Book Reviews

Quantitative Syntax Analysis

Reinhard Köhler
(Trier University)

Berlin and Boston: De Gruyter Mouton (Quantitative Linguistics series, edited by Reinhard Köhler, Gabriel Altmann, and Peter Grzybek, volume 65), 2012, x+224 pp, hardbound, ISBN 978-3-11-027219-2, €99.95, $140.00

Reviewed by
Chunshan Xu and Haitao Liu‡*
**Anhui University of Architecture, ‡Zhejiang University*

Quantitative linguistics (QL) is a discipline of linguistics, that, using real texts, studies languages with quantitative mathematical approaches, aiming to precisely describe and explain, with a system of mathematical laws, the operation and development of language systems. Later in this review, we will address the relationship between QL and computational linguistics. *Quantitative Syntax Analysis* is a recent work on QL by Reinhard Köhler that not only provides a comprehensive introduction to the work of QL on the syntactic level, but also sketches the theoretical grounds, the research paradigm, and the ultimate goals of quantitative linguistics in general.

In the first chapter, Köhler points to the vital role of syntax in language: Syntax enables language users to code structures instead of ideas as wholes. A text embodies a complex cognitive formation and meets several basic requirements in human communication, which implies that language is not autonomous, but a dynamic communicative system used by human beings. Hence, the ultimate understanding and explanation of syntax (and the whole language system) depends on usage-based investigation of the cognitive basis and the functional requirements of language, which is somewhat neglected in many mainstream syntactic studies. Even in those cases where explanatory power is acknowledged as the ultimate goal of linguistic investigation, the necessary knowledge is still required as to what a scientific linguistic theory is and how such a theory may be built. So far, it is rare for quantitative means to be used in syntactic study. One reason is that many syntacticians are too addicted to the enshrined traditional paradigms that have been proven to be somewhat inadequate when it comes to processing real texts. And that is why in computational linguistics, "devout executors of the belief in strictly formal methods as opposed to statistical ones do not have any chance to succeed" (page 4).

The second chapter, entitled "The Quantitative Analysis of Language and Text," begins with an explanation of the difference between quantitative linguistics and the formal branches of linguistics that have once been widely used in computational linguistics: QL is concerned with the quantitative properties important for understanding the development and the operation of linguistic systems, whereas the formal branches of linguistics use only qualitative mathematical means and formal logics to model structural properties of language, overlooking, in most cases, the aspects of systems that exceed structure, viz., functions, dynamics, and processes. Köhler points out that the successes of modern natural sciences (the exact, testable statements, the precise predictions, and the copious applications) all derive from their instruments and their advanced models. This implies that these instruments and models, for which the

quantitative parts of mathematics (probability theory and statistics, function theory, differential equations) are indispensable ingredients, are worth integrating into linguistics, which is the aim of QL.

Chapter 3, entitled "Empirical Analysis and Mathematical Modeling," reviews the important works of quantitative syntactic analysis. In the first section, Köhler gives a long list of the important syntactic units and properties defined within the frameworks of both phrase structure syntax and dependency syntax; this reflects the fact that researchers in both fields have been engaged in some fruitful quantitative studies. In Section 3.2, he defines *quantitation* of syntactic concepts as counting the objects under study, because syntactic analysis investigates only discrete objects. Section 3.4 is a detailed review of the important works on various syntactic phenomena within the frameworks of both phrase structure syntax and dependency syntax, including sentence length, probabilistic grammars and probabilistic parsing, Markov chains, Frumkina's law on the syntactic level, distribution of dependency distance, and distribution of dependency types, and so on. These quantitative models, which have been empirically corroborated with real texts, or sometimes treebanks and dictionaries (of various languages), can be linguistically, cognitively, or functionally interpreted—a rare achievement in the past statistical investigations of language. Apart from the models concerning probabilistic grammars and Markov chains, which have already been widely used in computational linguistics, there are some other works that may also have practical applications in various fields. For example, the mathematical model of sentence length, which describes the probability of neighboring length classes as a function of the probability of the first of the two given classes, may contribute to practical applications such as text classification and the measurement of text comprehensibility, and so forth. The frequency studies of word and syntactic constructions have obtained many results useful for language teaching, the construction of parsing algorithms, and the estimation of effort of (automatic) rule learning, and more. The syntactic studies on Frumkina's law, which is concerned with the number of text blocks with x occurrences of a given syntactic element or category, may benefit certain types of computational text processing if specific constructions or categories can be differentiated and found automatically by their particular distributions. One advantage of QL is that all its findings are mathematically formulated and linguistically interpreted, which at least makes it possible to be used in constructing models necessary for computational linguistics.

In Chapter 4, Köhler introduces his efforts to build a real "linguistic theory," for he believes that "there is not yet any elaborated linguistic theory in the sense of the philosophy of science" (page 21). Building such a theory begins with "plausible hypotheses," which may become laws when sufficiently attested and may then be further integrated into a coherent system. This is the process of setting up a scientific theory, as succinctly summarized in the title of this chapter: "Hypotheses, Laws, and Theories."

The first section of this chapter shows the first step toward a scientific linguistic theory, the process in which "plausible hypotheses" are deduced, interpreted, and empirically attested before finally becoming laws. In Section 2, Köhler introduces the foundation of his synergetic linguistics, which views language as a dynamic, self-organizing, and self-regulating system where the so-called *enslaving principle* and order parameters are the crucial elements. On this basis, the author builds a synergetic syntactic model in Section 4.2.7 with certain modeling principles. Within the framework of phrase-structure syntax, eight properties of syntactic constructions and four inventories are chosen to build this model, which are linked together by laws resulting from the verified hypotheses and subject to the regulation of some order parameters.

Quantitative linguistics, which is an unfamiliar field of study for many linguists, depends heavily on real texts and mathematical tools. Therefore we believe it is worthwhile to briefly clarify the differences and the relations between QL and corpus linguistics, on one hand, and between QL and computational linguistics, on the other hand. In comparison with QL, corpus linguistics is in fact more of a research methodology rather than an independent linguistic discipline, reflecting a shift of focus from competence to performance, from introspection to empirical study. This makes both the common ground and the difference between corpus linguistics and QL, which aims to quantitatively and mathematically explore, on the basis of real texts and treebanks, the fundamental laws governing the structure and evolution of language, and integrate them into a systematic theory capable of explanation and prediction.

Computational linguistics (CL) is an interdisciplinary field investigating the structure of natural languages from a formal, mathematical, and computational point of view. Compared with QL, CL seems to be more interested in research that can have direct applications in such fields as language understanding, language generation, machine translation, and so forth, than in the explanation of language structure, operation, and evolution. Traditionally, the mathematical models used in CL are derived from the linguistic theories via the process of formalization. Due to the limitations of the qualitative mathematical models, however, the statistical models, most of which so far fail linguistic interpretations, have recently become dominant in the field of computational linguistics. That is perhaps why Shuly Wintner (2009, page 641), in a "Last Words" article in this journal, called for "the return of linguistics to computational linguistics," implying that the advances in the field of CL may ultimately lie in the advances in the understanding of language itself.

Wintner's appeal does reflect the present situation of computational linguistics: a discipline in which many works are heavily oriented towards engineering and weakly grounded in linguistics. The formal linguistic models seem now outshone by the purely statistical paradigms, as the result of their inadequacy in processing real-world languages. Of course, this means no renouncement of the value of the traditional formal models. But it is obvious that considerable updating and enrichment are necessary for these formal models if they are to play significant roles in the future. In this regard, QL, which has a solid linguistic foundation, may help by providing quantitative cues (which are linguistically interpretable) to improve the performance in NLP, as has been illustrated in the case of probabilistic grammar that ingeniously integrates quantitative, statistical devices into qualitative linguistic models. QL has provided some useful models for CL. And it is reasonable to believe that it will continue to do so in the future, though some of its achievements seem now not ready to be directly used in CL.

But perhaps QL can do more than that. Though QL has not yet drawn much attention from computational linguistics, it is potentially a proper answer to Wintner's call for the return of linguistics to computational linguistics. CL aims to replicate in computers the patterns of human language behavior, whereas QL endeavors to mathematically and quantitatively reveal the laws and the principles that govern human language behavior—this is a relation between theory and practice. The success of a scientific discipline is usually based on precise models that are well-grounded in profound understanding of the object of study. Computational linguistics is no exception. Two features of QL are hence noteworthy. One is that it aims to explain, within a certain linguistic framework, the operation and the evolution of language by uncovering the systematic cognitive and functional regulations that underlie human languages. The other is that it tries to mathematically model, with systematic and precise quantitative laws, these regulations and the resulting mechanism of language. In view of these two features, we

believe that the success of QL will somehow and somewhat boost the studies in the field of CL and that the communication between QL and CL is and will be not only possible but also mutually beneficial. This is why we hold that it is worthwhile to recommend this book to researchers in the field of computational linguistics, a book presenting a panorama of QL in general and discoveries on the syntactic level in particular.

Reference

Wintner, Shuly. 2009. What science underlies natural language engineering? *Computational Linguistics*, 35(4):641–644.

Chunshan Xu is an assistant professor at Anhui University of Architecture, China. His research interests include syntactic parsing, syntactic complexity, and quantitative syntactic study. Xu's e-mail address is adinxu@yahoo.com.cn. *Haitao Liu* is a Qiushi distinguished professor of linguistics and applied linguistics at Zhejiang University, China. His current research interests include computational and quantitative linguistics, dependency syntax, and linguistic complex networks. Liu's address is School of International Studies, Zhejiang University – CHN-310058, Hangzhou, China; e-mail: lhtzju@gmail.com.