# The Leeds Corpus of English Dialects – Project

Juhani Klemola and Mark J. Jones

*Abstract*

This paper presents an introduction to the methods employed in the Leverhulme Trust funded *Leeds Corpus of English Dialects* – project. The aim of the project was to transcribe and edit the Survey of English Dialects tape-recordings for publication in machine-readable form. The paper is organised as follows: sections 1 and 2 provide some background information about the Survey of English Dialects (SED) and the SED tape-recordings; section 3 summarises the transcription conventions adopted in our project; section 4 gives some examples of how the transcription conventions were applied; section 5 discusses some examples of data culled from the tape-recordings; and section 6 offers some final remarks on the importance and value of the SED tape-recordings.

## 1. *Some background information about the SED*

The Survey of English Dialects is the only detailed nation-wide dialect survey which has ever been conducted in England. The idea of a nation-wide survey in England was developed by the Swiss dialectologist, Professor Eugen Dieth and Professor Harold Orton of Leeds. The fieldwork for the SED was undertaken in the 1950s in predominantly rural communities in England by 9 SED fieldworkers. The fieldworkers went through the same detailed questionnaire of over 1300 questions in 313 localities in order to collect comparable information on regional vocabulary, grammar, and pronunciation. The fieldworkers documented the informants' answers, together with other unsolicited information, into fieldworker notebooks in narrow phonetic transcription as the interviews progressed. The collected and edited transcriptions were published between 1962 and 1971 in twelve books – comprising

some 5500 pages – as the *Survey of English Dialects* (B)*: The Basic Material* (Orton et al. 1962-1971). The SED Basic Material have to date provided the data for hundreds of studies on English dialects, including eight different linguistic atlases, a dictionary, and scholarly articles and monographs (see Viereck 1991 for a selective bibliography of studies based on the SED Basic Material). Very many SED-based publications continue to be added to the list every year.

## 2. *SED tape-recordings*

Soon after the SED fieldwork began in the early 1950s, however, it was decided that the fieldworkers would – in addition to collecting the questionnaire data – also make tape-recordings of casual conversations with the informants. These tape-recordings were made in just under 300 of the 313 SED localities. In the SED publication programme Professor Orton lists a planned volume of phonetic transcriptions 'in both broad and narrow systems' of the tape-recorded material (Orton 1962: 21-22). This part of the SED publication programme was unfortunately never realised, and these unique tape-recordings have remained – stored in the basement of the School of English building at Leeds – for the most part unedited and untranscribed for over 40 years. The material has never been systematically used for research purposes.

The tape-recordings are in the typical dialect interview mould. Harold Orton describes the process of making these tape-recordings in the following terms:

> The material procured was never rehearsed, and, of course, never recited. It was spontaneous, and as a rule consisted of personal reminiscences or opinions, or discussed some task connected with the speaker's occupation, e.g. ploughing, harvesting, hedging, stacking, pig-killing, bread-making. The themes would crop up naturally – so it seemed to the informant – in the course of his conversation with the fieldworker, who, by further skilful management, would ensure that these informal and uninhibited remarks were secured on his tapes for permanent record.
>
> (Orton 1962: 19.)

The original recordings were made with a reel-to-reel tape-recorder. Tape in the 1950s, however, was very expensive, and thus it was not feasible to save the original tape-

recorded interviews in their entirety. In the words of Harold Orton: 'We ourselves felt unable, because of the high cost of tapes and of the lack of the appropriate storage, to preserve the tape-recordings intact. So it was decided to excerpt the best parts only and to re-record these on double sided 12 in. disks' (Orton 1962: 20).

The fact that large sections of the original tape-recordings thus had to be wiped out is, of course, a terrible loss for English Dialectology today. But, to try to find a brighter side to this sorry affair, it is highly likely that the technical quality of the selections from the original interviews transferred onto 78 rpm shellac discs has remained much better than it would have been had the original reel-to-reel tapes been kept instead.

What then are some of the advantages and disadvantages of the SED tape-recordings? In our opinion, the most important thing about the SED tape-recording corpus is that it represents

- **The only systematically collected nation-wide corpus of traditional dialect speech in mid-20th century England.**

We would further like to argue that another strength of the SED tapes is that

- **The informants are mainly NORMs (Non-mobile, older, rural, male)**

To consider this a strength may seem perverse to some, but a good case can be made for arguing that the informant choice that the SED project made is most fortunate, at least for historically orientated dialectology.

Thinking about the disadvantages the recordings have, the most serious drawback of the SED tape-recordings probably is that

- **The surviving individual recordings are relatively short, 8 to 10 minutes on average.**

Another possible disadvantage is that

- **The informants are mainly NORMs (Non-mobile, older, rural, male)**

Obviously, for sociolinguistically orientated research, the type of informant that the

SED aimed at is far from satisfactory.

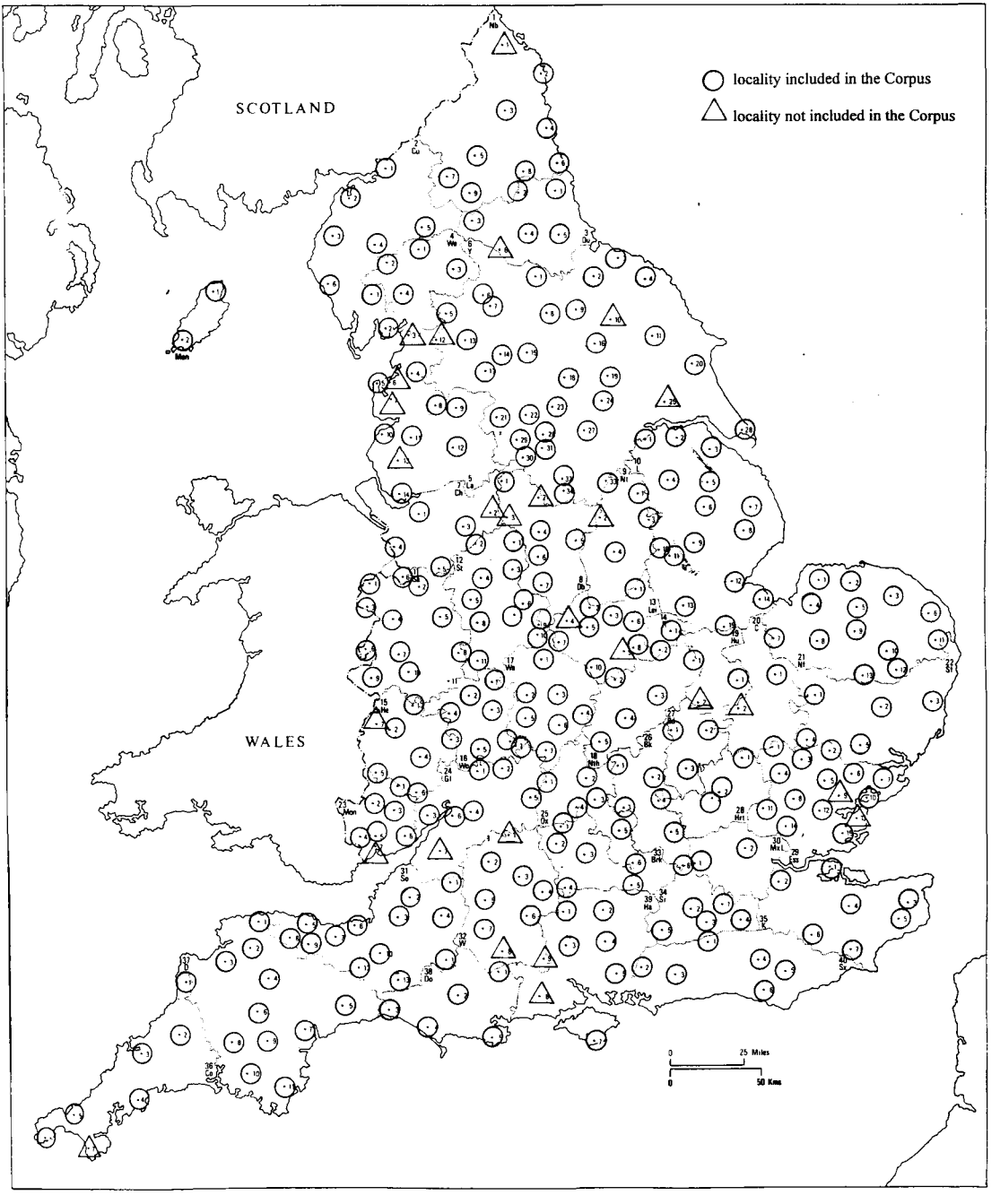
To get back to our present project, the aim of our Leverhulme Trust funded project was to transcribe and edit these tape-recordings for publication in electronic form. The project began in February 1997, when Mark Jones started his 12-month job of transcribing the recordings. The preliminary versions of the transcriptions were completed in December 1997, and the year 1998 was spent on checking the preliminary transcripts and making the necessary corrections. The transcriptions and the tape-recordings will now be published in CD-ROM format as *The SED-CDROM: The Spoken Corpus, recorded in England 1948-1973* (Klemola et al., *forthcoming*).

*Map 1* gives an indication of the scope of the Corpus. The 286 circles on this map represent localities where we have a usable recording from; the 27 triangles on the map represent gaps in the recordings. In some localities, no recordings were made, and in some cases the technical quality of the recordings is not good enough for inclusion in the Corpus. The total number of SED localities was 313, the surviving 286 recordings thus represent 91% of the total number of SED localities. And, as *Map 1* indicates, the gaps in the recordings are relatively few and far between.

## 3. *Transcription conventions*

When transcribing any language, and non-standard language in particular, a decision has to be made as to how to present the data. A phonetic transcription made using the IPA would allow one to represent the material in a form which did justice to its divergence from the standard and presented it in terms of a universally accepted and independently interpretable system. For the purposes of large scale comparison of lexical items and morphological/syntactic features phonetic transcription has very serious drawbacks in that even broad transcriptions make comparison between the same variable in divergent varieties practically impossible. In addition, accurate phonetic transcription takes a long time, and as it was intended that the recordings themselves would accompany any textual representation on CD-ROM – with the orthographic transcriptions aligned with the sound wave – phonetic transcriptions could be made when and where they were required by anyone using the package (cf. Sinclair 1995: 102).

It was decided therefore that the best way to proceed with transcribing the dialect recordings of the SED was to transcribe them in terms of standard English orthographical practices. There are at present no universally agreed conventions for transcribing non-standard speech in this way. A system of transcription conventions




*Map 1.* Localities included and not included in the *Leeds Corpus of English Dialects*

(CHAT conventions; **C**odes for the **H**uman **A**nalysis of **T**ranscripts) developed for transcribing child speech in the CHILDES project (MacWhinney 1995) was chosen as the starting point for our transcription conventions on the basis of its widespread use and relative unobtrusiveness. The CHAT conventions allow the transcriber to mark pauses and other discourse phenomena in a relatively easy way, which represents the sample of speech but does not cause undue difficulty to the reader. Standard punctuation is used to broadly reflect intonational units in the recording rather than morphosyntactic ones.

We have modified the CHAT conventions for use in the *Leeds Corpus of English Dialects* – project, but even with a set of transcriptional practices as a guideline, the task is not an easy one and ongoing revisions were made. The aim of the transcription has been to present the recorded material in as unadulterated a form as possible. Over the course of the transcription, many of the original conventions have been altered, especially those which it was felt imposed an undue amount of interpretation of the data. For the ease of analysis, several innovations have been made beyond the basic CHAT system which allow dialect material to be found as easily as possible. This has not been limited to dialect lexical items, but also involves standard English words used in a particular dialectal sense (e.g. *come* as a past tense form), where these would otherwise remain undistinguished and cause users of the system many hours of unnecessary labour.

When contracted forms appear, such as *he'd* for *he would*, the contracted section is preceded by a space before the apostrophe. This allows search routines to identify all the contracted forms en masse, without separate searches having to be conducted for *he'd*, *she'd* etc. The contraction is not glossed. Genitive case markings have no space preceding the apostrophe. The following are the transcription conventions used in the *Leeds Corpus*:

+... Indicates one of three possibilities

a) The speaker has trailed off and left a sentence unfinished before another speaker speaks.
b) as made a false start and leaves the sentence unfinished, continuing with another sentence
c) speech interrupted but continues after the interruption.

+/. The speaker has been forcibly interrupted.

| | |
|---|---|
| +" | Indicates the start of a quote (direct speech). |
| "+ | Indicates the end of a quote (direct speech). |
| [/] word [\] | The word/phrase bracketed has been unintentionally repeated. |
| [: word] | Replace by 'word'. Marks (a) the use of a standard English form in a non-standard way, e.g. *he come* [: came] *yesterday*, etc. or (b) the use of a non-standard form with the standard equivalent in brackets e.g. *they goed* [: went] *there.* |
| xxx | Speech unintelligible. |
| [*word*] | a) Marks a dialect term, e.g. *I* [*kenn*] *him.*<br>b) Marks a technical term, one which is used across large parts of the country with reference to a particular field of activity, e.g. the thatching term [*yelm*], a layer of cut straw, which occurs frequently across the country when thatching is discussed. Particular instances may not be synonymous. |
| [+word+] | Represents a dialect pronunciation of a particular word, when this is contrasted with a pronunciation reflecting the standard more closely, e.g. *No, I don't mean water, I mean* [+water+]. |
| spl- | Phonetically identifiable section of a word uttered. Occurs mostly when speaker stutters. Not followed in these instances by repetition marker [/]. |
| [!=a] | Indicates non-linguistic sounds affecting the conversation. Includes [!= clears throat], [!= coughs], [!= laughs], [!= sneezes], [!= sniffs], [!= spits], [!= yawns]. |
| ‖ | Indicates a discontinuity in the tape-recording (The 78 rpm disks only contain extracts from the original tape-recordings, and a surviving recording may consist of several extracts from the original tape-recording). |

## 4. *Transcription examples*

We have selected two short extracts from geographical extremes of the SED recordings to illustrate how the conventions work in practice. The first extract is from a recording from Heddon-on-the-Wall in Northumberland (SED 1 Nb 8). The recording was made by Stanley Ellis on March 21, 1953:

(1) <TM The blacksmith hoops that as +...
after you get all your fellies and things on.
You allow about # three quarters of an inch gape +...
# hole # uh # [*atween*] # all the points of the [/] the [\] fellies here,
all the cuts,
you know,
[*atween*] the spokes # where the # felly joints are,
you leave them open.
# xxx them fellies is all pinned and nailed you know,
and to keep them together.
# Well,
# that 's what they call # the allowance for the +...
when the ho- hoop 's put on,
hot,
it contracts you know. TM>
<SE Hmm. SE>
<TM They keep dousing it with cold water,
and draws all of your joints,
all your # [/] your [\] felly in. TM>

The second extract comes from a tape-recording from South Zeal in Devon (SED 37 D 6). The recording was made by Stanley Ellis on April 1, 1963:

(2) <TW You see,
# [*they 'm*] all in a # clique up there.
Choose ever anybody starts to go up # on Dartmoor,
you 've had it.
# Off they go.
There 's one fellow now,
# at the present moment,

I can # tell you a man,
# (a) young fellow,
# he been trying to keep up sheep up there,
and [*they 'm*] drive 'em # all over the place.
# A chap # doesn't know what to do with them.
# And he 's trying to # put sheep on the moor,
but # they won't let [*en*] go there,
and they ain't going to neither. TW>
<SE And is he really entitled to do it? SE>
<TW He 's entitled so much as I am and uh everybody as in the village. TW>

## 5. *Some examples of data*

In order to give some indication of the range of interesting linguistic features that can be found in the tape-recordings, we would finally like to discuss some examples of data culled from the corpus.

Examples (3) to (7) give an idea of the occurrence of a typically Northern feature of verb syntax in the corpus. The construction type exemplified under (3) to (7) was coined the 'Northern subject rule' by Ihalainen (1994). Other names for the phenomenon include 'the personal pronoun rule' (McIntosh 1983) and 'Northern Present-tense Rule' (Montgomery 1994).

(3) <BE uh a lot of fellows is +...
objects to them things because they 're
# the beasts when they lig down sometimes catches their knees on them and you get beasts with big knees.
But I don't that big knees are altogether # bl- +...
should all be +...
altogether blamed on that because they # [/] they [\] run hand in hand with abortions # chiefly. BE>
(We1 Great Strickland, Westmorland)

(4) <RM and uh # a lot of people thinks a # [/] a [\] pit man 's just a bloody duck egg.
But he 's not.
Oh no.

He knows his uh [/] he knows his [\] job. RM>
(Nb6 Earsdon, Northumberland)

(5) <FS But he had to pay for these chickens.
But [/] but [\] he 'll sell them him at uh +... FS>
<HS A shilling a week. HS>
<FS At four week old.
and they think they 've +...
because they don't eat much the first four weeks. FS>
(Y21 Heptonstall, Yorkshire)

(6) <DD Oh,
some takes four or five crops.
some three or four.
Some of them breeding spots.
Where they just has breeding ewes and breeds their own,
they 'll nobbut take three crops off 'em.
(Y13 Horton-in-Ribblesdale, Yorkshire)

(7) +" Keep your money in your pocket, "+
I said,
+" we don't breed 'em for you,
we breed 'em for ourselves. "+
(Y19, York, Yorkshire)

The Northern subject rule states essentially that, in the present tense, the verb takes the *-s* ending in all persons, singular and plural, unless it is adjacent to a personal pronoun subject (except for the third person singular, where the *–s* ending is used regardless of the type and proximity of the subject NP). Thus in (3) you find the inflected forms ***objects*** and ***catches*** on the second and third lines of the example, when the verb is not adjacent to a personal pronoun subject. On the other hand, when the verb **IS** adjacent to a personal pronoun subject, you find forms without the ***-s*** marker, as in ***they lig*** (line 3), ***you get*** (line 4), or ***they run*** (line 7).

Similarly, in (4), where the subject is a full NP, it is followed by a verb form in ***-s***, ***a lot of people thinks***, whereas in (5), with an adjacent personal pronoun subject, no ***-s*** marker is used in ***they think***. Examples (6) and (7), with the verb ***breed***, give further indication of the Northern subject rule operating in the Leeds

Corpus.

The phenomenon of the Northern subject rule is extremely interesting in terms of its geographical distribution, its history, and its typological rarity. The examples given here are just meant to give an indication of the fact that the SED tape-recordings provide a good source of data for the study of dialect syntax; for a more detailed discussion of the geographical distribution and history of the Northern subject rule construction, see Klemola *(in press)*.

The final set of examples, examples (8) to (10), are given here as an indication of the potential the SED tapes may have also for the study of the geographical distribution of various discourse markers in English dialects. Before we started transcribing the tapes, we were unaware of the geographical distribution of ***man*** as a discourse marker, associating it mainly with American English, especially African American Vernacular English. Therefore it was somewhat surprising to find the form being used in very traditional Northumberland speech, pronounced as /man/ rather than /mæːn/, however.

(8) as I say,
the [/] the [\] working man was just a bloody slave,
***man***.
# Aye.
(Nb6 Earsdon, Northumberland)

(9) The women never had no time to gan in,
# oh no,
the women's work was never done,
***man***,
poor buggers.
(Nb6 Earsdon, Northumberland)

(10) Badger.
# Oh,
there 's no dog can kill a badger,
***man***.
(Nb3 Thropton, Northumberland)

A preliminary analysis of the geographical distribution of the use of *man* as a discourse marker in the recordings indicates that this phenomenon is very definitely a

feature of the upper North. The 75 instances of *man* in this function found in the Corpus were confined to the Northern counties of Northumberland, Cumberland, Durham, and Westmorland.

## 6. *Final remarks*

We believe that the completion of the *Leeds Corpus of English Dialect* project will open up an exciting field of further research. The availability of the *The SED-CDROM: The Spoken Corpus, recorded in England 1948-1961*(Klemola et al., *forthcoming*) will make it possible for the first time to analyse an extensive corpus of spontaneous, comparable tape-recorded (English English) dialect speech and thus to study the regional variation in English dialect morphosyntax on a nation-wide scale. Access to the corpus of SED-tape-recordings will make it possible to study such areas as the syntax of negation, relative clauses, pronoun systems, etc., and thus to make important new discoveries about the grammatical structure of traditional dialects of English English. Combined with other historical corpora available, such as the *Helsinki Corpus of English Texts* and the *Corpus of Early English Correspondence*, it will also be possible to use the results of the synchronic analysis as the basis of a diachronic study of the morphosyntactic properties of non-standard vernacular varieties of early Modern English.

*The SED-CDROM: The Spoken Corpus, recorded in England 1948-1961* will be an extremely valuable source of data not only for dialectological research, but also for sociolinguists, historical linguists, and phoneticians interested in the study of language variation and change. Furthermore, it is expected that the *Corpus* will also be of interest to oral historians, to schools as source material for English language A-level courses, and to the general public interested in the richness of the English language.

**ACKNOWLEDGEMENT**

We gratefully acknowledge the support of the Leverhulme Trust (grant no: F/122/AT) which made possible the work reported here.

## REFERENCES

Ihalainen, Ossi. 1994. The dialects of England since 1776. In Robert Burchfield (ed), *The Cambridge History of the English Language. Volume 5: English in Britain and Overseas. Origins and Development,* pp. 197-274. Cambridge: Cambridge University Press.

Klemola, Juhani. in press. The origins of the Northern Subject Rule: a case of early contact? In Hildegard Tristram (ed), *Celtic Englishes II.* Heidelberg: Carl Winter Verlag.

Klemola, Juhani et al. forthcoming. *The Survey of English Dialects on CD-ROM: The Spoken Corpus Recorded in England 1948-1961.* London: Routledge.

McIntosh, Angus. 1983. Present indicative plural forms in the later Middle English of the North Midlands. In: Angus McIntosh, Michael L. Samuels, and Margaret Laing (eds.), *Middle English Dialectology: Essays on Some Principles and Problems*, pp. 116-22. Aberdeen: Aberdeen University Press.

MacWhinney, Brian. 1995. *The CHILDES Project: Tools for Analyzing Talk.* Second Edition. Mahwah, NJ: Lawrence Erlbaum.

Montgomery, Michael B. 1994. The evolution of verb concord in Scots. In Fenton, A. and D.A. MacDonald (eds), *Studies in Scots and Gaelic: Proceedings of the Third International Conference on the Languages of Scotland*, pp. 81-95. Edinburgh: Edinburgh University Press.

Orton, Harold. 1962. *Survey of English Dialects* (A): *Introduction.* Leeds: E.J. Arnold; repr. Routledge, London, 1998.

Orton, Harold, Michael V. Barry, Wilfrid J. Halliday, Philip M. Tilling, and Martyn F. Wakelin (eds). 1962-1971. *Survey of English Dialects* (B): *The Basic Material* (4 vols in 12 parts). Leeds: E.J. Arnold; repr. Routledge, London, 1998.

Sinclair, John. 1995. From theory to practice. In Geoffrey Leech, Greg Myers, and Jenny Thomas (eds), *Spoken English on Computer: Transcription, Mark-up and Application*, pp. 99-109. Harlow: Longman.

Viereck, Wolfgang (in collaboration with Heinrich Ramisch). 1991. *The Computer Developed Linguistic Atlas of England 1.* Tübingen: Max Niemeyer Verlag.

**Addresses**
Juhani Klemola
Department of English
University of Helsinki
PO Box 4 (Yliopistonkatu 3)
FIN-00014 University of Helsinki
Finland
<juhani.klemola@helsinki.fi>

Mark J. Jones
51 Lydgate Hall Crescent
Crosspool
Sheffield S10 5NE
England
<markjjones@hotmail.com>