Department of Human Studies
University of Salerno Via Giovanni Paolo II
132 – Fisciano (SA) 84084, Italy
[ciacobini@unisa.it]

Quantitative historical linguistics: A corpus framework. By Gard B. Jenset and Barbara McGillivray. Oxford: Oxford University Press, 2017. Pp. xiii, 229. ISBN 9780198718178. $88 (Hb).

Reviewed by Dirk Geeraerts, *University of Leuven*

The early twenty-first century has witnessed a major shift toward quantitative approaches in the methodology of linguistics. Specifically, whereas quantitative methods have long been a staple of sociolinguistic and psycholinguistic research, the past two decades have seen their expansion toward descriptive and theoretical grammar. In usage-based approaches to language in particular, like cognitive and probabilistic linguistics, a 'quantitative turn' has occurred that applies the statistical testing of hypotheses to data derived from text corpora. The central inspiration for Gard B. Jenset and Barbara McGillivray's book is the observation that this turn toward quantitative corpus studies has not yet penetrated historical linguistics to the same extent as some other subfields of linguistics. It accordingly sets out to introduce 'the framework for quantitative historical linguistics'. The seven chapters fall roughly into two parts. In Chs. 1 to 3, a general argumentation in support of quantitative historical linguistics is developed, whereas Chs. 4 to 7 deal with the implementation of the ensuing program. The discussion of 'why' thus leads naturally to a discussion of 'how'.

Two threads run through the first part of the text: a specification of the kind of quantitative historical linguistics that the authors intend to propagate, and an argumentation in favor of the model in question. Important features of this argumentation are a description of the actual situation in historical linguistics and a conceptual defense of the approach against potential objections. Organizationally, Ch. 1 introduces both threads, Ch. 2 develops the first thread, and Ch. 3 the second.

With regard to the first thread, the first chapter introduces the authors' notion of quantitative research in historical linguistics by means of a double contrast. On the one hand, quantitative research differs from the conventional use of evidence in historical linguistics that rests on example-based categorical judgments about the existence of specific linguistic phenomena but does not look into probabilistic, distributional data about trends of variation and change of the phenomenon in question. On the other hand, quantitative historical research needs to go beyond raw frequencies, in the sense that the multidimensional nature of language requires a multivariate statistical approach. In the second chapter, this conception is further developed in terms of the distinction between corpus-based and corpus-driven approaches. Whereas the former turn to corpora primarily for illustration and confirmation, the latter use corpus data at two stages of the empirical process: corresponding to the distinction between exploratory and confirmatory statistics, quantitative distributional evidence is initially used to generate hypotheses, and subsequently for testing them.

With regard to the second thread, the text provides quantitative data (appropriately, one could say) to the effect that such a method is less entrenched in historical linguistics than other fields of linguistics. This argumentation rests on a comparison of the 2012 volume of *Language* with six journals with a (not necessarily unique) focus on language change, such as *Diachronica*, *Folia Linguistica Historica*, and *Language Variation and Change*. As an explanation for the observation that historical linguistics seems to be lagging behind, the book invokes early negative experiences with glottochronology, plus the influence of structuralist and generative theories (though this is of course a factor that is not specific to historical linguistics). At the same time, it is demonstrated how the rise of quantitative linguistics goes hand in hand with the growing availability of electronic corpus materials—a trend that obviously creates an opportunity for historical linguistics just as for the other branches of linguistics.

Next to the 'the time is ripe, we shouldn't lag behind' argument, the plea for quantitative corpus research in historical linguistics includes a 'nothing is wrong with it' type of argumentation, in the form of a systematic rejection of potential objections. Section 3.7 skillfully refutes counterarguments from convenience, from redundancy, from scope limitations, from principle, and from pseudoscience. Crucially, it is argued that a quantitative approach is not incompatible with a categorial conception of linguistic structure. Although approaches that build variability into their core conception of language (roughly, all usage-based frameworks) are more prone to adopt quantitative methods than formalist approaches, there is nothing that prevents a quantitative perspective on the distribution of linguistic phenomena that are defined formally and categorially rather than probabilistically and continuously. In this respect, the framework defined in the book under review is explicitly presented as a theoretically neutral one.

The second half of the book goes into detail about the main dimensions of quantitative corpus research in historical linguistics: corpora, resources, and quantitative analysis. Ch. 4 deals with issues of historical corpus compilation, from both a philological and a technical point of view. Topics covered include markup languages, tokenization and part-of-speech tagging, and annotation standards. Ch. 5 discusses how other computational resources such as electronic dictionaries, treebank-based valency lexicons, or geographical and historical named-entity-based databases can be used alongside corpora, and how a linked data approach can help to integrate the various resources. Ch. 6 introduces multivariate statistical techniques (specifically, mixed-effects regression models) with two case studies, one on the cooccurrence of Latin spatial preverbs and specific argument patterns, and another on the rise of the existential *there* in Middle English. Ch. 7 presents a summary of the framework developed in the book, and illustrates it with a further case study on the variation between the third-person verbal endings *-s* and *-th* in Early Modern English.

Overall, this is a richly documented, well-informed, and convincingly argued book. It can be greatly recommended—but to whom exactly? One critical note that may indeed be formulated involves the intended audience of the book. Although the overall organization, going from the defense and the formulation of a specific research program to its realization, makes perfect sense in principle, one does get the impression that the two parts of the book are aimed at different audiences. The defense of a quantitative turn for historical linguistics in the first part seems to be directed primarily at historical linguists who are representative of the example-based methodology denounced in the first chapter and who are most likely unfamiliar with basic techniques of corpus linguistics and quantitative analysis. The second part, by contrast, with its detailed information about available corpora, resources, and tools, will be primarily useful for scholars who already master a firm basis of computational and statistical skills. To put it another way, traditional historical linguists who could be converted by the stringent methodological argumentation in the first part may need to cross a corpus-linguistic training gap before they can actually profit from the second part. Conversely, (synchronic) linguists who have already taken the quantitative turn may well appreciate the analytical systematicity of the first part but may not find many new insights there. In this respect, publishing the two parts of the book separately might have been an interesting alternative. The argumentative part could have been expanded with further case studies establishing the superiority of a quantitative corpus approach, and with a further exploration of the current state of historical linguistics (for instance, in terms of the representation of the quantitative corpus approaches in textbooks and teaching curricula). Similarly, the second part of the book could have been developed into a straightforward textbook-type publication, creating room for additional topics (like an overview of the state of art of quantitative studies in various subfields of historical linguistics) or an expansion of the languages covered beyond the current focus on English and the classical languages. As it stands, the book hesitates somewhat between a fundamental and a practical perspective. That does not make it less valuable, but different parts of the book will be valuable to different audiences.

Department of Linguistics
University of Leuven
Blijde Inkomststraat 21 / 3308
3000 Leuven, Belgium
[dirk.geeraerts@kuleuven.be]