Korpuslinguistik und interdisziplinäre Perspektiven auf Sprache

Band 5

Jost Gippert / Ralf Gehrke (eds.)

Historical Corpora

Challenges and Perspectives

narr VERLAG

Jost Gippert / Ralf Gehrke (eds.)

Historical Corpora

Challenges and Perspectives

narr
VERLAG

Bibliografische Information der Deutschen Nationalbibliothek

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.dnb.de abrufbar.

Internet: www.narr.de
E-Mail: info@narr.de

Redaktion: Melanie Steinle, Mannheim
Layout: Andy Scholz, Essen (www.andyscholz.com)
Printed in Germany

ISSN 2191-9577
ISBN 978-3-8233-6922-6

Contents

Preface

Historical Corpora – Challenges and Perspectives

The present volume contains most of the papers read at the international conference "Historical Corpora 2012", which was hosted by the LOEWE Research Cluster "Digital Humanities" of the State of Hesse at the University of Frankfurt on December 6-8, 2012. All in all, the conference comprised 27 individual papers, selected out of 45 applications in a meticulous peer-reviewing process by an international board, plus five keynote speeches, three of which have been kindly provided for publication in the present volume. It goes without saying that nearly all of the materials have been duly elaborated in the meantime in order to bring this volume up-to-date.

Both in arranging the conference program, which can be accessed on www.dhhe.de/historical-corpora, and in preparing the present volume, it became clear that the very title "Historical Corpora" opens a huge range of possible interpretations and, accordingly, topics, thus making it difficult beforehand to find a consistent order for the individual contributions. This is true, first of all, of the notion of "historical" itself. In many of the papers, this was taken to refer to older stages of given languages, be they "ancient", "old", "medieval", or just not contemporary. In other cases, it involved a perspective across different stages in the history of a language; for this perspective, which is mostly concerned with linguistic change in time, the term "diachronic" would be more appropriate in order to distinguish it from a consideration of individual stages in a language's past, which may be as "synchronic" as a study of contemporary language use. As the papers collected in the present volume show, the difference between these principles has a big impact on the structuring of corpora, their contents and their sizes. It may suffice here simply to mention a few points.

a) Contemporary corpora can be multimodal (comprising written, spoken, audiovisual, elicited and other types of data); the same may be true for historical corpora both under a synchronic and a diachronic perspective, but only if the time-depth in question does not exceed 120 years, given that the oldest recordings of spoken language hardly antedate the year 1900.

b) Contemporary corpora can always be thematically determined, provided the language in question is really "alive"; the same may be true of historical corpora both in a synchronic and a diachronic perspective, but only to a

certain extent, given that the further we go back in history, the fewer genres, registers, and text types we can expect to be represented in what has come down to us from the history of a given language.

c) For the same reasons, only contemporary corpora can be truly "balanced" in the sense that they cover all modes, registers, genres, etc. of a given language to an equal extent. The balancing of historical corpora is always restricted by the materials that have survived.

d) Contemporary corpora may be kept linguistically homogeneous, excluding, for instance, certain dialectal, sociolectal, or other strata. For historical corpora, this may be attempted as well, but only if they are synchronic; diachronic corpora can never be linguistically homogeneous as they cover linguistic change *per definitionem*.

e) Contemporary corpora may be kept orthographically homogeneous, depending on the consistency of the orthographical rules of a given language. For synchronic historical corpora, this may be attempted, too, but it will again depend on the language in question and the time depth envisaged (as orthography in the modern sense is a rather recent phenomenon); for diachronic corpora, this will mostly be impossible as orthography changes over time everywhere, partly in accordance with linguistic change and partly independently.

f) Contemporary corpora are "open" in the sense of being freely extensible at any time. For all kinds of historical corpora, however, extensibility is limited to what has been preserved, and the further we go back, the less material we can expect to find. Synchronic historical corpora can even be complete in the sense that they cover all the (written!) materials of a certain stage of a given language; an example of this is the TITUS corpus of Old Persian, http://titus.uni-frankfurt.de/texte/etcs/iran/airan/apers/apers.htm, a language that was written in cuneiform script between ca. 550 and 330 B.C. On the other hand, diachronic corpora cannot be complete if the time range to be covered includes contemporary usage.

In addition to the variety of perspectives on "historical corpora" emerging from these preliminary considerations, the contributions to the present volume differ, of course, with regard to the languages under concern. It is true that German – in nearly all its historical facettes – is the most widely addressed among them; however, the range of vernaculars treated extends far beyond

that, across the Romance languages into the Caucasus and from the recent past down into antiquity. Differences also concern the linguistic interests prevailing in the papers, which may focus on syntactic, semantic, pragmatic, lexicological or other phenomena. Beyond that, the program of the conference proves that historical corpora, in all senses of the term, have raised interest meanwhile not only in linguistics but also in neighbouring disciplines such as literary studies, history, philosophy, or theology, and we are delighted that some of the contributions to the present volume reflect views from outside linguistics. And of course, there are also contributions that are practically language-independent, dealing with more general issues of the structure of historical corpora (and the infrastructure required by them).

The arrangement of the papers in this volume tries to take these aspects into account as far as possible. There being no intrinsic principle that would be superior to others, we decided to let ourselves be guided by reader-friendliness, which simply presupposes that papers with related content should be placed close to each other. It goes without saying that there are no value judgements implied in the arrangement of any of the papers.

There are many persons and institutions to whom the editors of the present volume wish to express their gratitude: first of all, the keynote speakers, Gerhard Heyer, Karin Donhauser, Gerhard Lauer, Martin Durrell, and Anthony Kroch, who raised general topics of major interest and thus provided true highlights in a conference program of exceptional breadth and quality; second, the participants who sent their papers in for evaluation in due time and delivered them in Frankfurt, in some cases enduring long and unpleasant journeys; third, the peer reviewers, Pietro Beltrami, Anne Bohnenkamp-Renken, Nils Diewald, Karin Donhauser, Martin Durrell, Gerhard Heyer, Anthony Kroch, Gerhard Lauer, Henning Lobin, Anke Lüdeling, Rosemarie Lühr, Giovanna Marotta, Alexander Mehler, Cecilia Poletto, Andrea Rapp, Henning Reetz, Manfred Sailer, Maik Stührenberg, Marc van Oostendorp, Ulli Waltinger, and Helmut Weiß, who agreed to read and evaluate the papers submitted alongside their many other duties; fourth, the members of the staff of the LOEWE research cluster and the students of the Institute of Empirical Linguistics at the University of Frankfurt who helped us arrange and maintain the conference; fifth, the editors of the series "Korpuslinguistik und Interdisziplinäre Perspektiven auf Sprache – Corpus Linguistics and Interdisciplinary Perspectives on language (CLIP)", Harald Lüngen, Marc Kupietz and Christian Mair, as well as the Gunter Narr Verlag, Tübingen, for kindly accepting the present volume for

publication; sixth, the contributors of the present volume, who were ready to work through their papers again for publication after presenting them at the conference; and seventh, the Hesse State Ministry of Higher Education, Research and the Arts, which enabled us to work intensively on historical corpora and to organise the conference by granting the funding for our LOEWE Research Cluster.

Frankfurt, January 2015 Jost Gippert