Blind data vs. philology
Evidence in historical linguistics

Philipp Obrist
Eberhard Karls Universität Tübingen, Sonderforschungsbereich 441
philipp.obrist@uni-tuebingen.de

This contribution discusses some preliminary theoretical issues on corpus gathering for quantitative analysis of historical language data.

Corpus linguistics uses statistical methods to assure the accuracy of its research results. For example, when investigating a particular grammar issue in a determined corpus, analysis can be repeated on standardised text samples (e.g. 1000 words) until p-scores drop below a pre-determined benchmark, thus limiting the likeliness of an error to a merely statistical value (Biber, Conrad and Reppen 1998).

Due to theoretical and practical difficulties in corpus design, stratified corpora a largely preferred to representative ones, even in synchronic linguistics. If the ideal of exhaustiveness and representativity of corpora is in any case unattainable, this is particularly true for diachronic research. The historical linguist, unlike synchronic linguists, depends solely on a limited range of written documents: he can neither access a practically unlimited range of spoken and printed material, nor make use of the intuitions of native speakers. The consequences of such limitations on linguistic analysis have already been discussed in recent papers (cf. Fischer 2004).

In this contribution, we aim to demonstrate that methodological problems do not only appear on the level of linguistic analysis, but extend to the more basic one of linguistic description. Linguistic research on historical documents implies documentary and intertextual issues which – if universal and applying to any sort of text – are of a much larger scope in historical corpora than in contemporary ones.

Evidence will be given from the syntactical analysis of two texts in Old Spanish, the Poem of the Cid from the 11th/ 12th century and the Libro de Buen Amor, from the second quarter of the 14th century. While the former text is subject to a discussion of authorship and has possibly been partly rewritten over almost one century, the latter shows specific stylistic variation due to its utterly compilatory nature, including mimetic adaptations of different discourse traditions.

We will show that a syntactic analysis which proceeds subsets of data shows significant divergences to an all over analysis of the given texts, moreover, that determined subsets from different texts may show a more homogeneous syntactic structure com-

pared one to another than the texts as a whole. This suggests that philological methods, which orthodox structuralism widely banned from experimental linguistics, should be again taken into account in historical linguistic research.

References

Biber, D., S. Conrad and R. Reppen (1998). Corpus linguistics. Investigating Language Structure and Use. Cambridge University Press, Cambridge.

Fisher, O. (2004). What counts as evidence in historical linguistics? Studies in Language, **28**: 710-740.