Introduction to the ASU Corpus

a longitudinal oral and written text corpus of adult learner Swedish
with a corresponding part from native Swedes

Björn Hammarberg

Department of Linguistics, Stockholm University
ham@ling.su.se

Version 2010-11-16

CONTENTS

List of tables

1. What is the ASU Corpus?

The text corpus ASU consists of audiotaped and transcribed conversations and written essays in Swedish, produced by adult learners, as well as a comparable language material collected from native Swedes. The corpus is systematically constructed so as to serve as a basis for investigations of second language development and comparisons of learner and native language production. It documents the language of individual learners longitudinally at set intervals along a common time scale, so that it is possible to trace and compare stages of development within and between individuals. It is also intended to constitute a source for observing aspects of the acquisitional process in the second language.

The material was collected and edited within the project *Structural development of the second language (Andraspråkets strukturutveckling, ASU)* at the Department of Linguistics, Stockholm University, during the years 1990-93 and 1998. It can be accessed electronically from Språkbanken (The Swedish Language Bank), Department of Swedish Language, Gothenburg University with a special interface for searching and analysis developed in connection with the project *IT-based Collaborative Learning in Grammar (ITG).*

This introduction describes the design of the corpus in detail and explains the principles of transcription and tagging etc. It can be read both as a coherent account and for reference on specific matters. It serves two purposes:

- to provide basic documentation about the ASU Corpus to consult and refer to;
- to constitute the guide to the contents and structure of the corpus to which you need to have access while working with the corpus.

A separate practical manual (in Swedish), *Arbeta med ASU-korpusen*, explains how to work with the corpus in the ITG interface.

2. Directions for the design and a general characterization of the corpus

The construction of the corpus was guided by a number of demands and desiderata regarding the kind of material that was needed. The fundamental intention was to design a corpus which can document the dynamics and development of the learner language and its relation to the target language. A basic decision was also to focus on adult language. These premisses have given rise to several specific criteria which we have tried to satisfy. Those criteria at the same time describe some essential characteristics of this corpus.

Focus on individuals. The purpose of making the corpus longitudinal (in its learner part) motivates the requirement to document the language at the level of the individual. It should be possible to investigate the language of each individual, to compare a person with him/herself over time, and to compare the linguistic solutions and developmental profiles of different persons. Therefore, the corpus should contain relatively much material from relatively few persons – ”much from few” rather than ”little from many”.

Learner type. We should base the corpus on learners who have a clear motivation to acquire the language and who strive to communicate in the second language at their own intellectual level. Persons who aim at a developed second language are thus preferred to persons who remain using the second language for basic communicative needs.

Development. It is essential to get access to a kind of learner language which changes markedly over time and to observe it at short intervals, so that changes and stages can be located in time.

Span of stages. The corpus should ideally cover how an elaborate language evolves from early forms. The ambition is to document the elementary initial phase and the more advanced stages in a coherent span in the same persons. The starting-point should preferably be at the zero stage.

Speech and writing. What differences and similarities arise between spoken and written production? We should take the opportunity to register speech and writing in parallel in the same persons in a common time sequence. This will make it possible to study the relation between speech and writing per person and per point in time.

Native usage. A corresponding corpus part from native Swedes should be built up in a way similar to the learner part. It should as far as possible be comparable to the learner part, with comparable informants, the same methods of data collection and corresponding contents. One difference is of course that the native part does not register a language which develops over time, but represents a target language variety statically.

Internal comparability. As we know, learner languages are characterized by a complicated combination of systematicity and variation. A general purpose is to organize the corpus in such a way as to allow comparisons in different dimensions:

- Longitudinally: across acquisitional stages
- Interindvidually: between individual informants
- NNS/NS: between learners and native speakers
- Medium: between speech and writing
- Genre: between different activities in speech, or types of text in writing

3. The parts of the corpus – overview

The corpus is divided into ***four major parts*** according to informant category (***learners–native***) and medium (***oral–written***). This is shown in Table 1.

Within each main part, the text material is arranged according to ***person***, and for each person according to ***session***, ***text unit*** and ***chronological sequence in the text***. The oral corpus has one text unit (one recorded conversation) per session; the written corpus has two text units (essays) per session. See Table 1.

The texts are thus stored in the order of main part > person > chronology.

Table 1. The four main parts of the corpus, subdivisions and extent.

	Learner	***Native***	***Total Lnr+Nat***
Oral	10 persons x 10 sessions = 100 text units, ca 269,000 / 147,000 word tokens[1]	7 persons x 5 sessions = 35 text units, ca 149,000 / 98,000 word tokens [1]	ca 418,000 word tokens
Written	10 persons x 11 sessions x 2 texts = 220 text units, ca 50,000 word tokens	7 persons x 5 sessions x 2 texts = 70 text units, ca 25,000 word tokens	ca 75,000 word tokens
Total ASU			ca 493,000 word tokens

1 The numbers refer to: the whole dialogue / the informants' utterances

4. The learner part

4.1. Informants

Ten participants in the preparatory course in Swedish for foreign students at Stockholm University served as informants in the learner part of the corpus. They took part in repeated sessions from the start of the beginner course, but their participation took place outside the course and separately from it.

The students were recruited to the corpus project on the basis of (1) information which they had given before the course about their linguistic background, (2) talks with them in which the project and its purpose was explained, and (3) their own wish to take part. Those students who took part in the project were placed in the same class, an arrangement which kept the course factor constant and highly facilitated the organization of regular recording and essay-writing sessions. This class followed the normal schedule and routines of the preparatory course entirely, without any modifications due to the corpus project. It was also an explicit part of the arrangement that their participation and linguistic performance in the corpus project should not be used in the evaluation of them in the course.

General characterization of the learner category

Living at the time in the Stockholm area and receiving their linguistic input partly from the course and partly from the outside environment, these students can broadly be described as *'semi-formal learners'*. They were *'qualified learners'* in the sense that they all had secondary education, earlier experience with foreign languages, and a strong instrumental motivation to learn the language of the host country in order to proceed with the studies in their fields. Relatively speaking, they can be categorized as *'fast learners'*, since they advanced from the beginner stage through or close to the level required for university studies in Swedish within one to two years.

Information on individual informants

Information about the informants is summarized in Table 2. The individual learners are identified in the corpus by a code consisting of *letter+digit*, where the letter represents the learner's first language.

Aspects of homogeneity and heterogeneity

The group of informants is relatively *homogeneous* in some respects (in addition to being 'semi-formal', 'qualified' and 'fast' learners as defined above):

- *Age:* young adults, aged 19-20, median 20½.
- *Social class:* middle class.
- *Previous exposure to Swedish:* All except E2 and P1 can be regarded as pure beginners at the start of the project, having spent only few days in the country up to that point. (Even Q2 had had little contact with Swedish during his two months in Sweden before course start.) E1 and P1 had had some exposure to Swedish through contact with Swedes, which can be observed in the corpus, but were still placed in the beginner class.
- *Course progression:* All were enrolled in the same Swedish class during the first academic year, with the same teacher and course materials, following the same time schedule.
- *Prior L2s:* All had some prior knowledge of English, in accordance with general requirements for university studies in Sweden.
- *Study interests:* All had come to Sweden to pursue a professional education for a period of several years. Occurring study interests were economics, medicine, technology and film. None of them aimed at language studies after the preparatory course.

In some other respects the group of learners display *variation*:

- *First language:* The corpus shows an intentional spread from distant to close languages, in terms of the genetic, geographical and typological relation between the first language and Swedish. Three main groups are Chinese, Greek and Portuguese/Spanish speakers (together comprising the 8 learners who were absolute beginners at the start). In addition, there are a Polish and a bilingual German-English speaker. Note, however, that there has been no intention to provide for comparisons of learner groups defined by L1; rather, the corpus was designed for comparisons between individual interlanguages.
- *Cultural background:* While all informants were raised in urban environments, cultural backgrounds vary according to home country. Observations of language performance in the corpus suggest that a broad division between Europeans (E2, G2, G3, P1, Q1) and non-Europeans (C1, C2, C4, Q2, S1) is sometimes relevant. This appears to correlate with progress in Swedish.
- *Rate of progress in L2 and attained proficiency:* While all learners were recorded according to a common time schedule, there is a spread between more and less fast and successful learners. Both strong and weak learners are included among the informants. As shown in Table 2, some needed a third term of Swedish before passing the National Proficiency Test, the language requirement for learners of Swedish to be permitted to enter regular university or college courses (the test later named Tisus). In one case (Q2), this test was not passed within the first two years.

Table 2. Summary information about the native informants in the ASU Corpus.

Person	Sex	Age at start[1]	Raised in	First language	Prior L2 knowledge[2]	Length of stay in Sweden before start[1]	Course participation (terms)[3]	Passed Nat'l Proficiency Test
C1	F	22	China: Shanghai	Chinese (Shanghai & Mandarin)	English H Japanese L-M French L	24 days	A 90 S 91 A 91	Apr 92
C2	M	20	China: Shanghai	Chinese (Shanghai & Mandarin)	English M French L	19 days	A 90 S 91	Aug 91
C4	F	20	China: Beijing	Chinese (Mandarin)	English M	10 days	A 90 S 91 A 91	Oct 92
E2	M	28	Austria, India, Kenya	German & English	Swahili M	2 months	A 90 S 91	May 91
G2	M	22	Greece: Athens	Greek	English M Russian L	9 days	A 90 S 91	May 91
G3	M	19	Greece: Thessaloniki	Greek	English M French M	23 days	A 90 S 91	May 91
P1	F	20	Poland: Gdansk	Polish	English H Russian H German L Italian L	6 months	A 90 S 91	May 91
Q1	M	23	Portugal: Coimbra	Portuguese	English H French H Spanish M	10 days	A 90 S 91	Aug 91
Q2	M	21	Mozambique (urban)	Portuguese[4]	English M Spanish L French L	2 months	A 90 S 91 A 91	–
S1	M	20	Bolivia: Santa Cruz	Spanish	English M Portuguese L	10 days	A 90 S 91	Dec 91
Total	3 F 7 M	Mdn 20½						

1 "Start" refers to the course start on August 27, 1990.
2 Proficiency levels based on informants' statements: H = high; M = medium; L = low.
3 A = autumn term (Aug – Dec); S = spring term (Jan – May).
4 Q2 states only scanty knowledge of indigenous Mozambican languages.

4.2. Material

4.2.1. Method of data collection

The corpus material was collected separately from the language course during special recording and essay-writing sessions, and the corpus-constructing program was not connected with the coursework.

The ***oral*** material was collected through audio-recorded studio interviews with one informant at a time and one or two native Swedish interlocutors. The oral corpus units thus have the form of NS–NNS interaction sequences, the resulting texts containing a combination of learner and native production.

The interviews took place in the recording studio of the Stockholm University Language Laboratory (Lärostudion) on the Frescati campus. The speakers were seated at a round table with a microphone for mono recording hanging over it and the informant with her/his back toward the technician's window. The area at the table was lit and the periphery of the room toned down, thus providing a calm and undisturbed environment for the conversation. The sessions lasted 25 to 30 minutes and were recorded simultaneously on a DAT and a standard cassette.

The ***written*** material was collected during group sessions in a classroom. On each occasion, two hours were set off for the composition of two separate texts on given topics. This provided ample time for the essays that were written, and in most cases the writers used less time. The topic of each essay was handed out on a sheet of paper, with a given title and, where appropriate, a short instruction and/or some picture material. The essays were usually composed in a spontaneous way, without much reflection and without any extensive re-check. The informants wrote the texts by hand.

4.2.2. Extent and distribution over time

The oral part comprises 100 audio recordings, 10 with each of the 10 informants. The total time amounts to about 50 hours, 5 hours with each informant. The oral text measures ca 269,000 word tokens in total, out of which ca 147,000 constitute the learners' utterances.

The written part comprises 220 essays, 22 from each informant, written on 11 occasions, and totalling ca 50,000 word tokens.

Table 3 gives an overview of the distribution of the corpus units over time. The rounds of recordings and essay-writing are called *Time 1*, *Time 2* (*T1, T2* ...) etc. The units of the corpus which were collected on those occasions are labelled *M1*, *M2* etc. for the oral part (*M* for 'muntlig') and *S1*, *S2* etc. for the written part (*S* for 'skriftlig'). Tables 4 and 5 show the distribution in detail with the exact date of each oral and written session, respectively.

The first 9 occasions (T1–T9) extend over the academic year 1990/91, i.e. the period when all the informants participated in the same language course; T10 and T11 are follow-up sessions in the second and third spring term, respectively, T11 comprising only a written part. Thus T1–T10 constitute ten sets of oral and written units, collected in parallel, the oral part usually one or two weeks before the corresponding written part.

Table 3. Distribution of the learner corpus over time – overview.

Time	Aug Sep 90				Nov Dec 90	Feb 91			May 91	Mar Apr 92	Apr 93
Occasions (rounds)	T 1	T 2	T 3	T 4	T 5	T 6	T 7	T 8	T 9	T 10	T 11
Oral (M) and written (S) sessions											
Oral	M 1	M 2	M 3	M 4	M 5	M 6	M 7	M 8	M 9	M10	-
Written	S 1	S 2	S 3	S 4	S 5	S 6	S 7	S 8	S 9	S 10	S 11

Table 4. Time distribution of recordings per session and person. Dates (yymmdd).

Occasion	Person									
	C1	C2	C4	E2	G2	G3	P1	Q1	Q2	S1
M1	900905	900905	900905	900829	900829	900829	900829	900905	900905	900829
M2	900926	900926	901003	900919	900919	900919	900919	900926	900926	900919
M3	901017	901017	901024	901010	901010	901010	901017	901010	901010	901010
M4	901107	901107	901114	901031	901031	901031	901107	901031	901031	901031
M5	901128	901128	901128	901121	901121	901121	901121	901121	901121	901121
M6	910212	910205	910212	910204	910129	910129	910212	910205	910205	910204
M7	910311	910312	910311	910311	910305	910305	910305	910312	910312	910311
M8	910415	910416	910423	910423	910409	910409	910409	910416	910423	910415
M9	910513	910514	910513	910514	910507	910507	910507	910514	910514	910513
M10	920318	920331	920401	920424	920320	920318	920318	920320	920320	920318

Table 5. Time distribution of essays per session and person. Dates (yymmdd).

Occasion	Person									
	C1	C2	C4	E2	G2	G3	P1	Q1	Q2	S1
S1	900912	900912	900912	900912	900912	900912	900912	900912	900912	900912
S2	901003	901005	901003	901003	901003	901003	901003	901003	901003	901003
S3	901031	901026	901024	901026	901024	901024	901024	901024	901024	901026
S4	901114	901114	901114	901119	901114	901114	901114	901119	901114	901119
S5	901205	901205	901205	901205	901205	901205	901205	901205	901205	901205
S6	910211	910211	910211	910211	910211	910211	910211	910215	910215	910211
S7	910318	910325	910318	910318	910318	910318	910318	910318	910318	910318
S8	910422	910422	910422	910422	910422	910422	910422	910422	910422	910422
S9	910521	910521	910903	910521	910521	910521	910521	910527	910521	910521
S10	920327	920331	920401	920424	920327	920327	920409	920327	920327	920327
S11	930430	930430	930430	930430	930430	930430	930430	930430	930513	930430

4.2.3. Distribution of contents

The corpus has a strictly parallel design. All learners were given the same tasks during the corresponding M or S sessions. The purpose of this is to permit comparisons within the corpus, as outlined in Section 2 above.

Tables 6 and 7 summarize the activities and informants that were dealt with in the M and S sessions.

Table 6 provides an overview of the various types of activities performed during the recorded oral sessions. Various tasks alternated during the conversations:

- narration of picture stories (wordless cartoons)
- description and discussion of photos
- description of (the appearance and function of) physical objects, such as kitchen and office utensils
- interviews, mostly focusing on the informant's current situation, experiences and views
- discussion of newspaper articles; the informant was required to read the previous day's *Dagens Nyheter* in advance and to choose an article for discussion; the article formed the point of departure for a discussion which usually soon evolved into a freer conversation.

Some items, especially picture stories, were repeated on later occasions in order to permit diachronic comparisons. Likewise, the interviews sometimed returned to previous topics.

Table 7 gives a similar summary of the S sessions, i.e. the types and topics of the written essays. Each time, two separate texts were written on given topics. Here, too, some repetitions of topics were planned in. Thus, one pucture story was repeated twice in S sessions, two picture stories were used in both M and S sessions, and both topics of S11 were repetitions from S7.

Table 6. Overview of the contents of the oral learner part of the corpus. Distribution of activity types and topics

	Picture story	Photos	Objects	Interview	Newspaper
M 1	The dog	Class-teaching situations		Personal interview	
M 2	The fly	Students on campus Characteristic face		Visit to Skansen	
M 3	The party			Describing an imaginary person	
M 4	The dog Totte simulates being ill			My study interest	
M 5	Incident on the subway train			About the term that has passed	
M 6			Household utensils	About the Christmas vacation	Self-chosen article
M 7			Office utensils		Self-chosen article + The family page
M 8	The party				Self-chosen article
M 9	The dog			Plans for the future	Self-chosen article
M 10	The party		Office utensils	Things done during the past year Plans for the future The Swedish language	

Table 7. Overview of the contents of the written learner part of the corpus. Distribution of the types of essays and topics

	Picture story	Narration	Description	Discussion	Other
S 1			A guest student in Stockholm		Questions to a Swedish student
S 2	The fly	My first day in Sweden			
S 3	Adventure in a cinema	A vacation trip			
S 4			The University at Frescati	My study interest	
S 5	The subway: 1. The woman's story 2. The man's story				
S 6				If I was George Bush [discussion about the Gulf War 1991]	My newspaper reading
S 7			The family page in a daily newspaper	Raising children yesterday, today and tomorrow	
S 8	The fly			The migration and the Swedish society	
S 9				Argument for/against Sweden joining the EG	Advice to a countryman [about the encounter with Sweden]
S 10	The fly			The life of young people in Sweden and in my home country	
S 11			The family page in a daily newspaper	Raising children yesterday, today and tomorrow	

5. The native part

The native part was collected in 1998, after the learner corpus had been completed and edited. This facilitated the task of matching the native and learner corpus parts in terms of type of informants, method of data collection, and contents, since the research team could use their experience from the construction of the learner corpus as a frame of reference in building up the native part.

5.1. Informants

For the native part of the corpus, seven young undergraduate students at Stockholm University were recruited. Care was taken to obtain a type of native Swedish informants who corresponded as closely as possible to the type of young foreign students who served as informants in the learner part.

The native informants were all born and raised in Sweden with Swedish as their sole L1. They are identified as Z1 to Z7 in the corpus. Table 8 summarizes some individual data on these informants.

As Table 8 shows, there are four women and three men, aged 20–29, median 23. They spoke Stndard Central Swedish (although slightly coloured by a northern accent in the case of Z5 and Z7. They each possessed from two to four foreign or second languages to varying levels of proficiency. Their current study interests were in the areas of philosophy, literature, art history and social anthropology.

Tabell 8. Summary information about the native informants in the ASU Corpus.

Person	Sex	Age	Raised in	L2s (in order of proficiency)
Z 1	M	20	Stockholm	English, German, Spanish
Z 2	F	20	Stockholm	English, French
Z 3	F	26	Uppsala	English, Spanish, French
Z 4	F	20	Stockholm suburban area	English, German, Chinese
Z 5	M	23	Piteå	English, German, Russian, Finnish
Z 6	F	29	Stockholm suburban area	English, French, German, Icelandic
Z 7	M	25	Skellefteå	English, German
Total	4 K, 3 M	Mdn 23		

5.2. Material

5.2.1. Method of data collection

The corpus material was collected in a similar way as with the learner informants, as far as possible.

The oral material was collected through interviews with one informant at a time and two native Swedish interlocutors. The two interviewers were the same as with the learner informants. The interviews took place in a studio in the Phonetics Laboratory at the Department of Linguistics, Stockholm University. The participants were seated around a low table with one microphone for the informant and one for the interlocutors, for stereo recording. The informant was seated with her/his back toward the technician's window, so as not to get distracted. The sessions lasted 25 to 30 minutes and were recorded simultaneously on a DAT and a standard cassette.

The written material was collected during group sessions in a classroom. On each occasion, two hours were set off for the composition of two separate texts on given topics. The procedure was the same as with the learners (see 4.2.1 above).

Five oral and five written sessions were carried through in parallel, with intervals of one week.

5.2.2. Extent and distribution over time

Five audio recordings of ca 25–30 minutes were made with each of the seven native informants, totalling 35 recordings. The transcribed oral text measures ca 149,000 word tokens in total, out of which ca 98,000 words constitute the informant's utterances. The written part contains 10 essays from each informant, i.e. 70 altogether, amounting to ca 25,000 words running text.

The five oral and five written sessions with each informant are labelled *M1* to *M5* (oral) and *S1* to *S5* (written).

5.2.3. *Distribution of contents*

The native corpus part is strictly parallel internally in that all informants were given the same tasks on the corresponding M or S sessions. The contents in terms of activity types and topics is similar, and in part identical, to those in the learner corpus, but the native informants went through a shorter programme.

The activities and topics covered in the M and S sessions with the native informants are summarized in Tables 9 and 10. As can be seen, some topics were brought up both in the M and the S sessions.

Table 9. Overview of the contents of the oral native part of the corpus. Distribution of activity types and topics

	Picture story	Photos	Objects	Interview	Newspaper
M1	The dog	Pictures of two large families		Personal interview: personal data, study plans, interests	
M2	The fly		Kitchen utensils	Something done last year	Self-chosen article
M3	The party			Describing an imaginary person The Iraqi crisis	Self-chosen article
M4	Incident on the subway train		Office utensils	An interest for study or leisure time	Self-chosen article
M5	Totte simulates being ill	African story-telling scene		Thoughts about the Swedish language	Self-chosen article

Table 10. Overview of the contents of the written native part of the corpus. Distribution of the types of essays and topics

	Picture story	Narration	Description	Discussion	Other
S1	Adventure in a cinema		The University at Frescati		
S2		A vacation trip		Raising children yesterday, today and tomorrow	
S3	The fly			The migration and the Swedish society	
S4				My study interest	Advice to a foreign guest student
S5	The subway: 1. The woman's story 2. The man's story				

6. The appearance of the text in the computer-stored corpus

Transcription and grammatical tagging

The oral (***M***) and written (***S***) texts were typed on computer after collection. The principles adopted for the ***transcription*** are described in Section 7 below. The utterances by the informants, but not those by the interviewers, have been ***word-tagged*** according to a system described in Section 8 below.

Form of storage and user interface

The corpus is stored in an XML-based format and can be accessed from Språkbanken, Department of Swedish Language, University of Gothenburg, through a user interface (***the ITG interface***) which has been developed in connection with the project *IT-based Collaborative Learning in Grammar (ITG)*. This interface makes it possible, among other things, to work interactively with concordances and to save work results. How to work with the user interface is described (in Swedish) in a separate work manual ***(Arbeta med ASU-korpusen)***. The program requires for the time being that you have installed the software *Java* on your computer. (Regarding access, see below, Section 10.)

How to identify the texts in the corpus

Each transcribed recording and each essay constitutes a ***text unit***. The text units are identified by a letter-and-digit code which indicates ***Person + Medium + Session + Essay 1 or 2*** (the last only in the written corpus). Examples: C1M8 = Person C1, Oral, Session 8; G3S052 = Person G3, Written, Session 5, Essay 2. In the oral corpus, Session 10 is written as Roman X, e.g., E2MX = Person E2, Oral, Session 10.

The edited text

In the text, as it is presented to the user, some structuring information has been added:

The text has ***a fixed, permanent division and numbering of lines*** which can be cited in order to refer to ***locations*** in the corpus. The information on location is the necessary instrument for identifying different individuals, learner/native informants, speech/writing, the chronology of text units and the time sequence in the texts. Each line has ***line header*** in the left margin, which identifies the location by text unit and line in the text, e.g., C1M3 0067, Z3S041 016. You can thus read person, medium and chronological location directly in the line header.

Every new turn in the oral corpus is introduced by a ***speaker identification***, where ’I’ always indicates the informant in question, and ’B’ and ’E’ are the two native Swedish interlocutors (interviewers).

Comments have been added to the text on separate, inserted lines introduced by ’C’. These lines are not counted in the numbering of text lines.

The display of full texts (the function *Textvisning* in the ITG interface) also shows a ***text header*** before each text unit with information about the text unit, and for the oral texts also added ***sub-headings*** which distinguish different activities in the recorded dialogues and thereby structure the contents of the texts.

The *text header* which precedes the text proper contains the date of the recording or essay, as well as a given title for the essay or other information about the contents.

The *sub-headings* in the oral texts appear on inserted comment lines (C-lines) and are introduced by '>>>' in order to mark them as sub-headings in contrast to other comment lines. A number of ***terms for activities*** are used regularly in the sub-headings, so as to facilitate the identification and comparison of corresponding text passages across the text units. Such recurring terms are, e.g., BILDSERIE 'picture story', FOTO 'photo', FÖREMÅL 'object', INTERVJU 'interview', TIDNING 'newspaper' (cf. Tables 6 and 9 above). The particular object or topic spoken about is also mentioned here.

The text headers and sub-headings are only provided together with the display of full texts and are not included in the context field which appears with concordances. However, the *date* of each recording or essay can be seen in Tables 4 and 5 above, and *topics for activities and essays* are given in Tables 6 and 7 (learners) and 8 (native informants).

7. Principles of transcription

7.1. Transcription of the oral material

General principles:

A model has been devised for the needs of the oral ASU Corpus. It aims to capture the grammatical and lexical structure of the text as well as aspects of the utterance planning and the dialogue structure. It is *not* a transcription at phonetic level or a detailed rendering of the physical shape of the utterances. But it intends to account for the lexical and morphological form in which the words occur (in particular the occurrence of inflection) as clearly as possible.

A modified version of standard orthography is used throoughout for the Swedish text (and standard orthography is used for words in other languages). The modifications consist in (a) rendering (lexically/morphologically) deviant and reduced forms of words in the way they occur, and (b) writing the common 'spoken language forms' of words (to the extent that the speakers use them), such as e.g.:

de	for standard written	*det*	
dom		*de, dem*	
nåra		*några*	
e		*är*	
va		*vad, var*	
ja		*jag, ja*	
å		*och, att*	etc. etc.

Phonetic variants are not distinguished, but comments on prounciation are added on comment lines in some places where this is judged relevant. Prosodic speech features are not marked in the transcription, but prosody may certainly constitute a criterion for the syntactic or pragmatic interpretation of the utterance and thus influence the transcription.

Major structural units in the text:

Major types of units in the text which are set off by the transcription are the *word*, the *macrosyntagm* and the *turn*.

In the use of characters, a distinction is upheld between ***'words'*** and ***'non-words'***. 'Words' are the lexical units in the text. 'Non-words' include markings of syntactic boundaries, pauses and pausefillers and various pragmatic information (as shown in the key to symbols below). This distinction is achieved by the use of *lower-case letters* for 'words' and *other symbols than lower-case letters* for 'non-words'. Note, for example, that we avoid writing the pausefiller sound (a paralinguistic element) as 'eh' or 'öh' or the like, but represent it by the symbol '%'. The purpose of this principle is to make the *wording* of the speaker's text easier to discern when reading the transcripts.

Capital letters are used for inserted comments and subheadings (which are put on special C lines) and also for non-turnbreaking feedback insertions (back-channels) from an interlocutor (see the discussion of turns below). Capital letters are also used for comments of the type 'SKRATT', 'TYST', also locally within the turn. Text in capital letters is not included in word counts.

We have chosen to base the syntactic segmentation of the oral text on the concept of ***macrosyntagm (MS)*** (Loman & Jörgensen 1971). The MS is in its typical form a main clause with its embedded sub-clauses, if there are any. Coordinated main clauses are considered separate MSs, but with reduction of identical parts they are treated as one MS. Other types of MS are interjection and address MS, and sentence fragments. See Loman & Jörgensen (1971) for definitions and discussion. (Terms corresponding to macrosyntagms in analyses of English texts are 'T-units' (Hunt 1966; Richards et al. 1985); for spoken text 'C-units' (Biber et al. 1999: 1069).)

An MS is treated as *completed* if it does not lack any obligatory final part, and if the prosody does not suggest that it is uncompleted. The completed MS is demarcated by '?' (for question-MSs) or '.' (for other MSs). Syntactically *interrupted* sequences are demarcated by '/' at the interruption point. A characteristic type of sequence is *the completed macrosyntagm with its preceding interrupted starts (if any)*. This type of sequence (which may thus contain one or more '/', as well as pauses, pausefillers and other pragmatic elements) will reflect both features of the sentence planning and the resulting sentence. It can be regarded as the (unedited) counterpart in the spoken text to the (edited) sentence in the written text.

The ***turn*** is marked by starting on a new line introduced by a speaker identification (in a separate column).

The *division of turns* presents some problems. There are the following types of cases:

(a) *Regular turnshift*: new speaker, new turn.

(b) *Feedback insertion* from an interlocutor, non-turnbreaking, merely confirming or supportive, prompting the speaker to continue the turn, with no further semantic content: The insertion is written inside the turn, in parentheses ’( )’, in capital letters, with a letter identifying the interlocutor. For example, ’... (B: JAHA) ...’.

(c) *Intervention* by an interlocutor (with some semantic content, as opposed to a mere supportive feedback insertion of type (b)), *although the first speaker continues the turn*: This is written out as a turnshift. The interrupted turn is marked with ’\’ at the interruption point. The interlocutor’s utterance is put on a new line, as a separate turn. The first speaker’s continuation is then written out as a subsequent turn, introduced by ’\’. The ’\’ marks thus signal that there is a ’turn in the turn’.

There is no transcription device for simultaneous (overlapping) speech by two or more speakers. Such instances may be described separately on comment lines, but this is used restrictively. In uncomplicated cases of overlap of a turn-end and a turn-start, speakers’ turns are usually written out in succession.

Special symbols used in the M-transcription:

=	Empty pause.
==	Longer pause. (Long silence is noted on a C line.)
%	Pausefiller. (Replaces letter notations such as ’eh, ’ah’, ’öh’.)
%%	Long pausefiller.
xxx	Unidentifiable or murmured sequence. (Possible and plausible interpretations may be noted on a C line.)
-	After an interrupted word, or before a separate endpart of a word.
+	After a morphologically unclear form.
/	Syntactic interruption or break for re-planning or repair. Not used if identical forms are merely iterated.
\	Turn connector. Used at the end of the first part of a turn and at the beginning of the continuation, in case of an intervention by an interlocutor in the middle of an ongoing and continuing turn. (See turn division above.)
?	Question mark. After a completed question-macrosyntagm.
.	Period. After a completed macrosyntagm other than a question.
¿	Questioned expression. The speaker is seeking feedback on a word or phrase by an intonation indicating a metalinguistic question: ’Is this right?’
” ”	Quotes. Around direct quotations.

< > Around non-Swedish words or text sequences.

( ) Around capitalized text within speakers' turns: feedback insertions '(B: MHM)' or comments such as '(SUCKAR)', '(SKRATTAR)'.

7.2. Transcription of the written material

The transcription of the written corpus basically consists in copying the writer's handwritten text and adding the required C lines as described in section 7.1.1. However, a few normalizations have been applied:

. A period is always put at the end of macrosyntagms, unless the writer uses '?' or '!'. That is, periods are supplied if left out where there is obviously the end of an MS. Periods are removed from within macrosyntagms, for example when used with abbreviations. Thus, periods (as well as question and exclamation marks) are only used as MS delimiters.

xxx Unidentifiable sequence.

8. The tagging system

The informants' production has been word-tagged. That is, the words in the informants' part of the spoken dialogue and the entire written corpus has been supplied with morphological tags which indicate part of speech and certain grammatical sub-categories. The Swedish interlocutors' part of the dialogue has not been tagged. Punctuation marks have also been tagged, even in the interlocutors' part. In the ITG user interface it is possible to search for both word/punctuation forms and tags.

A general problem with pre-established tag systems is that users may wish to categorize their data according to special criteria motivated by the current research objective, criteria which a given tag system, even if detailed, is not likely to satisfy. ASU, being a limited and highly structured corpus, yields data which may be rewarding to analyse closely. The principle which is applied in the ASU Corpus is, on the one hand, to have a simple and rather broad tag division as a basis for searches. On the other hand, there are facilities for finer categorization and sorting of data according to freely chosen criteria during the interactive work with concordances in the ITG user interface.

The present ASU tag system comprises ca 50 different tags for parts of speech and certain grammatical and other categories and punctuation items. An overview is presented in Table 11. The specified list which then follows presents the tag categories with selected examples, and some comments on the tagging practice are given.

Table 11. The tags used in the ASU Corpus grouped according to grammatical categories. The tag categories are explained in the list below in the numerical order of the table with selected examples. The numbers in the table should be read from line to column, e.g. N = 1.1, EN = 1.2.

	.1	.2	.3	.4	.5	.6	.7	.8	.9	.10	.11
1. *Nouns*	N	EN									
2. *Pronouns & Articles*	P	PO	RP	ROBJ	FS	KP	DEM	PIF	REL	PÖ	ART
3. *Question words*	FRÅ	FRÅK									
4. *Quanti-fiers*	QU	OT									
5. *Verbs*	VS	VT	VINF	VSUP	VPC	VI	V	KOP	KOPT	KOPI	KOPÖ
6. *Verb particles*	PT										
7. *Infinitive mark*	IM										
8. *Adjective*	A										
9. *Adverb*	ADV	ADVG	ADVK	KADV							
10. *Con-junctions*	K	UK									
11. *Pre-positions*	PR										
12. *Inter-jections*	IJ										
13. *Funda-ment marker*	FM										
14. *Subject marker*	SM										
15. *Other wordtags*	U	I	XXX	X	Tag + X	Tag + Z					
16. *Punctu-ation*	del	syntBreak	pause								

Specified list of tags:

The closed classes are here rather richly exemplified in order to show how the tags are used, but this is not an exhaustive list. Note that grammatical function words are often polyfunctional, i.e. the may occur under more than one tag.

1. Nouns

1.1. **N** Nouns except proper names

1.2. **EN** Proper names

2. Pronouns and articles

2.1. **P** Personal pronoun
de dej dem den det dig dom du er han henne hon honom ja jag mej mig ni oss vi

2.2. **PO** Possessive pronoun
dens deras dess din dina ditt er era ert hans hennes min mina mitt vår våra vårt

2.3. **RP** Possessive reflexive pronoun
sin sina sitt själv

2.4. **ROBJ** Reflexive object pronoun
sej sig

2.5. **FS** Formal (dummy) subject
de det

2.6. **KP** Comparative pronoun
andra annan annat annorlunda ena fler flest likadan mer mest nästa olika samma sista sådan sådana sådant sån sånt

2.7. **DEM** Demonstrative pronoun
de den denna dessa det detta dom
NB! *den här* DEM U; *så här* KP DEM

2.8. **PIF** Indefinite pronoun
en inga ingen ingenting inget man någon någonting något några nån nånting nåra nåt sånt

2.9. **REL** Relativizer
som vilka vilken vilket

2.10. **PÖ** Other pronouns
egen eget egna enda varandra

2.11. **ART** Article

de den det dom en ett

3. Question words

3.1. **FRÅ** Interrogative adverb
hur när va vad var varför varifrån vart vem vilka vilken vilket
Used for question words in direct questions; cf. REL (2.9), FRÅK (3.2).

3.2. **FRÅK** Interrogative subjunction
hur när va vad var varför varifrån vart vem vilka vilken vilket
Cf. FRÅ (3.1).

4. Numerals and other quantifiers

4.1. **QU** Quantifier
all alla allt båda hela lite många mycke mycket några + cardinal numbers written in letters or figures.
Figures are tagged with QU when they stand for cardinal numbers (incl. years).
If they are ordinal numbers (e.g. *den 24 december*), the tag OT is used.
Dates written in figures: e.g. *1997-12-24* QU OT OT; *24/12 -97* OT OT QU.

4.2. **OT** Ordinal numbers
andra första tredje
Cf. QU (4.1).

5. Verbs

5.1. **VS** Verb present
blir bor bör finns får förstår går gör har kan mår måste ska skall står tror vet vill +er +ar

5.2. **VT** Verb past
blev fanns fick gick hade kom kunde skrev skulle stod såg tog ville +de +te

5.3. **VINF** Verb infinitive
bli bo få förstå ge gå se stå tro +a

5.4. **VSUP** Verb supine
fått gjort gått haft kommit köpt läst sagt sett skrivit tagit

5.5. **VPC** Verb participle
The tag VPC is used for both present participles and perfect participles.
The categories VPC and A (Adjectives, 8.1) are hard to separate; the tagging has been largely intuitive on the basis of the meaning in the context: if the word was felt to retain a ’verby’ character, VPC was used, otherwise A. Checking instances is recommended.

5.6. **VI** Verb imperative

5.7. **V** Verb ‘bare’ form

Used when the bare root form of the verb occurs, lacking e.g. an infinitive -a or a tense suffix. Ex: *<u>vänt</u> din brev* [an early-stage learner sentence].

5.8. **KOP** Copula present
e är

5.9. **KOPT** Copula past
va var

5.10. **KOPI** Copula infinitive
va vara

5.11. **KOPÖ** Copula other
varit vore

6. Verb particles

6.1. **PT** Particle
av bort fast fram ihop ihåg in me om på till upp ut

7. Infinitive mark

7.1. **IM** Infinitive mark
att å

8. Adjective

8.1. **A** Adjective
Cf. VPC (5.5).

9. Adverb

9.1. **ADV** Adverb
aldrig alldeles allti alltid alltså bara bredvid brevid då där egentligen endast faktiskt förstås förut genast heller här ibland idag igen inte ju kanske längre nere nu nästan också ofta precis redan sedan sen så till tillbaka ungefär uppe ändå ännu även

9.2. **ADVG** Adverbs of degree
ganska lite mycke mycket väldigt

9.3. **ADVK** Adverbial connector
annars då först nu sedan sen så

9.4. **KADV** Comparative adverb
för lika mer mera mest så

10. Conjunctions

10.1. **K** Coordinating conjunction
eller fast för men och så å utan

10.2. **UK** Subordinating conjunction (Subjunction)
att då därför eftersom fastän innan medan när om som trots än

UK is used for 3 types of *som*:
1. Comparative (*lika bra som*)
2. Predicative (*som studerande* ...)
3. A few instances where *som* has been used in the sense of *att*

UK is used together with U (see the comments on the tag U below) in phrase-formed subjunctions with *att*.
Examples of the distinction between K and UK in some cases:

K	UK U
för	*för att*
så	*så att*
-	*därför att*

11. Prepositions

11.1. **PR** Preposition
av bakom bland bredvid efter enligt framför från för genom hos i inom me med mellan mot om på ti till under ur utan utanför utav vid åt över
för ... sedan PR ...U

12. Interjections

12.1. **IJ** Interjection
adjö ah aha förlåt goddag hej hm hä ja jaha jaså jo mh mmh ne nej nå nä oj okej precis tack varsågod visst å åh

13. Topical adverb markers

13.1. **FM** Topical adverb marker
då så

14. Subject marker

14.1. **SM** Subject marker
som Ex: *jag vet vem som* ...

15. Other

15.1. **U** Unspecified
See comment on the tag U below.

15.2. **I** Interrupted word
Used for interrupted word parts, transcribed with ’-’ at the end.
Separate endparts of words, transcribed with ’-’ at the beginning, can usually be assigned to a grammatical category and are tagged accordingly.

15.3. **XXX** Unidentifiable
Used for unidentifiable sequences, transcribed as *xxx*.

15.4. **X** Untaggable
Used for audible/readable forms which are not possible to assign to a tag.

15.5 **Tagg+X** Uncertain tagging
If the category cannot be safely determined, the most plausible tag is chosen, and X is added to it (e.g. VSX = VS-uncertain).

15.6 **Tagg+Z** Non-Swedish word
Used in the *oral corpus* for words in other languages than Swedish, i.e. with phonologically and morphologically non-adapted code switches and stretches of text in another language than Swedish. Such words and sequences are surrounded by < > in the M transcripts. If possible, the non-Swedish words are tagged in a way corresponding to Swedish words, adding Z. Ex: <*airport*> is tagged NZ.

16.1 **del** Syntactic delimiter; turn connector; questioned expression
Delimiters . ! ? , ; : ” ( )
The parenthesis symbols here refer to parentheses in the writer's own text in the S corpus, not to parentheses around feedback insertions anc comments in the M corpus.
Turn connector \
Questioned expression ¿

16.2 **syntBreak** Syntactic break
/

16.3 **pause** Pause and pausefiller
= == % %%

*Comment on the use of the tag **U**:*

The tag U (Unspecified) is used din phrasal, hyphenated or otherwise divided expressions which we have regarded as fixed lexical phrases and subjected to *'synthetic tagging'* (as opposed to the normal *'analytic tagging'* in which each word receives a tag for a specific category). Such phrases are assigned as a whole to a category. Only the first word receives the specific tag, and U is used for the other word(s) in the phrase.

In the following examples the underlined words get the tag U, and the non-underlined words carry the specific tag shown on the left, which then represents the whole phrase. When the tag U is shown for a word, look for the nearest word on the left with a specific tag.

N *student rum, T-banan, bord-tennis, film star, del A, 1960-talet*

EN *saddam hussein, lars-åke, gula villan, kafe bojan, dagens nyheter, USA:s*

DEM *det här*

QU *en del, ett par, halv nio, tjugo tusen, (klockan) 8:50*

A *jätte svårt*

ADV *i morgon, i stället, till exempel, i alla fall, så här, hela tiden, då och då, för det mesta, för det andra, först och främst, framför allt, helt enkelt, där inne, mer eller mindre*

UK *därför att, för att, så att, som om*

PR *för ... sedan, i och med, på grund av, vad beträffar, tack vare*

IJ ja visst, god natt, hej då

9. The creation of the corpus, project funding and participant persons

The corpus was built up in its original form during the 1990s within the project *Andraspråkets strukturutveckling (ASU)* 'Structural development of the second language' at the Department of Linguistics, Stockholm University, led by Björn Hammarberg and with financial support from the Swedish Research Council for the Humanities and Social Sciences (HSFR) and the Faculty of Humanities, Stockholm University. The corpus has later undergone a thorough technical modernization in order to adjust it to modern standards for linguistic text corpora. This has taken place with the support of Magn. Bergvalls Stiftelse, Birgit & Gad Rausings Stiftelse för Humanistisk Forskning and Henrik Granholms Stiftelse.

The material for the learner part was collected in 1990–1993 and was transcribed, tagged and edited. In 1998 the corpus was supplemented by the material from the native Swedish informants, which was processed in the same way. Initially the corpus was stored in an ASCII format and was edited and prepared by means of the corpus software *PC Beta* (Brodda 1982, 1991). This corpus version was the basis for articles and dissertations in the 1990s and the first years of the present century.

During this phase the following persons participated in the construction of the corpus:

Niclas Abrahamsson (transcription of the oral learner part; tagging of the entire learner part; corpus editing; software development)
Maria Arnstad (transcription of the oral native part)
Dorothée Augustin (recruitment of native informants; studio assistance)
Christina Ericsson (transcription and tagging of the oral native part)
Björn Hammarberg (design and direction; interviews and essay tasks; transcription of the written learner part; corpus editing)
Eva Klingberg Merk (course instructor; interviews and essay tasks; transcription of the written learner part)
Ulrika Kvist Darnell (transcription of the written native part)

Benny Brodda served as consultant for the text processing. He supplied the software PC Beta with the tagging program PC Tagger, developed special program devices for the needs of the ASU Corpus and taught members of the project group to handle the software.

Kenneth Andersson (learner part) and *Hassan Djamshidpey* (native part) served as studio technicians during the recordings.

The technical modernization of the corpus has consisted in converting the entire corpus text into an XML format and connecting it to a user tool for searching and analysis (*the ITG interface*) which was developed in the project *IT-based Collaborative Learning in Grammar (ITG)*. The ITG project is affiliated with the Department of Linguistics and Philology at Uppsala University (*Anju Saxena*), and the ITG tool is based and developed at Språkbanken (The Swedish Language Bank) at Gothenburg University (*Lars Borin*). In this connection, user functions which meet the special needs of the ASU Corpus have been devised. Here *Lars Borin* has been responsible for the technological aspects and *Björn Hammarberg* for the SLA research needs. The programming has been carried out by *Camilla Bengtsson* (text conversion) and *Leif-Jöran Olsson* (the ITG interface and adjustments in the database).

10. Access to the ASU Corpus

How to access the corpus

The corpus is available for searches and analyses by means of the *ITG interface*, which is based at The Swedish Language Bank (Språkbanken), Department of Swedish Language, Gothenburg University. At present, the interface is available in Swedish only.
To access the ITG interface, proceed as follows:

- Download the software *Java* on your computer. (Free of charge from http://www.java.com.)
- To get access to the ASU Corpus via ITG, visit http://spraakbanken.gu.se/itg and follow the instructions there. (For the time being, they are in Swedish only.)
- When you have opened the ITG interface, expand the node *Korpus* on the main menu and click the sub-node *Korpussökning* in order to chose texts. (Other corpora, too, are accessible here.) This content manual and the work manual *Arbeta med ASU-korpusen* which you reach from the ITG interface, will be of use to you during your work.

What will you have access to?

You will be free to perform searches in the transcribed corpus, create concordances and get frequency information, study the search hits in context, edit and save concordances, and export and print out frequency lists, concordances and examples. You must comply with the conditions below.

Researchers who wish to access the transcribed full text are advised to contact Björn Hammarberg at ham@ling.su.se.
For questions about the ITG user interface, contact Språkbanken by e-mail sb@svenska.gu.se.

Conditions for using the ASU Corpus

The user is required to accept the following conditions:

- General: The ASU Corpus is intended for research and education. It is subject to copyright. The integrity of the participant informants has to be preserved.
- You are not allowed to distribute texts from the corpus for commercial purposes, or distribute them so that they can be used commercially. However, examples and context extracts may be reproduced in accordance with common scientific quotation practice.
- When material from the corpus is rendered, in publications or otherwise, the ASU Corpus, Department of Linguistics, Stockholm University shall be cited as source. (The present *Introduction to the ASU Corpus* can be referred to and quoted for orientation and detailed information about the corpus.)
- The identity of the informants has to be protected. They are anonymized in the corpus, but should clues to their identity still occur, these are not to be disclosed. The informants' personal dignity must always be respected.

References

Biber, D., S.Johansson, G. Leech, S. Conrad & E. Finegan (1999) *Longman Grammar of Spoken and Written English*. Harlow, Essex: Pearson Education.

Brodda, B. (1982) Problems with tagging – and a solution. *Nordic Journal of Linguistics*, 5:93-116.

Brodda, B. (1991) Do corpus work with PC Beta and be your own computational linguist. In *English Computer Corpora. Selected Papers and Research Guide*, ed. by S. Johansson & A.-B. Stenström. Berlin & New York: Mouton de Gruyter.

Hunt, K.W. (1966) Recent measures in syntactic development. *Elementary English*, 43:732-739.

Loman, B. & N. Jörgensen (1971) *Manual för analys och beskrivning av makrosyntagmer*. Lund: Studentlitteratur.

Richards, J., J. Platt, & H. Weber, (1985) *Longman Dictionary of Applied Linguistics*. Harlow: Longman.