Working Papers in Hungarian Sociolinguistics
No. 3, January 1998

FROM CARDS TO COMPUTER FILES: PROCESSING THE DATA OF THE BUDAPEST SOCIOLINGUISTIC INTERVIEW

TAMÁS VÁRADI

Linguistics Institute, Hungarian Academy of Sciences
H-1250 Budapest, P.O.Box 19., Hungary

Working Papers in Hungarian Sociolinguistics
No. 3, January 1998

FROM CARDS TO COMPUTER FILES: PROCESSING THE DATA OF THE BUDAPEST SOCIOLINGUISTIC INTERVIEW

TAMÁS VÁRADI

Linguistics Institute, Hungarian Academy of Sciences
H-1250 Budapest, P.O.Box 19., Hungary

ACKNOWLEDGEMENT

The publication of this working paper has been supported by grant T 025997 of Országos Tudományos Kutatási Alap.

ISSN 1418-2823

ISBN 963 9074 09 8

Contents

Preface

The present volume aims to give an overview of the way the data collected in the Budapest Sociolinguistic Interview (BSI) is processed. It will be readily appreciated that a large scale project like the BSI which generated an enormous amount of highly delicate linguistic data[1] collected through an array of various tasks, poses a challenge to data processing. It was felt that such a complex undertaking was worth devoting a separate volume to it. The core of the present paper was presented at the "Workshop on Spoken Language Analysis" held on 27-28 May 1994, in Venice.

After a brief overview of the BSI data, the paper will give a survey of our evolving attempts to accomodate the BSI data. During its history of more than ten years now, the BSI database system has undergone a major revision. This was precipitated by rapid technological advancement and necessitated by certain flaws in the design of the data model. Despite its recognized shortcomings, the earlier database version is also described in some detail because we feel it may still serve a useful purpose in showing how a first attempt may be made. If the tone of presentation is felt overly self-critical, it was a deliberate policy to help anyone embarking on a similar enterprise to avoid the pitfalls.

The emphasis will be on the data model and data entry – in line with the practical concerns of the BSI projects so far. However, Chapter 5 discusses data retrieval including a description of the multimedia edition of a complete interview on CD-ROM. The paper ends with a discussion of future work. It contains suggestions as to how the technology developed for the CD-ROM needs further elaboration in order to make all the BSI data accessible in a common graphic interface through the World Wide Web. While the bulk of the paper deals with the part of the BSI data that was numerically coded and processed in the database, Appendix A contains a detailed guide to the codes and conventions used in the transcription of the quided conversation. Familiarity with this part will prove indispensible for any serious study of the transcripts.

Budapest, January 1998

Tamás Váradi

[1] A detailed description of the BSI project and a comprehensive account of the data are published in other volumes in this series (see Kontra & Váradi 1997 and Váradi 1998, respectively).

References

Kontra, Miklós & Tamás Váradi. 1997. *The Budapest Sociolinguistic Interview: Version 3* (Working Papers in Hungarian Sociolinguistics No 2, December 1997). Budapest: Linguistics Institute of the Hungarian Academy of Sciences. 54 pp.

Váradi, Tamás. 1998. *Manual of The Budapest Sociolinguistic Interview Data* (Working Papers in Hungarian Sociolinguistics No 4, February 1998). Budapest: Linguistics Institute of the Hungarian Academy of Sciences. 101 pp.

1 Overview of the Budapest Sociolinguistic Interview Data

1.1 Brief historical background

The Budapest Sociolinguistic Interview (BSI) project is a long-term sociolinguistic project aiming to provide solid empirical data about the language varieties spoken in Budapest. A large body of tape recorded data was collected in a carefully compiled sociolinguistic interview which was administered to a representative sample of Budapest speakers. The BSI data collection took place in two phases. After an initial test version, termed BSI version 1, the first full scale investigation, BSI version 2, was conducted in 1987. Fifty pilot interviews were made with a quota sample of ten teachers over 50 years of age, ten university students, ten blue-collar workers, ten sales clerks, and ten vocational trainees aged 15-16.

This was followed in 1988-1989, by 200 tape-recorded interviews conducted on a subsample of informants of the 1000-strong national sample used for the pen-and-paper Hungarian National Sociolinguistic Survey (cf. Kontra 1995). This phase of the BSI project is known as BSI version 3[1].

1.2 Selection of the linguistic variables

The sociolinguistic variables involved in the survey included a range of phonological, morphological, syntactic and lexical phenomena. Their selection, which was partially based on suggestions made by linguist colleagues in response to a questionnaire, was motivated broadly by two reasons. (1) The majority of reseach topics were selected because they had been the subject of various statements in the literature without adequate empirical data adduced in support of the claims. Such issues included, for example, claims about the alleged effect of typewritten texts on vowel length[2]. (2) Some variables were included

[1] See Kontra & Váradi 1997 for a full description of the rationale and methodology of the BSI project.

[2] The previous standard for the Hungarian keyboard on typewriters lacked letters for the long vowels *í, Í, ű, Ű, ú, Ú*. This fact gave rise to the hypothesis that the shortening of these vowels may be due to the effect of typewritten texts. For a fuller discussion of this issue, see Pintzuk et al. 1995.

because they displayed variation that was little understood and thus called for empirical observation[3]. A further factor in the selection of variables was the suitability of the interview format to investigate the particular phenomenon. For the examination of some questions the face-to-face interview situation was simply unsuitable.

1.3 Classification of BSI tasks

The BSI used a number of different tasks.[4] They were devised with a view to creating situations in which informants produced data at different speech tempo and under varying levels of self-awareness.[5] Included below is a short summary of the tasks deployed in the interview. Kontra & Váradi 1997 gives a full account of the sociolinguistic variables and the methodology used, Váradi 1998 contains a detailed guide to all the data derived from BSI version 2. Table 1.1 shows a brief typology of the tasks according to the expected activity of the informants. This is followed by a short summary of the various types of BSI tasks. First the tasks administered with the aid of index cards are described and then the guided conversations will be discussed.

activity	description	self-monitoring
production	oral sentence completion (written clue)[6]	2
	reading passages (slow)	4
	reading passages (fast)	5
	reading minimal pairs	1
	reading word list	1
	word elicitation	2
	oral sentence completion (no clue)	2
	the reporter's test	4
	the staple remover test	5
judgement	"Same or different ..."	n/a
	"Which is correct ..."	n/a
	"How do YOU SAY IT?"	n/a

Table 1.1: BSI tasks by activities and level of self-monitoring

[3] The *-ba/-be* vs. *-ban/-ben* variation presents such a puzzling phenomenon (see Váradi 1994).

[4] The term *task* is an informal substitute for *instrument*, which is the technical term used in the BSI literature to refer to the particular task deployed in the interview to investigate a given variable.

[5] For a study of the effect of the presumed levels of self-monitoring on the (bVn) variable see Váradi 1995/1996.

1.4 Production tasks

1.4.1 Oral sentence completion (written clue)

In this type of tests informants were given a card which showed a sentence with a word missing and in the lower right corner it had a word printed separately. Informants were asked to insert the appropriate word-form in the sentence and read out the full sentence. For example:

Én tegnap nem	...	eleget.	ALSZIK
'Yesterday I did not	...	enough.'	'sleep'

1.4.2 Oral sentence completion (no clue)

The task is similar to the word insertion task above in that informants are asked to produce a full sentence by inserting a missing word but this time the test is done entirely orally. The field worker reads out the sentence frame and asks the informant to supply the missing word and pronounce the full sentence aloud. For example:

Sok mindenre emlékszem,	...	gyerekkoromban történt.
'I remember a lot of things	...	in my childhood happened'

The field worker pronounces the above sentence frame to the informant and encourages the informant to guess the missing word and repeat the sentence with the word inserted in the sentence i. e. *Sok mindenre emlékszem, ami gyerekkoromban történt.*

1.4.3 Word elicitation

This technique is familiar from traditional dialectologist field work. Lexical data are elicited by means of a question, which should be answered with a single word. For example:

Melyik az a szó, amelyiknek vécé a jelentése de k-val kezdődik?
'Which is the word that means *loo* but begins with a *k*?'

This task also includes sentence frames but they are read out to the informants by the field workers and the informants are expected to guess and produce only the missing word. For example:

A Földön már több mint 5 milliárd	...	él.
'There are already more than 5 billion	...	in the world'.

In this case, informants are supposed to utter the word *ember* 'man'[7].

[6]Level 2 awareness refers to the processing of the so-called primary variables. Level 3 awareness (not shown in the table) is attributed to the processing of the sentence frame itself. (See 2.1 on p. 15 for a discussion of the difference between primary and secondary variables.)

[7]Definite numeral determiners call for singular nouns in Hungarian.

1.4.4 Reading passages

A card with a typewritten passage was given to the informants. They were asked to read them through in silence then read them out as if to a friend of theirs who could not read because of a recent eye operation. Afterwards, they were asked to read the same passage again, this time as fast as they could. Altogether seven passages were used in this way. They were carefully made-up texts containing a high concentration of the variables that were tested in other parts of the interview as well. Efforts were made to ensure that they made coherent, natural flowing passages nevertheless.

1.4.5 Minimal pairs

Twenty cards each showing a minimal pair were given to the informants, who were asked to read them out. Not all the pairs constituted 'minimal pairs' in the technical sense of the term generally used in the linguistic literature such as for example *lombtalanít – lomtalanít*. Other pairs, such as *ezerszer – ezeregyszer* were featured for their suitability to elicit data on the *e – ë* variable in a highly compact manner. Whatever their technical status, however, all the pairs contained words that were very similar to each other.

1.4.6 Word list

Here again, informants were asked to read out separate words. The difference to the minimal pairs task was that the words were in groups of five or six written under each other on a card and they represented a rather mixed bag. One card, for example, contained the following words:

injekció, ember, erdőbe, bontsd föl, egyszer

1.4.7 The reporter's test

Informants were asked to give a running commentary of what the field worker was doing. They were trained in the test example to use verbs in the the present tense third person singular form. Field workers were instructed to carry out small actions like opening and then closing a window eliciting forms like *kinyitja/kinyissa* 'standard gloss: opens/should open'.

1.5 Judgement tasks

In these tasks informants were asked directly their opinion on matters of language use. As a matter of fact the first three in the list below did not even involve any speaking at all. Instead, their response consisted in filling out a questionnaire.

1.5.1 *"Same or different?"*

Informants were asked to listen to pairs of words recorded on cassette tape and played back to them on Walkman type cassette players through earphones. They had to fill out a questionnaire circling either the letter A if they thought the two words were identical or the letter K if they thought them different.[8]

1.5.2 *"Which is correct?"*

The set-up was identical to the previous task except that here informants had to decide which of the pair of words played on the tape was correct. They recorded their choices on the questionnaire sheet by circling the number 1 if they thought the first item heard was correct, number 2 otherwise.

1.5.3 *"How do* YOU *say it?"*

Again, pairs of words were played on the Walkman cassette player to the informants. They were asked to circle round number 1 if they used the first variant, number 2 if the second. All the items in this task investigated some phonological phenomenon.

1.5.4 "Demográfia"

This brief task is designed to record the perceived meaning of the word *demográfia*, which was currently undergoing modification in that a new meaning (birth control, family planning) was complementing or even supplanting the original sense of the word.

1.5.5 *The staple-removal test*

This test serves to document the birth of a word. The staple removal is a gadget that was practically unknown in Hungary at the time the interview was conducted. This part of the interview tested how people coped with naming a device they had not come across before. It consisted of a series of exchanges in which the field worker showed the device to the informants, tried to make them guess what its use was and finally got the informants to name the thing.

1.6 *Guided conversation*

Each BSI interview was required to contain at least 30 minutes of guided conversations. The BSI protocol contained a wide repertoire of *conversation modules*, i.e. conversations that revolved around a loose topic such as street crime, one's childhood etc. The set of conversation modules used in a BSI interview was largely left to the discretion of

[8] A for *azonos* 'identical', K for *különböző* 'different'.

the field workers except that there were some modules which had to be used in each interview. Some conversation modules even had to be introduced with the given words repeated verbatim. The field workers were instructed to engage the informants in natural flowing conversation that seemed to them spontaneous and conducted about topics that interested them. As far as the guided conversation part is concerned, each interview contains a core set and an unpredictible medley of conversation modules chosen out of the recommended set listed below.

1.6.1 List of conversation modules

Personal background (BIO)[9]

All the questions below were obligatory to raise:

1. Where were you born? Have you lived here throughout your life? (If not, where did you live and how long?)
2. Where were your parents born? Did they always live there? If not, where else and how long?
3. Where was your spouse born? Did s/he always live there? If not, where else and how long?
4. What's your occupation? Have you always had this job? If not, what else and when?
5. Your parents' occupation? If too numerous to list, when did they have what job?
6. Where exactly are we now? What's your address? (The question is used to elicit data for the *nyolc kerület – nyolcadik kerület* 'eight district – eighth distric') variation. If the interview takes place in a multifloor building the particular floor must also be asked to elicit data for the *öt emelet – ötödik emelet* 'five floor – fifth floor' variable.

Games (JÁT)

1. What was your favourite game in your childhood? How can one play that? What were the rules?
2. What do children play these days?
3. Can you recall a nursery rhyme?
4. Do you play any games these days? Do you play cards, parlour games, chess?

[9]The names of the conversation modules are followed by their three-letter codes in brackets. The modules which had to be introduced by repeating the introductory words verbatim are marked with XX.

Childhood (GYE)

What was your childhood like?

Fights, scuffles (VER)

1. As a child, you must have had a brawl sometimes. Can you recall a case when you had to fight for something?
2. Did you ever hit a man/woman? Why?
3. Do girls fight over here?
4. What is a fair fight and what is a mean one?
5. Have you ever been beaten up unjustly? What happened?
6. Have you ever beaten up somebody unjustly? What happened?

Dating (SZE)

1. How did you used to date in your days? How do you do it nowadays?
2. How do youngsters these days court? Do they court at all?

Marriage (HÁZ)

1. How did you meet your spouse?
2. How did you get married?
3. What was the wedding like?
4. What makes a good marriage?
5. Why do you think so many couples get divorced nowadays?

Danger of death (HAL) XX

Was there ever a moment in your life when it seemed that your life was in serious danger or that you might be seriously injured? When you thought "That's it. Curtains".

[If yes] What happened?

Fear of death (FÉL)

Surely, there were incidents in your life where something or somebody must have frightened you. What happened?

Dreams (ÁLM)

1. Can you recall a nice dream you had?
2. And a nightmare?

Family (CSA)

1. Tell me about your family.
2. And what about the family you were born in?

Religion (VAL) XX

When were you last asked what religion you had? What did you say? Is it important that somebody is religious or not? Why?

Friendship (BAR)

1. Tell me about your friends.
2. What makes a good friendship?
3. What makes a friendship go bad?
4. Has it ever happened that a good friend of yours turned out to be not so?

Street crime (BŰN)

1. Public security is continually worsening. What do you think is the reason for this?
2. What would you do if you were the police?
3. There are no brothels in Hungary but there is prostitution on the street and in hotels. Is this all right? Why?

School (ISK)

1. Did you like school when you were a child? Why?
2. Many people say that kids today don't learn even to read and write at school these days. Earlier they used to. Why?

Jobs, employment (MUN) XX

1. A lot of companies go bust nowadays. The bankrupcy is caused by the bad managers, yet it is the workers who get the sack. Is this right?
2. Women are often paid less in the same position doing the same job as men. Why is this so?

Abortion (ABO) XX

In Czechoslovakia women need no permit to have an abortion if they do not want to have a child. In Hungary, this is subject to a licence, therefore a woman can't freely decide whether to have a child or not. Which solution do you sympathise with, the Czechoslovakian or the Hungarian? Why?

Nuclear plants (ATO)

Are nuclear plants needed? Why?

Leisure time (SZI) XX

1. How much leisure time do you have?
2. What do you do in your free time?
3. Ten years ago did you have less free time or more?
4. What did you do then?

Jokes (HUM)

Do you like jokes? (If yes) Can you tell one you heard recently and think it's good?

Alternatively: Please tell me the joke that you consider the best you ever heard.

Ethnic minorities (ETN) XX

1. Very many people think that the Gipsies are doing too well in Hungary. Are they right?
2. Suppose you are to hire unskilled labourers and of the two candidates, who have equal qualifications, one is a Gipsy the other is Swabish, which one would you take on?

3. Do you know what CMÖ stands for? (If the answer is "no": *Cigánymentes övezet* 'Gipsy-free area'.) This abbreviation is often sprayed on bridges or walls of houses in Budapest, e.g. on Highway 3 there is a large graffiti. What do you think of this?

Language (NYE)

1. Did your teachers at school consider nice Hungarian speech important?
2. What rules did they stress often?
3. (For in-migrants or commuters) When you moved/started commuting to Budapest did you get commments about your accent from locals? What did they say? Were they right? Why?
4. Where do people speak nice Hungarian? Why?
5. (For informants who lived in the country for a long stretch of time) When you lived in the country[10] did you get comments from locals like you speak in a funny way? Once a schoolgirl from Budapest moved to Debrecen and their classmates told her she was putting on airs. Has anything like this happened to you? What exactly?
6. What's the language used in Budapest like? And that of the countryside?
7. XX Who do you think speak nice Hungarian of the following people?
   - leading politicians
   - elementary school teachers
   - shop assistants
   - teenagers
   - radio and television announcers
   - priests
8. (For in-migrants/commuters): Do you know words that you brought from home and locals here don't know or have learnt from you? Can you tell me some?
9. XX Has it ever happened that you were addressed with the formal/informal term of address (*te* vs. *maga*) and this was not right? What happened? Why was this not right?
10. Have you got any book on language or linguistics?
11. Have you got any dictionary? Bilingual dictionary? Which one? Do you use/read them?

[10] Field workers must find out where the informants lived.

12. Do you listen to or watch the programmes on language cultivation on the radio and television?

Informant's choice

Is there anything that I have not asked you about but you would have liked to talk about?

1.7 References

Kontra, Miklós. 1995. On current research into spoken Hungarian. *International Journal of the Sociology of Language* # 111:5-20.

Kontra, Miklós & Tamás Váradi. 1997. *The Budapest Sociolinguistic Interview: Version 3.* (Working Papers in Hungarian Sociolinguistics No 2, January 1998). Budapest: Linguistics Institute of the Hungarian Academy of Sciences. 54 pp.

Pintzuk, Susan; Miklós Kontra; Klára Sándor; Anna Borbély, 1995. *The effect of the typewriter on Hungarian reading style.* (Working Papers in Hungarian Sociolinguistics No 1, September 1995). Budapest: Linguistics Institute of the Hungarian Academy of Sciences. 30 pp.

Váradi, Tamás 1994. Hesitations between Inessive and Illative Forms in Hungarian (*-ba* and *-ban*). *Studies in Applied Linguistics* 1:123–140. [Debrecen]

– 1995/1996. Stylistic variation and the (bVn) variable in the Budapest Sociolinguistic Interview. *Acta Linguistica Hungarica* 43:295-309.

– 1998. *Manual of The Budapest Sociolinguistic Interview Data* (Working Papers in Hungarian Sociolinguistics No 4, February 1998). Budapest: Linguistics Institute of the Hungarian Academy of Sciences. 101 pp.

2 Transcribing the tape recorded material

2.1 What to transcribe

As the previous chapter demonstrated, we are faced with a large number of apparently disparate kind of data elicited through a battery of tasks which followed one another in a more or less set manner.

The tape recording contained a full record of what was said throughout the interview and apart from three tasks where answers were to be given by filling in questionnaire forms ("Same or different?" etc.), they included all the information that we were concerned with.

The first question is what to take down of this body of data.

A word by word transcript of the full taped interview does not seem a good idea. First of all, we don't need to take down everything because the interview contained set elements, i. e. frames. The most typical examples are the sentences with missing words in the oral sentence completion task. Even the reading passages can be regarded as fulfilling the same role on the discourse level. Their purpose was to provide a naturalistic context in which to elicit occurrences of the sociolinguistic variables that the BSI interview focussed on. On the other hand, part of the frame material also contained data that was elicited in other test items. For example, the very first test card

Ebben a ...	nem mehetsz színházba.	FARMER
In this ...	you cannot go to the theatre	jeans

contains three words with the (bVn) and the (bV) variables. One of them occurred in the word that was to be inserted into the sentence. This is what was presumably in the centre of attention of the informants, hence they were termed primary variables. The other two instances *ebben, színházba* were thought to engage the informants' attention to a lesser degree, hence they were termed secondary variables[1].

[1] For a full discussion of the value of this distinction with respect to the (bVn) data see Váradi 1995/1996.

2.2 In what form

Readers will recall that the whole interview was designed to elicit informants' use of certain linguistic items termed sociolinguistic variables. To a casual listener the tapes may sound just a stream of words, for us the tapes contain these nuggets of information scattered throughout the interview. Some of them are in predictable places (reading tasks), some are unpredictable (guided conversations). Among the data that we focus on, there will be recurrent patterns ie variants of the same variable. Again, it would be redundant and cumbersome to make a verbatim textual record of such items. Instead, it makes much more sense to code the different variants of the same variable with a number and enter only the number into the records.

IN SHORT, THE DATA SHOULD BE EXTRACTED IN A FORM THAT IS APPROPRIATE TO ITS CONTENT.

2.3 The structure of the input data

Let's now take a closer look at the data that we have to deal with. Immediately, we see a sharp division between the more or less free-form conversation part and the more closely structured card-based elicitation tasks. Accordingly, these two parts of the interview are processed in different ways. The conversations will be transcribed in the form of a more or less conventional transcript with some auxiliary information on the margins and some annotation interspersed in the text as will be described in the next section. The data from card-based elicitations will be entered into database tables in numerically coded form.

2.3.1 The conversation

The conversations were recorded according to a set of guidelines both as regards the form and the content of the transcript. Appendix A contains the full text of the annotation rules that were used in the transcription of the guided conversations. Following the text annotation methods employed by leading corpus projects at the time such as the LOB corpus (Garside et al. 1987) and the London-Lund Corpus (Svartvik 1990), the BSI transcripts used a fixed format line based approach. This means that each line is self-contained in the sense that it carries all the information necessary to uniquely identify it in the whole corpus. This information is encoded at the beginning of each line on character positions 1 - 16 which act as a virtual margin. This arrangement ensures that even if the corpus is subjected to a concordance search in the course of which the text lines may be jumbled up and sorted in alphabetical order of the query word, each line may be identified and the original context be looked up. See Figure A.1 in Appendix A for a sample page of transcription.

2.3.2 The card-based data

Itemise the data

The first task in processing the card based data is to itemise them, i. e. break them up into separate linguistic variables. At first blush, one would have thought that one card represents one item. For various reasons, however, this is not necessarily the case. Recall that in a frequently used task the informant is asked to produce the form of a word fitting the given context. The form itself that is produced may display a number of features which belong to different variables. Consider again the example cited in 2.1

Ebben a ...	nem mehetsz színházba.	FARMER
In this ...	you cannot go to the theatre	jeans

In order to insert the prompt word into the sentence, the informant has to make choices along two different sociolinguistic variables monitored by BSI: (1) vowel harmony (*farmerban* vs. *farmerben*) and (2) (bVn) variable (e. g. *farmerban* vs. *farmerba*). Therefore, just to encode the form of the prompt word in all aspects relevant to the BSI investigation, we need to enter it into two different records. In addition, as discussed in 2.1, the above test sentence frame contains a number of secondary variables[2] as well.

The first step, then, is to assign the maximum number of variables to each card. This yields a sequence of variables, some recurring, in the order that the informant is asked to produce them.

Setting up variants

The next task is to establish the range of variants of each variable. They are based on a more or less educated guess of what the informant is likely to, or can possibly, produce in the given context. Strictly speaking, this is anticipating and facilitating the coding of the data. There is no theoretical reason why a finite number of variants should be established beforehand. However, it does so happen that the spread of variant usage can be captured in a finite number of alternatives. In fact, we adopted the position that the number of variants would be a single digit figure and it very rarely proved insufficient. (Though a single such case was bad news enough!)

In practical terms, what happened was that before the data entry program was compiled, each card was carefully examined for all the potential forms that the given context might invoke. These were then arranged in decreasing order of likelihood of occurrence in that particular context and assigned a number. This order proved to be less insightful

[2]Such variables are termed secondary but this term is not meant to suggest any value judgement about their importance. On the contrary, one may argue that to the extent that the frame does admit different variants (it is mostly phonological, prosodic phenomena like assimilation, elision, liaison etc.) such variants provide more convincing evidence about the informant's vernacular than do primary variables.

than possible. As likelihood of occurrence varied with context, preference to this order meant that the same variant was not necessarily assigned the same numeric value. This problem only emerges when we are concerned about retrieval of data, a topic we thought we should face once we have sorted out all preceding stages. With hindsight, we can now conclude that it would have been wiser if the whole process of data collection, encoding, data entry and retrieval had been considered in its entirety from the beginning.

Having considered the structure of the data that served as input for our records, let's tackle the question of how to store them. We have briefly mentioned that they are stored in a database. The term database is often used fairly liberally to refer to any collection of data, however, it has a more restricted technical sense as well.

We'll be introducing the essential ideas of database systems as we go along in the discussion, but let's start by considering how the most popular type of database systems, the so-called relational databases work. Then we'll see how the BUSZI data could be arranged in terms of this scheme.

2.4 References

Garside, R. G., G. Leech & G. Sampson (eds.). 1987. *The Computational analysis of English. A corpus-based approach.* London: Longman.

Svartvik, Jan. (ed.) 1990. *The London-Lund Corpus of Spoken English. Description and Research.* Lund Studies in English 82. Lund: Lund University Press.

Váradi, Tamás. 1995/1996. Stylistic variation and the (bVn) variable in the Budapest Sociolinguistic Interview. *Acta Linguistica Hungarica* 43:295-309.

3 Relational database systems

A relational database system consists of a set of data tables. Each table is a set of data that have the same structure. A table is composed of an arbitrary number of *records*. A record is an elementary cluster of information relating to a single entity, typically containing a set of attributes called *fields*. Each entity *(record)* within a table must be characterised by the same attributes. This is just another way of saying that records must have the same structure. A relational database table is best conceptualised as a two dimensional table where records are the rows and fields are the columns. One can easily see that the number and sequence of columns must be exactly specified (otherwise we could not draw the table at all) whereas the number of records can easily vary (i.e. the table can be arbitrarily long[1]).

Let's see an example. Obviously, we'd like to store information of our informants. We want to record their personal details like name, address, age, sex etc. Already here the question arises of how to group this information. With names, for example, should there be two fields, one for family names, one for Christian names? And how should addresses be broken up? Well, what one must bear in mind in such cases is that it is relatively more difficult to access information from within a field than the contents of the whole field. Therefore, what is likely to be the target of a lookup either on its own or in combination with other items of information is best put in a separate field. So far, we may have the following scheme.

INFORMANTS

Surname	First name	Sex	Age	Education	Address

Figure 3.1: Details of informants divided into fields

The table in Figure 3.1 could be sufficient in itself but most of the time we are dealing with several tables which are related to each other. For example, associated with the answers would be the informant. Obviously, it would be hugely redundant to store the information on the informants in the table in which we keep the answers. (There are 250 informants, but up to 160 000 individual responses.) Instead, if we just stored the

[1] Tables can be arbitrarily wide as well, i.e. there is no limit in principle to how many fields a record may contain (though there may be one imposed by the particular software used) but it is important to note that the number and sequence of fields in a record of a table must be defined beforehand, and although it is possible to modify the table, we must keep to the current setting at all times.

name of the informant next to each response, we could use that as a pointer to look up further details of the person from the INFORMANTS table. More efficient than using the surnames or even combination of surname and first name would be to set up an ID field in the INFORMANTS table and use that as a key into the ANSWERS table. Note that in the INFORMANTS table there can't be two records of the same ID, whereas in the ANSWERS table the answers given by the same informant will be placed in separate records with the INF_ID field containing the same ID. Indeed, that is how 'relationships' are expressed between tables; through the identical content of fields whose function may be solely to act as link between tables. This is illustrated by Figure 3.2. The nature of the relationship may or may not be explicitely indicated in the field name or its content.

INFORMANTS

INF_ID	Surname	First name	Sex	Age	Education	Address

ANSWERS

INF_ID	Card_No	Answer	Transcriber_ID	Tape_counter	Checker_ID

Figure 3.2: Linking between two tables through INF_ID field

So much will be enough to convey the gist of how relational database systems are structured. Of course, these are merely ground rules enabling us to arrange our data in the required way but far from adequate to design an EFFICIENT system. In fact, these few rules are all that the relational database model imposes on the data. They leave practically limitless scope for arranging the data one is dealing with in any way that one finds suitable or relevant.

It should be emphasized that the key to success in constructing an efficient database system lies in the careful modelling of the relationships inherent in the data. This should be the first step and one that deserves meticulous analysis. It is basically a pencil and paper work, yielding an abstract scheme, independent of any software considerations.

The next step is to implement the data model on the computer. This typically means turning to a general purpose database management software package and using its programming language and facilities to develop a software system.

3.1 Structure of output data

Let's see in some detail a first attempt at arranging the output of the card-based part of the interviews[2].

Table 3.1 shows a sample of an answer file of production data.

[2]The system that the rest of the present chapter describes is no longer in operation and should be considered obsolete. It has been superceded by a revised system that is the subject of the next chapter.

INF_ID	Card No	Counter	R_Tr	R_Ch1	R_Ch2	Remark
B7211	1701	1a1932	1			
B7211	1701		3			
B7211	1802		1			
B7211	1903		2			I
B7211	2004		1			
B7211	2105		2			
B7211	2206		2			
B7211	2307		2			
B7211	2408		2			
B7211	2509		1			

Table 3.1: The contents of an answer file for production type of data

Field name	Description
Inf_ID	The name of the informant
Card_No	The ID number of the card used to elicit the answer
Counter	The tape counter reading[3]
R_Tr	Informant response noted down by the transcriber
R_Ch1	Informant response noted down by the first checker
R_Ch2	Informant reponse noted down by the second checker
Remark	Flag to indicate any remark (stored in a separate file)

Table 3.2: The structure of answer files for production type of data

The idea behind setting up the table in this way is to ensure that any particular record will reveal the essential points of who said what in response to which card. Note that the informant ID is carried in every record, apparently a huge redundancy. After all, one could argue, if we put the answers from the same informant in a file named B7211, we wouldn't have to set up a field for this. However, this would mean that once we remove the item from the file, we have no way of identifying the informant.

Each informant's answers were stored in separate files. Moreover, for each informant the answers given to the 26 different modules were again each stored in different files. This soon led to an explosion of the number of files generated as transcription work gathered momentum. This meant that the total set of results would have resided in 250 x 26 = 6500 files + the 250 text files. With hindsight, it was unnecessary to store the data in so many files. It put a substantial burden on the file storage system but, more importantly, it would have made data access from across different files extremely difficult.

This proliferation of files, however, was merely a nuisance, as the files could be collated into a single one without much difficulty.

[3]Obligatory only at the beginning of each module

Unfortunately, the structure of the answer files differed according to the type of modules they came from. Production data, judgement data and the answers to the staple remover test were processed differently.

One obvious difference between the production and the judgement data lies in the way the answers are coded (letters "*a*" for identical, and "*k*" for different, as against numbers for the production data) as well as in the reference system of the cards. Another, less trivial, problem was the following. Individual items in the judgement tasks were elicited not on separate cards, as in the production tasks, but on a sheet which contained 20 - 21 items. Therefore the field *Card No* could not be used to reference individual items as it served to identify the questionnaire sheet. To overcome this problem an auxiliary field (Item No) was devised to establish unique reference to individual items.

Accordingly, the judgement data were accommodated in the following type of tables:

Inf_ID	Card No	Item No	Counter	Response	Remark
B7301	9100	1	1a2610	a	
B7301	9100	2		k	
B7301	9100	3		a	
B7301	9100	4		k	
B7301	9100	5		a	
B7301	9100	6		a	
B7301	9100	7	1b0116	k	I
B7301	9100	8		a	
B7301	9100	9	1b0151	k	I
B7301	9100	10		k	

Table 3.3: Sample answers file for decision data

As it turned out later, the card numbering system was more seriously flawed.

First, card numbers were not unique. As pointed out in 2.1, the same card served to elicit a number of linguistic variables. Originally, the idea was to use consecutive numerical codes for the responses so that answer codes 1-3 covered lanuage problem *a*, codes 4-5 language problem *b* etc. To add to the problems, it was decided that a single digit number would be sufficient to record alternatives. This may have proved adequate for a single language variable but not when there were several variables all consecutively numbered. So it happened that occasionally the principle of consecutive numbering had to be broken anyway.

Secondly, and this was indeed most serious, the slow and fast reading data was assigned the same card number reference within the files. Again, this stems from an attempt to map closely the physical data and its representation in the database. As the same card was used for both the slow and the fast readings, the data was assigned to the same card number. True, they were put in different files and the filename did reflect what module

the data came from. However, this information was only good until the data was used in terms of files and not in terms of individual records. As soon as one wanted to pool the answers together, one would have been left with no means of distinguishing the fast and slow reading data.

In conclusion, we should note the following shortcomings in the data model:

- The data was fragmented and stored in too many separate files
- The structure of the data was not uniform across modules
- The numerical codes assigned to responses were not the same for the same linguistic phenomenon
- Most serious of all, the card numbering system was not consistent, i.e. card numbers were not unique
- In general, one may say that it failed to use the true potential of a relational database model. Far too much information was encoded externally, ie in filenames instead of being incorporated in the database tables.

3.2 Data entry

So far, we have been concerned with structuring the data so as to arrive at an optimal model. Optimal in the sense that all the relevant information should be recorded and be accessible in the most economical and efficient way. Implicitly, this also requires having regard to the way the information in the database is going to be processed but at this stage this should be a secondary consideration.

Let's now look at how the data was actually handled. Corresponding to the excessively fragmented and elaborate data structure, was a fairly complex program that controlled the data entry. It was bound to end up like that because most of the complexity of finding one's way among the numerous files was left to the data entry program to sort out. Fortunately, what went on behind the scenes was not apparent to the end users. They soon came to lament certain inflexibilities in the operation of the program, which were produced as a matter of design principles. These included the following decisions.

- The data entry program should be made as fool-proof as possible. This meant transcribers could only operate within the constraints of what the program allowed them to do.
- The entry of data should proceed in a strict chronological order. This proved to be the issue that was most resented by the staff as it deprived them of backtracking or recording just a select few items.

The operation of the system was fairly simple. Transcribers' work environment included a Sony BM 88 transcriber machine and a PC. What they heard on the cassette tape was entered directly into the computer system. The initial screen of the program is displayed in Figure 3.3. Having chosen the data entry function of the program, the transcribers saw the contents of the cards coming up on the screen one after another. Figure 3.4 is a screen shot of a data entry screen. Below the cards, the screen displayed the anticipated variants with their numerical codes and the transcribers were prompted to enter the variant they heard the informant say on the tape.

The smooth operation of the data entry was ensured by a database table. Recall that the whole interview was structured in terms of modules, some conversational, some card-based elicitation. The latter were listed in the table MODULES, whose contents is displayed in Tables 3.4 and 3.5.

MODULES serves as a source of data to control the procedure of the data entry as well as reflecting certain features of the data. The sequence of records here captures the chronological sequence of the modules and this aspect was used to control the order of the data entry. The advantage of using a table to determine the procedural aspects of the program lies in the flexibility this method provides. By replacing the contents or the sequence of the records, the same control program can be used to handle a different set of data.

The field *Modul* served to identify the source of the data that was displayed on the data entry screen. Again, these had been arranged in tables of the name registered in the *modul* field. This way, the input to the data entry screen could be manipulated easily and at any time without having to rewrite the program. Changes between BSI-2 and BSI-3 only call for changing these tables.

The contents of a sample input table (KARTYAK1 'cards1') is shown in Table 3.6. Fields *S1* and *S2*, two long lines of text, contain the text of the sentence frame into which the target word was to be inserted. Fields *V1* - *V9* contained the slots for the anticipated variants of the target form. The field *Lim* was designed to record the range of the numerical codes of the possible answers in response to the particular item. This served the double purpose of disallowing any mistaken entry by the transcriber and identifying which variable the answer referred to in case the same card served as the prompt for several variables. If none of the anticipated variants was actually used, the fom had to be entered in the memo field attached to the entry. Also, the memo notes were used to record any remark, paralinguistic or prosodic feature in the informant's speech that was relevant.

As mentioned earlier, the answers given by an informant in response to a module were stored in separate files the names of which were composed out of the informant's ID and the module ID as stored in the M_ID field. For example, items elicited with the first batch of cards (in the KARTYAK1 module) from, say, informant B7303 were stored in the file named B7303VL1.dbf. (Dbf is the standard extension name assigned by the program used, which was DBASE).

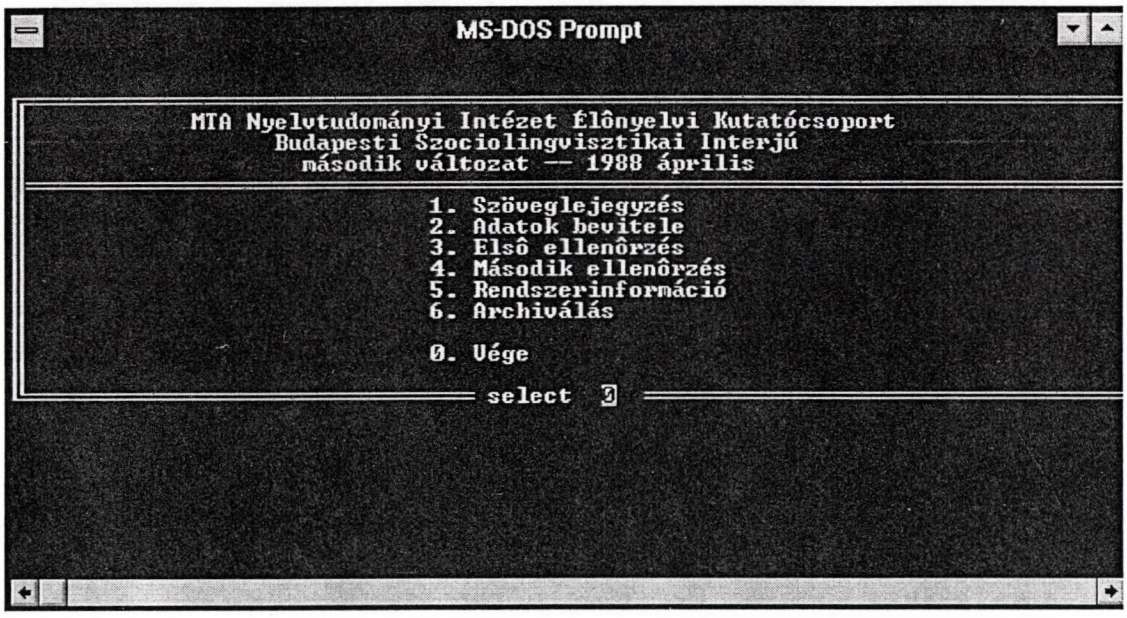

Figure 3.3: The main menu of the old user interface

1. Transcription of conversations
2. Data entry
3. First checking
4. Second checking
5. System info
6. Save data
0. Quit

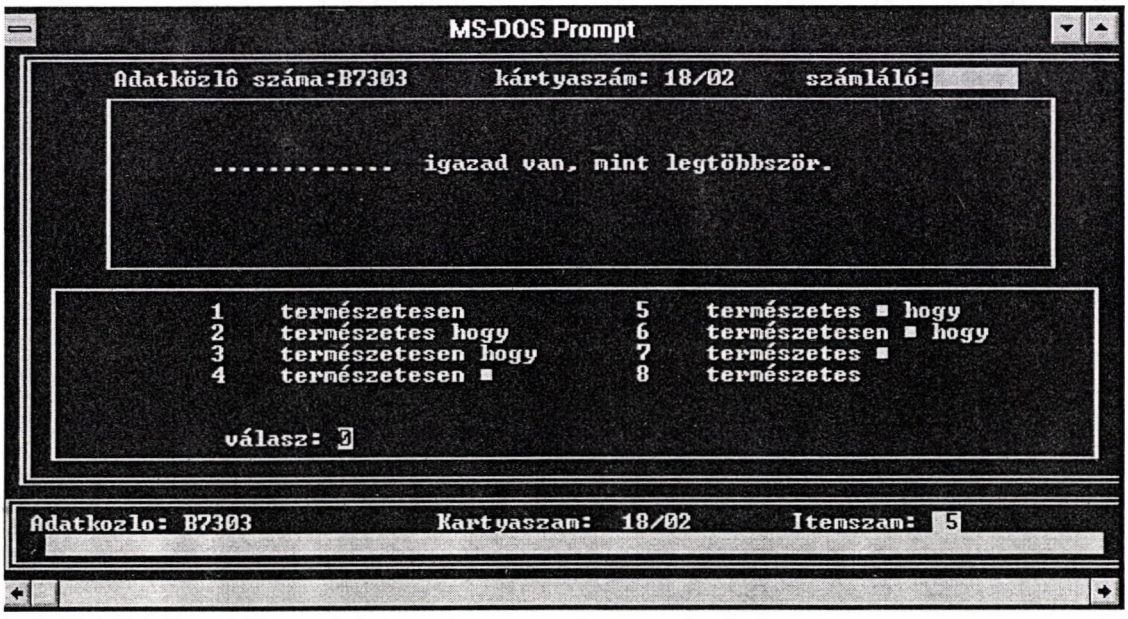

Figure 3.4: The data entry screen

Modul	M_ID	Program	From	Till	No of records
KARTYAK1	VL1	9	1701	4630	47
KARTYAK2	VL2	9	4931	7860	46
JOSKA	O1L	9	7901	7917	17
JOSKA	O1G	9	8001	8017	17
MEGHIRDE	O2L	9	8201	8218	18
MEGHIRDE	O2G	9	8301	8318	18
KARTYAK3	VL3	9	8501	8826	32
V.LAP_1	AK1	A	9100	9100	21
KARTYAK4	VL4	9	9701	10216	56
HATODIK	O3L	9	10301	10336	39
HATODIK	O3G	9	10401	10436	39
V.LAP_2	AK2	A	10500	10500	22
PISTA	O4L	9	10801	10833	28
PISTA	O4G	9	10901	10933	28
FELMERÜL	O5L	9	11001	11023	33
FELMERÜL	O5G	9	11101	11123	33
V.LAP_3	AK3	A	11200	11200	22
EZERSZER	O6L	9	11501	11527	27
EZERSZER	O6G	9	11601	11627	27
KARTYAK5	VL5	9	11801	13331	34
HOL_VAN	O7L	9	13401	13423	23
HOL_VAN	O7G	9	13501	13523	23
KARTYAK6	VL6	9	13601	14207	17
RIPORTER	RIP	9	14301	1430	46
DEMOGRAF	DEM	9	14500	14500	1
KISZEDO	KIS	K	14600	14600	1

Table 3.4: Contents of MODULES table

Field name	Description
Modul	The name of the module
M_ID	Three-letter ID used in the names of answer files
Program	An internal flag to indicate what type of data entry program to use
From	The number of the card the module starts
Till	The number of the card the module ends
No of records	The number of items in the module

Table 3.5: The structure of the MODULES table

Which modules belong to what type is recorded in the "Program" field. The two fields *From* and *Till* were meant to be looked up to see which module any particular card belongs to. The *Number of records* field served to establish when coding a module is complete.

In conclusion, the following seems to be a reasonable summary assessment of the data entry operation:

- It managed to do fairly smoothly and efficiently the extremely complex data handling operations imposed by the fragmented data model.
- One positive feature was the way the control program and the data were separated allowing easy update of the data without the need to rewrite the program.
- Shortcomings included artificial self-imposed limits to the number of alternatives, leading to quickfixes that made the system inconsistent.
- The rigid sequence imposed on the transcriber proved very unpopular with some of the staff who wanted to look into certain problems without having to transcribe whole modules. Also, lack of backtracking to review previously entered data turned out to be an unnecessary constraint.

Card No	**1701**
S1	Ebben a ..... jól nézel ki.
S2	ebben
V1	ebbe
V2	
V3	
V4	
V5	
V6	
V7	
V8	
V9	
Lim	12
Card No	**1701**
S1	Ebben a ..... jól nézel ki.
S2	farmerben
V1	farmerbe
V2	farmerban
V3	farmerba
V4	
V5	
V6	
V7	
V8	
V9	
Lim	36
Card No	**1903**
S1	Mari ..... egy ingemet tegnap.
S2	kimosta
V1	kimosott
V2	
V3	
V4	
V5	
V6	
V7	
V8	
V9	
Lim	12

Table 3.6: Contents of input card table KARTYAK1

4 A revised system

It was decided at the outset that after the system had been in operation for a while, it would be reviewed. Many of the shortcomings gradually became clear as transcription got under way. The development of hardware and software technology (the appearance of 386 machines on the one hand and the Windows operating system with a graphical interface and a sort of multitasking capability) has meant that a revision was felt necessary earlier than the completion of the first 50 interviews.

Although the original system was designed to be flexible enough to work even with a completely different set of data – due to the fact that the operation of the program was controlled by data stored in easily editable tables – the radical change in the software environment meant that the whole data entry program had to be abandoned and rebuilt from scratch.

4.1 Redesigned data model

Although the program has been given a completely different look and feel, more important is the way the data model was redesigned.

First of all, all the output data have now been brought to a consistent, uniform structure and are stored in a single file. In order to achieve this, the following changes were required:

1. Consistent and unique identifiers have been assigned to language items.
   Recall that the original system had very serious flaws in failing to distinguish *inside the data files* between slow and fast readings, between different variables using the same card etc. Now the whole scheme was converted to one where all the linguistic variables are numbered consistently whatever their source (module). This identifier, termed *item* here, now takes over the function of the card number reference. It replaces the card numbers but the link is kept in a table in the background so it will be possible to look up something by the card number reference, such as it is.

2. Alphabetical types of responses have been converted into numerical codes.

   This was a fairly trivial task. At the same time the consecutive numbering of responses to the same card was abolished.

3. Remarks have been incorporated into data files.

   Notes by transcribers and checkers are now kept together with the responses in the same record. They are kept in a memo field whose length can grow or shrink as desired.

A sample of the revised output data file is displayed in Tables 4.1 and 4.2.

1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
B7103	10	1		RA							
B7103	20	1		RA							
B7103	30	0		RA							Kommentárja: "Ja, hogy ban vagy ba ja értem."RA
B7103	40	0		RA							2., majd kétszer 1. RA
B7103	50	2		RA							
B7103	60	2		RA							
B7103	70	1		RA							
B7103	80	2		RA							
B7103	90	2		RA							
B7103	100	2		RA							
B7103	110	1		RA							

Table 4.1: Contents of the *Answers* data table

	Field name	Description
1	**Inf_ID**	Informant's ID
2	**Item**	Unique identifier of the language variable
3	**Ans**	Transcriber's coding of the response
4	**Date**	Date of above
5	**Tr_Name**	Transcriber's name
6	**Ch1**	Response value by first checker
7	**Ch1_Name**	Name of first checker
8	**Ch1_Date**	Date of first checking
9	**Ch2**	Response value by second checker
10	**Ch2_Name**	Name of second checker
11	**Ch2_Date**	Date of second checking
12	**Note**	Notes by any three

Table 4.2: The structure of the *Answers* table in Figure 4.1

The input data file for the data entry screen was also drastically redesigned. Earlier what was shown on the cards and the expected responses to a variable were rigidly

controlled by the way the data was structured. A maximum of 9 alternatives was allowed per card.

The major innovation here was to break this rigid mould and separate the invariant card header information from the alternatives. Accordingly, a separate table was set up to store the prompt skeleton information. This table called PROMPTS, stores two further pieces ofdata: 1) the physical card number, to maintain the link with theold reference system and 2) a field called STATUS to indicate whether the variable is explicitely focused on (primary variable) or not (secondary variable). The structure of the PROMPTS table is shown in Tables 4.3 and 4.4.

Item	**Card_No**	**Stat**	**S1**	**S2**
10	1601	2	Ebben a ..... jól nézel ki.	
20	1701	2	Ebben a ..... jól nézel ki.	
30	1802	1	..... igazad van, mint legtöbbször.	
40	1903	2	Mari ..... egy ingemet tegnap.	

Table 4.3: The contents of the *Prompts* table

	Field name	Description
1	**Item**	The unique ID number of the variable
2	**S2**	Optional second line of the prompt on the card
3	**Card_No**	The old card number reference
4	**Status**	The primary or secondary status of the variable
5	**S1**	First line of the prompt on the card

Table 4.4: The structure of the *Prompts* table

Anticipated possible responses are now stored in a table called OPTIONS. Its structure is displayed in Tables 4.5 and 4.6.

Each alternative is put in a separate record. This meant we didn't have to impose any artificial ceiling on the maximum number of alternatives. As each alternative carried the newly redesigned item reference number, they could be unambiguously attributed to the item they belonged to and the number of alternatives could be arbitrarily large or small. This simple but most powerful device, the central idea behind relational database systems, is illustrated in Figure 4.1. The arrows show how the records are linked by the identical content of a shared key field, ITEM.

The overall design of all the datafiles is displayed in Figure 4.2. In addition to the three central tables discussed so far it shows some auxiliary tables to store data about tape counter settings MAGNO, informants AKLISTA and transcribers LEJEGYZO, respectively.

Item	Order	Option	Value
10	10	ebben	1
10	20	ebbe	2
20	10	farmerben	1
20	20	farmerbe	2
20	30	farmerban	3
20	40	farmerba	4
30	10	természetesen	1
30	20	természetes hogy	2
30	30	természetesen hogy	3
30	40	természetesen	4

Table 4.5: Contents of Options table

	Field name	Description
1	**Item**	The unique ID number of the variable
2	**Order**	Sequence number of alternatives
3	**Option**	Form of the anticipated alternative
4	**Value**	Numerical code of the alternative

Table 4.6: The structure of the *Options* table

4.2 The graphical user interface

The revised data model allows a much simplified controlling program. Gone is the need to juggle with a number of different files as there are practically only the three tables above each stored in a single file. The windowing environment means that the transcription and the coding of the data could be done simultaneously in separate windows. The rest of the menu functions in the earlier system have been incorporated in the data entry screen which allows the user to choose between transcribing, first or second checking of data (*lejegyző* 'transcriber', *1. ellenőr*, '1st checker' and *2. ellenőr*, '2st checker' options). Also, there is an additional facility to revise one's own work (*Bevitel*, 'data entry', vs. *Javítás* options). Depending on what function is selected, the card next in line for that operation is displayed on the screen (i. e. if the transcriber and data entry buttons are selected, the screen will show the first card in row which is not yet transcribed from the given informant.) However, with the help of the vertical row of navigation buttons one can go back and forward, jump to the top or the bottom of the cards available for processing for the given operation. In addition, a given item may be jumped to directly by selecting its number in the drop-down window situated in the middle of the top row of small windows.

Item	Card No	Stat	S1	S2
10	1601	2	Ebben a ..... jól nézel ki.	
20	1701	2	Ebben a ..... jól nézel ki.	
30	1802	1	..... igazad van, mint legtöbbször.	
40	1903	2	Mari ..... egy ingemet tegnap.	

Item	Order	Option	Value
10	10	ebben	1
10	20	ebbe	2
20	10	farmerben	1
20	20	farmerbe	2
20	30	farmerban	3
20	40	farmerba	4
30	10	természetesen	1
30	20	természetes hogy	2
30	30	természetesen hogy	3
30	40	természetesen □	4

Figure 4.1: Linking the PROMPT and the OPTIONS tables

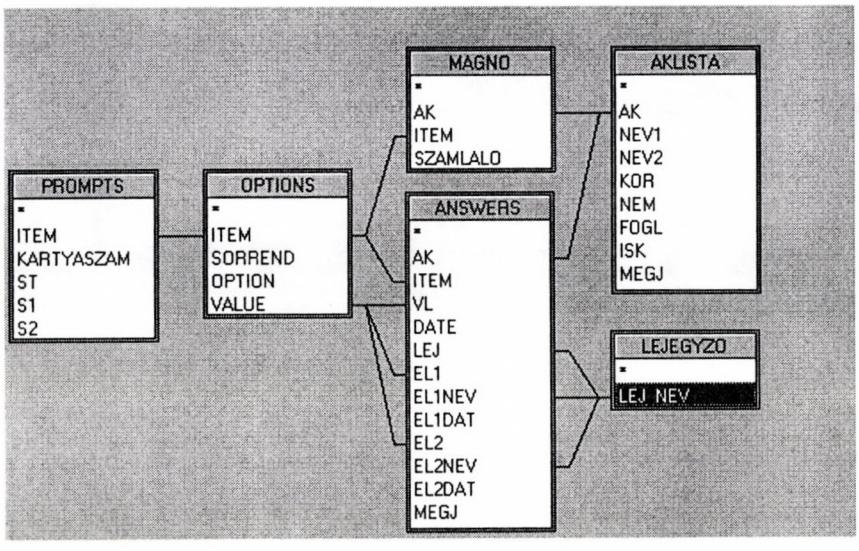

Figure 4.2: The linking of data tables in the revised system

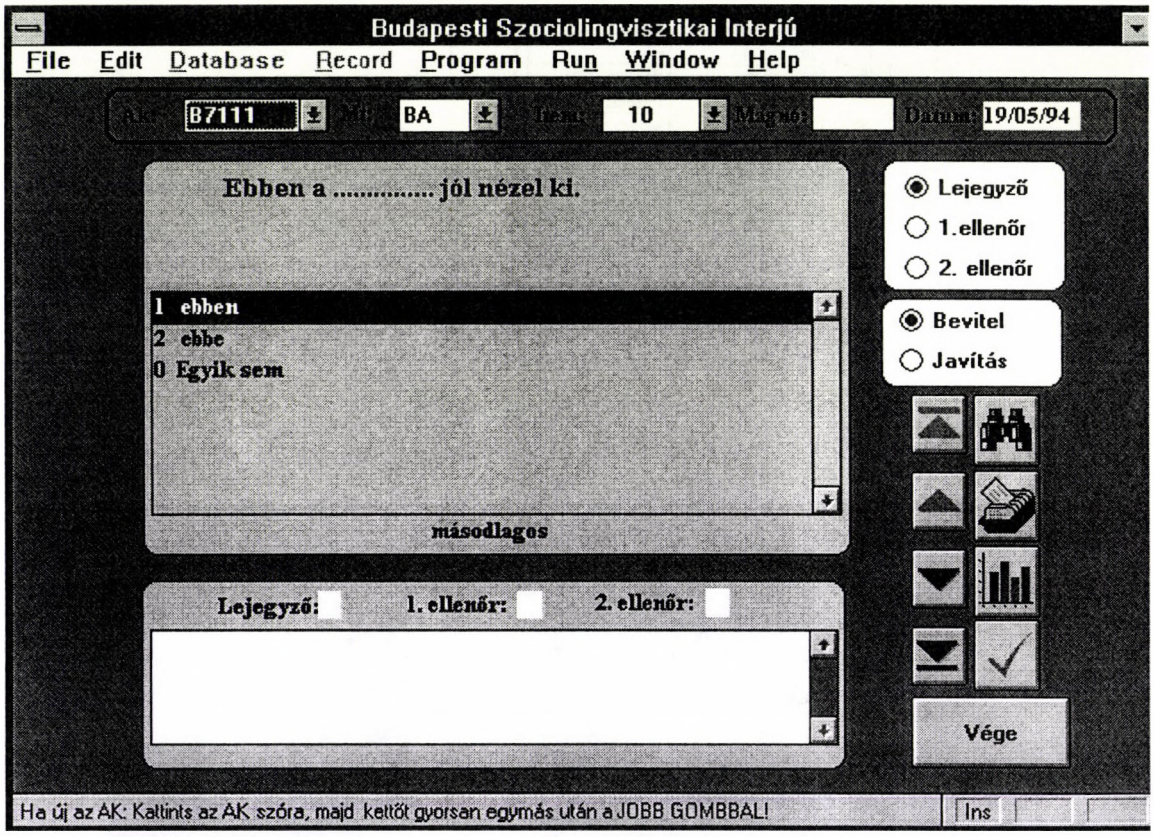

Figure 4.3: The revised data entry screen

5 Data retrieval

Most of the efforts in the BSI project so far have been contrentrated on recording as much data as possible. Given that data entry has proved an extremely labour intensive process[1], the major thrust of the work was to achieve a critical mass in the amount of recorded data before one can set about the obvious next phase, data retrieval[2].

5.1 Current limits

At the moment, there is no provision whatever to aid access to the data. (This is in accordance with the fact that the BSI archives are not yet open to the general public.) In the revised system, all of the informants' responses reside in the ANSWERS file. It is a standard Foxpro database file (see Table 4.2), it basically serves to tell which variant was used by a given informant in response to which item (who said what in response to which item) – all enshrined in numbers. In order to find out what the numbers stand for both as regards the variant and the item, one needs to link the ANSWERS table with the OPTIONS table to access the particular variant, as well as with the PROMPTS table to retrieve the prompt i. e. the text of the card. The linking is to be done through the VL ('response') and the ITEM fields in the way shown in Figure 4.2.

Lack of user friendly tools to access the data is merely a temporary constraint and implies no inherent shortcoming in the potential of the present arrangement of the data. A genuine limitation in the current data model, however, is the fact that at the moment the database comprises a collection of atomic observations of each informant's language use. In its present state it fails to reflect the fact that (1) the same test word occurs in many different tasks throughout the interview and (2) the same linguistic phenomenon (e. g. palatal assimilation, *-ik* conjugation, foreign words) is investigated in a number of different words. In short, the isolated atomic facts need to be aggregated along (socio)linguistically significant generalisations. This task will be taken up briefly in 6.1.

[1] It is estimated that one hour of conversation requires about 24 hours to transcribe.

[2] Paradoxically, in view of the enormous rate of advancement in software and hardware technology, this delay must have spared us the trouble of developing retrieval software that will have become obsolete by now.

5.2 A common multimedia interface

The first large-scale publication of BSI data was launched in late 1997. The full data plus the digitized sound of a complete BSI version 2 interview was published on a CD-ROM. This required the integration of the numerically coded data with the transcript in a unified system on the one hand, and the linking of the digitized sound files with the recoded data, on the other. As a result, it is now possible to very quickly access any part of the data and by simply clicking on an item to listen to it as well. Ten years ago when the BSI project was in its initial phase, all this sounded like a Utopean dream. This is no longer beyond the means of current software and hardware technology. The following sections give an overview of the process of how this was achieved.

5.2.1 Digitizing the tape recordings

Limitations of the tape recordings

The use of tape recording meant a revolutionary step in empirical linguistic analysis, and has clearly established itself as a basic research tool (only challenged perhaps by video tape recorders where their use was appropriate and feasible). However, both media have certain limitations in terms of handling.

1. The master tape cannot be reproduced without loss of quality,
2. its quality is subject to deterioration even if kept in very stringent storage conditions,[3]
3. positioning the tape to a precise location is a bit cumbersome.[4]

Making a digital recording

All these limitations can be overcome with the use of digital technology as digital recordings do not decay in time, exact duplicates can be produced even in chain copying and data can be looked up instantaneously. Ordinary (i. e. non-digital) tape recordings capture sound in terms of an analogue electric signal, variations in the level of voltage, basically. Digital recordings record sound as a stream of numbers which record the level

[3] The Linguistics Institute has never had a facility to store the tapes in the required conditions, therefore the master tapes are deposited in the archives of the Institute of Musicology of the Hungarian Academy of Sciences. They are accessed only occasionally if some critical check has to be made on the data. Otherwise the transcription is made on the basis of the duplicate cassette tape recording that was prepared of the master recording.

[4] The transcription of the guided conversations included tape counter settings on the margin which recorded the locations of the beginning and the end of each conversation module. Within the body of the modules the tape counter setting was recorded at every two minutes. This practice made some winding and rewinding of the tape almost always inevitable.

of the incoming signal strength measured at a set interval. In order to enjoy the benefits of digital technology, the original tape recordings had to be re-recorded digitally, a process referred to as digitization. This process no longer poses any challenge either as regards hardware or software. It only calls for a PC equipped with a sound card and a piece of sound editor software. True, it requires a lot of storage space but hard disk space and CD ROM disks have become relatively inexpensive.

The digital recording was made by playing the original tape recording into the input channel of the sound card which measured the signal stream at the required sampling frequency[5] and produced the stream of numbers which was stored in a file. The file could then be played back with the help of the sound card again, converting the numbers into analogue signals required by the speakers or earphones attached to the computer.

Linking the data with the sound files

One major design issue that had to be faced was how to break up the the original tape recording into separate digital sound files. The answer has a crucial bearing on the way the data is accessed. The important factor to consider is the way the digital sound is linked in with the data files. At the moment, it is unrealistic to expect a system where one could select any arbitrary stretch of the transcript and have the computer play back the corresponding sound bit.[6] Instead, what is accessed with a click of the mouse button will be a whole sound file. This means that all those parts of the whole interview which one would like to jump to need to be put in a separate file. It is also possible to navigate within the sound file, once it is retrieved, but that operation would be serial, i. e. would require winding and rewinding, in a virtual sense, of the data. In short, access is random to files, but serial within the files.

Therefore, the digitized sound was broken up into files in the following way. With the exception of the judgement data and the reading of the passages, each item of the card based part was recorded in a separate file. Note that making them accessible as individual units of the data and in full accoustic and textual richness detail (as against a mere number, with possibly a transcriber's comment) represents a significant advancement over the way the same data is treated in the BSI files.

Full-scale access to the sound of this part of the interview opens up the possibility to carry out future research on aspects of the data that were not monitored and therefore not preserved in any form either by the original BSI protocol.

The reading passages were not broken up into the isolated bits of data that were itemized in the database. Technologically, there was no problem in identifying and dissecting

[5] We used a sampling frequency of 22.1 kHerz/sec, i.e. the sound of the tape recording was measured 22,100 times every second. This was half the sampling rate used for CD quality recordings but was deemed adequate for the range of human speech.

[6] It is not implied that this is not feasible in principle in current state of the art technology but only that it is way beyond the means of the present project.

even individual sounds in the data stream. However, after a few tests it was obvious that an isolated recording of a word is insufficient to make judgement about phonological and prosodic characteristics of the data. (Length, for example, is relative to overall speech tempo, length of other similar units etc. One needs to listen to a longer stretch to make reliable judgement about individual units.) Therefore, each reading passage was put into separate files. Of course, the slow and fast readings of the same text were treated as different data.

As regards the guided conversations, each conversation module presented an obvious natural unit for a sound file. The only problem sometimes was withlong modules. They generated huge files[7] which took long to load and a longfile means it takes longer to find things through navigation (as againstinstant lookup). Therefore, it was decided that the modules that were longer than two minutes would be broken up into a series of one minute stretches.

5.2.2 Converting the data files

Processing the database files

As described in Tables 4.1 and 4.2 on p. 30, the data derived from the card based tasks of the interview were coded numerically. The data tables contain numbers which are impossible to interpret without linking them to the corresponding records in the data tables PROMPTS and OPTIONS in the way shown in Figure 4.2. These three tables can be linked and browsed with the help of a suitable database management program but this is not what is needed. What we need here is a plain text form of the answers given by the particular informant, all the records of the ANSWERS table that come from the given informant in a form where the text of the corresponding record from the PROMPTS table is supplied, together with the textual form of the answer as looked up in the OPTIONS table.

Handling the reading passages

The reading passages required a slightly different handling. First of all, to bring the database records more in line with the sound files, where the whole passage is heard continuously, the original passage is first shown in the same form as was handed over to the informant.[8] This version was followed by the forms of the passage where the standard forms were replaced by the variant forms actually used by the informant. However, only those items that were coded in the database were treated in this way. The rest of the

[7] The 22.1 kHerz/sec sampling rate meant that 22,100x2 bytes are needed to record one minute of data.

[8] This included slight deviations from the standard orthography in the case of passage 1 in that no *í*, *ú* and *ű* characters were used in them as if they were typewritten with a typewriter of the earlier Hungarian keyboard standard. This was done so as to allow the investigation of the possible effect of this keyboard standard on the length of the above listed vowels.

passages occur in standard orthographic form regardless of how it was uttered by the informants.

Adjusting the transcript

The transcript of the conversation modules required slight modifications only. The original format of the transcript was designed so that each line could be identified when a concordance is prepared of the data. For the present purposes, however, the invariant part of the left margin containing the ID of the informant and the particular module could be dispensed with entirely. Their role is either superfluous in the present context – as we are dealing with only one informant at the moment – or is filled by other navigational device. The tape counter settings on the right margin of the transcript can also be ommitted as irrelevant.

5.2.3 *Weaving everything together*

Having discussed the individual components of the system, it is time to give an overview of how they were integrated into a common user interface. Here again, we are fortunate to find that due to the enormous development of software technology since the start of the BSI project – which dates back to the age when home computers like the Commodore 64 ruled the day and the IBM PC XT was just on the horizon – we now have readily available technology to handle text, sound and pictures together in a simple and intuitive manner. This is provided by the hypertext technology which has now become an indispensible part of computer literacy as it is embodied in the help system of any windows-based operating system and, more conspicuously, in the Internet browser programs like Netscape or Internet Explorer.

The basic insight of the hypertext is as simple as ingenious. Ordinary text is basically a stream of characters, a flat two dimensional structure, which is processed in a serial manner. A hypertext adds another dimension to it in that it contains embedded links pointing to different parts of the same texts or other texts residing in different files (possibly in computers located on a different continent, even), allowing the user to follow different threads in reading the text. It also makes it possible to hide a lot of details, notes etc. that would only clutter up the main plane of the text, but it is just as powerful in linking in a lot of related information as well.

This mechanism alone will have proved a blessing for our purposes when we are dealing with such a richly structured set of data as the BSI. In addition, however, as a recent development, the technology has been developed to handle sound in the same way i. e. links embedded in the text can evoke the sound corresponding to it.[9]

In order to turn the BSI data into a hypertext system, we had to decide how to structure the data into separate units of reference (not necessarily residing in separate

[9] More precisely, any sound file that the link points to.

files but each accessible individually) and what navigational system to use to link them together. The structuring of the sound data has been discussed above in 5.2.1. The contents of the data files was structured into a menu system of three level depth. Figure 5.1 displays the main menu which provides access to the full data in chronological order (*Teljes anyag* 'full interview' option) as well as through the major task types *Irányított társalgás* 'guided conversation', *Kártyás feladatok* 'card based tasks', *Olvasási feladatok* 'reading tasks', *Ítéletalkotás* 'judgement tasks' options).

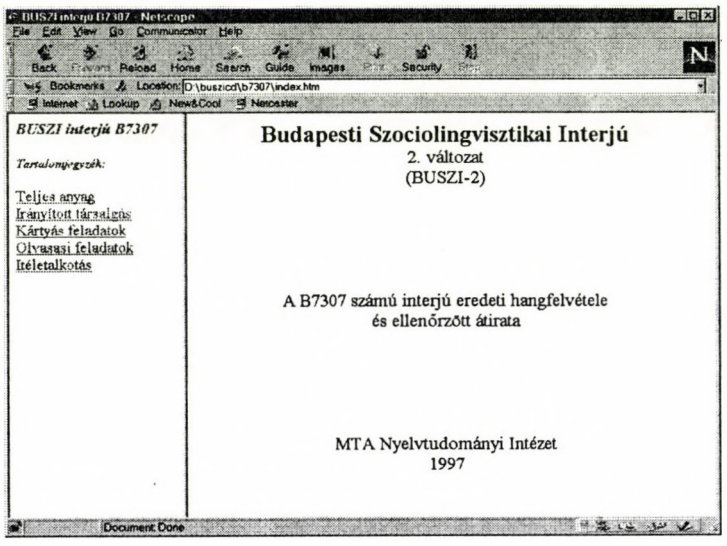

Figure 5.1: The title page and the main menu

If one clicks on the guided conversation module option, the left hand menu window is filled with the list of conversational modules that occurred in the particular interview and clicking further on any of them takes us to the beginning of the module in the transcript. This is displayed in Figure 5.2.

The following three figures each display the menu system and the format of various parts of the iterview. Figure 5.3 shows how the data tables are displayed. Figure 5.4 shows the same for judgement data and Figure 5.5 the menu and the different versions of a reading passage.

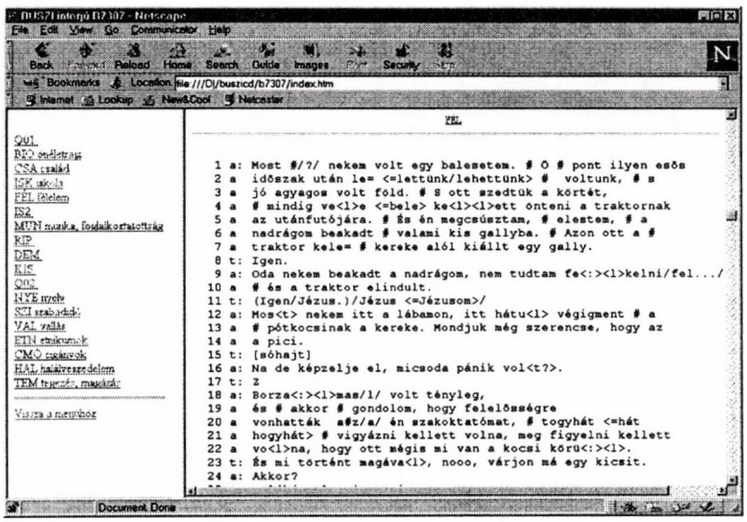

Figure 5.2: The menu system and format of the guided conversation modules

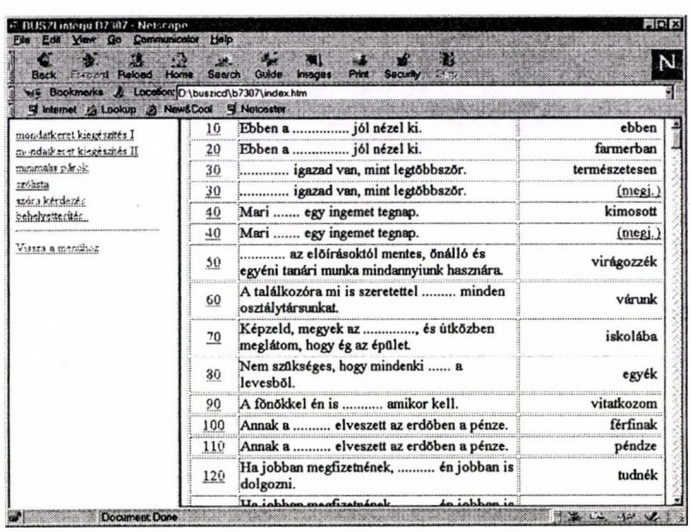

Figure 5.3: The menu system and format of the data tables

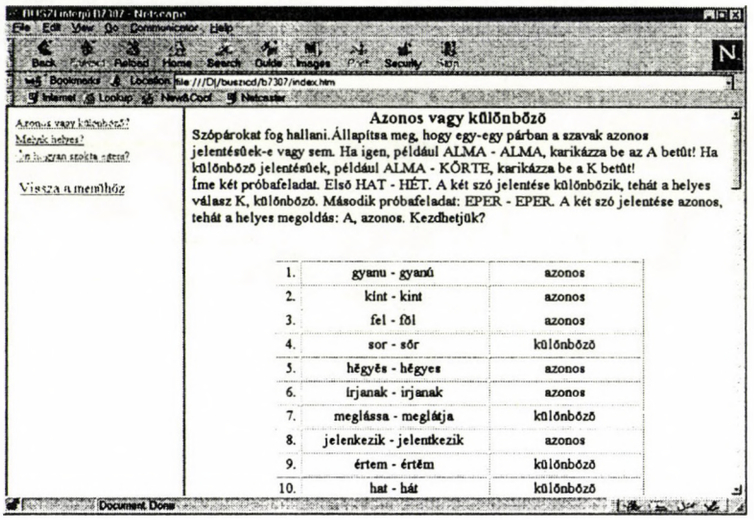

Figure 5.4: The menu system and format of the judgement data

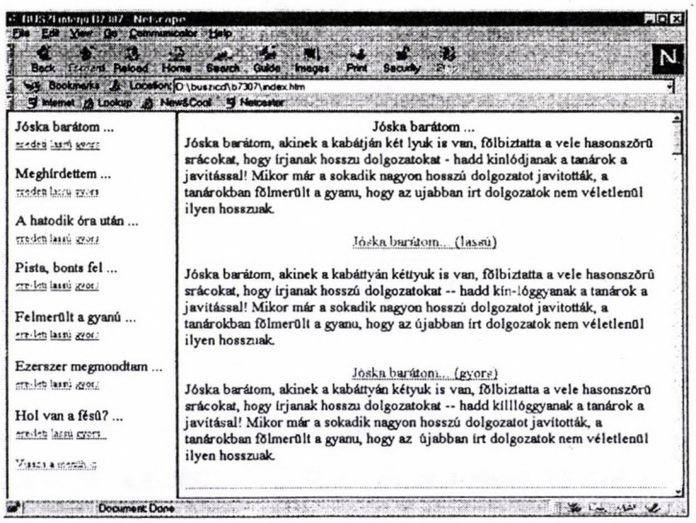

Figure 5.5: The menu system and format of the reading passages

6 Future work

6.1 Elaboration of the relational database system

The BSI database system in its current stage of development serves the purpose of accomodating the set of individual pieces of data recorded from each informant. This arrangement allows the preparation of certain summary statistics but when it comes to investigating linguistic questions in a general way, the data is difficult to handle. For example, in order to find out about a given research topic, say, *l*-deletion, one would have to cull all the items that this phenomenon was tested with from the list of the BSI items (as published in Váradi 1998) and build a query expression listing all the individual items collected. In addition, one would have to bear in mind not only which items relate to which research topic but also what the informants' numerically coded response actually stand for. In order to look up all the standard variants of a given sociolinguistic variable, one would have to know what number they were coded with for the particular item. (Given that there is a varying number of alternatives listed for the different items and they were coded in increasing order of likelihood of occurrence there is no guarantee that the standard forms would have the same numerical value.)

It follows from the above that in order to facilitate linguistically relevant queries the current database representation needs to be further elaborated. As a first step individual items will have to be assigned to the research topic(s) which they are meant to investigate. Next, the responses will have to be sorted out according to their linguistic relevance. At least, for those items where this dichotomy applies, the standard variants should be distinguished from the non-standard alternatives. These remarks are offered merely as suggestions of the work that lies ahead in this respect.

6.2 SGML coding

As described in Appendix A, especially in A.6, the transcript follows a rigid format where each line of text contains all the information necessary to identify the informant, the conversational module, the current speaker as well as the number of the line within the module. This essential reference information is adequate to unambiguously trace back any single line to its original context when the passages are input to a concordance program, which was the main application that the transcripts were anticipated to be

used with at the time that their format was designed.

In the meantime, the text formatting conventions have proved somewhat limiting on certain points. During the revision of the transcript it proved difficult to insert comments and corrections without disturbing the carefully laid out format. The format acts as a hindrance also when it comes to inserting grammatical annotations. Recall that the transcripts are riddled with codes that contain information about certain phenomena that the BSI project decided to collect in the conversation modules as well. (See Appendix A for a detailed list.) The guiding principle in devising the codes was that they should identify the research topic that the particular form exemplifies. For example, the code <*zuk*> is meant to indicate a hypercorrect use of the *-szuk/-szük* suffix. In this sense, the code system used in the transcript already goes some way in assigning actual variants to the linguistic phenomenon they belong to. Nevertheless, one can easily see the need to elaborate the system of cross-references between data and research topics, which would mean implanting further annotation in the text, something that can be done at the moment only at the expense of upsetting the layout of the transcript. More importantly, the BSI transcription rules represent in-house by-laws that require some effort to understand and to employ preventing the data from being portable, i. e. readily interpretable elsewhere.

Since the conception of the BSI project, however, guidelines have been worked out for the standardization of encoding of texts of all kinds, including spoken language. The recommendations which have been worked out as a result of several years' of international effort by experts from various fields, known as the Text Encoding Initiative (TEI) is now widely used (despite strong pockets of resistance with an obvious and respectible vested interest) and is quickly assuming the role of an international standard[1]. The TEI guidelines use a system of text annotation that has in fact been accepted as an international standard. This is SGML[2] (Standard Generalized Markup Language) which is quickly spreading in use as a world wide standard.[3]

SGML provides a simple but very powerful means of structuring the text into its logical components. One important principle is the separation of the logical structure of content from its layout and formatting, which is seen as a transient and replaceable surface feature. The logical structure of the text is defined in a separate file (called DTD, Document Type Description) much in the form of context-free grammar rules. The DTD specifies the main elements of a document and the hierarchical and sequential relationship between them. Each text file must obey the rules defined in the DTD that it belongs to or else the document is ill formed and will not be accepted by SGML processing tools.

The markup is inserted in the text in the form of tags which occur in pointed brackets. Tags are usually applied in pairs, one is used to mark the beginning and the other the end of the stretch of text that it refers to. For example,

[1]See C.M. Sperberg-McQueen & L. Burnard 1994).

[2]See C. F. Goldfarb 1990.

[3]HTML, a diluted SGML derivative, for example, is the language that made the current explosion of interest in the Internet possible.

<author>Aldous Huxley</author>
<title>Brave New World</title>
<genre>novel</genre>

Note the similarity between database fields as discussed in Chapter 3 and the units that occur in the above examples within tags. In fact, SGML files are actually textual database structures capable of representing highly complex hierarchical relationships in a flexible way.

One apparent drawback to SGML notation is lack of readability: a sufficiently rich SGML file may be so densely riddled with codes as to render the text completely beyond human consumption. However, with appropriate SGML processing software[4] the whole clutter of SGML annotation can be hidden or displayed at will. Also, because the formatting used in the source text will have no bearing on the final appearance of the document, the source text can be formatted in any way that reduces the problem. On the other hand, the same text can be assigned different layouts serving different purposes.

SGML provides the means and the mechanism of marking up the constituent parts of documents but leaves one free to decide how to apply the rules for a particular type of text. It is the TEI guidelines that contain recommedations as to how to structure a given type of text, what tags to use and how to relate the constituent units.

Adopting the TEI guidelines would bring obvious benefits in terms of portability of data. It would also introduce flexibility in revising and extending the trascript in that the transcript would not be subject to any rigid formatting constraint at all (except those relating to the syntax of the SGML tags themselves). At the same time, the SGML annotation would make it possible to describe the conversations in terms of their natural units, i.e. conversation turns. After all, text lines as units of transcription are arbitrary and artificial expediencies which can now be dispensed with altogether.

6.3 Implementing the database in a Client/Server setting

Converting the BSI data tables and integrating them in textual form together with the transcript in a common hypertext system as described in 5.2 is obviously a great help in providing access to the data. However, it has one crucial limitation. It gives a static picture of the whole of the data collected from a single informant. True, the hypertext navigation tools allow one to zoom in to any part of the data. Notice, however, that they can only take us to pages that are ready-made, prepared beforehand. What is lacking is the facility to make online queries and receive any kind of groupings of the data involving several informants or summary statistics computed on-line in response to a query.

Therefore, we need to develop the system to accommodate online queries. Fortunately, we can resort to the same technology described in 5.2 except that the hypertext system

[4] Apart from special SGML editors, WordPerfect 8 has a sophisticated SGML facility.

would function not merely to display static pages of data but also as an online interactive query tool mediating between the user and the data as well as displaying the result. The details of this process are too technical to go into here but the technique is widely used on the Internet. Consider, for example, how popular Internet search engines (like Excite, Yahoo etc.) are used. One fills in a form, submits it to the system and the result is displayed in the same browser window. What happens behind the scenes is that the request is forwarded to a program which then typically translates it into a database query, passes it to a database server, collects the response data and compiles the HTML files on the fly containing the data received from the database system. This chain of commuication is regulated by the so-called Common Gateway Interface, and the CGI programs mediate between the user, who typically uses an Internet client program (i. e. a browser), and the database system, operated as the server[5].

What remains to be done, then, is to set up the BSI database as a server and writing the CGI programs and the HTML pages that would allow querying the system through an Internet browser. We may use this setup to query the data that is available locally (on the same premises or on the same machine, for that matter). Using this technology even in such cases has obvious benefits. The user interface is familiar, intuitive, robust enough and, most importantly, comes free both for the developer and the user. At the same time, it also allows access to the data from any remote corner of the cyberspace at no additional programming expense.

[5]Hence the term Client/Server application.

A Transcription rules

A.1 Codes

A.1.1 Brackets

< > code

<= > explanation e.g. *közelibe<n> <=közeljövőben>*
More than one code are each put in separate < >.

[ ] extralinguistic remarks: the informant laughs, coughs, squints, tut-tuts, somebody enters the room etc. Also, tape counter setting to record beginning and end of long pauses (silence) and noises (see A.2.2).

A.1.2 The uncertainty of the transcriber

( ) Transcriber is not certain that s/he heard the form in ( ) e. g. *(Fönf) tanár úr.* This may occur inside word forms as well e. g. *sze(v)asztok.*
This code may be used in combination with others e. g. *(<d>)* means the transcriber is not certain s/he heard a case of *d* deletion, coded as <d> see A.1.12.

(.../...) Alternatives may be given e. g. *(Gondolom/tudom, hogy)*

Z Transcriber misses part of a word or an entire word. In case a sequence of words is not heard clearly, each misheard word should be marked with a Z, resulting in a string of Z's if necessary. E. g. *nem értem, hogy mi a Z mondasz.*

<?> Indicates inherent ambivalence in the data. The issue of what should be the standard form is unresolvable in the given context e.g. *ültettek egy diófát a kertbe<n?> Meghirdettek egy állást a Bécsikapu téri általános iskolába<n?>.*

A.1.3 Missing elements

<0 > Transcriber thinks the part following *0* should be obligatory in the standard variant e. g. *Persze <0az> az igazság, hogy még gyerek vagyok.*

<0a > article deletion

e.g. *Az volt a minihipotézisünk, hogy <0a> kontextus valamilyen módon befolyásolni fogja ezt a változást.*

<0e > *-e* interrogative particle

e.g. *Azt kell eldönteni, hogy a magyar köznyelvet akarjuk<0e> leírni.*

A.1.4 Pauses

□ Silent pause realised as a gap in the acoustic signal
e.g. *A □ másik dolog, amit nem tudtak □ megoldani □ a □ a □ köznyelvi □ gyűjtésben. □ ötvenhároméve megjelent nevezetes Bárczi-tanulmány*
It is marked inside the word form as well. e.g. *mi □t ö kutassunk*

ö For pauses realised as non-linguistic vocal (audible) phenomena, see hesitation Code (A.1.5)
Turn final pauses are not marked. The ":" at the head of a line (e. g. *t:*) indicates the beginning of an utterance and as such implies the presence of a pause. In case of continuous utterances stretching over several lines, this position is empty. If however, the speaker carries on without a pause after an overlapping speech, this is marked with a > instead of the colon (:).
Glottal stops are not taken down (either as pauses or when used literally).

A.1.5 Hesitation

ö short

öööö long

XXX lengthening as hesitation e.g. *aaalma, assszony, rrreális.* Whenever it is difficult to separate emotional lengthening from vocal hesitation as in ("kevésss, kevés"), they are to be classed lengthening as hesitation.

A.1.6 Non-conforming suffix

<H> Violation of vowel harmony rules.

e.g. *gyíknek <H> a farka*

A.1.7 Hypercorrect -ik *verb form*

<ik> *Nem szükséges, hogy a miniszter elvtárs minden kérdésre válaszoljék<ik>*

A.1.8 -suk/-sük, -szuk/-szük

<s> e.g. *Nem nyissa <s> <=nyitja> ki az ablakot.*

<suk> hypercorrection e.g. *Ne nyitja <suk> ki!*

<z> e.g. *Jóska felakassza <z> <=felakasztja> a kabátot.*

<zuk> hypercorrection e.g. *Ne akasztja <zuk> föl!*

N.B. Codes <s> (for *-suk/-sük*) and <z> (for *-szuk/-szük*) are obligatorily followed by explanation. Codes <suk> and <zuk> unambiguously stand for hypercorrect uses of-*suk/-sük* and *-szuk/-szük* respectively.

A.1.9 -nák

<nék> e.g. *én látnák<nék>*

A.1.10 -e *interrogative particle*

<e> *-e* not immediately following the predicate
e.g. *Nem-e <e> igaz, hogy*

<0e> missing *-e* (see A.1.3)

A.1.11 -ba/-be, -ban/-ben

<n> *-ba/-be* used instead of *-ban/-ben.* e.g. *Ebbe<n> az iskolába<n> tanítok.*

<ba> hypercorrection: *-ban/-ben* is used instead of *-ba/-be*
e.g. *Nem járok iskolában<ba>. Kijárok a temetőben <ba>*

A.1.12 l-, t-, d-*deletion*

<t> e.g. *mos<t>, jelen<t>kezik*

<d> e.g. *mos<d> meg*

<l> e.g. *jó<l> ismeri, kóstó<l>gatni, kolegái <=kollégái>*
if *l*-deletion results in compensatory lengthening, the lenghtening is marked as <:> except after long vowels where compensatory lenghtening is not transcribed. The code is not followed by any explanation. e.g. *fö<:><l>, ke<:><l><l>*

<z> As of BSI version 3 *z*-deletion is also marked, e.g. *szakszerve<z>eti.*

The shortening of phonologically long *l, t, d* is usually not transcribed, i. e. *kelett* is recorded in its standard form *kellett*, *nőtem* as *nőttem*. However, if the shortening results in a form that belongs to another lexeme, it is recorded in the shortened form and is followed by an explanation e. g. *halom* <= hallom>.

A.1.13 Consonant clusters

< - > In the card based TESTS cases of spelling pronounciations,[1] partial and full assimilation are all marked, e.g. *rácsszerű* <*cs-sz*>, *rácsszerű* <*c-sz*>,*rácsszerű* <*cc*>

In transcribing GUIDED CONVERSATIONS only spelling pronounciation is transcribed, and only word internally, e.g. *látja* <*t-j*>

<C> IN CLUSTERS OF THREE CONSONANTS

- AT MORPHEME BOUNDARIES deletion is marked, e.g.*min*<*d*> *négy*

- INSIDE MORPHEMES only lack of deletion, (i.e. spelling pronounciation) e. g. *mondta* <*d-t*>

The pronounciaton of foreign words (both in case of phonetic or spelling pronounciation) is explained and not standardized, e.g. *juice* <=*juice*>, *dzsúsz* <=*juice*>

A.1.14 Overlapping speech

Overlapping speech is transcribed within asterisks. The speech of the speaker who was speaking when the overlap began is transcribed till the end of the overlap. The beginning and end of the overlap is marked with an asterisk. Underneath follows the overlapping speech of the intervening speaker, also bounded by asterisks. If the second speaker takes over, his/her speech is transcribed continuously after the asterisk terminating the overlap. If the overlap is followed by the speech of the first speaker, then a new line is opened with the code of the speaker (*a* or *t*) followed by : if the first speaker paused or by > if s/he carried on without a pause, e.g.

```
a:   jók vo<:><l>tak a do<:><l>gozatok. □ Sza<l> ezér<t>,
a    *ezér<t> volt*
t:   *Igen*.
a>   nála kü<l>önösen *furcsa az, hogy*
t:   *Igen, □ igen*.
```

The * can be used word internally as well. Inside the word it is to be placed at syllable boundaries e.g.

```
t:   □ és ezzz □ nem volt megfelelő? □ Rosszul esett, □vagy
t    □ *nem tartotta megfelelőnek*?
a:   *Ez most □ a munkám*mal kapcsolatosan van, ugye?
```

[1] I.e. lack of consonant assimilation that takes place obligatorily and which is not marked by orthography.

If a word is broken up because of overlapping speech and is continued, both the ending of the first fragment and the beginning of the second is indicated with = e. g.

```
t: □ *nem tartotta megfelelõ*=
a: *Hát nem csak az*
t> =nek?
```

A.1.15 Slips, self-corrections, false starts

<= > Uncorrected slips e.g. *felmászott a látrán <=létrán>*

= a) Corrected slips e.g. *a fának a csa= csomója*

b) False starts, abandoned speech e.g. *akkor nyi= □ abból indulnék ki*

c) Correction of a suffix with another e.g. *kategóriánként =nak a*

... Abandoned phrases/sentences

a The structure is abandoned at a point where the last word is also abandoned: e.g. *abandoned word=...* or *abandoned word=...<=completed word>...*

b Completed word at the end of an abandoned phrase or sentence e.g.*ezek a gyerekek, hogy ...*

A.1.16 Response giving suffix only

-suffix e.g. *t: Az iskolától jössz?*
a: Nem, -ból.

A.1.17 Quotation

" " e.g. *valaki aszondja, hogy "kérem, van □ munkásnyelv Budapesten."*

A.1.18 Extralinguistic remarks

[ ] This is a rather mixed bag contaning information about (1) the informant' non-verbal behaviour (laughs, squints, coughs, tut-tuts), (2) events attending the interview (somebody enters, the phone rings, informant drops something etc.) (3) the length of long (>2 sec.) silences and noises (the tape counter setting marking the begining and the end of such stretches) as well as (4) technical remarks about the quality of the recording etc.

e.g. *mit ugatnak nekem azok a beszélők [nevet]*
a: Ha az a hülye szemüveges veszít is.
t:[nevet]

[nevet e] ... [nevet v] beginning and end point of laughter, e.g.
Hogyha valakinek van ideje me kedve a szabolcsi
[nevet e] ingázókat [nevet v]

The code [nevetve] ('laughing') is used to indicate that the word following the code is uttered in laughter.

A.1.19 Foreign words

<= > Foreign words are transcribed phonetically, if the form used by the informant does not agree with the orthographical form, it is explained e.g. *nyú jorki <=New York-i>*

A.1.20 Lenghtened variant consonants

They are standardized and not coded.

<ss> Except the lengthened *s* variant: the standardized form is followed by the code, e.g. *természetesen <ss>*

A.2 Instructions

A.2.1 Codes within words

The following codes can occur inside words: □, (), *

A.2.2 Long pauses

At the beginning and the end of long pauses and noises the tape counter setting must be recorded in [ ], see A.1.18.

A.2.3 Pronounciation variants: vowels

- In tests: transcribed
- In conversations: standardized including short/long variants
  e.g. *lakóság* → *lakosság*

  exceptions: – special words (see dictionary)
  – compensatory lenghtening (see A.1.20)
  – *e/ö* variants e.g. *fel - föl*
  – the *történetibe*-type.

A.2.4 Pronounciation variants: consonants

The following phenomena are standardized:
- shortening (see A.1.12)
- deletion (except *l, t, d*-deletion, see A.1.12)
- lenghtening (except *ss*, see A.1.20)

Compensatory lenghtening following vowel shortening (e. g. *szöllő, hüttő*) is not recorded.

A.2.5 Dialect speech

Distinctly dialectal features (such as diphthongization) should be recorded in the general profile of the informant. BSI transcripts only monitor *e/ö* usage.

A.2.6 □ö□ö□ö

□ö□ö□ö is recorded as many times as the informant utters it but continuous hesitation is transcribed as *öööö* (see A.1.5).

A.2.7 Syllable deletion

Deletion of one syllable is phonetically transcribed and then explained

e.g. *szöveki <=szövetkezeti>*

but: *szövetkeeti → szövetkezeti* (standardized and not explained).

BSI version 3 transcripts will transcribe not only syllable length deletion but also vowel deletion (including the concommittant deletion of neighbouring consonant(s) if any, e.g. *tulanképpen <=tulajdonképpen>*.

A.2.8 Pauses

keret□tet, but: *keret ö -tet* (pauses can be marked inside words, hesitation *ö* must be marked separately) see A.1.4 for how silence should be recorded.

A.3 Items to be standardized

1. Pronounciation variants (except A.2.3, A.2.4)
2. close *ë*; (except in card based test data)

A.4 Items not to be transcribed or standardized

1. every *-ja, -je* possessive suffix e.g. *ablaka-ablakja farka-farkja*
2. Mistakes in agreement

A.5 Dictionary (not to be standardised or explained)

- *ovoda, bölcsöde, kőrút, pósta, öntöde,*
- *mit tom én, asszem-asziszem, aszondja*
- *szal-szoal-sza-szoval*
- *kommonista, Ejrópa, inekció, Sofiane-Sofiané*
- *spré, sztressz*
- *gyün*
- *viszonlag*
- *má*
- *oszt <=aztán>*
- *mért, <=miért>* and derived forms (*mér, mé, miér, mié*).

A.6 Form conventions of transcribed text

A.6.1 Division of the transcription

Each conversation module forms a separate unit of text. Each unit has an identifier and a tape counter setting.

The identifier is made up of 8 characters, the first five of which is the ID of the informant, the rest is the three-letter code of the conversation module, e. g. B7307bio.

Important formal conventions:

- Each unit of text must be separated with (at least) one empty line.
- The first line of each unit must have the following data: Module Id (see above), tape counter setting of beginning and end of the module, transcriber ID, the dates when the transcription and checking were completed. This line should contain nothing else and each unit must be introduced with this header line. Each unit is numbered separately starting from the first line of the actual text, which is line 0001.

A.6.2 The format of text lines

Each line has 80 characters and they are used divided into the following fixed format:

columns	Content
1 – 5	identifier of the informant
6 – 8	identifier of the conversation module
10 – 13	line number within conversation module
15	identifier of current speaker[2]
16	continuity marker[3]
17 – 72	text
74 – 79	location on tape

Figure A.1 illustrate the above conventions. Transcribers were instructed to carefully observe the following points:

The body of transcribed text occupies character positions 17 – 72. The program breaks the lines automatically, so <ENTER> should only be used to insert empty lines to set off text units from each other.

Character position 16 is only indicated at the beginning of each turn. If the turn extends over several lines this position remains empty meaning there was no change of speaker.

Turns must not be separated with empty lines.

Transcribers only need to fill in the speaker and the continuity positions on the left margin. The identifier and the line numbers are supplied automatically. Tape counter setting should be recorded at roughly two minute intervals.

[2] *t* (terepmunkás) 'fieldworker' or *a* (adatközlő) 'informant'
[3] i.e. :=new turn, >=old turn continued

B7003bio 1a0042 RA 1988.07.18.
B7003bio 0001 a:Ott volt mint □ ö ált= állattenyésztési vezető, □ *s*
B7003bio 0002 t:*Igen.*
B7003bio 0003 a> aaa édesanyám pedig háztartásbeli volt. □ öö̈ö (a)
B7003bio 0004 a édesanyámnak a szakmája tanítónő volt □ valamikor, de
B7003bio 0005 a <0a> háború előtt tanított, □ <0a> háború után már nem
B7003bio 0006 a tanított.
B7003bio 0007 t:Igen, □ igen, □ értem. □ Namost, □ egészen tizennégy
B7003bio 0008 t éves koráig tehát akkor ott élt, ott lakott, ott járt
B7003bio 0009 t iskolába.
B7003bio 0010 a:Igen, igen.
B7003bio 0011 t:A Nyírségbe<n>, *ugye*? 1a0300
B7003bio 0012 a:*Igen.*

B7003vl1 1a0305 RA 1988.07.18.
⋮
1988.07.18.
B7003cmö 1a1529 RA 1988.07.19.
B7003cmö 0001 t:<0a> szomszédasszonyomnak már □ kinyitotta a táskáját.
B7003cmö 0002 a:Hát nekünk is □ az vo<l>t a szerencsénk, mer<t> a
B7003cmö 0003 a kolleganőnkkel mentünk az utcán,és □ aszondja nekem az
B7003cmö 0004 a Erika, hogy □ turkálnak a táskámba<n>. És hátranézek, és
B7003cmö 0005 a □ egy cigány férfi fogja a gyereke kezit, a másik
B7003cmö 0006 a kezivel az Erika táskájába nyú<l>ká<l>, a cigány nő az 1a1540
B7003cmö 0007 a ölébe<n> tartsa <s><=tartja> a gyerekét, és az én
B7003cmö 0008 a táskámba<n> *turká<l>*.
B7003cmö 0009 t:*Őrület*.

B7003vl2 1a1548 RA 1988.07.19.

Figure A.1: A sample page of transcription

Working Papers in Hungarian Sociolinguistics

No. 1: Pintzuk, Susan; Miklós Kontra; Klára Sándor; Anna Borbély. *The effect of the typewriter on Hungarian reading style*. September 1995.

No. 2: Kontra, Miklós & Tamás Váradi. *The Budapest Sociolinguistic Interview: Version 3*. December 1997.

No. 3: Váradi, Tamás. *From cards to computer files: Processing the data of The Budapest Sociolinguistic Interview*. January 1998.