Why ‘big data’ might not always be the best data in historical linguistics: a case study on the role of contact in structural changes of mediaeval English

It is widely-known that English underwent many drastic structural changes from Old English (OE) to late Middle English. Also widely-known is that Viking settlement-names at that time were very numerous, but confined to the Danelaw-area in northern and eastern England (cf. e.g. Stenton 1942; Cameron 1969: 75-86, 1971; Richards 1991: 33-36; Loyn 1994: 82-89; Crawford 2003: 59-60). And while it is generally uncontested that English borrowed quite a few and very basic words from Old Norse (ON) (cf. e.g. Hofmann 1955; Peters 1981; Thomason & Kaufman 1988; Townend 2002; Pons-Sanz 2007, 2013; Lutz 2012), there is much discord in the research-community regarding the role of ON/OE-contact in the structural changes mediaeval English underwent: outspoken affirmers of contact-induced or -driven grammatical change (e.g. McWhorter 2002; Townend 2002) face vehement deniers (e.g. Thim 2008, 2012; Lutz 2012).

This paper contributes to overcoming this stand-off by offering a case-study on English gender-development, a particularly puzzling instance of language-change: although more than a dozen studies on English gender-change precipitated since Körner’s (1888) initial treatise, this large body of research has produced “remarkably contradictory results” (Stenroos 2008: 451) regarding pathways of and reasons for gender-simplification. While a number of contributions found only chaos, i.e. absence of patterns (e.g. Lindelöf 1893, Ausbütel 1904, Markus 1988), there seems to be none yet that explicitly views the phenomenon in the context of OE/ON contact.

These contradictory and/or negative results have two likely causes: one is methodological. Large-scale studies typically analyse miscellanies of documents of strikingly dissimilar regional provenance, text-type/genre, topic, and even age (cf. Stenroos 2008: 451), all of which can exert a considerable confounding influence (cf. e.g. Szmrecsanyi 2016), “making it difficult to trace coherent patterns” (Stenroos 2008: 451). Conversely, though to a similar effect, smaller studies that base on a single document have no way of disentangling particularities of the document at hand from general trends and characteristics. The other reason is too narrow a theoretical focus, i.e. investigating only one type of potential predictor – say, phonology – while ignoring all or most others – such as morphological, semantic, cognitive, and contact effects, thus obtaining a partial and distorted picture (cf. e.g. Enger 2013 for an eloquent critique of such approaches).

As historical language-data is inherently limited, and as “[h]istorical documents survive by chance, not by design” (Labov 1994: 11), Labov described historical linguistics as “the art of making the best use out of bad data” (1994: 11). This paper hence suggests to neutralise or at least minimise the confounding influences sketched above by carefully selecting documents comparable in terms of text-type, style, and topic, drawing (where possible) on interdisciplinary evidence to ascertain age and provenance. Dividing the database thus
obtained into a north-eastern, OE/ON-contact-exposed cohort and a south-western, essentially OE/ON-contact-free cohort shows contact to quicken, possibly cause gender-simplification, but also show this process to underlie phonological, morphological, semantic and cognitive factors as well. This shows that the confounding effect of dissimilar documents in large datasets can be minimised by choosing a smaller, but more homogenous database instead.

References:


