VARIENG
research unit for variation, contacts and change in english

CORD

Corpus Resource Database

http://www.helsinki.fi/varieng/CoRD

General introduction

As the number of electronic corpora increases every year, it is becoming ever more challenging for linguists to be aware of, let alone familiar with the corpora available at any given time. The members of the Research Unit for Variation, Contacts, and Change in English (VARIENG) believe passionately that corpus linguistic research requires not only familiarity with the latest methodologies, but also with the details of each corpus. To that end, VARIENG has launched a free, open-access resource to facilitate the sharing of information on available linguistic corpora. In the first instance, we endeavor to collect information on all English language corpora.

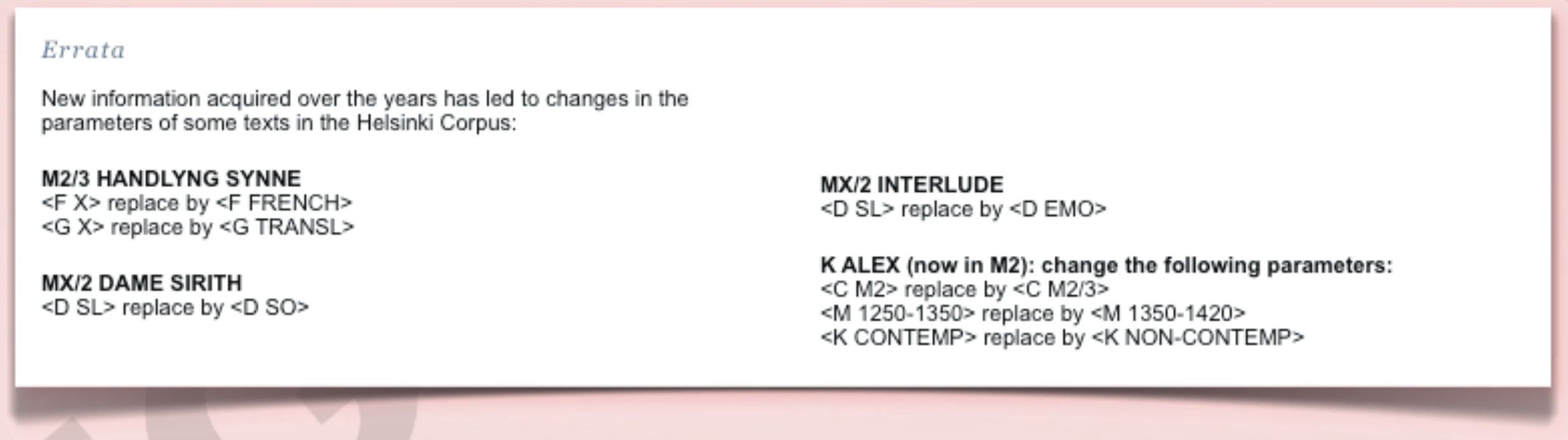
Compiling a corpus takes a long time. It is also hard work. The last thing one wants to do is compile a specialist corpus only to discover that a colleague did the same thing a few years earlier and would have been happy to share it with you. As CoRD grows, chances of accidentally overlapping compilations are diminished.

Authoritative information

All CoRD entries are written by the compilers of the respective corpus. No guess work, no interpretation. When you read a CoRD entry, you can trust the information to be accurate. Each CoRD entry comes with full details of when and by whom the information was submitted. A feedback form is made available for CoRD users to report any inaccuracies, and these can be communicated to the authors for clarification or explanation.

Up-to-date and accurate

CoRD entries can be updated to reflect the latest news about ongoing corpus projects. CoRD can also be used to make lists of errata available to corpus users. Stratifications originally envisioned may change due to new scholarship, and newly discovered information may change the classification of a particular text. Through the pages of CoRD, compilers can communicate the latest data to corpus users. Naturally, a full record will be kept of the update history, and CoRD pages always indicate when and by whom the page was last updated.



Errata

New information acquired over the years has led to changes in the parameters of some texts in the Helsinki Corpus:

M2/3 HANDLYNG SYNNE
<F X> replace by <F FRENCH>
<G X> replace by <G TRANSL>

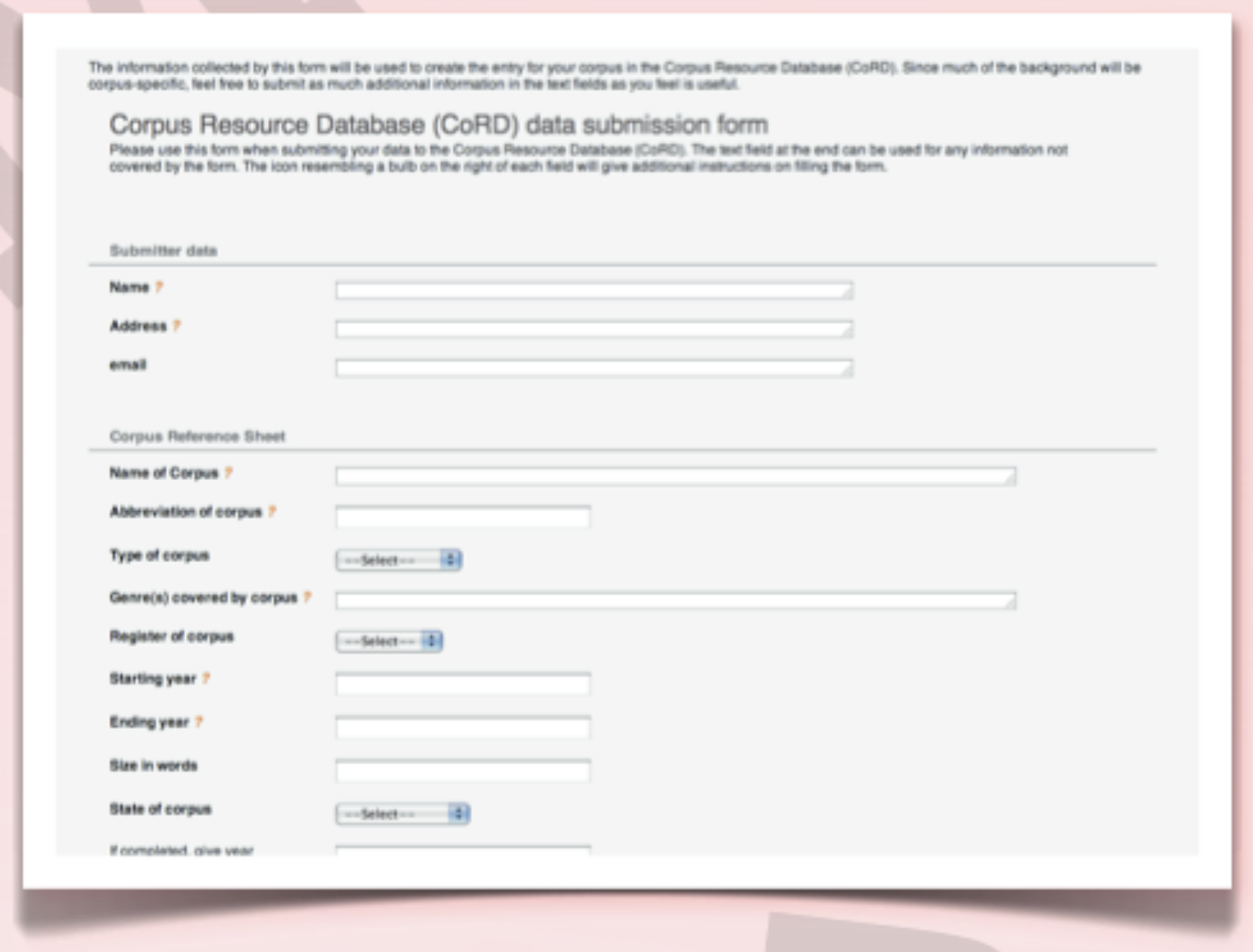
MX/2 DAME SIRITH
<D SL> replace by <D SO>

MX/2 INTERLUDE
<D SL> replace by <D EMO>

K ALEX (now in M2): change the following parameters:
<C M2> replace by <C M2/3>
<M 1250-1350> replace by <M 1350-1420>
<K CONTEMP> replace by <K NON-CONTEMP>

Submitting a corpus to CoRD

Corpus information can be submitted in one of two ways. The simplest option is to use the electronic form, available on the CoRD website. We take the information and manually convert it into HTML following the guidelines of the CoRD template.



The information collected by this form will be used to create the entry for your corpus in the Corpus Resource Database (CoRD). Since much of the background will be corpus-specific, feel free to submit as much additional information in the text fields as you feel is useful.

Corpus Resource Database (CoRD) data submission form

Please use this form when submitting your data to the Corpus Resource Database (CoRD). The text field at the end can be used for any information not covered by the form. The icon resembling a bulb on the right of each field will give additional instructions on filling the form.

Submitter data

Name ?

Address ?

email

Corpus Reference Sheet

Name of Corpus ?

Abbreviation of corpus ?

Type of corpus --Select--

Genre(s) covered by corpus ?

Register of corpus --Select--

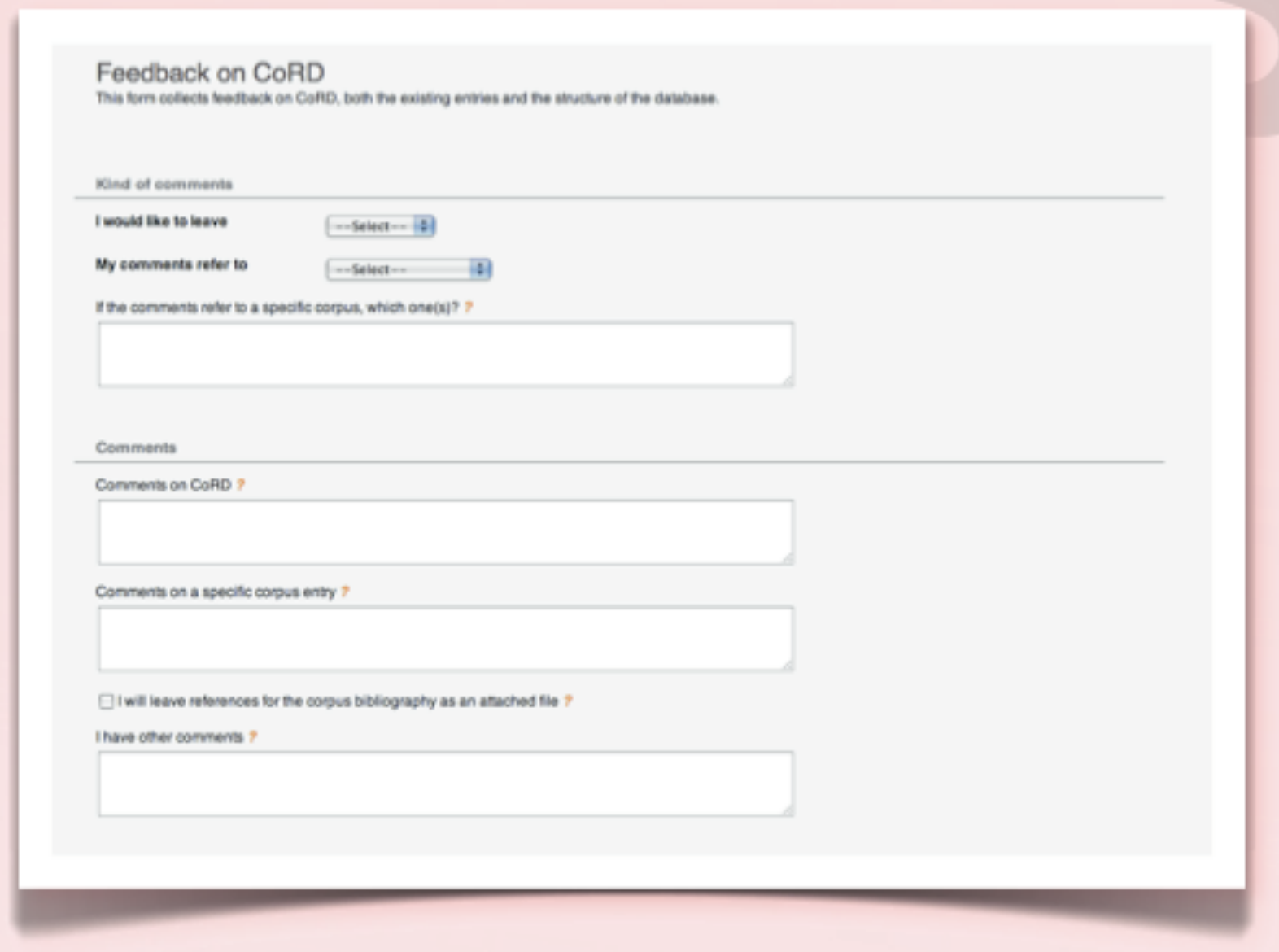
Starting year ?

Ending year ?

Size in words

State of corpus --Select--

If completed, give year



Feedback on CoRD

This form collects feedback on CoRD, both the existing entries and the structure of the database.

Kind of comments

I would like to leave --Select--

My comments refer to --Select--

If the comments refer to a specific corpus, which one(s)? ?

Comments

Comments on CoRD ?

Comments on a specific corpus entry ?

☐ I will leave references for the corpus bibliography as an attached file ?

I have other comments ?

Alternatively, if your corpus information is more complicated, or if you have specific requests about the layout and presentation, you can get in touch with us and we'll be happy to work with you in designing the CoRD entry. Naturally, you can add to, alter, or remove any information at any time.

What is CoRD for?

Reference information

A common problem in corpus linguistic scholarship is deciding on the correct reference line when citing a particular corpus. CoRD provides the recommended reference line for each corpus, written by the compilers themselves.

A quick source of information

Nearly all computers today are connected to the Internet. Through CoRD, users of corpora are one mouse click away from checking the basic data of a corpus such as its wordcount or that of its subcorpora, the annotation scheme, or perhaps the file format used in the distribution version.

Giving credit where credit is due

One of the shortcomings of conventional citation formats is that often only the leaders of research projects are mentioned by name. All the others, many of whom may have contributed extensive amounts of time and labour, are relegated to a footnote. CoRD allows corpus compilers to acknowledge all team members. Because CoRD entries are free-form and open-ended, the compilers can use as much space as necessary to explain the roles and responsibilities of each team member.

Today, funding bodies are also increasingly interested in public acknowledgement of their contributions to research. Long-running projects, as well as those with many project members, may have benefited from funding by a number of different organizations. Through CoRD, corpus compilation projects can credit sources of funding in as much detail as necessary.

Availability information

Some corpora are made available through open access academic repositories, while others are obtainable by subscription, membership in an organization, or directly from the compilers. CoRD tells you who to contact.

Teaching

Teachers of corpus linguistics can use CoRD in a number of different ways. In addition to referring students to CoRD for basic information on corpora, CoRD entries can serve as introductory material to the process of corpus compilation. Narrative accounts, such as those provided by the compilers of the Helsinki Corpus and the Helsinki Corpus of British English Dialects, can help current and future corpus compilers understand both the challenges and joys of compiling corpora.

Recording the living history of corpus linguistics

Although a relatively young discipline, corpus linguistics has in actual fact been around for quite some time. Over the years, corpus compilation methods have undergone considerable changes. CoRD is an online resource for recording and sharing not only hard facts, but also the stories and experiences of the compilers. Photographs, audio recordings and video clips can be used to liven up the presentation even further.

The Structure of a CoRD entry

CoRD entries are essentially open-ended in structure. Because the information comes directly from the compilers, it is as accurate as possible. While we suggest a basic two-fold structure for most entries, the details specific to each corpus will ultimately dictate how the information is arranged and presented.

In principle, CoRD entries comprise two main sections: *Basic Information* and *Background Information*. The *front page* of every CoRD entry gives basic information on the timespan and size of the corpus as well as on the genres and text types covered. Other information on the front page includes the names of the compilers and other team members, and the recommended reference line. Additional information on funding and details related to academic affiliations can also be included. Information on availability comes next, followed by up-to-date contact information. Naturally, hyperlinks to the compilers' own website or to other resources can be included as well.

Information on the structure of the corpus also goes under *Basic Information*. Charts and tables can be used to illustrate the organization of the corpus. This sub-section details the principles of temporal, sociolinguistic and genre-related stratification systems employed. VARIENG researchers believe that it is extremely important not only to know what kind of stratifications are used, but also what principles lie behind them. The stratification can be explained in detail, up to and including bibliographical information of text sources.

Any annotation schemes used can be explained and detailed. This will allow corpus users to quickly consult the annotations used in one or more corpora, thus saving plenty of valuable time and effort otherwise used in chasing down such information. Importantly, the compilers have as much space as they like to truly explain the reasons behind the choices they made. Facsimile images of early material can be included to illustrate issues arising in the editing of manuscripts and early printed material, while audio files can be used when describing the transcription of spoken data.

The *Background* section of a CoRD entry affords the compiler(s) the opportunity to narrate the details of the compilation process. This can include information both on the progression of the compilation itself, and on the human side of working on corpora. Cooperation with specialist consultants, such as historians or literary scholars, can be discussed. Compilers are encouraged to submit any and all multimedia content relevant to the process of corpus compilation. Photographs of equipment used and scanned images of notebook and journal entries can be included as historiographical evidence of the compiling process.

Finally, CoRD is envisioned to provide an up-to-date *bibliography* of research conducted using a particular corpus. This would not only allow members of the corpus linguistic community to know of and potentially contact colleagues interested in similar topics, but also provide corpus compilers with a single point of reference when evaluating the fruits of their labour. Like the corpus information, all bibliographical information will be submitted by the authors of the articles.

CoRD today

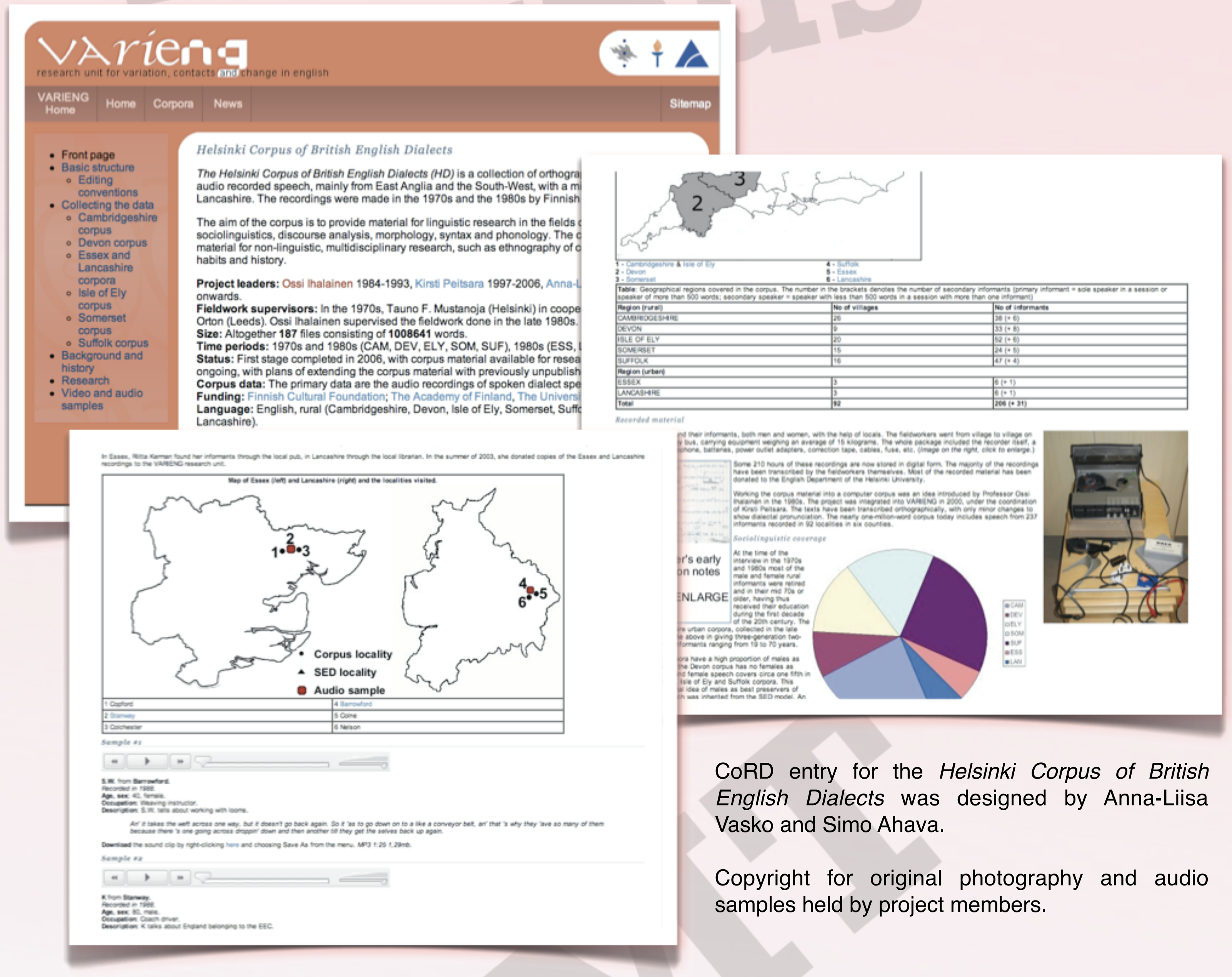
What's on CoRD right now?

CoRD includes the detailed descriptions of five corpora. In addition to the *Helsinki Corpus*, CoRD entries covers three corpora compiled in Helsinki: the *Corpus of British English Dialects*, The *Corpus of Early English Correspondence* (CEEC), and the *Corpus of Early English Medical Writing* (CEEM), the latter two with detailed sections on their subcorpora. In addition, CoRD includes entries on the *SCOTS* and *NECTE* corpora.

Multimedia on CoRD

Although still in an early stage, CoRD already includes a number of innovative features. In addition to standard visual aids like charts and maps, several entries feature photographs of the compilers and of the equipment used during the 1970's, 80's, and 90's. The entry on the *British English Dialects* corpus features audio clips to illustrate the recordings made in the 1970's and 80's on which the transcriptions are based. Entries on all corpora compiled in Helsinki come with a video presentation by the compilers: Matti Rissanen on *Helsinki Corpus*, Terttu Nevalainen on *CEEC*, Irma Taavitsainen on *CEEM*, and Kirsti Peitsara on the *British English Dialects corpus*.

Example: pages of The *Helsinki Corpus of British English Dialects* on CoRD



Helsinki Corpus of British English Dialects

The Helsinki Corpus of British English Dialects (HD) is a collection of orthogra audio recorded speech, mainly from East Anglia and the South-West, with a mi Lancashire. The recordings were made in the 1970s and the 1980s by Finnish

The aim of the corpus is to provide material for linguistic research in the fields c sociolinguistics, discourse analysis, morphology, syntax and phonology. The c material for non-linguistic, multidisciplinary research, such as ethnography of c habits and history.

Project leaders: Ossi Ihalainen 1984-1993, Kirsti Peitsara 1997-2006, Anna-L onwards.
Fieldwork supervisors: In the 1970s, Tauno F. Mustanoja (Helsinki) in coope Orton (Leeds). Ossi Ihalainen supervised the fieldwork done in the late 1980s.
Size: Altogether **187** files consisting of **1008641** words.
Time periods: 1970s and 1980s (CAM, DEV, ELY, SOM, SUF), 1980s (ESS, L
Status: First stage completed in 2006, with corpus material available for resea ongoing, with plans of extending the corpus material with previously unpublish
Corpus data: The primary data are the audio recordings of spoken dialect spe
Funding: Finnish Cultural Foundation; The Academy of Finland, The Universi
Language: English, rural (Cambridgeshire, Devon, Isle of Ely, Somerset, Suffo Lancashire).

Table: Geographical regions covered in the corpus. The number in the brackets denotes the number of secondary informants (primary informant = sole speaker in a session or speaker of more than 500 words; secondary speaker = speaker with less than 500 words in a session with more than one informant)

Region (rural)	No of villages	No of informants
CAMBRIDGESHIRE	26	38 (+ 6)
DEVON	9	33 (+ 8)
ISLE OF ELY	20	52 (+ 6)
SOMERSET	15	24 (+ 5)
SUFFOLK	16	47 (+ 4)
Region (urban)		
ESSEX	3	6 (+ 1)
LANCASHIRE	3	6 (+ 1)
Total	92	206 (+ 31)

Recorded material

nd their informants, both men and women, with the help of locals. The fieldworkers went from village to village on y bus, carrying equipment weighing an average of 15 kilograms. The whole package included the recorder itself, a phone, batteries, power outlet adapters, correction tape, cables, fuse, etc. (Image on the right, click to enlarge.)

Some 210 hours of these recordings are now stored in digital form. The majority of the recordings have been transcribed by the fieldworkers themselves. Most of the recorded material has been donated to the English Department of the Helsinki University.

Working the corpus material into a computer corpus was an idea introduced by Professor Ossi Ihalainen in the 1980s. The project was integrated into VARIENG in 2000, under the coordination of Kirsti Peitsara. The texts have been transcribed orthographically, with only minor changes to show dialectal pronunciation. The nearly one-million-word corpus today includes speech from 237 informants recorded in 92 localities in six counties.

r's early
on notes

Sociolinguistic coverage

At the time of the interview in the 1970s and 1980s most of the male and female informants were retired and in their mid 70s or older, having thus received their education during the first decade of the 20th century. The re urban corpora, collected in the late e above in giving three-generation two- formants ranging from 19 to 70 years.

ora have a high proportion of males as he Devon corpus has no females as d female speech covers circa one fifth in Isle of Ely and Suffolk corpora. This l idea of males as best preservers of h was inherited from the SED model. An

In Essex, Ritta Kerman found her informants through the local pub, in Lancashire through the local librarian. In the summer of 2003, she donated copies of the Essex and Lancashire recordings to the VARIENG research unit.

Map of Essex (left) and Lancashire (right) and the localities visited.

1 Copford	4 Barrowford
2 Stanway	5 Colne
3 Colchester	6 Nelson

Sample #1

S.W. from Barrowford.
Recorded in 1988.
Age, sex: 40, female.
Occupation: Weaving instructor.
Description: S.W. tells about working with looms.

An' it takes the weft across one way, but it doesn't go back again. So it 'as to go down on to a like a conveyor belt, an' that 's why they 'ave so many of them because there 's one going across droppin' down and then another till they get the selves back up again.

Download the sound clip by right-clicking here and choosing Save As from the menu. *MP3 1:25 1,29mb.*

Sample #2

K from Stanway.
Recorded in 1988.
Age, sex: 60, male.
Occupation: Coach driver.
Description: K talks about England belonging to the EEC.

CoRD entry for the *Helsinki Corpus of British English Dialects* was designed by Anna-Liisa Vasko and Simo Ahava.

Copyright for original photography and audio samples held by project members.

VARIENG

The Research Unit for Variation, Contacts, and Change in English (VARIENG) is a Center of Excellence funded by the Academy of Finland. VARIENG functions at the Universities of Helsinki and Jyväskylä, in affiliation with the Departments of English and Modern Languages, respectively.

CoRD team

The CoRD concept was envisioned by Terttu Nevalainen, Jukka Tyrkkö and Minna Palander-Collin. The basic structure for CoRD entries was developed by Jukka Tyrkkö, Matti Rissanen, Terttu Nevalainen, Irma Taavitsainen and Arja Nurmi, who also designed the electronic submission and feedback forms. The template for the CoRD website was designed by Tanja Säily and Jukka Tyrkkö, the present CoRD coordinator.

Contact information

To contact VARIENG, see http://www.helsinki.fi/varieng/