Corpus linguistics and dialectology

Lieselotte Anderwald and Benedikt Szmrecsanyi

1. Introduction

In contrast to sociolinguistics, dialectology and corpus linguistics have been rather uneasy bedfellows until relatively recently. In this article, we will focus on the use of modern corpora in the field of traditional dialectology – corpora, that is, which constitute principled, possibly computerized, and broadly representative collections of naturalistic spoken (and sometimes written) dialect material. This sets apart corpus-based dialectology from other approaches in empirical dialectology, which may be based on, e.g., dialect atlases and/or questionnaire data. The subject matter of traditional dialectology (as opposed to urban dialectology), then, is the distribution of linguistic features where, crucially, the primary parameter of variation is geographical distance and proximity between sampling locales. This means that this article will exclude from discussion corpora documenting variation between standard dialects (e.g. British English vs. American English), and we will also remain agnostic about corpus work on youth dialects and other sociolects, where variation is not stratified geographically but rather sociologically (but on this topic, cf. article 6 on corpora in sociolinguistics).

2. Aim and scope of traditional dialectology

The study of (non-standard) dialects has a venerable tradition going back at least to the nineteenth century (for a short overview, cf. Chambers/Trudgill 1998; for developments more specifically related to English, cf. Ihalainen 1994). Interest in rural varieties of a language that were (or appeared to be) far removed from the standard language were kindled by the Neogrammarian dictum of the 'exceptionlessness of sound change' (Ausnahmslosigkeit der Lautgesetze), a dictum dialect data could support (or, indeed, refute) – many theoretically interesting intermediate stages of proposed changes could actually be documented with the help of dialect data. Especially isolated, rural dialects were seen as unspoilt or uncorrupted varieties, representing diachronically removed stages of the language in a purer form. This more theoretical interest in (at least certain) non-standard varieties seems to have been superseded by an interest in the dialect *per se*, often by dialect speakers themselves – witness the establishment of many regional dialect societies e.g. in England at the end of the nineteenth century, the most prominent of which is perhaps the Yorkshire Dialect Society, founded in 1897 by the scholar Joseph Wright, himself a Yorkshireman and author of the monumental *English Dialect Dictionary* (1898-1905). (The society exists to this day, see

http://www.ydsociety.org.uk/.) As, naturally, dialects are as little static as any other living linguistic system, observable tendencies of change were perceived as threatening the 'purity' or authenticity of the 'real' dialect, leading to efforts to preserve or at least record their 'original' state. (Actual sociocultural changes on a scale unknown so far, like increasing industrialization, urban migration, improved education etc. must have played an important role in shaping this perception as well, but cannot be explored further within the scope of this article. For a short overview, cf. Beal 2004.) It is little wonder, then, that traditional dialectologists saw the speech of non-mobile older rural males as the only 'true' representative of the dialect in question, since we know from modern sociolinguistic studies that features like reduced mobility, higher age, non-urbanity, or male gender typically contribute to the speaker employing a more conservative variety linguistically. No doubt due to the predilection with sound change, phonological investigations were at the forefront of dialectological interest, closely followed by perhaps one of the most notable features of regional variation, namely differences in lexis. This focus on collecting and classifying lexical items has no doubt contributed to the image of dialectologists 'collecting butterflies', i.e. engaging in a time-consuming activity which is however perceived as ultimately irrelevant to the wider community of linguists.

In this article, we will try to show that modern dialectology has moved far beyond these more traditional concerns, not only in the choice of subjects, materials and methods, but also in its relevance to theory-building and indeed to wider linguistic concerns. Not least the method of employing corpus-based materials has contributed to the resurrection and rehabilitation of this discipline.

3. Traditional materials in dialectology

Traditional dialect studies have relied mostly on questionnaires to elicit lexical material and dialect phonology. The German Georg Wenker was one of the pioneers of the postal questionnaire: in 1876 and 1877 he sent out a questionnaire to schoolmasters all over Germany and had them translate his 40 sample sentences into the local dialect. To this day, the so-called "Wenker sentences" are used in research, and have recently been made available electronically at the University of Marburg, Germany (in the form of the so-called DiWA, the Digital Wenker Atlas, http://www.diwa.info/), over a century after they were first collected. Towards the end of the nineteenth century, trained fieldworkers were sent out all over the country to record the answers to typical questionnaire questions in some kind of phonetic script; the pioneer of this form of data collection was the Swiss Jules Gilliéron, who sent out

his fieldworker Edmond Edmont to cycle all over France. Edmont conducted around 700 interviews between 1886 and 1900. This type of nationwide dialectological fieldwork, especially from the first half of the twentieth century, typically resulted in linguistic atlases, and the French project (Gilliéron 1902-1910) became the benchmark for much subsequent work in Switzerland, Italy or Spain. Much material has also been published in the form of dialect dictionaries since the end of the nineteenth century (for an early specimen for British English dialects, cf. Wright 1898-1905 mentioned above; for a more modern example involving U.S.-American dialects, cf. the DARE project: Hall (1985-2009 [scheduled]), some volumes of which are still in the process of being published). None of the early atlas projects recorded larger stretches of discourse. This changed with the post-war nation-wide survey of England (The Survey of English Dialects, or SED; all informants' responses are published as Orton/Halliday 1962-1964; Orton/Wakelin 1967-1968; Orton/Barry 1969-1971; Orton/Tilling 1969-1971), as well as the various North American projects, many conceived before the second world war, but conducting field work mostly afterwards (for an overview cf. http://us.english.uga.edu/). Especially towards the second half of the century, tape recorders (or their precursors) were increasingly used to aid fieldworkers' jobs in the large atlas projects. Many of the tapes have survived, and they sometimes contain slightly longer stretches of informants' speech. Recently, some of this material from the SED has been transcribed and made accessible (the SED recordings, cf. Klemola/Jones 1999). These recordings probably constitute the oldest dialectological material that we possess today that can be called an (electronic) corpus of non-standard speech. Dialectological work in the 1970s and 1980s used recording technology as a matter of fact, but the recorded material as a rule remained with the researchers. Much work in recording dialects stemmed from university theses (for English dialects, especially from the centres of dialectology, Sheffield and Leeds, e.g. Shorrocks 1980, published as Shorrocks 1999, but recently also Newcastle, cf. below; interesting work has also been done at the University of Helsinki in Finland, for a recent example, cf. Vasko 2005); however, the material itself has not generally been made public.

One exception is a corpus that concentrates on one dialect area, the Newcastle Electronic Corpus of Tyneside English NECTE, cf. http://www.ncl.ac.uk/necte/. NECTE combines older dialect material, collected for the Tyneside Linguistic Survey in the late 1960s and 1970s, with recordings from the more modern project on Phonological Variation and Change in Contemporary English (PVC) (Milroy/Milroy/Docherty 1997), collected between 1991 and 1994 in the same area. This corpus has recently been published, making possible some diachronic studies for this dialect area.

Another resource worth mentioning that is just being finished is the Helsinki Corpus of British English Dialects (HD), which aims to gather together the materials collected for the purpose of university theses mentioned above (cf. http://www.eng.helsinki.fi/varieng/team3/1_3_4_2_hd.htm). Due to lack of manpower and general resources, it has taken about 30 years to complete, meaning that materials date from the 1970s and 1980s. The Helsinki dialect corpus (HD) contains material from Lancashire, Cambridgeshire, Suffolk, Essex, the Isle of Ely, Devon and Somerset, so that some regional comparisons are now becoming possible. The corpus contains almost one million words.

Recently, a new resource has been collected, the Freiburg English Dialect corpus (FRED) (Kortmann et al. 2005; Anderwald/Wagner 2007, cf. also http://www.anglistik.uni-freiburg.de/institut/lskortmann/FRED/). FRED contains almost 2.5 million words taken from Oral History Projects, conducted in the 1970s and 1980s, orthographically transcribed, from six large dialect areas in the British Isles. At least for English, this for the first time makes possible regional comparisons across Great Britain. A sizable subcorpus of FRED will be published on the third ICAME-CD (http://icame.uib.no/corpora.html).

As far as accessibility is concerned, the situation looks slightly better for German spoken language, because many researchers have deposited their material at the Institut für Deutsche Sprache (Institute for German Language) at Mannheim, Germany (http://www.ids-mannheim.de/), and corpora and their transcriptions are open to the public in the Datenbank Gesprochenes Deutsch (DGD, or database of spoken German). While this collection, too, contains situational or text-type specific discourse, it is also a good source for material from Lower German, Middle German, and Upper German dialects collected since 1950. Austrian dialect researchers have also utilized the Austrian Phonogramm recordings, held at the Austrian Audiovisual Research Archive (cf. http://www.pha.oeaw.ac.at/), samples of which are available commercially (Schüller 2003). Materials for most other languages seem to sadly lag behind the English (and, to a degree, German) vanguard, even where documenting the standard languages are concerned, not to mention non-standard varieties.

4. Examples of accessible dialect corpora

In the few corpora that are used in dialectology today we can distinguish several categories of materials that have been collected. In this section, we will discuss the use of traditional dialectological interviews, the use of more sociolinguistic materials, data from oral history projects, and participant interviews.

Dialectological "interviews": in contrast to recordings of spontaneous speech, as employed

in many sociolinguistic studies, dialectological corpora often consist of the recordings of more or less extended questionnaire sessions, as discussed above. The recording here was originally envisaged to aid fieldworkers' memories, but might constitute relevant material for phonetic and lexical investigations. A good example of this typical dialectological material are the SED recordings (however, commercial publication, originally intended, has not been resolved yet); for some first studies on this material cf. Klemola (2002, 2003).

Some material contains traditional sociolinguistic interviews: here, the borderline to sociolinguistics becomes fuzzy. If one defines variationist sociolinguistic studies, as Chambers and Trudgill do, as "urban dialectology" (e.g. Chambers/Trudgill 1998), only the rural – urban axis serves to distinguish the two fields, a somewhat artificial distinction. (However, as pointed out above, the choice of informants in dialectology is still mainly guided by the older traditional dialect speaker, so-called NORMs, whereas urban dialectology aims at socially stratified samples of speakers). An example containing this kind of dialectological-sociolinguistic material would be the Northern Ireland Transcribed Corpus of Speech (NITCS), cf. Kirk (1990).

Oral History Projects: as mentioned above, some dialect projects have used information collected for different purposes, e.g. Oral History Projects (cf. FRED mentioned above, and also Huber 2003, who plans a corpus based on South Wales oral history material, encompassing ca. 3 million words for that dialect area alone). An advantage here is that the participants' attention was genuinely on matters unrelated to language, thus avoiding the Observer's Paradox to some degree (cf. Labov 1972). On the other hand, interviewers will have been interested in specific content only, not necessarily paying particular attention to reducing the formal distance between interviewer and interviewee, or trying to elicit particularly informal styles, which might be desirable for the investigation of non-standard features.

Participant interviews/conversations: some dialectological and sociolinguistic work has consisted of dialect speakers, sometimes trained as fieldworkers, recording the speech of other dialect speakers. Thus for example, the PVC project encouraged dyadic interaction without the presence of the investigators (Milroy/Milroy/Docherty 1997).

Finally, recordings of regionally restricted, spontaneous conversations as in the Corpus of London Teenage Speech (COLT) are as amenable to dialectological investigations as to sociolinguistic ones, and studies taking these materials as a basis probably straddle the border between dialectology and sociolinguistics (Andersen 2001; Stenström/Andersen/Hasund 2002).

5. Analytical objectives in corpus-based dialectology

The analysis of dialect corpora – much as the analysis of other corpus data – can serve a variety of linguistic objectives. In what follows, we will illustrate this point by discussing a number of case studies seeking to explore dialect corpora for the primary aim of (i) functional-typological analysis, (ii) variationist analysis and probabilistic model-building, (iii) historical inquiry, and (iv) formalist/generative theory-building (or theory-rejection). In point of fact, the bulk of corpus-based studies in dialectology have addressed more than one objective at the same time, albeit with differing emphases; a survey of the literature reveals that most contemporary corpus-based dialectology does not stop at mere dialect description but offers some added theoretical and interpretational value. There are only some exceptions to this tendency: consider Jones (1985) for a corpus-based description of Tyneside English and Beal/Corrigan (2006) for a corpus-based descriptive follow-up on Tyneside negation specifically, or – in the realm of German dialectology – Patocka (1997) for a discussion of word order phenomena in Bavarian dialects spoken in Austria, drawing on a corpus covering assorted recordings in the Phonogramm archives at the Austrian Academy of Sciences (mentioned above). Still, dialect description – i.e. mapping the range and extent of variation, on the basis of naturalistic corpus data, of one or more dialect features in one or more dialect areas – necessarily precedes any interpretation of naturalistic dialect data in the light of a particular theoretical framework.

5.1. Functional-typological analysis

Functional-typological approaches to dialect data endeavor to marry functional typology to dialectological investigation. This means, in a nutshell, that the observable patterns of language-internal variation are analyzed and interpreted in terms of the same empirical and interpretational apparatus which is familiar from the typological study of large-scale cross-linguistic variation (cf. Kortmann 2004 for a collection of papers in this spirit).

Thus, Anderwald (2002, 2003) explores dialectal material in the spoken-demographic section of the *British National Corpus* (in which some texts are tagged for dialect area) to demonstrate that non-standard negation patterns are actually predicted by general cognitive and typological principles: the pervasiveness or invariant *don't* and *ain't* (e.g. *he don't/ain't like me*) in non-standard English, for example, is argued to bring non-standard English in line with cross-linguistic markedness criteria. In a quite similar vein, Herrmann (2005) presents evidence drawn from FRED that while standard English does not conform with

Keenan/Comrie's (1977) Noun Phrase Accessibility Hierarchy, English dialects do in that they allow subject gapping (e.g. *the man __ lived there*) in addition to object gapping (e.g. *the man __ I saw*). As it turns out, the Accessibility Hierarchy is also involved in dialectal pronoun systems and (pronominal) gender marking: Wagner (2004) discusses pronoun usage in traditional dialects in the Southwest of England as well as in Newfoundland, drawing on FRED for the Southwest of England and on a corpus of transcripts from the *Memorial University of Newfoundland Folklore and Language Archive* (MUNFLA). In the traditional dialects Wagner investigates, count nouns have historically been referred to by 'gendered' pronouns (*he, she*) whereas mass nouns have received *it*. This is a gender assignment system which is increasingly being crowded out by the system of Standard English. Crucially, Wagner shows that the way in which the traditional system is breaking down follows a path originally suggested by Ihalainen (for instance, 1991): Standard English forms start out from less accessible positions in the Accessibility Hierarchy and diffuse through the hierarchy to more accessible positions. This may point to a more general, possibly cross-linguistically valid, mechanism of language change. Also in regard to pronominal usage, Geyer (2003) draws on a relatively small corpus (covering about 40 minutes of transcribed speech) of Hetzlerisch Franconian, a dialect of German spoken in the village of Hetzel near Nuremberg, to document the inventory of phoric pronouns in that dialect (cf. colloquial German *der Mann kam in die Bäckerei, und* der *hatte einen Hut auf*). Geyer claims that phoric pronouns – and thus the interface between syntax and information structure – in Hetzlerisch Franconian ought to be viewed against the backdrop of crosslinguistic, typological variation.

Corpus-based dialectology can thus be embedded in a theoretical framework that seeks to predict, and explain, dialectal variation by well-known cross-linguistic parameters of variation.

5.2. Variationism and probabilistic model-building

Corpus studies in this line of dialectological inquiry explore corpus data of traditional dialects with quantitative techniques similar to those of urban dialectology and sociolinguistics (cf. article 6). This means that the main theoretical interest is in the (probabilistic) constraints that govern the choice between two variant forms, or between a dialectal variant and a standard variant. Thus, Hernández (2002) is concerned with untriggered *self*-forms (as in *for somebody like myself* instead of *for somebody like me*) in non-standard English. Investigating the *Northern Ireland Transcribed Corpus of Speech* (NITCS) and the spoken section of the *British National Corpus* as well as questionnaire data,

Hernández establishes that the phenomenon is not, as has been previously claimed in the literature, a feature that is typical merely of (dialectal) Irish English and that the choice between untriggered *self*-forms and their standard alternatives is governed by a hierarchy of language-internal and language-external constraints.

As a genuine study in probabilistic grammar, Pietsch (2005) addresses verbal agreement in (traditional) northern dialects of England. Since Middle English times, Northern English dialects have been displaying the so-called Northern Subject Rule: invariant verbal *–s* occurs anywhere (e.g. *birds sings*) except when the verb is directly adjacent to a simple personal pronoun (e.g. **I sings*). This system used to be categorical in traditional dialects but is now highly variable and competes with Standard English verbal concord. On the basis of dialect atlas material from the SED, on the one hand, and on corpus data – FRED and the NITCS – on the other hand, Pietsch aims to uncover the factors governing the occurrence or non-occurrence of verbal *–s* in these dialects. In addition to qualitative analysis, Pietsch marshals multivariate analysis methods (more specifically, Variable Rule Analysis) to investigate the probabilistic, intralinguistic constraints that govern the inherent variability of the variable in corpus data. It turns out that verbal concord in Northern varieties in the British Isles is a hybrid system that is best interpreted in terms of usage-based models of grammatical competence, such as cognitive grammar.

Exploiting the pervasive variation in dialect data, among other data sources, as a research site for probabilistic and psycholinguistic model building, Szmrecsanyi (2006) explores morphosyntactic persistence, i.e. the tendency of speakers to re-use linguistic material that they have used or heard before. The phenomenon, which is partly psycholinguistic and partly discourse-functional in nature, plays particular methodological havoc with those corpus-based approaches that rely on naturalistic data where dialect speakers potentially echo standard variants used by their interlocutors. By way of multivariate analyses of a number of corpora of English, among them the dialect corpus FRED, with regard to a number of well-known alternations in the grammar of English, Szmrecsanyi presents evidence that persistence is indeed a major probabilistic constraint on the way (dialect) speakers make linguistic choices. Needless to say, as a psycholinguistic constraint persistence is not specific to English, or English dialects: for example, in Romance dialects where plural expression is variable – such as Brazilian Portuguese (cf. Scherre/Naro 1991) and Puerto Rican Spanish (cf. Poplack 1980), corpus data exhibit similar parallelism effects, and Travis (2007) draws on corpus material to demonstrate that null subjects in two dialects of New World Spanish, Colombian Spanish and New Mexican Spanish, are more likely when another null subject

expression was recent.

With a similar interest in recency effects and the dialogic interdependence between interviewer and interviewee utterances, Hollmann/Siewierska (2006) offer a methodological outline of how accommodation theory and, therefore, the concept of sociolinguistic salience can be brought to bear on dialect corpus data (more specifically, in a corpus of oral history transcripts in the Lancashire dialect). Straightforwardly enough, Hollmann/Siewierska suggest, first, to determine the first variable form in the text, and then to calculate whether the likelihood of the interviewee using the dialectal variant increases significantly towards the end of the text; if it does, it is safe to assume that the interviewee accommodates to the interviewer.

Making extensive use of methods and explanatory patterns along the lines of modern variationist sociolinguistics, Sali Tagliamonte and her co-workers, in a series of recent studies, have sought to explore corpus data sampling traditional dialects - with a particular focus on relic areas - in a number of locales in Great Britain and the Americas (note, though, that the corpora subject to analysis are generally not available to the wider research community). This line of research, which rigorously combines the methodological machinery of urban dialectology with traditional dialect data, has investigated the following phenomena in English dialect data:

- *was/were* variation (e.g. *You were hungry but he were thirsty*) (Tagliamonte 1998; Tagliamonte/Smith 2000);
- the habitual past (e.g. *he would always dance* vs. *he used to always dance*) (Tagliamonte/Lawrence 2000);
- *come/came* variation (e.g. *when I come home that day* vs. *when I came home that day*) (Tagliamonte 2001);
- NEG/AUX contraction (e.g. *he isn't going* vs. *he's not going*) (Tagliamonte/Smith 2002);
- markers of stative/possessive meaning (e.g. *He has/has got/got a car*) (Tagliamonte 2003);
- zero complementation (e.g. *he shows that/∅ he can do it*) (Tagliamonte/Smith 2005);
- variation between relative markers (e.g. *the man that/∅/as/who … I saw*) (Tagliamonte/Smith/Lawrence 2005);
- (t,d) deletion (e.g. *I was told* vs. *I was tol∅*) (Tagliamonte/Temple 2005).

Characteristically, quantitative findings - especially factor weights in Variable Rule analysis and the resulting constraint rankings - are often interpreted in terms of historical or comparative research questions, where similar constraint rankings in different locales are taken to indicate genetic relatedness (cf. Tagliamonte/Temple 2005, 84). In addition, Tagliamonte and associates also occasionally interpret findings in terms of grammaticalization theory or in regard to how and why incoming forms may (or may not) diffuse through the community (for instance, Tagliamonte/Smith/Lawrence 2005). Sometimes, the variable portfolio includes external variables such as speaker age, which allows for tracking changes in apparent time (for instance, Tagliamonte 1998). In Tagliamonte/Temple (2005), (t,d) deletion is discussed against the backdrop of phonological theory.

In sum, dialect corpora can serve as a rich resource for establishing and benchmarking probabilistic grammars across different geographic locales. The patterns that this approach yields can be interpreted in terms of genetic relatedness, or along the lines of more general processing-related constraints that leave their mark on languages and dialects.

5.3. Historical linguistics

Given dialectology's traditional orientation towards historical linguistics and diachronic explanation, it should surprise no one that much corpus-based dialectology is still interested, to varying degrees, in the diachronic evolution and synchronic areal diffusion of linguistic forms. Exactly along these lines, Klemola (1996) investigates non-standard affirmative periphrastic *do* (e.g. *we did always go to school*) in traditional dialects in the Southwest of England, drawing on an oral history corpus and a corpus of SED fieldworker material. Klemola shows that unstressed *do* periphrases do not actually differ from simple present tense forms but that past tense *did* often carries habitual aspect. Klemola then takes these corpus findings as a starting point to discuss the historical development of *do*-support in English. Jones/Tagliamonte (2004) further add to our knowledge about the history of periphrastic *do* by considering the constraints operating on preverbal *did* in two corpora, one sampling Samaná English (a variety of English spoken on the Northeastern peninsula of the Dominican Republic, a community which was originally settled by African American ex-slaves), and the other sampling Somerset English, the traditional dialect which is spoken in England's Southwest. Jones/Tagliamonte establish that the internal constraints governing periphrastic *do* variation work in a curiously similar way in the two dialect corpora and that, moreover, the

constraint ranking is exactly the same. Jones/Tagliamonte thus suggest that even relic forms, such as preverbal *did*, follow "diachronic patterns in systematic linguistic conditioning" (2004: 119), and that preverbal *did* "continues to maintain a complex set of constraints [...] that can be traced in the history of English" (2004: 119).

In much the same historical-evolutionary spirit, Pusch (2001) follows up on a historical-evolutionary research question (cf. Pusch 2000 for a partial summary in English), exploring enunciative particles (such as preverbal *que*) in varieties of Gascony Occitan on the basis of the *Corpus Occitano-Gascon*. Through a quantitative and qualitative analysis of the synchronic distribution of the phenomenon in corpus data and by additionally presenting evidence from other languages, Pusch seeks to illuminate the grammaticalization processes which the family of enunciative particles has been undergoing. Pusch argues that the genesis of the phenomenon is best explained in terms of functional and communicative pressures.

Dialect corpora have also been used to shed light on the socio-historical genesis of dialects and language varieties: with an overarching interest in the linguistic consequences of industrialization as a social phenomenon, Grosse et al. (1987), for example, present a historical corpus sampling a variety of dialect materials (such as personal letters, postcards, etc., dating primarily from the second half of the nineteenth century) in order to explore the evolution of Ruhrdeutsch (Ruhr area German).

In sum, what all these case studies have in common is that they draw on naturalistic dialect data to document the genesis, evolution, and/or the resulting synchronic layering of non-standard – or former non-standard – linguistic forms. Thus, corpus-based dialectology is employed to pursue *per se* historical research questions. It seems worth pointing out in this connection that such dialectological inquiry may also serve to trace the consequences of colonial transplantation: Elsig/Poplack (2006), for example, rely on corpora of Quebec French (the *Récits du français québécois d'autrefois* and the *Ottawa-Hull French Corpus*) to document the history of question formation in Québec dialects of French vis-à-vis French French.

5.4. Formalist/generative theory building

If dialectology and corpus linguistics can be said to be strange bedfellows, formalist approaches and corpus linguistics are a methodically even more outlandish pairing. Thanks to the recent theoretical interest in syntactic microvariation (cf. some of the papers in Barbiers/Cornips 2002 and Cornips/Corrigan 2005), however, the past few years have seen a number of formalist linguists exploiting naturalistic dialect corpus data as a source for

authentic examples. To name but a few representative examples: Vangsnes (2005, 221) searches *the Oslo Corpus of Tagged Norwegian Texts* for certain *wh*-pronouns; Westergaard (2003) studies a corpus sampling child and adult data of the Tromsø dialect to demonstrate that this dialect exhibits both V2 and V3 word order; Ledgeway (2005) draws on a corpus of Southern Italian dialect texts to shed light on the dual complementizer system of those dialects (this particular study even features some quantitative analyses). Other formal analysts have relied on dialect corpora as a means to check on the reliability of dialect atlas data (cf., for instance, Cornips 2002 on variation between infinitival complementizers in Dutch dialects).

As for formalist theory rejection (an exercise which certainly comes more easily to most corpus linguists), of course, corpus data have been used to demonstrate *ex negativo* that formalist accounts of dialect phenomena are insufficient. For example, Pietsch (2005) can be seen as an extended empirical argument that formal approaches to the Northern Subject Rule cannot explain the observable range of variability in corpus data, and Anderwald (to appear) uses non-standard past tense forms to argue for a usage-based model of language processing.

6. Some issues in dialect corpus analysis

In a number of ways, corpus-based dialectology is subject to methodological and technical challenges that arise out of the particular nature of the data studied. For instance, mapping dialect data to digitized text is not a trivial task.

In the absence of general guidelines many idiosyncratic solutions exist (but cf. the sensible guidelines in Tagliamonte 2006). A purely phonetic transcription, while feasible in principle, is prohibitively labor-intensive to generate. In addition, a narrow phonetic transcription would make impossible the automatic retrieval of most phenomena – something that corpus linguistics is, after all, designed to do. A corpus search for a particular form would become extremely tedious, if not impossible, because each potential phonological or phonetic variant form would have to be considered separately, many of which may not be known to the researcher without extensive supplemental information. Indeed, for many studies this kind of detail is not only unnecessary, but might be a direct hindrance. Not surprisingly, then, most dialect corpora do not contain a phonetic or phonemic transcription of the data (but cf. the questionnaire-based Dutch SAND project, which works with several tiers, one of which is a phonetic transcription, cf. Barbiers et al. 2005). Researchers are thus usually left with the other bad alternative: creating some kind of orthographic representation. The general problem is that this means adapting the standard orthography to some considerable degree. It is clear

that the accepted codified orthography of any standard language has been optimized for the standard, although this optimal fit may be some past stage, as the largely fossilized orthographies e.g. of French or English testify (also cf. article 30 on this issue).

Differences in lexis may constitute a first problem, as for purely dialectal words a consistent spelling may never have been devised. Researchers involved in transcribing dialect data have therefore also always paid particular attention to possible dialect literature, where lexical items may be documented (if not consistently). A general guideline has been to exclude phonetic and phonological information in the dialect representation, but to include morphological information. (However, this distinction might require some in-depth knowledge of the phenomenon under investigation.)

Finally, only those phenomena should be included that are genuine dialect features, rather than very general features of allegro speech on the one hand, or simple "eye dialect" features on the other. Eye dialect features are meant to somehow convey, by means of orthography, the special quality and subjective phonological "feel" of speech (consider, in the case of English, <Whatcher thinkin?> for *What are you thinking?*). Thus, normal processes of spoken language (not necessarily dialectal) are rendered orthographically, something which does not contribute to linguistic heuristics, but only serves to degrade and stigmatize the speaker in question (on this last topic, cf. especially Preston 2000).

Ideally, all decisions made during the transcription process should be documented in a transcription protocol and published with the material at hand, although this is rarely the case. It is fair to say, then, that current corpus technology – as employed in corpus-based dialectology – is not especially well geared to deal with dialect data.

Another concern in corpus-based dialectology is sampling. For somewhat pragmatic reasons (that is, to obtain speech characterized by a sufficiently large number of dialect features) and to ensure comparability to older materials, many English dialect corpora, as pointed out above, heavily rely on sampling non-mobile old rural males as the most traditional, broadest dialect speakers. Yet, this design choice plays havoc with a number of sociolinguistic research questions that would require a balanced, representative, stratified database – which, however, would be less likely to yield a sufficient number of dialect features.

This leads us to a final, more general issue. All the limitations and troubles that are inherent to any corpus approach to language are, in a number of ways, even more clearly apparent in corpus-based dialectology. Thus, for instance, many phenomena theoretically interesting to dialectologists are excruciatingly rare in corpus data. Note that the bulk of

corpus-based dialectology research has sought to investigate dialect data from a morphosyntactic perspective. It is well-known that the study of morphological or syntactic phenomena necessitates much larger databases of naturalistic data than the study of, e.g., phonology or lexis. A concrete example may illustrate this problem: double modal constructions are known to recur in a number of non-standard varieties of English, for example in Scottish English (cf. Brown 1991). Yet, the FRED corpus, after all containing 2.5 million words from all major dialect areas in Great Britain including Scotland and the North, yields but one (!) clear example of a double modal construction (*it* might should *tell you on the tickets how much luggage you're allowed to take*, FRED MLN_005), and one fairly questionable example (*because you said a Italian* would could *stab you in the back*, FRED DEV_003). With the pragmatic contexts that license double modals being so rare, then, the conclusion is that a thorough analysis of double modal constructions would probably require a dialect corpus spanning several hundreds of millions of words – a size which is, needless to say, illusionary. Similar frequency issues (which are by no means particular to corpus-based dialectology; cf. article 37) bedevil the corpus-based study of other interesting but rare dialect features as well: Hollmann/Siewierska (2006), for instance, note that material from no less than five corpora (among them, FRED and the BNC) is still insufficient to conclusively study the contexts of ditransitive verb complementation (of the type *He gave it me* or *He gave me it*) in the Lancashire dialect of English; in the same vein, Tagliamonte (1998) is unable to quantitatively investigate collective subjects as a determinant of *was/were* variation in York English since in spite of the considerable total size of her corpus, collective nouns are not sufficiently attested.

7. Conclusion

Corpus-based dialectology can certainly constitute a very important first step in the study of micro-variation (intra-, rather than inter-linguistic variation). Hollmann/Siewierska (2006) argue that corpus-based dialectology must be complemented by other methods such as elicitation tasks and attitude questionnaires. In the case of double modals, for example, it is often claimed that elicitation tasks and questionnaires tapping speakers' intuitions are needed to determine the exact syntactic, not to mention pragmatic circumstances determining the possible structure, combinations as well as the use of these constructions (cf. Montgomery 1998 for an overview). To date, however, follow-up studies employing a different methodology after a corpus investigation are rarely carried out. Nevertheless, it should be pointed out that despite methodological drawbacks and theoretical problems, the use of

corpora in dialectology is still an under-developed and under-researched area that merits much more attention. We are only just beginning to tap this rich resource of natural language data, which might enable us to increase our knowledge of the constraints and parameters of linguistic variation considerably. Especially the use of sufficiently large corpora of dialect speech might allow us to consider intra-linguistic variation in a much less haphazard and idiosyncratic way than has been done so far, and might thus constitute one of the most promising avenues for further research.

8. Literature

Andersen, G. (2001), *Pragmatic Markers and Sociolinguistic Variation*. Amsterdam/Philadelphia: Benjamins.

Anderwald, L. (2002), *Negation in Non-standard British English: Gaps, Regularizations, Asymmetries*. (Studies in Germanic Linguistics.) London/New York: Routledge.

Anderwald, L. (2003), Non-standard English and Typological Principles: The Case of Negation. In: Rohdenburg, G./Mondorf, B. (eds.), *Determinants of Grammatical Variation in English*. Berlin/New York: Mouton de Gruyter**,** 507-530.

Anderwald, L. (to appear), *The Morphology of English Dialects: Verb-formation in Non-standard English*. Cambridge: Cambridge University Press.

Anderwald, L./Wagner, S. (2007), FRED - the Freiburg English Dialect Corpus. In: Beal, J./Corrigan, K. P./Moisl, H. (eds.), *Creating and Digitizing Language Corpora*, Vol. 1: *Synchronic Databases.* London: Palgrave Macmillan, 35-53.

Barbiers, S./Bennis, H./De Vogelaer, G./Devos, M./van der Ham, M./Haslinger, I./van Koppen, M./van Craenenbroeck, J./van den Heede, V. (eds.) (2005), *Syntactic Atlas of the Dutch Dialects (Sand).* Amsterdam: Amsterdam University Press.

Barbiers, S./Cornips, L. (2002), *Syntactic Microvariation*. Electronic publication of the Meertens Instituut, available at: http://www.meertens.knaw.nl/books/synmic/.

Beal, J. C. (2004), *English in Modern Times, 1700-1945*. London: Arnold.

Beal, J. C./Corrigan, K. P. (2006), *No, Nay, Never*: Negation in Tyneside English. In: Iyeiri, Y. (ed.), *Aspects of English Negation*. Amsterdam: Benjamins, 139-157.

Brown, K. (1991), Double Modals in Hawick Scots. Dialects of English: Studies in Grammatical Variation. In: Trudgill, P./Chambers, J. (eds.), *Dialects of English: Studies in Grammatical Variation.* London/New York: Longman**,** 74-103.

Chambers, J. K./Trudgill, P. (1998), *Dialectology*. Cambridge: Cambridge University Press.

Cornips, L. (2002), Variation between the Infinitival Complementizers *om/voor* in Spontaneous Speech Data Compared to Elicitation Data. In: Barbiers/Cornips 2002, 75-96.

Cornips, L./Corrigan, K. P. (eds.) (2005), *Syntax and Variation: Reconciling the Biological and the Social*. Amsterdam/Philadelphia: Benjamins.

Elsig, M./Poplack, S. (2006), Transplanted Dialects and Language Change: Question Formation in Québec. In: *Penn Working Papers in Linguistics* 12(2), 77-90.

Geyer, K. (2003), *Hetzlerisch: Dokumentation spontansprachlicher Texte und grammatische Analyse der phorischen Pronomina im ostfränkischen Dialekt des Dorfes Hetzles*. München: Lincom Europa.

Gilliéron, J. (1902-1910), *Atlas linguistique de la France*. Paris: Champion.

Grosse, S./Grimberg, M./Hölscher, T./Karweick, J./Kuntz, H. (1987), Sprachwandel und Sprachwachstum im Ruhrgebiet des 19. Jahrhunderts unter dem Einfluß der Industrialisierung. In: *Zeitschrift für Dialektologie und Linguistik* 54(2), 202-221.

Hall, J. H. (ed.) (1985-2009 [scheduled]), *The Dictionary of American Regional English*. Cambridge, MA: Harvard University Press.

Hernández, N. (2002), A Context Hierarchy of Untriggered Self-forms in English. In: *Zeitschrift für Anglistik und Amerikanistik* 50(3), 269-284.

Herrmann, T. (2005), Relative Clauses in English Dialects of the British Isles. In: Kortmann et al. 2005, 21-124.

Hollmann, W./Siewierska, A. (2006), Corpora and (the Need for) Other Methods in a Study of Lancashire Dialect. In: *Zeitschrift für Anglistik und Amerikanistik* 54(1), 21-34.

Huber, M. (2003), The Corpus of English in South-east Wales and its Synchronic and Diachronic Implications. In: Tristram, H. L. C. (ed.), *The Celtic Englishes III*. Heidelberg: Winter**,** 182-200.

Ihalainen, O. (1991), On Grammatical Diffusion in Somerset Folk Speech. In: Trudgill, P./Chambers, J. (eds.), *Dialects of English: Studies in Grammatical Variation*. London/New York: Longman**,** 104-119.
Ihalainen, O. (1994), The Dialects of England since 1776. In: Burchfield, R. (ed.), *Cambridge History of the English Language*. Vol. 5: *English in Britain and Overseas: Origins and Development.* Cambridge: Cambridge University Press, 197-274.
Jones, M./Tagliamonte, S. (2004), From Somerset to Samaná: Preverbal *did* in the Voyage of English. In: *Language Variation and Change* 16, 93-126.
Jones, V. (1985), Tyneside Syntax: A Presentation of Some Data from the *Tyneside Linguistic Survey*. In: Viereck, W. (ed.), *Focus on England and Wales.* Amsterdam: Benjamins, 163-178.
Keenan, E./Comrie, B. (1977), Noun Phrase Accessibility and Universal Grammar. In: *Linguistic Inquiry* 8(1), 63-99.
Kirk, J. M. (1990), *Northern Ireland Transcribed Corpus of Speech*. Colchester: Economic and Social Research Council Data Archive, University of Essex.
Klemola, J. (1996), Non-standard Periphrastic *do*: A Study in Variation and Change. Unpublished PhD thesis. Essex: University of Essex, Department of Language and Linguistics.
Klemola, J. (2002), Continuity and Change in Dialect Morphosyntax. In: Kastovsky, D./Kaltenböck, G./Reichl, S. (eds.), *Anglistentag 2001 Wien*. Trier: Wissenschaftlicher Verlag Trier, 47-56.
Klemola, J. (2003), Personal Pronouns in the Traditional Dialects of the South West of England. In: Tristram, H. L. C. (ed.), *The Celtic Englishes III*. Heidelberg: Winter**,** 260-275.
Klemola, J./Jones, M. J. (1999), The Leeds Corpus of English Dialects-project. In: *Leeds Studies in English* 30, 17-30.
Kortmann, B. (ed.) (2004), *Dialectology Meets Typology: Dialect Grammar from a Cross-linguistic Perspective*. Berlin/New York: Mouton de Gruyter.
Kortmann, B./Herrmann, T./Pietsch, L./Wagner, S. (eds.) (2005), *A Comparative Grammar of English Dialects*. Berlin/New York: Mouton de Gruyter.
Labov, W. (1972), *Sociolinguistic Patterns.* Philadelphia: Philadelphia University Press.
Ledgeway, A. (2005), Moving through the Left Periphery: The Dual Complementiser System in the Dialects of Southern Italy. In: *Transactions of the Philological Society* 103(3), 339-396.
Milroy, L./Milroy, J./Docherty, G. (1997), *Phonological Variation and Change in Contemporary Spoken British English: Final Report to the UK Economic and Social Research Council, grant No. R000234892*. Department of Speech, University of Newcastle-upon-Tyne.
Montgomery, M. B. (1998), Multiple Modals in LAGS and LAMSAS. In: Montgomery, M. B./Nunnally, T. (eds.), *From the Gulf States and Beyond: The Legacy of Lee Pederson and LAGS*. Tuscaloosa/London: University of Alabama Press, 90-122.
Orton, H./ Barry, M. V. (eds.) (1969-1971), *Survey of English Dialects. The West Midland Counties*. Leeds: Arnold.
Orton, H./Halliday, W. J. (eds.) (1962-1964), *Survey of English Dialects. The Six Northern Counties and the Isle of Man.* Leeds: Arnold.
Orton, H./Tilling, P. M. (eds.) (1969-1971), *Survey of English Dialects. The East Midland Counties and East Anglia.* Leeds: Arnold.
Orton, H./Wakelin, M. F. (eds.) (1967-1968), *Survey of English Dialects. The Southern Counties.* Leeds: Arnold.
Patocka, F. (1997), *Satzgliedstellung in den bairischen Dialekten Österreichs*. Frankfurt am Main: Peter Lang.
Pietsch, L. (2005), *Variable Grammars: Verbal Agreement in Northern Dialects of English.* Tübingen: Niemeyer.
Poplack, S. (1980), The Notion of the Plural in Puerto Rican Spanish: Competing Constraints on (s) Deletion. In: Labov, W. (ed.), *Locating Language in Time and Space*. New York: Academic Press, 55–67.
Preston, D. (2000), 'Mowr and Mowr Bayud Spellin': Confessions of a Sociolinguist. In: *Journal of Sociolinguistics* 4, 614-621.
Pusch, C. D. (2000), The Attitudinal Meaning of Preverbal Markers in Gascon: Insights from the Analysis of Literary and Spoken Language Data. In: Andersen, G./Fretheim, T. (eds.), *Pragmatic Markers and Propositional Attitude*. Amsterdam: Benjamins, 189-206.
Pusch, C. D. (2001), *Morphosyntax, Informationsstruktur und Pragmatik: Präverbale Marker im gaskognischen Okzitanisch und in anderen Sprachen*. Tübingen: Gunter Narr.
Scherre, M. M. P./Naro, A. J. (1991), Marking in Discourse: "Birds of a Feather". In: *Language Variation and Change* 3(1), 23-32.
Schüller, D. (ed.) (2003), *Tondokumente aus dem Phonogrammarchiv der Österreichischen Akademie der Wissenschaften: 'Dazähl'n'*, [CD]. Wien: Veröffentlichungen der Österreichischen Akademie der Wissenschaften.
Shorrocks, G. (1980), *A Grammar of the Dialect of Farnworth and District.* Sheffield: University of Sheffield.
Shorrocks, G. (1999), *A Grammar of the Dialect of the Bolton Area*. Frankfurt am Main etc.: Peter Lang.

Stenström, A-B./Andersen, G./Hasund, I. K. (2002), *Trends in Teenage Talk: Corpus Compilation, Analysis and Findings.* Amsterdam/Philadelphia: Benjamins.
Szmrecsanyi, B. (2006), *Morphosyntactic Persistence in Spoken English: A Corpus Study at the Intersection of Variationist Sociolinguistics, Psycholinguistics, and Discourse Analysis*. Berlin/New York: Mouton de Gruyter.
Tagliamonte, S. (1998), *Was/were* Variation across the Generations: View from the City of York. In: *Language Variation and Change* 10(2), 153-191.
Tagliamonte, S. (2001), *Come/came* Variation in English Dialects. In: *American Speech* 76(1), 42-61.
Tagliamonte, S. (2003), 'Every Place Has a Different Toll': Determinants of Grammatical Variation in a Cross-variety Perspective. In: Rohdenburg, G./Mondorf, B. (eds.), *Determinants of Grammatical Variation in English*. Berlin/New York: Mouton de Gruyter**,** 531-554.
Tagliamonte, S. (2006), *Analysing Sociolinguistic Variation*. Cambridge: Cambridge University Press.
Tagliamonte, S./Lawrence, H. (2000), 'I used to Dance, but I don't Dance Now': The Habitual Past in English. In: *Journal of English Linguistics* 28, 324-353.
Tagliamonte, S./Smith, J. (2000), Old *was*, New Ecology: Viewing English through the Sociolinguistic Filter. In: Poplack, S. (ed.), *The English History of African American English*. Malden, MA: Blackwell**,** 141-171.
Tagliamonte, S./Smith, J. (2002), 'Either it isn't or it's not': NEG/AUX Contraction in British Dialects. In: *English World-Wide* 23(2), 251-281.
Tagliamonte, S./Smith, J. (2005), 'No Momentary Fancy!' The *Zero* 'Complementizer' in English Dialects. In: *English Language and Linguistics* 9(2), 289-309.
Tagliamonte, S./Smith, J./Lawrence, H. (2005), 'No Taming the Vernacular!' Insights from the Relatives in Northern Britain. In: *Language Variation and Change* 17, 75-112.
Tagliamonte, S./Temple, R. (2005), New Perspectives on an Ol' Variable: (t,d) in British English. In: *Language Variation and Change* 17, 281-302.
Travis, C. E. (2007), Genre Effects on Subject Expression in Spanish: Priming in Narrative and Conversation. In: *Language Variation and Change* 19(2), 1-35.
Vangsnes, Ø. A. (2005), Microparameters for Norwegian *Wh*-grammars. In: *Linguistic Variation Yearbook* 5(1), 187-226.
Vasko, A-L. (2005) *UP CAMBRIDGE. Prepositional Locative Expressions in Dialect Speech: A Corpus-based Study of the Cambridgeshire Dialect.* Helsinki: Société Néophilologique.
Wagner, S. (2004), 'Gendered' Pronouns in English Dialects - a Typological Perspective. In: Kortmann 2004, 479-496.
Westergaard, M. R. (2003), Word Order in *Wh*-questions in a North Norwegian Dialect: Some Evidence from an Acquisition Study. In: *Nordic Journal of Linguistics* 26, 81-109.
Wright, J. (1898-1905), *The English Dialect Dictionary*. Oxford: Frowde.