Originalveröffentlichung in: Gippert, Jost/ Gehrke, Ralf (Hrsg.):
Historical Corpora. Challenges and Perspectives.–
Tübingen: Narr, 2015. S. 12-33. (Korpuslinguistik und interdisziplinäre Perspektiven auf Sprache 5)

MARTIN DURRELL

'Representativeness', 'Bad Data', and legitimate expectations

What can an electronic historical corpus tell us that we didn't actually know already (and how)?

Abstract

The availability of electronic corpora of historical stages of languages has been welcomed as possibly attenuating the inherent problem of diachronic linguistics, i.e. that we only have access to what has chanced to come down to us – the problem which was memorably named by Labov (1992) as one of "Bad Data". However, such corpora can only give us access to an increased amount ot historical material and this can essentially still only be a partial and possibly distorted picture of the actual language at a particular period of history. Corpora can be improved by taking a more representative sample of extant texts if these are available (as they are in significant number for periods after the invention of printing). But, as examples from the recently compiled GerManC corpus of seventeenth and eighteenth century German show, the evidence from such corpora can still fail to yield definitive answers to our questions about earlier stages of a language. The data still require expert interpretation, and it is important to be realistic about what can legitimately be expected from an electronic historical corpus.

1. Introduction

My primary aim in this paper is to take the opportunity of standing back and taking a look at what we expect from historical linguistic corpora, consider the possibilities they provide and re-assess their inherent limitations, in particular in the light of the kind of caveats which have been voiced eloquently over the years by Rissanen (1989; 2008). These observations will be chiefly based on our recent experience in Manchester over the past few years of compiling a historical corpus of Early Modern German, the GerManC corpus – and in my own case coming relatively new to the whole field of corpus structure, compilation and design.

2. The problem of 'Bad Data'

The obvious starting point is to consider the data of historical linguistics. In effect, historical linguistics has always been based on a corpus, although we haven't always used the term. We have quite simply a body of data which is,

first of all, inherently finite – quite obviously so in the case of older languages like Gothic or Runic Norse. In this respect it is like any body of historical data in that we are wholly dependent on what has chanced to come down to us and we have to make sense of it, interpret it and make inferences from it on the basis of explicit (and hopefully sound) theoretical principles. Lass (1997: 42) summarizes the problem succinctly as follows:

> The past is not directly knowable or independently available to us as such. But it is knowable through inference, which depends on theoretically directed interpretation and evaluation of witnesses, and where necessary on the actual construction of missing witnesses, which then become part of the record.

As Labov (1992) points out there is no avoiding the fact that all historical linguists have a limited set of accidentally preserved "Bad Data", and in the field of corpus linguistics this has been taken up and emphasized by Nevalainen (1999). We cannot control it, nor can we appeal to native speaker intuitions. Essentially, as Lass (1997: 24) says, we cannot reconstruct the past but only "encounter it only indirectly, through theoretical judgements about what we take to be its witnesses", in other words we must simply make hypotheses on the basis of the imperfect data we have in the light of our knowledge about language in general and the particular language at that point in its history.

Now, on the face of it, electronic corpora seem to offer us a way out of this dilemma. In the case of languages like those of medieval and early modern Europe, which are better attested than, say, Gothic, they seem to offer the prospect of affording easy access to an unprecedented amount of data. Instead of spending weeks or months in libraries ploughing through texts hunting for examples of a particular form, construction or vocabulary item, it is all available with a few keystrokes in the comfort of the scholar's own study. As Cantos (2012: 102) says:

> [...] corpus linguistics can fruitfully contribute to overcome the obstacles of the bad data problem; by allowing researchers to process simultaneously almost all the texts that have survived from a given period, corpus linguistics partly solves the fragmentary nature of historical material, and ensures that early varieties can be reliably reconstructed, [...].

However, concealed within this apparently positive claim are a number of very indicative hedges. Apart from the fact that, even with our present technology, it hardly seems a realistic prospect to "process simultaneously almost all the

texts that have survived" from, say, seventeenth century German, to say that "corpus linguistics can fruitfully contribute to overcome the obstacles of the bad data problem" is possibly still some way from overcoming it, and if it "partly solves the fragmentary nature of historical material", the solution can still only ever be partial. The crucial point is that what we have is still written language data which has been preserved by chance. We may be able to access more of it more quickly and more simply, but it still has all the inherent qualities which led Labov to refer to it as "Bad Data". It might not necessarily provide better insights than we already have, or give us a much clearer picture of the language at the particular point in time we are investigating. However much data we have, in historical linguistics, as in any historical discipline, we can only ever be dealing with "Bad Data". An apposite example here would be the recent account by Jones (2009) of the passive auxiliary in the older Germanic languages, notably Old High German. Using extant electronic corpora he was able to propose a convincing and more comprehensive analysis of the distinction in function of the two passive auxiliaries in terms of *Aktionsart* of a kind which had eluded earlier scholars. However, the methods, procedures and theoretical foundations of his account were very much those of traditional philology and historical linguistics (and crucially his expertise in Latin and Greek, as well as in older Germanic), and there is no inherent reason why earlier scholars should not have been able to arrive at the same analysis without the benefit of electronic corpora. Electronic procedures simplified the searches and comparison of the examples, but what was crucial was that Jones (2009) simply asked the right questions within an adequate theoretical framework. On the other hand, where the data are insufficient, we will still lack adequate evidential base for a convincing account. An obvious example would be the still intractable question of whether aspect was a fully functioning grammatical category in the Gothic (and Germanic) verb, similar to Slavonic, with its exponence in prefixation, in particular through the prefix *ga-* (cf. Leiss 1992: 54-71). With or without the benefit of electronic corpora, we can only ever put forward well-founded hypotheses to understand and explain the data we have and try to evaluate them comparatively on the basis of our linguistic expertise.

3. Representativeness in historical corpora

For more recent historical periods, especially after Gutenberg, like that which was the basis of the GerManC corpus of Early Modern German, the amount of available material naturally increases exponentially and it is probably unrealistic to suggest that all the available material could be digitized, and even if that were possible the corpus could then run the risk of becoming unmanageable or inherently skewed. This means that we have to address the familiar issues of size, balance and representativeness (cf. Hunston 2008: 160). A large corpus obviously seems desirable, but with that, two things must be borne in mind. First, as mentioned earlier, however much material is included, we are still only dealing with what happens to have come down to us by chance, and a large corpus cannot solve *per se* the fundamental problem of "Bad Data". Secondly, any corpus is in essence an artefact and entails all the kind of provisos and caveats indicated by Rissanen (2008: 64-67). It is a subset of the language as it existed at a particular time and it cannot answer the kind of questions which we are able to put to living speakers. We must beware of confusing a corpus with "the language" and of assuming that it gives us some kind of access to the grammar of a native speaker. And in this context it is important always to remember that we are dealing with written data, and the relationship between the varieties used in speech and writing may be rather problematic, especially after the development of widespread literacy or a widespread literary culture and the incipient stages of linguistic standardization (cf. Hennig 2009). Nevertheless, it is by no means certain that the assertion by Hunston (2002: 23) that "a statement about evidence in a corpus is a statement about that corpus, not about the language or register of which the corpus is a sample" is wholly tenable, since ultimately we have no real choice in historical linguistics but to extrapolate knowledge about the development of the language from such samples. As Rissanen (2008: 64-67) says though, no corpus, and especially no historical corpus, can be truly representative in a strictly statistical sense. Similarly, Wegera (2013: 64) points out that we can never know precisely what the relationship is between the sample and the language as a whole, and he refers to Köhler (2005: 5) who puts this very clearly:

> Keine Stichprobe kann repräsentative Sprachdaten in dem Sinne liefern, dass in dem in der Statistik üblichen Sinne gültige Schlussfolgerungen auf die Population, auf das „Sprachganze", möglich wären. Kein Korpus ist groß genug, um die Diversität der Daten im Hinblick auf Parameter wie Medium, Thematik,

> Stilebene, Genre, Textsorte, soziale, areale, dialektale Varietäten, gesprochene vs. geschriebene Texte etc. repräsentativ abzubilden. Versuche, das Problem durch Erweiterung der Stichprobe zu lösen, vergrößern nur die Diversität der Daten im Hinblick auf die bekannten (und möglicherweise noch unbekannte) Variabilitätsfaktoren und damit die Inhomogenität.

If Leech (1991: 27) says that a corpus is representative if "findings based on its contents can be generalized to a larger hypothetical corpus", that ultimately begs the question of how we can ever be in a position to establish how that hypothetical generalization can be carried out. Nevertheless, as Leech has said more recently (2007: 143-144), the debate about balance, representativeness and comparability might lead people

> [...] to reject these concepts as being ill defined, problematic and unattainable. My attitude is different from this. [...] these are important considerations, and even if we cannot achieve them 100 per cent, we should not abandon the attempt to define and achieve them. We should aim at a gradual approximation to these goals, as crucial desiderata of corpus design. It is best to recognise that these goals are not an all-or-nothing: there is a scale of representativity, of balancedness, of comparability. We should seek to define realistically attainable positions on these scales, rather than abandon them altogether.

Nevertheless, the question still remains of how criteria might be established to assist us in seeking to define these positions.

4. The GerManC corpus

For historical periods after the introduction of printing by the use of movable type in Europe, the structure and design of any corpus will ultimately be determined by underlying research questions, i.e. what does the researcher want to know about the particular language (or language variety) at that stage in its development. In the case of the GerManC corpus, the primary objective was to provide a research resource which could be exploited to trace the process of standardization in German between the (conventionalized) end of the Early New High German (ENHG) period in 1650 and the relatively final stages of the process of codification at the end of the 18th century. For the period up to 1650 the *Bonner Frühneuhochdeutschkorpus* (Bonn corpus of Early New High German) is available, but standardization was in that period still in the process of selection of variants, and codification had hardly begun. Thus, much more variation still existed in the mid-seventeenth century than, say, in English or

French, and, characteristically for the history of German, this variation had a marked regional dimension. Despite any caveats that one might have about representativeness, it was obviously desirable to have available an electronic corpus which would provide as broad and balanced a picture as possible of the language during this period. The selection of texts was thus modelled on the notion of representativeness developed by Biber for the ARCHER corpus of English (ARCHER = "A Representative Corpus of Historical English Registers", cf. Biber/Finegan/Atkinson 1993). An additional and important reason for this decision was the fact that David Denison, a colleague in the Department of English Language and Linguistics at the University of Manchester, was co-ordinating the team developing further versions of ARCHER, and a number of postgraduate students had been investigating comparative developments in English and German, for example Auer (2009) and Storjohann (2003). They had used the ARCHER corpus as their resource for historical material in English, but were hampered by the lack of a comparable systematic data collection for German. Using ARCHER as a model we thus considered that at least a greater degree of representativeness could be achieved by including in the first place a wider span of registers. These could not be identical with those of ARCHER because of differences in the types of texts which have been preserved for German, but the following registers were found to provide sufficient material: newspapers, narrative prose (not only fiction), drama, legal texts, sermons, personal letters, scientific texts and texts on humanities-based topics. The time-span of 150 years was divided into three sub-periods of 50 years (following the model of the Bonn Early New High German corpus), and given the continued importance of regional variation in German, the German-speaking area was divided into five major regions: North, West Central, East Central, South-West (including Switzerland) and South-East (including Austria). This proved adequate to cover the level of variation still present in the language. For the completed corpus three 2,000-word text samples were selected for each subdivision in terms of register, sub-period and region, and the whole thus contains nearly a million words. This is a relatively small corpus, but it reflects what could practically be achieved given the time and resources available (cf. Durrell et al. 2011).

It became clear from the earliest results that many previous investigations of the development of the forms and structures of the language in this period had indeed not been fully representative, since they had rarely taken the range of user-based or usage-based variation into account, tending to concentrate on

the prestige literary variety and consider almost exclusively the developments in that register. This reflected on the one hand the ideology of standard (cf. Milroy/Milroy 1999), as this register was commonly equated with the language as such, but on the other, of course, it was precisely those texts which were most readily accessible in the days before digitization. However, taking evidence from a single register clearly runs the risk of presenting a skewed picture of developments in the language as a whole. It was not just Bad Data, but an artificially restricted set of Bad Data which excluded a lot of the material which has actually come down to us.

4.1 The order of finite and non-finite verbs in subordinate clauses

A characteristic example of the sort of limitations this meant for research into the development of German is provided by the issue of the relative order of finite and non-finite verbs in subordinate clauses. Although this has long been regarded as one of the most interesting issues in the syntax of German (and other Germanic languages) from a theoretical perspective, it is noteworthy that the recent study by Sapp (2011) only covers the period up to 1650 and from 1800, as corpus data were not available for the intervening period, and the most comprehensive older study (Härd 1981) concentrates exclusively on developments in literary texts, as does the more recent account of developments in the intervening century and a half by Takada (1994).

Taking subordinate clauses with two verbs, the order of verbs within these groups was still fairly free in ENHG and three possible sequences are all relatively common:

(a) [...] **FINITE + NON-FINITE:**
[...], *dass du es heute* [...] ***sollst machen***

(b) [...] **FINITE** [...] **+ NON-FINITE:**
[...], *dass du es* [...] ***sollst*** *heute* [...] ***machen***

(c) [...] **NON-FINITE + FINITE:**
[...], *dass du es heute* [...] ***machen sollst***

We considered examples with a modal auxiliary rather than the periphrastic perfect tense, since, as will be shown later, the perfect auxiliary is often omitted in subordinate clauses at this time. The only acceptable sequence in modern standard German is (c), and according to Härd (1981) this was already established as the dominant norm by 1600. The study by Lühr (1985) also estab-

lished that Luther used this sequence in nearly 90% of possible instances (cf. Fleischer 2011: 166). However, preliminary data from the GerManC corpus, given in Table 1, show that the older sequences do persist after 1650, even though they are relatively infrequent, with the highest proportion being in northern texts where they account for some 14% of cases in the first period.

	1650-1700			1700-1750			1750-1800		
	a	b	c	a	b	c	a	b	c
North	260	26	16	319	9	8	215	1	
WCG	234	7	4	148	2	2	168		
ECG	321	8	10	258	8	6	159		
WUG	177	12	4	174	5	5	185		
EUG	172	13	4	156	10	3	174	1	3

Sequence (a): [...], *dass du es heute* [...] *machen* ***sollst***
Sequence (b): [...], *dass du es heute* [...] ***sollst*** *machen*
Sequence (c): [...], *dass du es* [...] ***sollst*** *heute* [...] *machen*

Table 1: Sequence of finite and non-finite verbs in subordinate clauses (Two-part sequences)

The picture is similar with three-part groups, as shown in Table 2 on the basis of passive constructions with a modal auxiliary.

	1650-1700		1700-1750		1750-1800	
	a	b	a	b	a	b
North	15	15	16	15	67	2
WCG	16	10	15	14	65	6
ECG	27	10	33	15	35	
WUG	19	16	13	10	14	2
EUG	12	7	11	12	10	4

Sequence (a): [...], *dass es* [...] *gemacht werden* ***soll***
Sequence (b): [...], *dass es* [...] ***soll*** *gemacht werden*

Table 2: Sequence of finite and non-finite verbs in subordinate clauses (Three-part sequences)

These sequences are naturally less frequent, and variation continues over a considerably longer period. As Fleischer (2011: 167) points out, though:

> Die Datensituation in Bezug auf die historische Entwicklung ist [...] widersprüchlich. Nach Härd (1981: 89) geht im 17. Jahrhundert „das finite Hilfsverb den infiniten Konstituenten voran." Dagegen schließt Takada (1994: 215) aus einer Korpusanalyse von Texten des 17. Jahrhunderts, dass sich die Nachstellung des Finitums auf Kosten der Voranstellung ausbreitet. Je nach analysiertem Korpus kommt man also zu verschiedenen Schlüssen.

Takada (1994) and Härd (1981) both use relatively limited sets of material with no allowance for representativeness, and it is perhaps not altogether surprising that their findings show marked differences. Härd (1981), unlike Takada (1994), was not using an electronic corpus, and his material shows a quite dramatic change after 1700, with almost total dominance of final position after that date, as in the modern standard (cf. Härd 1981: 170). By contrast, our corpus, which unlike these earlier studies includes material from a range of registers, shows a rather different picture, with variation persisting much longer and the two sequences in three-part groups evenly balanced until 1750, with the exception of East Central German – significantly the region whose usage, especially in literary genres, had high prestige and tended to function as a model for the developing standard. Nevertheless, the varying findings demonstrate that the problem of Bad Data in relation to the diachronic development of this feature is probably insoluble, since it is unlikely to be possible to find enough instances of these relatively rare constructions to provide an absolutely definitive picture of the process by which the variant which has become the modern norm was finally selected.

4.2 Genre-related variation in the order of finite and non-finite verbs

In practice, the problem of inadequate data even occurs with the two-place constructions. If we separate out the figures by genre, we find that a strikingly high proportion – in fact a majority – of the attestations for sequences with the finite verb first are in dramas, especially in North German.

Nothing comparable has been noted in earlier studies, despite the fact that we are dealing with a literary genre. The fact that most Period 1 dramas are in verse may be an additional complicating factor. However, verse is the norm for dramas of this period, and although we were aware of the problems this might

entail, we felt that we could not represent the genre properly if verse dramas were excluded. In fact even if it had been felt that prose dramas were to be preferred, it could have been difficult to find sufficient for our samples.

	1650-1700			1700-1750			1750-1800		
	a	b	c	a	b	c	a	b	c
North	25	23	9	34	4	2	11	1	
WCG	1	3		15			31		
ECG	40	6	2	19	1		19		
WUG	6	10	1	17	4	2	35		
EUG	8	6		7			29	1	

Sequence (a): [...], *dass du es heute* [...] ***machen sollst***
Sequence (b): [...], *dass du es heute* [...] ***sollst machen***
Sequence (c): [...], *dass du es* [...] ***sollst*** *heute* [...] ***machen***

Table 3: Sequence of finite and non-finite verbs in subordinate clauses (Two-part sequences in Drama texts)

Nevertheless, the high proportion of instances of the finite verb being placed first cannot simply be explained by the exigences of rhyme or metre. First of all, this order must clearly still be grammatical, since ungrammaticality is not acceptable even in verse. It is also notably predominant in North German, and to a lesser extent in West Upper German, although with such small figures one hesitates to draw any firm conclusions. Interestingly, Takada's (1994) data also show a relatively high proportion of sequences with the finite verb first from northern texts. He actually refers to these as *Niederdeutsch*, but it is not Low German, but High German written by North Germans. However, it is not impossible that the order is calqued on Low German dialect. Equally we could here have, in an orally-oriented genre such as the drama, a reflection of general spoken norms, with persistence of variation, such as Hennig (2009) found in her data from *Nahesprache*. Finally, it would seem significant that the proportions in East Central German are quite different, with second position clearly predominant even in drama written in verse. And this is the region with the most highly developed degree of literary culture and whose language is most prestigious and often dominant in the selection of the variant which is ultimately selected as standard.

It would seem to be the case here that the effect of acquiring additional data in an electronic corpus has actually been to raise more questions than it answers. The picture provided by traditional selection of texts, such as Härd (1981) undertook, was fairly straightforward, with a relatively early selection of the variant which was eventually codified for the written standard. What we have found in our corpus, which includes texts from a wider variety of genres, could be a more rounded picture of developments in the language as a whole in that it shows variation persisting much longer and differing according to genre and region. But the picture is clearly much more complex, and the reasons underlying the variation and its persistence are more difficult to explain, such that one can only draw very tentative conclusions which need to be corrobarated with further evidence, possibly of another kind altogether. In practice, we appear simply to have acquired more Bad Data, if not even Worse Data.

4.3 The a-finite construction

The problem of representativeness and data has come to light again recently through research currently being undertaken in Manchester using the GerManC corpus on the so-called a-finite construction of older German – the ellipsis of the auxiliary (most commonly the perfect auxiliary) in subordinate clauses, as in the following extract from the "Extraordinari Europæische Zeitung" No. 77, published in Hanau in 1701:

> Es hat sich auch dieser Prælat solcher Commission aquitiret, ist aber darinnen so glücklich nicht gewesen/ als er wohl gewünschet hätte/ weil der Hr. Cardinal Bedencken träget ferner etwas an die Stände dieses Königreichs gelangen zu lassen/ ehe und bevor dieselbe ihres Sentements auff das letzte Königl. Patent und sein Schreiben **so er dabey an dieselbe abgelassen**/ entdecket und kund gemacht haben werden.

This construction is of considerable interest for general syntactic theory, as Breitbarth (2005) shows in her study of the feature. It emerged in late Middle High German and became frequent in Early New High German. Breitbarth's study is based on five texts of roughly 9,000 words each from the *Bonner Frühneuhochdeutschkorpus* for each of the periods covered by that corpus, and her findings show a rapid decline in the occurrence of the construction in the eighteenth century.

Breitbarth (2005) points out the potential limitations of her data sources, but claims that her figures are broadly in line with those obtained in earlier studies by Admoni (1967) and Härd (1981). Effectively, even though she does acknowledge that her data do not prove anything for the language as a whole, and that they can only be taken for what they are, i.e. the output of individual speaker's grammars, she does claim that she has been able to show a general tendency in the language and that the construction becomes much less frequent after 1700 and has pretty well disappeared by 1800.

	CLAUSE TYPE					
	RELATIVE		ADVERBIAL		ARGUMENT	
sub-period	percent	number	percent	number	percent	number
1450-1500	2.6	229	4.6	245	1.2	215
1500-1550	16.8	255	19.7	257	9.5	235
1550-1600	48.2	434	54.0	420	26.4	179
1600-1650	66.9	565	68.9	478	52.7	237
1650-1700	60.8	392	65.7	488	44.9	176
1700-1800	17.9	163	6.6	145	25.2	76

Table 4: The a-finite construction (Data from Breitbarth 2005)

However, these are actually quite broad conclusions, as emphasized in the title of her thesis, but ultimately they rely on a rather small number of actual texts which may lack adequate representativeness. Recent work by Thomas (2012) on the basis of the GerManC corpus, on the other hand, has revealed a very different picture. Even though she has initially only investigated texts from a single genre (Humanities) in a single region (West Central German) over the period 1650-1800, she found that, far from declining in the eighteenth century, the incidence of the a-finite construction actually increased markedly after 1700 and still accounted for a majority of instances in the second half of the eighteenth century.

These data are naturally still only preliminary, but they correspond closely to initial observations by the inputters, including the present author, in the rest of the corpus, and they may well actually be representative of general written usage (and it is important to bear this latter point in mind, since it is questionable

whether the construction was ever current in speech). Even so, it is by no means out of the question that when all genres and regions have been investigated systematically, the results may be closer to those obtained by Breitbarth (2005).

Period	finite	afinite	Total	percentage a-finite
1650-1700	44	38	82	46.34%
1700-1750	24	87	111	78.38%
1750-1800	15	18	33	54.55%
TOTAL	83	143	226	63.27%

Table 5: The a-finite construction (Data from *Humanities* texts in the GerManC corpus)

Even if that were to be the case, though, we see here very forcibly the validity of Rissanen's (1989; 2008) caveats mentioned earlier, in particular that a corpus cannot be equated with "the language". If we found the assertion by Hunston (2002: 23) that "a statement about evidence in a corpus is a statement about that corpus, not about the language or register of which the corpus is a sample" rather too limiting, it does still flag up the potential risks which must always be borne in mind of basing broad conclusions on data whose representativeness is by no means assured and which cannot be queried directly by reference to actual language users. What findings we obtain and what conclusions we might draw are only tentative hypotheses based on what is inherently Bad Data.

4.4 Difficulties with identifying and marking "words"

Despite these caveats, a clear advantage of electronic corpora is that sophisticated tools can be (and have been) developed which enormously facilitate linguistic analysis and data collection (cf. McEnery/Hardie 2012). Corpora can be tagged, annotated for morpho-syntactic categories and parsed so that instances of particular forms and constructions can be found more quickly (and more reliably) than when laborious searches needed to be made in the original documents, even if the actual procedures of linguistic analysis are still essentially the same.

However, the exploitation of such tools, too, may not be as straightforward as one might wish, and Denison (2013) has shown that there are inherent problems involved in annotation which he has characterized (Denison 2010) as "WYSIWYTCH", i.e. "What You See Is What Your Theory Can Handle". As Denison (2013: 17) says:

> [...] for grammatical mark-up, with few exceptions a given scheme must privilege one particular analysis for each word, sentence or other unit of analysis. [...] Grammatical mark-up remains essentially a matter of synchronic analysis, and the guiding principle is to be as specific as possible; tagsets routinely deploy a much finer set of distinctions than traditional word classes.

Denison argues that these principles can be problematic since certain forms may not allow of unambiguous allocation to a particular tag, with a particular problem in English being the porousness of word-class boundaries, especially in a diachronic context (cf. also Denison in prep.). Similar problems were encountered in the process of annotating GerManC (cf. Scheible et al. 2011). One striking example involves one of the oldest and most fundamental theoretical problems in linguistics, i.e. the question of what constitutes a "word". For instance, as Denison (2013: 25-27) points out, the two most recent authoritative grammars of English differ on the analysis of complex prepositions like *on behalf of*, with Quirk et al. (1985: 670-673) claiming that eight out of nine indicators support a complex preposition analysis, i.e. as a single word, whilst Huddleston/Pullum (2002: 620-622) find no syntactic grounds for recognising such strings as complex prepositions. And, as Denison (2013: 27) points out, the British National Corpus tags every occurrence of *on behalf of* in two different ways at different levels of XML mark-up, i.e.:

(a) $[_{PRP}$ on$]$ $[_{NN1}$ behalf$]$ $[_{PRF}$ of$]$
(b) $[_{PRP}$ on behalf of$]$

That is, in (a) the three words are tagged individually as PRP (preposition) + NN1 (singular common noun) + PRF (preposition *of*), whereas in (b) the whole string is treated as a "multiword token" and tagged as a single preposition.

Historical stages of German throw up numerous instances of such intractable problems which affect tokenization, normalization, lemmatization and tagging. For instance, pronoun cliticization is very prevalent, especially (perhaps unsurprisingly) in the verse dramas, for example, from J.R. Karsten's "Christrühmendes Schau-Spiel" (Frankfurt/Main 1668):

Au! au! der Arm! du Hund! hast du ihn uns verrenkkt/
So ***wirstu*** *ohne Gnad an Galgen aufgehenkkt!*

and we felt there was no alternative but to solve this by tokenizing the two elements as distinct words. However, this is clearly not entirely satisfactory since it makes searches for individual cliticized forms less straightforward. An even more pervasive problem is one which besets modern German, i.e. whether "words" should be written separately or together, and variation in respect of this is rampant in the period before codification of the orthography. To take a frequent example, there is much variation, even within single texts, between writing the infinitive particle *zu* separately from the verb or prefixed to it, e.g. *zugewarten* or *zu gewarten*. We decided after considerable discussion that the particle and the verb had to be consistently tokenized separately, not least to avoid potential confusion with *zu* as a verb prefix, and the simplex verb was required as input to normalization and lemmatization.

In practice, though, it is more frequent in Early Modern German for what are now seen as compound words to be printed separately – or, if they were written together, with internal capitalization or hyphens. Eventually we felt there was little alternative to take the forms as we found them, i.e. assigning compound nouns written as a single word (or with a hyphen) to a normalized lemma corresponding to the modern standard form, i.e. with no internal capitals or hyphens. Thus *Südseeinsel* would be used as the normalized form of *SüdSeeInsel* or *Südsee-Insel* or any other variant on these. However, compounds written in the text as two (or more) words were tagged as individual words, so that *Süd See Insel* is tagged as three words. This seemed the only practical solution, although it is clearly less than wholly satisfactory. A similar problem arose with verb prefixes written separately from the verb, e.g. *wahr nahm*, which is still very frequent in seventeenth century texts, whereas in modern usage they would be written together. However, in practice this turned out to be a rather less serious issue, because the verb prefix *wahr* could still be tagged as such, i.e. PTKVZ using the Stuttgart-Tübingen tagset (STTS, cf. Schiller et al. 1999), a possibility provided by the fact that in modern German prefixes can be separated from the verb and may thus be allocated a distinct tag.

As we saw, Denison (2013: 27) showed how the British National Corpus attempts to solve this kind of problem with two separate tags at different levels of XML mark-up, but this of course makes considerable demands in terms of time and resources. However, Early Modern German presents a more complex

variant of this problem with some conjunctions, in that it is not unusual for conjunctions which in the modern language are clearly single words, like *ob-gleich*, to appear as separate words in texts of this period. It seems straightforward to tag these as separate words using STTS tags, i.e. *ob* KOUS [...] *gleich* ADV, following the model provided for in the STTS tagset guidelines for tagging two word conjunctions of modern German like *als ob*, i.e. KOKOM *als* KOUS *ob* – despite the fact that it is perhaps not entirely satisfactory, since these "words" are clearly operating as a single semantic or syntactic "multiword" unit. However, in older German the parts of such "multiword" conjunctions are frequently separated by anything up to four words in the subordinate clause, as in the following example from "Drey Bücher Der Magnetischen Artzney-Kunst" by Guillelmus Maxvellus (Frankfurt 1687):

> *Er purgieret allein unter sich/ man darff sich auch keiner Salivation befahren/* ***ob*** *man sich ihme* ***gleich*** *bey erfordernder Noth etlich mal gebrauchet*

Clearly these can still be tagged in the same way, with *ob* identified as the conjunction proper KOUS, and *gleich* as an adverb ADV, but it would seem that in one plausible analysis we are still dealing with a multi-word token which requires some appropriate identification which we have not (yet) been able to assign satisfactorily, especially as that could be the most helpful to the corpus user attempting to trace the development of this conjunction – and this is a criterion which must always be borne in mind, since a corpus is in the first instance ultimately a resource for researchers.

However, such cases only serve to further illustrate the central issue being addressed here. The existence of such constructions has long been known in German historical linguistics, so what has the corpus told us that we didn't know already? Are the problems just outlined simply a product of the difficulty of devising optimally efficient tools by means of which we can access and analyse the large amount of Bad Data which an electronic corpus may provide us with? Are we just engrossing ourselves in the fascination of the complex technology and the challenge of compiling programs to solve problems to which we may already know the answer (cf. Wegera 2013: 58)? Naturally, there may be examples of this, but it is evident that good practice must be to remain aware of these dangers and to always remember that the compilation of a corpus and the challenges of designing tools are not ends in themselves.

5. Conclusions

On balance, though, the existence of electronic datasets has facilitated huge steps forward in understanding language history and language change. Not only have sophisticated tools made it possible to ask questions which simply could not be considered previously, for example the research into complex patterns of change in usage reported in Hilpert (2011), with motion charts of development, even if such do depend on very big datasets of a kind to which we can only have recourse for fairly recent periods, or the work on linguistic networks in Late Latin by Mehler et al. (2013). The simple fact of the increased accessibility to data by large numbers of scholars has been immensely beneficial. It is no longer the case, for instance, that doctoral students have to laboriously and time-consumingly compile datasets, and this process has to be repeated by every new researcher. In this way, to return to the example of *obgleich*, we may have been aware that the construction existed, but we can now have a much clearer picture of when it emerged or how frequent it was in comparison to the compound, and whether it was used more in one genre or one region than another.

Nevertheless, it is still vital to be clear what one can legitimately expect from such corpora. We still do not have access to the whole of the language, but only what has chanced to come down to us, and that this is written language which may have autonomous norms at some remove from those of the spoken language, not least because of the development of standardized prescriptions. The a-finite construction discussed earlier may be an example of the problems entailed by this latter issue, since it is perhaps doubtful whether this was ever current in spontaneous spoken production. Even though it is important, in dealing with a period with a relatively large amount of preserved material, to sample what we have as widely as possible taking the variables we are aware of into account, we can only make statements in relation to those parameters, effectively formulating hypotheses on the basis of the Bad Data which we still have access to, in the light of our own (possibly limited) competence as historical linguists, our overall knowledge of the diachrony of the language involved and the circumstances in which the texts were produced, inasfar as these are known – and these can be very limited, as for example in the case of early newspapers (cf. Durrell/Ensslin/Bennett 2008).

An electronic corpus means, first and foremost, that we can store very large datasets and access and query them very quickly. But you do have to know what you are looking for; even with a large electronic dataset it is still the case that the real work starts when the counting stops. Any findings from a corpus need to be carefully investigated and elucidated in the light of what else we know about the language in question at the period in question, and, as Rissanen (1989: 16-17) says:

> In the analysis, synthesis and conclusions, the machine does not replace the human brain. We will be able to ask the right questions, draw inferences and explain the phenomena revealed by our data only if we develop a good overall mastery of the ancient language form we are studying.

References

Admoni, Wladimir G. (1967): Der Umfang und die Gestaltungsmittel des Satzes in der deutschen Literatursprache bis zum Ende des 18. Jahrhunderts. In: Beiträge zur Geschichte der deutschen Sprache und Literatur (Halle) 89: 144-199.

Auer, Anita (2009): The subjunctive in the age of prescriptivism. English and German developments during the Eighteenth Century. Basingstoke: Palgrave Macmillan.

Bennett, Paul/Durrell, Martin/Scheible, Silke/Whitt, Richard J. (eds.) (2013): New methods in historical corpus linguistics. (= Korpuslinguistik und interdisziplinäre Perspektiven auf Sprache / Corpus Linguistics and Interdisciplinary Perspectives on Language (CLIP) 3). Tübingen: Narr.

Biber, Douglas/Finegan, Edward/Atkinson, Dwight (1993): ARCHER and its challenges. Compiling and exploring a representative corpus of historical English registers. In: Aarts, Jan/de Haan, Pieter/Oostdijk, Nellehe (eds.): English language corpora. Design, analysis and exploitation. Amsterdam: Rodopi, 1-13.

Breitbarth, Anne (2005): Live fast, die young. The short life of Early Modern German auxiliary ellipsis. PhD dissertation, Tilburg University. http://www.lotpublications.nl/publish/articles/001464/bookpart.pdf (last accessed December 6, 2013).

Cantos, Pascual (2012): Corpora for the study of linguistic variation and change. Types and computational applications. In: Hernández Campoy/Silvestre, Conde/Camilo, Juan (eds.) (2012): The handbook of historical sociolinguistics. Malden etc.: Wiley-Blackwell, 99-122.

Denison, David (2010): Category change in English with and without structural change. In: Traugott, Elizabeth Closs/Trousdale, Graeme (eds.): Gradience, gradualness and grammaticalization. Amsterdam/Philadelphia: John Benjamins, 105-28.

Denison, David (2013): Grammatical mark-up. Some more demarcation disputes. In: Bennett et al. (eds.), 17-35.

Denison, David (in prep.): English word classes. (= Cambridge Studies in Linguistics). Cambridge: Cambridge University Press.

Durrell, Martin/Ensslin, Astrid/Bennett, Paul (2008): Zeitungen und Sprachausgleich im 17. und 18. Jahrhundert. In: Zeitschrift für deutsche Philologie 127: 263-279.

Durrell, Martin/Bennett, Paul/Scheible, Silke/Whitt, Richard J. (2011): Investigating diachronic grammatical variation in Early Modern German. Evidence from the GerManC corpus. In: Konopka, Marek/Kubczak, Jacqueline/Mair, Christian/Šticha, František/Waßner, Ulrich H. (eds.): Grammatik und Korpora 2009. (= Korpuslinguistik und interdisziplinäre Perspektiven auf Sprache / Corpus Linguistics and Interdisciplinary Perspectives on Language (CLIP) 1). Tübingen: Narr, 539-549.

Fleischer, Jürg (2011): Historische Syntax des Deutschen. Eine Einführung. (= Narr Studienbücher). Tübingen: Narr.

Härd, John Evert (1981): Studien zur Struktur mehrgliedriger deutscher Nebensatzprädikate. Diachronie und Synchronie. (= Göteborger Germanistische Forschungen 21). Göteborg: Acta Universitatis Gothoburgensis.

Hennig, Mathilde (2009): Nähe und Distanzierung. Verschriftlichung und Reorganisation des Nähebereichs im Neuhochdeutschen. Kassel: Kassel University Press.

Hilpert, Martin (2011): Dynamic visualizations of language change: Motion charts on the basis of bivariate and multivariate data from diachronic corpora. In: International Journal of Corpus Linguistics 16: 435-461.

Huddleston, Rodney/Pullum, Geoffrey K. (2002): The Cambridge grammar of the English language. Cambridge: Cambridge University Press.

Hunston, Susan (2002): Corpora in applied linguistics. Cambridge: Cambridge University Press.

Hunston, Susan (2008): Collection strategies and design decisions. In: Lüdeling/Kytö (eds.), Vol. 1: 154-168.

Jones, Howard (2009): *Aktionsart* in the Old High German passive with special reference to the *Tatian* and *Isidor* translations. (= Beiträge zur germanischen Sprachwissenschaft 20). Hamburg: Buske.

Köhler, Reinhard (2005): Korpuslinguistik. Zu wissenschaftstheoretischen Grundlagen und methodologischen Perspektiven. In: LDV-Forum 20: 1-16.

Labov, William (1992): Some principles of linguistic methodology. In: Language in Society 1: 97-120.

Lass, Roger (1997): Historical linguistics and language change. Cambridge: Cambridge University Press.

Leech, Geoffrey (1991): The state of the art in corpus linguistics. In: Aijmer, Karin/Altenberg, Bengt (eds.): English corpus linguistics. Studies in honour of Jan Svartvik. London: Longman, 8-30.

Leech, Geoffrey (2007): New resources, or just better old ones? The Holy Grail of representativeness. In: Hundt, Marianne/Nesselhauf, Nadja/Biewer, Carolin (eds.): Corpus linguistics and the web. Amsterdam: Rodopi, 134-149.

Leiss, Elisabeth (1992): Die Verbalkategorien des Deutschen. Ein Beitrag zur Theorie der sprachlichen Kategorisierung. (= Studia Linguistica Germanica 31). Berlin/New York: de Gruyter.

Lüdeling, Anke/Kytö, Merja (eds.) (2008-2009): Corpus linguistics. An international handbook. 2 vols. Berlin/New York: de Gruyter.

Lühr, Rosemarie (1985): Zur Syntax des Nebensatzes bei Luther. In: Sprachwissenschaft 10: 26-50.

McEnery, Tony/Hardie, Andrew (2012): Corpus linguistics. Method, theory and practice. Cambridge: Cambridge University Press.

Mehler, Alexander/Schwandt, Silke/Gleim, Rüdiger/Ernst, Alexandra (2013): Inducing linguistic networks from historical corpora. Towards a new method in historical semantics. In: Bennett et al. (eds.), 257-274.

Milroy, James/Milroy, Lesley (1999): Authority in language. Investigating language prescription and standardisation. 3rd. ed. London: Routledge & Kegan Paul.

Nevalainen, Terttu (1999): Making the best use of "bad" data. Evidence for socio-linguistic variation in Early Modern English. In: Neuphilologische Mitteilungen 100: 499-533.

Quirk, Randolph/Greenbaum, Sidney/Leech, Geoffrey/Svartvik, Jan (1985): A comprehensive grammar of the English language. London/New York: Longman.

Rissanen, Matti (1989): Three problems connected with the use of diachronic corpora. In: ICAME Journal 13: 16-19.

Rissanen, Matti (2008): Corpus linguistics and historical linguistics. In: Lüdeling/Kytö (eds.), 53-68.

Sapp, Christopher D. (2011): Verb order in subordinate clauses from Medieval to Modern German. Amsterdam/Philadelphia: John Benjamins.

Scheible, Silke/Whitt, Richard J./Durrell, Martin/Bennett, Paul (2011): Evaluating an 'off-the-shelf' POS-tagger on Early Modern German text. In: Proceedings of the ACL-HLT 2011 Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH 2011). Portland, Oregon. http://www.aclweb.org/anthology/W/W11/W11-1503.pdf (last accessed December 6, 2013).

Schiller, Anne/Teufel, Simone/Stöckert, Christine/Thielen, Christine (1999): Guidelines für das Tagging deutscher Textcorpora mit STTS. Technical Report. Stuttgart: Institut für maschinelle Sprachverarbeitung.

Storjohann, Petra (2003): A diachronic constrastive lexical field analysis of verbs of human locomotion in German and English. Frankfurt a.M. etc.: Peter Lang.

Takada, Hiroyuki (1994): Zur Wortstellung des mehrgliedrigen Verbalkomplexes im Nebensatz im 17. Jahrhundert. Mit einer Beantwortung der Frage, warum die Wortstellung von Grimmelshausens 'Simplicissimus' geändert wurde. In: Zeitschrift für Germanistische Linguistik 22: 190-219.

Thomas, Victoria (2012): Report of pilot search: auxiliary ellipsis in German embedded clauses 1650-1800. Unpublished working paper, School of Arts, Languages & Cultures, University of Manchester.

Wegera, Klaus-Peter (2013): Language data exploitation. Design and analysis of historical language corpora. In: Bennett et al. (eds.), 55-73.