European Science Foundation

Standing Committee for the Social Sciences (SCH)

ESF SCSS EXPLORATORY WORKSHOP

Corpora In Phonological Research

Scientific Report

Amsterdam, Netherlands, 15 - 17 June 2006

Convened by:

Gjert Kristoffersen[①], Marc van Oostendorp[②], Jacques Durand[③], and Chantal Lyche[④]

[①] University of Bergen, Bergen University Research Foundation (Unifob AS)
[②] Royal Netherlands Academy of Arts and Sciences
[③] Université de Toulouse - Le Mirail, ERSS
[④] Department of Literature, Area Studies and European Languages, University of Oslo

Executive Summary

Main Objectives of the Workshop

The aim of the workshop was to bring together researchers working with oral corpora in different areas of phonological research with experts on corpus building and text coding in order to explore the possibilities of establishing general standards for transcription and annotation of phonological corpora. The development of such standards will allow both for re-use of the data within other projects and for interoperability between the corpora that are built according to these standards.

Participants and venue

The workshop was attended by 17 participants from 10 different European countries. In addition the ESF-representative took part in the proceedings. We had originally planned for a somewhat bigger group (20-25 participants) but due to last minute cancellations, the actual number was reduced. Even if 10 countries were represented, especially the East European countries were underrepresented. The workshop took place on the premises of the Meertens instituut in Amsterdam.

Proceedings

The programme consisted of three main sections:[1]

1. Presentation of relevant work at the participants' home institutions. 13 presentations were given, all summarized below.
2. A short introduction to XML-coding by Tone Merete Bruvik, former executive officer of the Text Encoding Initiative Consortium.
3. Group sessions with ensuing plenary discussions where two main topics were discussed. On Friday we discussed how general guidelines for coding of metadata as well as structure and content can be established, disseminated and adhered to, and on Saturday morning we discussed future cooperation

Before the presentations started, Gjert Kristoffersen welcomed the participants on behalf of the convenors. The ESF representative then gave a short presentation of the European Science Foundations, and finally Jacques Durand situated the theme of the workshop within the larger scientific fields of theoretical phonology, phonetics and corpus linguistics.

Results

In general, the participants agree that the workshop was a success that promises future results in the form of a European network as well as the emergence of future common projects. The most robust result of the workshop is the participants' commitment to form a network through

[1] See below for the complete programme.

which they can continue cooperation. There was a shared feeling that we have both common research interests and a common problem in the fact that general standards for encoding phonological corpora are lacking today. The commitment to continue cooperation is therefore a commitment to develop such standards. A group of coordinators was named, and the whole group agreed to meet again in July 2007 in order to develop more concrete ideas for scientific cooperation.

Scientific content of workshop

Presentation of relevant work at the participants' home institutions

The majority of the participants had in good time before the workshop been asked to present work and resources at their home institutions relevant for the central topics of the workshop. John Nerbonne of the University of Groningen had been invited to participate at the workshop, but was unable to attend the full workshop due to other obligations. Since he was in Amsterdam at the time of the workshop, he was invited to present his work on Friday morning. The presentations were followed by questions from the audience, but the main discussion was postponed to Friday afternoon and Saturday morning. The discussion is therefore summarized in the following section.

Hans Basbøll, University of Southern Denmark

Basbøll gave a presentation of some Danish speech corpora that can be used for phonological analyses. *CorDiale* is a corpus of Danish dialects comprising 1 million running words (170 texts, 150 geographical locations), collected (1934-94) by the former Department of Danish Dialectology (University of Copenhagen), and *BySoc*, a corpus of Urban Sociolinguistics (Copenhagen) comprising 1.3 million running words in free-style speech connected to a search engine, collected and transcribed under the Danish Research Council initiative on Spoken Danish in its variations.[2]

The Danish National Research Foundation Centre for *Language Change in Real Time* (LANCHART), directed by Frans Gregersen, University of Copenhagen, is preparing a large corpus of recorded natural speech from all over Denmark. In many cases speakers have been recorded twice within a distance of several decades.[3] A coding system is under development. Phonetic tags mainly regard length and vowel quality, and these can be combined with very detailed metadata on interaction and genre. This project has the potential of offering unique possibilities for accounting for patterns of phonological variation and change within Danish.

DanPASS – A Danish Phonetically Annotated Spontaneous Speech Corpus, directed by Nina Grønnum, University of Copenhagen, consists of monologues and dialogues involving 22 speakers and about 70.000 running words.[4] Analysis and annotation is performed in

[2] See http://www.id.cbs.dk/~ pjuel/

[3] See http://dgcss.hum.ku.dk

[4] See http://www.cphling.dk/~ng/danpass_webpage/danpass.htm

PRAAT. The acoustic signal is segmented into syllables, words and prosodic phrases; and different tiers represent orthography, part-of-speech tagging, phonological notation, phonetic transcription, the pitch relation between stressed and post-tonic syllable, and phrasal intonation. *DanPASS* will be fully available in the beginning of 2007.

The *Odense Twin Corpus* consists of speech in natural settings from 6 families with twins (within the age range 0;5-6;2, most of them from 0;9 until 4+), in total about 2 millions running words of which ca. 345.000 are transcribed and 285.000 coded. For analysis they use *OLAM*, a semiautomatic system for coding and searching, developed by Claus Lambertsen, Berlin, and Hans Basbøll and Thomas O. Madsen at the Center for Child Language in Odense.[5] There are at the moment tiers for phonology, morphology, and orthography. Phonological information available in the system include distinctive features, segment patterns, syllable structure, stress patterns, stød patterns, and length patterns, and the system is also designed to account for phonology-morphology (-orthography) interaction. It is presently used in an acquisitional project in cooperation with W.U. Dressler, Vienna, Steven Gillis, Antwerp and Dorit Ravid, Tel Aviv.

Jacques Durand, Université de Toulouse;, Bernard Laks, Universié de Paris X; Chantal Lyche, University of Oslo; Noël Nguyen, Université de Provence ; Atanas Tchobanov, Universié de Paris X (Joint presentation of the project Phonologie du Français contemporain)

The main goals of The Phonology of Contemporary French (PFC) project aims at describing the phonological systems of contemporary French while taking into account geographical, social and stylistic variation. It involves over thirty researchers from a variety of countries working with a common protocol for the recording, partial transcription and analysis of over 500 speakers from the francophone world.

For each speaker, the corpus consists of two reading tasks, a word-list of 94 items including 10 minimal pairs and a small text. In addition, the speaker is recorded during a guided conversation where he answers questions from an investigator he is not familiar with, and during a free conversation with a person close to him/her. A minimum of five minutes of each conversation are transcribed.

For the analysis of schwa, liaison and prosody, an alphanumerical system has been devised and the coding is performed in Praat on individual tiers. The whole text and each conversation are coded for schwa and liaison (3 minutes for schwa, 5 for liaison). Prosody is coded on portions of the text and on a few non-contiguous portions of the conversation for a few selected speakers only. The coded transcriptions and the sound files for a given speaker are available online at the project site.[6] The site has a comprehensible search system and browsing facilities for the speaker database.

Semi-automatic speech processing tools have been developed to locate the position in the speech signal of the target vowels as produced in the word lists, and to extract formant

[5] See http://www.humaniora.sdu.dk/boernesprog

[6] http:// www.projet-pfc.net

frequencies at the mid-point of each vowel. Large-scale comparisons have been performed between both intrinsic and extrinsic vowel normalization methods proposed in the literature, to determine which of them allows speaker-dependent variations in the vowels' spectral shape to be factored out while preserving regional variations to the largest possible extent.

Anders Eriksson, University of Gothenburg

Erikson presented two databases based on recordings of 107 Swedish dialects spoken in Sweden and the Swedish speaking parts of Finland. The larger one, intended for research, consists of the entire recorded material from the Swedia 2000 project.[7] In this material each of the 107 dialects is represented by three speakers from each of the four categories of speakers in the database – older women, older men, younger women and younger men. Four types of speech material was recorded – spontaneous speech, a wordlist from which the phoneme inventory of the dialect may be derived, a wordlist from which the realization of the quantity system may be derived and a word list from which the realization of the tonal accent distinction may be derived. The database is searchable at several different levels. The total database holds about 800 hours of speech material (6–9 hours of speech per dialect).

Ulrike Gut, University of Freiburg

Gut presented the recently completed LeaP corpus, a fully text-to-tone aligned and extensively phonologically annotated speech corpus of learner English and learner German with a length of more than 12 hours.[8] It includes 359 recordings with 131 speakers with 32 different native languages as well as 18 recordings with native speakers and comprises four different speech styles: reading of a nonsense word lists, reading passage, retellings of the story and free speech in an interview. The corpus is extensively annotated on the phonological and prosodic level, including six tiers with manual annotation and two automatically annotated tiers. It is encoded in the XML-based data format TASX, which supports data annotation, format exchange and corpus searches. The talk focused on the multilevel annotation and the reliability of the manual phonetic annotations in the corpus as well as the advantages of the XML-based data format. Its main points were that the reliability off manual transcriptions in the areas of phonetics/phonology and prosody depend on the complexity of the coding schemas involved. It was suggested to standardize the annotation process for phonological corpora. The advantages of the XML-based data format were seen to lie in the easy convertibility of transcriptions made in the various mostly proprietary formats offered by speech analysis programmes.

[7] See http://swedia.ling.umu.se/ (Swedish only)

[8] See http://www.phonetik.uni-freiburg.de/leap/LeaPCorpus.pdf

Maria-Rosa Lloret, Universitat de Barcelona

Lloret presented the Catalan *Corpus Oral Dialectal* (COD).[9] The content of this corpus were collected with computerization in mind through a questionnaire of approximately 600 phonetic and morphological items and recordings of 10-minute samples of casual speech in 86 county towns of the whole Catalan-speaking area. 2-3 speakers were interviewed in each town. The use of a third speaker was designed in order to be able to select the majority, more frequent form in cases with variation where data computerization required a single answer. The informants were 30-45 years old, middle class speakers, with a minimum amount of formal education. The selection of localities and speakers was done with the purpose of recording the common mode of speaking of the inhabitants of more urban areas.

The results of the questionnaire have been systematized in databases which now contain around 162.493 phonetic and morphological items, which are going to be published in a CD at the end of 2006. The databases contain phonetic and orthographic transcriptions as well as the morphological segmentation of each item. Up to now, 50 free-speech samples have been orthographically and phonetically transcribed and aligned with their corresponding sound files. The corpus is currently being used in phonological, morphological and phonetic studies, focusing on language variation and linguistic change. The data from the questionnaire are also used to develop dialect-grouping techniques based on a multivariate analysis, in line with dialectometrics and the cluster analysis. The content of the databases is going to be made accessible through computerized maps, which will incorporate the corresponding sound files.

Inger Moen, University of Oslo

Moen introduced the *Electronic database of Norwegian speech sounds,* which is currently under development at the University of Oslo.[10] It is intended to be a resource for phonetic research and practice. Beside its theoretical contribution as a research tool and as a basis for cross-linguistic comparison, one of the primary goals is for the database to be a tool that can assist in the treatment of speech and language impaired subjects with articulatory disorders. The database will include data from eight Norwegian speakers, with all consonantal phonemes in relevant vocalic contexts. A number of registration techniques are used: the primary tools are simultaneous EPG and EMA recordings, registering oral passive and active articulation, respectively. This is complemented by measurements of vocal fold activity, intra-oral air pressure, and nasal and oral emission of airflow. With the addition of the acoustic picture, provided by spectrography, all major aspects of speech production are covered and may be displayed in an explicit manner. The corpus will be accessible in three different formats, of which the research database with unprocessed raw data is the most important from the perspective of the current workshop.

[9] See http://www.ub.edu/lincat

[10] See http://www.hf.uio.no/iln/forskning/forskergrupper/klinisk-ling/prosjekter/database/ (Norwegian only)

Hermann Moisl, University of Newcastle upon Tyne

Moisl presented The Newcastle Electronic Corpus of Tyneside English (NECTE), which is a corpus of dialect speech from Tyneside in North-East England.[11] It is based on two pre-existing corpora, one of them collected in the late 1960s by the Tyneside Linguistic Survey (TLS) project, and the other in 1994 by the Phonological Variation and Change in Contemporary Spoken English (PVC) project. NECTE amalgamates the TLS and PVC materials into a single Text Encoding Initiative (TEI)-conformant XML-encoded corpus and makes them available in a variety of aligned formats: digitized audio, standard orthographic transcription, phonetic transcription, and part-of-speech tagged. Two issues need to be resolved if the full research potential of the corpus is to be realized. These are availability of reliable transcriptions and the establishment of representational standards.

Phonetic / phonological transcription of speech presents three well known problems, all of them attributable to manual production by human transcribers: (i) it is hugely time-consuming, (ii) it is subjective in the sense that different transcribers typically produce different representations for a given speech segment, and (iii) as the size of the corpus grows, so does the difficulty of maintaining consistency of practice across the transcription. What is required is automation of the transcription process using computational tools. The project has begun to look at a way of doing this using orthographic transcription to disambiguate the audio signal.

As to representational standards, it has long been standard practice for corpus creators to define their own standards, with the result that they are not easily portable. Various standards have been proposed in response. The currently dominant one is the Text Encoding Initiative. There are two problems with TEI for representation of phonetic / phonological corpora, however. On the one hand, the TEI recommendation for linguistics corpora is vestigial, and needs to be developed if it is to be useful for any but the most basic representational purposes. On the other, the plethora of XML tags makes TEI-encoded corpora difficult to use directly, and requires development of XML-based analytical applications.

John Nerbonne, University of Groningen

Nerbonne presented work currently being done at the University of Groningen on techniques for measuring and analysing pronunciation similarity, focusing in particular on the sort of material found in dialect atlases. Therese Leinonen, a graduate student in Groningen, collaborated in the preparation of the talk. The pronunciation comparison is implemented using a modified version of string edit distance. In the talk the basic techniques and also results for Dutch, German, Norwegian, American English, Sardinian, and Bulgarian were reviewed, including some recent work in which the effect of normalizing for fast speech rules was explored, and some ongoing work seeking to determine regular pronunciation differences. So far, this work has *not* been successful in demonstrating that more refined notions of pronunciation similarity in segments provide more sensitive measures of overall

[11] See http://www.ncl.ac.uk/necte/

dialectal affinity. This has inspired the team to survey some other areas of the language sciences. The conclusion is that the notion 'similar in pronunciation' has not received a great deal of attention. An (early) overview of the work, "Edit Distance and Dialect Proximity", was published as the introduction to the reissue of Sankoff and Kruskal's classic text on string comparison, *Time Warps, String Edits and Macromolecules: Introduction to the Theory and Practice of Sequence Comparison*, available at http://www.let.rug.nl/nerbonne/papers/tw-se-mm.pdf.

Marc van Oostendorp, Meertens Instituut

In his talk, van Oostendorp discussed the so-called GTP (Goeman-Taeldeman Project), which has resulted in a database of approximately 1.000.000 words from 613 Dutch and Frisian dialects in the Netherlands, Flanders and France, which in turn has led to two paper atlases: the *Fonologische Atlas van de Nederlandse Dialecten* and the *Morfologische Atlas van de Nederlandse Dialecten.*[12] The database project started in 1978, and had the ambition to collect dialect data which would (also) be relevant to contemporary linguistics. In the presentation, the atlases and the database were evaluated in this respect, and it was shown that some questions might indeed be answered by using the database. An example is the issue of Feature Economy (Clements 2003), which predicts that phonological features should be used maximally economically in a given linguistic system. The predictions of this theory can be tested on a dialect database, but some problems, mainly relating to the transcriptions also arise: Despite the establishment of a strong protocol, transcribers may differ among each other, causing certain perturbations in the data. In addition, transcriptions seldom are at exactly the right level of analysis.

Gjert Kristoffersen, University of Bergen

Kristoffersen introduced *Talesøk*, a tool for automated search in recorded speech.[13] A basic goal of this project is to construct transcriptions of speech that can serve as general search tools so that the need for explicit coding of content (not of structure) is minimized. This will provide a means for obtaining efficient access to recorded and digitized speech data. At the centre of the project is a corpus of Norwegian speech data in the form of socio-linguistic interviews with speakers of different varieties of Norwegian. At present, the corpus contains about 500.000 words. Basic requirements of a transcription that will meet these goals, are that it must give rapid and efficient access to the primary data, i.e. the recorded sound, and at the same time secure maximal consistency and cost efficiency, given the fact that transcription is time consuming and therefore expensive. In the second part of the talk, the distinction between coding of structure and coding of content was discussed. While coding of structure (segments, syllables, morphology, expected tonal melody etc) is necessary to the extent that searches that one may want to do in the corpus presupposes structural information, there are several arguments against coding of content, e.g. specific variants of a certain variable: First,

[12] See http://www.meertens.knaw.nl/projecten/mand/GTPintroE.html

[13] See http://spraktek.aksis.uib.no/projects/17

any coding of content is an interpretation of data (like e.g. phonetic transcription) where different potential sources of error may result in the data not being represented correctly and consistently. Second, coding of content will often be project specific and of less interest to later users of the corpus, and third, it may make a corpus unwieldy and difficult to use for later users. On the other hand, content must be coded somewhere, so the real question is whether it should be done inside the corpus or externally, e.g. in a database program or a statistical package with no links to the corpus. If coding is done within the corpus, the data used within a given project will be more accessible for others. It is all the same preferable to hold structure and content apart, e.g. in separate tiers. Division into tiers may also be a desirable option when different variables or units are to be coded.

Dimitris Papazachariou (in collaboration with Angela Ralli), University of Patras

In this paper, a spoken corpus on dialectal data from six different Greek dialects was presented, with emphasis on the rationale behind the choice of methods, tools, and techniques. In particular, the corpus consists of data from six different research projects, including dialects from Northern Greece (Lesvos and Kilkis), from Southern Greece (Patras), as well as Greek dialects from abroad (in Italy and in Turkey). The recordings consist of more than 300 hours of casual and friendly conversations, taken from more than 400 informants (unevenly distributed). The ethnographic methodology used for data collection and recording was presented, with particular emphasis on the rationale of the methods chosen. Arguments were put forward in favour of orthographic transcription, compared to phonetic and phonemic transcription. The reasons for using Praat in the transcription and annotation of the digitized recordings were discussed, along with the problems that were encountered using Praat, such as the fact that it does not incorporate Greek Unicode characters. Furthermore, the methodology for handling the transcription and annotation of the casual speech of different speakers within each recording was discussed, and finally, it was demonstrated how scripting in Praat has proved to be a useful tool, e.g. in obtaining measurements of vowel formants as part of a phonetic analysis of the vowel system of Modern Greek.

Anthi Revithiadou, University of the Aegean

Revithiadou presented a corpus on Ofitika Pontic (OP), a Greek dialect once spoken in the area of Trapezounta in the coast line of the Black Sea, and now spoken in the village of Nea Trapezounta in the Perfecture of Pieria (Macedonia, North Greece) by approximately 1500 speakers. There is no previous research done on this Pontic dialect. In May 2004, the cultural club ‘Ypsilantis’ invited a group of linguists and students from the University of the Aegean to do field-work and record the few remaining speakers of OP. The group consists of two linguists, two students of linguistics and a local member of the linguistic community who is a ‘bilingual’ himself. The group has visited the village three times. Some of the basic goals are to record the dialect in its current form and at the same time to extract linguistic (and social) information by means of participant observation, to collect linguistic data that would allow the analysis of patterns of language use by community members who belong to different age groups, so that they could be juxtaposed to the structure of their personal networks, and to get

a clear picture of the phonological (and morpho-syntactic) features of the dialect and determine the degree of influence of Standard Greek. The data are recorded in a digital form and phonemically transcribed by two appropriately trained students. The recordings consist of 14 hours of conversations with 18 speakers, all of them members of a network that resides within the definable territory of the village. 11 speakers were born between 1916 and 1940, 5 speakers were born between 1941 and 1969 and 2 speakers were born between 1974 and 1982. Unfortunately, the linguistic data have not been organized in a database due to lack of funding.

R. also introduced briefly an ongoing project on Greek contact varieties with Turkish such as the Greek spoken by the Muslim Community of Rhodes. The research follows ethnographic data collection procedures that allow the team to obtain a realistic picture of patterns of language use, patterns of the informal social organization, i.e. networks, operating in the community, as well as patterns of contrast between two environments, the urban and the suburban environment.

Anne Catherine Simon, Université catholique de Louvain

Simon represents the *Research Centre VALIBEL* (VAriétés LInguistiques du français en BELgique), founded in 1988, which pursues the building of a large data set of spoken French within the French speaking community of Belgium.[14] The main principles adopted for the analysis of variation in spoken speech are:

1. *Metadata*: The analysis of speech variation (regional, stylistic, language change…) requires detailed information about the speaker, the speech setting, as well as the methodology used in data collecting and transcriptions.
2. *Sound files and transcription*: Transcription cannot substitute the speech recording. Any database must therefore allow for easy access to sound files which remain the core data for the analyst. Secondary data (transcripts, coding…) are built up using the Praat software. (Note that text grid is the format used for archiving transcription and coding.)
3. *Prosodic variation analysis*: The syllable (alternatively, the phoneme) is the basic unit for analysing prosodic structure and variation. Morpho-syntactic annotation is needed when analysing prosody, whatever the theoretical model.

VALIBEL takes part in various projects that meet the scientific points targeted by the Corpora in Phonological Research workshop. Three of the projects were outlined, all having a special emphasis on the analysis of prosody.

1. *MOCA*: A *Multimedia Oral Corpora Administration* system for analysing spoken (including conversational) data in a sociolinguistic perspective (VALIBEL & CENTAL, UCL; Peter Gilles, University of Freiburg).

[14] See http://valibel.fltr.ucl.ac.be/

2. *IVTS*: (*Intonation Variation Transcription System*). A manual transcription system for studying intonation variation (adapted from IViE; B. Post, U. of Cambridge; E. Delais-Roussarie, Paris 7; A.C. Simon)
3. Multilayered prosodic annotation for *spontaneous speech* combining various levels of analysis:
    a. semi-automatic phonetic alignment and syllabification (J.-P. Goldman, U. of Geneva);
    b. semi-automatic morpho-syntactic labelling of spoken French (Anne Dister, CENTAL/UCL);
    c. a multilayered prosodic annotation system for studying prosodic phenomena (vowel lengthening, intonation contours, accentuation…) while taking into account the aspects of spontaneous speech delivery (interruption, self-repair, hesitation,…) (F. Poiré, U. Western Ontario; P. Mertens, KULeuven; A.C. Simon). This annotation system combines with the automatic intonation transcription PROSOGRAM, by P. Mertens, see http://bach.arts.kuleuven.be/pmertens/**prosogram**/

A short introduction to XML-coding

Tone Merete Bruvik of the University of Bergen was responsible for this part of the programme. She is one of the world's leading experts on text encoding, and acted as Executive Director of Text Encoding Initiative Consortium from 2001 to 2004.[15] The reason why this topic was made part of the programme is that interoperability between corpora presupposes a common standard of encoding metadata as well as structure. The Text Encoding Initiative (TEI) represents the most commonly used standard today. Knowledge of text encoding and TEI differed among the participants of the workshop, and Ms. Bruvik's presentation was therefore very useful as an introduction to the ensuing discussion.

After having presented some examples of what a coded text may look like and what standards a digitally encoded text should meet, B. briefly introduced different text encoding languages before XML was presented in more detail, including the concepts of DTDs and schemas. TEI was then introduced and described in considerable detail, including examples of how linguistic content in a given text may be coded. Finally some problems with TEI, such as overlapping and discontinuous elements were discussed, in addition to more general questions such as the choice of what to code, the use of separate tiers and the need for a community that make basic decisions with respect to standards etc.

Group sessions with ensuing plenary discussions,

The discussions were divided by topic into two independent sessions, the first on Friday afternoon on coding and coding standards and the second on Saturday morning on future cooperation. Both Friday and Saturday the participants were divided into two groups.

[15] See www.tei-c.org

Coding

The topics for the discussion Friday afternoon was stated as follows: The usefulness of and potential of corpora in phonological research seem to be established beyond doubt in the presentations Thursday afternoon and Friday morning. The most pressing question now is how general guidelines for coding of metadata as well as structure and content can be established, disseminated and adhered to, and how we can build research projects on the European level where these guidelines can be put into practice.

There was a general agreement that a speech corpus designed for phonological research must as a minimum consist of the following:

- A sound file
- An orthographic transcription aligned with the sound file
- A set of standardized metadata that defines the corpus

Transcription: For cost as well as reliability reasons, the basic transcription of the sound file must be orthographic. But even if orthographic transcriptions are less costly and more reliable, defining standards for consistent transcription of speech by means of standard orthography is not trivial, and must be addressed. Transcription and sound must be aligned, so that the sound corresponding to a specific part of the transcription can be easily accessed.

Depending on the goals of a specific project, other types of transcriptions, such as phonetic or phonemic transcription, may be added, but they should not supplant the orthographic transcription. Different projects will have different needs for phonological tiers, depending on different kinds of use. The number of tiers is in principle limited, but a recommended list of relevant tiers might be useful.

A basic problem with all transcriptions is that they are products of interpretation. The result is that people don't trust each others transcriptions. Within a more long term perspective, the possibility of automatic transcriptions, which will make transcriptions at least more objective, (but not necessarily more correct), should be investigated.

Metadata: We need standards for coding of metadata in order to be able to work on each other's databases, and to avoid inventing the wheel over and over again.

Two basic questions are:

- What are the relevant metadata?
- How should they be coded?

We in other words need a specification of the relevant metadata before we can decide how to code them. Here the question arises whether it is possible to define a set of metadata that is relevant for all projects, and whether project specific need to code additional metadata should

be catered for by means of a set of general guidelines. It was pointed out that the IMDI (ISLE Meta Data Initiative) already offers a standard for different kinds of metadata.[16]

As to the coding itself, XML should be recommended. It is flexible, and allows users to define their own tags. An important question is whether only standards for coding metadata should be recommended, or whether the coding standards should be extended to linguistic content as well. The latter position implies that tags will reflect theoretical positions.

Future cooperation

The topic of the discussions Saturday was future cooperation, and the possibility of creating common projects with a realistic chance of obtaining funding. There was general agreement that the outcome of the workshop should be, as a first step towards a more permanent network, the formation of a European network consisting of the participants of the workshop. We also want to include those who were invited to the Amsterdam workshop, but for some reason were not able to participate. In addition, we want to broaden our contacts with Eastern Europe, which is underrepresented in the present group, both with respect to countries and to language families.

We agreed to call the network *European Corpus Phonology Group (CorPho)*. It will assemble researchers and research teams interested in combining insights from theoretical phonology, both diachronic and synchronic, linguistic variation studies, phonetics and corpus linguistics. The responsibility of launching the network was delegated to a group of three coordinators: Ulrike Gut (Freiburg), Marc van Oostendorp (Amsterdam) and Gjert Kristoffersen (Bergen).

We agreed on the following immediate goals for the network:

- Organize a website where information about the group and about the Amsterdam workshop will be published, along with pdf-files of the presentations given in Amsterdam.
- Organize a discussion group on the net
- Create a database of the corpora in use within the network
- Take steps towards establishing standards for metadata as well as coding
    - Evaluation of the IMDI scheme along with metadata structures already adopted for major speech corpora
    - Approach TEI in order to explore the possibility of defining an appropriate tagset as part of the TEI
- Take steps to create contacts and cooperation with groups working on corpus based speech technology.
- Explore the possibility of creating a handbook of speech corpora and corpus phonology
- Explore the possibilities of finding funding for a summer school in 2008

[16] See http://www.mpi.nl/IMDI/

We agreed on meeting again in Toulouse on 5 July 2007. An important topic for the meeting should be discussion in more detail of a scientific theme for the group, which can serve as a framework for the development of a major research project with European funding. At the workshop this year, we agreed on a first approximation: *Imposing uniform representations on non-discrete speech*.

Assessment of Results

The most robust result of the workshop is the participants' commitment to continue cooperation through the CorPho network. This reflects the shared feeling that emerged from the workshop that we have both common research interests and a common problem in the fact that common standards for encoding phonological corpora are lacking today. The commitment to continue cooperation is therefore a commitment to develop such standards. This will be part of the more comprehensive endeavour to create big research infrastructures on the European level, as manifested by the Research Infrastructures program in the RP7 and the ESFRI initiative.[17] The direction that the cooperation will take will mainly be towards basic research in accordance with the principles that forms the basis of the Ideas chapter of RP7, but the possibility of developing more practically oriented projects that will better meet the requirements of other categories within RP7 will also be considered.

Workshop programme

Thursday 15 June 2006

14.00	**Gjert Kristoffersen:** *Welcome*
14.10	**Presentation of the European Science Foundation (ESF)** **Karl Pajusalu** (Dept. of Estonian and Finno-Ugric Languages, University of Tartu) (Standing Committee for the Humanities)
14.20	**Jacques Durand**: Opening remarks
14.40 - 18.00	**Presentation of relevant work at the participants home institutions**
14.40	**Hans Basbøll**, University of Southern Denmark
15.00	**Project presentation: Phonologie du Français contemporain**. **Chantal Lyche/Jacques Durand/Bernard Laks/Noël Nguyen/Atanas Tchobanov**
15.40	**Ulrike Gut**, University of Freiburg
16.00	*Coffee break*
16.30	**Anders Eriksson**: University of Gothenburg

[17] http://cordis.europa.eu/fp7/capacities.htm#1 and http://cordis.europa.eu/esfri/

16.50	**Hermann Moisl**, University of Newcastle upon Tyne
17.10	**Maria-Rosa Lloret**, University of Barcelona
17.30	**Inger Moen**, University of Oslo
18.00	*End of day 1*

Friday 16 June 2006

09.00	(Invited speaker): **John Nerbonne**, University of Groningen
09.20	**Marc van Oostendorp**, Meertens institute
09.40	**Gjert Kristoffersen**, University of Bergen
10.00	*Break*
10.30	**Dimitris Papazachariou**, University of Patras
10.50	**Anthi Revithiadou**, University of the Aegean
11.10	**Anne Catherine Simon**, Université catholique de Louvain
11.30	*Break*
11.45	**Tone Merete Bruvik**, University of Bergen: Short introduction to XML-coding
12.30	*Lunch*
14.00 - 18.00	**Group session and plenary discussion 1**

Saturday 17 June 2006

09.00 – 12.30	**Group session and plenary discussion 2**
12.30	*End of workshop*

Final list of participants

In addition to the invited participants, Professor Karl Pajusalu (Dept. of Estonian and Finno-Ugric Languages, University of Tartu) attended the workshop as representative of the Standing Committee for the Humanities of the ESF.

List of participants

Hans Basbøll
University of Southern Denmark
Denmark

Inger Moen
University of Oslo
Norway

Tone Merete Bruvik
University of Bergen
Norway

Hermann Moisl
University of Newcastle upon Tyne
UK

Jacques Durand
Université de Toulouse - Le Mirail
France

Anders Eriksson
University of Gothenburg
Sweden

Ulrike Gut
University of Freiburg
Germany

Gjert Kristoffersen
University of Bergen
Norway

Bernard Laks
Université de Paris X
France

Maria-Rosa Lloret
Universitat de Barcelona
Spain

Chantal Lyche
University of Oslo
Norway

Noël Nguyen
CNRS & Université de Provence
France

Marc van Oostendorp
Royal Netherlands Academy of Arts and Sciences
The Netherlands

Dimitris Papazachariou
University of Patras
Greece

Anthi Revithiadou
University of the Aegean
Greece

Anne Catherine Simon
Université catholique de Louvain
Belgium

Atanas Tchobanov
Université de Paris X
France

Statistical information of participants

Age bracket

No precise information have been collected on the age of the participants, but the convenors consider the age of the participants to be well distributed across the different age brackets, with the youngest participants in their thirties and the oldest above sixty.

Countries of origin

Ten countries were represented among the participants: Belgium, Denmark, France, Germany, Greece, the Netherlands, Norway, Spain, Sweden and the United Kingdom. Despite efforts made by the convenors during the preparatory stages of the workshop, we did not manage to identify suitable participants from East Europe.

Gender

Among the participants were seven women and ten men.