The Workshop Programme

14.30 Opening

Session 1: Corpus compilation and (orthographic) transcription

14.35 The Corpus of Spoken Israeli Hebrew (CoSIH); Phase I: The Pilot Study
− Shlomo Izre'el and Giora Rahav, Tel-Aviv University

15.00 The ICSI Meeting Corpus: Near-field and far-field, multi-channel transcriptions for speech and language researchers
− Jane Edwards, International Computer Science Institute, and Institute of Cognitive Studies, UC Berkeley

15.25 Orality and difficulties in the transcription of a spoken corpus
− Ana González Ledesma, Guillermo De la Madrid Heitzmann, Manuel Alcántara Plá, Raúl De la Torre Cuesta, and Antonio Moreno-Sandoval, Universidad Autónoma de Madrid

15.50 Processing spoken language data: The BASE experience
− Sarah Creer and Paul Thompson, University of Reading

16.15 – 16.45 Coffee break

Session 2: Corpus annotation

16.45 A "toolbox" for tagging the Spanish C-ORAL-ROM corpus
− José Guirao and Antonio Moreno-Sandoval, University of Granada and Universidad Autónoma de Madrid

17.10 Towards the creation of an electronic corpus to study directionality in simultaneous interpreting
− Claudio Bendazzoli, Cristina Monti, Annalisa Sandrelli, Mariachiara Russo, Marco Baroni, Silvia Bernardini, Gabriela Mack, Elio Ballardini and Peter Mead, University of Bologna

17.35 Developing a dialogue act coding scheme: An experience of annotating the Estonian Dialogue Corpus
− Tiit Hennoste, Mare Koit, Andriela Rääbis, and Maret Valdisoo, University of Tartu

18.00 – 18.15 Short (15-minute) break

Session 3: Extending corpus parameters

18.15 WinPitch Corpus. A text to speech analysis and alignment tool for large multimodal corpora
− Philippe Martin, Université Paris 7

18.40 Automatic annotation of speech corpora for prosodic prominence
− Fabio Tamburini and Carlo Caini, University of Bologna

19.05 Towards dynamic corpora
− Daan Broeder, Hennie Brugman, Nelleke Oostdijk, and Peter Wittenburg, Max Planck Institute for Psycholinguistics and University of Nijmegen

19.30 Closing remarks

Workshop Organisers

Nelleke OOSTDIJK	University of Nijmegen
Gjert KRISTOFFERSEN	University of Bergen
Geoffrey SAMPSON	University of Sussex

Workshop Programme Committee

Daan BROEDER	Max Planck Institute
Emanuela CRESTI	University of Florence
Gjert KRISTOFFERSEN	University of Bergen
Tony MCENERY	University of Lancaster
Nelleke OOSTDIJK	University of Nijmegen
Pavel IRCING	University of Western Bohemia
Geoffrey SAMPSON	University of Sussex
Antonio Moreno SANDOVAL	University of Madrid
Jean VERÓNIS	Université de Provence

Table of Contents

Author Index

The Corpus of Spoken Israeli Hebrew (*CoSIH*); Phase I: The Pilot Study

Shlomo Izre’el and Giora Rahav
Department of Hebrew and Semitic Languages; Department of Sociology
Tel-Aviv University
IL-69978 Tel-Aviv, Israel
{Izreel; grrhv}@post.tau.ac.il

Abstract

The Corpus of Spoken Israeli Hebrew (CoSIH) is, to the best of our knowledge, the first corpus designed to integrate both demographic and contextual variables in its compilation of texts. The suggested design is culturally dependent to suit the structure of the Israeli Hebrew speech community, yet the principles governing this design are such that they would service study of many other speech communities, to the extent that the design itself may be employed in the compilation of other language corpora with the necessary, culture-dependent modifications. A detailed description of the design can be found in Izre’el, Hary & Rahav (2001).

In the paper offered for the workshop, we describe the pilot study of *CoSIH*, its procedures and some of its lessons. The results of the pilot study will bring about some changes in the final model of *CoSIH* and in some procedural strategies. We will address a few of the key issues involved in the construction of the corpus in order to achieve the analytical model we have designed. These are: (1) Demographic sampling and recruiting informants; (2) Evaluation of sequential longitudinal recording: technical matters and ethical issues; (3) Contextual sampling: long- and short-term time sampling, speech sampling; (4) The concept of ‘cell’. Lastly, the issue of transcription and annotation will be addressed briefly.

Introduction: The Corpus of Spoken Israeli Hebrew (*CoSIH*)

The objective of the *CoSIH* project is at creating a corpus of spoken Israeli Hebrew in order to facilitate research in a range of disciplines concerned with the Hebrew language, the sociolinguistics of the Hebrew speaking community in Israel, and with the general methodology of Corpus Linguistics. The corpus will be disseminated publicly in multimedia format.

A detailed description of the *CoSIH* project has been published in Izre’el, Hary & Rahav (2001). In this paper we describe the pilot study of *CoSIH*, its procedures and some of its lessons. The results of the pilot study will bring about some changes in the final model of *CoSIH* and in some procedural strategies. This discussion should be viewed as an appendix to the above-mentioned paper, and each section below will be referring to the respective section in that paper (abbreviated IJCL’01). A short summary of the *CoSIH* project is nevertheless in order.

With the outburst of corpus linguistics and the tremendous advance in the use of computers, many spoken corpora have been compiled and disseminated. Some of them have been compiled with attention to demographic and contextual varieties (useful gateways to reviewing such efforts are, among others, the SFB 411 one or the Gateway for Corpus Linguistics on the Internet; further references and some discussion can be found in *CoSIH*’s website and in IJCL’01). *CoSIH* is, to the best of our knowledge, unique in its method to combine both kinds of variables into a single model.

CoSIH is designed to include a representation of most varieties of spoken Hebrew as it is used in Israel today. *CoSIH* will consist of two complementary corpora: a main corpus and a supplementary corpus. The main corpus, which will comprise about 90% of the entire collection, will be sampled statistically. For analytical purposes it will use a conceptual tool in the form of a multidimensional matrix combining demographic and contextual tiers. The supplementary corpus will include about 10% of the collected data, and will add to the statistically sampled corpus some targeted demographically sampled texts and a contextually designed collection.

Daylong recordings of 950 informants and 50 other linguistic events (mostly from the media) will be collected within one year along with respective sociolinguistic data. These recordings will be evaluated, and a sample from each will be transcribed, to set up a five-million-word corpus.

CoSIH will be a basis from which research in many diverse areas will be launched, including, inter alia, theoretical and applied linguistics, sociolinguistics and cultural studies, communication studies, corpus linguistics, computational linguistics, translation studies, and many more.

While these are mostly long-term objectives, our immediate objectives are: Analysis of the Israeli linguistic community: its ethnolinguistic distribution, its linguistic and sociolinguistic behavior and attitudes; the study of demographic and contextual varieties (‘dialects’ and ‘registers’) as related to corpus compilation. Setting up a spoken corpus constitutes the initial phase of a major change in our view of language, i.e., as a multi-variant and dynamic continuum. Looking at language differently, as a multi-variant dynamic continuum is a primary target that 21st century linguistics should adopt, and the compilation of corpora is a necessary initial stage for such an endeavor.

The Corpus of Spoken Israeli Hebrew (*CoSIH*) is, to the best of our knowledge, the first corpus designed to integrate both demographic and contextual criteria in its compilation of texts. The design is highly innovative in this respect, and it is expected that its implementation will be a significant contribution to the discipline of corpus linguistics.

The suggested design is culturally dependent to suit the special structure of the Israeli Hebrew speech community and thus includes both native and non-native speakers of Hebrew. Yet the principles governing this design are such that they would service study of many other speech communities, to the extent that the design itself may be employed in the compilation of other language corpora with the necessary, culture-dependent modifications.

CoSIH Phase I: Pilot study

The pilot study, which included also the first steps of a pretest, aimed at achieving the following goals:

(1) To review a variety of linguistic groups among the Israeli Hebrew speech community in terms of linguistic and sociolinguistic behavior.
(2) To study issues involved in random sampling of the population.
(3) To study procedures involved in recruiting informants, eliciting natural recordings and sociolinguistic data.
(4) To study differences in attitude towards cooperation of informants from different sections in the population.
(5) To study issues involved in a sequential longitudinal recording by informants.
(6) To study technical tools, data recording techniques, and transcription issues.
(7) To make preliminary observations as regards linguistic contexts as related to different types of population.

Procedures taken were as follows:

(1) Recruiting informants in quota sampling and getting their preliminary consent to take part in this research.
(2) Instructing each informant as regards recording.
(3) Sequential recording by informant in a variety of time spans.
(4) Tapes collection from informants; inquiry about settings and conditions of the recordings made.
(5) Conducting a sociolinguistic interview.
(6) Questioning the informants as regards technical issues and problems encountered.
(7) Signing consent forms granting us permission to use the recordings.
(8) Preliminary organization of raw data (recordings and written forms).
(9) Evaluation of recordings: quality, language use, sufficient data, etc.
(10) Time sampling and selection of recorded samples to be included in the corpus.
(11) Data organization, registration in database, storage.
(12) Hebrew transcripts.
(13) Selection of segments for expanded analyses: phonetic transcription; glossing; English translation.
(14) Analysis of recordings for evaluation of distinct linguistic varieties for demographic and contextual variation.
(15) Evaluation of Phase I as a whole.

Some Key Issues: Goals, Alternatives and Lessons Gained

In this part of our paper we would like to address a few of the key issues involved in the construction of the corpus in order to achieve the analytical model we have designed. These are: (1) Demographic sampling and recruiting informants; (2) Evaluation of sequential longitudinal recording; (3) Contextual sampling; (4) The concept of 'cell'.

(1) Demographic Sampling and Recruiting Informants (IJCL'01: §§5.1.1, 5.2.1)

While the representative informants for *CoSIH* will be recruited by a probabilistic procedure, we have used quota sampling for the pilot study, trying to reach a wide coverage of the main socio-demographic groups in the population. Recruiting informants for this phase was made by three data collection agencies (a university associated agency and two well recognized, reputable commercial agencies). Each of the agencies was asked to collect data from 16 informants according to the demographic categories presented in Table 1:

Age	Edu-cation	Ashke-nazi	Mizra-hi	Arabs	Special groups
young	≤high school				
	>high school				
old	≤high school				
	>high school				

Table 1: Demographic categories

The three first groups (=columns) were set to fit, mutatis mutandis, the major demographic sections of the Israeli Hebrew speaking community: Jews of European or other Western ethnic origin ('Ashkenazi'); Jews of Asian or African ethnic origin ('Mizrahi'); non-Jews, of which the majority are Arabs, comprising ca. 20% of the Israeli population. The fourth column, 'special groups', was set to consist of three demographic sections for which we hypothesized to show significant differences in their use of language and in their linguistic structure: ultra-orthodox, soldiers and members of other security forces, and recently-arrived immigrants. Each agency was assigned one of these latter groups.

Of the three major ethnic groups, each agency was assigned to recruit four informants: two young (<20) and two old (>50), two with high education, two without. Lastly, each agency was instructed to recruit men and women in equal numbers, irrespective of any of the other criteria.

By choosing to hire a data collection agency we followed the procedure of BNC. We hired three agencys at the pilot phase in order to study procedures and pave the way to select one or more for the larger project, and we now have some idea about the pluses and minuses of each.

This decision has proven right. Academics in general, and linguists in particular, are not the ideal people to knock on doors and persuade people to join them in their research. Survey employees have enough patience and experience to do that, given that they themselves are persuaded by the need to conduct such a research. Money too is a factor in recruiting informants, although not of any kind. The rich or yuppies would not be tempted to be exploited for less than $50. Others would be too shy to do so in any case, or too short of self-confidence. People like me, who tend to get annoyed from answering commercial phone calls, will also tend to decline this generous offer…

The rate of consent to take part in such a burdening undertaking is an important factor for achieving a reasonable representative sample of the population. The B.I. and Lucille Cohen Institute for Public Opinion Research at Tel-Aviv University conducted a telephone survey for us on this issue, asking the following question:

- Would you be willing to take part in the future in a unique research in which you will be asked — for payment — to record all your daily activities during one day?

This survey was conducted on 1170 people, who consisted a representative sample of the Jewish population. A representative sample of 1170 individuals was surveyed. 40% of the respondents answered this question positively. However, as the response rate was about 55% of the households (or apartments), this means that the rate of positive responses may be as low as 22% of the whole population. Among those, only half, viz., 11%, are expected to eventually agree to full cooperation.

Among the Arab population, the consent rate was 24% out of 150 people who were asked. This means a much lower rate of consent than in the Jewish population. As expected, there are differences in consent tendencies among various sections in the population. For example, consent is lower among men than among women in the Jewish sector, while it is higher among men than among women in the Arab sector (the difference seems to be lower among Arab women with high education). There is further a problem of language use among Arab women of lower education, as they do not tend to use Hebrew in daily life, if they speak the language at all. Arabs tend in general to be less open to take part in research of the type proposed, due to their more prominent concern regarding privacy, as well as due to some political anxiety. Concern for privacy is shared by other sectors in the population, notably ultra-orthodox and soldiers. Political anxiety may also be found among new immigrants from the Former Soviet Union. We therefore expect problems in sampling in some sectors of the population, and may need to resort to quota sampling if the random sampling will result in lesser representation of some sectors.

(2) Evaluation of Sequential Longitudinal Recording (IJCL'01: §5.2.2)

(a) Technical Matters

We used Sony TCD-D100 DAT recorders with Sonic Studios stereophonic DSM-1S/L microphones. Each cassette has a capacity of four hours of quality recording. The acoustic output is excellent. However, the recorders seem to have caused difficulties in technical handling, especially at the point of replacing cassettes. Therefore, our first recruited informants from each agency recorded 8 hours each, and the last got to a full 24-hour span, with four or five cassettes each. This enabled the agency's representative to study the technical issues and the ways to overcome them with the informants. Unfortunately, many of the cassettes came back either empty or not fully recorded. In other cases, the sound of the recording person, i.e., our informant, who is closest to the microphones, was distorted. This has proved to be an especially unhappy situation, since it seems that the blame went to our representatives, who failed either to instruct or to supervise quality recordings, with the result being that our targeted informant was not recorded properly. One other crucial point was the physical connection between the microphones and the recorder, which caused distorted recording and at times even loss of some. One last problem is power supply. Two internal lithium batteries are good for some six or seven hours of recording. Still, for the informant's convenience and in order to ensure recording continuity, their replacement should have been made along with the replacement of a cassette. We used instead a battery sled assembly, which holds four C-type batteries. This power supply is good for 24 hours, so that batteries would not need to be replaced during any single recording session. This usually proved to be the case, yet informants have complained on their weight.

Fortunately, time heals in this case, and recent developments in digital recorders will enable us to use quality long-term hardware recorders with no operational complications. At this time, 6-hour sequential recording seems feasible, but when we get to the larger project, we may be able to use still better, more convenient equipment. Hardware recording with no mechanics may further lead toward some other solution regarding power supply.

(b) Ethical Issues

During a whole day our informants meet with people, with which they may have more or less meaningful conversations. The recording equipment was put into a pouch that was carried on the belt or in a bag. The microphones were attached to a device that was carried on the informant's neck, so that the microphones were located one at each respective side of the informant's head, close to the ears. The cable connecting between the microphones and the recorder was hidden beneath the informant's clothes. This way, the recording hardware would not attract any attention of either interlocutors or the surrounding people. This, indeed, proved to be the case in most instances.

The Israeli law does not prevent recording of a third party by a person who either takes part in the conversation or where it is clear that the speaking individual is aware of the attendance of that person. Whereas the recording informant signs a consent form allowing the *CoSIH* project to use the recorded data for research purposes, the other recorded people do not. Still, we are concerned with keeping the privacy not only of our informants but also of their interlocutors. Therefore, our own obligation, expressed explicitly in the consent forms as in other written forms handles to our informants, is to erase personal names and other betraying data of either the informants or their interlocutors from both the transcripts and the respective sound data.

Eliminating personal names in transcripts is an easy task, and the procedure taken is replacing the names with other names that are similar in form and in their socio-cultural setting. Being a multi-cultural nation, Israel has diverse traditions of name giving. Also, name giving to the newborn is a matter of changing fashion, and can indicate age and origin of the person carrying that name. As for name elimination in the sound files, this is a more complicated matter. One way of doing this is putting a weak beep instead of the name. However, this cause problems in understanding, especially in discourse passages that include many names. Therefore, we have devised an alternative method in which only the consonants are eliminated, so that both the syllable structure and the prosody remain intact. This method is still under examination.

Apart from this procedure, we allow informants to object to the inclusion of any part of the recording retroactively, as well as refraining from handing down to us anything they deem sensitive, or even all the recorded materials.

The last issue to be dealt with in this section is awareness to recording. This is an important matter to look at, be it on the part of the informants' interlocutors, in case they know about the recording, and especially on the part of the informants themselves. Change in speech form can take place in front of any microphone, all the more so if the recorded person knows that the goal of the research is linguistic study. We tried to overcome this latter problem by avoiding preliminary awareness of the linguistic goals of the research. When an informant is approached by our representative, s/he is being told that the goal of the research is "recording the daily life of Israeli inhabitants". Although this is not the whole truth, it is the truth, and nothing but the truth. When our representative comes to collect the recordings and before working on the sociolinguistic questionnaire, then our representative tells the informant that the recordings will be used for the compilation of *CoSIH* and requests the informant's consent to use the data.

Our impression is that in most cases speech style and language use is very similar all the way. This will, however, have to be checked in a thorough linguistic research. We had asked our informants to try tell us orally during the recording any information we could use later about the interlocutors or the setting of the recording at any new session. Only some informants kept to this procedure. We will have to double-check the wisdom of this procedure. Of course, whenever informants refer to their being recorded or recording, wherever meta-language is used to describe settings and circumstances of the recording or the recorded interlocutors, this not what we would like to see as part of the natural linguistic behavior of our informants. It is a matter to decide whether such chunks can be included as an integral part of the corpus, albeit in a separate section.

(3) Contextual Sampling (IJCL'01: §5.2)

CoSIH has been designed to be a fully representative corpus, integrating both demographic and contextual variables into a single database. Representativeness is achieved by sampling, and *CoSIH*'s design plan included a main corpus comprising 90% of the data of which both speakers' population and speech events will be selected randomly.

Obtaining a representative sample of the individuals in a group, or society, is known and commonly used. The various forms and methods of survey sampling pride a good representation of the individuals in society. For *CoSIH* this will be done by the use of a statistical sample of the Israeli population (IJCL'01: §5.2.1). However, reaching the goal of having a fully representative corpus in contextual terms too is still a vastly unexplored area (for some examples of spoken corpora aimed at representativeness not only in demographic terms but also in contextual terms see IJCL'01: §5.2.1). By 'contextual sampling' we mean sampling time and speech situations in order to get a representative sample of speech events in various environments. Thus, contextual sampling involves three issues: (i) long-term time sampling; (ii) short-term time sampling; (iii) speech sampling.

(i) Long-term time sampling

By 'long-term time sampling' we mean sampling of the recorded sets, i.e., all daylong recordings made by our informants, throughout the data collection period. Season may well influence language use, definitely in the lexical domain, but also in other domains. This is notably expected to occur in the holidays seasons. As the data-collection period is expected to last throughout a whole year (IJCL'01: 175), we expect long-term time sampling to come as a byproduct of this procedure. Eventually, we may nevertheless have a slightly imbalanced sample.

One particular problem is recordings on Saturday, the Jewish Sabbath, and on religious holidays. First, Saturdays and religious holidays are not working days in Israel. Therefore, we will have problems in asking our representatives to go and ask people to start recordings on Saturday or on a holiday. Also, a high percentage of Israeli Jews (estimated to be anywhere between 20% and 50%) would not operate a recorder on Sabbath or on a holiday because of religious constraints or out of respect to tradition. Even if we eventually find techniques to overcome these problems, we should expect under-representation of Sabbath and holiday recordings, definitely among Jews with religious or traditional restrictions.

(ii) Short-term time sampling

By 'short-term time sampling' we mean drawing a sample of 5,000-word units from each of the recruited daylong recordings. Language use change along the day, notably due to change in environment and interlocutors, but perhaps also due to other reasons, which one cannot predict at this time, like fatigue, attentiveness, and even mood.

The sampling procedure of recorded segments will be a statistically representative selection of one-hour recorded segments from each 24-span recording made by each individual informant. This will follow a procedure of elimination of long silent periods and long unintelligible speech passages (IJCL'01: §5.2.2). Hopefully, as with long-term time sampling, time distribution among the hundreds of daylong-recorded sets will produce a good sample of time within the day. However, if on a large-scale pretest (which we aim at conducting at the beginning of our data-collection year) we will see that this procedure results in imbalance, we will try the following alternative sampling procedure: We will sample time points along all raw daylong recordings to see whether they are located in the midst of a substantial speech event. In case the answer is negative, we will try another time point, until we find one that fits our demands. This or another procedure will have to be checked in the pretest.

(iii) Speech sampling

This sampling procedure can result either in hour-long speech events, or, in the majority of cases, in recorded segments shorter than an hour. While these segments may include substantial materials for inclusion in the corpus. In order to obtain our one-hour recorded segments we will need, in these cases, to make another step in order to reach this goal. This will be done by collapsing shorter speech segments into a single one by further elimination of silent periods that are too short to be eliminated in the first procedure. By using this latter procedure we will have samples of both long and short speech events, which may well represent different types of speech and language patterns. It should be born in mind that distinction is made between sampling and analytical procedures, so that the requirements from sampling, although they may converge with the requirements set for compiling the analytical unit, viz., the cell, are not the same.

(4) The Concept of 'Cell' (IJCL'01: §5.1)

As an end product, *CoSIH* will consist of "cells". A "cell" is an analytical unit. A cell is the basic sociolinguistic unit of *CoSIH*. It should aid the user to conduct research based on sociolinguistic data supplied in the *CoSIH* database, data that include both demographic features and contextual settings of the textual data included and compare it to data of other cells.

Word count is a basis upon which the size of corpora is usually defined. We kept to this tradition, and in our initial design of *CoSIH* a cell was defined as a recorded segment designated to include 5,000 words of coherent continuous text. Each cell was meant to consist of one or more texts produced by one or more speakers classified according to both demographic and contextual criteria. To illustrate this, it was said that "a cell may include a single 5,000-word text extracted from a university lecture given by a female 50-year-old native Israeli speaker of Western-European origin or two face-to-face conversations between two 20-year-old soldiers of Russian origin, one comprising 2,000 words, the other 3,000; or a cell may consist of several shorter phone conversations between a boss and employees. In all of these cases, each of the included sections will be a coherent continuous text" (IJCL'01: 190). The selection procedure of textual data to be included within a cell was to be made from the one-hour recorded segments extracted from the daylong recordings at the sampling stage.

Based on our experience gained so far, the above setting looks unachievable. In our initial design we did pay attention to speech rate and to uneven distribution of contextual setting and text types among different types of the population. However, we did not give enough thought to the fact that simple sampling procedures will not yield the 5,000-word segments to conform to our strict demographic and contextual criteria, and that we will need more complex sampling procedures. Furthermore, our aim was to have each cell consist of the speech of individual speakers about whom we can supply precise and accurate sociolinguistic data, viz., our recording/recorded informants. Linguistic materials gained from the speech of any other recorded individuals, be they interlocutors of our informants or other people, although they may be suitable for general linguistic analyses, are not good enough for lectal investigations, either linguistic or sociolinguistic.

In most sampled one-hour segments from our pilot recordings, the informant (i.e., the recording person) did not speak enough to lend us our 5,000-word cell we strived for. In order to achieve this target of having 5,000 words from a single informant, we would have to sample much longer recorded segments. Needless to say, one does not speak in empty space, and having the context of one's speech is part and parcel of any speech event. Since the corpus will eventually present the texts in both sound and transcription, we will therefore need to transcribe a lot more than originally expected. Anyone who has ever experienced natural spontaneous language transcription will know that this is not an achievable goal.

We have therefore changed our definition of cell to include segments of speech events of the same contextual category consisting of 5,000 words by all people taking part in these speech events. This definition is subject to one restriction: any cell must include at least 1,000 words in substantial speech uttered by *CoSIH*'s recording informant (or informants sharing the same demographic criteria). By 'substantial speech' we mean that the speech of the informant will not include only brief replicas with no linguistic significance. This change still fits the requirements set in the original design as regards cell capacity in terms of enabling linguistic and sociolinguistic research. As referred to in our IJCL paper (p. 194 n. 5), Biber, relying on linguistic-feature counts conducted on 1,000-word textual sub-samples of three of the early English corpora (both written and spoken), concluded that "the 2,000-word and 5,000-word texts in the standard corpora are reliable representatives of their respective text categories for analyses of this type" (Biber, 1990: 261). It may be noted at this juncture, that although spoken Hebrew is more analytic than the written medium, still the highly synthetic nature of Hebrew and its concise written structure will result in larger chunks than its English parallel by 25% or so.

There are some sets of recordings in the pilot sample that do show an intensive participation of the informant in many of the recorded speech events. From such sets we expect to have at least 2,000 word of the informant within a 5,000-word sample. In the long run, we may consider having a subcorpus of *CoSIH* with all cells consisting a minimum number of 2,000 words of their informants. If so, we will aim at achieving representativeness also in this subcorpus. Table 2 will serve to illustrate some types of speech events with varying percentage of participation of the recording informant. All recorded segments are of an identical length of 30 minutes each.

Recorded segment	Speakers	Total turns	Informant turns	informant turns %	Total words	informant words	informant words %
1	3	591	139	23.5%	4301	657	15.3%
2	4	616	121	19.6%	3625	756	20.1%
3	4	329	120	36.5%	2788	1102	39.5%
4	3	513	223	43.5%	4038	1667	41.3%

Table 2: Participation of informants in speech events

One last issue that has not been investigated yet is representativeness in terms of contextual variables. It has been mentioned above that *CoSIH* has been designed to be a fully representative corpus, integrating both demographic and contextual criteria into a single database. Contextual sampling as described above will show the distribution of speech events of varying types among the Israeli population. The collection of recordings as sampled is expected to result in deficiency in contextual categories, which will manifest itself during the procedure of allocation of texts into cells (IJCL'01: §5.3). This may lead to a decision to enhance the corpus by over-representation of some texts of varying contextual variables. This will provide better representation of contexts at he cost of being less representative of times or speakers.

Our contextual categories include three main variables and two secondary ones. The main variables are:

(a) Interpersonal relations: intimacy vs. distance
(b) Discourse structure: role driven vs. non-structured interaction
(c) Discourse topic: personal vs. impersonal

The secondary variables are:

(i) Active participants: monologue vs. dialogue
(ii) Medium: phone vs. face-to-face

A very brief survey of our pilot study already suggests that among dialogues we may expect a fair distribution of texts according to our designed main variables. However, we will probably be short of monologues. While the definition of naturally occurring monologue may be a matter for discussion, we do expect the need to over-represent monologues in some way.

As regards phone conversations, in most cases the person on the other side of the line is not heard at all. Special recording techniques may be used in some cases, e.g., with informants whose work involves many phone conversations. In other cases, a small subset of our corpus may be designed to bring forth telephone conversations.

Transcription and Annotation

CoSIH's designed size, five million words, requires some serious limitations as regards project duration, human power and financing, since five million words is a large corpus in terms of spoken corpora (Blanche-Benveniste 2000: 63). The texts will be recorded in natural settings, which means an often noisy environment and many overlaps between speakers, just to mention two of the most conspicuous problems for transcription. Existing corpora of similar size and scope are all transcribed in the standard orthography, and may include some additional notations, primarily of conversational features or intonation (e.g., Svartvik and Quirk 1980; Du Bois et al. 1992; 1993). From both our experience in the pilot study and from experience of others we note that one needs many dozens of hours to transcribe one hour of a spoken conversation recorded in a natural setting. At this point, our estimate is an average of 250 hours of transcription labor per one hour of recording. Given the above, and since *CoSIH* will present its texts to the user in both sound and transcript, we have decided to have *CoSIH* transcribed not in a phonetic transcription of any kind but in the standard orthography. Still, in order to illustrate phonetic variation of spoken Israeli Hebrew, we aim at including small samples from each cell in phonetic transcription (IPA). For some notes on transcripts in Hebrew orthography see Izre'el (2004).

The method of transcription follows in principle the one developed by Du Bois et al. (1992, 1993), in that it makes a visual representation of intonation units and includes some annotation of final tones. We are still considering the best annotation system and transcription principles for our texts. One should recall at this juncture that *CoSIH* aims at offering a synchronic presentation of sound and transcription in multimedia format. Some samples of transcribed texts have been published in Izre'el (2002). A preliminary analysis of Hebrew intonation units and final tones is presented in Izre'el (in press) and in Amir, Silber-Varod & Izre'el (2004a).

Finally, Hebrew standard orthography goes from right to left and, more prominently, does not include full and unambiguous representation of vowels. This last feature poses a serious obstacle to automatic analysis of the transcribed text. We are still contemplating the ways to overcome this problem in order to enable automatic analysis that will bring about a possible tagging service for *CoSIH*. A possible solution may eventually be found by adding a parallel pseudo-phonemic, broad transcription to the Hebrew text. A sample of a transcription of this kind can be viewed in Amir, Silber-Varod & Izre'el (2004b). This sample further includes glossing and English translation, two additional features of *CoSIH* that we have not yet given serious consideration.

A Final Word: What Is Next

The pilot phase of *CoSIH* is only the first step in a long road until our ambitious project is disseminated. Our next step is a pretest that will implement the lessons gained by the pilot study, examine their effectiveness, and study issues involved in large-scale informant recruiting and data collection. Unlike the pilot-collected data, data collected in the pretest phase will form part of the final corpus. We plan to use data from *CoSIH* Phase I, our pilot, to compile a mini-corpus on its own with its ca. 45 informants.

Acknowledgements

The pilot study was designed by the authors together with Benjamin Hary. It has been followed and supervised by the other launching members of the *CoSIH* project: Mira Ariel, John Du Bois, Yaakov Bentolila, Shmuel Bolozky, Otto Jastrow, Geoffrey Khan, Elana Shohamy and Eliezer Ben-Rafael. Vered Silber Varod has been a tremendous help in preparing the statistical data. We thank all those who, with care and enthusiasm, have taken part in this initial endeavor.

References

Amir, N., Silber-Varod, V. & Izre'el, Sh. (2004a). Characteristics of Intonation Unit Boundaries in Spontaneous Spoken Hebrew: Perception and Acoustic Correlates. In Speech Prosody 2004. [Scheduled for publication, March 2004.]

Amir, N., Silber-Varod, V. & Izre'el, Sh. (2004b). Characteristics of Intonation Unit Boundaries in Spontaneous Spoken Hebrew: Perception and Acoustic Correlates. A Sound Sample.
<http://www.tau.ac.il/humanities/semitic/sp2004.html>

Biber, D. (1990). Methodological Issues Regarding Corpus-based Analyses of Linguistic Variation. Literary and Linguistic Computing, 5, 257-269.

Blanche-Benveniste, C. (2000). Transcription de l'oral et morphologie. In M. Guille & R. Kiesler (Eds.), Romania una et diversa: Phololologische Studien für Theodor Berchem zum 65. Geburstag. Band 1: Sprachwissenschaft (pp. 61-74). Tübingen: Gunter Narr.

Du Bois, J. W., Cumming, S., Schuetze-Coburn, S. & Paolino, D. (1992). Discourse Transcription. Santa Barbara Papers in Linguistics, 4. Santa Barbara, CA: Department of Linguistics, University of California, Santa Barbara.

Du Bois, J. W., Cumming, S., Schuetze-Coburn, S. & Paolino, D. (1993). Outline of Discourse Transcription. In: J. A. Edwards & M. D. Lampert (Eds.), Talking Data: Transcription and Coding in Discourse Research (pp. 45-89). Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Izre'el, Sh. (2002). The Corpus of Spoken Israeli Hebrew (CoSIH): Textual Samples. *Leshonénu* 64, 289-314. (In Hebrew.)

Izre'el, Sh. (2004). Transcribing Spoken Israeli Hebrew: Preliminary Notes. In D. Ravid, & H. Bat-Zeev Shyldkrot (Eds.), Perspectives on Language and Language Development. Dordrecht: Kluwer. [scheduled publication: 2004].

Izre'el, Sh. (in press). From Speech to Syntax — from Theory to Transcription. In M. Bar-Asher & Ch. Cohen (Eds.), Aaron Dotan Anniversary Volume. (In Hebrew.)

Izre'el, Sh., Hary B. & Rahav, G. (2001). Designing CoSIH: The Corpus of Spoken Israeli Hebrew. International Journal of Corpus Linguistics, 6, 171-197.

Svartvik, J. & Quirk, R. (1980). *A Corpus of English Conversation.* Lund: Lund University Press.

Websites

BNC. The British National Corpus: The Spoken Component. <http://www.natcorp.ox.ac.uk/what/spok_design.html>

CoSIH: The Corpus of Spoken Israeli Hebrew. <http://www.tau.ac.il/humanities/semitic/cosih.html>

Gateway for Corpus Linguistics on the Internet. <http://www.corpus-linguistics.de/corpora/corp_spoken.html>

SFB 411. <http://www.sfb441.uni-tuebingen.de/c1/corpora-engl.html>

The ICSI Meeting Corpus: Close-talking and Far-field, Multi-channel Transcriptions for Speech and Language Researchers

Jane A. Edwards
International Computer Science Institute, and
Institute of Cognitive Studies, UC Berkeley
edwards@ICSI.Berkeley.EDU

Abstract

The recently-completed ICSI Meeting Corpus is available through the LDC. It consists of audio and transcripts of 75 research meetings, ranging in size from 3 to 10 people, with an average of 6 people. The meetings were recorded by means of both close-talking (headset or lapel) microphones and far-field (table-top) microphones. The close-talking microphones enable separation of each person's audible activities from those of every other participant. The far-field microphones provide a view of the meeting as a whole. The transcripts preserve words and other communicative phenomena, displayed in musical score format, time-synchronized to the digitized audio recordings. The corpus is intended as a resource for both speech researchers and language researchers. This paper describes the methods used to prepare the corpus, some interesting challenges and solutions, and the benefits of using both close-talking and far-field microphones.

1. Introduction

The "ICSI Meeting Corpus" was recently completed and is now available through the Linguistics Data Consortium (LDC). It consists of audio recordings and transcripts of 75 naturally-occurring research meetings. The goal was to produce a high-quality resource for use by both speech researchers and language researchers.

All meeting participants were recorded both by close-talking microphones (usually a headset), and by far-field microphones (arranged along the tabletop). The close-talking microphones enable separation of each person's audible activities from those of the other participants. The far-field microphones provide a view of the meeting as a whole. The resulting audio recordings fill 9 DVD's.

The meetings were "natural" (not contrived): they would have occurred regardless of whether or not they were recorded. The meetings recorded were, in large part, regular weekly meetings of ICSI research groups (5 main groups), and usually lasted about an hour.

They ranged in size from 3 to 10 participants (with average size of 6). This is much larger than most multi-party interactions recorded in other corpora. The corpus contains 5 main types of meetings and 53 unique speakers.

The meetings differed in the degree to which they followed an agenda, and the degree to which power was centralized or distributed evenly among the participants. The participants knew each other well for the most part, and cared about the matters under discussion. This led to some degree of overlapping speech, from very little to very much, depending on the group.

Standard procedures were observed in terms of Human Subjects requirements for informed consent. The Consent Form asked participants for permission to use their data in the corpus, and let them know they would have access to the transcripts and audio prior to public release of the data, and that things would be excised from their speech in meetings if they requested such excisions.

This paper describes some of the methods used in preparing the ICSI Meeting Corpus. (For information on other aspects of the corpus, please see Morgan, et al., 2001 and 2003; Janin, et al., 2003). ICSI's is the first meetings corpus to be released with audio and transcripts for public use. It is important to mention in passing that corpora of meetings are also being prepared by CMU and NIST, among others. This is an active interest area and certain to become more so in the future.

2. The main goals of transcription for the ICSI Meeting Corpus

Even if ongoing international efforts toward increasing standardization of data encoding methods at least in their broad outlines (e.g., TEI, EAGLES, CES, MATE), it is still the case that projects are necessarily unique in certain ways, dictated by their specific goals, and intended audience.

This corpus was designed for use by two distinct research communities: speech recognition researchers on the one hand and language researchers (linguistics and discourse researchers) on the other hand.

The main goal was to produce a word-level transcript of each speaker's channel, time synchronized to the digitized audio recording. Non-word events were also captured, and comments were added concerning aspects which might be relevant to either speech recognition (e.g., voice quality or non-canonical pronunciation), or discourse research (e.g., situational comments). The transcription conventions were chosen to be as theory-neutral and as minimalistic as possible. Among other things, there was no attempt made to capture "short" vs. "long" pauses. Pauses could be noted if they stood out to a transcriber, but perceived pause length, which is found in virtually all discourse corpora, was beyond the scope of this project. Prominence was treated in a similar manner, as was intonation. Several annotation efforts are already underway, but the most immediate goal was to provide the most accurate basic transcript possible, to be embellished later as needed.

The use of individual microphones for everyone at the meeting was invaluable in disentangling the many overlaps which occurred during the meeting. In addition, it made it possible to capture such things as whispered comments, very quiet laughs, and the sudden inbreaths which occur prior to attempting to gain the floor -- most of which would be impossible with less sensitive and/or shared microphones.

In some cases, every inhale and exhale could be heard on a particular channel. But such breathing patterns were not preserved in the transcript due to being predictable and uninformative. In contrast breathing patterns which were potentially communicative were preserved (e.g., the sudden outbreath of frustration, a sudden inbreath of surprise, or a yawn). Although English is full of words such as “sigh,” “wheeze”, and “gasp”, these were usually not appropriate, because they seemed either overly negative or inappropriately dramatic. Instead, more neutral descriptions were used.

3. Multi-channel audio and visual representation

If an interaction has only a couple of participants, many transcription methods are viable (as discussed in Edwards, 2002). In the ICSI Meeting Corpus, however, overlaps could include as many as 10 people (if everyone laughed at a joke) and 3- or 4-way overlaps were not uncommon. In such cases, musical score notation has obvious natural advantages over other transcription methods (e.g., Edwards, 1992; Ehlich, 1993), since it enables simultaneous or partially overlapping events to be displayed one above the other, with reference to a common time line.

The musical score notation for this corpus needed furthermore to be time-sychronized with the audio recording for each speaker. This was accomplished by use of a computer interface called "Channeltrans" (www.icsi.berkeley.edu/Speech/mr/channeltrans.html). It is an extension of the "Transcriber" interface (Barras, Geoffrois, Wu, and Liberman, 2000). Both are available free of charge.

Both Transcriber and Channeltras preserve events and the time bins in which they occurred, and both of them do so in XML format. Channeltrans differs from Transcriber in that it preserves the channel number in addition to the time and event. That is, unlike Transcriber, which has only one display ribbon for speech, Channeltrans has as many display ribbons as there are participants. In addition, Channeltrans allows the time bins on each ribbon to be totally independent of those on all other ribbons. Both properties -- multiple ribbons and independent time segmentation -- were essential for the Meeting Corpus data, due to the large number of participants and great amount of overlapping speech in these meetings.

The basic strategy used in transcribing the data was to view each display ribbon as capturing the actions of a particular meeting participant, heard over the close-talking microphone which he or she was wearing (i.e., the dominant speaker on that channel). In cases of crosstalk, other speakers might be heard on the same channel, but only the events produced by the dominant speaker were transcribed on that speaker's ribbon. That is, even if an utterance could be heard on several channels, it was transcribed only on one channel, i.e., the channel corresponding to the person who spoke that utterance.

4. Some time-saving strategies in first-pass transcripts

The basic task of transcribing the data involved identifying the boundaries of an event (e.g., utterance, noise, happenstance) and transcribing the nature of the event itself.

Because the meetings often had so many participants, it was impractical to accomplish the time bin segmentation in a strictly manual way (i.e., having transcribers do all the segmentation into time bins). For an hour meeting with ten participants, for example, it would have required ten hours to listen to each channel exhaustively to find the time bins which required encoding. Viewing the energy waveform for activity might seem an effective solution, but in fact, it was not very reliable.

A highly effective approach turned out to be to apply a speech-nonspeech detector to the audio recordings to generate a preliminary segmentation into time bins to be adjusted later by human transcribers. (For details see Pfau, Ellis & Stolcke, 2001).

Undergraduate transcribers were encouraged to correct, adjust, or add new segmentations as needed. The time bins were intended simply as units of a manageable size with clean breaks on either side (i.e., no truncated words). Utterances might be contained in a single time bin or they might extend across several time bins. The time bins were not to intended as definable discourse units or prosodic units but simply as manageable units, which could be made more precise later if needed.

The project also used professional transcription agencies for some first-past transcripts. The presegmented versions were processed in such a way that all of the time bins which the presegmenter identified as containing events were strung together in a linear fashion, and recorded onto a cassette together with sequence numbers to prevent duplication or omissions of segments. The professional transcribers then transcribed each time bin chunk, together with its sequence number, and the resulting chunk-wise transcript was re-assembled at ICSI and double checked by the student transcribers.

5. Checking the transcripts for word-level accuracy

After a transcript was completed, it was submitted to a spell-checker, and then reviewed in its entirety while the checker listened to the audio recording. After this was completed, the process was repeated by one of two senior researchers.

Even though the data had by this time been seen by at least two and often three pairs of carefully trained eyes, these "read-throughs" by the senior researchers led to a number of corrections at the word and utterance level. This reflects two aspects of the meetings: the highly technical nature of the discussions, and the fact that many meeting participants were non-native speakers of English. The senior researchers have technical backgrounds which gave them an advantage over both the linguistically-trained student checkers and the professional transcribers. This experience is no doubt familiar to anyone who has ever prepared a transcript, and is a clear reminder of the extent to which a linguistic message is underdetermined by its acoustics and of the importance of context, intonation, pragmatic conventions and world knowledge in filling in the gaps.

In about twenty cases, there were words or phrases which seemed acoustically very clear but remained incomprehensible even to senior researchers. In these cases, the actual speakers were asked to listen and demystify them. Here are two examples:

> (1) .. now that of course we have sort of started to lick blood with this, {QUAL editor's note, speaker explained that "lick blood" is a German idiom meaning "having started something, and wanting more of it"}
>
> (2) From Michael Strube, I've heard very good stuff about the chunk parser that is done by FORWISS, {QUAL editor's note, speaker-verified names}

Once checking was completed by a senior researcher, the transcripts were made available for correction by the participants themselves. Very few errors were detected in this way.

6. Benefits of using both Close-talking and Far-Field Microphones

On the surface it would seem that close-talking microphones alone would be sufficient in that they provide a sensitive record of each person’s speech. However, there are several situations in which the far-field microphones are invaluable.

a) Compensating for some glitches on the close-talking recording

When a word was unclear or even the speaker’s identity who spoke it, the far-field microphones often provided a sufficiently different “ear” on the situation to be able to clarify it.

b) Tracking discourse when some participants are out of the room

In some meetings, a participant left the room while still wearing the microphone (e.g., to photocopy something for the meeting, or to arrange something with the administrative staff). If a transcript had been based only on the close-talking microphone, the outcome of these intermingled conversations would have been confusing or even bizarre. The situation became immediately clear when listening to the far-field microphone.

The next two are somewhat more subtle and will require more sophisticated approaches to use the far-field data, but are possible in principle, and not far from what is being done already.

c) Distinguishing self-oriented subvocalizations from shared communicative behavior

Where should the researcher draw the line between that which is probably audible only on the close-talking microphone and that which was probably heard by others? This is almost a Heisenberg uncertainty problem. With poorer quality recordings in the past, the researcher could be sure that if he or she heard something, the other people in the interaction no doubt heard it too. But the Meeting Corpus includes some degree of "over-precision", that is, vocalizations which speakers may make for their own purposes, with no intention that they be part of the meeting as a whole. For example, there are cases in which a speaker says a word or two to him- or herself at low volume. And there is even a case in which a particular speaker said "uh-huh" an unexpectedly large number of times, but at such a low volume that it was inaudible to other participants at the meeting. If a backchannel occurs in a meeting and no one else hears it, is it still communicative?

The high quality far-field microphones provide the raw data which can in principle be used to estimate what the others may have heard.

d) Determining the best mix of channels to represent the meeting as a whole

This is an extension of the previous problem. During the calibration of equipment at the beginning of the meeting, the technician often boosted the recording volume of close-talking microphones for "soft-talkers" relative to people who normally speak more loudly. When these channels are simply combined, the person with the soft voice will be louder than he or she was in the actual meeting relative to the other participants.

In contrast, the high quality far-field microphones record multiple participants without adjustments to the loudness of the participants individually, and when combined, can give a better approximation of the relative loudness of speakers at the meeting. Some work would be needed to match up the levels of the different tabletop microphones, but this is possible in principle. Without the far-field microphones, the problem would be data-limited and therefore unsolvable.

7. Conclusions

This paper has discussed the general structure of the Meeting Corpus, and some of the procedures it developed in meeting the challenges of transcribing 75 actual (rather than contrived) meetings in musical score format, time-synchronized to digitized audio recordings.

It also briefly described a minimalist approach to transcription which was chosen to serve the needs of two very different research communities: speech and language research.

Finally, it discussed the benefits of having both close-talking and far-field microphones, mentioning several types of problems which would be insurmountable without the use of both types of microphones.

8. Acknowledgments

Data gathering and preparation of the ICSI Meeting Corpus was made possible by the following sponsors: DARPA (through Communicator and a “ROAR” seedling), IBM, and the Swiss National Science Foundation (through NCCR’s IM2 project). In addition, I wish to thank Nelson Morgan, Barbara Peskin, and the LREC reviewers for their helpful comments on an earlier draft of this manuscript.

9. References

Barras, C., Geoffrois, E., Wu, Z. and Liberman, M. (2000). Transcriber: development and use of a tool for assisting speech corpora production. *Speech Communication* special issue on Speech Annotation and Corpus Tools, Vol. **33**, No 1-2.

Edwards, Jane A. (1992). Design Principles for the Transcription of Spoken Discourse. In J. Svartvik (Ed.) *Directions in Corpus Linguistics: Proceedings of the Nobel Symposium 82, Stockholm, August 4-8, 1991* (pp. 129--147). NY: Mouton de Gruyter.

Edwards, Jane A. (2002) The Transcription of Discourse. In D. Tannen, D. Schiffrin, and H. Hamilton (eds). *The Handbook of Discourse Analysis*. NY: Blackwell (pp. 321-348).

Ehlich, K. (1993) HIAT: A Transcription System for Discourse Data. In J. A. Edwards & M. D. Lampert (Eds.) *Talking data: Transcription and coding in discourse research.* (pp. 123-148). Lawrence Erlbaum Associates, Inc; Hillsdale, NJ.

Janin, A., Baron, D., Edwards, J., Ellis, D., Gelbart, D., Morgan, N., Peskin, B., Pfau, T., Shriberg, E., Stolcke, A. & Wooters, C. (2003). The ICSI Meeting Corpus. *ICASSP-2003*, Hong Kong, April 2003. <ftp://ftp.icsi.berkeley.edu/pub/speech/papers/icassp03-janin.pdf>.

Morgan, N., Baron, D., Edwards, J., Ellis, D., Gelbart, D., Janin, A., Pfau, T., Shriberg, E., and Stolcke, A. (2001) The Meeting Project at ICSI. *Human Language Technologies Conference*, San Diego, March 2001

Morgan, N., Baron, D., Bhagat, S., Carvey, H., Dhillon, R., Edwards, J., Gelbart, D., Janin, A., Krupski, A., Peskin, B., Pfau, T., Shriberg, E., Stolcke, A., and Wooters C. Meetings About Meetings: Research at ICSI on Speech in Multiparty Conversations. ICASSP-2003, Hong Kong, April 2003. <ftp://ftp.icsi.berkeley.edu/pub/speech/papers/icassp03meetings.pdf>

Pfau, T. Ellis, D. P. W. & Stolcke, A. (2001), Multispeaker Speech Activity Detection for the ICSI Meeting Recorder. *Proc. IEEE Automatic Speech Recognition and Understanding Workshop*, Madonna di Campiglio, Italy.

Orality and Difficulties in the Transcription of Spoken Corpora

González Ledesma, Ana; De la Madrid Heitzmann, Guillermo; Alcántara Plá, Manuel; De la Torre Cuesta, Raúl & Moreno Sandoval, Antonio

Laboratorio de Lingüística Informática
Universidad Autónoma de Madrid
e-mail address{ana; guille; manuel; raul; sandoval}@lllf.maria.uam.es

Abstract

This paper analyses the effects of certain oral features on the process of transcription of spontaneous speech recordings. On the basis of the statistical analysis of the data obtained from the C-ORAL-ROM corpus, it will be shown empirically that transcription difficulties vary according to the communicative situation, the degree of formality and the number of participants.

1. Introduction

This paper is the result of an experiment carried out by a group of transcribers at the Laboratorio de Lingüística Informática (LLI) at the Universidad Autónoma de Madrid, once the recording and transcribing phases of the C-ORAL-ROM project were over.

The goal of the experiment was to confirm certain hypothesis which had arisen after the transcription process as the team attempted to determine which communicative interactions caused more difficulties in the transcription phase and why.

The original hypothesis consisted in relating these transcription problems to the frequency of occurrence of two kinds of linguistic phenomena typical of spoken interactions:

- **Production features,** such as *fragmented words, supports, retractings,* etc.
- **Interaction features**, such as the *number of turns* or the *overlapping* (Llisterri, 1997).

This would lead to the conclusion that the more frequent these phenomena were in a spoken interaction, the more time and effort needed by the linguist in the process of transcription.

However, it was the team's aim to rationalize these impressions and confirm the causes in an empirical way. Thus, the following objectives were stated:

- Definition of orality.
- Development of a computational tool which could help to establish a relation between conversational genres and orality features.
- Showing how these features vary inside the corpus depending of the register.
- Data analysis and verification of how orality is related to the difficulties present in the transcription process.
- Establishing a typology of transcription problems based on the results of the analysis.

However, before attempting further explanation of the experiment and analysis of the results, it is necessary to discuss some features of the corpus used, focusing on those which are related to its design and distribution.

2. Description of the corpus

C-ORAL-ROM is a multilingual spontaneous speech corpus (Cresti et al., 2002) of the four main roman languages: French, Italian, Portuguese and Spanish. Each subcorpus consists of around 300,000 words. With the aim of enabling comparability between the different subcorpora, several sampling criteria concerning the distribution of the corpus were established: as long as each of the variation parameters is fully present in the corpus, the linguistic variation will be well represented (Moreno, 2002). In that sense, two elements are to be considered as basic elements: on one hand, the *characteristics of the speakers* and, on the other hand, the *context of use*. As far as the speakers are concerned, age, sex, education, occupation and geographical origin were taken into account. As for the contexts of use, a basic distinction was made between the dialogic structure (monologues and dialogues or conversations) and kind of situation (familiar or public).

A second important distinction was made between formal and informal speech. Each is represented in the corpus by 50% of the texts. Inside the informal part, a distinction was made between the familiar and the public domains: the first is represented by 75% of the texts, while the public domain accounts for the other 25%. As for the formal speech, the distribution of the texts was made following a thematic criterion: the *natural context formal speech* area (43% of the texts) is formed by recordings such as conferences, political debates, political speeches, sermons, professional explanations and texts dealing with business, law and teaching. In the same way, the texts which are part of the *formal speech in media* section (40% of the texts) are grouped in the following categories: interviews, meteo, news, reportages, scientific press, sports and talk-shows. Finally, inside the formal speech part, a section made up of phone recordings (17 % of the texts) is included.

Other relevant criteria concerning the corpus design are: acoustic quality of the samples (all are digital recordings), legal status (recording, transcription and publishing were done after the written authorization of all participants) and spontaneity of the recordings (no previous scripts were used and there were no restrictions in the use of the language and the expression of opinions).

<table>
<tr><td rowspan="4"><u>*Informal*</u></td><td rowspan="2">*Familiar/Private*</td><td>*Monologue*</td></tr>
<tr><td rowspan="2">*Dialogue*</td></tr>
<tr><td rowspan="2">*Public*</td></tr>
<tr><td>*Conversation*</td></tr>
</table>

Formal		
Formal in natural context	*Media*	*Telephone*
political speech	news	private conversation
political debate	sport	phone call services (man interaction)
preaching	interviews	phone call services (machine interaction)
teaching	meteo	
Professional explanation	scientific press	
conference	reportage	
business	talk shows political debate	

Figure 1: Distribution of the C-ORAL-ROM corpus

3. The notion of orality

It is well known that **spoken language** is not always a synonym to **orality**, if we understand orality as the presence of linguistic, paralinguistic and interactive phenomena, such as *retracting* or *overlapping,* which are not present in the written register. The registers in spoken language vary depending on the communicative situation. For instance, a text being a transcription of a sermon will differ significantly from a private conversation between friends, as far as the subject, the communicative context, the goals and the relation between participants are concerned (Romaine, 1996).

These differences are present not only at a morpho-syntactic, lexical and discoursive levels, but also at a more basic level which has to do with discourse production and which we will refer to as **degree of orality**.

The goal of this paper is to study that degree of orality considering the different conversational genres established in C-ORAL-ROM, in such a way that the hypothesis stated at the beginning -the presence of certain spoken features makes the transcription process much more difficult- can be confirmed.

Those phenomena chosen as the object of this experiment are typical features of spontaneous speech: overlapping, retractings, dialogic turns, speaking speed, fragmented words ("psicolog" instead of "psicología", for example) or supports, coded in C-ORAL-ROM as *&ah* and *&eh*.

4. Orality and transcription problems: the original hypothesis

In order to find out what kind of relation there is between orality and linguistic registers, two **scales of transcription difficulty** were stated taking into consideration the following two parameters:

4.1. Degree of formality (Scale 1)

Two ends can be considered when dealing with the texts in terms of transcription difficulty: on one end, the most complex, those texts distinguished as *private*; on the other end, the easiest, those texts classified as *formal*.

```
               informal      media    formal
+ difficult --------------------------------- - difficult
```

4.2. Number of speakers (Scale 2)

This parameter affects only those texts classified as *informal* (the most complex according to Scale 1) and considers as most complex those texts with a higher number of participants (three or more), while those with one or two participant imply a lower degree of difficulty.

```
               conversation     dialog      monolog
+ difficult -----------------------------------------    - difficult
```

5. The computational tool

The C-ORAL-ROM corpus is tagged with XML. Using the information included in the tags, we developed a program which automatically calculate the frequency of occurrence of each of the following features: overlapping, retracting, number of dialogic turns, speaking speed, fragmented words and supports. These frequencies were calculated for each class of texts.

Thus, the results show the average number of words between two occurrences of a phenomenon, except in the case of *speaking speed*, where the figures correspond to the number of words per second. The higher the number of words, the less important is the phenomenon in the class of text in question. In order to facilitate the reading of the figures, only one decimal was used in the final results.

6. Textual typology and transcription problems: analysis of the data

In this section, the results obtained by the program are analyzed. The analysis procedure has always been the same for each of the linguistic phenomena studied:

First, the relation between frequency of occurrence of the features and textual typology is stated.

Second, we evaluated whether this relation confirms the original hypothesis, which states that certain kinds of texts are harder to transcribe, according to the scales of difficulty.

6.1. Number of dialogic turns

The first feature to be analyzed is the **number of dialogic turns**, understood as the number of times a speaker replaces another in the conversation. According to the original hypothesis, in the analysis of the data it is assumed that there is a direct relation between the number of turns and the effort needed in the transcription process.

Below, it is shown how this feature is reflected in the mentioned classifications from a quantitative point of view.

In *Figure 1*, which analyses the **degree of formality (scale 1)**, it can be observed how the participants in informal texts produce shorter turns, while those turns belonging to formal texts are longer and those turns produced in media texts have an intermediate length, closer to formal texts than to informal ones.

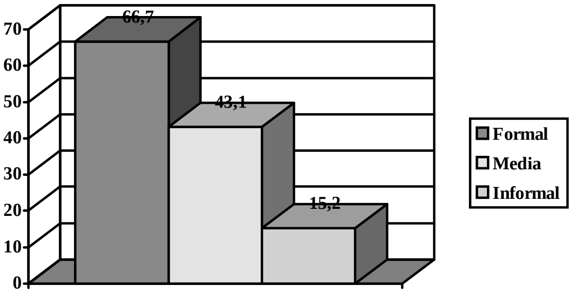


Figure 1: Words per turn in Scale 1.

In the groups dealing with **number of speakers (Scale 2)**, apart from the obvious conclusion about monologues, those turns belonging to dialogues are almost two and a half words longer than those belonging to conversations.

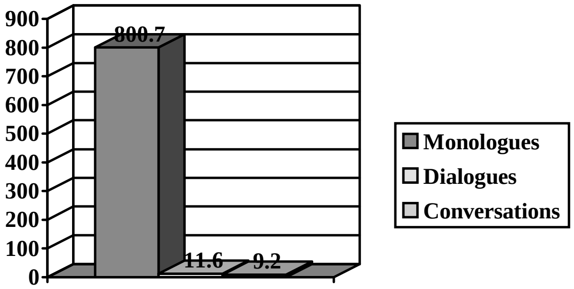


Figure 2: Words per turn in Scale 2

These results confirm the original hypothesis, that is to say: the higher the number of speakers, the shorter are the turns (considering the number of words per turn) and therefore the bigger is the effort necessary in the transcription. Furthermore, it has been proved that shorter turns are typical of informal texts and so it is in this area of the corpus where the transcriber will find more difficulties.

6.2. Overlapping

This second feature is directly related to the previous one and represents, according to C-ORAL-ROM transcribers, one of the most important difficulties in the transcription task: **overlapping**. Again, the results are obtained dividing the number of words by the number of overlapping cases (except in monologues, where there is obviously no overlapping):

Figure 3 is the confirmation of *Figure 1*. As expected, overlapping is less frequent in the formal and media genres than in the informal one. In the informal genre, as shown in *Figure 4*, the difference between dialogues and conversations is an average of almost ten words. These data prove that overlapping is prototypical of the informal genre and, furthermore, of the conversation subgenre. As far as the transcription task is concerned, this fact puts the conversation subgenre on the furthest end in terms of transcription difficulty.

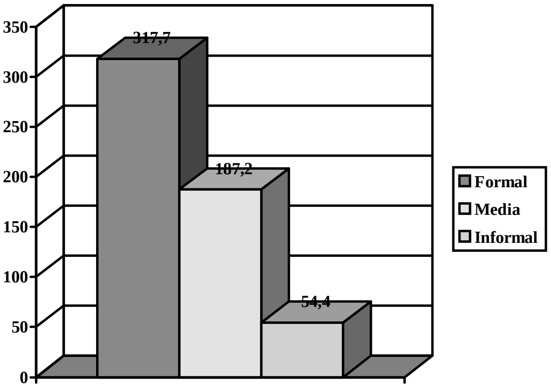


Figure 3: Wods per overlapping in Scale 1

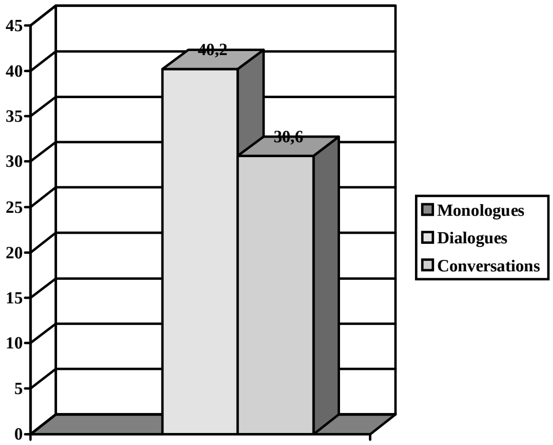


Figure 4: Words per overlapping in Scale 2.

6.3. Speaking speed

Another important feature for the transcriber is the speed at which the participants speak. These are the results obtained for C-ORAL-ROM:

This feature, expressed in words per second, confirms once more how, in terms of speaking speed and given that the faster a participant speaks the harder the is to transcribe, the informal genre and the conversational subgenre are the most laborious in the transcription task. As we can see in the figures, a prototypical participant in an informal conversation utters approximately three and a half words per second, while for formal texts and informal monologues the average is 2.6-2.7 words per second.

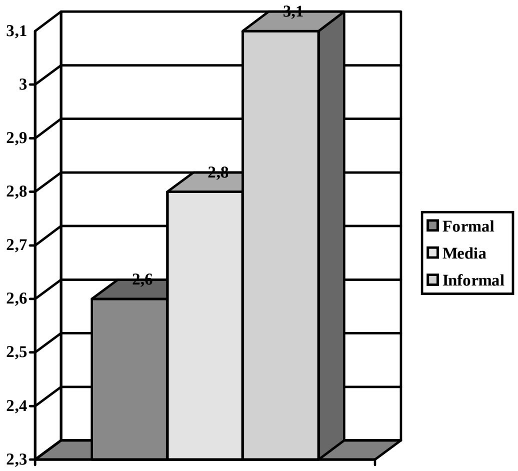


Figure 5: Words per second in Scale 1

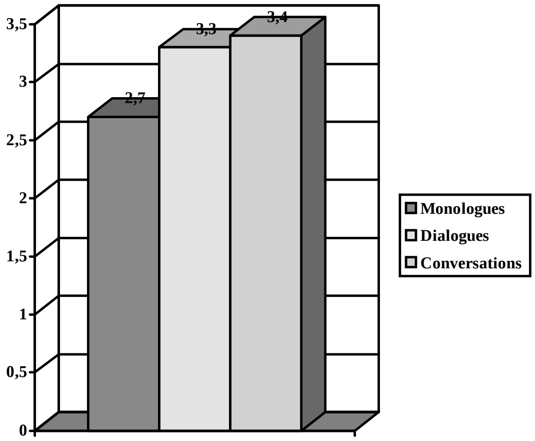


Figure 6: Words per second in Scale 2

6.4. Fragmented words.

So far, the data has confirmed the original hypothesis. However, in regards to the frequency of fragmented words, the hypothesis was not supported. A fragmented word occurs when a speaker does not complete the utterance of the word.

In *Figure 7,* it becomes obvious that most participants in the media genre are speaking *professionals.* Even though they speak at a higher speed than those appearing in formal texts (*Figure* 5), the frequency of occurrence of fragmented words in this kind of texts is much lower than it is in other genres, which share almost the same ratio. On the other hand, the formal genre is characterized by a high number of fragmented words.

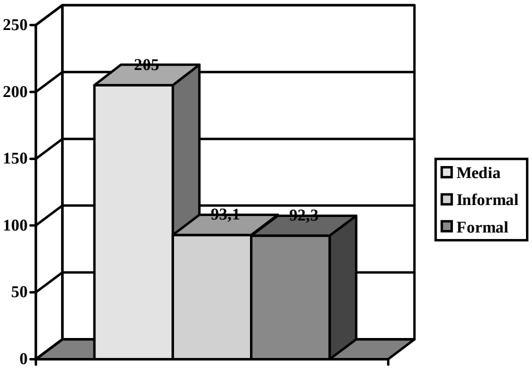


Figure 7: Words per fragment in Scale 1

Also, unexpectedly, *Figure 8* shows that the number of fragmented words is higher in dialogues than it is in monologues and conversations. The fact that dialogues are not in an intermediate position (as it happens in the rest of the results) leads to the conclusion that there is not a direct relation between number of participants and frequency of occurrence of fragmented words, an hypothesis that should be confirmed with further data.

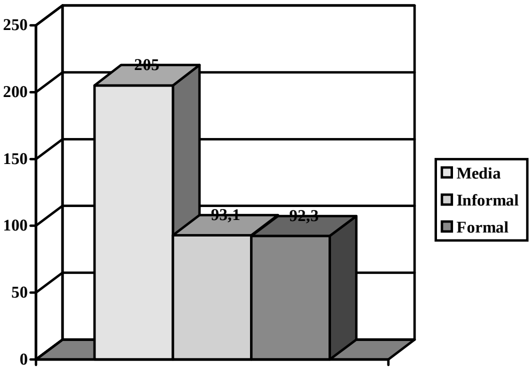


Figure 7: Words per fragment in Scale 1

Also, unexpectedly, *Figure 8* shows that the number of fragmented words is higher in dialogues than it is in monologues and conversations. The fact that dialogues are not in an intermediate position (as it happens in the rest of the results) leads to the conclusion that there is not a direct relation between number of participants and frequency of occurrence of fragmented words, an hypothesis that should be confirmed with further data.

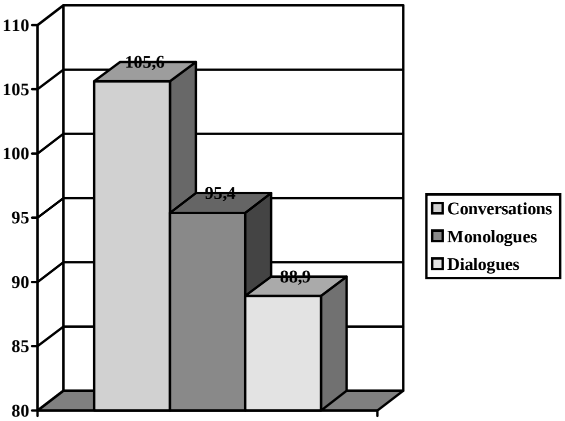


Figure 8: Words between fragments in Scale 2.

All this would show how, in the transcription process, fragmented words are not perceived by the transcriber as an added difficulty, given that, in the difficulty scale (*Scale 1*), the formal genre is the easiest to transcribe.

6.5. Supports

The following analysis corresponds to the frequency of occurrence of **supports.**

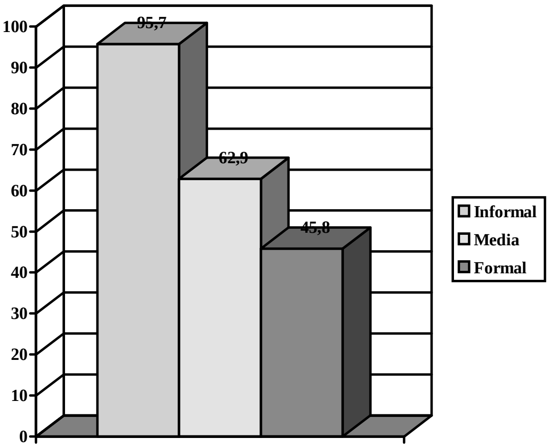


Figure 9: Words between supports in Scale 1

Figure 9 shows an opposite arrangement to the difficulty order originally suggested, where formal texts were classified as the easiest ones. Similarly to the case of fragmented words, the prototypical participant in a formal recording resorts to supports every 46 words, in order to sustain his discourse. This contrasts with the ratio in informal texts, where the average is of almost 96 words.

The information presented in this figure is quite unexpected, especially when, in this sense, the formal genre is made up of communicative interactions such as conferences or lessons in an academic context, which are quite close to texts following some kind of script. This helps to understand this phenomenon not as a symptom of lack of planning, but as a support which participants in this kind of recordings find useful or necessary.

These data could somehow be connected to the number of participants, that is to say, the more the speakers in a conversation, the less supports are used, due to the dynamics of the interaction. In order to confirm this hypothesis, we can look at *Figure 10*, where the number of participants is one of the parameters.

Nevertheless, this table shows how conversations represent the texts with the lowest frequency of occurrence of supports. In fact, the data prove that, as the number of participants increases (and, if we look at *Figure 2*, the turn is longer), so do the frequencies of supports.

Therefore, this is again a revealing finding. First of all, for the description of the different linguistic registers of the spoken corpus and further studies on this field. Secondly, it is also important for the transcribers, as supports do not seem to constitute an added difficulty in the transcription process.

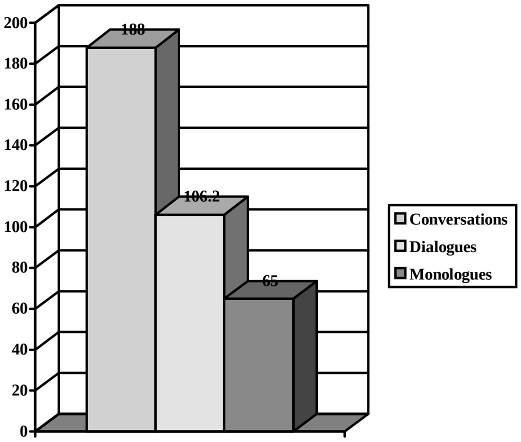


Figure 10: Words between support in Scale 2.

6.6. Retractings

Finally, figures relating the frequency of occurrence of **retractings** were analyzed in the six groups, obtaining the following results.

Again, the scale of difficulty is inverted. Even though the frequency of occurrence of retractings increases in the informal texts, and this matches the predictions made, it is interesting to observe how the formal and media genres invert their positions with respect to the figures. This leads to important conclusions regarding the differences between these two genres, which are in principle quite similar, given that both are planned and are characterized by a register which is close to written language.

As for the results in *Figure 12*, retracting is a characteristic phenomenon of informal monologues, which again raises the question of the motive behind this phenomenon. It is interesting to remark that this feature presents the highest frequencies in a kind of text where there is no interaction at all between participants.

Regarding the transcription problems, again the original hypothesis is inverted. Contrary to what was expected, there is not a direct relation between the presence of retractings and the degree of difficulty in the transcription process, as the *informal monologue* is the

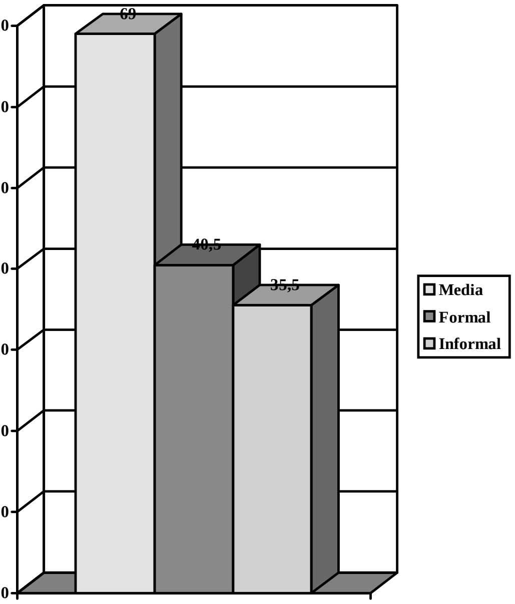


Figure 11: Words between restarts in Scale 1

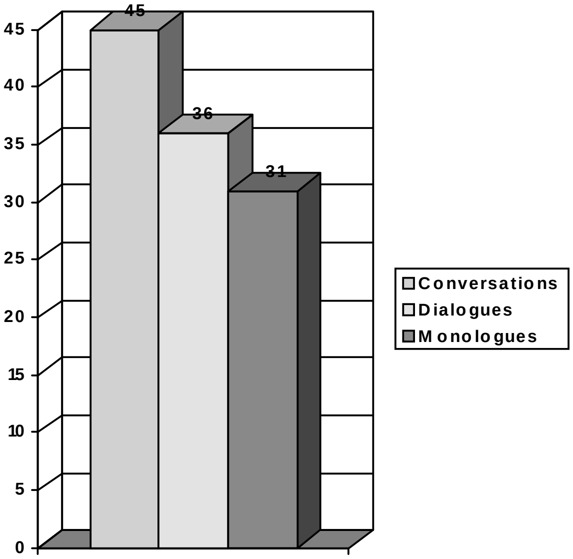


Figure 12: Words between restarts in Scale 2

last in the scale of difficulty (Scale 2) introduced in section 3.

7. Conclusions.

Sometimes, the obvious facts has to be proven in order to question its value of truth, and this is exactly what has been accomplished in this paper. Starting from a apparently natural hypothesis, which consists in relating the presence of certain spoken features to a special difficulty in the transcription process, it has been deduced from the analysis of the results that this hypothesis is not always true because there are some spoken features such as supports, whose frequency of occurrence is higher in those texts which, as it is the case with formal texts and informal monologues, are not an added obstacle for the transcriber.

All this would lead to the classification of the features of the corpus into two groups:

- **Interactional features:** number of words per turn, frequency of overlapping and speaking speed (*Figures 1-6*).
- **Production features:** frequency of occurrence of fragmented words, supports and retractings (*Figures 7-12*).

The distribution of the types of text in the case of group 1 matches exactly the intuitive difficulty scales presented as *scales 1 and 2*.

However, the cases in group 2 (production features) show a distribution which is even opposite to scales 1 and 2 in some of its aspects: *media* is the genre with less influence coming from fragmented words and retractings, and the *formal* genre is the one with a highest ratio of fragmented words and supports (this last feature shows its lowest ratio in *informal* texts).

As for the second scale (informal texts), observing the second group of features (production features), *conversations* appear as the less affected subgenre, while *monologues* stand out as the richest in supports and retractings. The scales, if only the production features were taken into account, would be as the following, from less to most difficult:

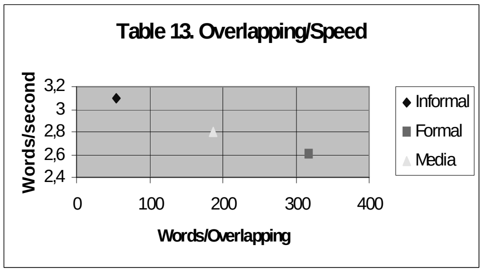
- difficult ----------------------- + difficult
media informal formal

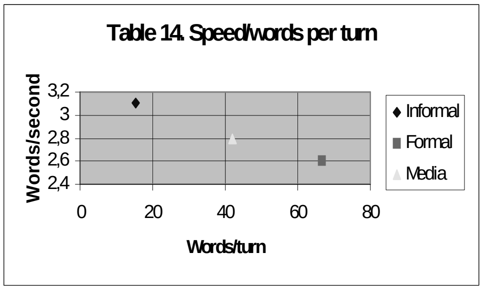
Figure 2: Scale 1 and production features.

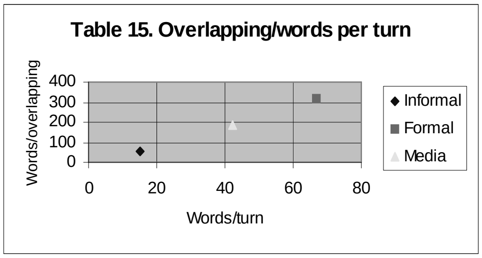
- difficult ----------------------- + difficult
conversations dialogues monologues

Figure 3: Scale 2 and production features.

The fact that *interactional features* match exactly the intuitions made at the beginning and that production features are almost the opposite (the highest difficulty end corresponds to formal texts and informal monologues) gives way to the conclusion that the difficulty perceived by the transcriber comes from the features in the first group (interactional). This is shown below in the correlation of the data for interactional features in *Scale 1* (that is, the data on figures 1 to 6 for informal, formal and media). It is clearly shown on the figures how the order informal-media-formal is kept at all the times. The three figures correspond to relations between **speed and overlapping** (*Figure 13*), **speed and words per turn** (*Figure 14*) and **overlapping and words per turn** (*Figure 15*).







Those features belonging to the second group (*production features*) are problematic in the *establishment of the text* (Benveniste, 1998) phase, as it has to be decided what is going to be the written representation for that kind of recording; however, the transcriber does not consider these features as obstacles in the transcription process.

These empirical conclusions should be confirmed by applying the same analysis on new texts, as it will be the case with a previously created corpus in LLI-UAM: CORLEC (Moreno, 2002). CORLEC does not follow exactly the same transcription criteria as C-ORAL-ROM (mainly because it was recorded and transcribed 10 years before), but it has the advantage of being three times larger in terms of the number of words.

Nonetheless, the main conclusion (the complexity of a transcription derives from the interaction features and not from the production features) is fully justified by the representative character of the data used as a basis. More specifically, there were 429 different speakers, some of them participating in different recordings, which means a total of 554 participants. The number of texts is not very high (169 recordings) but its great variety should be highlighted (as shown in the distribution chapter). Finally, below is a summary of the data used:

Total data in C-ORAL-ROM (general data)				
Texts	Speakers	Participants	Turns	Words
169	429	554	15595	312597

Total data in C-ORAL-ROM (features)			
Overlapping	Retractings	Fragmented words	Supports
4307	7860	3084	4807

Table 16: Absolute data for C-ORAL-ROM.

Further investigation applying this methodology, extended to other criteria, might include characterizing the different spoken registers included in C-ORAL-ROM at all the linguistic levels. Besides, an optimum result of applying this methodology would lead to a prediction of the typology of a given spoken text, based on quantitative data not in qualitative ones.

8. References

Blanche-Benveniste, C. (1998). Estudios lingüísticos sobre la relación entre oralidad y escritura. Barcelona, Gedisa.

Briz, A. (1996). El español coloquial: situación y uso. Madrid, Arco Libros.

Cresti, E. et al. (2002). The C-ORAL-ROM project. New methods for spoken language archives in a multilingual romance corpus. In Proceedings of LREC 2002. Las Palmas de Gran Canaria.

Gallardo, B (1993). La transición entre turnos conversacionales: silencios, solapamientos e interrupciones. In Contextos, XI/21-22, (pp. 189--220).

Halliday, M.A.K. (1985). Spoken and Written Language. Oxford University Press.

Listerri, J. (1997). Transcripción, etiquetado y codificación de corpus orales. In Fundación Duques de Soria. Seminario de Industrias de la Lengua. http://liceu.uab.es/~joaquim/publicacions/FDS97.html

Martí, M.A (Ed.) (2003). Tecnologías del lenguaje. Barcelona, UOC.

Moreno, A. (2002). La evolución de los corpus de habla espontánea: la experiencia del LLI-UAM. In Actas de las II Jornadas en Tecnologías del Habla, diciembre de 2002.Granada.

Tusón, A. (1997). Análisis de la conversación. Barcelona, Ariel.

Rodríguez L.J**.,** I. Torres & A. Varona (2001). Annotation and analysis of disfluencies in a spontaneous speech corpus in Spanish. In Proceedings of the workshop on Disfluency on Spontaneous Speech. Scotland, University of Edimburgh.

Romaine, S. (1996). El lenguaje en la sociedad. Una introducción a la sociolingüística. Barcelona, Ariel.

Processing Spoken Language Data: The BASE Experience

Sarah Creer & Paul Thompson University of Reading, UK
S.M.Creer@reading.ac.uk; P.A.Thompson@reading.ac.uk

Abstract

Transcription and mark-up of spoken language data should ideally present as accurate, full and impartial a representation of the original speech event as is possible, but processing of the data record is subject to a number of compromises between the pull of competing forces, such as the demand for *user readability* while also aiming for *computer readability*, and the requirement (for purposes of interchangeability) for *conformity to existing standards* vs the *accurate description of the particularities of the data*. This paper presents problems that we have encountered during the process of creating a corpus of orthographically transcribed spoken language data for the British Academic Spoken English corpus. Some limitations in the recommendations of the TEI Guidelines are also discussed.

1 Introduction

The British Academic Spoken English (BASE) corpus is a collection of recordings and marked-up transcripts of academic lectures and seminars that is being developed at the Universities of Warwick and Reading. It is designed to be a British counterpart to the Michigan Corpus of Academic Spoken English (MICASE), a corpus that is to some degree representative of spoken events in academic settings in the USA.

Unlike MICASE it does not include speech events other than lectures and seminars, and the majority of the recordings are on digital video rather than audio tape. The corpus currently consists of 160 lectures and 38 seminar recordings, equally spread over four broad academic domains, of which three quarters have already been transcribed, and 40 lectures have been marked up (as of 20/3/04).

The main motivation for developing this corpus is to create a resource for research into spoken academic discourse, and it is targeted chiefly at EAP researchers and practitioners. The corpus should provide a wealth of evidence of naturally occurring language in specific contexts, and the intention is to make it highly accessible, by creating a web interface for interrogation of the corpus, and by tailoring the interface to the needs of potential end-users.

In the early stages of the project, the recordings of the lectures were transcribed by a variety of paid student and volunteer transcribers, for reasons of expedience, and this resulted in a degree of variability in the quality and accuracy of the transcriptions. Once funding was secured, however, we have worked to achieve consistency in the rendering of spoken language in an orthographic representation, and we have also aimed to make the corpus compatible (and thus interchangeable) with other corpora, in particular the MICASE corpus. We therefore follow the TEI Guidelines (as does MICASE) and employ the set of TEI elements for mark-up of spoken language data shown in Table 1.

Transcription and mark-up of spoken language data should ideally present as accurate and impartial a representation of the speech event as possible, but this ideal can never be achieved, as Cook (1995) persuasively argues. The process of data capture and transcription is subject to a number of powerful tensions, repeatedly forcing the transcriber away from the ideal and towards compromise, and this paper describes some of the questions that we have had to deal with during the process of creating our corpus.

Tag	Description
<u>	*An utterance is a discrete sequence of speech produced by one participant, or group of participants, in a speech event. The tag contains transcription of lexical items.*
<pause/>	*Indicates a perceived pause, either between or within utterances, of at least 0.2 seconds duration.*
<vocal/>	*A non-lexical vocal event such as laughter, coughing.*
<kinesic/>	*A non-vocal communicative event such as putting hand up, handing out paper, etc.*
<event/>	*An occurrence, not necessarily communicative, usually non-verbal, noted because it affects comprehension of the surrounding discourse. For example fire alarm, playing of audio tape, etc.*
<shift/>	*A marked change in voice quality for any one speaker*
<writing>	*A passage of written text revealed to participants in the course of a spoken text.*
<distinct>	*Used for words or phrases in languages other than present-day British English. This includes earlier forms of English but does not include proper names.*
<sic>	*Used when a speaker makes a mistake without self-correcting, and the error might otherwise appear to be a transcribing error.*
<trunc>	*Used when a word is truncated.*
<gap/>	*Used to indicate omissions in the text. Also used when names referred to in the recording are withheld at the request of the participant(s).*
<unclear>	*Used when the transcriber is uncertain of exact word(s)*

Table 1: List of elements used in mark-up of the BASE corpus (attributes are not shown)

2 Practical Issues

2.1 Time and Money

The greatest practical concerns for anyone involved in the development of a corpus of spoken language data are the intertwined issues of time and money. Processing spoken language data is an extremely time consuming and expensive process and this places constraints and limitations on what can be achieved. On the BASE project we have calculated that one hour's worth of recording takes, on average, at least 10 hours to transcribe, 3-4 hours to check, and then more than 8-15 hours to mark up. Decisions have to be made over the level of detail that can be incorporated in the mark-up, while at the same time allowing for a large number of recordings to be transcribed and annotated. This is one of the primary tensions underlying decisions over the mark-up of the data: **breadth vs depth**. The following sections describe a range of other tensions affecting and constraining the corpus developers' decision-making processes.

2.2 Corpus Design - Meeting End-Users' Requirements

Design is continually adjusted and reviewed throughout the process of building a corpus and it is important to base these decisions on the perceived usefulness of that particular design feature for the end-users. Ideally corpora would be designed to provide useful information for all areas of speech research, but the constraints of time and money militate against this. The BASE project is funded by a Resource Enhancement grant from the Arts and Humanities Research Board of England, and deadlines for completion of the project must be met, as well as targets for the numbers of transcripts to be completed.

Equally powerful, however, is the notion of the end-user community: the corpus is designed to constitute a set of resources for language teachers and researchers into English for Academic Purposes, and it is these end-users who must be kept in mind throughout the design process. It is for this reason that the chosen form of representation in the corpus is orthographic transcription. This decision is not without its problems, as will be shown in the following sections, but it is fundamental to the project that the corpus should be non-threatening, in its presentation, for the language teacher.

2.3 Conformity

The spoken language data representation has to be readable not only for humans (**user readability**) but also for computers (**computer readability**). At a basic level, this simply requires a high level of systematicity in mark-up, but at a broader level it argues the case for standardisation, and the importance of inter-changeability between corpora. The BASE project has chosen to make its corpus a sister to MICASE, and it is planned that end-users should observe a regularity in annotation and in structuring between the two corpora. On a broader level, the BASE project aims at conformity to the TEI Guidelines. However, as discussed below, there are problems in the guidelines. A further dynamic tension at play here, then, is that between **compliance** with existing guidelines and the need for **customisation** of mark-up to account for the features of the actual data, not all of which have been satisfactorily treated for within the guidelines.

2.4 Data-Gathering - Quality of Recordings

Records of naturally occurring spoken language events are inevitably to some degree partial. At the same time, the corpus developer needs to find ways to produce accounts of these events that, while partial, are of approximately equal degrees of partiality. This principle can be seriously challenged at the first stage: that of data collection in natural settings.

From our experience, the quality of digitised recordings of natural speech events is clearly superior to that of an audiocassette and therefore the transcription itself is more accurate. The BASE recordings were made on minidisc and on digital video in different classrooms and lecture theatres throughout the two universities. In an ideal world, one could get recordings of equal quality in any of these different settings. In reality, however, we encountered problems with the equipment not picking up all of the speech data due to technical difficulties. In some lecture theatres, for example, bulbs in the overhead projectors interfered with the wireless lapel microphones and every time lecturers moved close to the projector to change a transparency, they became inaudible. This kind of data loss can also lead to an inconsistency in the transcripts – gaps in a transcript can be due to a number of factors, such as recording problems, the poor levels of articulation of a speaker, possibly also the need to edit a section for reasons of anonymity (see 2.5 below), and decisions have to be taken over whether, and how, these gaps are to be indicated in the transcript.

Another potential source of inconsistency is the mode of the recordings themselves. The BASE project recordings are mostly on video but some are purely audio. The audio-only recordings do not contain the same level of detail of context and paralinguistic activity that are captured by the video recordings, and this poses a problem for the transcriber who will have to choose between keeping the level of detail restricted to what can be discerned on the audio tracks or having varying levels of detail depending on what is available on video, on the one hand, and audio, on the other. Allied to this dilemma is the practical issue of what software the transcriber should use: a simple audio transcription aid with the use of a VCR for viewing video as a supplementary source of information, or a dedicated computer programme interface, such as the freeware programme, Soundscriber, which permits viewing of the video on the same screen as the text editor and audio controls. The size of the image on the larger monitor may make certain visual features of the recording more salient than they would appear to be when viewed on a small window on a computer screen.

2.5 Constraints on Authenticity

Ideally, we aim to capture naturally occurring speech in an authentic situation, but there is always the question of how authentic the event recorded is when a participant is aware of being observed. Example 1 illustrates an extreme of the observer's paradox – the lecturer, by way of a joke, adds a spoof French translation of a car name for the benefit of the international students (not present at the lecture) who may listen to the recording later as part of their EAP course.

Example 1

```
it's just oscillating off it goes just
oscillates and for the foreign language
students very much like a Citroen Diane
suspension it's getting towards that or
le Diane de Citroen or whatever it
whatever it is am i supposed to interact
with this or should this be a
```

RL027 pt2 09.07-09.46

Another potential threat to the authenticity of the data is the need to protect the identities of the participants. Participants must be aware of the recordings being made, what the data is to be used for and who it will be made available to. This will involve some editing and suppression of the data. A difficult issue is to what extent this suppression should be made. A degree of editing will reduce the authenticity of the data but if the corpus is to be made freely available it is possible that people might make use of data in ways not previously anticipated. For example, the corpus could be used as a source of reference for information that is taken out of context and was not agreed to by the participants. With this in mind it is an ethical decision whether to let the participants look at the transcriptions before they become part of the corpus. This would ensure that all the speakers in the recordings agree to make the information contained in the data available to others and that any information which they see as being problematic or not acceptable in a wider context can be removed. This may be an important ethical issue but it can also seriously compromise the authenticity of naturally occurring data.

3 Representation and Interpretation – Degrees of Partialness

3.1 Orthographic Transcription

Choices on how to represent data are inextricably linked with how the dataset is going to be used and who is going to use it. This choice determines to some extent what information is going to be captured and what is going to be lost. There is a tension between using IPA transcription and capturing the intricacies in the realisation of the speech produced and maintaining user readability and ease of transcription. An orthographic transcription may lose the details in pronunciation but takes into account time and money constraints allowing transcription by and readability for non-experts in IPA.

Using written language (orthographic transcription) to represent spoken language creates problems in that this privileges the written form and to some extent the standardised English form. Such a representation does not fully convey the diversity of spoken English but, again, sacrificing this information allows for a more comprehensive searching capacity. It provides a means for collecting together words that do not necessarily have the same sound profile but which carry what is interpreted as the same meaning characterised by that surface representation, providing there is consistency in spelling and representations of codified and non-codified words.

As explained above, the BASE project is aimed primarily at EAP researchers and practitioners, and so the decision was taken to use orthographic transcription. In the following sections, the focus is therefore on problems that have been encountered in creating a consistent and accurate orthographic representation, and a balanced and detailed mark-up of the data.

3.2 Spelling

A primary difficulty in providing an orthographic transcription of the data is that of spelling. It seems reasonable to assume that there should be standards and conventions available for the representation of language, and transcription guidelines often make reference to a particular spell-checking system used, or cite dictionaries from which spelling rules have been taken. Speech, however, is a transient entity which could provide problems in standardisation and interchangeability of corpora in the short and long term. Some of these problems are:

- Variation in spellings between dictionaries.
- Variation in spellings of words that are all deemed acceptable in English, e.g. *categorise/categorize*, and *per cent, percent* and *per-cent*.
- Technical terms and neologisms not yet codified, which can run the danger of idiosyncrasy.
- How to represent made up words, e.g. *skaboodle*. The lack of a one-to-one phoneme-to-grapheme relationship in English provides inconsistency across corpora and ambiguity in pronunciation.

Decisions have to be made early on in the transcription process to ensure consistency, which, in turn, allows comprehensive and precise searches, and accurate word frequency lists to be drawn from the corpus. For the BASE project a list of standard conventions will be provided for reference alongside the transcriptions.

On whose authority?

In the BASE corpus, there is a lecture on Pericles. The name is spelled *Pericles* in the Oxford Reference Online database, but the notes given by the lecturer use the spelling *Perikles*. By definition the lecturer is an expert in his field, which raises the question of whether to follow the spelling of the expert or to use that of the standard language.

3.3 Definitions of a ‘Word’ and Orthographic Representation of Non-Words

It is a generally recognised fact that spoken and written language differ in various respects, and these differences create a problem for the representation of spoken language following written standards. Spoken language is a continuous process, a complex interaction of articulators, whereas written language is a discrete representation of segments at the word level. What is written in dictionaries does not define what and how all things are said. If the continual stream of speech is segmented into words using, for example, the definition that a word is the minimum unit in writing to which meaning can be assigned, items such as truncated words and noises do not fit into this categorisation. Questions about which individual noises are meaningful enough to warrant word status presents another set of choices to be made by the transcriber.

In Example 3, taken from a lecture on mechanical engineering, the question is whether “zing” is to be represented as a word, or marked with a <vocal> tag which would make it a non-lexical item.

Example 3

```
so i want the arm to come in as quick as
possible to to that point <shift
feature="pitch" new="high"/> zing <shift
feature="pitch" new="normal"/> like that
do the <vocal desc="buzzing noise"
iterated="y" dur="1"/> do the welding and
move away again now if it's overdamped
the case we saw before it will not come
in zing it will come in <shift
feature="tempo" new="ll"/> zing <shift
feature="tempo" new="normal"/> and
eventually get there if it's as the case
we're going to move on to it goes past it
overshoots then of course it would go
zing
```

RL027 pt2 01.24-01.50

In this extract, there are two distinct types of noise made: one is a buzzing noise, which has been tagged here as <vocal/>, and the other is a noise which is here represented as ‘zing’. One influence on the decision to represent it thus is that it consists of sounds of English which can be put together to make a lexical item, and a second factor is that it can be found in a dictionary. In the first three instances of the sound in this extract it performs a role similar to that played by the buzzing noise in describing the sounds made by parts of a welding machine, and it could be argued that the two types of noise should be marked up as being similar, rather than one being <vocal/> and the other lexicalised with a <shift/> tag employed to show the marked prosody of the first three instances. The fourth instance, however, suggests that the sound is performing a more conventional ‘word’ role in the utterance. An alternative is to try to represent the buzzing noise as a word, in the same way as ‘zing’, although there is no clear means (‘buzz’ is not an adequate representation of the sound made in this instance) of doing so.

Another example from the corpus is that of the lecturer who frequently used the phrase “all right” after statements throughout the lecture. The intonation of this remained the same through the lecture but the realisation turned into “mm” rather than using the actual words. This is an example of how a simple transcript of the lecture would hinder the understanding of the speech event. The written transcript can only adequately provide lexical information about the event rather than the paralinguistic and contextual information that the participants are using. This argues the case for linking the transcript to a prosodic transcription or to the audio files, which would help to disambiguate the meaning.

3.4 Visual Representation of Data above the Word Level

The visual representation has an influence on the interpretation and perceptions of the event by the reader. The layout and punctuation provides interpretations of how the spoken language was produced. Edwards (1993) discusses the role of the transcript in spoken language research providing a set of maxims for readability of transcripts which the transcriber should keep in mind to limit misinterpretation:

- Proximity of related events.
- Visual separability of unlike events.
- Time-space iconicity.
- Logical priority.
- Mnemonic marking.
- Efficiency and compactness.

Where written language is delineated by punctuation, the equivalent for spoken language is the complex interaction of prosodic marking. To represent one level of the prosody, and give clues to the phrasing of the data, all pauses have been marked in the BASE corpus and precisely measured. This makes the transcript more difficult to read for the end-user but also provides information about the realisation of the speech. This is an example of the kind of compromise that has to be made in the processing of spoken language data, between the **need for readability**, and the importance of **contextual and paralinguistic information**.

3.5 Interpretation: The Transcriber’s Dilemma

Any transferral of data from a recording of raw speech into written form will be an interpretation. To represent what is heard means that some information is inevitably lost, and this will restrict the options for further interpretation of the data by others.

3.5.1 Relevance

A difficult issue for the transcriber is that they are responsible for deciding what details are relevant to the event. The transcript is necessarily selective and a matter of interpretation. Having visual accompaniments to the transcript can aid the user but it does not overcome the problem that the observer is not a participant in the event. What is deemed relevant by the transcriber may not have been given the same relevance by the participants in the event itself. Often the transcriber does not have all the information, such as just having audio recordings where non-vocal events can be heard but decisions on their relevance have to be made without all the information necessary. Audio recordings do not necessarily capture all non-vocal information that is relevant and a degree of editing has therefore been done without the transcriber being aware of that decision.

3.5.2 Realisation: Surface Form vs Underlying Form

Decisions about whether to represent a fully or partially articulated word as its phonetically realised form or as its underlying form provide further problems of interpretation. If a word is to be represented orthographically, then the representation depends on whether the word appears in a codified form, for example ‘till’ is a separate codified shortened version of ‘until’ but there is no separate dictionary entry for ‘`kay’ as the partially articulated version of ‘okay’. A further example is the *learnt*/*learned* difference. The past and past participle of *learn* can be either *learnt* or *learned*, where *learnt* is preferred in British English, particularly when the word functions as a participial adjective, according to Fowler (1999). In a single recording it is possible to find a speaker switching between a pronunciation of the word which ends in a /t/ sound, and one which ends in /d/. The transcriber then has to interpret whether this distinction exists in the mind of the speaker (they actually thought ‘learnt’ in the first case, and ‘learned’ in the second), and whether or not this should be made clear through the orthography. To make the transcript comprehensively searchable however, the orthography has to be standardised. Representing what is said is a matter of interpretation and the challenge is to constrain the degrees of variability of this interpretation as far as possible.

3.5.3 Interpretation of Homophones

Making decisions about how to represent the spoken form as written language is particularly apparent in the problem of homophones. The transcriber has to interpret the context within which the word is spoken to assign it the most probable meaning and written form. In cases such as “they’re”, “their” and “there” and “see” and “sea”, reliance on the context can usually disambiguate these forms but the transcriber cannot know that that particular form chosen to represent the occurrence is that form which the speaker intended to utter. There are instances where the context does not provide a weighting towards one or the other of the possible representations due to the non-fluent production of speech.

Example 4A

```
this is a problem with property a
companies always have to have somewhere
to live
```

Example 4B

```
this is a problem with property a
company's always have to have somewhere
to live
```

RL031 pt2 07.19-07.26

Making a decision about which of these two forms (‘company’s’ and ‘companies’) is the one that the speaker intended is making assumptions about cognition and pre-planning of utterances. The aim of making a non-theory dependent corpus then has to be compromised to make a representation of what has been uttered.

3.5.4 Correction

The transcriber must transcribe what is heard and not correct the speech to make it grammatical in standard written English. The observer cannot know whether what has been uttered fulfils the intention of the speaker. Correcting what would be seen as grammatical errors in standard written English also implies that the form spoken by the speaker is not acceptable and is a form of judgement on dialectal or idiolectal features of the individual’s speech.

4 Difficulties Using TEI Guidelines

Spoken communication differs from written communication in that it is not only a lexical and syntactic event, but also consists of the context in which the event occurs and of additional paralinguistic information (Cook 1995). To become a participant in a spoken communication event, presence at the point in time at which the communication is delivered and membership of the audience at which the communication is aimed are required. The written record of the words spoken is insufficient to capture all that is required to interpret a spoken event. The TEI guidelines on mark-up of spoken language data provide means for making the representation of the event more comprehensive. Problems in the practicalities of adapting tags for spoken language events from linear written language and trying to represent these events to provide an effective interpretation are discussed in this section.

4.1 Representation

Trying to accurately represent the speech event, particularly the temporal aspects of speech, while conforming to the TEI guidelines, provides a difficulty for the encoder. Representation of an event influences the interpretation by the user as they can only interpret the event from the information provided and how it is presented.

4.1.1 Variability

One of the biggest tensions in providing an accurate representation of spoken language data using the TEI guidelines is between customisation and conformity. Where there are too many potential ways to encode events (for example, ways to mark time alignment) in the TEI guidelines, there will be ambiguity, confusion and this will limit consistency both within and across corpora.

4.1.2 Speech Event as a Temporal Entity

Representing spoken language in written form imposes linearity on speech. To a certain extent this exists due to the constraints on a human’s ability to produce more than one word at a time. However, the TEI guidelines impose linearity on the whole speech event and the events contained within it. Speech is a temporal phenomenon and on levels above and below the representation of the word, other non-linear events, such as prosody, are being constrained by the linear demands of the encoding system. The linear representation also only allows events to occur at certain break points in the transcripts, the word boundaries. These artificial boundaries in a stream of continual speech are enforced by the orthographic transcription system, not necessarily correctly representing the actual temporal events.

This tension between accurately representing the temporal nature of the event and conforming to the TEI standards can be illustrated by <u>, the dividing unit of utterance. It is defined as “a stretch of speech usually preceded and followed by silence or by a change of speaker”. However, a speaker does not necessarily finish their utterance when another interrupts and the option for the transcriber is then to either indicate the temporal aspect of the interruption, breaking the utterance artificially or mark the end of the utterance as a whole entity. The overlap can then be indicated either with time stamps or by using style sheets. If the speaker who interrupts fails to gain the floor, his or her utterance then becomes embedded. The difficulty then is to decide where to mark that embedded utterance. The choice is either to break the utterance of the interrupted speaker, following Edwards’ (1993) principle of proximity, which implies in the transcript that there was a break, or to violate this principle allowing the loss of temporal information. This becomes more complicated if other speakers try to interrupt as it then can become unclear, when represented linearly, which speaker is being interrupted. Using <u> in the representation of multi-speaker events (such as small group discussions in interactive lectures, or a heated exchange of opinion during a seminar) becomes highly problematic. Without time stamps or a link to the recorded event itself, the user can only interpret from the information presented and the manner in which it is presented.

4.1.3 Illustrations

A usual component of academic lectures is that there will be audio-visual aids and illustrations, which can be central to the communication of the information contained in the lecture. The decision is to what extent and how to represent them. The tag provided for such an event is the <writing> tag, which contains “a passage of written text revealed to participants in the course of a spoken text”. The main issue with this tag is how the items being revealed are represented. It is only the written text that is specified to be encoded, again privileging the written form. Items such as diagrams, formulae and illustrations can be just as communicative, particularly in disciplines such as Meteorology and Economics that make frequent use of symbolic representation, and should also be included. Further questions such as how far this can be done within the text and to what extent it is to be reproduced in its physical appearance, such as the font and size, need to be addressed.

The <writing> tag does not deal with how to represent other types of illustrations such as pronunciation examples. Example 6 comes from a lecture in which different accents of English are described demonstrating different pronunciations of the word “through”. In the orthographic transcription this would not be marked but as it affects the understanding of the lecture as an illustration of a point, it would be useful to have a phonetic description of what is pronounced. In the BASE corpus, we have chosen to use the <distinct> element with the ‘lang’ attribute set to a ‘sampa’ value, and then to represent the sound using the Sampa system of phonetic transcription.

Example 6A orthographic transcription

```
so if i start saying if i start changing
my vowel in through to through to through
or something like that which many English
people do
```

Example 6B with suggested pronunciation mark-up

```
so if i start saying if i start changing
my vowel in <distinct lang="sampa">
[Tru:]</distinct> to <distinct
lang="sampa">[Tr}]</distinct> to
<distinct lang="sampa"> [TrY]</distinct>
or something like that which many English
people do
```

RL032 pt1 06.58-07.08

4.2 Interpretation

A general problem in the TEI guidelines in the processing of spoken language data is that many of the tags involve a great deal of interpretation by the encoder. For example, the difference between the tags <event/> and <kinesic/> is that events marked as kinesic are communicative. Not only does the encoder have to decide whether the event is relevant, an interpretation in itself, but also whether it has a communicative function.

Inconsistencies in encoding data also occur across corpora due to the individual needs and interpretation of the encoder. For example the TEI definition of <pause/> is “a pause either between or within utterances”. In the BASE project a pause is empirically defined as a period of silence from 0.2 seconds long, narrowing down the TEI definition. This definition of <pause/> means that the DTD would have to be altered as it does not allow for a pause to occur within <distinct>, an element which identifies words and phrases that are distinct in some way from the surrounding language.

Example 7

```
Passy himself has given instance of th-,
in-, instances of this <pause dur="0.5"/>
# <distinct lang=”french”> dans un parler
tant soit peu langue on
distinguera</distinct> <pause dur="0.4"/>
<distinct lang=”french”>trois petites
roues</distinct> <pause dur="0.2"/>
that's # three little wheels
```

RL011 pt1 05.54-06.04

The <distinct> tag was originally created to describe a feature of written, rather than spoken, language but the need for it can be demonstrated in Example 7 above.

The extendable nature of the TEI guidelines allow for this customisation which can be noted in the DTD. Problems occur, however, when customisation leads to ambiguity in interpretation.

The TEI definition of <pause/> leaves scope for a range of interpretations, some of which will be primarily impressionistic. Customising the definition of <pause/> in turn affects the definition of <u> marking utterances, where the attribute, 'trans', which describes the transition between utterances, has the possible values for beginning the following utterance:

- *smooth* - without unusual pause or rapidity.
- *latching* - with a markedly shorter pause than normal.
- *overlap* - before the previous one has finished.
- *pause* - after a noticeable pause.

Interpretations of the terms 'unusual' and 'normal' then have to be made. Taking the definition of a pause of the BASE project, the attribute smooth becomes redundant. A normal pause would be defined as one that is 0.2 seconds or longer and shorter than normal would be less than that, fitting into the categories of *pause* and *latching* respectively. This compromises the degree of interchangeability of corpora as these definitions depend on a definition of a pause decided by the individual encoder. The dependency on the individual interpretation illustrates the tension between the freedom to customise to the needs of the individual and the ability to conform to conventions and standards to make interchangeable corpora.

4.3 Gaps in the TEI Guidelines for Spoken Language Data

There are gaps in the TEI guidelines which, if filled, could make it far easier to create corpora that have a strict set of standards and conventions. This would create a higher degree of interchangeability between corpora, and reduce the onus on the individual transcriber.

4.3.1 Quoted or Read Speech

One of the main gaps in the guidelines is that there is no satisfactory way of marking speech that is not the speakers' own, such as directly read or quoted material (a common feature of academic discourse). The TEI guidelines suggest that "reading" could be the description of an event but this seems to be unsatisfactory as it refers to the action of reading as a separate event to the language and speech being produced. This would cause difficulties such as the encoder needing visual information about when the reader was looking at the text to read it and making interpretive decisions about whether the text was being read or not. It is relevant for users to know that a portion of language is written language communicated through speech. If marked with an empty <event/> tag, the read speech would not be easily suppressed from the rest of the corpus if so wished by the user. For these reasons, this is not how read or quoted speech is marked in the BASE corpus. A decision remains to be taken as to whether to adapt an existing TEI-specified tag, that roughly approximates, or to introduce a new tag for use in our corpus, and move for it to be included in future versions of the TEI Guidelines. The working definition used for the tag is "text which can be attributed to an identifiable source when it is being quoted and not referenced where the whole text being quoted is at the non-finite clause and above level". This is not a definition without problems but it captures the larger portions of speech which are definitely read or directly quoted. At the same time, there are a number of complications, for example, the extent to which individual words read or quoted would be marked, and how to deal with idioms, proverbs and sayings, and placement of the tags in a stretch of speech with mistakes in the reading or breaks for embellishment. A distinction would also have to be made between text that is being read and text that uses part of the same language to refer to the contents of the text previously read. Information about the source and whether it is present at the time of the speech event or whether it is quoted from memory is also to be considered.

4.3.2 Other Gaps

The TEI guidelines do not address speech impediments providing no guidance for encoders. The TEI guidelines also do not fully represent the whole range of possible occurrences of unintelligible fragments of speech such as truncated words.

4.4 <shift/>

The <shift/> tag exemplifies the problems discussed of both interpretation and representation in the TEI guidelines. The <shift/> tag "marks the point at which some paralinguistic feature of a series of utterances by any one speaker changes". There has to be an initial interpretation by the non-participant encoder to select these segments of speech that demonstrate this change. For example, voice quality is noted as a possible shift in speech but creaky voice is a feature of speech that occurs frequently at the end of phrases. The issue here is whether this would only be marked in places where it is not expected and conveying non-lexical information. This means that the encoding process would then become tied to one particular theory of paralinguistic behaviour selected and interpreted by the encoder.

On top of interpretation, the further question then is how the shift is represented. The TEI guidelines use individual features such as voice quality, pitch range and rhythm and represent changes in the features as relative discrete changes. Speech is a continuous stream of moving articulators which do not necessarily move in the discrete way that this definition requires. The imposition of a linear structure on speech events also only allows the tags to be placed at word boundaries, which makes the decision about where to place the tags difficult for the encoder and the output inaccurate.

The design of the <shift/> tag does not allow more than one feature change to be noted in one tag, even though it is not one feature but that combination of changes in the parameters that creates the perceived shift. The way that this is designed means that each feature can end at different points in the speech and also be picked out for separate analysis. However, it also creates a problem of consistency. The individual changes in each feature may not be enough in themselves to create a perceived shift but are crucial components in combination for creating the shift. The one feature design implies that wherever there is that amount of a change in that feature, there will be a tag indicating it, which is not the case.

Another option of how the change could be represented is by defining the shift as a description of the function e.g. <shift desc=”mimicking a child’s voice”/>. The problem here is that this is an interpretation made by a non-participant in the speech event making non-empirical impressionistic judgements. It would privilege the section because of its function rather than its actual realisation.

Neither of these options provides a satisfactory way of encoding paralinguistic information and it raises the question of whether these sections should be marked at all if there is no accurate and consistent way of describing them. The attempt to capture valuable information about the details of the event is undermined by interpretative and inconsistent representation. It would seem that a comprehensive prosodic transcription of the text or a link to the audio file would be the only non-interpretive option here. However this would not allow direct searching and retrieval of sites of particular prosodic interest. Questions concerning how to deal with shifts in paralinguistic features using the <shift/> tag therefore remain unanswered.

5 Conclusion

In this paper we have discussed some of the issues involved in processing spoken language data, and some of the difficulties we have experienced in using the TEI guidelines for the encoding of the BASE corpus.

Basing a system of transcription and encoding for spoken language data on standards for written language raises a number of problems. It is assumed that standards exist for orthographic transcription in codified language but the reference texts that exist are not entirely comprehensive (and indeed cannot be). This results in difficulties for maintaining consistency both within a corpus and across corpora, which in turn restricts interchangeability.

Problems in the TEI Guidelines for the encoding of spoken language data have been identified from the experience of compiling the BASE corpus. The attempt to provide standards across mark-up and therefore interchangeability of corpora will be successful only if there are tags that can encompass all that the encoder wishes to capture, and standardisation of the usage and interpretation of these tags.

The TEI guidelines cater for a primarily linear representation of speech events which places constraints on the interpretation. Using time markers and linking the marked up transcripts to the recordings that have been made could provide a more comprehensive, although still not a non-interpretive representation of the event. A time-aligned multichannel audio and visual representation of the event alongside the marked up transcript may be able to capture more of the contextual and paralinguistic events, thus relieving the transcriber/encoder of much of the burden of interpretation, and, at the same time, better taking into account the temporal nature of spoken language.

The importance of providing a resource that is rich in information for the user while also limiting the degree of transcriber interpretation has been stressed. Guidelines for the transcription and tagging of spoken language can regard spoken language as a subset of text encoding, as long as there is an accurate reflection of the temporal, non-linear aspects of spoken language rather than constraining it by the non-temporal, linear constraints imposed on it by the written form.

References

BASE corpus URL: http://www.rdg.ac.uk/AcaDepts/ll/base_corpus/

Cook, G. (1995) Theoretical issues: transcribing the untranscribable, in G. Leech, G. Myers and J. Thomas (eds.) *Spoken English on computer*, Harlow, Longman: pp 35-53.

Edwards, J.A. (1993) Principles and contrasting systems of discourse transcription, in J. A. Edwards and M. D. Lampert (eds.) *Talking data: transcription and coding in discourse research*, Hillsdale, New Jersey, Lawrence Erlbaum: pp3-31.

MICASE Michigan corpus of academic spoken English URL: www.hti.umich.edu/m/micase/

Oxford Reference Online, Oxford, OUP. URL: www.oxfordreference.com/views/GLOBAL.html

Pocket Fowler's Modern English Usage. (1999) R. Allen (ed.) Oxford University Press, *Oxford Reference Online*.
URL: www.oxfordreference.com/views/
BOOK_SEARCH.html?book=t30

Text Encoding Initiative URL: www.tei-c.org/P4X/index.html

A “toolbox” for tagging the Spanish C-ORAL-ROM corpus

José M. Guirao† Antonio Moreno-Sandoval‡

† Universidad de Granada, Spain
‡ Universidad Autónoma Madrid, Spain
jmguirao@ugr.es sandoval@maria.lllf.uam.es

Abstract

The goals of this paper are to present the tagging procedure for a Spanish spoken corpus, and to show a tool developed for helping human annotators in the process. Some tagging problems especially relevant in spoken corpora, although found also in written texts, will be introduced first. The paper will summarise the experience of the group in tagging one of the currently largest spontaneous speech corpora (over 300.000 transcribed words)

1 Introduction: problems in tagging a spoken corpus

1.1 The multi-word tagging

The Spanish C-ORAL-ROM corpus consists of 312597 tagged tokens. Every tag marks a lexical unit, regardless the number of graphical words it is made of. That is, words and multi-words are considered as a unit or token. For instance, “hola” and “buenos días” are counted (and tagged) as one token each. Accordingly, amalgams of two lexical units in a single graphical word, such as “al” or “del”, are split into two tokens: “a” “el”, “de” “el”. This assumption is important for understanding the tagging procedure and the tagger evaluation. Moreover, a tagger which cannot analyse multi-words will produce very poor results, at least for a Spanish spoken corpus, where very frequent multi-words will be tagged incorrectly:

Correct	Incorrect
“o sea” Discourse Marker	“o” Conj. + “sea” V
“en lugar de” Prep.	“en” Prep + “lugar” N + “de” Prep
“por ejemplo” D M	“por” Prep + “ejemplo” V

Table 1. Multiword tagging.

1.2. A tag for Discourse Markers

Discourse Markers (DM), whose frequency is lower in written texts, are especially relevant in spontaneous speech. Many tagsets and taggers do not include them, and they are usually considered as adverbs, adjectives or nouns. In our annotation, there is a distinction between discourse markers and other POSs. As a consequence, new cases of ambiguity arise:

“bueno” ADJ “Juan es bueno” Juan is *good*	“bueno” DM “bueno / espero que te guste” *Well* I hope you like it.
“hombre” N “Juan es un hombre bueno” Juan is a good *man*	“hombre” DM “hombre / no te enfades” Don’t be mad, man!

Table 2. Discourse Markers and POS ambiguity

Sometimes it is difficult to decide whether the proper tag is a DM or other category. The intonation, or the pragmatic context can help the trained annotator, but it is impossible to formalise in the disambiguation grammar. DM are responsible for a residual uncertainty.

1.3. Tokenization

To segment the stream in tokens presents several differences with respect to the same task in a written corpus. Not only the recognition of multi-words or amalgams, but also the prosodic tags.

Contrary to written texts, where punctuation marks help to delimit analysis units as clauses, sentences and paragraphs, in spoken transcriptions prosodic marks are used instead. Transcriptions are divided in dialogic turns, and turns have tone units, retracting marks, overlapping marks, disfluencies marks, etc. All these types of phenomena fragment the utterance and introduce additional difficulty in tagging: “agrammatical” sequences are quite frequent in spontaneous speech.

1.4. Unknown words

Every tagger will have to deal with words that are not in its lexicon or in its training corpus. In spontaneous speech there are several sources of unknown words:

- **Neologisms**: Spoken language includes words which are not in the dictionaries or in written texts. New words invented by speakers, which are not incorporated yet to the common language.
- **Pronunciation** mistakes: speakers hardly use the proper word. However, the transcription has to reflect the actual use.
- **Derivatives**: The use of appreciative derivation (prefixes or suffixes) is quite common in spontaneous speech. As a result, a common word as “agua” (water) can be said as “agüita” (literally “little water”). Rules for handling derivation are needed.

On the other hand, **proper names** recognition is not a problem in spoken language: since the transcription does not follow the written language rules, only proper names start with a capital letter.

2 The tagger

The main goal is to provide a complete morphological and POS tagging, including lemmatisation. These tasks have been performed automatically and validated by expert annotators. For the automatic tagging, a hybrid rule-based/statistical tagger has been used. The procedure is divided in three steps:

1. **Word analysis**: a morphosyntactic analyser provides all possible tags for a specific token.
2. **Disambiguation phase 1**: a feature-based Constraint Grammar resolves some of the ambiguities
3. **Disambiguation phase 2**: a statistical tagger (the TnT tagger) resolves the remaining ambiguous analyses.

Human annotators have access interactively to the three phases, and can manually change the annotations. In order to validate the human annotation, the whole tagging system is run and the final results are compared against the human-annotated corpus. Evaluation results are reported in Moreno et al. (forthcoming). It is important to stress that the evaluation experiment on a 50.000 words test corpus did show both a few mistakes in the human annotation and some incorrect rules in the disambiguation grammar. The mistakes were fixed while some problems in the grammar are intrinsically unsolved. The precision rate in the evaluation was 95.6. The figure is quite good compared to similar taggers, if we take into account that some of them do not deal with multi-words and discourse markers are not in their tagsets.

The tagging procedure and its evaluation is described in Moreno & Guirao(2003). Here we will briefly provide the main points.

2.1. Word analysis

For the morphological analysis we use GRAMPAL (Moreno 1991; Moreno & Goñi 1995) which is based on a rich morpheme lexicon of over 50.000 lexical units, and morphological rules. GRAMPAL is a symbolic model based on feature unification grammar The system is reversible: same set of rules and same lexicon for both analysis and generation of inflected wordforms. It is designed to allow only grammatical forms. The most prominent feature is its linguistic rigour, which avoids both over-acceptance and over-generation, providing at the same time all the possible analyses for a given word. This system has been successfully used in language engineering applications as ARIES (Goñi, González and Moreno 1997).

With respect to the original system, developed for analysing written language, new modules have been incorporated to handle specific spoken language features:

1. A new tokenizer, for identifying utterance boundaries by means of dialog turns and prosodic tags.
2. A derivative morphology recogniser, including rules and lexicon entries for over 240 prefixes and suffixes.

The analysis procedure consist of five parts:

1. *Unknown words detection*: after the tokenizer segments the transcription in tokens, a quick look-up for unknown words is run. The detected new words are added to the lexicon
2. *Lexical pre-processing:* here the program splits portmanteau words (“al”, “del” → “a” “el”, “de” “el”) and verbs with clitics (“damelo” → “da” “me” “lo”).
3. *Multi-words recognition*: the text is scanned for candidates to multi-words. A lexicon, compiled from printed dictionaries and corpora, is used for the task.
4. *Single words recognition*: every single token is scanned for every possible analysis according to the morphological rules and lexicon entries. Approximately 30% of the tokens are given more than one analysis, and some of them are given up to 5 different analyses.
5. *Unknown words recognition*: those remaining tokens that are not considered new words, pass through the derivative morphology rules. If some tokens still remain without any analysis (because they were not included in the lexicon nor were recognised by the derivative rules), they will wait until the statistical processing, where the most probable tag, according the surrounding context, is given.

2.2. Disambiguation grammar

POS disambiguation has been solved using a rule-based model. In particular, an extension of a Constraint Grammar using features in a Context-Sensible PS. The output of the tagger is a feature structure written in XML. Here is shown the possible tags for the token “la”, as an article and as a pronoun.

```
<pal    cat=″ART″    lema=″el″    gen=″fem″
num=″plu″> la </pal>

<pal cat=″P″ lema=″la″ pers=″p3″ gen=″fem″
num=″plu″> la </pal>
```

The formalism allows several types of context sensitive rules. The most basic and frequent rule is as follows:

```
"word" → <cat="X"> / _• <cat ="Y">
"word" → <cat="Z"> / <cat ="W"> •
```

Here are some rules for disambiguating the token “la”:

```
"la" -> <cat="ART"> / •_<cat="N" gen="fem">
"la" -> <cat="P">   / •_<cat="V">
"la" -> <cat="ART"> / • <cat="ADJ">
"la" -> <cat="P">   /  yo _•
"la" -> <cat="P">   /  tú _•
```

The grammar writer tries always to provide as much particular rules as possible for a given ambiguous case. The goal is to get the higher level of precision in disambiguation in this phase.

2.3. Statistical disambiguation

For the remaining unresolved ambiguities, an statistical tagger (the Tnt tagger, Brants 2000) is applied. The statistical model has been obtained from a 50000 words training corpus, which is a subset of the whole spoken corpus. The training corpus has been verified by linguists. The statistical part is applied at the end of the process, when the competence-based knowledge (the grammar and lexicon) is not able to provide a precise and discrete analysis. This way, in case no appropriate analysis is found, always the likeliest tag is assigned.

3. The "toolbox"

In order to help human annotators, an xml-based interface has been developed, which allows the interactive edition of the lexicon, the disambiguation grammar and the annotated text. This "tool box" integrates the different modules of the system: the tokenizer, the morphological analyser, the rule-based and the statistical disambiguation. The interface has resulted to be a useful tool for controlling the complex process of enriching and modifying the mentioned modules.

In this section, we will show some screenshots to give an idea of the benefits of employing such an interface tool. This section will also show how the annotator faces different type of problems.

3.1. Editing the annotated texts

The most basic tool is an editor-concordancer which allows to search for problems and wrong analyses. The experienced annotator usually knows which are the problematic cases. In Spanish the most frequent and hard problem is the disambiguation of "que", as a RELative and as a Conjunction. "Que" is the most frequent token in spoken Spanish, and it appears in so many contexts that it is impossible to write disambiguation rules for every case, and statistical models do not resolve either (at least in the current state of training). Careful verification by hand is needed.

This option allows to search for occurrences of "que" in some problematic contexts (see Figure 1). After finding a wrong tag, the annotator has the option to directly write the correct tag, and saving the result.

3.2. What if the word is not in the lexicon?

No lexicon (nor a statistical model) is complete. As a consequence, a method for adding new entries is needed. Spontaneous speech presents words which are not usually found in written texts or printed dictionaries. This option edits the GRAMPAL lexicon, allows to introduce and modify entries and saves the enriched lexicon (Figure 2).

3.3. The disambiguation grammar editor

In the interactive process of revising the annotated texts, the linguist wants to add new rules for disambiguation in specific contexts. As a typical grammar writing process, the linguist has to test the new rule. This option allows to edit the grammar file, compile it and try a utterance (Figure 3). The last is especially useful for checking whether the new rule is working properly or not, without running the whole tagging process on the file.

4. Conclusions and future work

Quality tagging of corpus requires, in addition to a good and complete tagger, a human verification of the annotated text. If the corpus is intended to be used as a reference data resource, as it is the case, then a linguist-controlled annotation is a must. A friendly interface that integrates the different modules is a clearly useful tool.

This paper has also shown the experience of tagging an spontaneous speech corpus. In many ways, the procedure is similar to tagging a written corpus, but some differences have also been exposed.

We expect to enrich the current tool box and the current corpus with new layers of annotations, syntactic and semantic information (Alcántara 2003).

5. References

Alcántara, M. (2003). Semantic Tagging System for Corpora. Poster presented at the V International Workshop on Computational Semantics, Tilburg (Holland).

Brants, T. (2000). TNT – a statistical part-of-speech tagger. In Proceedings of ANLP, Seattle, USA.

Goñi, J.M., González, J.C. and Moreno, A. (1997). ARIES: a lexical platform for engineering Spanish processing tools. In Natural Language Engineering, 3(4), pp. 317-345.

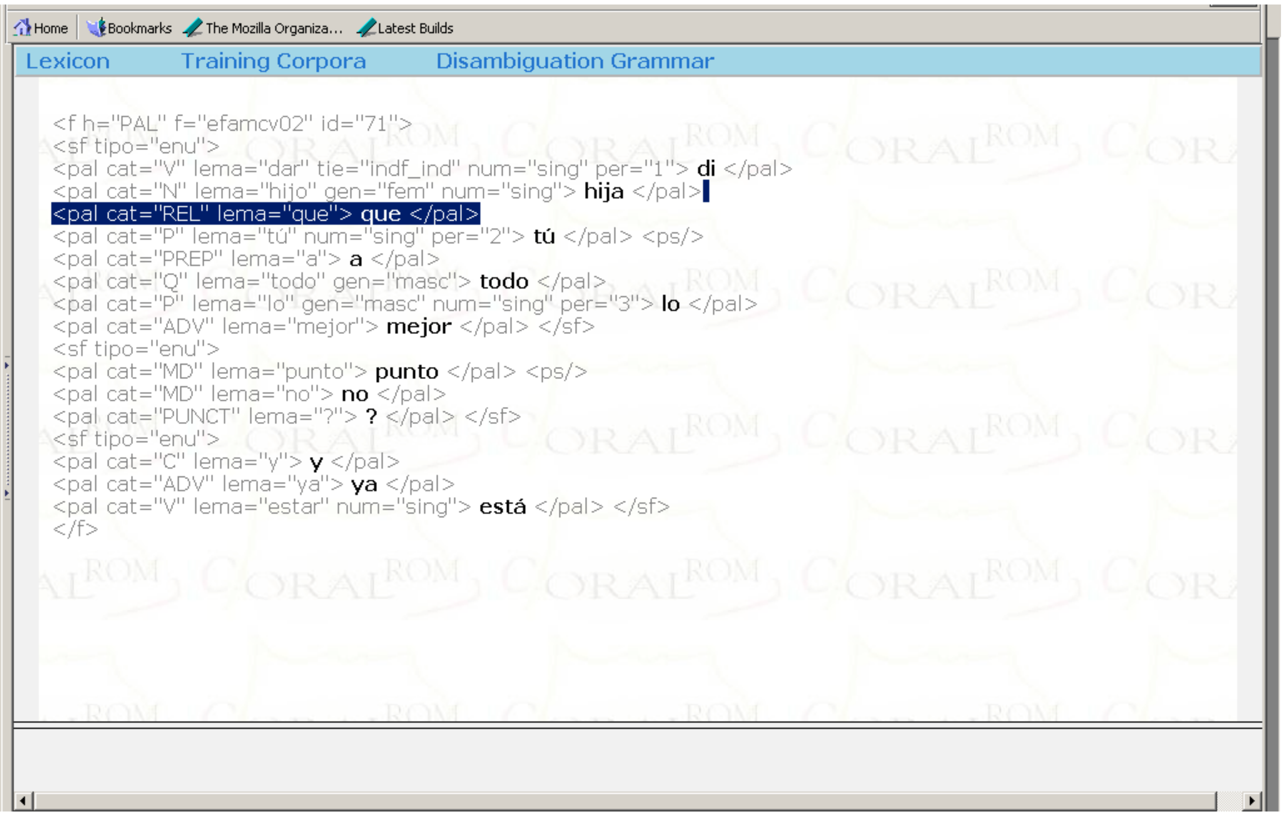
Moreno, A. (1991). Un modelo computacional para el análisis y generación de la morfología del español. Ph.D. Thesis. Universidad Autónoma de Madrid, Spain.

Moreno, A. and Goñi, J.M. (1995). GRAMPAL: a morphological processor of Spanish implemented in Prolog. In Proceedings of Joint Conference on Declarative Programming (GULP-PRODE' 95). Marina di Vietri, Italy.

Moreno, A. and Guirao, J.M. (2003). Tagging a spontaneous speech corpus of Spanish. In Proceedings of RANLP 2003. Borovets, Bulgaria.

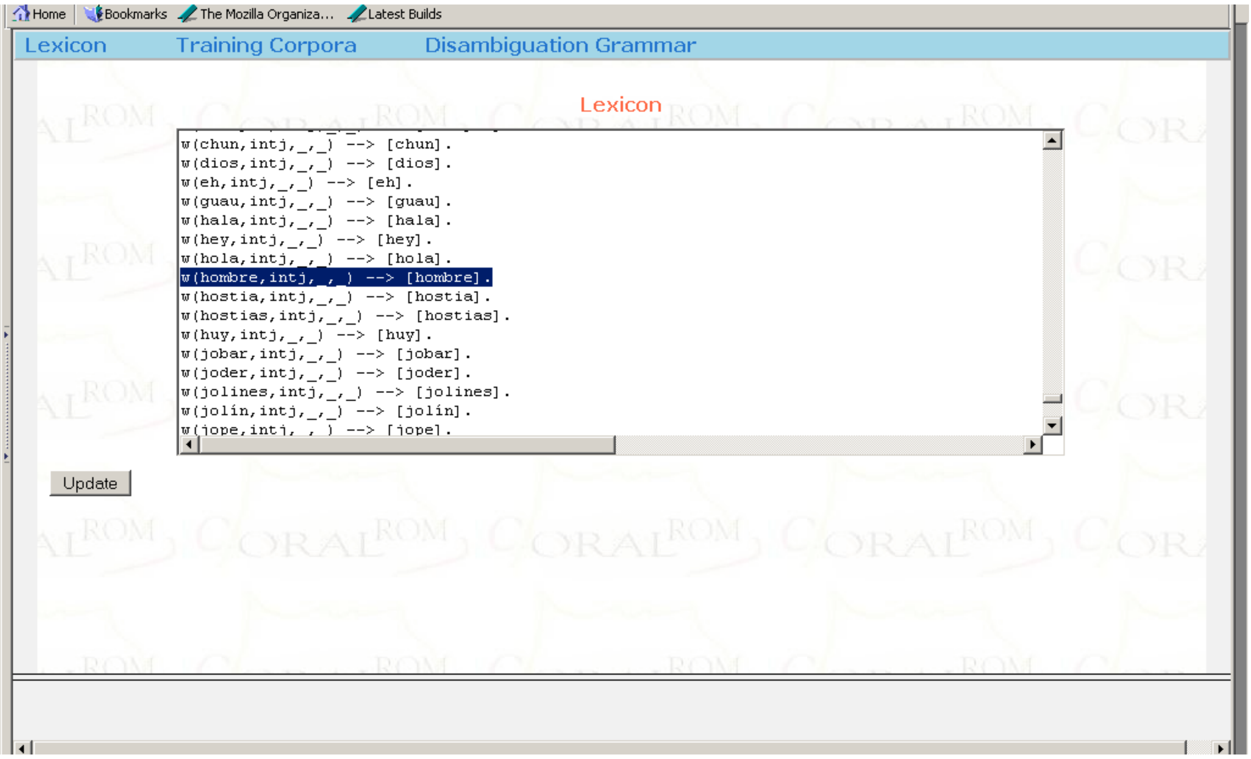
Moreno, A., De la Madrid, G., Alcántara, M., González, A., Guirao, J.M. and De la Torre, Raúl (forthcoming). Notes on the Spanish Spoken corpora and linguistic studies. To be published as a chapter in the C-ORAL-ROM book.

6. Appendix: The Screenshots



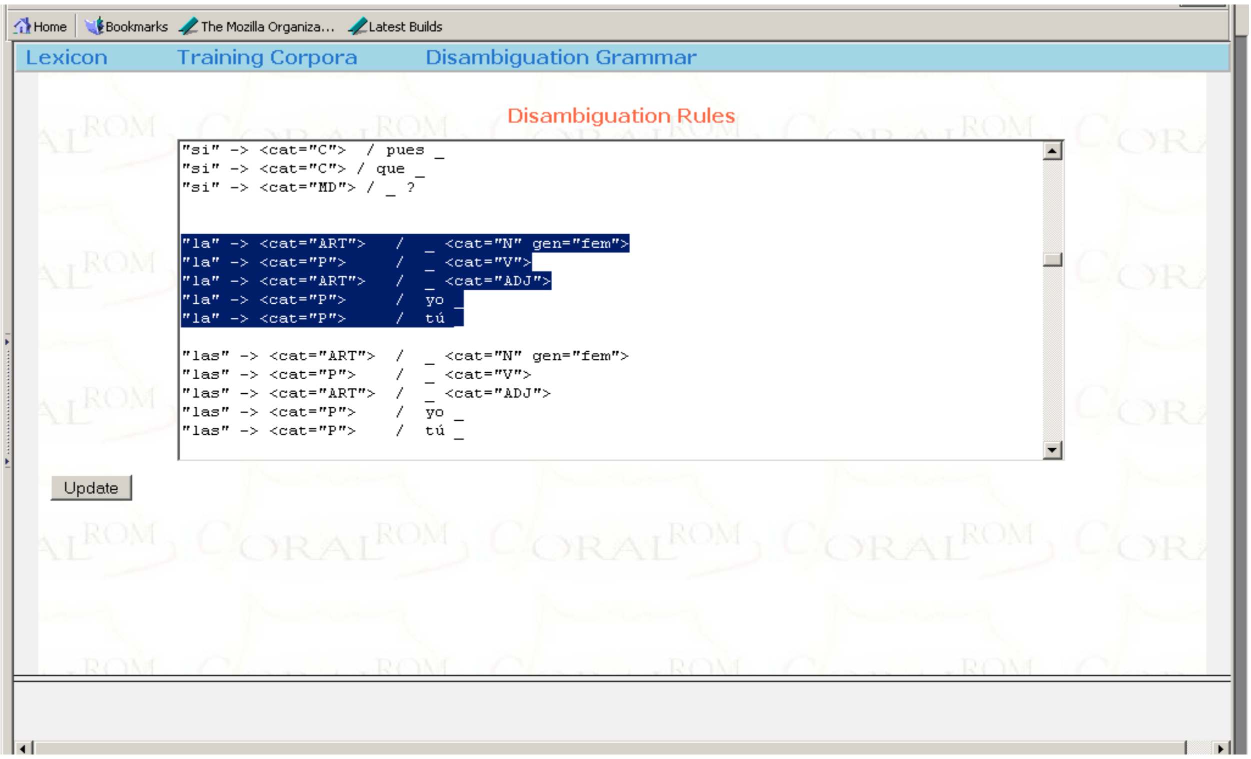
```
<f h="PAL" f="efamcv02" id="71">
<sf tipo="enu">
<pal cat="V" lema="dar" tie="indf_ind" num="sing" per="1"> di </pal>
<pal cat="N" lema="hijo" gen="fem" num="sing"> hija </pal>
<pal cat="REL" lema="que"> que </pal>
<pal cat="P" lema="tú" num="sing" per="2"> tú </pal> <ps/>
<pal cat="PREP" lema="a"> a </pal>
<pal cat="Q" lema="todo" gen="masc"> todo </pal>
<pal cat="P" lema="lo" gen="masc" num="sing" per="3"> lo </pal>
<pal cat="ADV" lema="mejor"> mejor </pal> </sf>
<sf tipo="enu">
<pal cat="MD" lema="punto"> punto </pal> <ps/>
<pal cat="MD" lema="no"> no </pal>
<pal cat="PUNCT" lema="?"> ? </pal> </sf>
<sf tipo="enu">
<pal cat="C" lema="y"> y </pal>
<pal cat="ADV" lema="ya"> ya </pal>
<pal cat="V" lema="estar" num="sing"> está </pal> </sf>
</f>
```

Figure 1. Editing the tagged file



```
w(chun,intj,_,_) --> [chun].
w(dios,intj,_,_) --> [dios].
w(eh,intj,_,_) --> [eh].
w(guau,intj,_,_) --> [guau].
w(hala,intj,_,_) --> [hala].
w(hey,intj,_,_) --> [hey].
w(hola,intj,_,_) --> [hola].
w(hombre,intj,_,_) --> [hombre].
w(hostia,intj,_,_) --> [hostia].
w(hostias,intj,_,_) --> [hostias].
w(huy,intj,_,_) --> [huy].
w(jobar,intj,_,_) --> [jobar].
w(joder,intj,_,_) --> [joder].
w(jolines,intj,_,_) --> [jolines].
w(jolín,intj,_,_) --> [jolín].
w(jope,intj,_,_) --> [jope].
```

Figure 2. Editing the lexicon



```
"si" -> <cat="C">  / pues _
"si" -> <cat="C"> / que _
"si" -> <cat="MD"> / _ ?


"la" -> <cat="ART">    /   _ <cat="N" gen="fem">
"la" -> <cat="P">      /   _ <cat="V">
"la" -> <cat="ART">    /   _ <cat="ADJ">
"la" -> <cat="P">      /   yo _
"la" -> <cat="P">      /   tú _

"las" -> <cat="ART">  /   _ <cat="N" gen="fem">
"las" -> <cat="P">    /   _ <cat="V">
"las" -> <cat="ART">  /   _ <cat="ADJ">
"las" -> <cat="P">    /   yo _
"las" -> <cat="P">    /   tú _
```

Figure 3. Editing the disambiguation grammar

Towards the Creation of an Electronic Corpus to Study Directionality in Simultaneous Interpreting

Claudio Bendazzoli, Cristina Monti, Annalisa Sandrelli, Mariachiara Russo, Marco Baroni, Silvia Bernardini, Gabriele Mack, Elio Ballardini, Peter Mead

Dipartimento di Studi Interdisciplinari su Traduzione, Lingue e Culture (SITLeC), University of Bologna
direzionalita@sslmit.unibo.it

Abstract

Spoken corpora have long been awaited in the field of simultaneous interpreting studies. Small scale attempts have so far provided a variety of theories and results which need scientific validation. Our research project aims at creating an electronic parallel corpus for study of simultaneous interpretation from and into different languages (Italian, English and Spanish), including both source and target texts. The corpus will serve as a basis to observe the strategies implemented by professional interpreters depending on the languages involved. The following paper presents the focus of our study, the steps undertaken to select, collect and transcribe the material and the analysis to be carried out at different levels.

1. Introduction

The main object of our research project is the study of simultaneous interpreting (SI), as an activity influenced by the language pairs involved and the language direction in which interpreting is carried out. The aim is to create an electronic corpus as a means of investigating linguistic and textual strategies used by interpreters (e.g. generalisation as a way to overcome lexical difficulties within the time constraints of SI), the role of such strategies being briefly outlined at the end of this introductory section.

In the field of translation studies, electronic corpora have already been recognized as a promising tool which can be employed in different ways and for different purposes (Aston, 2001; Laviosa 1998). Yet, very little has been done so far in the field of conference interpreting.

It is probably because of difficulties in developing a sound methodology that there seem to be no examples of electronic SI corpora in the literature. The attempts so far have involved small scale research projects (Kalina, 1994). The collected SI data were stored in computers, but not in a machine-readable format, limiting the possibility to consult and analyse them. The creation of an electronic corpus of simultaneously interpreted data has been long awaited. As highlighted by Armstrong (1997), it would undoubtedly offer a wide range of advantages to SI researchers. In this respect, Shlesinger highlights

> *the potential to use large, machine-readable corpora to arrive at global inferences about the interpreted text* (1998: 487)

Shlesinger (489) argues that bilingual and parallel corpora provide a sound basis for testing of hypotheses (for example, about interpreting strategies) and reviewing the results of previous studies which have limited empirical validity.

As an inter-linguistic activity consisting in the oral translation of speech while it is being delivered, SI is an object of study with different investigation paradigms depending on the research focus – cognitive, linguistic, pragmatic and so on.

The interpreter's output is an oral target text (TT), produced at the same time as s/he is listening to an oral source text (ST) uttered by a speaker in another language. Against this background, the TT can be considered as a secondary text, meaning that it depends on the ST in both content and pace of delivery. The interpreter's TT presents specific features that characterize it and distinguish it both from written translation (since it is oral and is simultaneously delivered) and natural oral speech (since it is determined by the ST).

However, the study of SI and the study of spoken language share many common methodological issues, as the interpreter's output shares a distinctive element with ordinary speech and spoken language, namely orality. This is why in the creation of a corpus of interpreted texts our difficulties are in many ways the same as those encountered by linguists and researchers engaged in the compilation of spoken language corpora, ranging from data collection and their subsequent transcription to more complex levels of annotation.

In short, our SI corpus analysis will have to adjust to its needs the best of existing methodology for both written and spoken language corpora. The procedures, tools and techniques already developed in these fields must be taken into account for reference and inspiration. A corpus of STs and TTs will provide a useful basis to study interpreting strategies and to observe how these are applied as a function of different variables, like the nature of the ST (read vs. off-the-cuff etc.), directionality and language pairs (cognate languages vs. non-cognate languages).

The term "directionality" implies considering the language combinations in which an interpreter works. His or her languages are usually classified as follows:

> *Active languages: A: the interpreter's native language (...), into which the interpreter works from all her or his other languages (...). B: a language other than the interpreter's native language, of which she or he has a perfect command and into which she or he works from*

> *one or more of her or his other languages. (...) Passive languages: C: languages of which the interpreter has a complete understanding and from which she or he works.*
> (AIIC, Art. 7)

Therefore, the possible language directions in SI are: from A – or possibly C – to B (active), and from B or C to A (passive).

SI has been defined by Riccardi (1999: 170) as a problem solving activity. Problems are due to its distinctive features, namely listening and talking at the same time, the use of two different languages, the production of a message that comes from another speaker, etc. To perform this task, the interpreter has to develop some general strategies. In addition to these, there are specific strategies which depend on the languages involved (Riccardi, 1999). The lexical and syntactical structures of one language cannot be merely transferred into another language, as illustrated by the differing word order for sequences of simple units like substantives and adjectives in a given language pair (e.g. English-Italian). Interpreting from or into German provides the even more evident problem of verb position. The interpreter must thus take into account the specific features of the two languages within the time constraints imposed by SI. This is why textual features of TTs and STs can be studied with a view to detecting the implementation of interpreters' strategies.

The next sections describe our project, focusing on the research material, the methodology for data collection, transcription and the different levels of analysis.

2. Materials and Methods

Collecting recordings of STs and TTs as they are actually produced is the first major obstacle we faced. In the field of interpreting studies this is a widely acknowledged problem. Sylvia Kalina (1994: 224) explains, in this respect, the reluctance of conference organizers and contributors to make recordings available, because of confidentiality issues; similarly, she notes that interpreters may feel self-conscious if they know they are being recorded.

In addition to this problem, we will have to manage the quantity and homogeneity of data. We need to compile a large corpus in order to be able to base our results on as many interpretations as possible.

With this aim in mind, we decided to collect interpretations from the European Parliament (EP). The TV channel "EbS" (Europe by Satellite, http://europa.eu.int/comm/ebs/index_en.html) broadcasts the plenary sessions of the EP in all of the official languages, thus making it possible to record many hours of data. This channel allows a choice of language when SI in different languages is provided. It has already been used in other studies, as it can offer a variety of communicative events in many languages with high quality interpreting (de Manuel Jerez, 2003; Turrini, 2002).

The plenary sessions provide a specific setting in which the following variables are controlled:

- all interpreters are qualified and experienced professionals;
- most interpreters work into their mother tongue[1], i.e. from B and/or C to A;
- all speeches share a number of features: the same context of production, where a strictly institutionalized procedure is followed; they are all political speeches; they are usually prepared in advance by the speakers (who only have a limited amount of time available); and they generally have a formal register and a similar content pattern.

In these EP sessions interpreters work in particularly difficult conditions, due to the variety of specific terminology and problems of interpreting read texts, which often reflect the standard of written rather than spoken language and may also be badly delivered (Marzocchi & Zucchetto, 1997: 82). These are recurrent problems faced by interpreters in other contexts as well. However, as Marzocchi & Zucchetto (ibid.) point out:

> [these] *phenomena are so extreme in this setting, that the interpreter's intuitive, subjective limit of what can actually be interpreted is sometimes reached. The plenary seems therefore to provide suitable conditions for research in view of the intensity reached by such phenomena.*

This source thus provides an excellent opportunity to see how professionals interpreting into their A language rise to the different challenges, namely which strategies they employ, when and how.

For a full assessment of directionality, interpretations into B languages must be collected as well. There are thus plans to record material from conferences on the Italian private market, where interpretation in this direction is common practice.

As far as methodology is concerned, we intend to subdivide our study into three stages:

- collection of the material through video-recordings, digitisation and editing;
- transcription of the material and creation an electronic corpus;
- analysis of the corpus.

2.1 Collection of Material

Given the spoken nature of the study material, it has to be collected in the form of recordings. At present, all the plenary sessions of the European Parliament to be held this year (2004) are being video-recorded from the satellite TV channel *EbS*.

In our study, a first set of recordings will focus on Italian, English and Spanish, to be able to compare

[1] According to the so called "mother tongue principle", interpreters working for the European Parliament generally work only into their native language, with the exception of those language combinations which cannot be otherwise covered (Marzocchi & Zucchetto, 1997: 74). The latter are expected to increase as a consequence of the enlargement process.

interpretations between cognate vs. non-cognate languages. These recordings will then be digitised: the original speeches will be digitised as video files, while the interpretations will be converted into audio files (wav or mp3 format). The aim of this considerably time-consuming "collecting stage" is to create an archive of digital recordings on DVD and CD to be used in future studies as well (Shlesinger, 1998; Gile et al., 2001).

An entire file, i.e. a full morning or afternoon recording, will be subsequently edited following the sequence of the speakers taking the floor during each parliamentary session. Editing digital audio files is much easier than editing audiotapes and can be done with dedicated software programmes, such as *CoolEdit* or *Wavelab*.

Each file will feature a reference code, in order to provide instant information on the time, mode of delivery and the languages involved. The reference code system is structured as follows:

e.g. 09-02-04-m-001-int-en-it
day month year (m) morning / (p) afternoon session progressive number (org) original / (int) interpreted version source language target language

Each text will be integrated with a header containing metatextual information, thus making it possible to group and query the files also on the basis of different entries. The header is to be structured as in the following example:

date: 10-02-04-m
speech number: 033
language: en
type: org
duration: 02.30
number of words: 356
mean speed: 142,4
text delivery: off-the-cuff/written to be read/mix
speaker: Byrne
gender: M
origin: Irish
political function: Commissioner, DG: Health and Consumer protection
specific topic: European Centre for Disease Prevention and Control

Original tracks (i.e. STs) will be stored alongside their interpretations into the other languages (i.e. TTs) in the same directory. This will be identified by the reference code explained above.

Each digital clip is complemented by its transcription, so that every speech is available in two versions, i.e. spoken and written. It is worth emphasising once again that the actual data are the recorded speeches to be stored, while the transcripts are the means by which it is possible to carry out computer-based analysis of the corpus. The corpus will probably include a collection of written texts, i.e. the written translated versions of the speeches. These translations are available on the Internet website of the European Parliament and will be used in the future stages of our research, with a view to studying interpreted texts vs. both original and translated texts.

The next section outlines the transcription criteria.

2.2. Transcription and Annotation

There is general consensus that transcribing is a time-consuming and labour-intensive activity, which cannot be outsourced. Since one of our objectives is to create a significantly large corpus of interpreted texts, the time factor must be taken into account.

In order to speed up the transcribing process, speech-recognition software programmes are proving extremely useful. The two programs employed in the study are *Dragon Naturally Speaking* and *IBM Via Voice*. As voice recognition software requires the use of the same voice to be able to recognize it, the transcriber trains the system to recognize his/her voice first. Then, s/he listens to the recording and performs what conference interpreters refer to as "shadowing", i.e. repeating aloud every word s/he is listening to (Schweda Nicholson, 1990; Lambert, 1992). This way, a rough draft of the transcript can be obtained in a relatively short time (as fast as the interpreters' oral performance); finally, the transcript is revised while annotations are made.

It must be pointed out that the transcribing process is an integral part of the analysis, as the transfer from the spoken to the written mode involves a deep processing of the material.[2] In this respect, focusing on the main object of one's study is as important as choosing clear transcribing conventions.

Literature on several notation systems used in transcriptions from other studies was reviewed, and conventions employed in final undergraduate dissertations from the "Scuola Superiore di Lingue Moderne per Interpreti e Traduttori in Forlì" (University of Bologna) were assessed.

The Jeffersonian system[3] proved to be the most efficient one for our purposes, in that it is well established and widely accepted in the research community (Orletti & Testa, 1991: 254; O'Konnel & Kowal, 1994). Besides choosing a notation system as a reference procedure, transcription conventions were adapted with a view to using the transcripts for automatic analysis. Since this called for even greater simplicity, basic symbols for speech features (in STs and TTs) were carefully selected. As Armstrong suggests:

> *The first level of annotation should aim at recording a minimal amount of reliably identifiable information.* (Armstrong, 1997: 158)

[2] See also Orletti & Testa (1991: 277), who suggest that "[transcription] be considered as a moment for theoretical reflection and not as a mere translation from spoken to written". [our translation]

[3] This system was first developed by Gail Jefferson and then adapted for a variety of research purposes, such as conversational analysis and interpreting studies.

After long discussions, it was decided that the first level of notation should be limited to the lexical level by adopting a standard orthographic transcription, following the rules of the Interinstitutional style guide[4]. Other levels, such as prosodic, paralinguistic and extralinguistic features, appeared to be beyond the present scope of our study and too complex to be analysed simultaneously. It therefore seemed advisable to concentrate on a few selected aspects at a time.

Our basic transcriptions (annotated verbal elements) can provide the basis for further work. At a later stage other annotations could be added through the use of dedicated electronic tools of proven scientific validity. One example concerns the distribution of pauses in the speeches under analysis. Pauses could be added systematically to the available transcriptions by employing IT tools to measure the duration of pauses and to produce graphic representations of variations in wavelength and pitch (e.g. *CoolEdit*, *WinPitch*).

The following transcription conventions were established for the first group of selected features in our STs and TTs which can be indicative of interpreters' processing difficulties:

#	incomprehensible lexical elements
:	vowel/consonant lengthening
=	latching
-	word truncation
uhm	preceeds word truncation
/	mispronunciation and disfluencies
…	silent long pause
ehm	filled pauses
h	sigh
( )	general comments (as a header to the transcript)
[ ]	specific comments (to point out specific features within the transcript)

Table 1: our notation system

Initially, punctuation was used as a prosodic marker. However, this method proved to be both unreliable and too time-consuming without the aid of electronic tools. As was explained earlier, in the initial stage of our study there was no need to note prosodic features. These could be studied later, through the use of proper software tools.

To facilitate automatic analysis, the text needs to be spelt "properly". Thus, all the words uttered in ways that deviate from accepted standards (i.e. latching, vowel or consonant lengthening, truncated words and so on, see the list in Table 1.) and which needed to be transcribed using notation symbols were normalised in the transcripts.

Each normalised item is followed, between angular brackets < >, by the same item as it was actually uttered featuring the notation symbol concerned. Depending on the kind of analysis to be carried out, the words in brackets can be included or excluded automatically, as in the following example:

thank you President ehm well I have been asked a number of <o:f> uhm <quest-> questions by all of the speakers and I will endeavour to answer as many </may/> as possible on the time available to <to:> me

When the transcripts are completed, they can be grouped and analysed in different ways. In order to allow a wide range of possible analyses, much care was taken in choosing basic ASCII characters, in order to keep computer readability problems to a minimum.

The manually transcribed data are then tagged and lemmatized. For English, we use the TreeTagger (Schmid, 1994). For Spanish, we use FreeLing (Carreras et al., 2004). For Italian, we use the tagger combination described by Baroni et al. (2004) and we lemmatize using a morphological lexicon that is still under development at SSLMIT. These data are then converted to XML format used by the IMS Corpus Work Bench - CWB (Christ, 1994), as in the following example:

```
<speech date="10-02-04-m" id="022" lang="en"
type="original" duration="16.00" ... "
function="Commissioner, DG: Health and Consumer
protection" topic="Asian bird flu">
...
I          PP     I        I
have       VHP    have     h:ave
been       VBN    be       been
supplying  VVG    supply   supplying
...
</speech>
```

The XML attributes specified for each speech allow the user to restrict queries according to several parameters, e.g., whether the speech is original or interpreted, according to topic, duration, etc.

The four columns inside each <speech> element correspond to the normalized wordform, part of speech, lemma and transcription of each token. The wordform stream contains a normalized orthographic transcription of the token. The transcription stream, while not a full-fledged phonetic transcription, marks phenomena such as lengthenings, hesitations, misproductions, etc.

The users interested in lexical aspects of original and interpreted speeches can rely on the normalized wordform (which was also used for tagging and lemmatization), whereas those who are also interested in the phonetic aspects can issue queries based on the transcription stream.

The corpus can be queried on the web using the powerful CQP language of CWB. We also provide a web interface to cwb-scan-corpus, where the output of a query is automatically sent to the issuer (cwb-scan-corpus is a

[4] Available from the website of the EU: http://publications.eu.int/code/en/en-000400.htm

CWB tool that produces frequency lists for ngrams matching a certain pattern).

Current work on automated annotation and encoding is focusing on how to align speeches and on how to encode alignment data. For up-to-date information on the status of encoding and the available interfaces to the corpus, please visit http://sslmit.unibo.it/~direzionalita/sslimint.html.

3. Levels of Analysis: Present and Future Developments

As pointed out in section 3.2., the corpus can be analysed at different levels and transcriptions can be complemented at any time with further levels of annotation for specific purposes at different stages in time. At an initial stage, the corpus can be analysed to highlight a variety of features, such as word frequency, grammatical constructions or lexical density (Shlesinger, 1998: 488).

When working on a parallel corpus (Shlesinger, *ibid.*), STs can be compared with corresponding TTs in different languages, as shown in Figure 1. This type of analysis may show language and direction-related strategies in interpreting. In other words, the analysis may reveal certain translational patterns and how difficulties are handled depending on the languages involved.

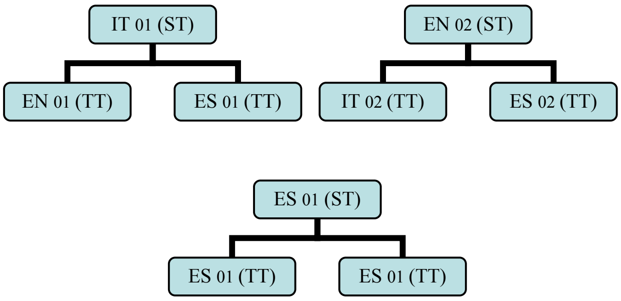


Fig. 1. parallel corpus
(EN = English, ES = Spanish, IT = Italian)

Overall, the creation of an electronic corpus of interpreted texts may provide a significant amount of interesting insights into SI, in that

> *Access to this data on a computer is the first step for a systematic study of the regularities within a given language or across languages. The researchers can more easily identify recurrent patterns that might represent the individual interpreter's strategy in a given situation or for a particular language combination.*
> (Armstrong, 1997: 159)

Since the project is in its initial stages, the corpus is still not large enough to warrant meaningful results. However, in order to test the notation conventions in use, a small-scale analysis was conducted with some automatic analysis systems.

The Unix Command Line proved to be a flexible and powerful tool for preliminary analysis. It allows users to carry out many operations that would otherwise be impossible or very difficult with a graphic interface (Church, 1994). The advice of experts in the field of computational linguistics has been crucial at this stage of the study.

Unix provides many generic utilities to manipulate and query the text files. It relies on many tools and commands that can be combined for the task at hand in a flexible way.

A future stage of our study entails the analysis of the material through what Baker defines as comparable corpora:

> *La méthodologie employée ici ne consiste pas à comparer des textes sources à leur traductions, mais plutôt à comparer des textes originaux et des traductions dans une même langue et dans des domaines apparentés. Les deux ensemble de textes, sous forme électronique (...), sont appelés « corpus comparables »*
> (Baker, 1998: 2)

All the STs in each language can be compared with all the TTs interpreted into the same language and their written translations (WT) (Shlesinger, 1998: 3), as can be seen in Figure 2.

This will provide a basis for analysing mode-specific strategies and general interpreting strategies. An analysis of this kind may yield results related to the differences between modes of delivery (original speech vs. interpreted speech vs. written translation), showing different styles in SI[5] and strategies specific to interpreting or to written translation.

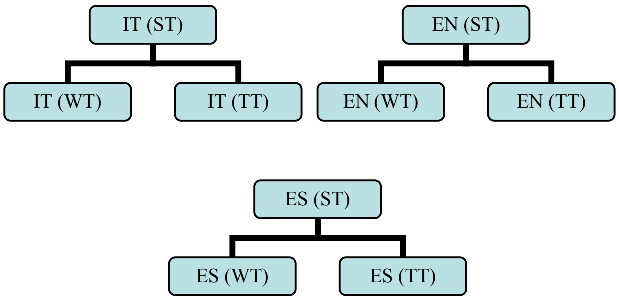


Fig. 2. comparable corpus
(EN = English, ES = Spanish, IT = Italian)

Next, a further level of analysis will involve the alignment of the texts, which may provide even more information on interpreters' strategies and techniques. However, the alignment of interpreted texts with their corresponding STs may prove to be more complex than the alignment of written STs and translated TTs, owing to considerable variations in interpreters' overall management of ST inputs and *décalage*[6].

Among the possible solutions to the problems posed by *décalage* and alignment, there is the possibility to

[5] For examples of this in translation see: Marmaridou (1996).

[6] In conference interpreting, *décalage* (or ear voice span) is the time elapsing between the speaker's and the interpreter's output. In fact, SI is not 100% *simultaneous*, as interpreters begin their delivery slightly after what speakers say.

visualise the transcripts in a score notation, so as to have different lines for each text. The first line displays the ST, while the lines below show the TTs in different languages. Several software programs are currently being evaluated, such as *Exmaralda* (Schmidt, 2001; 2003a; 2003b) and *SyncWriter* (Ehlich, 1993; Meyer, 1998, 2000) to try out the various features they offer and gauge their suitability to the kinds of analysis we wish to carry out.

However, the use of these tools will be a further step in the present research, and we are open to suggestions and support from anyone who has an interest in interpreting studies and the relatively pioneering use of corpora within this field.

To conclude on a positive note, we are confident that corpus linguistics can offer useful tools for the analysis of interpreted data, but at the same time that these data can be of interest to computational linguists as well. Indeed, study of SI can provide significant insights in many other disciplines, ranging from psycholinguistics to language production and acquisition, in that it can be considered an extraordinary "laboratory of [...] experimentation [...] both for the control of external variables and for the 'artificiality' of the task" (Flores D'Arcais, 1978: 393).

References

AIIC Language Classification in http://www.aiic.net/ViewPage.cfm/article118.htm#langues

ARMSTRONG, Susanne (1997) "Corpus based methods for NLP and translation studies". *Interpreting* 2:1/2, 141-162.

ASTON, Guy (ed.) (2001) *Learning With Corpora*. Houston TX: Athesian.

BAKER, Mona (1998) "Réexplorer la langue de la traduction: une approche par corpus". *Meta*, 43: 4, 1-7 in
www.erudit.org

BARONI, Marco, BERNARDINI, Silvia, COMASTRI, F., PICCIONI, Lorenzo, VOLPI, Alessandra, ASTON, Guy & Marco MAZZOLENI (2004) "Introducing the La Repubblica Corpus: a large, annotated, TEI (XML)-compliant corpus of newspaper in Italian". LREC 2004.

CARRERAS, X., CHAO I., PADRÓ, L. & M. PADRÓ (2004) "Freeling: an open-source suite of language analyzers". LREC 2004.

CHURCH, Ward Kenneth (1994) "Unix^{TM}. for poets". *Notes of a course from the European Summer School on Language and Speech Communication, Corpus Based Methods*, July.

CENCINI, Marco (2000) *Il Television Interpreting Corpus (TIC) - Proposta di Codifica Conforme alle Norme TEI per Trascrizioni di Eventi di Interpretazione in Televisione*. SSLiMIT, University of Bologna, unpublished undergraduate dissertation.

CHRIST, O. (1994) "A modular and flexible architecture for an integrated corpus query system". COMPLEX 2004.

DE MANUEL JEREZ, Jesús (2003) "El canal Ebs en la mejora de la calidad de la formación de intérpretes: estudio de un corpus en vídeo del Parlamento Europeo". In Kelly, D. (ed.) *Forum on Directionality in Translating and Interpreting*. Granada, Spain 14-15 November 2002. Amsterdam: John Benjamins.

EHLICH, Konrad (1993) "HIAT: a transcritpion system for discourse data". In Edwards, J.A. & M.D. Lampert (eds.) (1993), 123-148.

FALBO, Caterina, RUSSO, Mariachiara e STRANIERO SERGIO, Francesco (a cura di) (1999) Interpretazione simultanea e consecutiva. Problemi teorici e metodologie didattiche. Milano: Hoepli.

FLORES d'ARCAIS, G.B. (1978) "The contribution of cognitive psychology to the study of interpretation" in GERVER, D. & H.W. SINAIKO (eds.) (1978), 385-402.

GERVER, D. & H.W. SINAIKO (eds.) (1978) *Language Interpretation and Communication*. New York & London: Plenum Press.

GILE, Daniel et al. (2001) *Getting Started in Interpreting Research*. Amsterdam: John Benjamins.

KALINA, Sylvia (1994) "Analyzing Interpreters' Performance: Methods and Problems". In Cay Dollerup and Annette Lindegaard (eds.) *Teaching Translation and Interpreting 2: Insights, Aims, Visions*. Amsterdam: John Benjamins, 225-232.

LAMBERT, Sylvie (1992) "Shadowing". *The Interpreters' Newsletter*, 4: 15-24.

LAVIOSA, Sara (ed.) (1998) *L'approche basée sur le corpus. The Corpus-Based Approach. Meta* 43:4 (special issue).

MARMARIDOU, S.S.A. (1996) "Directionality in translation processes and practices". *Target* 8: 1, 49-73.

MARZOCCHI, Carlo e Giancarlo, ZUCCHETTO (1997) "Some considerations on interpreting in an institutional context: the case of the European Parliament". *Terminologie et Traduction*, 3, 70-85

MEYER, Bernd (1998) "What transcritpions of authentic discourse can reveal about interpreting". *Interpreting*, 3: 1, 65-83.

MEYER, Bernd (2000) "The computer-based transcritpion of simultaneous interpreting". In

DIMITROVA, E. Birgitta & Kenneth, HILTENSTAM (eds.), 151-158.

O'CONNEL, Daniel C. and KOWAL, Sabine (1994). "Some Current Transcription Systems for Spoken Discourse: A Critical analysis". *Pragmatics* 4: 81-107.

ORLETTI, Franca e Renata TESTA (1991) "La trascrizione di un corpus di interlingua: aspetti teorici e metodologici", *Studi italiani di linguistica teorica e applicata* 20:2, 243-283.

RICCARDI, Alessandra (1999) "Interpretazione simultanea: strategie generali e specifiche" in FALBO, Caterina, RUSSO, Mariachiara e STRANIERO SERGIO, Francesco (a cura di) (1999), 161-174.

SCHMIDT, H. (1994) "Probabilistic part-of-speech tagging using decision trees". *International Conference on New Methods in Language Processing*.

SCHMIDT, Thomas (2001) "The transcription system EXMARaLDA: An application of the annotation graph formalism as the Basis of a Database of Multilingual Spoken Discourse". In Proceedings of the IRCS Workshop on Linguistic Databases, 219-227, in http://www.rrz.uni-hamburg.de/exmaralda/de/dokumentation.html

SCHMIDT, Thomas (2003a) *A short introduction to the EXMARaLDA Partitur-Editor,* in http://www.rrz.uni-hamburg.de/exmaralda/de/dokumentation.html

SCHMIDT, Thomas (2003b) "Visualising Linguistic Annotation as Interlinear Text". In *Arbeiten zur Mehrsprachigkeit*, Serie B (46) Hamburg, in http://www.rrz.uni-hamburg.de/exmaralda/de/dokumentation.html

SCHWEDA NICHOLSON, N. (1990) "The role of shadowing in interpreter training". *The Interpreters' Newsletter*, 3: 33-40.

SHLESINGER, Miriam (1998) "Corpus-based interpreting studies as an offshoot of corpus-based translation studies". *Meta* 43:4, 486-493.

TURRINI, Cinzia (2002) *L'interpretazione del linguaggio non letterale al Parlamento Europeo*. SSLiMIT, University of Bologna, unpublished undergraduate dissertation.

Developing a Dialogue Act Coding Scheme: An Experience of Annotating the Estonian Dialogue Corpus

Tiit Hennoste, Mare Koit, Andriela Rääbis, Maret Valdisoo

University of Tartu
Liivi 2, 50409 Tartu, Estonia
{tiit.hennoste, mare.koit, andriela.raabis, maret}@ut.ee

Abstract

The paper gives an overview of the dialogue act coding scheme that we have been developing with the goal of annotating the Estonian dialogue corpus. Our primary task is to analyse Estonian spoken dialogues with the further aim to model human-computer interaction in Estonian. We have studied various coding schemes and tried to adopt the best features of these schemes. The paper describes our experience of analysing and annotating the Estonian dialogue corpus.

1. Introduction

When we started to build the Estonian dialogue corpus (EDiC), we set up two tasks: 1) to study human-human spoken dialogues, and 2) to model human-computer interaction in Estonian. Our goal is to model natural dialogue on the computer, i.e. the computer as a dialogue participant must follow the norms and rules of human-human communication.

We analyse how various types of dialogue acts are used in a special domain – calls for information (information offices, travel bureaux, etc.), and how it depends on Estonian cultural space. Therefore, the study of human-human conversations is essential.

The main part of the EDiC is made up of dialogues taken from the Corpus of Spoken Estonian of the University of Tartu (Hennoste, 2003a) – 205 calls for information and 115 face-to-face interactions, altogether 320 transcribed texts with the total length of 100, 000 running words. The transcription of conversation analysis (CA) is used.[1] Every text is provided with a header that gives background information.

The remaining part of the EDiC – 21 written information dialogues – were collected during computer simulations using the Wizard of Oz method (Valdisoo et al., 2003).

We have decided to develop our own typology of dialogue acts because no coding scheme seemed to fully correspond to our needs. We tried to adopt positive sides of various schemes. Below, we will describe the principles of our typology and an experience of using the typology for annotating the EDiC. It can be mentioned that 306 spoken dialogues (84,000 running words) and all the 21 simulated dialogues were annotated separately by two different persons using our coding scheme and then unified. The kappa value is 0.74 (computed on 45 spoken information dialogues). Our typology contains 126 dialogue acts.

2. Theoretical Sources

The first well-known typology of dialogue acts based on the study of real conversations was worked out by John Sinclair and Malcolm Coulthard (Sinclair & Coulthard, 1975). This typology was further developed by Anna-Brita Stenström (Stenström, 1994). Several researchers have considered practical problems of determining dialogue acts in the last decade – corpus linguists, discourse and conversation analysts, language technologists (Allwood et al., 2001; Stolcke et al., 2000; Jokinen et al., 2001).

We started our work by studying various annotation schemes with the goal of finding a suitable scheme for annotating dialogue acts in our dialogues. A very good overview of coding schemes is given in the MATE report (Dybkjær, 2000). We were looking for general, not domain-restricted schemes (e.g. DAMSL, SWBD-DAMSL, Traum's scheme). On the other hand, we studied dialogue acts used in CA (e.g. Stenström, 1994), and formal theories of dialogue (e.g. Dynamic Interpretation Theory which departs from the analysis of spoken information dialogues, Bunt, 1999).

Why do we have developed a new typology? In our opinion, the existing treatments face at least the following problems:

1) The categories used by most of the typologies are too general. Wide categories join phenomena that behave differently in actual texts.

2) The acts used in dialogues are typically divided into two groups: information acts and dialogue managing acts. The most important one is repair that means solving all the communication and/or linguistic problems (cf. Schegloff et al., 1977). Repair is considered together with other dialogue management acts, not as a separate phenomenon. Different repairing means and methods are considered, instead of repair as a unitary process having its own beginning and ending. Human-human communication can not be fluent in principle. Similarly, the computer must be able to differentiate problem-solving acts from information acts in human-computer interaction. It is essential because some information acts and repair acts have similar form (e.g. almost all initiations of repairs are questions). This differentiates our typology from the existing typologies: the dialogue managing acts must be divided into 1) fluent conversation managing acts and 2) acts for solving communication problems, or repair acts.

3) Most of the studies are based on English texts but it is not always possible to transfer the results into other languages and cultures.

4) Most of the existing typologies are logic-centred. A typology has been construed from a small number of dialogue examples and/or theoretical conceptions, and then adjusted in the process of actual analysis of dialogues. The main problem with this approach is that a

[1] See Appendix 1.

typology contains a small number of acts, and the acts have wide borders. The analysis of actual conversation needs a more detailed typology.

A typology of dialogue acts must satisfy the general principles of classification. There are several requirements for developing a dialogue act system for dialogue analysis. First, the acts that people use in actual conversations must be found. Secondly, the acts system must make it possible to differentiate various functions. Thirdly, the typology must make it possible to differentiate utterances that have similar linguistic realisation but different functions.

The principles underlying our typology are the same as for the other coding schemes (Edwards, 1995). Three types of principles are considered: 1) category design, 2) readability, 3) computer manipulation.

There are three **category design** principles: categories must be systematically discriminable, exhaustive, and systematically contrastive.

The first principle means that it must be clear for every event in the data and every category if the category is applicable or not.

Exhaustibility means that there must be a fitting category for every particular case in the data (even if only 'miscellaneous'). For that reason, every type in our typology contains a subtype 'other' which is used for annotating the items we are not interested in at the moment, or are not able to determine exactly.

Contrastivity needs some more discussion. If the categories are considered as exclusive alternatives then they determine partially each other's boundaries. If the categories are not mutually exclusive, like dialogue acts, then an implicit contrast exists between the presence and absence of every individual conceptual property. A researcher who is making a choice of a set of descriptive coding categories must apply the contrastivity of categories, that is, (s)he must choose such categories which most likely reveal the necessary properties (Edwards, 1995: 21-22).

Readability means that the typology must be clearly arranged and understandable for users. A readable coding scheme is re-usable, and has an effect on the kappa value of annotation.

The **computer manipulation** principles are systematicality and predictability.

Systematicality means that variability must be avoided (e.g. pronouncing variability, capital letters, word gaps). Variability causes problems in spoken language.

Predictability means that general rules must be found that make it possible to pre-determine the codes.

3. Conversation Analysis as a Basis of EDiC Dialogue Act Typology

Our typology departs from the point of view of conversation analysis that focuses on the techniques used by people when they are actually engaged in social interaction. This is an empirical, inductive analysis of conversation data (Hutchby & Wooffitt, 1998). The main idea underlying the analysis is that conversation is the collaboration of participants based on three mechanisms: turn taking, repair, and adjacency pairs (AP). An advantage of this approach is that CA departs from empirical data, i.e. it tries to find the explicit markers in the text that allow to determine the functions of utterances. In our opinion, it is especially important for human-computer interaction.

On the other hand, CA implements only microanalysis, it does not use a previously ready-made typology of dialogue acts but tries to analyse every dialogue act as if it were unique.

The departing point of the CA is that a partner always has to react to a previous turn regardless of his/her own plans and strategies. That's why the analysis of relations between two turns is central in this approach.

People follow implicit and explicit norms in their conversation. Still, violations of norms are possible. In this case, signals to the partner must be given. We suppose that the computer as a dialogue participant must follow the norms and recognise signals of their violations by the partner.

A serious problem with the CA approach is that no lists of dialogue acts are formed, and no co-occurrences of acts are treated. Only one action is considered at a moment.

On the basis of empirical microanalysis of dialogues, CA has established three means of conversation organisation: turn taking, adjacency pairs, and repair organisation.

3.1. Turn Taking

A dialogue consists of turns. A turn is a continued speech of one speaker. There are four problems here.

1. The hearer must recognise when a turn has ended. Intonation-dependent, grammatical and pragmatic marks exist that signalise turn endings (Ford & Thompson, 1996).

2. The hearer must understand who will speak next. Turn taking rules are used (Sacks et al., 1974:704).

3. Turns consist of turn constructional units (TCU). Their boundaries are intonation-dependent and/or grammatical-pragmatic. In the ideal case, the holding entity of a dialogue act is the least linguistic unit used for performance of an action. The main holding unit is TCU but it is not a rule. A part of a TCU, or many TUCs together can hold one act. Sometimes, a phrase or a single word can be an individual act. Therefore, no definite linguistic constructions exist that could be called as holding units of dialogue acts. The study of empirical data is needed for determining different dialogue acts and their linguistic realisations.

4. Can one unit hold many functions at the same time? It is possible in our dialogue act system. Acts can be classified on several grounds. One classification criterion is formal (for example, wh-questions, yes/no questions, etc.). Another criterion is informal - the function that an act performs in conversation (e.g. adjustable questions). Therefore, an utterance can have more than one tag in the annotated dialogue corpus.

3.2. Adjacency Pairs

The fact that conversation is based on a system of APs is most fundamental for information dialogues.

Some classes of dialogue acts conventionally form pairs where the production of the first part makes the second part relevant (Hutchby & Wooffitt, 1998: 39-43):

- the parts of AP are ordered - there are differences between the first and second act. Given the first part, a fixed second part is required (e.g. question requires answer)

- the second part has a fixed relevant place in dialogue. Ideally, two parts are located next to each other (e.g. a question requires an answer). There may be insertion sequences between these parts but the second part remains relevant even if it is not produced in the next turn.

The computer must be able to differentiate the first part of an AP (which expects a reaction) from acts that do not require a reaction, e.g. questions from narrative, or real questions from rhetorical ones. Thus, the first principle of the act typology is: the acts forming APs must be differentiated from the acts that do not form APs.

Such a system makes it possible to relate antecedents and consequences and to analyse such types of turns /utterances that are located between question and answer (insertions sequences).

This is different from the logical approaches. An advantage of the CA approach is that dialogue can be viewed as act sequence. When analysing a dialogue we can suppose that the first part of an AP always presupposes the second part.

3.3. Repair

Human-human communication is not fluent in principle. A special system is used to signal and solve communication problems. CA explains that a universal system – repair – is applicable to different conversation types. The repair system is language-dependent to a certain degree.

Spoken conversation is linear, non-reversible. Therefore, repair is a text sequence. Repair initiations and repair performance means can be found. Rights, obligations and actions of participants in a repair sequence can be analysed.

4. EDiC Typology of Dialogue Acts

Based on the above principles we get the following main typology of dialogue acts (cf. Hennoste et al., 2003a; 2003b).

The acts are divided into two big groups – adjacency pair acts (AP acts) and single acts (non-AP acts). On the other hand, the acts are divided into dialogue managing acts and information acts.

Dialogue managing acts are divided into fluent conversation managing acts and problem solving acts. All the sub-groups contain the type 'other'. This type contains the remaining acts of the sub-group. Such an approach gives us an opportunity to extend the typology when needed.

Let us consider the groups in greater detail.[2]

4.1. AP acts

4.1.1. Communication management acts

There are two sub-types of **fluent** communication management AP acts: conventional acts and topic change acts.

Conventional acts are Greeting, Wish, How-are-you, etc. These acts are linguistically expressed as formulas, they can be given as lists. The acts form certain APs and occur in certain parts of dialogue (mostly at the beginning and at the end).

Topic change acts are used to start a new topic or sub-topic.[3]

```
V: öheksa ´null ´neli.
nine zero four    DIJ: GIVING INFORMATION
H: mhmh.
mhmh  VR: ACKNOWLEDGEMENT: NEUTRAL
 ja=siis ´üks küsimus [´veel]=et
 and then another question    YA: ADVANCE
NOTE
V:                  [{-}]         YA:    UN-
INTERPRETABLE
H:   kas   ´telefoninumbrid   mis   algavad
numbritega  viis  ´kaks,  (.)  kas  need  on
´Rakvere või: ´Pajusti omad.
are the phone numbers that begin with the
numbers five two for Rakvere or for Pajusti
KYE:  ALTERNATIVE  QUESTION  |  TVE:  TOPIC
CHANGE INITIATION
```

Problem solving AP acts are used for other-initiated repairs and contact control (cf Hennoste et al., 2003b).

We differentiate three types of **repair initiations**. In the first two types, the hearer who recognises a problem in the previous text initiates a repair, and the partner who caused the problem carries out repair. These two types are *clarification* and *non-understanding*.

The third type is *reformulation* (candidate understanding in CA) where the hearer initiates the repair and suggest her own interpretation of the problematic place. The partner agrees with, or rejects this interpretation.

```
H: a:ga kallis see `tööluba on.
how much does this work permit cost  KYE:
WH-QUESTION
(0.5)
-> V: kuidas
Pardon       PPE: NON-UNDERSTANDING | KYE:
WH-QUESTION
-> H: kallis `tööluba on.
how much does the work permit cost  PPJ:
REPAIR | KYJ: GIVING INFORMATION | KYE: WH-
QUESTION
V: ei, tö- `tööluba ei=ole=`vaja.
no  work  permit  is  required  KYJ:  GIVING
INFORMATION
```

The second type of problem solving AP acts is the group **Contact control** acts. The speaker checks the functioning of the communication canal (*can you hear, hallo*). These acts occur typically in phone conversations where certain phrases are used – formulas can be given as lists.

4.1.2. Information acts

Another group of AP acts is formed by Information acts. There are 3 sub-groups depending on the function and construction of the first part of AP: directive, question, opinion (cf. Hennoste et al., 2003b).

[2] The full list of dialogue acts is given in Appendix 2. Act names are originally in Estonian. An act token consists of two parts: the first two letters are an abbreviation of the act group name in Estonian (e.g KK = kontakti kontroll 'contact control'). The third letter is only used for AP acts: the first (E) or the second (J) part of an AP act. The second part of a token is full name of an act.

[3] Transcription marks are listed in Appendix 1.

Most dialogue act typologies contain similar groups. The differences concern only determining the borderlines between directives and questions. Sometimes questions and directives are differentiated on the basis whether the user needs some information (then it is question) or whether (s)he wants to influence the hearer's future non-communicative actions (then it is directive). Our departing point is that it is not important for dialogue continuation whether the hearer must to do something outside of the current dialogue or not. She must react to both a question and a directive because both are the first parts of APs. The second part of an AP can be verbal or non-verbal (an action). It can immediately follow the first part of AP or occur later. In addition to that, the answer can influence the course of dialogue (e.g. determine the structure of the partner's next turn). The main difference between directives and questions is formal – questions in Estonian have special explicit form (interrogatives, intonation, and special word order) whereas directives do not.

The first parts of **directive** APs in our typology are 1) request, 2) proposal, and 3) offer (Hennoste et al., 2003a). The second parts are fulfilling directive: giving information / missing information / action, agreement with directive, refusal of directive, postponing the answer to directive, restricted fulfilling of directive, restricted agreement with directive.

In the following example H informs V that he wants to speak with a certain person.

```
H: tere `päevast,
good afternoon RIJ: GREETING
`ütelge palun (.) `hambaravi telefoninumbrit.
please tell me dentist phone number   DIE: REQUEST
V: kaheksakend neli kuus kaheksa kaheksa.
eighty-four six eight eight   DIJ: GIVING INFORMATION
```

Proposals and offers differ from requests because they expect a different second part. Requests expect giving information. Suitable reactions to requests are fulfilling directive: giving information or fulfilling directive: missing information. Proposals and offers expect agreement or refusal (agreement with directive and refusal of directive in our typology). Therefore, requests are similar to open questions and proposals and offers are similar to closed yes/no questions.

Offers are differentiated from proposals. In the first case, the action originates with the author (offer: *I'll send you the programme*); in the second case, it originates with the partner (proposal: *please come tomorrow, call me later*), cf the following example.

```
H: ää kas te oskate öelda kui palju se ´pilet maksab.
could you tell me how much the ticket would cost  KYE: WH-QUESTION
V: kahjuks ´piletite=inda meil ei=ole.
unfortunately we do not have ticket prices KYJ: MISSING INFORMATION
te peate sealt küsima=
you must ask there  DIE: PROPOSAL
ma=võin ´numbri anda kui ´soovite.
I can give you the number if you wish  DIE: OFFER
H: mt ee võite anda ´küll jah?
mm you can give yes  DIJ: AGREEING | DIE: REQUEST
V: see on kaks ´kolm, (0.5)
it is two three  DIJ: GIVING INFORMATION
```

The sub-group **Question** contains acts that are expressed in the form of questions in Estonian dialogues (the first parts of APs) and answers to questions (the second parts).

There are three question types that depend on the expected reaction:

- questions that expect giving information: wh-question, open yes/no question
- questions that expect agreement/refusal: closed yes/no question, question that offers answer
- questions that expect the choice of an alternative: alternative question.

We differentiate two sub-groups in the first and third group because on the one hand they have formal specific features and, on the other hand, there are particular problems with determination of their boundaries (see also Hennoste, 2003b). Both open yes/no question and closed yes/no question have similar form but they expect different reactions from the answerer (e.g. *Are you open in winter?* expects the answer yes or no, but by asking *Is there a bus that arrives in Tallinn after 8?* the questioner really wants to know the departure times of buses).

The second type of question (expecting agreement/refusal) can be divided into two sub-types: closed yes-no question, and question that offers answer (e.g. *see ´seitseteist kolmkümend on kõige ´ilisem või /is the seventeen thirty the latest/*). The questioner has an opinion, a hypothesis and (s)he expects the partner to confirm it. These sub-types can be differentiated on the basis of different linguistic realisations in Estonian.

Certain questions are closely connected with corresponding answers in APs:

- wh-questions and open yes/no-questions => open answers: giving information / missing information;
- closed yes/no-questions and questions that offer answer => closed answers: yes / no / agreeing no / other yes/no-answer;
- alternative questions => alternative answers: one / both / third choice / negative / other alternative answer.

The third sub-group of information AP acts, **Opinion,** contains the following first part acts: assertion, opinion, other. Assertion expresses a reliable knowledge while opinion expresses a belief. The second parts are: accept, reject, limited accept, refusal, other.

4.2. Non-AP acts

The second big group of acts is formed by Non-AP acts. Similar sub-types as in the AP acts group are distinguished: Communication management acts and Information acts.

4.2.1. Communication management acts

Conventions form an important sub-group of fluent communication managing acts: remission, introduction, recognition, contact, call. They do not expect a fixed second part.

Responses (considered as feedback by some researchers) are cases where the hearer gives a reaction signal of her own accord. The central dialogue particles that express reactions to previous turns are *mhmh, jah* and *ahah* in

Estonian (correspondingly, *hem*, *yes*, *oh*). They perform different functions in conversation: *mhmh* shows that a participant is hearing and distancing; *jah* indicates agreement; *ahah* shows that previous information was new for the hearer. Dialogue particles can form a turn alone or start a new, longer turn. There are neutral and evaluative responses - continuer: evaluative, continuer: neutral, acknowledgement: evaluative, acknowledgement: neutral, etc.

```
M: `kolmsada kolmkend=`seitse.
three  hundred  and  thirty  seven     KYJ:
GIVING INFORMATION
(0.8)
O: {* mhmh *}
mhmh VR:   ACKNOWLEDGMENT: NEUTRAL
 (0.8) see on nor`maalne.
this  is  all  right     VR: ACKNOWLEDGMENT:
EVALUATIVE
```

Non-AP problem solving acts are self-repair and other-repair. Unlike other-initiated repair, the speaker or the hearer himself/herself carries out the repair.

4.2.2. Information acts

Non-AP information acts are divided into two sub-groups: primary single acts and additional information.

Primary single acts give information, express opinion etc., thus they carry on conversation and add some information but they are not the first parts of adjacency pairs. These acts are advance note, narration, promise, statement, border-mark, rhetorical question, rhetorical answer, quotation, non-verbal, other. The act 'un-interpretable' that we use for coding the un-interpretable utterances belongs to this group too.

Additional information acts are used by the speaker to modify or complement his/her preliminary information - specification, explication, account, inference, conclusion, etc. In most cases, this act follows immediately the main act.

An extended example of an annotated spoken dialogue is given in Appendix 3.

When comparing our typology with that of DAMSL, a number of similarities can be mentioned. DAMSL has 4 annotation levels (Communicative Status, Information Level, Forward Looking Function, Backward Looking Function). The same levels are expressed in our typology (directly or indirectly).

(1) Communicative Status

DAMSL has 3 tags on this level: un-interpretable, abandoned, self-talk.

We have the act 'un-interpretable' in the group of non-AP primary single acts; abandoned and self-talk are a part of the act 'self-repair' (in the group of non-AP repair acts).

(2) Information Level

DAMSL has 4 sub-levels on this level: Task (doing the task), Task management (talking about the task), Communication management (maintaining the communication), Other level.

Act groups Question, Directive, Opinion and non-AP additional information acts correspond to the Task level in our typology.

The task management level is not expressed in our typology.

Conventional, Contact control, Other-initiated repair, and non-AP responses (acknowledgement, continuer) correspond to the Communication management level in our typology.

The label 'Other' in every group expresses other level.

(3) Forward Looking Function is expressed by the first parts of APs and (non-AP) primary single acts.

(4) Backward Looking Function is expressed by the second parts of APs and non-AP responses.

5. Conclusions and Further Work

We currently use self-written software for annotating our corpus. The annotator sees a dialogue acts tree in the left window pane and a dialogue text in the right one. (S)he fixes a place in the dialogue and chooses a suitable dialogue act in the tree. Thus the process of annotating proceeds manually at the moment. A detailed annotation manual accompanies the software, dialogue acts definitions and annotation examples help and simplify the annotation process. Our next goal is to write a program for automatic recognition of dialogue acts. Morphological and syntactic features can be found that make it possible to recognise dialogue acts. Some acts (e.g. conventions) can be given as lists.

Some modules for the automatic processing of Estonian are already available – morphological analyser and generator, syntactic analyser, text-to-speech synthesiser. Some other modules are under work. Our dialogue system will bring together all these modules. Our current work is a step toward a dialogue system that interacts with the user in Estonian following norms and rules of human-human communication.

Acknowledgement

This work was supported in part by the Estonian Science Foundation under Grant No. 5685.

References

Allwood, J. & Ahlsen, E. & Björnberg, M. & Nivre, J. (2001). Social activity and communication act-related coding. In J. Allwood (Ed.), Gothenburg Papers in Theoretical Linguistics 85. Dialog Coding – Function and Grammar. Göteborg Coding Schemas (pp.1--28), Goteburg.

Bunt, H. (1999). Dynamic Interpretation and Dialogue Theory. In M.M. Taylor & F. Neel & D.G. Bouwhuis (Eds.). The Structure of Multimodal Dialogue II (pp. 139--166). John Benjamins Publishing Company, Philadelphia/Amsterdam.

Dybkjær, L. (2000). MATE Deliverable D6.2. Final Report http://mate.nis.sdu.dk/about/deliverables.html (variant 20.03.2004).

Edwards, J. (1995). Principles and alternative systems in the transcription, coding and mark-up of spoken discourse. In G. Leech & G. Myers & J.Thomas (Eds.), Spoken English on Computer. Transcription, Mark-up and Application (pp. 19--34).. London: Longman.

Ford, C. E. & Thompson S. A. (1996). Interactional units in conversation: syntactic, intonational and pragmatic resources for the management of turns. In E. Ochs & E. Schegloff & S. Thompson (Eds.), Interaction and Grammar (pp. 134--184). Cambridge UP

Hennoste, T. (2003a). Suulise eesti keele uurimine: korpus. In Keel ja Kirjandus 7, 481--500.

Hennoste, T. (2003b). Question-answer adjacency pair relations in information dialogues: Estonian case. In P. J. Henrichsen (Ed.), Nordic Research on Relations between Utterancies. Proceedings of the NordTalk Symposium at CMOL (CBS) December 2002. Copenhagen Working Papers in LSP 3-2003 (pp. 171--185).

Hennoste, T. & Koit, M. & Rääbis, A. & Strandson, K. Valdisoo, M. & Vutt, E. (2003a). Directives in Estonian Information Dialogues. In V. Matousek & P. Mautner (Eds.),Text, Speech and Dialogue. 6th International Conference TSD 2003 (pp. 406--411). Springer.

Hennoste, T. & Koit, M. & Rääbis, A. & Strandson, K. Valdisoo, M. & Vutt, E. (2003b). Developing a Typology of Dialogue Acts: Some Boundary Problems. In Proceedings of the 4th SIGdial Workshop on Discourse and Dialogue, Sapporo , 5-6 July, 2003 (pp. 226-235).

Hutchby, I. & Wooffitt, R. (1998). Conversation Analysis. Principles, Practices and Applications. Polity Press.

Jokinen, K. & Hurtig, T.& Hynnä, K. & Kanto, K. & Kaipainen, M. & Kermanen, A. (2001). Self-Organizing Dialogue Management. In Proceedings of NLPRS Workshop, Tokyo.

Levinson, S.C. (1982). Pragmatics. Cambridge UP.

Sacks, H., & Schegloff, E. & Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. In Language 50 (4), 696--735.

Schegloff, E. & Jefferson, G. & Sacks, H. (1977). The preference for self-correction in the organization of repair in conversation. In Language 52 (2), 361-382.

Sinclair, J.M. & Coulthard, R.M. (1975). Towards of Analysis of Discourse: The English used by Teachers and Pupils. London: Oxford UP.

Stolcke, A. & Coccaro, N. & Bates, R. & Taylor, P. & Van Ess-Dykema, C. & Ries, K. & Shriberg, E. & Jurafsky, D. & Martin, R. & Meteer, M. (2000). Dialogue act modeling for automatic tagging and recognition of conversational speech. In Computaional Linguistics 26 (3), 339–373.

Stenström, A.-B. (1994). An Introduction to Spoken Interaction. London and New York: Longman.

Ten Have, P. (1999). Doing Conversational Analysis. Sage.

Valdisoo MN & Vutt EV & Koit ME. (2003). On a method for designing a dialogue system and the experience of its application. In Journal of Computer and Systems Sciences International, 42(3), 456-464.

Appendix 1. Transcription marks

falling intonation	point
fall not to low	comma
raising intonation	?
short interval (max 0.2 sec)	(.)
timed interval	(2.0)
begin of overlap	[
end of overlap	]
latching at end of utterance	word=
latching at beginning	=word
drawling	::
stress	` at the beginning of the stressed syllable
glottal cut off	do-
in-breath	.hhh
item in doubt	{text}
unreachable text	{---}

Appendix 2. EDiC Typology of Dialogue Acts

Names of acts are originally in Estonian. Below one can find their translations into English. Some acts have more than one translation variant yet.

The name of an act consists of two parts: 1) an abbreviation of two or three letters: the first two letters give abbreviation of the name of act-group in Estonian (e.g. KK = kontakti kontroll 'contact control'); the third letter is used only for adjacency pair acts - the first (E) or second (J) part of an AP act; 2) full name of the act

I Adjacency pair acts
1.1 Dialogue managing acts
1.1.1 Fluent communication
1.1.1.1 Conventional

Greeting
RIE: Greeting
RIJ: Greeting

ClosingGoodbye
RIE: ClosingGoodbye
RIJ: ClosingGoodbye

Wish
RIE: Wish
RIJ: Gratitude, Grateful
RIJ: Wish

How-are-you
RIE: How-are-you
RIJ: How-are-you

Thanking
RIE: Thanking
RIJ: Thanking Here you are

Please
RIE: Please
RIJ: Thanking

Apology
RIE: Apology
RIJ: Apology

Introduction/presentation
RIE: Introduction/presentation
RIJ: Introduction/presentation
RIJ: Evaluation

Summons
RIE: Summons
RIJ: Answer

Pre-closing
RIE: Pre-closing
RIJ: Accept
RIJ: Reject

Conventional: other
RIE: other
RIJ: other

1.1.1.2. Topic change
TVE: initiation
TVE: other
TVJ: accept
TVJ: reject
TVJ: other

1.1.2 Problem solving
1.1.2.1 Other-initiated repair
PPE: reformulation
PPE: clarification
PPE: non-understanding
PPE: other
PPJ: repair
PPJ: other

1.1.2.2 Contact control
KKE: initiation
KKE: other
KKJ: affirmation/confirmation
KKJ: other

1.2 Information acts
1.2.1 Directives
DIE: request
DIE: proposal
DIE: offer
DIE: wait/hang on
DIE: other
DIJ: giving information
DIJ: missing information
DIJ: refusal
DIJ: doubting/unconfident/maybe
DIJ: agreeing/accepting
DIJ: disagreeing
DIJ: restricted agreeing
DIJ: action
DIJ: adjournment/postponement/deferral/holding before answer
DIJ: other

1.2.2 Question
KYE: closed yes/no question
KYE: open yes/no question
KYE: alternative question
KYE: wh-question
KYE: offering/proposing answer
KYE: specifying
KYE: adjusting the conditions of answer
KYE: other
KYJ: yes
KYJ: no
KYJ: agreeing no
KYJ: other yes/no answer
KYJ: alternative answer: one
KYJ: alternative answer: both
KYJ: alternative answer: third choice
KYJ: alternative: negative
KYJ: alternative: other
KYJ: action
KYJ: giving information
KYJ: missing information
KYJ: refusal
KYJ: adjournment /postponement /deferral/holding before answer
KYJ: alternate
KYJ: doubting/unconfident/maybe
KYJ: other

1.2.3 Opinion
SEE: assertion, argument
SEE: opinion
SEE: other
SEJ: agreeing/accepting
SEJ: reject
SEJ: partial/limited accept
SEJ: refusal
SEJ: other

II Non-AP acts
2.1 Dialogue managing acts
2.1.1 Fluent communication
2.1.1.1 Conventional
RY: remission
RY: Introduce/present/acquaint
RY: recognition
RY: contact
RY: call
RY: other

2.1.1.2 Responses
VR: continuer: evaluative
VR: continuer: neutral
VR: acknowledgement: evaluative
VR: acknowledgement: neutral
VR: change of state: evaluative
VR: change of state: neutral
VR: bounder/delineate: evaluative
VR: bounder/delineate: neutral
VR: repair evaluation
VR: other

2.1.2 Problem solving
2.1.2.1 Repair
PA: self repair
PA: other repair
PA: other

2.2 Information acts
2.2.1 Primary single acts
YA: advance note
YA: narration
YA: promise
YA: statement
YA: border-mark
YA: rhetorical question
YA: rhetorical answer
YA: quotation/reported
YA: other
YA: non-verbal
YA: un-interpretable

2.2.2 Additional information
IL: specification
IL: explication
IL: account/justification
IL: inference
IL: conclusion
IL: emphasise
IL: softening

IL: assessment
IL: other

Appendix 3. Dialogue Annotation: an Example

```
1. ((kutsung))
((summons))  RIE: SUMMONS
2. V: ´Estmar=´info,
R: Estmar=information RIJ: ANSWER | RY:
INTRODUCING
3. ´Leenu=kuuleb
Leenu speaking RY: INTRODUCING
4. tere
good afternoon RIE: GREETING
5. H: tere.
C: good afternoon RIJ: GREETING
6. (0.8) {Leenu.}
Leenu YA: OTHER
7. V: ja?
R: yes VR: CONTINUER: NEUTRAL
8. (0.5)
9. H: rotilõks.
C: rat trap DIE: REQUEST
10. (1.8)
11. V: kuidas?
R: pardon PPE: NON-UNDERSTANDING | KYE: WH-
QUESTION
12. H: rotilõks.
C: rat trap PPJ: REPAIR | KYJ: GIVING
INFORMATION | DIE: REQUEST
13. (0.8)
14. V: jah, rotilõks
R: yes, rat trap VR: ACKNOWLEDGEMENT:
NEUTRAL | PPE: NON-UNDERSTANDING
15. H: {´andke ´kõik.}
C: give all PPJ: REPAIR | DIE: REQUEST
16. V: kuidas?
R: pardon PPE: NON-UNDERSTANDING | KYE: WH-
QUESTION
17. H: {´kõik kus ma saan ´osta.}
C: all where I can buy PPJ: REPAIR | KYJ:
GIVING INFORMATION | DIE: REQUEST
18. (2.2)
19. V: e ma arvan et seda saab teha
majapidamistarvete ´kauplustest.
R: I suppose that you can do this in
household commodities shops SEE: OPINION
20. (0.5) saan teile neid ´pakkuda,
soovite.
may I suggest you any if you wish DIE:
OFFER
21. (1.0)
22. H: {---} ´majapidamistarvete
kauplusest.
C: in household commodities shop PPE:
CLARIFICATION | KYE: QUESTION OFFERING
ANSWER
23. V: ma arvan ´küll jah.
R: yes I think so PPJ: REPAIR | KYJ: YES
24. (...)
25. H: mitte ´hiirelõksu.
C: not mouse trap IL: SPECIFICATION
26. mul on kurat=ee noh (.) päris ´võikad
´elukad.
I have bloody hideous creatures IL: ACCOUNT
27. (0.5)
28. V: jah ma ´usun.
R: yes I believe VR: ACKNOWLEDGEMENT:
NEUTRAL
29. (0.5) m:a arvan et neid saab ka
´majapidamistarvete ´kauplusest,
I suppose that you can get them too in
household commodities shop SEE: OPINION
30. ma ei oska teile küll midagi ´muud
´pakkuda.
I can not suggest anything else SEE:
ASSERTION
31. H: noh?
C: well DIJ: AGREEING | DIE: REQUEST
32. (0.8)
33. V: jaa üks hetk?
R: yes one moment DIJ: POSTPONEMENT BEFORE
ANSWER
34. (...) ee ´Meltoni ´äri.
Meltoni äri DIJ: GIVING INFORMATION
35. H: jah?
C: yes VR: CONTINUER: NEUTRAL
36. (0.5)
37. V: neli kolm null, (.) kaheksa viis
kaheksa.
R: four three zero, eight five eight DIJ:
GIVING INFORMATION
38. (1.2) ´Eritreid.
Eritreid DIJ: GIVING INFORMATION
/---/
56. (...) ´Ristiku kauplus, (1.0) neli
seitse üks, (.) kolm viis kolm.
Ristiku shop, four seven one, three five
three. DIJ: GIVING INFORMATION
57. (1.0)
58. H: no aitab.
C: that'll do DIE: REQUEST
59. (.)
60. V: jaa=palun?
R: you are welcome DIJ: AGREEING | RIE: YOU
ARE WELCOME
61. H: aitäh.
C: thank you RIJ: THANKING
```

The client (H) is calling with the aim to get information about shops where he can bay a rat trap.

The dialogue begins with an introduction where the participants greet each other. Then the client presents his request (row 9). The request is too fuzzy, and the information officer cannot recognise the act. A series of repairs follow to solve communication problems in co-operation. Turn 17 shows that client understands the problem, he repairs his previous information, and makes a new request. The officer gives an answer in form of belief, and makes an offer. The answer causes a problem for the client. The reason is that he is sure that rat traps cannot be bought in household commodities shops, there are only mouse traps there (row 25). The officer is convinced that she is right (row 29). She offers the client to finish this topic and to return to giving information (row 30). Client accepts the offer and makes a request to get information (row 31). A new sub-part begins where information is given /received. The hearer is using continuer repeatedly (e.g. row 35). This sub-part finishes with a directive – proposal to finish the subject (row 59). The partner accepts the proposal, and the conversation ends.

This example demonstrates four large parts of the dialogue – two formal (conventional beginning and ending) and two informal parts. The first one deals with solving a communication problem, and the other gives direct information.

WinPitch Corpus

A Text to Speech Alignment and Analysis Tool for Large Multimodal Corpora

Philippe Martin

UFRL Université Paris 7 Denis Diderot
92, Avenue de France, 75013 Paris, France
philippe.martin@linguist.jussieu.fr

Abstract

WinPitch Corpus is an innovative software program for computer-aided alignment of large corpora. It provides a method for easy and precise selection of alignment units, ranging from syllable to whole sentences in a hierarchical storing system of aligned data. The method is based on the ability to link visually a target with the perception of corresponding speech sound played back. Listening to slower speech, an operator is able to select with a mouse click a segment of text corresponding to the speech sound perceived, and generate by this action bidirectional speech-text pointers defining the alignment. This method has the advantage on emerging automatic processes to be effective even for poor quality speech recordings, or in case of speakers' voice overlap. A recent version of the software handles multimedia files and is capable to display the corresponding video streams at slower speed.

1. Introduction

Large spontaneous speech corpora are becoming essential for the continuing development of fundamental research in linguistics and language engineering. Commercially or experimentally available automatic recognition software is usually delivering poor performance for ordinary non scripted speech. This is mainly due to the use, at the higher stages of the recognition, of syntactic models built from read speech data and not from spontaneous discourse analysis. Likewise, speech synthesis from text should strongly benefit from the emergence of prosodic models inferred from real life data. A better understanding of prosodic interactions linked to spontaneous conversations would also bring spectacular improvement to second language teaching, which is still based for a great part on read material.

These examples demonstrate the importance of spontaneous speech corpora at both levels of text and speech. Although put a speech corpus together seems to be a simple task per se, involving speech to text transcription and alignment, this ceases to be the case when the volume of data becomes large. Problems inherent to transcription, even for well recorded speech data, are not trivial (Blanche-Benveniste, 2002), and the "manual" transcription and alignment of just one hour data becomes quickly cost prohibitive, even with the use of modern signal editing software. The development of adequate and user friendly tools is thus essential for the elaboration of large speech corpora.

2. Text to speech alignment

Text to speech alignment establishes a bi-univocal relationship between units of speech and units of text. In its simplest implementation, each unit of text (be syllable, word, syntagms or sentence) receives a temporal index corresponding to the time position of its equivalent in the sound file. When this process is achieved, an operator can select an aligned unit of text and listen to the corresponding speech segment. Acoustical analysis of the speech sound, such as melodic curve and spectrogram, can also be displayed at the same time. Conversely, the selection of a speech segment will highlight the corresponding segment of text, in its orthographic or phonetic transcription.

Text to speech alignment is frequently used in multimedia language learning software, where the user can easily listen to the sound corresponding to a specific word or sentence merely by clicking on the appropriate text segments. Other important applications are found in fundamental research in phonetics, in synthesis by rule developments, or in speech recognition validation tests. In all these domains, elaboration of large corpora is becoming an essential activity for fundamental research and language engineering.

2.1 Spectrographic alignment

Most students in experimental phonetics having received some training in spectrographic reading are capable of segmenting speech sounds accurately with the visual cues displayed by acoustical analysis. Fricative consonants and stops present easily recognizable graphic features, as do oral vowels. However, nasals followed by vowels however are harder to segment, but the overall process generally leads to a high quality, although labor intensive, segmentation.

In reality, units of text and units of speech cannot correspond exactly, as phonetic and phonological units are defined by human perception on the axis of continuous articulatory transitions, whereas speech signal segments are defined as physical entities. Alignment and segmentation can therefore be only approximations, and the physical time limits of speech segments must be positioned somewhere during articulatory transitions of speech sounds.

Nevertheless, transitions between speech sounds are used in some automatic segmentation algorithms, such as (Cosi, 1997). These processes utilize the spectral discontinuities on the time axis, and give acceptable results in otherwise visually clear cases (and will fail for a sequence such as nasal consonant followed by a vowel for example).

As in all processes of acoustical analysis of speech, reliability of this approach relies on the validity of the hypotheses implied in the process, the most important one assuming the presence of only one sound source in the signal. Background noise and other speech sources will of course lead to disappointing results in the segmentation.

2.2 Automatic alignment with hidden Markov models

Another automatic or semi-automatic method for text to speech alignment utilizes algorithms used for speech recognition (often based on parameters obtained by a HMM Hidden Markov Model applied to the speech data). This approach appears as a subset of the general speech recognition problem, since the text as already known. The limits of speech segments are then found from a phonetic or orthographic transcription (Talin and Wightman 1994, Fohr, Mari, et Haton 1996).

Although attractive, systems based on automatic speech recognition suffer from the same limitations as speech recognition itself: somewhat high error rate (15% to 20%) without the use of a syntactic model at a higher stage of the process, and the difficulty to train the system with the speaker voices (which must be made from samples of the corpus itself). Again, good results are to be obtained only if the speech signal presents a high signal to noise ratio, and if the voices to align do not differ too much from the models used to train the algorithm. Overlapping speech constitutes of course a very difficult case for these systems.

2.3 Automatic alignment by synthesis

Another automatic method of alignment proceeds by comparing the spectral variations of the signal along time with another speech signal, generated by a speech synthesizer fed by the text to align (Malfrère and Dutoit, 2000). The advantage here stems from the fact that it is easier to align successive spectra on two distinct time scale (by dynamic warping) than segment sounds from automatic recognition of segments.

The limits of this approach however are similar to those of the use of HMM: poor signal to noise level, deviant characteristics of speaker's voices (compared to the models used in the synthesis process) and again speech sources overlapping constitute difficult problems to this approach.

2.4 Limits of automatic alignment

In summary, automatic text to speech alignment processes present the following recurrent limitations:

1. Their performance depends on speaker's voice characteristics, which cannot be to different from the models implied in these methods;

2. The recording signal to noise ratio must be high enough to reduce the error rate to an acceptable level. The radio and TV broadcasts generally meet this requirement, but spontaneous speech recordings made in various public environments (street, public transportation, etc.) present hardly these characteristics. Echo in the speech signal is another aggravating factor;

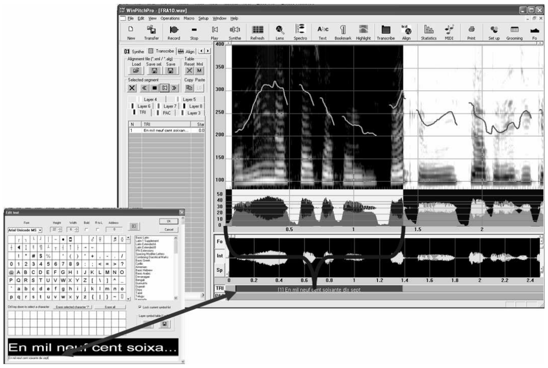
3. Speaker's voices overlapping, frequently found in spontaneous dialogs constitutes an aggravating factor.

Furthermore, use of recordings for phonetic and general linguistic research requires an acceptable quality of the recording itself, such as a good frequency response curve and a low phase distortion.

All these considerations seem to indicate that a human operator is presently required to obtain a reliable text to speech alignment. All the problems mentioned above are then transferred to the operator, who, with appropriate and well ergonomically designed tools, should performed better than by correcting manually the errors made by an automatic system of alignment.

3. Alignment and transcription

Text to speech alignment can be executed in two modes, depending if the text preexists or not. In the first mode, the text must be created, and the operator proceeds by selecting segments of speech in sequences (which can be played back at reduced speed to enhance intelligibility) and type the corresponding text perceived, either orthographically or in phonetic transcription (WinPitch Corpus allows the use of any font defined in Unicode). During this process, a database is automatically updated, and will be later saved directly in Excel® or XML format.



WinPitchPro - [FRA10.wav]

N	TRI	Star
1	En mil neuf cent soixan...	0.0

[1] En mil neuf cent soixante dix sept

En mil neuf cent soixa...

En mil neuf cent soixante dix sept

File Edit View Operations Macro Setup Window Help
New Transfer Record Stop Play Synthe Refresh Lens Spectro Text Bookmark Highlight Transcribe Align Statistics MIDI Print Set up Grooming Fo
Synthe Transcribe Align
Alignment file (*.xml / *.alg) Load Save sel. Save
Table Reset Mnl
Selected segment
Copy Paste
Layer 4 Layer 5 Layer 6 Layer 7 Layer 8 TRI PAC Layer 3
400 350 300 250 200 150 100
50 40 30 20 10 0
0.5 1 1.5 2
Fo Int Sp
0 0.2 0.4 0.6 0.8 1 1.2 1.4 1.6 1.8 2 2.2 2.4
TRI
Edit text
Font Height Width Bold R to L Address
Arial Unicode MS
OK Cancel
Lock current symbol list
Erase all

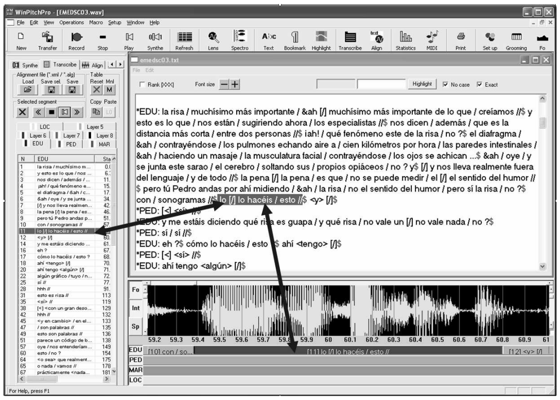
Figure 1: *Simultaneous transcription and alignment. The user sequentially defines segment of speech and enters the corresponding text.*

Unicode characters can be entered 1) merely by clicking on the appropriate cell of the Unicode table (the user can setup a special purpose table of up to 48 selected characters to avoid navigating into many pages of the Unicode character map) or 2) directly from the keyboard using Microsoft® Multilanguage interface available in many target languages (mandarin, Thai, Tibetan, etc.).

4. Computer assisted alignment

Experimental studies have shown that coordination between visual spotting of words and positioning of a mouse on a computer screen could be obtain by slowing down speech playback by a suitable factor, depending on the size of the text object to spot (larger chunks of text require less processing time, and thus allow a faster speech playback rate in the process). WinPitch Corpus assisted text to speech alignment is based on this principle.

In the second mode of operation, the text preexists, and is displayed dynamically in a window while the corresponding speech sound is played back at a slower speed (which can be adjusted continuously on the fly). At each identification of a speech unit to segment and align (be a syllable, word, syntagms or sentence), the operator clicks with the computer mouse on the text segment perceived. The program records the position of the cursor on the text window (which defines the end of the text segment to align) and the time of the click (remapped on the real time scale of the speech wave). This process generates continuously a database of pointers linking segments of text and segments of speech.


WinPitchPro - [EMEDSC03.wav]
File Edit View Operations Macro Setup Window Help
New Transfer Record Stop Play Synthe Refresh Lens Spectro Text Bookmark Highlight Transcribe Align Statistics MIDI Print Set up Grooming Fo
emedsc03.txt
File Edit
Rank [XXX] Font size
Highlight
No case
Exact
Fo
Int
Sp
59.2 59.3 59.4 59.5 59.6 59.7 59.8 59.9 60 60.1 60.2 60.3 60.4 60.5 60.6 60.7 60.8 60.9 61
EDU
PED
MAR
LOC
For Help, press F1

*EDU: la risa / muchísimo más importante / &ah [/] muchísimo más importante de lo que / creíamos //$ y esto es lo que / nos están / sugiriendo ahora / los especialistas //$ nos dicen / además / que es la distancia más corta / entre dos personas //$ ¡ah! / qué fenómeno este de la risa / no ?$ el diafragma / &ah / contrayéndose / los pulmones echando aire a / cien kilómetros por hora / las paredes intestinales / &ah / haciendo un masaje / la musculatura facial / contrayéndose / los ojos se achican ...$ &ah / oye / y se junta este sarao / el cerebro / soltando sus / propios opiáceos / no ? y$ [/] y nos lleva realmente fuera del lenguaje / y de todo //$ la pena [/] la pena / es que / no se puede medir / el [/] el sentido del humor //$ pero tú Pedro andas por ahí midiendo / &ah / la risa / no el sentido del humor / pero sí la risa / no ?$ con / sonogramas //$ lo [/] lo hacéis / esto //$ <y> [/]$
*PED: [<] <sí> //$
*EDU: y me estáis diciendo qué risa es guapa / y qué risa / no vale un [/] no vale nada / no ?$
*PED: sí / sí //$
*EDU: eh ?$ cómo lo hacéis / esto //$ ahí <tengo> [/]$
*PED: [<] <sí> //$
*EDU: ahí tengo <algún> [/]$

[10] con / so... [11] lo [/] lo hacéis / esto // [12] <y> [/]

Figure 2: *Assisted alignment by slowing down speech playback. At each mouse click on a unit of text perceived at slower speed (top right window), bidirectional pointers are generated automatically between the corresponding speech segment (bottom right window) and a database (left window).*

Various tools are provided to backtrack, fine tune speech segment limits (with the help of a displayed spectrogram), dynamically modify limits of overlapping voices, etc.

4.1 Slowing down speech

Variable rate speech playback is the central engine of the assisted text to speech aligner. It uses a modified version of the PSOLA algorithm (Moulines et Charpentier, 1990) and allows high quality re-synthesis of natural speech. This quality is strongly dependent on precise pitch marking, and therefore on reliable fundamental frequency analysis. Errors in pitch marking (missing markers, double markers) induce an echo effect due to the misalignment of pitch chunks added in the PSOLA algorithm. Fo estimation is obtained by the spectral comb method (Martin, 1980), and the speed playback factor can vary from 7 to ½ (speech played back at double speed). This rate is dynamically adjustable by the user while the alignment is processed, allowing operations on very large files and the continuous speed adjustment as required by the operator.

4.2 Automatic layer assignment

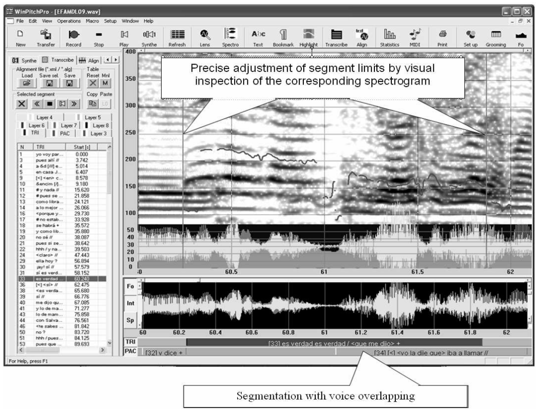
Preexisting text can be organized (following a simple convention for naming speakers turns) so that segments are automatically assigned to their corresponding layers. The user does not have to worry about speaker's turns while aligning, as the program will put segmented text in the appropriate layer assigned to each speaker (8 layers are presently available, but future extension will provide for unlimited number of speakers).

4.3 Automatic layer assignment

Preexisting text can be organized (following a simple convention for naming speakers turns) so that segments are automatically assigned to their corresponding layers. The user does not have to worry about speaker's turns while aligning, as the program will put segmented text in the appropriate layer assigned to each speaker (8 layers are presently available, but future extension will provide for unlimited number of speakers).

4.4 Fine tuning and speaker overlap

Once the assisted alignment session ends, the program displays automatically the text under the corresponding speech segments, represented by their acoustic analysis (spectrogram, intensity and melodic curves, waveform). The user can then adjust precisely the segments by dragging their limits with the mouse, with the help of visual inspection of the corresponding spectrogram (or other available acoustical information). Overlapping segments can then easily and precisely be defined.


WinPitchPro - [FAMDL09.wav]
File Edit View Operations Macro Setup Window Help
New Transfer Record Stop Play Synthe Refresh Lens Spectro Text Bookmark Highlight Transcribe Align Statistics MIDI Print Set up Grooming Fo
Fo
Int
Sp
TRI
PAC

Precise adjustment of segment limits by visual inspection of the corresponding spectrogram

Segmentation with voice overlapping

Figure 3: *Fine tuning of speech segments limits with the help of a simultaneously displayed spectrogram (which allows precise segmentation in case of speaker's overlapping).*

4.5 Text and segmentation correction

Limits defined by clicking on the last unit (word) of segments can be easily edited (end of segment markers can be inserted, deleted and moved inside the text with simple keyboard commands.

4.6 Prosodic morphing

The PSOLA engine is also used for prosodic morphing of any part of the speech data, allowing the user to modify with very simple graphic commands the prosodic parameters or the original data (fundamental frequency, intensity, segment duration, pauses).

4.7 Multimodal alignment with WinPitch Corpus

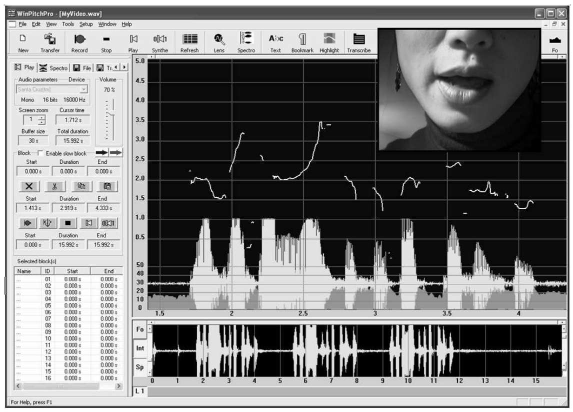

WinPitchPro - [MyVideo.wav]
File Edit View Tools Setup Window Help
New Transfer Record Stop Play Synthe Refresh Lens Spectro Text Bookmark Highlight Transcribe Fo
Play Spectro File
Audio parameters Device Volume
Mono 16 bits 16000 Hz
Screen zoom Cursor time
Buffer size Total duration
Block Enable slow block
Start Duration End
Selected block(s)
Name ID Start End
Fo
Int
Sp
For Help, press F1

Figure 4: *Simultaneous display of acoustic and video information of a multimedia file.*

WinPitch Corpus can read and process most multimedia file formats, and offers the same capabilities when displaying video with the sound data. Alignment of multimodal files can be done the same way as with sound files. The program also allows real time recording and analysis of speech data. This feature is particularly useful to monitor recording from the real time spectrographic display. Problems such as low signal level, echo, saturation can immediately be noticed and fixed.

4.8 A large scale application

WinPitch Corpus has been extensively used in the elaboration of the C-ORAL-ROM project (C-Oral-Rom, 2004), in which about 1.2 million word were aligned with spontaneous speech recorded in various conditions.

5. Conclusion

Computer assisted text to speech alignment, as being much faster than conventional and automatic methods, thanks to the use of ergonomically well designed tools, allows the development of large corpora of various languages, which in turn will, among other benefits, induce a better comprehension of the relationship between syntax and intonation. The method of slowing down speech used in WinPitch Corpus permits to handle corpora of very variable sound quality, which would prevent the use of automatic methods based on speech recognition.

In this software, alignment is executed by an operator clicking on the units to segment (syllables, words, syntagms, sentences) displayed on a program window while the corresponding speech sounds is played back at reduced speed. This slow down process allows the psychometrical coordination needed for the process, which can be executed in one pass et does not require any particular expertise in phonetics. The process is of course much faster than the traditional approach where a trained phonetician has to align the sound segments one by one by shifting a time window along the speech signal. It is also more reliable that emerging automatic methods, which require reasonably good quality recordings and exclude too large variations or voice and pronunciations characteristics of the speakers.

Numerous functions of WinPitch Corpus allow precise adjustment a segment limits, as well as the exact definition of overlapping segments of speech, thanks to simultaneously displayed acoustic information (spectrogram, waveform, intensity and melodic curves). The user can also edit the transcription on the fly, which any font provided in Unicode. Phonetic transcription, syntactic tagging, and any other information of interest can be easily added on one of the 8 layers available for transcription.

Other currently available software (such as Praat, 2004) do not presently have a set of features optimized for text to speech alignment of large corpora, which make their use only suitable for a limited amount of data.

References

Blanche-Benveniste, C. (2002) « Réflexions sur les transcriptions de corpus de français parlé », *Revue PArole*, 22-23-24, 2002, pp. 91-117.

C-Oral-Rom (2004) "Integrated reference corpora for spoken romance languages" (IST-2000-26228 - Shared-cost RTD) http://lablita.dit.unifi.it/coralrom/

Cosi, P. (1997) "SLAM v1.0 for Windows : a Simple PC-Based Tool for Segmentation and Labelling", *Proc. Of ICSPAT-97, Int. Conf. On Signal processing Applications and Technology*, San Diego, CA, Sept. 1997, pp. 1714-1718.

Fohr, D., Mari, J.-F. et Haton, J.-P. (1996) « Utilisation des modèles de Markov pour l'étiquetage automatique et la reconnaissance de BREF80 », *Actes des XXIèmes Journées d'Etude sur la Parole*, Avignon, pp. 339-342.

Malfrère, F. et Dutoit, T. (2000) « Alignement automatique du texte sur la parole et extraction de caractéristiques prosodiques », in *Ressources et évaluation en ingénierie des langues*, Chibout, Mariani, Masson, Néel ed., De Boeck et Larcier, Paris, pp. 541-552.

Martin, J.C. and Kipp, M. (2002) "Annotating and Measuring Multimodal Behaviour – Tycoon Metrics in the Anvil Tool", *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC'2002), Las Palmas, Canary Islands, Spain, pp. 29-31 may 2002.*

Martin, Ph. (1981) "Mesure de la fréquence fondamentale par intercorrélation avec une fonction peigne", *Actes des XIIèmes Journées d'Etude sur la Parole*, Montréal, juin 1981.

Martin, Ph. (2000) « Peigne et brosse pour Fo : Mesure de la fréquence fondamentale par alignement de spectres séquentiels », *Actes des XXIIIèmes XXI Journées d'Etude sur la Parole*, Aussois, France, juin 2000, pp. 245-248.

Moulines, E. & Charpentier, M. (1990) "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones", *Speech Communication*, Vol 9, pp. 453-467.

Talin, D. and Wightman, C.W. (1994) "The Aligner: Text-to-Speech Alignment using Markov Models and a Pronunciation Dictionary", *Second ESCA/IEEE Workshop on Speech Synthesis*, pp. 89-92.

Praat, 2003. http://www.praat.org

WinPitch (1996, 2004) http://www.winpitch.com

Automatic Annotation of Speech Corpora for Prosodic Prominence

F. Tamburini and C. Caini

University of Bologna
f.tamburini@cilta.unibo.it ccaini@deis.unibo.it

Abstract

This paper presents a study on the automatic detection of prosodic prominence in continuous speech, with particular reference to American English, but with good prospects of application to other languages. Perceptual prosodic prominence is supported by two different prosodic features: pitch accent and stress. Pitch accent is acoustically connected with fundamental frequency (F0) movements and overall syllable energy, whereas stress exhibits a strong correlation with syllable nuclei duration and mid-to-high-frequency emphasis. This paper shows that a careful measurement of these acoustic parameters, as well as the identification of their connection to prosodic phenomena, makes it possible to build automatic systems capable of identifying prominent syllables in utterances with performance comparable with the inter-human agreement reported in the literature without using any kind of information apart the acoustic parameters derived directly from speech waveforms.

Introduction

The study of prosodic phenomena in speech is a central topic in language investigation and it is generally agreed that it represents one of the main streams for improving the performances of speech processing systems. Speakers tend to focus the listener's attention on the most important parts of the message by means of prosodic markers and signal its correct interpretation by means of intonation, pauses, prominences, ...

Automatic Speech Recognition systems can take advantage of software modules devoted to prosody management enhancing the global classification performances (Hastie, *et al.* 2001; Hieronymous, *et al.* 1992; Shriberg & Stolcke, 2001), as well as can do Automatic Speech Understanding systems (Beckman & Venditti, 2000; Nöth, *et al.* 2000; Shriberg, *et al.* 1998). Prosodic modules can enhance the fluency and adequacy of automatic speech-generation systems (Bulyko, *et al.* 1999; Portele & Heuft, 1997; Wightman, *et al.* 2000) and it may be extremely useful for solving ambiguities in natural language parsing (Hirschberg & Avesani, 2000; Warren, 1996).

One of the most interesting applications of automatic techniques for handling prosodic phenomena is that of language resource annotation, such as prosodically tagged speech corpora, both for research purposes and for language teaching (Beckman & Venditti, 2000; Campione & Veronis, 1998; Hirst, 2001). In this field the request of prosodically annotated resources is increasing and the difficulty and the prohibitively high costs for a manual production often limit, and have limited, the design of such resources.

One of the most important prosodic features is prominence. “Prominence is the property by which linguistic units are perceived as standing out from their environment” (Terken, 1991). Following Beckman's (1986) phonological view, further developed by other scholars, for example Bagshaw (1993; 1994), syllables that are perceived as prominent either contain a pitch accent or are stressed. On the acoustic/phonetic side, the accomplishment of such features is strictly correlated with particular behaviour of acoustic parameters, either considered as single features or, more commonly, as combinations of them. As well as the works already cited, there are many other studies suggesting that some of the main acoustic correlates of prominence are pitch movements, overall syllable energy, syllable duration and spectral emphasis (Sluijter & van Heuven, 1996; Streefkerk, 1997; Taylor, 2000).

This paper presents a study on the relationships between prosodic prominence and acoustic features with the aim of designing a system for the automatic detection of prominence in speech using only acoustic/phonetic parameters and cues. Our work has been developed restricting the information sources to the utterance waveform, avoiding any other resource that might not be always available, it would certainly be expensive to build, and permanently bound the system to one specific language. The method we will present do not rely on additional phonetic information, such as phone labelling and/or utterance transcriptions as well as the use of complex techniques requiring heavy training phases on manually annotated data such as hidden Markov models, neural networks or similar methods. This study performs an analysis of the correlation between prominence and a set of acoustic features to identify the best acoustic correlates of prosodic prominence and use such information to build a system capable of identifying prosodic prominence in continuous speech.

Despite the quantity and quality of studies on this topic, it seems that the automatic and reliable detection of prosodic prominence, without providing phonetic information, such as utterance transcription or training corpora composed of segmented utterances, is still an open question.

The data set used in our experiments is a subset of the DARPA/TIMIT acoustic-phonetic continuous speech corpus, consisting of thousands of transcribed, phone-segmented and aligned sentences of American English. In this study, the TIMIT annotations are used only for testing and measuring system performance, not as additional information for the prominence detection algorithms.

The rest of the paper is organised as follows. Section 2 presents syllable nuclei identification for duration measures. Section 3 outlines the computation procedures for the other parameters involved in prominence detection. Section 4 presents a study about the combination and relationships of these acoustic features to identify prosodic features such as pitch accents, stress and prominence, and section 5 discusses the automatic detector of prosodic prominence presented in this study. Section 6 draws the conclusions of the work, comparing and discussing the results obtained with the literature examined.

Syllable nuclei identification and duration measures

The linguistic theories of prosodic prominence mentioned above agree in considering syllable duration as one of the fundamental acoustic parameters for detecting syllable stress, certainly in American English, but also in many other languages. Unfortunately, the automatic segmentation of the utterance into syllables is a challenging task; even defining the syllable concept in continuous speech is often misleading. Resyllabification phenomena and ambisyllabic units contribute to making syllables an entity with fuzzy boundaries. Moreover a lot of studies have made clear that the main contribution of prominence to syllable lengthening is concentrated in the vocalic part of it, mainly increasing the syllable nucleus duration (Greenberg, *et al.* 2003; Silipo & Greenberg 1999; van Bergem, 1993; van Kuijk & Boves, 1999). The relevant conclusion, interesting for the present prominence study, is that we can reliably replace the syllable duration measure, necessarily affected by large measurement error whenever obtained by automatic procedures, with the measure of syllable nucleus duration as in (Jenkin & Scordilis, 1996; Waterson, 1987), which can be automatically obtained with a higher accuracy level.

To reliably identify the syllable nuclei in the utterance and measure their duration we applied a modified version of the convex-hull algorithm proposed by Mermelstein (1975) to the utterance energy profile. This was computed after band-pass filtering (300-900 Hz) the speech-samples, as suggested in (Howitt, 2000), to filter out energy information not belonging to vowel units which forms the syllable nucleus. The segmentation points were restricted the to the ones derived from the algorithm proposed by Andre-Obrecht (1988) that detects regions of spectrally quasi-stationary speech in the utterance. The duration parameter is then normalised by considering the mean duration of the syllable nuclei in the utterance. This is a standard technique for Rate-Of-Speech normalisation, described, for example, in (Neumeyer, 1996).

All the subsequent measurements of acoustic parameters will be referred to the syllable-nucleus intervals computed using the method described above.

Other acoustic parameters

Fundamental frequency (F0) contour

The extraction of F0 contour, or pitch contour, is another demanding task. Most of the complexity of pitch extraction process resides in candidate selection and post-processing optimisations. Stops and glitches often tend to distort the contour, introducing spurious changes in the profile. Other typical problems in obtaining a correct pitch profile derive from octave jumps, where the pitch frequency computed by the algorithm, in a specific speech frame, is found to be double (or half) the correct pitch frequency. A post-processing procedure to smooth out such variations is often required in order to obtain more reliable results.

To extract pitch contour we used the ESPS get_f0 program, based on the algorithm presented in (Talkin, 1995), that is considered in literature the best pitch-tracking method. To obtain a continuous profile, the post-processing phase involves octave-jump removers and profile smoothers, derived from the ones proposed in (Bagshaw, 1994), applied both at voiced interval and sentence level, and a final interpolation between voiced regions.

Energy measures

Differently from the parameters presented in the previous subsections, the third acoustic parameter considered here, namely the syllable nucleus energy (or intensity), can be automatically computed in various ways without any particular difficulty. Here we refer to RMS energy, defined as:

$$E_j^{RMS} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} a_{ji}^2}$$

where N is the number of samples per frame and $a_{1..N}$ are the speech samples in the j-th frame. The nucleus energy is successively normalised to the mean nucleus energy over the utterance. This reduces the energy variation across different utterances and different speakers.

In the recent literature, and in particular in the influential work of Sluijter & van Heuven (1996), it has been claimed, that mid-to-high frequency emphasis is a useful parameter in determining stressed syllables. To verify this hypothesis, each nucleus segment spectrum was divided into three bands, making use of band-pass FIR filters, namely from 0 to 300 Hz, from 300 to 2200 Hz and from 2200 to 4000 Hz. The RMS energy of each segment/band pair was successively computed. By examining the distributions of prominent and non-prominent syllable energies in the frequency bands considered the two bands 0-300 Hz and 2200-4000 Hz show a clear overlapping between prominent and non-prominent syllable distributions, while the central band from 300 to 2200 Hz exhibits a clear separation between the two classes. These quantitative results confirm the dependence of syllable prominence to vowel mid-to-high frequency emphasis, the frequency band where the main vowel formants reside. Thus, agreeing with the hypothesis suggested by Sluijter & van Heuven, with a view to identifying stressed syllables, we will consider that the spectral emphasis is measured by the energy of this specific frequency band.

Prosodic parameters

This section examines the prosodic quantities, stress, pitch accent and prominence, that are the object of the study, and their acoustic correlates. As already mentioned in the introduction, syllables that are perceived as prominent either contain a pitch accent or a stress accent, or both. Thus, prominence can be described by relying on two different prosodic parameters, stress and pitch accent, both sufficient to identify a prominent syllable, but none of them necessary to mark a syllable as prominent.

The data used in the following sections are derived from the TIMIT corpus and every syllable was manually classified as prominent or non-prominent. It emerges quite clearly in the following subsections that being able to classify these syllables with respect to the two different phenomena, namely stress and pitch accent, instead of classifying them with respect to prominence, would have been preferable, for both the qualitative analysis that we will carry out in this section and the design of the final detector. Unfortunately it is very difficult for humans to distinguish between stress and pitch accents when listening to an utterance. It is only possible to reliably perceive if a syllable is prominent or not with respect to the surrounding context. This lead to a certain degree of overlapping in the study of the involved phenomena.

Stress

The main correlates of syllable stress reported in literature are syllable duration and energy (Bagshaw, 1993; 1994; Streefkerk, 1997; 1999). On this topic Sluijter & van Heuven (1996) have introduced a further refinement, confirmed also in later studies (Heldner, 2001), casting some light on the exact correlation between the different acoustic parameters. Their studies clearly divided the two phenomena involved in supporting prominence perception, pointing out that the most reliable correlates of syllable stress are syllable duration and mid-to-high frequency emphasis.

In figure 1 two sets of prominent and non-prominent syllables are depicted as a function of both log-normalised nucleus duration and log-normalised RMS energy in the 300 to 2200 Hz band. There is a clear evidence supporting Sluijter & van Heuven's ideas: prominent syllables exhibit a longer duration and greater energy in the vowel mid-to-high-frequency band.

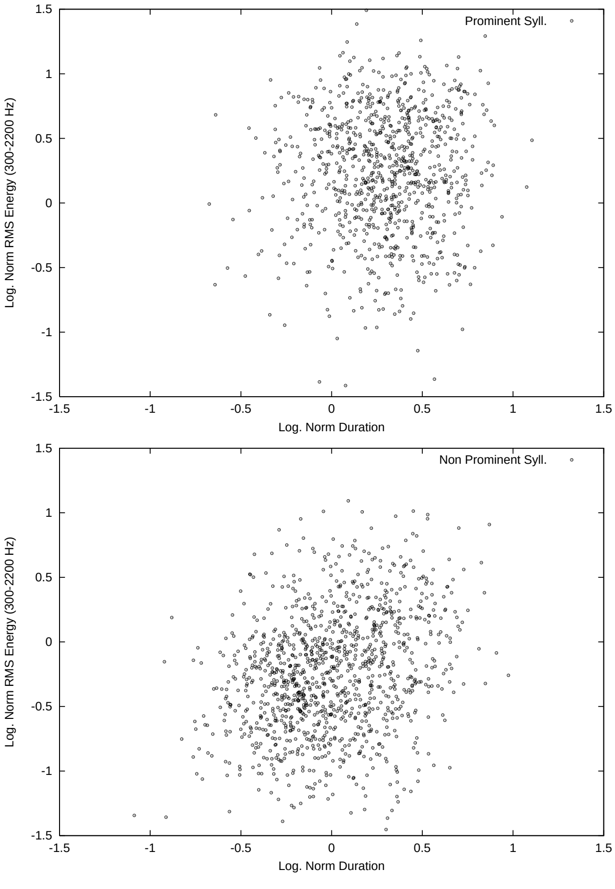


Figure 1: Prominent and non-prominent syllables as a function of log-normalised nucleus duration and log-normalised nucleus energy in the spectral band from 300 to 2200 Hz.

Pitch accent

There is a long tradition of studies dealing with intonation profiles and pitch accents (Pierrehumbert, 1980; Beckman, 1996; Campione & Veronis, 1998). Unfortunately, the categorisations introduced in these studies, as well as the famous ToBI labelling scheme (Pitrelli, *et al.* 1994), appears to be difficult to implement in an automatic system. Taylor (1995; 2000) proposed a different view of intonation events. Starting from a rise/fall/connection (RFC) model, he defined a set of parameters capable of uniquely describing events in the pitch contour. This set, called TILT, consists of five parameters defined as:

$$tilt_{amp} = \frac{|A_{rise}| - |A_{fall}|}{|A_{rise}| + |A_{fall}|} \qquad tilt_{dur} = \frac{D_{rise} - D_{fall}}{D_{rise} + D_{fall}}$$

$$tilt = \frac{|A_{rise}| - |A_{fall}|}{2 \cdot (|A_{rise}| + |A_{fall}|)} + \frac{D_{rise} - D_{fall}}{2 \cdot (D_{rise} + D_{fall})}$$

$$A_{event} = |A_{rise}| + |A_{fall}| \qquad D_{event} = D_{rise} + D_{fall}$$

where A_{rise}, A_{fall}, D_{rise}, D_{fall} are respectively the amplitude and the duration of the rise and fall segments of the intonation event.

Our implementation for the extraction of the pitch shape follows Taylor’s proposal. The F0 contour is first converted into an intermediate RFC model. To do that the contour is segmented into frames 0.025 second long and the data in each frame are linearly interpolated using a Least Median Squares method. Then every frame interpolating line is classified as rise, fall or connection, depending on its gradient, as suggested in (Taylor, 1993), and subsequent frames with the same classification are merged into one interval. The duration and amplitudes of the rise and fall sections are measured to finally derive the TILT parameter set and assign them to the intonational events in the F0 contour extracted from the RFC representation. As described by Taylor (2000), an intonational event that can be considered as a good candidate for pitch accent exhibits a rise followed by a fall in the pitch profile. There are different degrees of such profiles and, in general, rise sections appear to be more relevant for prominence.

Sluijter and van Heuven suggested that the pitch accent can be reliably detected by using overall syllable energy and some measure of pitch variation. As far as pitch variation is concerned, the event amplitude, which is one of the TILT parameters, can be considered as a proper measure, being the sum of the absolute amplitude of the rise and fall sections of a generic intonational event. However, a further refinement can be obtained by multiplying the event amplitude (A_{event}) by its duration (D_{event}) to reduce the significance of spike errors. Figure 2 shows a plot of prominent and non-prominent syllables as a function of overall syllable nucleus energy and the product of the event parameters on a log scale. Qualitatively, a clear correlation emerges among these parameters when identifying prominent syllables.

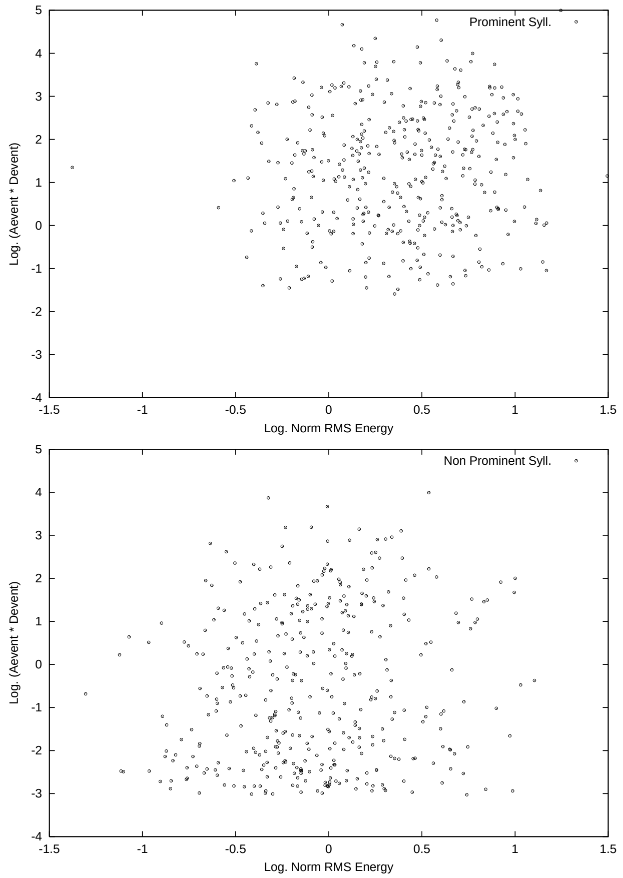


Figure 2: A plot of prominent and non-prominent syllables as a function of overall syllable nucleus energy and intonational event parameters.

Prominence

We have established some qualitative relationships between acoustic parameters and some prosodic quantities, in particular stress and pitch accent. As suggested in the literature and confirmed by our experiments, metrical stress strictly depends on syllable nuclei duration and energy in a specific spectral band: the longer the duration and the higher the energy in the syllable nucleus, the greater the stress perception. In the same way, high overall nucleus energy and wide pitch movement produce the strongest pitch accent.

The two phonological concept, namely stress and pitch accent, considered in this study as in intermediate level, will help us to relate the acoustic/phonetic parameters with prominence. As we will see in the next section, the relationships between these phenomena and the qualitative observations described before will be useful in defining the behaviour of the prominence detector.

Prominence detector

According to Taylor (2000), all the prosodic parameters involved in prominence study should be considered as continuous quantities, avoiding any kind of categorisation. On the other hand, for testing the reliability of an automatic system, hand-tagged categorical data have to be used. For these reasons we chose to describe and manage the prosodic parameters presented above as continuous values, and successively introduce some provisional categorisations to compare the behaviour and performance of the automatic process with the hand-tagged data.

Bearing in mind the qualitative relationships among the acoustic parameters outlined above, it seems possible to combine them properly to build a "prominence function" able to derive a continuous value of prominence directly from the acoustic features of every syllable nucleus. Our proposal for such a function is:

$$\mathrm{Prom}^i = \max\left\{ en^i_{300-2200} \cdot dur^i,\ en^i_{ov} \cdot (A^i_{event} \cdot D^i_{event}) \right\}$$

where $en_{300\text{-}2200}$ is the energy in the 300-2200 Hz frequency band, *dur* is the nucleus duration, en_{ov} is the overall energy in the nucleus and A_{event} and D_{event} are the parameters derived from the TILT model. All the parameters refer to a generic *i*-th syllable nucleus in the utterance examined. Although the *Prom* function definition is somewhat arbitrary and tentative, it has a rationale, as it was derived in such a way as to mathematically express the fact that a prominent syllable is usually stressed or pitch accented or both and that these prosody parameters can be successfully derived from the acoustic parameters that appear in the formula. This continuous approach is fully justified by considering that the classification into prominent or not prominent cannot be carried out, at least in an optimal way, if the context of the neighbouring syllables is neglected.

As pointed out before, to evaluate the system by comparing it with hand-tagged data, it is necessary to introduce some kind of categorisation, by considering the prominence level of the syllable compared with its neighbours. Following Terken perspective, identifying prominent syllables implies the search for the local maxima of the *Prom* function defined above. Therefore, in our classifier the prominence value of every syllable nucleus is compared with the two neighbours and, if it represents a maximum, then it is considered prominent. However, it is neither impossible nor rare for consecutive syllables to be prominent, for example whenever two successive monosyllabic words are both emphasised. The two syllables would certainly present a different "level" of prominence, but, in a dichotomic-classification perspective (prominent or non-prominent), levels of prominence cannot be taken into account. To partially overcome this problem, the peak picking algorithm was enhanced to tackle this relatively frequent case. Whenever two subsequent syllables differ only by 15% of their prominence value, the test is performed by ignoring the neighbours with similar prominence and by considering instead the next syllable nuclei. Moreover, syllables that have a very high prominence value, greater than 70% of the maximum peak in the utterance, are also considered as prominent, independently of the context. A plot of prominence function for a sentence taken from the TIMIT corpus is shown in figure 3.

Numerical results show that by making use of the *Prom* function and the enhanced peak picking method described above, it is possible to design a reliable prominence detector. The model was tested using a subset of TIMIT utterances, composed of 5708 syllables taken from 384 utterances spoken by 51 different speakers of American English. The prominence detector correctly classified 80.80% of the syllables as either prominent or non-prominent, with an insertion rate of 8.22% (false alarms) and a deletion rate of 10.98% (missed detections).

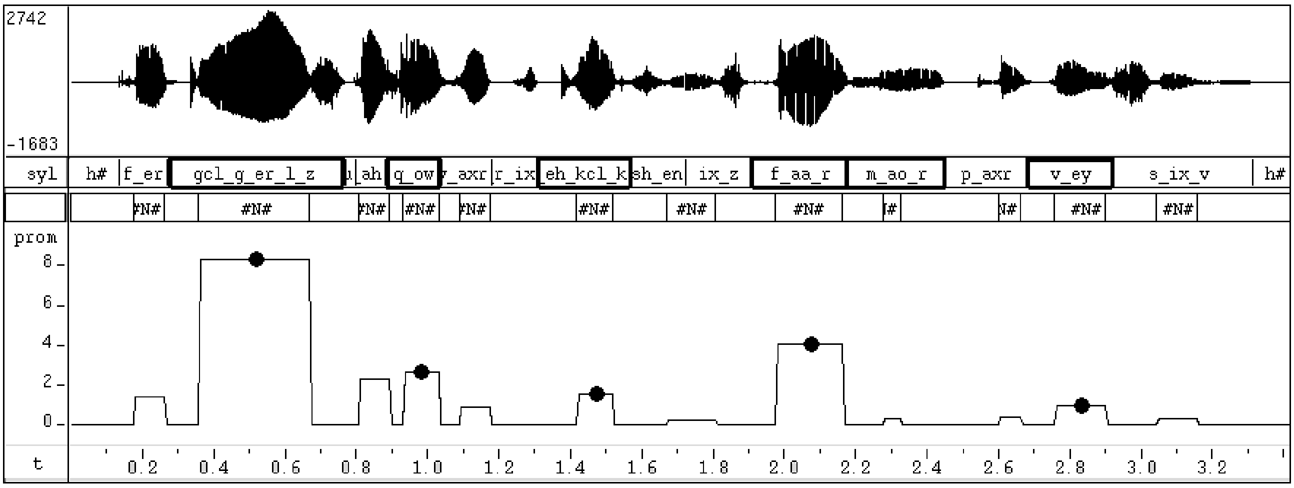


Figure 3: Prosodic prominence function values for the utterance "For girls the overprotection is far more pervasive". Proceeding from the top, we have: the waveform plot, the syllable segmentation (only for comparison purposes), the syllable nuclei as detected by the system (marked by #N#), and finally the prominence values for every nucleus identified by the segmentation procedure. The prominent nuclei, as identified by the automatic system, are marked by a dot on the function profile, while prominent syllables, as classified by a human listener, are indicated by a thick box in the syllable segmentation track ("syl").

As pointed out before, this is an unsupervised system, thus there is no need for any training phase.

Conclusions and discussion

It is widely accepted in the literature that inter-human agreement, when manually tagging prominence in American English continuous speech, is around 80-90% according to the different number of prominence classes chosen for the annotation (Pickering, *et al.* 1996; Jenkin & Scordilis, 1996). The prominence detector presented here exhibits an overall agreement of 80.80% with the data manually tagged by a native speaker, without exploiting any information apart from acoustic parameters derived directly from the utterance waveform. As these results are in the same range of those obtained by human taggers, the prominence detector can be seen as a possible alternative to manual tagging for building large resources of speech annotated with prominence information.

Previous studies tend to use different approaches. Bagshaw (1993) built a prominence detection system for computer aided pronunciation teaching, thus using the utterance transcription to guide the segmentation and the detection process obtaining a 61.6% of agreement with human-tagged data, that is much less than the one obtained by the systems presented in our work. Jenkin & Scordilis (1996) implemented and compared three different system for prominence detection, all based on theoretical models that require training phases. The most performant is based on neural networks and achieved a correct classification on 81-84% of cases. All systems presented in their study require a complex training phase and additional tagged data to do it. Similar considerations can be made about the results obtained by Wightman & Ostendorf (1994) with their system, based on a model that uses decision trees similar to a discrete HMM and an Automatic Speech Recognition module. The model is trained using maximum likelihood estimation and achieves 85-86% of correct classification when applied to prominence detection. All these methods make an heavy use of additional information such as phonetic and orthographic transcriptions, segmentation information or ASR systems.

It would be interesting to test the validity of our approach with different languages. Theoretically, different languages involve different combinations of acoustic parameters or different weightings among them, but the methods presented here should be easily adapted to cope with these inter-language variations. A study in this direction is presently under way considering the Italian language.

References

Andre-Obrecht, R. (1988). A New Statistical Approach for the Automatic Segmentation of Continuous Speech Signals. IEEE Transactions on Acoustics, Speech and Signal processing, 36(1), 29-40.

Bagshaw, P.C. (1993). An investigation of acoustic events related to sentential stress and pitch accents, in English. Speech Communication, 13(3-4), 333-342.

Bagshaw, P.C. (1994). Automatic prosodic analysis for computer-aided pronunciation teaching. PhD thesis, University of Edinburgh.

Beckman, M.E. (1986). Stress and non-stress accent. Dordrecht, Holland: Foris Publications.

Beckman, M.E. & Venditti, J.J. (2000). Tagging prosody and discourse structure in elicited spontaneous speech. In Proc. of Science and Technology Agency Priority Program Symposium on Spontaneous Speech (pp. 87-98), Tokyo.

Bulyko, I., Ostendorf M. & Price P. (1999). On the Relative Importance of Different Prosodic Factors in Improving Speech Synthesis. In Proc. of ICPhS '99 (pp. 81-84), San Francisco.

Campione, E. & Veronis, J. (1998). A multilingual prosodic database. In Proc. of ICSLP '98 (pp. 3163-3166), Sydney.

Greenberg, S., Carvey, H., Hitchcock, L. & Chang S. (2003). The Phonetic Patterning of Spontaneous

American English Discourse. In Proc. of ISCA/IEEE Workshop on Spontaneous Speech Processing and Recognition, Tokyo.

Hastie, H.W., Poesio, M. & Isard, S. (2001). Automatically predicting dialog structure using prosodic features. Speech Communication, 36(1-2), 63-79.

Heldner, M. (2001). Spectral Emphasis as an Additional Source of Information in Accent Detection. In Proc. of Prosody 2001: ISCA Tutorial and Research Workshop on Prosody in Speech Recognition and Understanding (pp. 57-60), Red Bank, NJ.

Hironymous, J.L., McKelvie, D. & McInnes, F.R. (1992). Use of acoustic sentence level and lexical stress in HSMM speech recognition. In Proc. of ICASSP '92 (pp.225-227), San Francisco, California.

Hirshberg, J. & Avesani, C. (2000). Prosodic disambiguation in English and Italian. In A. Botinis (Ed.), Intonation (pp. 87-95), Kluwer Academic Publisher.

Hirst, D.J. (2001). Automatic analysis of prosody for multilingual speech corpora. In E. Keller, G. Bailly, J. Terken & M. Huckvale (Eds.), Improvements in Speech Synthesis. Chichester, UK: Wiley.

Howitt, A.W. (2000). Automatic Syllable Detection for Vowel Landmarks. PhD Thesis, MIT.

Jenkin, K.L. & Scordilis M.S. (1996). Development and comparison of three syllable stress classifiers. In Proc. of ICSLP '96 (pp. 733-736). Philadelphia.

Mermelstein, P. (1975). Automatic segmentation of speech into syllabic units. Journal Acoustical Society of America, 58(4), 880-883.

Neumeyer, L., Franco, H., Weintraub, M. & Price, P. (1996). Automatic Text-Independent Pronunciation Scoring of Foreign Language Student Speech. In Proc. of ICSLP '96 (pp. 1457-1460). Philadelphia.

Nöth, E., Batliner, A., Kießling, A., Kompe, R. & Niemann, H. (2000). VERBMOBIL: The Use of Prosody in the Linguistic Components of a Speech Understanding System. IEEE Transactions on Speech and Audio Processing, 8(5), 519-532.

Pickering, B., Williams, B. & Knowles, G. (1996). Analysis of transcriber differences in SEC. In Knowles G., Wichmann, A. & Alderson, P. (Eds), Working with speech (pp. 61-86). London: Longman.

Pierrehumbert, J.B. (1980). The phonetics and phonology of English intonation. PhD thesis, MIT.

Pitrelli J., Beckman M. & Hirschberg J. (1994). Evaluation of prosodic transcription labeling reliability in the ToBI framework. In Proc. of ICSLP '94 (pp. 123-126). Yokohama, Japan.

Portele, T. & Heuft, B. (1997). Towards a prominence-based syntesis system. Speech Communication, 21(1-2), 61-72.

Shriberg, E., Baker, R., Stolcke, A., Taylor, P., Jurafsky, D., Ries, K., Coccaro, N., Martin, R., Meteer, M. & van Ess-Dykema, C. (1998). Can Prosody Aid the Automatic Classification of Dialog Act in Conversational Speech? Language and Speech, 41(3-4), 439-487.

Shriberg, E. & Stolcke, A. (2001). Prosody modeling for automatic speech recognition and understanding. In Proc. of ISCA Workshop on Prosody in Speech Recognition and Understanding (pp. 13-16), Red Bank.

Silipo, R. & Greenberg, S. (1999). Automatic transcription of prosodic stress for spontaneous English discourse. In Proc. of ICPhS '99 (pp. 2351-2354), San Francisco.

Sluijter, A. & van Heuven, V. (1996). Acoustic correlates of linguistic stress and accent in Dutch and American English. In Proc. of ICSLP' 96 (pp. 630-633), Philadelphia.

Streefkerk, B.M. (1997). Acoustical correlates of prominence: a design for research. In Proc. of Inst. of Phon. Sciences, University of Amsterdam, 20, 131-142.

Streefkerk, B M., Pols L.C.W. & ten Bosch L.F.M. (1999). Acoustical features as predictors for prominence in read aloud Dutch sentences used in ANN's. In Proceedings of Eurospeech '99 (pp. 551-554), Budapest.

Talkin, D. (1995). A robust algorithm for pitch tracking (RAPT). In W.B. Kleijn & K.K. Paliwal (Eds.), Speech coding and synthesis (pp. 495-518). New York: Elsevier.

Taylor, P.A. (1993). Automatic Recognition of Intonation from F0 Contours using the Rise/Fall/Connection Model. In Proc. of Eurospeech '93 (pp. 789-792), Berlin.

Taylor, P.A. (1995). The rise/fall/connection model of intonation. Speech Communication, 15(1-2):169-186.

Taylor, P.A. (2000). Analysis and Synthesis of Intonation using the Tilt Model. Journal Acoustical Society of America, 107 (3):1697-1714.

Terken, J. (1991). Fundamental frequency and perceived prominence. Journal Acoustical Society of America, 89 (4):1768-1776.

van Bergem, D. (1993). Acoustic vowel reduction as a function of sentence accent, word stress and word class on the quality of vowels. Speech Communication, 12(1), 1-23.

van Kuijk, D. & Boves L. (1999). Acoustic characteristic of lexical stress in continuous telephone speech. Speech Communication, 27(2), 95-111.

Warren, P. (1996). Prosody and Parsing: an introduction. Language and Cognitive Processes, 11 (1/2):1-16.

Waterson, N. (1987). Prosodic phonology: The theory and its application to language acquisition and speech processing. Grevatt and Grevatt: Great Britain.

Wightman, C.W. & Ostendorf, M. (1994). Automatic labelling of prosodic patterns. IEEE Transaction on Speech and Audio Processing, 2 (4): 469-481.

Wightman, C.W., Syrdal, A.K., Stemmer, G. & Conkie, A. (2000). Perceptually based automatic prosody labeling and prosodically enriched unit selection improve concatenative text-to-speech synthesis. In Proc. of ICSLP 2000 (pp. 71-74), Beijing.

Towards Dynamic Corpora

Daan Broeder, Hennie Brugman, Nelleke Oostdijk+, Peter Wittenburg
Max-Planck Institute for Psycholinguistics, Nijmegen, The Netherlands
daan.broeder@mpi.nl
+Dept. of Language and Speech, University of Nijmegen

Abstract

In this paper, the idea of a "Dynamic Corpus Environment" is taken up. After we specify and discuss the functional design of such an environment, the idea is further elaborated upon by illustrating its implementation as it is envisaged, for example, for the Spoken Dutch Corpus and the DOBES Corpus.

1. Introduction

Traditionally, language resources were created predominantly for a specific purpose within the context of a specific project. This implied the careful selection and definition of all relevant parameters including the structure of the corpus. Characteristically, the creation of language resources was perceived of as a finite undertaking: once finished, it was not normally expected that the corpus would be extended with new data, let alone with types of annotations that were not already present in the corpus. Examples of such static corpora are abundant. For English, they include the Brown and LOB corpora and their modern counterparts (Freiburg Brown and LOB), the COLT corpus, and the Penn Treebank. Extending a corpus that was originally conceived as a static corpus is often not unproblematic: the addition of data may disturb the usually well-balanced corpus design. Moreover, for corpora that were designed essentially as text corpora (if only because for a long time the addition of media data was hampered due to the fact that storage was rather expensive), the linking of audio and /or video to other representations of the data is very labour-intensive.

Often a corpus does not offer any specific exploitation tools or a dedicated environment, and the user is expected to create his own scripts. While for the purpose of corpus exploitation it is helpful to have a corpus tool environment, for the incorporation of new data such an environment is an absolute necessity. Some corpora, including for example CHILDES [1], do offer a way of adding new data and new annotation tiers. However, relations between tiers are limited to certain predefined types, and then the relation is on a serial token-to-token basis.[1]

A further issue in the context of corpus exploitation is that if the addition of data is permitted, the exploitation environment should allow for creating and exploiting sub-corpora. In the definition of these sub-corpora the metadata play a crucial role as they immediately give access to the parameters that were operants in the implementation of the corpus design.

2. A Dynamic Corpus Environment

The idea of a dynamic corpus environment is that a corpus should be made flexible in so far that new projects could be set up and carried out with the aim of adding further data and/or new types of information to the already existing data. While static corpora have the advantage that upon completion it is clear what data and annotations are involved, a big disadvantage is that from the day they are completed they are at risk of becoming obsolete as they fail to incorporate novel insights and do not mirror the latest state of the art. Also upon the discovery of errors or omissions, corpus maintenance is severely hindered if there is no environment in which, preferably under controlled conditions, corrections can be carried through. In case additional annotations exist, they exist independently alongside the original corpus and are generally speaking not easily accessible. In such a configuration, they cannot be used to enrich the original data.

Over the past decades, we have seen that time and again with new corpora also new exploitation tools were created. In terms of investment this procedure cannot be seen as advantageous. Also as developers keep re-inventing what was already available in a variety of tools for use in a number of highly idiosyncratic environments, users are withheld an optimal environment with a maximum of functionality and ease of use. The annotation and exploitation of corpora could be made much more transparent if there were to be an environment in which corpora could be annotated, maintained and extended, and in which a suite of tools and resources were available for the user to apply.

2.1. Requirements

A necessary condition for a dynamic corpus is that one should be able to replace or add more material while maintaining the design, i.e. it should be possible to extract the original corpus or a sub-corpus with a different design. To this end access to available metadata is essential. It is the metadata that provide the key to making selections for inclusion in a sub-corpus or sub-corpora from a larger corpus. Thus it becomes possible to meet the user's needs with customized (sub-)corpora, At the same time, allowing for all data (original data and additions to these) to be made accessible through one and the same

[1] The authors are aware that the new CHILDES XML annotation format may alleviate these restrictions.

environment, holds the advantage that the user can at all times exploit further data and /or annotations. From the perspective of resource management and the investments involved it amounts to a (possibly huge) optimalisation.

The addition of whole subcorpora containing also different types of resources and annotation-types demands that the metadata set used to describe the resources is general enough to describe all resources in the corpus but can also be more specific and specialized for accurate descriptions of these added subcorpora. The IMDI metadata set [2] that we propose to use for our implementation allows for extensions to a core set of descriptors.

What is also required and what constitutes a greater challenge is to be able to add new annotation types that interrelate both with each other and with the existing types. For this a very general annotation model is needed. Adding new types of primary data with associated annotations can be another problem. This is the case when for instance video is added or annotation of two-dimensional objects is required. These points do not only influence the design of the annotation model but also demand flexibility and configuration possibilities of the exploitation tool environment.

Certainly not the least important component of a dynamic corpus environment, would be a versioning system. A user may decide on his own authority to change or replace some existing annotations in a private installation of the corpus. This is something he may later regret. Therefore, reverting such an act should be easy. The same problem may occur at the installation site of the central maintainer of a corpus where the scientific justification for some changes may be doubted after some period of time. Although standard versioning systems are available [6] special care must be taken to provide for the ways in which the annotations are interrelated. The way private installations of a corpus, with modifications and additions, should relate to an authorized central one is another interesting challenge for such a versioning system. The possibility to synchronize changes via the Internet is something that looks feasible and will be explored.

Clearly not all users would be allowed to perform all possible actions and therefore an authentication/ authorization mechanism is also needed The requirements for such an authorization mechanism depend of course on the corpus availability and usage scenario's. At one end of the spectrum, for publicly available corpora that can be accessed on the Web there is no need for any protection at all while commercially interesting data that needs to be accessible via the Web will be protected by strong authentication and watermarking. Another desirable scenario is when groups of cooperating researchers share access to each others data and administrate access rights amongst themselves without burdening the people responsible for the corpus storage. A protection scheme is in all cases demanded for update and extension of corpora. The most flexible scheme is that privileged users can after automatic validation of the formats of the annotation files and resources add them to the corpus. The fact that a versioning mechanism is in place may make this policy more acceptable since any change can be revoked afterwards. In a research environment it is likely that people will need to access archives at different sites in the world. It will be very helpful if they could authenticate themselves using the same identification with each of these archives. So a central authentification site or sites exchanging authentication information would be desirable. Discussions amongst archives to come to such a system are under way.

3. Proposed Testbeds

3.1. The Spoken Dutch Corpus

For the Spoken Dutch Corpus (Corpus Gesproken Nederlands, CGN) [3] it was decided at the start of the project that a corpus exploitation tool ("COREX") [4] should be created. Moreover, it was stipulated that the tool environment should be capable of importing into the corpus additional data, although of the same type and format as already existed in the corpus. When the CGN project came to and end (January 2004), a new phase started, viz. corpus maintenance[2], In this phase it would be possible to adopt a highly restrictive view in which maintenance would amount to and be limited to keeping copies of the annotation data and audio recordings available and distributing them to users. However, a more challenging approach is possible and will be pursued in which we shall attempt to create a tool environment that makes the CGN corpus a dynamic one.

In the process of creating a dynamic corpus several steps can be distinguished. First of all, it should be possible to insert new recordings with associated annotations. This should be easy to do, while the addition of new material should not in any way affect existing structures. COREX accesses the CGN corpus resources through IMDI metadata records that bind together the recordings with their annotations and metadata in resource bundles. New IMDI metadata records can easily be added to the existing ones. IMDI metadata also defines the different browsable access paths to these resource bundles. By means of new tools from the IMDI tool environment that will be incorporated in the CGN tool environment, these trees can be extended with new resource bundles and new tree structures can be added.

There are no limitations to extending the corpus structure with new data. The IMDI tools allow users to define their own view on the corpus by defining new private nodes that allow access to any sub-corpus of the original corpus and any new additions, i.e. an extension to the total corpus does not influence the working environment of the individual user, which is very important. The existing browse, metadata - and content

[2] . Of course, it is also possible to look upon the correction of errors in the production process as maintenance, but that is not intended here. In the case of the CGN, parts of the corpus were released in the form of intermediate release at more or less regular intervals and errors in a release were corrected in a next release. Thus the correction of errors effectively coincided with the production of new data.

search components of COREX are independent of any extensions and will continue to function. [3]

In order to add new types of annotations within the current annotation model used by CGN, it is required that they conform to the standard CGN XML annotation format. This is a limitation in the sense that all annotations can only be linked to tokens on the orthographic tier. Simple annotations of this type can be made visible in the COREX annotation viewer as a new annotation layer without any problems. Sometimes, however, these annotations are not all that simple and the annotations are given extra meaning through the use of attributes. This is already the case for the prosodic and phonetic annotations. Here we have to try and find a way to configure the annotation viewer in such a way that special attributes can be displayed in an appropriate manner. The same procedure applies when making these annotations searchable with or without special attributes. If we want more flexibility with respect to annotation structures, we may have to convert the annotations to a more general model and format like EAF [5]. This will require major adaptations of the annotation viewer but will not affect the metadata part.

Adding annotations that are interrelated to other annotations as is the case for all annotations in the CGN corpus puts high demands on validation of the ingested annotations. General automatic validation is feasible as far as compliance with the defining annotation format schema is concerned. Also some automatic checks may be done using specific knowledge of the CGN annotation model that is not captured by the annotations XML-Schema. All this helps guarding the consistency of the corpus but does not help judge the content of the new annotations. For that a human operator must take a sample an use methods appropriate for the annotation type. It would be helpful if the versioning mechanism would allow showing the updated corpus initially only to relevant users: "updater" and the "evaluator".

3.2. The DOBES corpus

The DOBES corpus (Documentation Bedrohter Sprachen) [7] is a corpus of field-linguistic multi-media resources which has been under construction over the last 4 years. At the moment it contains 700 hand-crafted IMDI sessions for video and audio files, mostly with accompanying annotation files. Requirements differ somewhat from the CGN corpus in that:

1. There is more emphasis on active and passive access through the Internet because it is being used and created by scattered groups of field-linguists all over the world. Also there will be demand for robust and trustworthy mechanisms where researchers are able to administrate access rights for the resources themselves. It is expected that parts of this functionality can be implemented in line with the policies advocated by DELAMAN [8], an organization of archives for endangered language and music data.

2. The annotations are currently in a format (EAF) where all annotations tiers are present in one single file. So after ingesting new annotations checking consistency between the different annotation tiers is less of a burden since we assume this was already done by the tool used to create the annotation file.

3. DOBES has a flexible corpus structure that is locally formed according to the views of the researcher responsible for that specific part of the corpus. So a requirement for DOBES is the possibility of creating different structures for different parts of the corpus.

4. Infrastructural aspects

We expect that in the near future passive exploitation, manipulation and addition of resources in corpora via the Internet will be commonplace. Tools that will enable users to do this can be (a) standard web-browsers or (b) specialist tools that will be able to do some extra validation of the resources and metadata to be ingested. Both ways have their specific advantages. Also, in both cases, these tools will present users a logical view on the corpus that decouples them from the actual physical storage of the data. This is required when we want to achieve a situation where we use a system of resource identifiers that is not dependent on the actual physical location of the resource via URLs such as is currently the state of affairs. Such a system of abstract resource identifiers has been discussed for Web resources in general intensively [9] since it solves the problem of dead links that occur so often when people move resources to another location. Currently there are two different implementations available: PURL [10] and The Handle System [11], we have chosen for the handle system for reasons of robustness and support from major organisations.

[3] Note, however, that after data has been added new index files need to be created to speed up search.

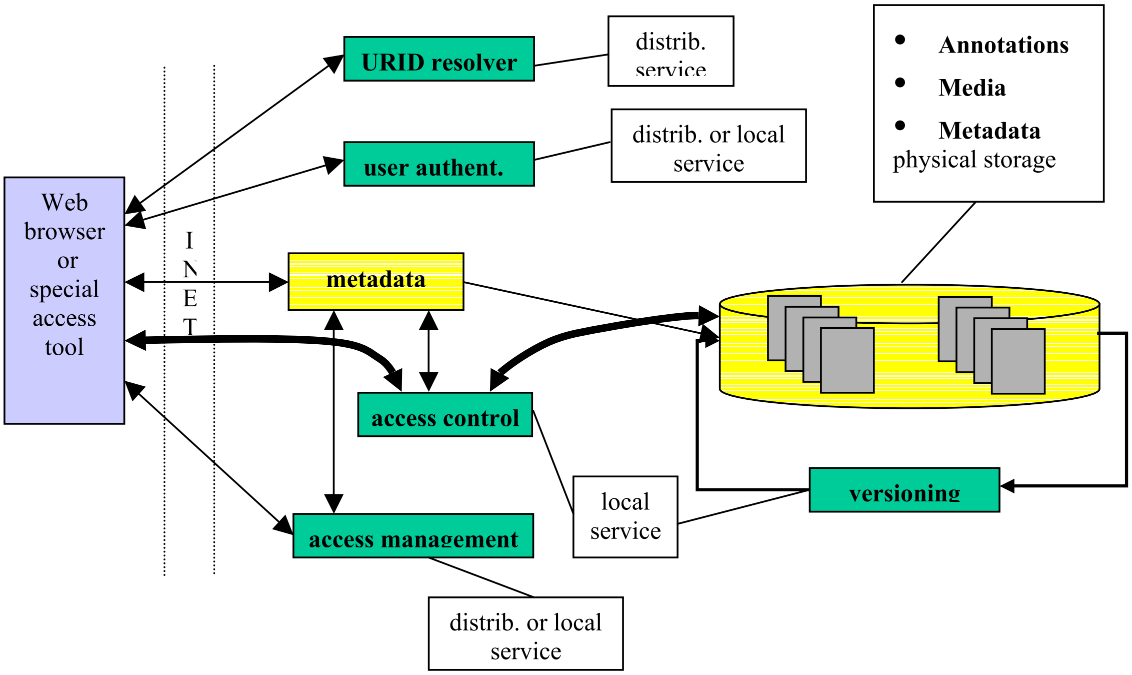


Figure 1 Components of a Dynamic Corpus Environment

5. References

[1] CHILDES: Child Language Data Exchange System; http://childes.psy.cmu.edu

[2] Isle Metadata Initiative IMDI, http://www.mpi.nl/IMDI

[3] Oostdijk, N. 2000. The Spoken Dutch Corpus. Outline and first evaluation. In: M. Gravilidou, G. Caravannis, S. Markantonatou, S. Piperidis, and G. Stainhaouer (eds.), Proceedings of the Second International Conference on Language Resources and Evaluation (LREC). 31 May-2 June 2000. Athens, Greece. Vol. 2: 887-894.

[4] Oostdijk, N. and D. Broeder. 2003. The Spoken Dutch Corpus and its Exploitation Environment. In *Proceedings of the 4th International Workshop on Linguistically Interpreted Corpora (LINC-03)*. 14 April, 2003. Budapest, Hungary.

[5] Eudico Annotation Format, H. Brugman, EMELD workshop on Digitizing & Annotating texts & field recordings, June 2003, http://emeld.org/workshop/ 2003/brugman-paper.html

[6] For example see: Concurrent Versions System CVS, http://www.cvshome.org/docs/manual/

[7] DOBES: http://www.mpi.nl/DOBES

[8] DELAMAN: http://www.delaman.org/

[9] OpenURL http://library.caltech.edu/openurl/

10] PURL http://purl.oclc.org/

[91] Handle System: http://www.handle.net/rfc/rfc3650.html