[Review of:] **Gard B. Jenset, Barbara McGillivray.** ***Quantitative historical linguistics: A corpus framework.*** Oxford: Oxford University Press, 2017. 256 p. ISBN 9780198718178.

Dmitry S. Nikolaev

Russian State University for the Humanities, Moscow, Russia; Russian Presidential Academy of National Economy and Public Administration, Moscow, Russia; Stockholm University, Sweden; dnikolaev@fastmail.com

Acknowledgements: Work on the paper was supported by the project "Texts and practices of folklore: Typology, semiotics, new research methods" at the Russian State University for the Humanities.

Дмитрий Сергеевич Николаев

Российский государственный гуманитарный университет, Москва, Россия; Российская академия народного хозяйства и государственной службы, Москва, Россия; Стокгольмский университет, Швеция; dnikolaev@fastmail.com

Благодарности: Работа написана при поддержке проекта РГГУ «Тексты и практики фольклора: Типология, семиотика, новые методы исследования».

DOI: 10.31857/S0373658X0008306-5

The publication under review is a comparatively rare specimen in contemporary linguistics: it is essentially a book-length argument in favour of a particular approach to doing historical-linguistics research. The authors aim "to introduce the framework for *quanitative historical linguistics*, and to provide some examples of how this framework can be applied in research" (p. 1; emphasis in the original); and then they do precisely this. Along the way, however, they also spend a great deal of effort to persuade the reader that their framework is actually the best possible way of doing historical linguistics and to refute alternative takes on the matter.

Jenset and McGillivray begin their argument by positing that the family of statistical models developed in corpus linguistics must be adopted by the historical-linguistics community. They note that even though historical linguistics is known to be highly "data-centric", "quantitative corpus methods are still underused and often misused in historical linguistics, and an overarching methodological structure inside which to place such methods is missing" (p. 4). The book therefore endeavours to show "what it means to be empirical in historical linguistics research and how to go about doing it." (ibid.).

The authors then pose and resolve several methodological questions, the most important of which are

Why should historical linguistics be corpus-based and quantitative? (Because otherwise it is impossible to reproduce other people's research, properly formulate and refute claims, and compare models.)

and

Why should historical linguistics be probabilistic? (Because rigid symbolic models tend to be vulnerable to linguistic variation and performance factors. Jenset and McGillivray underline that it is possible to adhere to strict symbolic models of grammar on the theoretical level but still investigate their realisations using probabilistic methods.)

The scholars also note that the methods used to analyse corpus data must be adequate to the task. This boils down to the postulates that (i) presenting uncontextualised raw frequencies of occurrence of different phenomena is not enough; and that (ii) as historical-linguistic trends are usually shaped by an array of factors, researchers should use multivariate methods to model them (multivariate models also being useful to directly estimate explanatory power of competing hypotheses).

Jenset and McGillivray then explore a sociological angle. They survey the current state of the art in historical lingusitics by counting the number of quantitative and corpus-based articles

in the latest issues of several historical-linguistics journals. They then compare the proportion of quantitative articles in each journal with the proportion of quantitative articles in *Language*, used as a baseline representing best practices in general linguistics. The scholars note that publications in *Language* tend on average to be more quantitative and empirical in nature than those from historical-linguistics journals and conclude that historical linguisics is still not a truly empirical, data-driven discipline.

They contextualise this issue using the Moore-ian technology-adoption life cycle. In this perspective, the adoption of corpus-based quantitative historical linguistics has reached a perilous "chasm" between the "early adopter" and "early majority" stages. The failure to cross this adoption threshold due to the general community's refusal or hesitance to embrace empirical methods may become lethal to the discipline or at least seriously set back its development.

In order to push quantitative historical linguistics forward at this crucial juncture and propel it over the chasm, in Chapter 2 Jenset and McGillivray propose a new framework in which to conduct research in historical linguistics.

First, they solidify the terminology needed for such a framework. The following are regarded as the foundational terms:

— **Evidence**: things that can be independently observed and verified by different researchers. Evidence can be **quantiative** (i.e. count-based) or **distributional** in nature; both types of evidence should be quantified in a way that makes independent verification feasible.
— **Claim**: any statement based on the evidence, which does not repeat the evidence itself. Claims can be used as constituent elements for making further claims.
— **Probability**. The researchers argue in favour of following the Bayesian approach, where probabilistic statements reflect the degree of their authors' certainty, as this approach "is explicitly made contingent on our knowledge and our argumentation in a manner that is different and better than in the [frequentist] case" (p. 41).
— **Historical corpus**: a machine-readable systematically sampled collection of natural-language texts representative of some state of the language. The scholars note that non-systematic samples, such as collections of examples, can be biased and should not be regarded as corpora.
— **Linguistic annotation scheme**: a consistent way to annotate texts from a corpus.
— **Hypothesis**: a claim that can be empirically verified.
— **Model**: a representation of some linguistic phenomenon derived from statistical verification of hypotheses on corpus data.
— **Trend**: a directional change in the probability of some linguistic phenomenon over time detectable and verifiable using statistical methods on corpus data.

Then the scholars state several guiding **principles**:

1. **Consensus**: "[T]o achieve the aim of quantitative historical linguistics research, it is necesssary to reach consensus among those scholars who accept the premises of quantitative historical linguistics" (p. 45).
2. **Conclusions**: "All conclusions in quantitative historical lingusitics must follow logically from shared assumptions and evidence available to the historical linguistics community" (p. 46).
3. **Almost any claim is possible**: "Every claim has a non-zero probability of being true, unless it is logically or physically impossible" (p. 47).
4. **Some claims are stronger than others**: "There is a hierarchy of claims from weakest to strongest" (p. 47).
5. **Strong claims require strong evidence**: "The strength of any claim is always proportional to the strength of evidence supporting it" (p. 48).
6. **Possibly does not entail probably**: "The inference from 'possibly' to 'probably' is not logically valid" (p. 49).

7. **The weakest link**: "The conclusion is only as strong as the weakest premise it builds on" (p. 49).
8. **Spell out quantities**: "Implicitly quantitative claims are still quantitative and require quantitative evidence" (p. 50).
9. **Trends should be modelled probabilistically**: "Quantitative historical linguistics can rely on different types of evidence, but only quantitative evidence can serve as evidence for trends" (p. 50).
10. **Corpora are the prime source of quantitative evidence**: "Corpora are the optimal sources of quantitative evidence in quantitative historical linguistics" (p. 51).
11. **The crud factor**: "Language is multivariate and should be studied as such" (p. 51).
12. **Mind your stats**: "Quantitative analyses of language data must adhere to best practices in applied statistics" (p. 52).

The scholars also specify what they consider to be necessary elements of the presentation of data and analyses in research papers (such as references to resources used, the size of the corpus, the desciption of the annotation schema, details of different analyses performed including those that did not lead to the desired results, etc.). They argue for making data and code available on dedicated resources such as Figshare or Github and for publishing datasets separately in specialised journals such as *Scientific Data* or *Research Data Journal for the Humanities and Social Sciences*.

Finally, they discuss the possibility of combining the data-driven approach with theory-based approaches, which tend to produce categorical, non-probabilistic claims. On one hand, they note that theoretical research can be used to formulate testable hypotheses and that, generally, "exploratory approaches to historical linguistics analyses need access to domain knowledge and need to be theoretically grounded" (p. 63). On the other hand, however, exploratory analyses can be used to let the model emerge from the data.

Chapter 3 is almost purely argumentative. Jenset and McGillivray trace the history of several types of quantitative techniques in historical linguistics and point out that even though glottochronology crashed and burned this does not disqualify the whole enterprise. They survey several classes of counter-arguments levelled against quantitative historical linguistics (impracticality, redundancy, limited scope, general irrelevance) and refute them one by one saying, essentially, that one can do almost all kinds of historical-linguistics research using data-driven quantitative methods, and what one cannot do in this way one probably should not do at all.

Chapter 4, by contrast, is entirely practical. It demonstrates the basic structure and typical building blocks of corpus annotation on the part-of-speech, syntax, semantics, and pragmatics levels. It includes listings of XML- and table-formatted text excerpts, encoded parse trees from the Early Modern English Treebank and the Latin Dependency Treebank, and some other examples. It also contains a valuable list of existing historical corpora of different languages.

Chapter 5 builds on this overview in order to show how to combine different datasets into collections of "linked data" and how to build resources on top of other resources (corpus-driven lexica being a prominent example). Problems of working with lower-quality linguistic resources such as collections of raw archival texts are also discussed.

Chapter 6 presents a sample of statistical techniques that Jenset and McGillivray consider most appropriate for historical-linguistics research. In practice, these techniques boil down to multivariate mixed-model generalised linear models (basic linear regression and logistic regression) and multiple correspondence analysis for multivariate categorical data. Two case studies based on these methods are presented: an elucidation of the factors influencing the argument structure of Latin prefixed verbs and an analysis of the rise of existential *there* in Middle English.

Chapter 7 is meant to present a piece of research conducted wholly in accordance with the proposed framework. In order to study the factors behind a shift in the English verbal morphology — the third person singular ending *-(e)s* replacing *-(e)th* as the dominant variant in the time period from 1500 to 1700 CE, — the scholars go through the following steps:

1. Relevant sentences were extracted from the Penn-Helsinki Parsed Corpus of Early Modern English treebank using a Python script.
2. Present-tense verbs were lemmatised in order to compute lemma frequencies for three sub-periods (1500–1569, 1570–1639, 1640–1710).
3. Data were collected into a data-frame format suitable for multivariate-regression analysis.
4. Several exploratory analyses were conducted showing the prevalence of different endings in different time periods and their dependence on particular combinations of values of categorical variables (such as the gender of the author and the phonological context).
5. A series of mixed-effects logistic regression models estimating the probability of switching from *-(e)th* to *-(e)s* given different sets of fixed-effects predictors with genre used as a random effect were fit for each sub-period (the model for the whole corpus turning out to be ill-behaved), and the best model was selected using binned-residual plots.

Coefficients of the three resulting models were then inspected in order to validate or refute some of the claims presented in the literature as to the factors that influenced the shift. The data and the code are duly available on the Github repository https://github.com/gjenset, which makes it possible to reproduce the analyses and try out alternative approaches. The scholars finish the chapter and the book by reiterating their belief in the usefulness of corpus-based probabilistic approaches to empirical research in historical linguistics and express hope that the adoption of a common statistical framework may improve cross-discsiplinary communication in the wider field of the study of language.

Given that the author of this review is a quantitative linguist himself, it is hard for him to assess the crucial merit of the book: its potential to bring quantitative corpus-based historical linguistics to a wider audience and to trigger a paradigm shift in the discipline. I can try, however, to check the internal coherence of the framework and the argumentation as well as the overall quality of the presentation.

Firstly, it must be pointed out that some parts of the framework seem unnecessary complicated and dogmatic. This is mostly due to the use of widespread scientific terms in a somewhat surprising sense.

For instance, the definition of **trend** as a "directional change in the probability of some linguistic phenomenon over time detectable and verifiable using statistical methods on corpus data" is counterintuitive: probabilities cannot be observed, and what researchers detect and mostly have to reason about are relative frequencies. Moreover, while relative frequencies can be straightforwardly computed, probabilities are model dependent. For example, when a negative trend is observed that culminates with some phenomenon being completely lost, we can note that after some point in time its relative frequency is zero. The probability we assign to it, however, will be different dependent on which priors we use in a Bayesian setting, advocated by the authors, or will not even be well defined in a frequentist one. The issue of **verification** of this probability is of course even more thorny.

A similar issue arises due to the demand that a **hypothesis** must necessarily be empirically verifiable. It may be pointed out that conjectures about the past cannot be empirically verified in the strict experimental sense. We can only amass observations that do not contradict them. It would do better to demand that hypothesis be **empirically falsifiable** (in the sense that they have consequences that make particular patterns in the data impossible or highly unlikely), but even that demand can hardly be baked into the term's definition.

The discussion of the notion of **consensus** is even more troubling in that it extends beyond the mere methodological common ground, which is indeed indispensable: "[T]he effort of creating consensus without a common ground of fundamental principles is probably going to be futile" (p. 46). However, the effort is necessarily bound to be futile anyway; therefore, the researchers make the following qualification: "[T]he principle cannot be understood as an injunction to *achieve* consensus, only to *seek* it, since consensus by definition must involve more than one researcher" (ibid.; emphasis in the original). It is hard to understand why more than one

researcher are unable to achieve consensus, but the demand to actively seek it is clearly counterproductive, if only for the well-known fact that when a measure becomes a target it ceases to be a good measure. Whatever consensus there is in the research community must be a logical consequence of common methodology and data. Finally, the thesis that "[t]o challenge the consensus is to seek its amendment" (ibid.) makes the whole discussion void of substance: whatever one does, the consensus is inescapable.

Another related idea proposed by Jenset and McGillivray, that "on the whole we must assume that [experts'] beliefs and claims are accurate, *given the current state of knowledge in the field*" (ibid.; emphasis in the original), is only relevant either for the consumers of scientific knowledge (it is safer to trust the established scholars) or for the people who have to promulgate radical new ideas (it is better at least to pretend that you are trying to modify the currently held opinions and not to show that they are altogether worthless). It is hard to see how it can be applied in the actual process of doing empirical research and presenting its results in a compelling way.

The discussion above may look like nitpicking, but one must bear in mind that the high-level ideas presented in Chapter 2 are the cornerstone of Jenset and McGillivray's original contribution. The authors do not propose new concrete methods for doing historical linguistics but try to provide the scholars with a general vector, and it behooves us to seriously test its validity. An uncharitable observer might also point out that some components of this vector are lifted directly from a treatise on Bayesian approach to history by R.C. Carrier, which is referenced 11 times in Chapter 2. This should not necessarily be troubling in itself, after all Carrier may be an underappreciated source of relevant methodological knowledge. However, having declared their strict Carrier-inspired adherence to the Bayesian approach to truth and probability, the scholars then suddenly abandon it completely and in the rest of the book use frequentist methods, making the framework somewhat incoherent.

Jenset and McGillivray find themselves on a much firmer ground refuting claims about limitations and general inapplicability of quantitative historical linguistics in Chapter 3, which makes for an amusing and satifying read. Hardened detractors of quantitative approaches to historical linguistics are unlikely to be swayed by the argument, but to convert them would be a very tall order.

Turning to the general structure of the book, one must note that, unfortunately, it does not seem to hang very well together. The introductory Chapter 1 includes a good share of the argumentation presented later in greater detail in Chapters 2 and 3. Given that the point the scholars are trying to make is neither very deep nor strictly original, readers may be advised to skip these chapters altogether and go straight to Chapter 7, where the framework is presented in the form of a checklist in the section 7.2 "Core steps of the research process", followed by a detailed case study.

Chapters 4 and 5 are unnecessary for the main argument; however, they present a very useful overview of the approaches to corpus annotation and of the available resources in this area. The discussion of the structure of XML markup looks rather out of place after high-level methodological discussions ("Within the scope of the <body> tag, we see two instances of the tag <text>, which indicates that <text> is nested inside <body>. The text in double quotes contained in the tag <text> is an attribute…", p. 106), and one has to assume that after giving a philosophical argument in favour of the quantitative approach the scholars then wanted to give a "taste" of it by exposing some technical minutiae of the corpus methodology.

Chapter 6 is also introductory in nature and serves as a short tutorial on statistical methods in historical linguistics based on two case studies. The analyses in the first one, on Latin prefixed verbs, are not detailed, and it looks like it was used as a pretext to quickly introduce linear regression, logistic regression, and multiple correspondence analysis. The second one, on the rise of existential *there* in Middle English, is rather exhaustive and provides a nice example of the application of logistic regression to linguistic data. The choice of fixed and random effects is discussed, and a careful analysis of the resulting models using R^2, Harrel's C index, and binned-residuals plots is presented. The reader, however, will then discover that the same types of models

(logistic regression and MCA) are applied again in very much the same way to a different dataset in an even more detailed case study in Chapter 7. Having more worked-out examples is never a bad thing, but this repetition of similar material under different rubrics is definitely puzzling.

Overall, *Quantitative Historical Linguistics* is hardly a book to be read from beginning to end. Go to Chapters 1 and 7 for the framework as a practical guide; add Chapter 6 if you need pointers on statistics (in addition to definitions and examples, it contains many useful references); take stock of the existing resources in Chapters 4 and 5 and of the polemics around the field in Chapter 3. And maybe try out Chapter 2 if you are unsure as to what "claim", "evidence", or "corpus" is.

Получено / received 29.08.2019 *Принято / accepted 17.09.2019*

Вопросы языкознания
научный журнал Российской академии наук
(свидетельство о СМИ ПИ № ФС77-77284 от 10.12.2019 г.)

Оригинал-макет подготовлен С. С. Белоусовым

Адрес редакции: 119019, Москва, ул. Волхонка, 18/2,
Институт русского языка им. В. В. Виноградова РАН, редакция журнала «Вопросы языкознания»,
тел.: +7 495 637-25-16, e-mail: voprosy@mail.ru

Подписано к печати 30.01.2020 Формат 70×100 1/16 Уч.-изд. л. 15
Тираж 340 экз. Зак. 4/1а Цена свободная

Учредители: Российская академия наук, Институт русского языка им. В. В. Виноградова РАН

Издатель: Российская академия наук

Исполнитель по контракту № 4У-ЭА-040-19 ООО «Интеграция: Образование и Наука»
105082, г. Москва, Рубцовская наб., д. 3, стр. 1, пом. 13–14

Отпечатано в ООО «Институт информационных технологий»