Joseph Wright's *English Dialect Dictionary* (1898-1905) Computerised: architecture and retrieval routine

Manfred Markus
University of Innsbruck

Dagstuhl, *Digital Historical Corpora*, 3-10 Dec, 2006

0. Brief self-introduction of the author

MM: (a) compiler and (b) user of various historical corpora

(a) Compiling ('ICAMET'):
- ☐ Innsbruck Full-Text Database of Middle English Prose (1100-1500)
- ☐ Innsbruck Corpus of ME and EmodE Letters (1388-1700)
- ☐ Wright's *English Dialect Dictionary* (1898-1905): co-directors: Reinhard Heuberger and Alexander Onysko

(b) Preparation and use of corpora
- ☐ Normalisation (of the Letter Corpus): demonstration
- ☐ Tagging (of the Letter Corpus) for specific purposes (e.g. pragmatics)
- ☐ Combined use, e.g. Innsbruck corpora plus *OED*
- ☐ Use of search machines such as *WordCruncher* and *WordSmith* for various retrieval procedures

Comment on Chapter 0: By way of an introduction on this conference of relatively heterogeneous participants – philologists and computer scientists being the two main groups – I would like to emphasise the difficulty of finding a common language in the interdisciplinary dialogue. While English is, of course, the *lingua franca* amongst the participants, there is still a risk of non- and misunderstanding. As I see it, computer philologists tend to produce theoretical concepts of how problems can be solved with the help of software, but there is a limited competence in identifiying the problems. Philologists, on the other hand, with a good knowledge of the textual, historical and cultural embeddedness of their objects of research, often know the problems at issue, but may not be aware of the new computer-assisted tools and strategies waiting for them to solve their problems.

Given this basic difference of competence, we should all try to „meet halfway", i.e. to cooperate. The first step towards cooperation is to clarify one's own position. I, therefore welcome the invitation of the organisers of the Dagstuhl conference to present a short self-introduction.

Most corpus-linguists either compile a corpus or use corpora compiled by others. My own academic past has caused me to do both, compiling and analysing. Details of my corpus-compiling activity can be seen from the survey given above. Wright's *English Dialect Dictionary* is the last in a series of projects of corpus-compilation.

As regards the application of corpora, it seems to me that both the historicity and complexity of corpus texts are often underestimated. The larger and the older the texts of a corpus are, the more are they likely to be based on variable and erstwhile norms which disallow computer accessibility of the texts in their original shape. Editorial practices (if editions are used for compilation) and the now unimaginable irregularities of spelling in at least medieval and Early Modern English texts are the two main arguments for the necessity of normalisation. If a word comes along in shapes that are incalculable at hoc, the results of query routines are bound to be misleading. For more detailed arguments along these lines see my eralier publications Markus 1997 and 2000.

Normalisation is closely connected with tagging, at least automatic or semi-automatic tagging. The reason is obvious: the computer can only identify strings if their spelling is consistent. One of the reasons why historical English studies have so far very much been limited to studies of words is the non-availability of appropriately tagged corpora. Syntactic, semantic and pragmatic features cannot be found in untagged word strings. I have, therefore, tagged the *Innsbruck Letter Corpus* (1386 to 1698) in a tentative procedure to allow the questions that I was interested in (cf Markus 2001). There is no one and only kind of tagging. The tags, naturally, depend on the research aim. As I see it, the ideal tagger would allow for different levels of features, from spelling idiosyncrasies to pragmatics and text linguistics, to be faded in and out, according to the scholar's needs.

Apart from normalisation and tagging, I would also like to plea for the flexible use of dictionaries, such as the *OED* and the *MED*, in addition to, and complemetation of, special corpora, whether these are self-compiled or not. Also, index and concordance programs, such as WordSmith, are extremely helpful, since they allow contextual queries and, thus, the retrieval of phrases and connotations even in untagged texts.

0.1. Normalisation: example

$I   [T]o the moost noble & Worthiest Lordes, moost ryghtful
$N   To the most noble and worthiest Lords, most rightful
$I & wysest conseille to owre lige Lorde the Kyng, compleynen, if
$N and wisest council to our liege Lord the King, complain, if
$I it lyke to yow, the folk of the Mercerye of London, [as] a
$N it likes to you, the folk of the Mercery of London, as a
$I membre of the same citee, of many wronges subtiles & also
$N member of the same city, of many wrongs subtles and_abb also
$I open oppressions, ydo to hem by longe tyme here bifore
$N open oppressions, done to them by long time here before
$I passed.
$N passed.
$I   Of which oon was where the eleccion of Mairaltee is to be to
$N   Of which one was where the election of Mayoralty is to be to
$I the fre men of the Citee bi gode & paisible auys of the wysest
$N the free men of the City by good and_abb peaceable advise of the wisest
$I & trewest at o day in the yere frelich, there nought-withstondyng
$N and truest at one day in the year freely, there notwithstanding

Since normalisation has been a controversial issue and is on the agenda of this conference, I would like to use the opportunity of demonstrating the macro routine which has been developed by an undergraduate team member (Mag. Christoph Praxmarer) of my present project at Innsbruck. The example above shows an „Input line" ($I) and a normalised line in an interlinear arrangement ($N). If the macro is activated, the first line of WORD icons provides two extra buttons which allow a separate fading in and out of the two lines, according to the user's preference. This enables us to use a normalised text version of the *Innsbruck Letter Corpus* and yet to keep in constant touch with the original. This close proximity to the original text is not only in line with the philological ethos of transparency, but also a simple necesssity when questions of spelling, phonology and morphology are at issue. It, thus, comes to light that in the last paragraph *one* is spelt in two different ways, <oon> and <o>. For syntactic questions, on the other hand, the input line can be faded out. Needless to say that, in order to allow such syntactic queries, normalisation as we understand it is limited to spelling and lexis (words unknown in present-day English are translated and marked by an asterisk); normalisation must not affect syntax.

1. Introduction

Comment on Chapter 1:
Introduction to Wright's *EDD*

Picture 1: Wright on a painting (foto taken by MM)
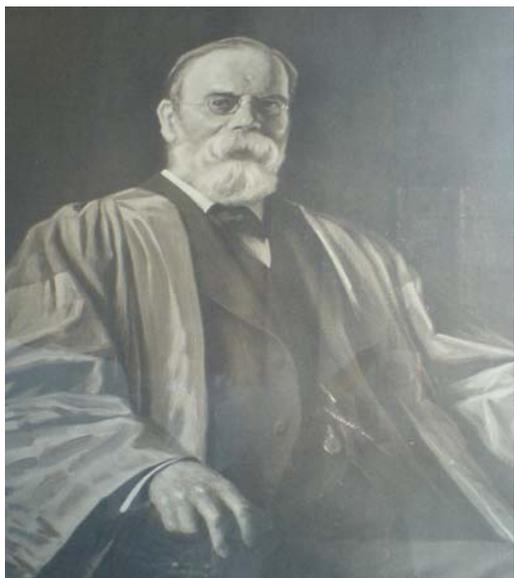
Picture 2: The six volumes of the English Dialect Dictionary (foto taken by MM)

Details on the history of research that Wright's *English Dialect Dictionary* (1898 to 1905) has to be seen in, as well as on Wright's personal circumstances before and during his compilation of the dictionary, can be found in Markus 2007 (forthcoming) and in Heuberger/Markus 2007 (forthcoming). Suffice it to say that Joseph Wright was an admirable researcher and research manager and that his comprehensive dialect dictionary, with a total of more than 5000 pages in six volumes, has wrongly been neglected, even ignored. According to Wright's own Preface, the dictionary is the 'complete vocabulary of all dialect words still in

use, or known to have been in use during the last two hundred years' (i.e. 1700-c1900) and 'the largest and most comprehensive dialect dictionary ever published in any country'.

In addition to these obvious assets of the *Dictionary* from Wright's own point of view, there are three further advantages in hindsight:

1. The *Dictionary's* historical range of 200 years, namely back to 1700, covers the phase of what is now usually referred to as Late Modern English – a generally neglected period of the history of the English language.

2. Wright used an admirably precise and scholarly method of linguistic description, from phonetics to the citations.

3. In line with the policy of the *English Dialect Society*, the *Dictionary* shows a concern with details of cultural history, namely "superstitions and practices in relation to religion, death, witchcraft, apprenticeship, courtship and the like" (Holder 2004: 258); this is well in line with the now increasing interest of university staff and students in cultural studies.

2. Why an electronic version of the *EDD*?

(a) Historical dialectology
(b) (Historical) Spoken English
(c) Historical linguistics, in particular, lexicology

Comment on Chapter 2:

To produce and distribute an electronic version of Wright's *EDD* is not just filling the gap in English historical linguistics, but it will – in contrast to the *OED* – particularly respond to the needs of three special branches in historical English linguistics.

(a) The first one is historical dialectology. While this has been done to quite an extent in view of Old and Middle English, a historical dialectology of Late Modern English hardly exists. The main reason for this could well be the fact that the twentieth century turned scholars' main interest to the system of languages rather than to their varieties – just a good decade after Wright's English Dialect Dictionary de Saussure's *Cours de Linguistique Générale* was published. Another factor could well be the continuation of the eighteenth century's preference of norms, visible in the focussing on the „English Southern standard" and on „King's English" in the earlier nineteenth and also in the twentieth century. Finally, the increase of urbanisation, in line with industrialisation, from about 1840 on has motivated many recent scholars, e.g. Romaine (1998: 14), to favour sociolectal parameters at the cost of regionalectal ones.

Given that the present situation of British dialects is one of survival, the **revival** (bringing to life) of relatively recent dialectal features should be welcome. It is true that dialects in Britain have not died out since 1900, which is what some scholars of Wright's kind and time portended prematurely. However, it can be expected that Wright's *Dictionary*, written soon after the introduction of compulsory education (1870) and the invention of new traffic systems, was fairly close to the original dialect situation, still unaffected by dialect levelling as a result of new modes of communication.

(b) Secondly, Wright's *Dictionary* in electronic form will help us to go a good step forward in the study of historical spoken English, which again hardly exists (I can think of only one exception: Bøgholm, *English Speech from an Historical Point of View,* 1939). Most of the *EDD's* entries have a phonetic transcription, which, while not identical with the IPA transcription now well-known (though fairly similar to it) reliably mirrors the phonemic pronunciation and word accent in more than 100,000 dialectal words (the figure is the sum of entry words plus derivations and compounds within the entries). At the moment my

team members of the Wright project („SPEED") and myself have still the problem of the special phonetic symbols and our database ORACLE to be incompatible with one another. But we will try to make the transcriptions of the lemmata accessible to researchers as soon as possible, if needs be, in a coded version of the special characters.

Apart from phonetics, phonology and prosody, Wright's dialect material will allow us to topicalise issues of collocation, colloquialisms, patterns of repetition and deviation.

(c) The third point mentioned above refers to the contribution of Wright's *Dictionary* to English historical linguistics, in particular, lexicology. English lexis has recently and with growing interest been studied in view of the international distribution and importance of words (cf, e.g., Gramley 2001). But the role and survival of words in different British dialects has widely escaped attention. This is not just a compilation of long lists of words, but also a question of discovering dialectal patterns and preferences in the use of both individual words and phrases, as well as idioms and metaphors (cf Markus 2008, in progress, on the tendency of British dialects towards diminutive suffixes, such as –*ie* or –*et*).

3. Wright's *EDD*: structure

Picture 3: a random page from the *EDD*
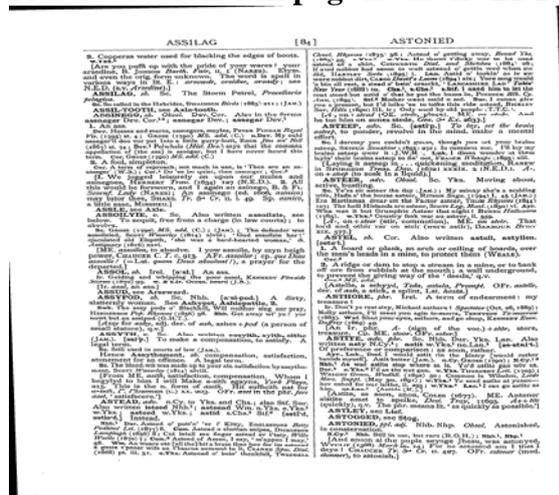
Picture 4: An extract of the same page in close-up:

The two entries on ASTEEP and ASTEER illustrate the basically eight slots, or parameters used in the entries (though not always all of them at a time).

3.1. Survey of the structure of the *EDD's* entries
1. *Lemma*, or *headword*
2. *part of speech*, such as *v.* (for *verb*)
3. *usage label*, such as *obs.* (for *obsolete*);
4. dialect *regions* and *counties*
5. *phonetic transcription* (not the IPA, but simliar to it)
6. *definition*
7. *citations* with their sources
8. *comments* or cross references

3.2.   Parts of speech: syntactic and pragmatic functions

After the beginning of the project, it soon turned out that the eight paramters were less precisely defined than seemed to be the case at first sight. The second parameter (parts of speech), for example, mostly refers to word classes, but there are also a great many other, in particular, syntactic and pragmatic implications (cf the following two tables):

Table 1: Parts of speech: syntactic function

form	explanation	domain
also in comb.	also in combination	syntax
also in phr.	also in phrases	syntax
also used advb.	also used adverbially	syntax
attrib.	attributive, -ly	syntax
comb.	combination	syntax
improperly used as inf.	improperly used as infinitive	syntax
obj.	object	syntax
pass.	passive, -ly	syntax
pred.	predicative, -ly	syntax

Table 2: Parts of speech: pragmatic function

also used as a familiar term of address	also used as a familiar term of address	pragmatics
also used as a fencing term	also used as a fencing term	pragmatics
also used as a term of contempt	also used as a term of contempt	pragmatics
also used as a term of endearment to (X)	X = (an infant, children)	pragmatics
also used as an epithet of contempt	also used as an epithet of contempt	pragmatics
and in gen. colloq. use	and in general colloquial use	pragmatics
euphonic	euphonic	pragmatics

in colloq. use	in colloquial use	pragmatics

3.3. Dialect: "translation" of fuzzy data

Dialects are not only referred to on different levels of size (county, region and nation), but sometimes also in fuzzy terms. The fuzziness may be semantic („different counties") or syntactic (in the case of unclear reference: „in some parts of East Riding and West Yorkshire"). As the latter examle shows, dialectal areas are also referred to by way of directions (N, S, W, E, NW, NE, SW, SE). We have filtered out all dialectal phrases and explained or „translated" them into their components, trying to keep those of a fuzzy quality („in some parts of x) apart from the more precise ones. While the list of fuzzy phrases has become longer and longer, here are a few examples:

Table 3: "Translation" of fuzzy dialect data

form	explanation	region
(x) & (y) counties	(n., e., s., w., sw., se., ne., nw.) & (n., e., s., w., sw., se., ne., nw.) counties	x_part y_part
(x) also (y)	region x/region y	x_tot y_tot
different counties	different counties	gen_part
in gen. use throughout dial. exc. in (x)	in general use throughout dialects except in (region)	gen_part X_not

4. The query mask

Since the project team at Innsbruck does not need any time for the dialectologist's usual initial task, which consists in conceiving the questionnaires and in having them filled in by field workers (cf Francis 1983: 78-103), we have focussed our interest so far on scanning and parsing the dictionary as well as on the structure of the query mask.Though the picture below is still provisional, is is clear enough that the mask will – in addition to the trivial „go-for-a-string-routine" – allow access along two main options: search areas, such as headwords, heads (the main parts of the entries) or full text, and search filters (on the right hand side). The filters will work in the way of run-down menus: you click on counties and then get the complete list, within which counties, regions and states can be activated according to the Boolean operators.

Picture 5: A provisional query mask for the *EDD*

Joseph Wright's English Dialect Dictionary

Export   Print   Image   Comment   County Map   Region Map   About   Help

search areas: [   ] [go]
☑ headwords   ☐ heads   ☐ last search   [simple search]
☐ variants   ☐ compounds   ☐ derivations
☐ phrases   ☐ full text   ☐ definitions
☐ citations

search filters:
regions/counties   part of speech   phonetic   etymology   source
usage label   morphemic   time span   [reset search]

Index list

[close] ☐ OR ☑ AND

Regions

☐ America   ☐ Australia   ☐ Canada
☐ East Anglia   ☐ East Country   ☐ East Ireland
☐ East Scotland   ☐ England   ☐ Ireland
☐ Lakeland   ☐ Midlands   ☐ Munster
☐ New South Wales   ☐ New Zealand   ☐ Newfoundland
☐ North Country   ☐ North East Scotland ☐ North Ireland
☐ North Scotland   ☐ North Wales   ☐ Scotland
☐ South Country   ☐ South Ireland   ☐ South Scotland
☐ South Wales   ☐ South West Scotland ☐ Wales
☐ West Country   ☐ West Scotland   ☐ Ulster
☐ U.S.A.   ☐ West Ireland

Counties

☐ Anglesea   ☐ Angus   ☐ Antrim
☐ Argyll   ☐ Armagh   ☐ Ayrshire
☐ Buchan   ☐ Buckinghamshire   ☐ Bedfordshire
☐ Banffshire   ☐ Brecknockshire   ☐ Berkshire
☐ Buteshire   ☐ Berwickshire   ☐ Caithness
☐ Cavan   ☐ Cardiganshire   ☐ Cheshire
☐ Clare   ☐ Clackmannanshire   ☐ Clydesdale
☐ Cambridgeshire   ☐ Cornwall   ☐ Cork
☐ Carlow   ☐ Cromarty   ☐ Caernarfonshire
☐ Carmarthenshire   ☐ Cumberland   ☐ Derbyshire
☐ Devonshire   ☐ Dumbartonshire   ☐ Dumfriesshire
☐ Denbigh   ☐ Donegal   ☐ Dorset
☐ Dublin   ☐ Durham   ☐ Down

Since a picture is worth a thousand words, we are thinking of converting outputs of dialect queries into maps. We are considering an overlay technique: a geographical map, for example, could be laid over the map showing dialect distribution. This would have to work along the lines of the Google Earth approach. An Innsbruck dissertation in progress (Praxmarer) is concerned with the visualisation of dialect data. At present no further information can be given. To visualise, however, Wright's concept of dialectal areas, drawn from his references to dialect both in the *EDD* and also in his *English Dialect Grammar* (1905), we have tentatively produced a first map:

Picture 6: Regions and counties of E, Sc and Wales, according to Wright (self-made map)

The Counties of Great Britain

Insular North
North Scotland
North East Scotland
East Scotland
West Scotland
South West Scotland
South Scotland

North Wales
South Wales

North Country
Lakeland
Midlands
East Country
East Anglia
West Country
South West Country
South Country
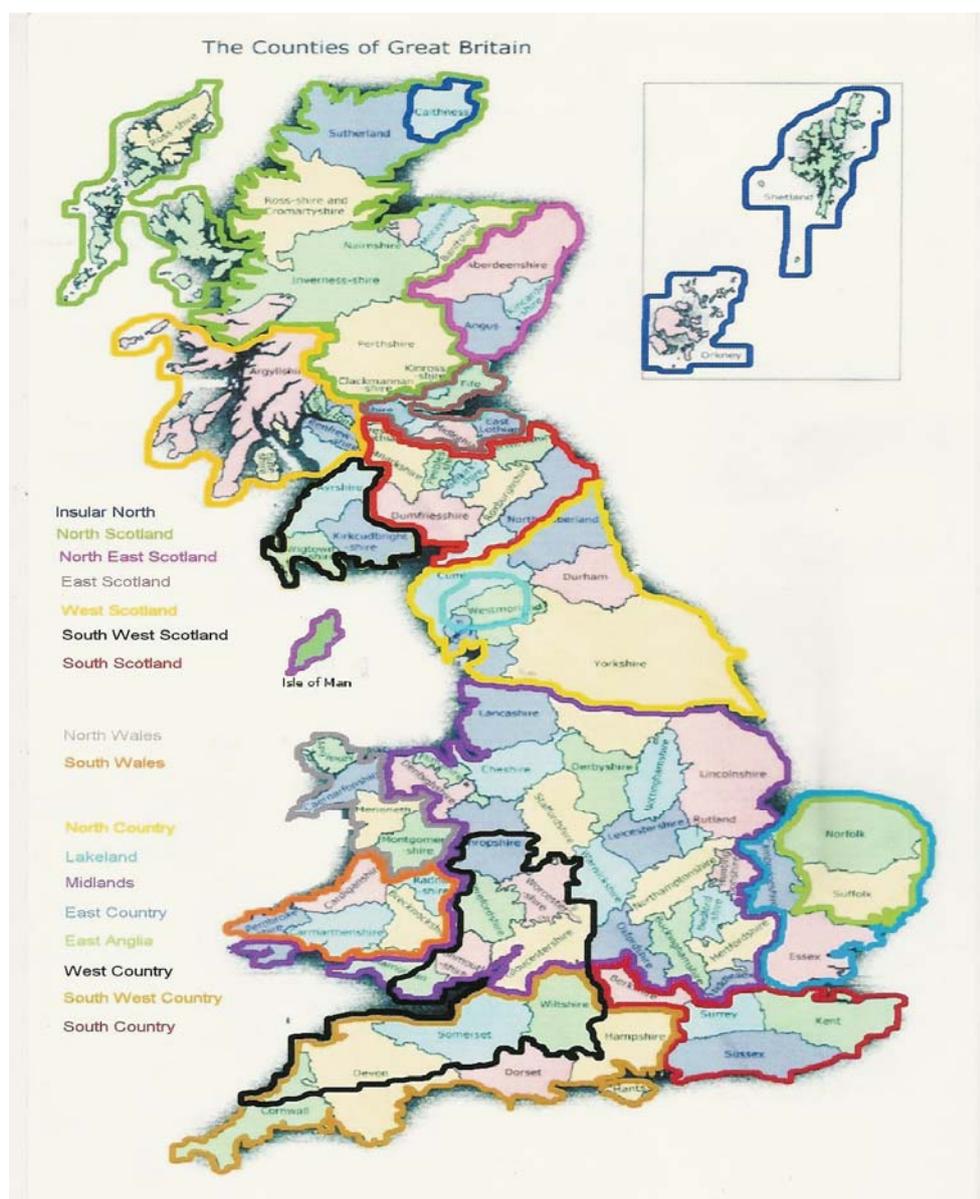
5. Repeated parsing

Our work during 2006 (and well into 2007) had to focus on parsing, the correct structural attribution of the different parts of an entry. Given that the entries are sometimes extremely long and complex, the „devil" has been in the detail. The long paragraphs with citations, for example, while attributable *in toto* to the parameter „citations", also contain „sub-parameters" – sources, meanings, dialect data. Meaning, for its part, is not only found in the slot mainly concerned with the definition of a lemma, but also elsewhere in an entry. On the other hand, the slot of „definition" often lists new lemmata, attributable to the head lemma due to its pattern of word formation (compound or derivation) or in the way of a phrase list. Given that italics are no valid feature of distinction for our scanner, it is not so easy for the parser to know phrases from definitions. The following quotation is an example:

Picture 7: Mixture of features in the field of *definition*

4. In phr. *Abune a'*, beyond reason ; *aboon-a-bit*, exces-
sively; *aboon the breath*, across the forehead; *abone-broe*, see
quot.; *aboon grees*, upstairs ; *to get aboon hands*, to become
supreme, get the 'upper hand'; *aboon with oneself*; *aboon
plum*, drunk; *ower* (over) *and aboon*, (1) entirely, alto-
gether, (2) into the bargain.

All in all, most of the eight parameters initially distinguished have revealed a certain multifunctionality, cf Picture 8:

Picture 8: Relations between entry fields and the menu mask
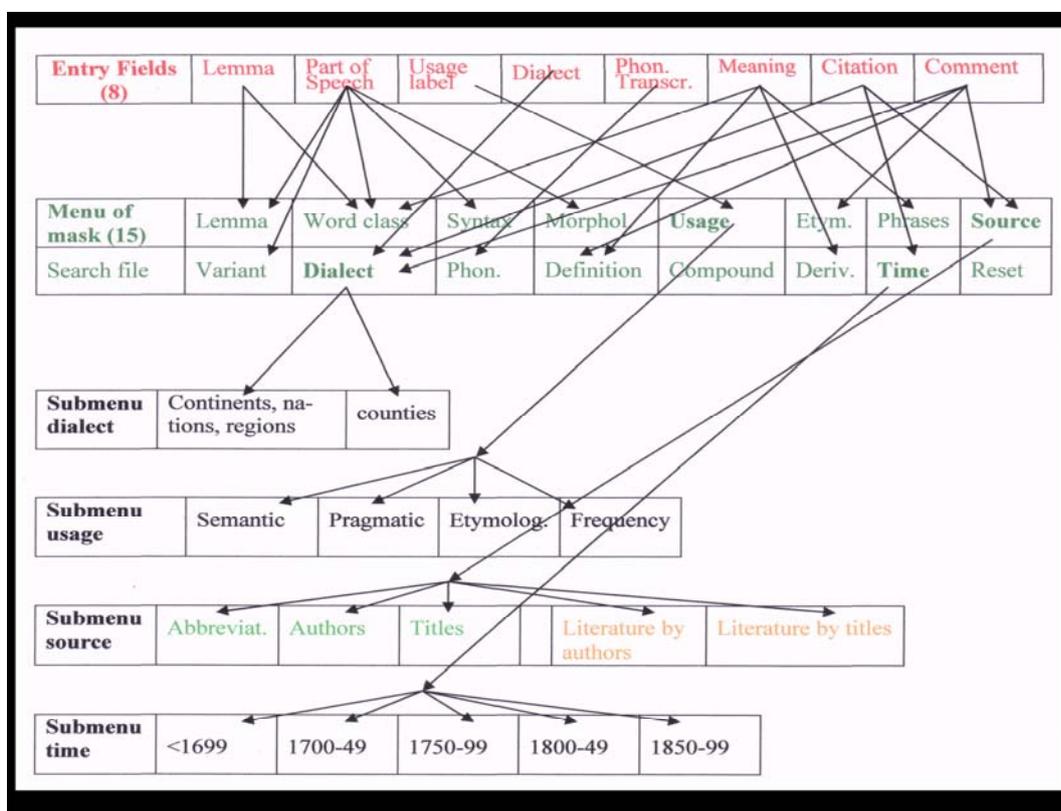
6. Summing up

Name of the Innsbruck project:	SPEED (Spoken English in Early Dialects)
Duration:	1 July 2006-middle of 2009
Main aim:	electronic version of Wright'sEDD; linguistic research on that basis
Size of the EDD:	5000 pages, based on 12.000 questionnaires; ca. 100.000 dialect words from Britain and overseas
Value of the EDD:	authoritative on Late Modern English dialect up to 1900
Perspective for research:	hides additional dialect material on word formation, phrases, semantics, pragmatics

Visualisation strategy: computer maps instead of traditional paper maps
Financial aspects: Innsbruck government-funded project from 2006 to 2009
Copyright: No copyright problems. Licence given by Oxford UP.
Future access: Planned: Online distribution (beta version 2007)

Any comments and cooperation welcome!

References

Bøgholm, N. 1939. *English Speech from an Historical Point of View.* Copenhagen etc: Nyt Nordisk Forlag Arnold Busck.

Francis, W. Nelson. 1983. *Dialectology. An Introduction.* New York: Longmans.

Gramley, Stephan. 2001. *The Vocabulary of World English.* London: Arnold: 2001.

Heuberger, Reinhard, and Manfred Markus. 2007 (in print). "The Architecture of Joseph Wright's English Dialect Dictionary: Preparing the Computerised Version." *International Journal of Lexicography.*

Holder, R.W. 2004. *The Dictionary Men. Their lives and times.* Bath: Bath University Press.

Markus, Manfred. 1997. "Normalization of Middle English Prose: Possibilities and Limits." *Corpus-based Studies in English. Papers from the Seventeenth International Conference on English Language Research on Computerized Corpora, Stockholm, May 15-19, 1996.* Ed. Magnus Ljung. Amsterdam, Atlanta, GA: Rodopi: 211-226.

Markus, Manfred. 2000. "Normalizing the Word-forms in Ayenbite of Inwyt." *Placing Middle English in Context. Topics in English Linguistics* 35. Eds. Irma Taavitsainen, Terttu Nevalainen, Päivi Pahta and Matti Rissanen. New York, etc.: Mouton de Gruyter: 181-198.

Markus, Manfred. 2001. "The Development of Prose in Early Modern English in View of the Gender Question: Using Grammatical Idiosyncrasies of 15th and 17th Century Letters." *Early Modern English Text Types. European Journal of English Studies* 5.2. Ed. Lilo Moessner. Lisse: Swets & Zeitlinger Publishers: 181-196.

Markus, Manfred. 2007 (forthc.) "Wright's English Dialect Dictionary Computerised: Towards a New Source of Information." *Online publication VARIENG Helsinki.*

Markus, Manfred. 2008 (in progress). "Diminutives in British dialects (based on Wright's *English Dialect Dictionary*)".

Romaine, Suzanne ed. 1998. *The Cambridge History of the English Language, IV: 1776-1997.* Cambridge: UP.

Wright, Joseph 1898-1905. *The English Dialect Dictionary.* 6 vols. Oxford: Henry Frowde.

Wright, Joseph 1905. *The English Dialect Grammar.* Oxford, 1905, repr. 1968.