Descriptiveness, Activity and Nominality in Formalized Text Sequences

Peter Zörnig

in cooperation with

Kamil Stachowski
Ioan-Iovitz Popescu
Tayebeh Mosavi Miyangah
Panchanan Mohanty
Emmerich Kelih
Ruina Chen
Gabriel Altmann

2015
RAM-Verlag

Studies in quantitative linguistics

1. U. Strauss, F. Fan, G. Altmann, *Problems in quantitative linguistics 1*. 2008, VIII + 134 pp.

2. V. Altmann, G. Altmann, *Anleitung zu quantitativen Textanalysen. Methoden und Anwendungen*. 2008, IV+193 pp.

3. I.-I. Popescu, J. Mačutek, G. Altmann, *Aspects of word frequencies*. 2009, IV +198 pp.

4. R. Köhler, G. Altmann, *Problems in quantitative linguistics 2*. 2009, VII + 142 pp.

5. R. Köhler (ed.), *Issues in Quantitative Linguistics*. 2009, VI + 205 pp.

6. A. Tuzzi, I.-I. Popescu, G.Altmann, *Quantitative aspects of Italian texts*. 2010, IV+161 pp.

7. F. Fan, Y. Deng, *Quantitative linguistic computing with Perl*. 2010, VIII + 205 pp.

8. I.-I. Popescu et al., *Vectors and codes of text*. 2010, III + 162 pp.

9. F. Fan, *Data processing and management for quantitative linguistics with Foxpro*. 2010, V + 233 pp.

10. I.-I. Popescu, R. Čech, G. Altmann, *The lambda-structure of texts*. 2011, II + 181 pp

11. E. Kelih et al. (eds.), *Issues in Quantitative Linguistics Vol. 2*. 2011, IV + 188 pp.

12. R. Čech, G. Altmann, *Problems in quantitative linguistics 3*. 2011, VI + 168 pp.

13. R. Köhler, G. Altmann (eds.), *Issues in Quantitative Linguistics Vol 3*. 2013, IV + 403 pp.

14. R. Köhler, G. Altmann, *Problems in Quantitative Linguistics Vol. 4*. 2014, VIII+148 pp.

15. Best, K.-H., Kelih, E. (eds.), *Entlehnungen und Fremdwörter: Quantitative Aspekte*. 2014. VI + 163 pp.

16. I.-I. Popescu, K.-H. Best, G. Altmann, G. *Unified Modeling of Length in Language*. 2014, VIII + 123 pp.

17. G. Altmann, R. Čech, J. Mčutek, L. Uhlířová (eds.), *Empirical Approaches to Text and Language Analysis dedicated to Luděk Hřebíček on the occasion of his 80th birthday*. 2014. VI + 231 pp.

18. M. Kubát, V. Matlach., R. Čech,, *QUITA Quantitative Index Text Analyzer*. 2014, VII + 106

19. K.-H. Best, *Studien zur Geschichte der Quantitativen Linguistik.* 2015. III + 158 pp.

ISBN: 978-3-942303-31-6

RAM-Verlag
Stüttinghauser Ringstr. 44
D-58515 Lüdenscheid
RAM-Verlag@t-online.de
http://ram-verlag.de

Preface

In the present book we study characteristics of language based on formalized text sequences. The study of text as a sequence of various entities is rapidly developing in form of articles, omnibus volumes and monographs. In fact, our linguistic study can be considered as a part of a very fertile interdisciplinary research activity devoted to the analysis of information sequences. Such sequences occur also in computational biology (e.g. in form of DNA strings), in coding theory and data compression. While qualitative linguistic analysis searches for rules which are important for language learning, quantitative analysis tries to capture hidden mechanisms which are not necessary for the understanding of language. Except for certain poetic phenomena, e.g. rhythm which can be produced consciously, these mechanisms cannot be learned and do not represent the core of standard linguistics.

In the present book, a group consisting of mathematicians and linguists – specialists for a certain language – attempts to discover textual phenomena which may seem to be strange for the "normal" linguistics but whose deciphering may help to reveal candidates for laws. Laws are the highest aim of science because without them no theories and no explanations are possible. Unfortunately, in linguistics the testing of a hypothesis is never finished, one can at most validate it to a certain degree. In practice, this validation will never terminate because one would be forced to analyze all languages and, in case of text laws, as many texts as possible. Here no corpuses can help because none of them contains the complete history of language, the evolution of an individual speaker or a complete collection of text sorts.

Hence our attempts merely reveal a few of the infinite number of facets of a text. We try to collect data, find models of their behavior in form of hypotheses, test them, compare the results in texts of eleven languages available to us and try to create a research domain which will never be satisfactorily explored.

We present all observed data in order to enable other researchers to analyze them applying other methods or other characterizations, and to formulate and test other hypotheses. We reduced the whole field to specific phenomena of description, activity and specifying, otherwise the study would be too extensive. Nevertheless, we show at some places the possibility of going into the depth of the hierarchy of phenomena.

Peter Zörnig

Contents

1. Introduction

Every linguistic entity has an uncountable number of properties. Their number does not depend on the entity itself – as has been supposed for centuries in the philosophy – but on the *status quo* in linguistics as a science. The researchers define the entities, establish some classifications according to the aim of their research, search for the links between the properties and seek the forces that bring them about. Usually the links between properties are substantiated linguistically – as shown in synergetic linguistics – and are based on the assumption that language is a dynamic system. The text, as the most complex linguistic entity, has the most properties of all, comprising both those of hierarchically lower composing entities and its own ones. While lower entities (except for clause and phrase) are static or local constructions that can be found in dictionaries, the text is in addition a *sequence* of lower units and is able to display a special aspect of the course of any given property.

The fact that texts are written differently because they follow different (conscious or unconscious) aims is well known. There are disciplines like text-type and style classification, language development based on texts of the youth, frequency dictionaries, metrics, speech act, psycholinguistics, sociolinguistics, etc. following quite different aims. Some of these disciplines – or better, some aspects – have already been partially quantified and some mathematical models can be found in this research (cf. e.g. Janda 2013). The history of quantifying linguistic phenomena with mathematical models is more than one hundred years old and the bibliography is very extensive (cf. Köhler 1995). However, mathematical models are no bearers or warrants of truth; they merely reflect our striving for more understandable and more exact capturing of the research object, and yield us the possibility of operating formally with the "facts" discovered. Disciplines using mathematics develop faster than other ones.

Here we shall restrict ourselves to two domains: the expression of text descriptiveness vs. its expression of activity concerning only adjectives and verbs, and the nominality vs. predicativity/specification which is restricted here to the comparison of noun, adjective and verb occurrences. Descriptiveness is expressed by the use of adjectives specifying a noun, and some adverbs specifying both the adjectives and the verbs. Here adverbs and adverbial expressions will be omitted. The adjectives are usually parts of the nominal phrase (*the nice girl*) but they can be added also to the verb (*the girl is nice*; Hungarian: *a szép kislány; a kislány szép*; Russian: *krasivaja devuška; devuška krasivaja*) with or without copula according to the grammar of the given language. Activity is expressed (mostly) by verbs and can even be scaled. We shall not do it here and take into account all forms of the verb "to be" only if it is expressed overtly, e.g. in Indonesian, in stressed forms one uses *ada*, otherwise it does not exist; in other languages it may be quite complex, e.g. the personal forms of *to be* exist but as copula it is omitted. We omit also the modal and the other auxiliary verbs if they accompany the main verb. A text translated from an Indo-European language into Indonesian would be here automatically less active if we counted also “to be”. The cases of Odia and Turkish are described below. Here we are not interested in language typology but in text properties. Verbs consisting of several parts, e.g. in sentence like Slk. *Bol by som býval chcel urobiť*, E. *I would like to work*, Hu. *Szerettem volna megcsinálni* will be considered as 1 verb. Gerunds, gerundives and participles may be interpreted according to the official grammar. In some languages they have different forms, e.g. Slk. *tancoval spievajúc* (he danced singing) but *spievajúci muž tancoval* (the singing man danced). In the first case there are two verbs, in the second one there is an adjective and a verb. In some languages a decision will be necessary in several cases. For a survey of English see Krug (2001), Quirk et al. (1985).

Nominality is both a matter of style and text sort, perhaps also a matter of language. One can express the same subject either by a mere verb, e.g. *I inform you* or one can express this subject using also a noun, e.g. *I convey to you the information,* as is usual in information-theoretical texts. As to nouns, we consider nominal compounds as one noun even if they contain a blank or a conjunction or other joining morphemes and ignore the rest, e.g. *United States; light velocity; bottle filling machine; Natur- und Kulturschutz,* full personal names (*Franz Liszt*), titles (*Der Vorsitzende des internationalen Kommittees*), etc.

In general, one supposes that lyrical poetry is rather descriptive and epical rather active but this need not be the case (cf. Popescu, Čech, Altmann 2013). Further, one supposes that scientific and judicial text-types are rather nominal than active, but this must be tested separately.

It has been shown in the literature that these three word classes may give a text a special character: the adjectives emphasize the descriptiveness, the verbs show the activity, and the nouns may be characteristic of the nominalized expression, e.g. in scientific or judicial texts. The numbers of occurrences of these classes may be combined, their sequences can be scrutinized and help to disclose some aspects of the text dynamics.

The study of predicativity/specification could be continued taking into account logical predicates of second, third, … order, e.g. adverbs are predicates of both adjectives and verbs, but this way of seeing the text has not been studied up to now. In the same way, the trees developed in some grammars (dependency and generative grammar) may be reinterpreted in this sense: for each word the downward number of steps in the hierarchy (tree) will be stated and an indicator can be constructed taking into account the numbers obtained. It would be more appropriate to speak about specification because it is easier to state semantically which word specifies another word than to get problems with the philosophical concept of predicate. The problem may be considered also from the topic-comment or thema-rhema points of view.

A slightly more complex task is the scaling of word classes; at a deeper level even the entities of an individual class may be scaled; for example, the verbs according to the degree of the activity they express, e.g. *to run* expresses more activity than *to sleep*; or to the history of the rise of an activity in the biological development of Man, e.g. *eat, feel, move, play, think*, *speak* arose in different periods of our development – but this task needs the cooperation of biologists and anthropologists; the adjectives may be scaled according to the level of the properties (e.g. *nice, pretty, beautiful*, *magnificent*, *splendid*, etc.) or by gradation expressed grammatically or lexically. Nouns can be scaled according to the abstract/concrete scale, specific/generic scale, imagery (cf. e.g. Darley, Sherman, Siegel 1959; DeVito 1967; Flesch 1950; Paivio 1979; Pikas 1966; Kisro-Völker 1984; Ballmer, Brennenstuhl 1986), etc. The same holds for all other word classes. Some of the categories have been scrutinized by psycholinguists, child language specialists, grammarians, semanticists, etc. In general linguistics, it is rather a task for the future, even if one finds a great number of trials both in books and on the Internet.

In the present book we shall directly analyze or take into account the results concerning some languages, even if the counting had been performed using different principles. We restrict ourselves to the given aspects and shall not search for their interrelations with other viewpoints. Such an enterprise would be infinite and must be left to future research. It can be performed only stepwise. We consider merely modern journalistic texts; automatically, one could extend the research to the development of journalistic texts historically or scrutinize other text types.

Quite different approaches to sequences in texts can be found in Mikros, Mačutek (2015).