Tanja Säily: Sociolinguistic Variation in English Derivational Productivity. Studies and Methods in Diachronic Corpus Linguistics. Helsinki: Société Néophilologique 2014 (Mémoires de la Société Néophilologique de Helsinki Nr. XCIV). VII + 284 Seiten.

Lea Kawaletz

In recent decades, it has become almost impossible to investigate the productivity of word formation processes through corpora without taking into account statistical measures of some kind. In the field of historical linguistics, however, this has been problematic due to small corpus sizes and the lack of suitable methods. This book proposes such methods, making a strong point for data-driven approaches while exploring the diachronic and sociolinguistic variation of productivity in word formation. Specifically, the study focuses on the native suffix *-ness* and its Romance-based counterpart *-ity*, analysing data from 15th to 18th letters, 18th century trial proceedings, and the BNC. In order to obtain valid results, the author combines statistical and visualization methods which are already established in other research traditions, and complements these with original software. Apart from presenting interesting findings in the field of historical sociolinguistics, she provides methodological insights which are relevant for research based on historical (or otherwise problematic) corpora, especially for but not restricted to those interested in productivity and word formation.

The book consists of three parts. Part I summarizes the theoretical background and previous research as well as the statistical methods employed, and does so in detail but not excessively. The frame is set by five well-motivated research questions, four of which are methodological:

1. Is there sociolinguistic variation and change in the productivity of *-ness* and *-ity* in the history of English?
2. How can we study productivity in small corpora which contain a great deal of spelling variation?
3. How can we study variation and change in corpora which may not be completely comparable over time and across genres?
4. Are the productivity measures proposed in previous research valid in and applicable to sociolinguistic data of this kind?
5. What are the requirements for a usable tool for studying variation in productivity in data of this kind?

Given that the book is a cumulative dissertation, Part II is comprised of six distinct papers, each dealing with one or more of these research questions. For instance, chapters 7 and 8 concentrate largely on research question (3) and thus on methodology, while chapter 11 has a linguistic focus and tackles research question (1) by applying the methods described in the preceding chapters to 18th century letters. It is obvious that this dissertation is quite complex, basically dealing with any problem you might encounter when trying to answer research questions such as (1) by consulting historical corpora. However, with Part I setting the scene and the otherwise independent chapters arranged in a way that they build logically upon each other, the reader can keep track of the big picture. Part III then ties up all loose ends, evaluating both the linguistic results and the

draft do not quote

methods used. With this frame provided by parts I and III the reader can also get an overview of the whole subject matter while being able to quickly identify which chapters are relevant to read up on specific details. The book is completed by two appendices: a very helpful glossary of statistical terms and a list of sociolinguistic parameters applied.

The author makes a number of valuable methodological observations. For instance, she shows that hapax-based measures of productivity are not reliable when applied to small corpora, and proposes a combination of statistical methods and visualization to overcome this problem. Thus, the rather strict method of resampling by permutation testing is combined with type and hapax accumulation curves, which allow the researcher to discover near-significant tendencies visually. In collaboration with computer scientists, the author has developed an open source program which not only computes the upper and lower bounds of these accumulation curves by means of Monte Carlo sampling, but also provides a built-in measure of statistical significance. The author weighs the pros and cons of every method applied and conclusion reached, and argues convincingly for the usefulness of introducing unconventional methods into a traditional field. Apart from the open access software, complete analysis files are also available online, aiding in the transparency of her conclusions and offering others the possibility to build on her research. Her findings are relevant primarily for diachronic studies, but they are also applicable to synchronic research faced with similar problems, such as large but messy corpora compiled from the Internet.

Given the solid methodology it can be safely said that the author makes the best out of the available historical data and presents a number of interesting linguistic findings: Overall, she finds an influence of social factors (gender, social rank and participant relations) on the use of both *-ness* and *-ity*. Furthermore, diachronic change in productivity can mostly be asserted for *-ity*, while the use of *-ness* has remained very stable. To pick out one especially interesting finding with regard to social influences, the author shows that gender plays a central role: Throughout history and up until today, women appear to use *-ity* less productively than men, while there is no gender difference with *-ness*. Why this is the case, however, remains to be seen.

Overall, both the results and the methodology are convincing, and my points of criticism are few: First, although the author takes great care to introduce all statistical methodology and terminology to the uninitiated reader, some terms remain undefined (e.g. 'NP-complete problem'). Second, the social variables investigated are not always completely transparent. Thus, 'domicile', a variable investigated in chapter 6 which turns out as not significant, is not defined there, nor introduced in the background section of the book. It could be interpreted as referring to the parameter 'Place of birth: main domicile' which is listed in Appendix II. However, the subcategory 'other', which figures in chapter 6, is not mentioned there, so that it remains unclear what exactly is meant by this term.

These marginal issues hardly diminish the overall good quality of this book. I believe that anyone basing his/her research on historical (or comparably problematic) corpora can greatly benefit from this work, especially when dealing with word formation processes and/or the influence of sociolinguistic factors. Although this book is heavy on statistics, every method is so well explained that, in combination with the provided glossary of statistical terms, anybody who has a basic knowledge of statistics can follow with ease.

draft do not quote

References

The British National Corpus (BNC). 2007. Version 3 (BNC XML Edition). Distributed by Oxford University Computing Services on behalf of the BNC Consortium. http://www.natcorp.ox. ac.uk/ (accessed 01 April 2015).

draft do not quote