The *Helsinki Dialect Corpus*: Characteristics of Speech and Aspects of Variation

Kirsti Peitsara and Anna-Liisa Vasko

1.The *Helsinki Dialect Corpus* (Kirsti Peitsara)

The *Helsinki Dialect Corpus of British English (HD)* is a collection of orthographically transcribed tape-recorded speech, mainly from rural East Anglia and the South-West, with a minor collection from Lancashire. The recordings were made in the 1970s and the late 1980s by Finnish Post-Graduates, advised in this work by Professor Orton and other Leeds scholars. The aim was to create a computer corpus consisting of enough continuous speech to provide material for the study of dialectal morphosyntax and thus to supplement the *Leeds Survey of English Dialects*. For this reason, the informants were chosen according to the criteria used in the *Leeds Survey*: elderly native speakers in rural occupations, who had lived locally most of their lives. The average age of the informants was 60-70 years (the eldest being 90 years) at the moment of recording.

In Finland, the work was first supervised by Professor Tauno F. Mustanoja and, after his retirement, by Professor Ossi Ihalainen, under whose leadership the first computer version was prepared. At present the project for completing the computer version is supervised by Kirsti Peitsara. The machine-readable corpus (Word Cruncher version) will include nearly 0.8m words consisting of seven sub-corpora, named after their collectors:

The Ihalainen Somerset Corpus (c.165,000 words)
The Stigell Devon Corpus (c. 85,000 words)
The Tammivaara-Balaam Isle of Ely Corpus (c. 84,000 words)
The Ojanen-Vasko Cambridgeshire Corpus (c. 210,000 words)
The Pasanen Suffolk Corpus (c. 200,000 words)
The Kerman Essex Corpus (c. 34,000 words)
The Kerman Lancashire Corpus (c. 55,000 words)

At present the fieldworkers have full copyright of their material, but some of it can be used for research at the English Department research centre in Helsinki by mutual agreement (contact: anna-liisa.vasko(at)helsinki.fi [contact information changed 15 February 2011]).

As has been shown in pilot studies made by Kirsti Peitsara and her students, a corpus of continuous speech can be used to get more information on aspects of dialectal morphosyntax through quantitative research. It also provides a possibility to observe the variation of the features in their linguistic contexts and with different speakers, both male and female. Though the method of free interview did not perhaps ensure enough data on some linguistic aspects (e.g. the past tense tends to predominate over the present tense), the *HD* is an improvement in some respects that the *SED* has been criticized for (e.g. its questionnaire-based method only produced emphatic variants of words in answers).

The main way of using the *HD* is for morphosyntactic study. The late Professor Ossi Ihalainen's studies on his Somerset Corpus contributed, for example, to our knowledge of the uses of the auxiliary *do*, negation (*idn', wadn')* and the pronominal system (*'n* 'it') in Somerset English. Anna-Liisa Vasko's pilot studies on her Cambridgeshire Corpus indicated the potential of the corpus for the study of prepositions, the topic of her Doctoral Thesis (2005 [updated 10 March 2011]). My studies on the Pasanen Suffolk Corpus show that it provides an important addition to the information in the *SED*, which was based on relatively few informants in this area. For example, for the subject relativizer with an animate antecedent, the *SED*-based atlases divide Suffolk in three areas with the zero relativizer prevailing in central Suffolk, *who* in eastern and western parts of the area. According to my study (2002 [updated 10 March 2011]), the zero relativizer is predominant in the Suffolk dialect, *that*, *what* and *who* being about equal as secondary variants. The East Anglian corpora in the *HD* also prove that personal pronoun *that* (*that is raining*, etc.) was a common feature in this area, not only a relic feature in north-eastern Norfolk. Work on other morphosyntactic features is in progress.

Preliminary lexical studies on the *HD* show that such items as variants for 'anyway' (*anyway*, *anyhow*, *anyrate*, *anyroad*) may be both regionally and idiolectally distributed, while the use of discourse elements, such as *you know/you see/see/you know what I mean* or *like* (*we puts in catch crops, you know, like*), would rather seem to be idiolectal than dialectal features. Conclusive studies on these and other aspects are still to be made.

The orthographical transcription of the *HD* allows some types of phonological study, since *h*-dropping and other types of elision (*'leven* 'eleven', *'company* 'accompany', etc.), as well as intrusive *r* and other intrusive sounds (e.g. *Januwary*) are indicated. So are conspicuous diagnostic sounds, as far as this is possible using the ordinary orthography (e.g. fricative voicing in Somerset: *zometimes*, etc.). Pilot studies show that *h*-dropping seems to be about equally frequent (somewhat more than 30% of the *h*-words) in the Somerset, Devon and Isle of Ely corpora, though the prevailing maps (after *SED*) classify Somerset as *h*-retaining and the other two areas *h*-dropping. In the case of Suffolk (7%), the corpus results agree with the prevailing idea of it as an *h*-retaining area. Hypercorrection (*hengineer*, *hopen*, etc.), again, seems to occur particularly in the Isle of Ely and Devon dialects. (Here I am indebted to Susanna Saarela who conducted a frequency study on *h*-dropping and hypercorrection as a course paper.) The details of *h*-dropping and hypercorrection (the phonetic surroundings, idiolectal preference, etc.) have not yet been examined.

2. Research on spoken language: Characteristics of speech, with a focus on the dialect of Cambridgeshire (Anna-Liisa Vasko)

To study speech, it is necessary to use the written form of language, the transcriptions. Transcriptions necessarily lack some phonetic and prosodic information to fully represent the complexity of spoken language. The forms of speech and writing (acoustic-vocalic vs. graphic; auditive vs. visual) are completely different and, thus, inevitably contain features present in one form only which cannot be fully represented in the other. One of these is a situation in which speakers are talking simultaneously, which in writing has to be presented in a linear form.

Naturally, the accuracy of transcription depends on the purpose of study. In a corpus intended mainly for the study of morphosyntactic features, orthographic transcription is usually sufficient. It may, however, not be so easy to decide the spelling of a certain construction. Consider example (1):

(1) *Lovely, [wɔnt], Mum?* (BR, Bassingbourn)

In order to decide how to write [wɔnt], which sounds like standard English (henceforth StE) *want,* more context is needed, of course:

> [WIFE:] *Chittlins.*
> [BR:] *An' that good many folk used to = ha' these bellies, clean them, cook them, an' eat them. Lovely, [wɔnt], Mum?*

The context would imply the StE interpretation 'They (i.e. chitterlings, pigs' intestines) were lovely, weren't they, Mum?' Another interpretation could be: **'**It was lovely, wasn't it, Mum?' In that case *Chittlins* would be considered a collective expression. The third possibility might be a syntactic blend: 'They were lovely, wasn't it, Mum?' A syntactic blend (or anacoluthon) is not uncommon, as dialect speech tolerates a freedom of syntactic structure that would generally be regarded as unacceptable in writing.

The spelling *Lovely, wan' 't, Mum?* was chosen, because question tags usually have a subject and the pronoun *they* doesn't usually occur in the form *t* in the Cambridgeshire dialect. On the other hand, the form *wan'* [*wɔn*] is frequently used for both StE *weren't* and *wasn't*.

Another problem connected with the study of spoken language is how to recognize speech units, meaningful entities in the speech of individual speakers. In grammar, the sentence has traditionally been treated as the fundamental structural unit.[1] In dialect speech, like in conversational language in general, such a unit does not really exist. There does not seem to be any reliable method for defining sentences in spoken language in terms of their syntactic form or semantic content. It is true that orthographic transcriptions conventionally contain sentence-final punctuation marks (periods, question marks, and exclamation marks). They are, naturally, inserted by the transcriber as cues reflecting features such as change of topic, falling intonation and pause in the stream of speech. It is, however, typical of dialect speech to proceed in a long sequence of paratactic units without any indication of such features. Thus, punctuation in dialect

[1] In discourse analysis researchers go beyond the sentence. (Quirk *et al*. 1985 Ch. 11, 19.)

transcription does not necessarily follow the rules of the standard; its function is mainly to help the reader.

In the short extract of an interview below a few typical features of the Cambridgeshire dialect are illustrated. Some of these features can be found in informal spoken English, and others are characteristic of dialect speech only. The participants in this part of the interview are the informant [ES] from Willingham, an 82-year-old local man, and the interviewer [MH]. The latter, also a local man, is bilingual, using the standard with standard speakers and the local dialect when talking to speakers of the local dialect. In the text the informant's words are italicized, people referred to are anonymized by using an * for every syllable of the name, pauses are indicated with = and unclear passages with %---%.

> [MH:] Oh yes. Yeah, yeah. They dug = they dug gravel in the in the sandpits then, din't they?
> [ES:] *Yeah, well, I'll tell you what they uset, us ol' us ol' boys up Cut- Cutter End they used to call us Cutter-eenders up there.*
> [MH:] Yeah yeah.
> [ES:] *That were always Cutter Eend. I tell you a lot about ol' ** give me a ride out Farmer's Fen what we'll we'll get on 'bout the gravel diggin'=*
> [MH:] Right.
> [ES:] *= what ** ** and * * * used to dig gravel then, they = 'ey started in the middle.*
> [MH:] *Yeah.*
> [ES:] *In the middle o' field.*
> [MH:] Yeah
> [ES:] *Well, us ol' boys we used to go in there dinnertime to school = an' they used to get us in this pit = with their shovel, an' they used to say, "Now then let's see who can chuck the furthest." They did you er, "Who = which which you o' boys he's = let's see which is the best man or %---% to see which one can chuck 'at the furthest." They didn't use to say the farthest chuck 'at the furthest* [LAUGHING]. *And = an' an' = an' afterwards = at sharp winter = them what were out o' work they had to dig right there weren't no ground outside. They had one fourpence a day, had eight bob a week, 'ad twopence an hour = digging gravel, only leave them high trees down the bottom end.*
> [MH:] Oh Yeah.
> [ES:] *Yes they war. But = if you got out o' work that time o' day when you were a young chap …*

The informant [ES] begins with two one-word inserts (*Yeah, well*). He starts telling about gravel digging (*I'll tell you what they uset*), but changes the topic (*us ol' us ol' boys…*). Then he remembers he was going to tell about gravel digging (*we'll we'll get on 'bout the gravel diggin'*). The knowledge shared by the informant and the interviewer is expressed in different ways (e.g. *I tell you a lot about ol' ** give me..*, meaning 'I have told you many times').

Examples like (2) and (3) show us the 'core' sentences, the body of the speaker's message.

(2) *us ol' us ol' boys up Cu- Cutter End they used to call us Cutter-eenders up there.*

(3) *us ol' boys we used to go in there dinnertime to school = an' they used to get us in this pit = with their shovel*

These 'core' sentences are preceded by the preface, 'initial dislocation', *us ol' boys (up Cu- Cutter End)*. This sentence-type shows the general tendency in dialect speech to express first the thought that is foremost in the speaker's mind at the moment of speaking. In StE grammar, example (2) could also be interpreted as showing the extraposed object *us ol' boys up Cu- Cutter End* followed by a pleonastic pronoun (*us*) after the verb *call*. Example (3) again would exemplify the use of a pleonastic pronoun with the subject.

On the other hand, if the speaker has little chance to plan the structure as he proceeds, there is a need to modify the message retrospectively, that is, to 'tag on' as an afterthought, like *with their shovel* in example (3).

Sentences are often left grammatically incomplete, like in example (4):

(4) *I tell you a lot about ol' ** give me a ride out Farmer's Fen what*

Incompletion of this type is natural, because speech is created at the moment of speaking and the structures are not fully premeditated.

Ellipsis of various sorts is frequent.

(5) *them what were out o' work had to dig right there [there]weren't no ground outside*

In StE grammar, ellipsis is 'missing' out of one or more words. In dialect speech nothing may be felt as missing. The speaker adapts his speech to the situation and the understanding of the hearer.

Correction or reformulation is possible only through hesitations, false starts, and other non-fluencies:

(6a) *Who = which which you o' boys he's = let's see which is the best man*

(6b) *us ol' us o' (boys)*

(6c) *Cut- Cutter End*

(6d) *er*

Unlike in written language, this ‘erroneous’ or corrected part cannot be eliminated but remains to be seen in transcription.

In addition to the features illustrated by the extract above, the *Cambridgeshire Corpus* provides data for various other constructions, e.g. the type seen in examples (7) and (8):

(7) *The chap lived in the vil = down Rices Road = what used to work there.* (EW, Swaffham Prior)

(8) [Q:] You used to drive the sheep to market, din’t you?
[SC:] *No, no, I din’t = the man did what used to grind = what see after them = he used to drive ‘em on the road* (SC, Lt. Eversden)

There are two explanations for cases like these: the relativizer does not immediately follow the antecedent, as would be expected in StE. The relative clause is a kind of ‘after-thought’. Thus, one interpretation for example (7) may be ‘The chap who used to work there lived…’ . In dialect speech the relativizer and the antecedent are frequently separated, for example, by an adverb or a parenthesis. So for example in the Suffolk dialect: *Well, I went with ‘im, to this here pub HERE what I had*; *I done some work there = a lot on the house, YOU SEE, what they were building* (Peitsara 1985: 353). Another possibility is that *what* is used ‘independently’ in the sense ‘the one who’. This latter interpretation is reinforced in examples like (9) and (10):

(9) *We used to go Sundays or what* (i.e. ‘those who’) *looked after the ’orses.* (AS, Harston)

(10) *But it een’t a shovel, is it? What* (i.e. ‘The one [that]’) *I’m*[2]_*described* (SS, Willingham)

To sum up, generally speaking, the characteristics of informal spoken English can be detected in dialect speech as well. However, some of them are ‘taken a step further’ and occur in ‘more advanced’ forms. In the typical on-going flow of dialect speech clausal and non-clausal elements are ‘woven’ together even more freely than in more formal varieties of the spoken language. Dialect speech is created in real time, ‘on the spot‘. Speakers proceed without a premeditated plan and may often change the structure in the middle of a sentence. Consequently, dialect speech is characterized by pauses, hesitations and repetitions. To save time and energy, speakers aim to reduce the length of what they have to say by not expressing the words not essential for understanding, which often results in ellipsis from the point of view of the standard language. Prefaces and tags are common as well. The effect of such devices is to eliminate complex structures from the ‘core’ sentence. When analysing dialect speech, it is, thus, necessary to see the structure in a context large enough to find out patterns typical of individual speakers.

[2] Notice that *‘m* is the auxiliary of the perfect tense, not of the passive.

At present, the system of analyzing dialect speech is still, to a large extent, based on that of the written language. It is, however, doubtful to what extent the tools of grammatical analysis developed in and for the study of the written language are applicable to the spoken language. Should we use the terminology of the written language and speak about sentences or clauses, for example? Or could we analyze speech at unit level, i.e. at the level of meaningful entities in the speech of individual speakers?

E-mail: anna-liisa.vasko(at)helsinki.fi (Kirsti Peitsara's email is not publically available; she can be reached through the editor or her co-author.) [updated 10 March 2011]

REFERENCES

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad & Edward Finegan (eds) (1999) *Longman Grammar of Spoken and Written English.* London: Longman.

Brown, Gillian (1990[1977]) *Listening to Spoken English*. London/New York: Longman.

Ihalainen, Ossi (1985) "*He took the bottle and put 'n in his pocket:* the object pronoun *it* in present-day Somerset." In Viereck, Wolfgang (ed) *Focus on: England and Wales*. Varieties of English around the World GS 4. Amsterdam: John Benjamins. 153-161.

Ihalainen, Ossi (1987) "Towards a grammar of the Somerset dialect." In Kahlas-Tarkka, Leena (ed) *Neophilologica Fennica*. Mémoires de la Société Néophilologique de Helsinki 45. Helsinki: Société Néophilologique. 71-86.

Ihalainen, Ossi (1990) "A source of data for the study of English dialect syntax: the Helsinki Corpus." In Aarts, Jan & Willem Meijs (eds) *Theory and Practice in Corpus Linguistics.* Amsterdam: Rodopi. 83-103.

Ihalainen, Ossi (1991) "Periphrastic *do* in affirmative sentences in the dialect of East Somerset." In Trudgill, Peter & J.K. Chambers (eds) *Dialects of English: Studies in Grammatical Variation*. London: Longman. 148-160.

Ihalainen, Ossi (1991) "A point of verb syntax in south-western British English: an analysis of a dialect continuum." In Johansson, Stig & Anna-Brita Stenström (eds) *English Computer Corpora. Selected Papers and Research Guide.* Topics in English Linguistics 3. Berlin: Mouton de Gruyter. 201-214.

Ihalainen, Ossi (1994) "The dialects of England since 1776." In Burchfield, Robert (ed) *The Cambridge History of the English Language. Volume V: English Language in Britain and Overseas. Origins and Development.* Cambridge: Cambridge University Press. 197-274.

Ojanen (Vasko), Anna-Liisa (1985) "Use and non-use of prepositions in spatial expressions in the dialect of Cambridgeshire." In Viereck, Wolfgang (ed) *Focus on: England and Wales*. Varieties of English around the World GS 4. Amsterdam: John Benjamins. 179-212.

Orton, Harold (1962) *Survey of English Dialects: Introduction.* Leeds: E.J. Arnold & Son.

Orton, Harold, Stewart Sanderson & John Widdowson (1978) *The Linguistic Atlas of England*. London: University of Leeds.

Peitsara (Kekäläinen), Kirsti (1985) “Relative clauses in the dialect of Suffolk.” *Neuphilologische Mitteilungen* 86. 353-357.

Peitsara, Kirsti (1996) “Studies on the structure of the Suffolk dialect.” In Klemola, Juhani, Merja Kytö & Matti Rissanen (eds) *Speech Past and Present, Studies in English Dialectology in Memory of Ossi Ihalainen*. Bamberger Beiträge zur Englischen Sprachwissenschaft 38. Frankfurt am Main: Peter Lang. 284-307.

Peitsara, Kirsti (2000) “The prepositions ON and OF in partitive and temporal constructions in British English dialetcs.” *Neuphilologische Mitteilungen* 101:2. 323-332.

Peitsara, Kirsti (2001) ”Englannin kielen vaihtelu ja muutos: Puhutun kielen tutkimus.” Tieteen päivät 11.1.2001. Unprinted.

Peitsara, Kirsti (2002) “Relativizers in the Suffolk dialect.” In Poussa, Patricia & Magnus Lundberg (eds) *Relativisation on the North Sea Littoral*. Lincom. 167-180. [reference updated 10 March 2011]

Quirk, R., S. Greenbaum, G. Leech and J. Svartvik (1985) *A Comprehensive Grammar of the English Language.* London/New York: Longman.

Vasko (Ojanen), Anna-Liisa (2005) *Up Cambridge. Prepositional locative expressions in dialect speech: a corpus-based study of the Cambridgeshire dialect. Mémoires de la Société Néophilologique de Helsinki* LXV. Helsinki: Société Néophilologique. [reference added 10 March 2011]