this chapter is that it is not one single model which is adequate for all languages (at least for all syllabic languages, one should say) as proposed by Fucks in the 1950s, but, in line with the general theoretical assumptions, different models. However, it is again the hyperpoisson distribution which is favored as "some kind of a basic model" (p. xi), and which, in a way, may thus be called the implicit hero of the whole book.

With these general tendencies in mind, let us now turn to the individual studies present in part II. This part can, in fact, be subdivided into a number of subchapters, depending on the type of linguistic unit under study:

II.1 FD of morph length
II.2 FD of syllable length
II.3 FD of word length
II.4 FD of rhythmic units
II.5 FD of sentence length
II.6 Studies on the Zipf-Mandelbrot law.

II.1 Without a doubt, the quantitative study of morphs and of FD of morph length (measured by the number of phonemes per morph), is innovative.

According to the relevant study by K.-H. Best ("Zur Länge von Morphen in deutschen Texten," pp. 1–14), based on 21 German journalistic texts, the hyperpoisson distribution is a good model. This result is particularly interesting because the very same texts are studied again in subchapter II.2, when the frequency of syllable lengths is studied.

II.2 In this subchapter, 2 studies are represented analyzing syllable length in German journalistic texts. The study by Cassier ("Silbenlängen in Meldungen der deutschen Tagespresse," pp. 33–42) is based on 82 German news items; it is an extension of the first study (Best: "Silbenlängen in Meldungen der Tagespresse," pp. 15–32), which, in turn, is based on the 21 news items mentioned in II.1, which are analyzed with regard to syllable length distribution this time. Both studies arrive at the result that the Conway-Maxwell-Poisson distribution is a good model. Thus both morph and syllable length frequencies can be modeled on the basis of the general theoretical approach, albeit resulting in two different models.

III.3 The subchapter on word length distribution can be regarded as the heart of the book; in it, nine studies on texts from different languages are presented: A. Ahlers (different types of Low German texts); S. Ammermann (German letters over a period of 500 years); K.-H. Best and I. Kaspar (Faeroese letters); K.-H. Best and J. Zhu (Chinese texts and dictionaries); J. Drechsler (Gaelic email messages); A. Kiefer (texts from the Pfalz, a German Rhenish-Franconian dialect); S. Kuhr (songs and fables by Martin Luther); N. Rheinländer (Dutch letters); and A.B. Stark (Swiss German letters).

It is not the place, here, to present the individual results in detail. Let us summarize general tendencies arising from theses studies instead.

1. In all studies, FD models can be found which are in coincidence with the theoretical assumptions mentioned above;
2. The FD models diverge with regard to languages, at least, and also with regard to language varieties and dialects; other factors are not systematically controlled.

II.4 Also innovative is K.-H. Best's study on the FD of rhythmic units in texts. The question at stake is how many unstressed syllables can be found between 2 stressed syllables, and how the FD of these distances can be modeled. Based on German language material by Marbe from the beginning of the 20th century, Best finds the hyperpoisson distribution to be a good model; he is well aware of the fact, however, that his study is based on rather

thin ice and cannot be more than (one should say: not less) a hint at research to be done in future.

II.5 Studies on sentence length represent the second important block in the whole volume. In this part, 4 analyses can be found, 2 on German texts (K.-H. Best, M. Wittek), the others on Chinese (Zh. Jing) and Russian (M. Roukk).

Following Altmann's (1988) general postulate that sentence length distribution is supposed to follow the negative binomial distribution, if a sentence is measured by the number of clauses it contains, and the hyperpascal distribution, if measured in words per sentence, empirical findings of the last years have led to contradictory results.

For the sake of demonstration, let us look at the relevant studies on German texts. After Niehaus' (1997) analyses there seemed to be good reason to assume that the frequency distribution of sentence length (measured in clauses) indeed follows the negative binomial distribution. Her subsequent study (Niehaus, 2001) on the frequency distribution of sentence length measured in the number of words per sentence on the basis of the very same texts, however, showed that the FD also follows the negative binomial distribution: the hyperpascal distribution is a good model only if classes with range 5 (1–5, 6–10, etc., words per sentence) are built and analyzed; still, under this condition, the negative binomial distribution also turns out to be a good model.

Now, in the volume under discussion, K.-H. Best, in his contribution "Wie viele Wörter enthalten Sätze im Deutschen?" (pp. 167–201), again finds the negative binomial distribution to be the best model for sentence length FD measured in words, with and without class contractions. Best also refers to reanalyses of the Niehaus 1997 study, which, as he reports (p. 198), show that the hyperpoisson distribution leads to even better results than the negative binomial distribution; according to Best (p. 198), this is also in line with findings by Strehlow (1997) and Wittek (1995). Curiously enough, however, Wittek himself presents these results in the volume under discussion ("Zur Entwicklung der Satzlänge im gegenwärtigen Deutschen" (pp. 219–247), and it is not the hyperpoisson distribution, but the positive Poisson distribution, which turns out to be a good model for his 80 geographical texts in the German language. This fact, however, is not really astonishing, since, under specific conditions ($r \to \infty, q \to 0, rq \to a$), the zero-truncated ("positive") negative binomial distribution converges against the positive Poisson distribution (Wimmer & Altmann, 1999, p. 541) – a fact interestingly enough observed by Best and Altmann (1996, p. 176) with regard to the FD of word length in

German texts. The hyperpoisson distribution, however, turns out to be a good FD model for the special case of sentence length of German proverbs (Grzybek & Schlatte, 2002). It seems most reasonable to assume that influencing factors (such as text length, text type, authorship) come into play, the role of which is completely unknown so far.

In this respect, M. Roukk's study on sentence length in Russian texts (pp. 211–218), is relevant; here, the author studies the FD of sentence length in 22 Russian texts (11 literary, 9 journalistic, and 2 scientific texts), sentence length being measured in the number of words per sentence, with classes of size 5 (1–5, 6–10, etc.). Again, there is a related study by the same author on sentence length FD for sentences being measured in the number of clauses contained (Roukk, 2001), in which the author claimed the positive negative binomial distribution to be a good model (also referring to the hyperpoisson distribution as possibly coming into play however). Now, with sentence length measured in number of words per sentence, the hyperpoisson distribution turns out to be a good model, as the author claims (p. 212). This claim should not remain uncontradicted however, since a re-analysis of Roukk's data shows that her data can also be modeled well with the negative binomial distribution, the hyperpascal distribution coming into play only when data are contracted into more encompassing class sizes; as opposed to the negative binomial distribution, the hyperpoisson distribution – which indeed is a good model for Chekov's short stories, too – seems no good as a model for longer Russian texts (cf. Grzybek & Kelih, 2002; Kelih, 2002, who have done some systematic study on the influence of text length, sentence definition, and class size).

Thus, on the one hand, the negative binomial distribution in fact seems to be of the utmost importance for the modeling of sentence length distribution, be it on the number of clauses or the number of words per sentence. Yet it seems reasonable to side with Best's (p. 199) conclusion that, just as in the case of word length distribution, we are concerned here not with one specific model valid for any text from whatever language, but with a number of distributions, depending on factors such as language, functional style, authorship, etc. Therefore studies like Jing's on sentence length distribution in Chinese texts (pp. 202–210) are so important, since most studies concentrate on texts from Indo-European languages; interestingly enough however, the distribution of sentence length (measured in the number of characters) in Chinese also follows the hyperpoisson distribution.

As a major result of this whole chapter, then, one may draw the conclusion that the FD of sentence length, just like that of word length, follows particular

laws, which might well be in line with the theoretical approaches outlined by Wimmer et al. (1994) and Wimmer and Altmann (1996). The studies available thus far, however, have not systematically enough studied those factors which come into play as influencing factors.

II.6 A. Knüppel's contribution "Untersuchungen zum Zipf-Mandelbrot-Gesetz an deutschen Texten" (pp. 248–280), in which rank-frequency studies of word forms are related to word length studies, is highly innovative indeed: Based on 20 German letters, Knüppel analyzes how the parameters a and b of the Zipf-Mandelbrot formula – $P_x = P \cdot (x+b)^{-a}$ with $P^{-1} = \sum_{i=1}^{n} (j+b)^{-a}$ – can be interpreted, in particular, how they can be related to word length. As a result, she finds that a and b seem to be closely correlated, and that both parameters are influenced by word length.

Best rounds out the volume by a short biographical note on S.G. Čebanov (1897–1966), a Russian medical doctor who, in 1947, published an article on the importance of the Poisson distribution, thus anticipating subsequent assumptions by W. Fucks. Best also offers an annotated bibliography with works from the "Göttingen project" (pp. 284–310), which is also available as a permanently updated internet file (http://www.gwdg.de/~kbest/project.htm). This bibliography is extremely helpful for any scholar interested in quantitative linguistics, and everyone working in this field should help the compiler by sending him up-to-date information about relevant works.

Summarizing then, one can say that by the publication of the present book, Karl-Heinz Best has presented convincing results on the question of frequency distributions in texts. The volume also makes it clear that there are many open questions to be tackled in future studies. Let us name some of them:

1. From many studies, the status and adequacy of the fitted distributions do not become completely clear – cf. expressions such as "since *xy* has turned out to be a good FD model in case of . . . , it has initially been tested" (p. 6), "in analogy to . . ., *xy* was fitted to. . ." (p. 36), "*xy* is fitted to the word length distribution" (p. 50), "the fitting of *xy* resulted in. . ." (p. 67), "it seemed reasonable to fit *xy* to . . ." (p. 133), and so forth. In other words, it does not become clear, if only a particular FD model was exclusively checked, or fitted to the material, the selection of the model being based on prior experience or theoretical assumptions, or if only those results are reported which, for the given study, represent "the best fit" (i.e., fitting both the empirical findings and theoretical assumptions).

2. Given that a particular model turns out to be adequate for texts of a given language still, the parameters of that model are not constant, as we know. A problem unsolved then, is the linguistic interpretation of these parameters. One good first approach might be to see how 2 (or more) parameters are interrelated (cf. pp. 12, 30, 99, or 261). Also, the derivation of particular characteristics, such as Ord's criteria, may be helpful (cf. pp. 28, 83ff., 97); in this specific case, one should be aware of the fact, however, that linear regression models will be effective only with regard to particular FD models, and that other approaches like concentration ellipses or related techniques might turn out to be more appropriate. Finally, the relation between the parameters of a given model and external factors is of utmost importance; a good example for this approach is Knüppel's above-mentioned study, relating the parameters a and b of the Zipf-Mandelbrot distribution to word length as an influencing factor (for further factors see below).
3. A further problem is that of the level of analysis: as is well known, by now, the analysis of text corpora may result in artificial results, since, in that case, the material represents some kind of "quasi text," that is, a text mixture. Yet, in a particular sense, this tendency continues, if one considers an analysis of particular texts (or of a particular text type) to be representative for a given language. As long as no systematic studies are available, this focusing might be rather dangerous, and this danger has clearly to be seen, if the letter as one specific genre is considered to be "an ideal text type" – not only because there are quite a number of different types of letter (cf. Ermert, 1979), but, more importantly, because information on possible genre specifics (and related influencing factors) gets lost.
4. It should have become obvious, both from the review above and from the problems pointed out in the ruminations ahead, that a number of factors are likely to influence the frequency distribution of linguistic units (be it that these factors influence the type of FD model, or "only" the parameters of a given model. Among these factors – which most probably interact in specific ways – are, to name but a few:

- text length,
- text type (genre),
- authorship (which gives an opportunity to understand individual style as the individual variation of supra-individual regularities),
- language (which might result in a specific kind of language typology),

- language layer (dialect vs. standard language, oral vs. written communication, etc.),
- analytic level within a given language (e.g., sentences measured in clauses, words, syllables, etc.), and
- temporal aspects (diachronic aspects of language evolution and language change, of literary history, etc.).

Hopefully, some answers to these questions raised will soon be yielded by QuanTA, the Graz project on Quantitative Text Analysis: http://www-gewi.uni-graz.at/quanta. On the basis of texts from three Slavic languages (Croatian, Russian, Slovenian), this project, which is financially supported by the Austrian Fond für wissenschaftliche Forschung, specifically focuses on factors influencing the frequency distribution of word lengths (Grzybek & Stadlober, 2002).

Peter Grzybek, Universität Graz, Institut für Slawistik, Merangasse 70, A-8010 Graz. E-mail: grzybek@uni-graz.at

REFERENCES

Altmann, G. (1988). Verteilung der Satzlängen. In K.-P. Schultz (Ed.), *Glottometrika 9* (pp. 147–169). Bochum: Brockmeyer.

Best, K.H. (1998). Results and perspectives of the Göttingen project on quantitative linguistics. *Journal of Quantitative Linguistics, 5*, 155–162.

Best, K.H., & Altmann, G. (1996). Project report. *Journal of Quantitative Linguistics, 3*, 85–88.

Ermert, K. (1979). *Briefsorten. Untersuchungen zu Theorie und Empirie der Textklassifikation.* Tübingen: Niemeyer.

Grzybek, P., & Kelih, E. (2002). *Sentence length distribution in Russian texts: The influence of sentence definition and class size* (in press).

Grzybek, P., & Schlatte, R. (2002). Zur Satzlänge deutscher Sprichwörter – Ein Neuansatz. In E. Piirainen & I. Piirainen (Ed.), *Phraseologieforschung in Raum und Zeit* (pp. 287–305). Baltmannsweiler: Scheider.

Grzybek, P., & Stadlober, E. (2002). *QuanTA – The Graz Project on Quantitative Text Analysis.* Manuscript submitted for publication.

Kelih, E. (2002). *Untersuchungen zur Satzlänge in Slowenischen und Russischen Prosatexten.* Graz: Diplomarbeit.

Niehaus, B. (1997). Untersuchung zur Satzlängenhäufigkeit im Deutschen. In K.-H. Best (Ed.), *Glottometrika 16* (pp. 213–275). Trier: WVT.

Niehaus, B. (2001). Die Satzlängenverteilung in literarischen Prosatexten der Gegenwart. In L. Uhlířová, G. Wimmer, G. Altmann, & R. Köhler (Eds.), *Text as a linguistic paradigm: Levels, constituents, constructs. Festschrift in honour of Ludek Hrebícek* (pp. 196–214). Trier: WVT.

Roukk, M. (2001). Satzlängen in Texten von A. Tschechow. *Göttinger Beiträge zur Sprachwissenschaft, 5*, 113–120.

Strehlow, M. (1997). *Satzlängen in pädagogischen Fachartikeln des 19. Jahrhunderts.* Göttingen: Staatsexamensarbeit.

Wimmer, G., & Altmann, G. (1996). The theory of word length: Some results and generalizations. *Glottometrika 15* (pp. 112–133). Trier: WVT.

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

Wimmer, G., Köhler, R., Grotjahn, R., & Altmann, G. (1994). Towards a theory of word length distribution. *Journal of Quantitative Linguistics, 1*, 98–106.

Wittek, M. (1995). *Zur Entwicklung der Satzkomplexität im gegenwärtigen Deutschen.* Göttingen: Staatsexamensarbeit.