Introduktion till ASU-korpusen

en longitudinell muntlig och skriftlig textkorpus av vuxna inlärares svenska
med en motsvarande del från infödda svenskar

Björn Hammarberg

Institutionen för lingvistik, Stockholms universitet
ham@ling.su.se

Version 2010-11-16

INNEHÅLL

Förteckning över tabeller

1. Vad är ASU-korpusen?

Textkorpusen ASU består av ljudinspelade och transkriberade samtal och skrivna uppsatser på svenska producerade av vuxna inlärare, och därtill ett jämförbart språkmaterial insamlat från infödda svenskar. Korpusen är systematiskt uppbyggd för att tjäna som underlag för undersökningar av andraspråksutveckling och jämförelser av inlärares och inföddas språkproduktion. Den dokumenterar språket hos individuella inlärare longitudinellt med bestämda tidsintervall, så att man kan spåra och jämföra stadier i utvecklingen inom och mellan individer. Den ska också kunna tjäna som källa för att observera aspekter av tillägnandeprocessen i andraspråket.

Materialet insamlades och bearbetades inom projektet *Andraspråkets strukturutveckling (ASU)* vid Institutionen för lingvistik, Stockholms universitet åren 1990-93 och 1998. Det är tillgängligt elektroniskt via Språkbanken, Institutionen för svenska språket, Göteborgs universitet med ett särskilt anpassat sök- och analysverktyg utarbetat i anslutning till projektet *IT-based Collaborative Learning in Grammar (ITG)*.

Den här introduktionen beskriver korpusens inriktning och uppbyggnad i detalj och förklarar principerna för transkription och taggning m.m. Den är tänkt att kunna läsas både sammanhängande och uppslagsvis. Den tjänar två syften:

- att vara en grundläggande dokumentation om ASU-korpusen att konsultera för orientering och att referera till;
- att utgöra den vägledning till korpusens innehåll och utformning som man behöver ha till hands under arbetet med korpusen.

En separat bruksanvisning, *Arbeta med ASU-korpusen*, förklarar hur man arbetar praktiskt med korpusen i ITG-gränssnittet.

2. Riktlinjer för uppbyggnaden och allmän karakteristik av korpusens inriktning

En rad krav och önskemål har fått vara vägledande för hur korpusen skulle läggas upp. I botten ligger det grundläggande syftet att få fram en korpus som inriktas på att dokumentera inlärarspråkets dynamik och utveckling och dess förhållande till målspråket. Ett grundläggande val har också varit att den ska inriktas på vuxna personers språk. Ur detta ger sig ett antal specifika kriterier, som vi har sökt tillgodose. De kan samtidigt sägas beskriva några väsentliga karakteristika för den här korpusen.

Individinriktning. Ur syftet att göra korpusen longitudinell (i sin inlärardel) ger sig kravet att den ska dokumentera språket på individnivå. Det ska i möjligaste mån gå att få en bild av språket hos varje individ för sig, att jämföra individerna med sig själva över tiden och att jämföra olika personers språklösningar och utvecklingsprofiler. Korpusen bör därför innehålla relativt mycket material från relativt få personer – ”mycket från få” hellre än ”lite från många”.

Inlärartyp. Korpusen bör baseras på inlärare som har en klar motivation att tillägna sig språket och som strävar att kommunicera på andraspråket på sin egen intellektuella nivå.

Personer som har sikte på ett utvecklat andraspråk föredras alltså framför personer som stannar vid att utnyttja andraspråket för elementär baskommunikation.

Utveckling. Det gäller att få tillgång till ett inlärarspråk som förändras påtagligt över tiden, och att observera det vid flera tillfällen med korta intervaller, så att förändringar och stadier kan tidsfästas.

Stadieomfång. Det är ett mål att söka täcka framväxten av ett utvecklat språk från tidiga former. Ambitionen är att kunna belägga den elementära initiala fasen och de mer avancerade skedena i ett sammanhängande spann hos samma personer. Startpunkten bör ligga på noll-stadiet.

Tal och skrift. Vilka skillnader och överensstämmelser uppstår mellan talad och skriven produktion? Här ges ett värdefullt tillfälle att registrera muntlig och skriftlig produktion parallellt hos samma personer i en gemensam ordnad tidssekvens. Därigenom kan relationen tal/skrift studeras per person och per tidpunkt.

Införddas språkbruk. En motsvarande korpusdel från införda svenskar bör byggas upp på likartat sätt som inlärardelen. Den bör så långt möjligt göras jämförbar med inlärardelen, med jämförbara informanter, samma insamlingsmetoder och motsvarande innehåll. En skillnad är att den införda delen ju inte registrerar ett språk som utvecklas över tiden, utan statiskt representerar en målspråksvarietet.

Intern jämförbarhet. Inlärarspråk präglas ju av en komplicerad kombination av systematik och variation. En genomgående målsättning är att organisera korpusen så, att det blir möjligt att inom den göra jämförelser i olika dimensioner:

- longitudinellt över inlärningsstadier
- mellan olika individer
- mellan inlärare och införda
- mellan tal och skrift
- mellan olika aktiviteter i tal och mellan olika slags text i skrift

3. Korpusens delar – översikt

Korpusen är indelad i ***fyra huvuddelar*** efter informantkategori (***inlärare–införda***) och efter medium (***muntligt–skriftligt***). Tabell 1 visar de fyra huvuddelarna.

Inom varje huvuddel är textmaterialet indelat efter ***person***, samt för varje person ***kronologiskt*** efter ***tillfälle, textenhet*** och ***tidsföljd i texten***. Den muntliga korpusen har en textenhet (ett inspelat samtal) per tillfälle; den skriftliga korpusen har två textenheter (uppsatstexter) per tillfälle. Se tabell 1.

Texterna är således lagrade i ordning efter huvuddel > person > kronologi.

Tabell 1. Korpusens fyra huvuddelar och deras underindelning och omfång.

	Inlärare	***Infödda***	***Summa Inl+Inf***
Muntligt	10 personer x 10 tillfällen = 100 textenheter, ca 269 000 / 147 000 löpord[1]	7 personer x 5 tillfällen = 35 textenheter, ca 149 000 / 98 000 löpord[1]	ca 418 000 löpord
Skriftligt	10 personer x 11 tillfällen x 2 texter = 220 textenheter, ca 50 000 löpord	7 personer x 5 tillfällen x 2 texter = 70 textenheter, ca 25 000 löpord	ca 75 000 löpord
ASU totalt			ca 493 000 löpord

1 Uppgiften avser: hela dialogen / informanternas yttranden.

4. Inlärardelen

4.1. Informanter

Som informanter i inlärardelen tjänstgjorde tio deltagare i preparandkursen i svenska för utländska studenter vid Stockholms universitet. De medverkade vid återkommande tillfällen alltifrån starten av nybörjarkursen, men deras medverkan skedde utanför kursen och fristående från denna.

Studenterna rekryterades till korpusprojektet på basis av (1) den information de givit före kursstarten om sin språkliga bakgrund, (2) samtal där projektets syfte och uppläggning förklarades, samt (3) deras egen önskan att medverka. De studenter som ingick i projektet placerades i samma klass, ett arrangemang som höll kursfaktorn konstant och i hög grad förenklade organiserandet av regelbundna inspelnings- och skrivsessioner. Denna klass följde helt och hållet preparandkursens vanliga schema och rutiner, utan några modifikationer på grund av korpusprojektet. Det var en uttrycklig del av arrangemanget att studenternas deltagande och språkliga prestationer i korpusprojektet inte skulle användas i bedömningen av dem i kursen.

Allmän karakteristik av inlärarkategorin:

I och med att informanterna bodde i Stockholmsområdet under projekttiden och fick sin språkliga input dels från kursen och dels från den omgivande språkmiljön, kan de grovt beskrivas som *”semi-formella inlärare”*. De var *”kvalificerade inlärare”* i den meningen att de alla hade gymnasieutbildning, tidigare erfarenhet av främmande språk, och en stark instrumentell motivation att lära sig värdlandets språk för att kunna bedriva studier inom sina fackområden. Relativt sett kan de kategoriseras som *”snabba inlärare”*, i och med att de avancerade från nybörjarstadiet till eller näst intill den nivå som krävs för högskolestudier på svenska inom ett till två läsår.

Information om individuella informanter:

Information om informanterna sammanfattas i tabell 2. De enskilda inlärarna betecknas i korpusen med en kod bestående av bokstav+siffra, där bokstaven representerar inlärarens förstaspråk.

Aspekter av homogenitet och heterogenitet:

I några avseenden är informantgruppen relativt *homogen* (förutom att alla är ”semi-formella”, ”kvalificerade” och ”snabba” inlärare enligt definitionerna ovan):

- *Ålder:* unga vuxna, 19-28 år, median 20½.
- *Samhällsklass:* medelklass.
- *Tidigare vistelse i svensk språkmiljö:* Alla utom E2 och P1 kan betraktas som rena nybörjare i svenska vid kursstarten, då de dittills bara hade varit några få dagar i Sverige. (Även Q2 hade haft mycket ringa kontakt med svenska under sina två månader i Sverige före kursstarten.) E2 och P1 hade haft en del informell svensk input genom kontakt med svenskar, vilket kan observeras i korpusen, men kursledningen bedömde ändå att de hörde hemma i en nybörjarklass.
- *Kursprogression:* Alla placerades i samma svenskklass under det första läsåret, med samma lärare, kursmaterial och schema.
- *Tidigare L2-kunskaper:* Alla hade en viss behärskning av engelska, i enlighet med vad som generellt krävs för högskolestudier i Sverige.
- *Studieintressen:* Alla hade kommit till Sverige för att genomgå en professionell utbildning under en period av några år. Förekommande studieintressen var ekonomi, medicin, teknologi och film; ingen av de tio var inriktad på språkstudier efter preparandkursen.

I några andra avseenden uppvisar inlärarna en *variation* sinsemellan:

- *Förstaspråk:* Korpusen representerar avsiktligt en spridning från avlägsna till nära språk med hänsyn till den genetiska, geografiska och typologiska relationen mellan förstaspråket och svenskan. Tre huvudgrupper bildas av kinesiska, grekiska och spanska/portugisiska förstaspråkstalare, vilka tillsammans utgör de 8 inlärare som var rena nybörjare vid starten. Dessutom finns en polskspråkig inlärare (P1) och en inlärare (E2) med tyska hemifrån och en under barndomen förvärvad tvåspråkighet tyska-engelska. *Observera dock* att korpusen *inte* är avsedd att ge underlag för att jämföra grupper av inlärare definierade per L1; korpusen är designad för jämförelser mellan individuella inlärarspråk.
- *Kulturell bakgrund:* Samtidigt som alla informanterna var uppvuxna till väsentlig del i urbana miljöer, så varierar den kulturella bakgrunden alltefter hemland. Iakttagelser av språkliga drag i korpusen ger vid handen att en grovindelning mellan européer (E2, G2, G3, Q1, P1) och utomeuropéer (C1, C2, C4, Q2, S1) ibland är relevant; en sådan gruppering tycks kunna korrelera med framstegstakten i svenska.
- *Framstegstakten i svenska och uppnådd färdighet:* Under det att alla informanterna observerades med likvärdiga tidsintervall under en gemensam projektperiod, finns det en spridning mellan mer och mindre snabba och framgångsrika inlärare. Både starka och svaga inlärare finns representerade i informantgruppen. Som framgår av tabell 2, behövde några en tredje termin för att klara rikstestet, språkkravet för utländska studenter för att få behörighet för högskolestudier på svenska (det test som senare har benämnts Tisus). I ett fall (Q2) klarades inte rikstestet under de två första åren.

Tabell 2. Summariska uppgifter om inlärarna i ASU-korpusen

Person	Kön	Ålder vid start[1]	Uppvuxen i	Förstaspråk	Tidigare L2-kunskaper[2]	Vistelselängd i Sverige före start[1]	Kurs-deltagande (terminer)[3]	Klarat rikstest
C1	K	22	Kina: Shanghai	Kinesiska (Shanghai & Mandarin)	Engelska H Japanska L-M Franska L	24 dagar	H 90 V 91 H 91	Apr 92
C2	M	20	Kina: Shanghai	Kinesiska (Shanghai & Mandarin)	Engelska M Franska L	19 dagar	H 90 V 91	Aug 91
C4	K	20	Kina: Beijing	Kinesiska (Mandarin)	Engelska M	10 dagar	H 90 V 91 H 91	Okt 92
E2	M	28	Österrike, Indien, Kenya	Tyska & Engelska	Swahili M	2 månader	H 90 V 91	Maj 91
G2	M	22	Grekland: Athen	Grekiska	Engelska M Ryska L	9 dagar	H 90 V 91	Maj 91
G3	M	19	Grekland: Thessaloniki	Grekiska	Engelska M Franska M	23 dagar	H 90 V 91	Maj 91
P1	K	20	Polen: Gdansk	Polska	Engelska H Ryska H Tyska L Italienska L	6 månader	H 90 V 91	Maj 91
Q1	M	23	Portugal: Coimbra	Portugisiska	Engelska H Franska H Spanska M	10 dagar	H 90 V 91	Aug 91
Q2	M	21	Moçambique (stad)	Portugisiska[4]	Engelska M Spanska L Franska L	2 månader	H 90 V 91 H 91	–
S1	M	20	Bolivia: Santa Cruz	Spanska	Engelska M Portugisiska L	10 dagar	H 90 V 91	Dec 91
Summa	3 K 7 M	Mdn 20½						

1 ”Start” avser kursstarten den 27 augusti 1990.
2 Kunskapsnivåer baseras på informanternas egna uppgifter: H = hög; M = medel; L = låg.
3 H = hösttermin (aug/sept – dec); V = vårtermin (jan – maj).
4 Q2 uppger bara helt rudimentära kunskaper i inhemska moçambikiska språk.

4.2. Material

4.2.1. Insamlingsmetod

Deltagarnas språkproduktion samlades in separat från språkkursen under särskilda inspelnings- och skrivsessioner, och korpusuppbyggnaden var inte knuten till inlärarnas arbete i kursen.

Det ***muntliga*** materialet samlades in genom ljudinspelade intervjusamtal med en informant åt gången och en till två infödda svenska samtalspartner. Den muntliga korpusen har således formen av interaktioner mellan infödda och inlärare (NS–NNS), och de resulterande texterna innehåller därmed en kombination av produktion från infödda talare och inlärare. (I analyser av den färdiga korpusen kan man skilja mellan att arbeta med dialogtexten eller med inlärarens produktion separat.)

Intervjuerna ägde rum i Lärostudions inspelningsstudio, Stockholms universitet. De samtalande satt kring ett runt bord med en mikrofon för monoinspelning hängande från taket över bordet och med informanten med ryggen vänd mot fönstret till teknikerrummet. Området vid bordet var upplyst och rummets periferi var nedsläckt, så att en lugn och störningsfri miljö för samtalet bildades. Samtalen varade 25–30 minuter och inspelades samtidigt på DAT och standardkassett.

Det ***skriftliga*** materialet insamlades under skrivsessioner gruppvis i en lärosal. Vid varje tillfälle var två timmar avsatta för att skriva två uppsatstexter om givna ämnen. Det innebar gott om tid för de uppsatser som kom att skrivas, och i de flesta fall använde skribenterna mindre tid. Ämnet för varje uppsats gavs på en lapp, med en given rubrik och i förekommande fall en kort instruktion och/eller ett bildunderlag. Uppsatserna författades vanligen på ett spontant sätt, utan mycket funderande och utan någon omfattande efterkontroll. Informanterna skrev texterna för hand.

4.2.2. Omfång och fördelning över tiden

Den muntliga delen omfattar 100 ljudinspelningar, 10 med var och en av 10 informanter. Den sammanlagda tiden uppgår till ca 50 timmar, 5 timmar med varje informant. Den muntliga texten omfattar totalt ca 269 000 löpord, varav ca 147 000 utgör inlärarnas yttranden.

Den skrivna delen omfattar 220 uppsatser, 22 från varje informant, skrivna vid 11 tillfällen, och uppgår till ca 50 000 löpord.

I tabell 3 visas översiktligt korpusens fördelning över tid. Inspelningarna och uppsatserna producerades parallellt i tiden i omgångar som betecknas som *Tillfälle 1, Tillfälle 2* (*T1, T2* ...) etc. De korpusenheter som producerades vid dessa tillfällen betecknas *M1, M2* etc. för de muntliga inspelningarna och *S1, S2* etc. för de skriftliga uppsatserna. Tabell 4 och 5 visar fördelningen i detalj med exakt datum för varje muntlig respektive skriftlig session.

De första 9 tillfällena (T1–T9) utsträcker sig över läsåret 1990/91, dvs den period då alla informanterna följde samma språkkurs; T10 och T11 är uppföljningstillfällen under den andra respektive tredje vårterminen; vid tillfälle T11 gjordes bara en skriftlig del. Således har vi i T1–T10 tio omgångar av muntliga och skriftliga texter, parallellt insamlade så att det muntliga tillfället vanligen ägde rum en till två veckor före det skriftliga.

Tabell 3. Fördelningen av inlärarkorpusen över tiden – översikt.

Tidpunkt	Aug Sep 90				Nov Dec 90	Feb 91			Maj 91	Mar Apr 92	Apr 93
Tillfällen (omgångar)	T 1	T 2	T 3	T 4	T 5	T 6	T 7	T 8	T 9	T 10	T 11
Korpusenheter (muntliga och skriftliga tillfällen)											
Muntliga	M 1	M 2	M 3	M 4	M 5	M 6	M 7	M 8	M 9	M10	-
Skriftliga	S 1	S 2	S 3	S 4	S 5	S 6	S 7	S 8	S 9	S 10	S 11

Tabell 4. Tidsfördelning av inspelningar per tillfälle och person. Datum (ååmmdd).

Tillfälle	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

Tabell 5. Tidsfördelning av uppsatser per tillfälle och person. Datum (ååmmdd).

Tillfälle	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. Fördelning av innehållet

Korpusen har en strikt parallell uppläggning. Alla inlärarna fick utföra samma uppgifter vid motsvarande M- eller S-tillfällen. Syftet med detta är att möjliggöra jämförelser inom korpusen, enligt de intentioner som beskrevs i avsnitt 2 ovan.

Tabell 6 och 7 sammanfattar de aktiviteter och ämnen som ingick vid M- och S-tillfällena.

Tabell 6 ger en översikt av de olika slags aktiviteter som togs upp under de inspelade samtalen. Olika moment alternerade under samtalen:

- berättelser av ordlösa bildserier;
- beskrivning av foton och samtal kring dessa;
- beskrivning av olika föremåls utseende och funktion;
- intervjuinslag, oftast med fokus på informantens aktuella situation, erfarenheter och uppfattningar;
- diskussion av tidningsartiklar: informanten hade läst föregående dags Dagens Nyheter och valt ut en artikel för diskussion; artikeln fick bilda utgångspunkt för en diskussion som vanligen snart övergick i ett friare samtal.

Några inslag, speciellt några av bildberättelserna, upprepades vid senare tillfällen för att möjliggöra diakroniska jämförelser. Likaså återvände intervjuerna ibland till ämnen som man talat om tidigare.

Tabell 7 ger en motsvarande översikt av S-tillfällena, dvs de skriftliga uppsatsernas typer och ämnen. Varje gång skrevs två texter om givna ämnen. Även här planerades upprepningar av några uppsatsämnen in. Sålunda togs en bildberättelse upp tre gånger vid S-tillfällen, två bildberättelser utnyttjades vid både M- och S-tillfällen, och båda ämnena vid tillfället S11 var upprepningar från S7.

Tabell 6. Innehållsöversikt över ASU-korpusens muntliga inlärardel. Fördelning av aktivitetstyper och ämnen.

	Bildserie	Foton	Föremål	Intervju	Tidning
M 1	Hunden	Bilder av undervisnings-situationer		Personintervju	
M 2	Flugan	Studenter på campus Karaktärsfullt ansikte		Skansenbesök	
M 3	Festen			Beskriva en tänkt person	
M 4	Hunden Totte			Mitt studie-intresse	
M 5	Tunnelbanan			Om terminen som varit	
M 6			Husgeråd	Om juluppehållet	Självvald artikel
M 7			Kontors-artiklar		Självvald artikel + Familjesidan
M 8	Festen				Självvald artikel
M 9	Hunden			Framtidsplaner	Självvald artikel
M 10	Festen		Kontors-artiklar	Gjort under året Framtidsplaner Svenska språket	

Tabell 7. Innehållsöversikt över ASU-korpusens skriftliga inlärardel. Fördelning av uppsatsernas typer och ämnen.

	Bildserie	Berättelse	Beskrivning	Diskussion	Övrigt
S 1			En gäststudent i Stockholm		Frågor till en svensk student
S 2	Flugan	Min första dag i Sverige			
S 3	Äventyr på bio	En semester-resa			
S 4			Universitetet i Frescati	Mitt fackintresse	
S 5	Tunnelbanan: 1. Kvinnans berättelse 2. Mannens berättelse				
S 6				Om jag var George Bush [diskussion om gulfkriget 1991]	Min tidnings-läsning
S 7			Familjesidan i en daglig tidning	Barnuppfostran igår, idag, i morgon	
S 8	Flugan			Migrationen och det svenska samhället	
S 9				Argument för/mot svensk EG-anslutning	Råd till en landsman [om mötet med Sverige]
S 10	Flugan			Ungdomars liv i Sverige och i mitt hemland	
S 11			Familjesidan i en daglig tidning	Barnuppfostran igår, idag, i morgon	

5. De inföddas del

Korpusen med infödda svenskar insamlades 1998, sedan inlärarkorpusen hade fullbordats och redigerats. Detta underlättade uppgiften att matcha de båda delkorpusarna med avseende på typ av informanter, metoder för datainsamling och ämnesinnehåll, eftersom forskarlaget kunde utnyttja sina erfarenheter från konstruktionen av inlärarkorpusen som utgångspunkt när delen med infödda byggdes upp.

5.1. Informanter

För korpusdelen med infödda rekryterades sju unga studenter i grundutbildning vid Stockholms universitet. Strävan var att få tag i en typ av infödda svenska informanter som så långt möjligt motsvarade den typ av unga utländska gäststuderande som utgjorde informanter i inlärardelen.

De infödda informanterna var alla födda och uppvuxna i Sverige med svenska som sitt enda L1. De identifieras med beteckningarna Z1 till Z7 i korpusen. Tabell 8 sammanfattar några individuella data för dessa personer.

Som tabell 8 utvisar, ingår fyra kvinnor och tre män i ålder 20–29 år, median 23 år. De talade en centralsvensk standardvarietet (även om en lätt nordsvensk prägel i uttalet märktes hos Z5 och Z7). Deras kunskaper i andra språk varierade mellan två och fyra språk, med skiftande färdighetsnivå. Deras aktuella studieintressen vid tiden för medverkan i korpusprojektet låg inom områdena filosofi, litteraturvetenskap, konsthistoria och socialantropologi.

Tabell 8. Summariska uppgifter om de infödda informanterna i ASU-korpusen.

Person	Kön	Ålder	Uppväxtort	L2-kunskaper (i ordning efter färdighet)
Z 1	M	20	Stockholm	engelska, tyska, spanska
Z 2	K	20	Stockholm	engelska, franska
Z 3	K	26	Uppsala	engelska, spanska, franska
Z 4	K	20	Stockholmsförort	engelska, tyska, kinesiska
Z 5	M	23	Piteå	engelska, tyska, ryska, finska
Z 6	K	29	Stockholmsförort	engelska, franska, tyska, isländska
Z 7	M	25	Skellefteå	engelska, tyska
Summa	4 K, 3 M	Mdn 23		

5.2. Material

5.2.1. Insamlingsmetod

Materialet från de infödda samlades in på liknande sätt som från inlärarna, så långt det var möjligt.

Det muntliga materialet insamlades genom ljudinspelade intervjusamtal med en inlärare och två infödda samtalspartner åt gången. De två intervjuarna var desamma som för inlärarna. Samtalen ägde rum i en studio i Fonetiklaboratoriet, Institutionen för lingvistik, Stockholms universitet. Deltagarna satt kring ett lågt bord med en mikrofon för informanten och en för intervjuarna, för stereoupptagning. Informanten satt med ryggen vänd mot teknikerfönstret, för att inte distraheras. Samtalen varade 25–30 minuter och spelades in samtidigt på DAT och standardkassett.

Det skriftliga materialet insamlades under gruppsessioner i en lärosal. Vid varje tillfälle avsattes två timmar för två uppsatser om givna ämnen. Proceduren var densamma som för inlärarna (se 4.2.1 ovan).

Fem omgångar med parallell muntlig och skriftlig datainsamling genomfördes med en veckas mellanrum.

5.2.2. Omfång och fördelning över tiden

Fem ljudinspelningar gjordes med var och en av de sju infödda informanterna, alltså totalt 35 inspelningar. Den transkriberade muntliga texten omfattar ca 149 000 löpord, varav ca 98 000 ord utgör informanternas yttranden. Den skrivna delen omfattar 10 uppsatser från varje informant, dvs totalt 70 texter med sammanlagt ca 25 000 löpord.

De resulterande korpusenheterna från dessa fem omgångar betecknas *M1* till *M5* (muntligt material) resp. *S1* till *S5* (skriftligt material).

5.2.3. *Fördelning av innehållet*

Korpusen från de infödda är strikt parallell internt, i och med att alla informanterna fick utföra samma uppgifter vid motsvarande sessioner. Innehållet, med avseende på aktivitetstyper och samtals- och uppsatsämnen är likartat, och delvis identiskt, med innehållet i inlärarkorpusen, men de infödda informanterna genomgick ett kortare program under sina fem M- och fem S-sessioner.

De aktiviteter och ämnen som ingick i M- och S-sessionerna med de infödda informanterna sammanfattas i tabell 9 och 10. Som framgår där, togs några av ämnena upp vid både muntliga och skriftliga tillfällen.

Tabell 9. Innehållsöversikt över ASU-korpusens muntliga del med infödda informanter. Fördelning av aktivitetstyper och ämnen.

	Bildserie	Foton	Föremål	Intervju	Tidning
M1	Hunden	Bilder av två stora familjer		Personintervju: persondata, studieplaner, intressen	
M2	Flugan		Husgeråd	Något du gjort förra året	Självvald artikel
M3	Festen			Beskriva en tänkt person; Irak-krisen	Självvald artikel
M4	Tunnelbanan		Kontorsartiklar	Ett studie- eller fritidsintresse	Självvald artikel
M5	Totte	Afrikansk berättarscen		Tankar om svenska språket	Självvald artikel

Tabell 10. Innehållsöversikt över ASU-korpusens skriftliga del med infödda informanter. Fördelning av uppsatsernas typer och ämnen.

	Bildserie	Berättelse	Beskrivning	Diskussion	Övrigt
S1	Äventyr på bio		Universitetet i Frescati		
S2		En semester-resa		Barnuppfostran igår, idag, i morgon	
S3	Flugan			Migrationen och det svenska samhället	
S4				Mitt fackintresse	Råd till en utländsk gäststudent
S5	Tunnelbanan: 1. Kvinnans berättelse 2. Mannens berättelse				

6. Textens form i den datorlagrade korpusen

Transkription och grammatisk taggning

De muntliga (***M***) och de skriftliga (***S***) texterna har skrivits in på dator efter insamlingstillfället. Principerna för ***transkriptionen*** beskrivs i avsnitt 7 nedan. Informanternas yttranden, men inte intervjuarnas, har ***ordtaggats*** enligt ett system som beskrivs i avsnitt 8 nedan.

Lagringsform och användargränssnitt

Korpusen är lagrad i ett XML-baserat format och är tillgänglig via Språkbanken, Institutionen för svenska språket, Göteborgs universitet genom ett användargränssnitt ***(ITG-gränssnittet)*** som har utarbetats i anslutning till projektet *IT-based Collaborative Learning in Grammar (ITG)* och är anpassat till korpusens individbaserade och longitudinella struktur. Bland annat finns här möjlighet att arbeta interaktivt med konkordanser och spara arbetsresultat. Hur man kan arbeta med gränssnittet beskrivs i en separat bruksanvisning ***(Arbeta med ASU-korpusen)***. ITG-programmet kräver i dagsläget att programvaran *Java* är installerad på datorn. (Beträffande åtkomst, se nedan, avsnitt 10.)

Hur identifieras texterna i korpusen?

Varje transkriberad inspelning och varje uppsatstext utgör en ***textenhet***. Textenheterna identifieras med en bokstavs- och sifferkod som betecknar ***Person + Medium + Tillfälle + Uppsatstext 1 eller 2*** (det sista bara i skriftkorpusen). Exempel: C1M8 = Person C1, Muntligt, Tillfälle 8; G3S052 = Person G3, Skriftligt, Tillfälle 5, Uppsats 2. I den muntliga korpusen skrivs tillfälle 10 som romerskt X, t.ex. E2MX = Person E2, Muntligt, Tillfälle 10.

Den redigerade texten

I den form texten har för användaren finns information inlagd, som strukturerar texten:

Texten har en bestämd, ***permanent radindelning*** som kan citeras för att ange ***beläggställen*** i korpusen. Uppgiften om beläggställe är det nödvändiga instrumentet för att identifiera olika individer, inlärare/infödda informanter, tal-/skriftmaterial, textenheternas kronologi och den löpande tidsföljden i texten. Varje rad har i vänsterkanten ett ***radhuvud***, som anger beläggstället med hjälp av textenhet och löpande rad i texten, t.ex. C1M3 0067, Z3S041 016. I radhuvudet kan man alltså direkt läsa av person, medium och kronologisk lokalisering.

Varje ny tur i den muntliga korpusen inleds med en ***talaridentifikation***, där ”I” alltid betecknar den aktuella informanten, och ”B” och ”E” betecknar de båda infödda svenska interlokutörerna (intervjuarna).

Kommentarer har lagts till i texten på särskilda rader inledda med ”C”. Dessa rader räknas inte i radnumreringen.

I visningen av kompletta texter (funktionen *Textvisning* i ITG-gränssnittet) ingår också ett ***texthuvud*** före varje textenhet med uppgifter om textenheten, samt i de *muntliga* texterna

tillagda ***mellanrubriker*** som särskiljer olika aktiviteter i de inspelade samtalen och därigenom strukturerar upp innehållet i texterna.

Texthuvudet som föregår själva texten innehåller datum för inspelning/uppsats, samt given uppsatsrubrik eller annan innehållsinformation.

Mellanrubrikerna i de muntliga texterna står på egna kommentarrader (C-rader) och inleds med ”>>>” för att markera mellanrubrik till skillnad från andra kommentarrader. Ett antal ***beteckningar för aktiviteter*** används genomgående i mellanrubrikerna, så att man lättare kan identifiera och jämföra motsvarande textpassager tvärsöver textenheterna. Sådana återkommande beteckningar är bl.a. BILDSERIE, FOTO, FÖREMÅL, INTERVJU, TIDNING (jfr tabell 6 och 9 ovan). Dessutom anges här vilket objekt eller ämne man talar om.

Texthuvud och mellanrubriker anges bara till visningen av fulltext och tas inte med i textfältet till konkordanser. Dock kan *datum* för varje inspelning resp. uppsats i *inlärardelen* utläsas ur tabell 4 och 5 ovan och *ämnen för aktiviteter och uppsatser* ur tabell 6 och 7 (inlärarna) och 8 och 9 (infödda).

7. Principer för transkriptionen

7.1. Transkription av det muntliga materialet

Allmänna principer:

En transkriptionsmodell har utvecklats för den muntliga ASU-korpusens behov. Den syftar till att fånga upp textens grammatiska och lexikala struktur och aspekter av yttrandeplaneringen och dialogstrukturen. Det är *inte* en transkription på fonetisk nivå eller en detaljerad återgivning av yttrandenas fysiska form. Men den avser att redovisa den lexikala och morfologiska form som orden uppträder i (i synnerhet förekommande böjning) så noga som möjligt.

En modifierad version av svensk standardortografi används genomgående för den svenska texten; för ord på andra språk används språkets standardortografi. Modifikationerna består i (a) att avvikande och reducerade former av orden skrivs så som de uppträder, och (b) att talspråksformer av orden återges i den mån talarna använder dem; exempelvis:

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

Transkriptionen är lexikalt och morfo-syntaktiskt grundad. *Fonetiska varianter* i uttalet särskiljs inte, men kommentarer till uttalet kan ges på separata kommentarrader i fall där detta bedömts relevant. *Prosodiska drag* i talet markeras inte i transkriptionen, dock har prosodin vägts in som ett kriterium för den syntaktiska eller pragmatiska tolkningen av yttrandena och därigenom varit vägledande för transkriptionen.

Huvudtyper av strukturella enheter i texten:

De huvudtyper av enheter i texten som markeras genom transkriptionen är *ordet*, *makrosyntagmen* och *turen*.

I teckenbruket upprätthålls en distinktion mellan ***”ord”*** och ***”icke-ord”***. ”Ord” är de lexikala enheterna i texten. ”Icke-ord” innefattar markeringar av syntaktiska gränser, pauser, pausfyllare och olika slags pragmatisk information (som anges i teckenförteckningen nedan). Skillnaden upprätthålls genom att *gemena bokstäver* genomgående används för ”ord” och *andra tecken än gemena bokstäver* används för ”icke-ord”. Observera till exempel att vi undviker att skriva pausfyllarljudet (ett paraspråkligt element) som ”eh” eller ”öh” eller liknande, utan representerar det med tecknet ”%”. Ett syfte med denna princip är att göra *ordalydelsen* i talarens text lättare att urskilja när man läser transkriptionen.

Versaler används dels för tillagda kommentarer och rubriker (på C-rader), dels för icke-turbrytande inpass (uppbackningar) från en samtalspartner (se beträffande turindelningen nedan). Versaler används också för kommentarer av typen ”SKRATT”, ”TYST”, även lokalt inom turen. Text i versaler tas inte med vid ordräkning.

För den syntaktiska segmenteringen av den muntliga texten har vi valt att utgå från begreppet ***makrosyntagm* (*MS*)** (Loman & Jörgensen 1971). Makrosyntagmen är i sin typiska form en huvudsats med sina tillhörande bisatser, om sådana finns. Samordnade huvudsatser klassas som separata MS; men med reduktion av identiska delar behandlas de som en MS. Andra typer av MS är interjektions- och tilltals-MS samt satsfragment. Se Loman & Jörgensen (1971) för utförligare definitioner och diskussion. (I analyser av engelska motsvaras MS av ”T-units” (Hunt 1966; Richards m.fl. 1985); för talad text ”C-units” (Biber m.fl. 1999:1069).)

En MS behandlas i ASU-korpusen som *fullbordad* om den inte saknar någon obligatorisk del i slutet, och om prosodin inte tyder på att den är ofullbordad. Den fullbordade MS:en avgränsas med frågetecken ”?” (för fråge-MS) eller punkt ”.” (för övriga MS). Syntaktiskt *avbrutna* sekvenser avgränsas med ”/” vid avbrottsstället. Sådana avbrutna sekvenser förekommer typiskt vid successiv planering och produktion av en makrosyntagm. En karakteristisk typ av sekvens i texten är följaktligen *den fullbordade makrosyntagmen med sina eventuella föregående avbrutna ansatser*. Denna typ av sekvens (som alltså kan innehålla ett eller flera ”/” såväl som pauser, pausfyllare och andra pragmatiska element) reflekterar både drag i satsplaneringen och den resulterande satsen. Den kan ses som den oredigerade motsvarigheten i talet till den redigerade satsen i skriven text.

Turen markeras genom ny rad inledd med en *talaridentifikation* (i egen kolumn).

Turindelningen erbjuder vissa särfall för transkriptionen att hantera; följande typer kan särskiljas:

(a) *Reguljär turväxling:* ny talare, ny tur.

(b) *Inpass* från en samtalspartner, icke-turbrytande. Hit räknas feedback i form av uppbackningar, som enbart bekräftar, stödjer eller uppmanar till fortsättning, utan att ha något ytterligare semantiskt innehåll. Inpasset skrivs inuti turen, inom parentes ”( )”, i versaler, med en identifikation av interlokutören, t.ex. ”... (B: JAHA) ...”.

(c) *Intervention* från en samtalspartner (med något semantiskt innehåll, i motsats till inpassen i fall (b)), *men där den första talaren fortsätter sin tur*. Detta transkriberas som en turväxling. Den avbrutna turen markeras med ”\” vid avbrottet. Interlokutörens yttrande sätts på ny rad, som en separat tur. Den första talarens fortsättning kommer sedan utskriven som en följande tur, inledd med ”\”.

Transkriptionen innehåller ingen beteckning för samtidigt tal av två eller flera talare. Sådana fall kan beskrivas på kommentarrader, men detta har tillämpats sparsamt. I okomplicerade fall, där ett turslut och en turbörjan överlappar, har talarnas turer vanligen skrivits ut efter varandra.

Särskilda tecken som används i M-transkriptionen:

=	Tom paus.
==	Längre paus. (Lång tystnad noteras på C-rad.)
%	Pausfyllare. (Ersätter bokstavsbeteckningar som ”eh”, ”öh” etc.)
%%	Lång pausfyllare.
xxx	Oidentifierbar sekvens. (Möjliga och plausibla tolkningar kan noteras på C-rad.)
-	Efter ett avbrutet ord, eller före en separat slutdel av ett ord.
+	Efter en morfologiskt otydlig form.
/	Syntaktiskt avbruten sekvens, eller avbrott för omplanering eller reparation. Används inte om identiska former upprepas.
\	Tursammanbindare. Används vid slutet av en tur och början av samma talares nästa tur, om en samtalspartner har intervenerat mitt i talarens yttrande och denne fortsätter sin tur. (Se beträffande turindelning ovan.)
?	Frågetecken. Efter avslutad frågemakrosyntagm.
.	Punkt. Efter en avslutad makrosyntagm som inte är en fråga.
¿	Befrågat uttryck. Talare söker feedback beträffande ett uttryck genom en intonation som signalerar en metaspråklig fråga: ”Är det här rätt?”
” ”	Citattecken. Kring direkta anföringar.
< >	Kring ord eller sekvenser på andra språk än svenska.
( )	Kring text i versaler inom talarens tur: uppbackningar t.ex. ”(B: MHM)” eller kommentarer som ”(SUCKAR)”, ”(SKRATTAR)”.

7.2. Transkription av det skriftliga materialet

Transkriptionen av skriftkorpusen har i princip bestått i att skriva av skribenternas handskrivna texter och lägga till erforderliga C-rader så som beskrivs i avsnitt 7.1. I några bestämda avseenden har texten dock normaliserats:

. En punkt sätts alltid ut vid slutet av makrosyntagmen, om inte skribenten själv använder ”?” eller ”!”. Det innebär att en punkt suppleras om den saknas på ställen som uppenbart är slutet av en MS. Likaså avlägsnas punkter inom makrosyntagmer, t.ex. vid förkortningar. Därmed kommer punkt (såväl som fråge- och utropstecken) enbart att förekomma vid slutet av en MS.

xxx Oidentifierbar sekvens.

8. Taggningssystemet

Informanternas produktion i korpusen är ordtaggad. Det vill säga, orden i informanternas del av den muntliga texten och hela den skriftliga uppsatstexten har försetts med en morfologisk taggning som anger ordklasser och vissa grammatiska underkategorier. De svenska intervjuarnas del har inte ordtaggats. Skiljetecken har också försetts med taggar, även i intervjuarnas del. I ITG-gränssnittet kan man söka på ordens/skiljetecknens former och/eller taggar.

Ett generellt problem med förvalda taggsystem är att forskaren kan vilja indela sina data efter speciella kriterier som betingas av den aktuella forskningsuppgiften, kriterier som ett givet taggsystem, även om det är finindelat, inte tillgodoser. ASU som är en liten och strukturerad korpus ger ett material som kan vara tacksamt att analysera i detalj. Principen i ASU-korpusen är att å ena sidan ha en enkel och ganska grov taggindelning som grundval för sökningar. Å andra sidan ges det möjligheter att finkategorisera och finsortera data efter själva kriterier genom interaktivt arbete med konkordanser i det tillhörande användargränssnittet.

ASU-korpusens nuvarande taggsystem omfattar ett 50-tal olika taggar för ordklasser och vissa grammatiska kategorier och andra markeringar samt skiljetecken. En översikt av dem ges i tabell 11. I den specificerade lista som sedan följer presenteras taggkategorierna med valda exempel, och en del kommentarer ges till taggningspraxis.

Tabell 11. ASU-korpusens taggar grupperade efter grammatiska kategorier.
I den efterföljande uppräkningen förklaras taggkategorierna i tabellens nummerordning med valda exempel. Numreringen utläses i tabellen från rad till kolumn, t.ex. N = 1.1, EN = 1.2.

	.1	.2	.3	.4	.5	.6	.7	.8	.9	.10	.11
1. *Substantiv*	N	EN									
2. *Pronomen & Artiklar*	P	PO	RP	ROBJ	FS	KP	DEM	PIF	REL	PÖ	ART
3. *Frågeord*	FRÅ	FRÅK									
4. *Kvantifikatorer*	QU	OT									
5. *Verb*	VS	VT	VINF	VSUP	VPC	VI	V	KOP	KOPT	KOPI	KOPÖ
6. *Verbpartiklar*	PT										
7. *Infinitivmärke*	IM										
8. *Adjektiv*	A										
9. *Adverb*	ADV	ADVG	ADVK	KADV							
10. *Konjunktioner*	K	UK									
11. *Prepositioner*	PR										
12. *Interjektioner*	IJ										
13. *Fundamentmarkör*	FM										
14. *Subjektsmärke*	SM										
15. *Övriga ordtaggar*	U	I	XXX	X	Tagg + X	Tagg + Z					
16. *Skiljetecken*	del	syntBreak	pause								

Specificerad lista över taggar:

Ord ur de slutna ordklasserna exemplifieras här i urval för att visa hur taggarna används. Det är alltså inte en uttömmande ordförteckning. Observera att grammatiska funktionsord ofta är polyfunktionella, dvs återfinns under flera taggar.

1. Substantiv

1.1. **N** Substantiv utom egennamn

1.2. **EN** Egennamn

2. Pronomen och artiklar

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

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

2.3. **RP** Possessivt reflexivt pronomen
sin sina sitt själv

2.4. **ROBJ** Reflexivt objekt
sej sig

2.5. **FS** Formellt subjekt
de det

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

2.7. **DEM** Demonstrativt pronomen
de den denna dessa det detta dom
Obs! *den här* DEM U; *så här* KP DEM

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

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

2.10. **PÖ** Pronomen övriga
egen eget egna enda varandra

2.11. **ART** Artikel
de den det dom en ett

3. Frågeord

3.1. **FRÅ** Frågeadverb
hur när va vad var varför varifrån vart vem vilka vilken vilket
Används för frågeord i direkta frågor; jfr REL (2.9), FRÅK (3.2).

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

4. Räkneord och andra kvantifikatorer

4.1. **QU** Kvantifikator
all alla allt båda hela lite många mycke mycket några samt grundtal skrivna med bokstäver eller siffror.
Siffror taggas med QU när de betecknar grundtal (inkl. årtal). Om de betecknar ordningstal (t.ex. *den 24 december*), används taggen OT.
Datum skrivna med siffror taggas så här: *1997-12-24* QU OT OT;
24/12 -97 OT OT QU.

4.2. **OT** Ordningstal
andra första tredje
Jfr QU (4.1).

5. Verb

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

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

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

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

5.5. **VPC** Verb particip
Taggen VPC används för både presens och perfekt particip.
Kategorierna VPC och A (Adjektiv, 8.1) är svåra att separera, och taggningen har gjorts ganska intuitivt på grundval av den betydelse kontexten ger: om ordet uppfattades ha kvar en ’verbisk’ karaktär, valdes VPC, annars A. Uppmärksamhet på enskilda fall rekommenderas.

5.6. **VI** Verb imperativ

5.7. **V** Verb ‘naken’ form
Används när den rena rotformen av verbet uppträder, utan t.ex. ett infinitiv-*a* eller en tempusändelse. Ex: *<u>vänt</u> din brev*.

5.8. **KOP** Kopula presens
e är

5.9. **KOPT** Kopula preteritum
va var

5.10. **KOPI** Kopula infinitiv
va vara

5.11. **KOPÖ** Kopula övriga
varit vore

6. Verbpartiklar

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

7. Infinitivmärke

7.1. **IM** Infinitivmärke
att å

8. Adjektiv

8.1. **A** Adjektiv
Jfr 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** Adverb graderande
ganska lite mycke mycket väldigt

9.3. **ADVK** Adverbiell konnektor
annars då först nu sedan sen så

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

10. Konjunktioner

10.1. **K** Samordnande konjunktion

eller fast för men och så å utan

10.2. **UK** Underordnande konjunktion (Subjunktion)
att då därför eftersom fastän innan medan när om som trots än

UK används för 3 typer av *som*:
1. Komparativt (*lika bra som*)
2. Predikativt (*som studerande ...*)
3. Några ställen där *som* uppträder hos inlärarna som målspråkets *att*

UK används tillsammans med U (se kommentaren till taggen U nedan) vid frasformiga subjunktioner med *att*.
Exempel på distinktion mellan K och UK i några fall:

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

11. Prepositioner

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. Interjektioner

12.1. **IJ** Interjektion
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. Fundamentmarkörer

13.1. **FM** Fundamentmarkör
då så

14. Subjektsmärke

14.1. **SM** Subjektsmärke
som Ex: *jag vet vem som ...*

15. Övriga

15.1. **U** Ospecificerad
Se kommentaren till taggen U nedan.

15.2. **I** Avbrutet ord
Används för avbrutna ord, transkriberade med ”-” i slutet.
Separata slutdelar av ord, transkriberade med ”-” i början, kan vanligtvis hänföras till en grammatisk kategori och taggas därefter.

15.3. **XXX** Otydbar
Används för oidentifierbara sekvenser, transkriberade *xxx.*

15.4. **X** Otaggbar
Används för hörbara/läsbara former där en tagg inte är möjlig att bestämma.

15.5 **Tagg+X** Osäker taggning
Om ordets grammatiska kategori inte helt säkert kan bestämmas, väljs den mest plausibla tolkningen, följd av X, t.ex. VSX = VS-osäkert.

15.6 **Tagg+Z** Icke-svenskt ord
Används i den *muntliga korpusen* för text på andra språk än svenska, bl.a. vid icke-anpassade kodväxlingar, dvs när en fonologiskt och morfologiskt icke svenskanpassad växling till ett annat språk uppträder. Sådana ord och sekvenser transkriberas inom < > i den muntliga korpusen. Om möjligt taggas de icke-svenska orden på motsvarande sätt som de svenska, med tillägg av Z. Ex: <*airport*> taggas NZ.

16.1 **del** Syntaktisk gränsmarkör; tursammanbindare; befrågat uttryck
Gränsmarkörer . ! ? , ; : ” ( )
Parentestecknen avser parenteser i informantens egen text i skriftkorpusen, inte parenteser kring inpass och kommentarer i den muntliga korpusen.
Tursammanbindare \
Befrågat uttryck ¿

16.2 **syntBreak** Syntaktiskt brott
/

16.3 **pause** Paus och pausfyllare
= == % %%

*Kommentarer till bruket av taggen **U**:*

Taggen U (ospecificerad) används i frasformiga, bindestrecksförsedda eller andra grafiskt delade uttryck som vi har betraktat som fasta lexikala fraser eller sammansättningar. På dessa har vi tillämpat *”syntetisk taggning”*, varvid frasen i sin helhet hänförs till en kategori, i motsats till normal *”analytisk taggning”*, där varje ord får en tagg för en specifik kategori. Bara det första ordet i frasen får den specifika taggen, och U används för övriga ord i frasen.

I följande exempel får de understrukna orden taggen U, medan de icke-understrukna orden får den specifika tagg som visas till vänster, och denna tagg representerar då hela frasen. När taggen U uppträder i korpusen, bör man således orientera sig vänsterut till det samhörande ordet med en specifik tagg.

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. Korpusens tillkomst, projektfinansiering och medverkande personer

Korpusen byggdes upp i sin ursprungliga form under 1990-talet i projektet *Andraspråkets strukturutveckling (ASU)* vid Institutionen för lingvistik, Stockholms universitet, under ledning av Björn Hammarberg och med finansiellt stöd från Humanistisk-Samhällsvetenskapliga Forskningsrådet och Stockholms universitets humanistiska fakultet. Korpusen har senare genomgått en grundlig teknisk modernisering för att överensstämma med vad som idag vinner giltighet som standard för språkliga textkorpusar. Detta har skett med stöd från Magn. Bergvalls Stiftelse, Birgit & Gad Rausings Stiftelse för Humanistisk Forskning samt Henrik Granholms Stiftelse.

Materialet till inlärardelen samlades in 1990-93 och transkriberades, taggades och redigerades. Vintern 1998 kompletterades korpusen med materialet från infödda informanter, vilket behandlades på motsvarande sätt. Korpusen lagrades från början i ASCII-format och redigerades och bearbetades med hjälp av korpusprogramvaran *PC Beta* (Brodda 1982, 1991). Denna korpusversion har legat till grund för artiklar och avhandlingar under 1990-talet och de första åren av 2000-talet.

Under detta skede medverkade följande personer i tillkomsten av korpusen:

Niclas Abrahamsson (transkription av den muntliga inlärardelen; taggning av hela inlärardelen; korpusredigering; programutveckling)
Maria Arnstad (transkription av den muntliga infödda delen)
Dorothée Augustin (rekrytering av infödda informanter; studioassistans)
Christina Ericsson (transkription och taggning av den muntliga infödda delen)
Björn Hammarberg (uppläggning och ledning; intervjuer och uppsatsuppgifter; transkription av den skriftliga inlärardelen; korpusredigering)
Eva Klingberg Merk (inlärarnas lärare i preparandkursen; intervjuer och uppsatsuppgifter; transkription av den skriftliga inlärardelen)
Ulrika Kvist Darnell (transkription av den skriftliga infödda delen)

Benny Brodda fungerade som rådgivare för textprocessningen. Han tillhandahöll programvaran PC Beta med taggningsprogrammet PC Tagger, gjorde specialanpassningar för ASU-korpusens behov och lärde medlemmar av projektgruppen att använda programmen.

Kenneth Andersson (inlärardelen) och *Hassan Djamshidpey* (infödda delen) fungerade som inspelningstekniker vid materialinsamlingen.

Den tekniska moderniseringen av korpusen har inneburit att hela korpusen har konverterats till ett XML-baserat lagringsformat och kopplats till ett sök- och analysverktyg (*ITG-gränssnittet*) som har utarbetats i projektet *IT-based Collaborative Learning in Grammar (ITG)*. ITG-projektet är knutet till Institutionen för lingvistik och filologi, Uppsala universitet (*Anju Saxena*), och ITG-gränssnittet administreras och utvecklas vid Språkbanken, Göteborgs universitet (*Lars Borin*). I ITG-gränssnittet har särskilda användarfunktioner skapats för att tillgodose ASU-korpusens speciella egenskaper. Härvid har *Lars Borin* haft det språkteknologiska huvudansvaret och *Björn Hammarberg* huvudansvaret för L2-forskningens användningsbehov. Programmeringen har utförts av *Camilla Bengtsson* (textkonverteringen), *Leif-Jöran Olsson* (utveckling av gränssnittet och justeringar av textkonverteringen) samt *Elena Volodina* (justeringar av textkonverteringen).

10. Tillgång till korpusen

Hur kommer man åt korpusen?

Korpusen är tillgänglig för sökning och analys med hjälp av *ITG-gränssnittet*, som handhas av Språkbanken, Institutionen för svenska språket, Göteborgs universitet. Gå tillväga i följande steg:

- Ladda ner programvaran *Java* på din dator. (Gratis från http://www.java.com.)
- För att få tillgång till ASU-korpusen via ITG, gå in på http://spraakbanken.gu.se/itg och följ instruktionerna där.
- När du öppnat ITG-gränssnittet, gå till rubriken *Korpus* på menyn och klicka på underrubriken *Korpussökning* för att välja texter. (Även andra korpusar än ASU är tillgängliga här.) För arbetet har du sedan nytta av denna innehållsmanual samt arbetsmanualen *Arbeta med ASU-korpusen*, som nås från ITG-gränssnittet.

Vad får man tillgång till?

Tillgången innefattar att man fritt kan göra sökningar i den transkriberade korpusen, skapa konkordanser och få frekvensuppgifter, studera sökträffarna i begränsad kontext, redigera om och spara konkordanser samt exportera och göra utskrifter av frekvenslistor, konkordanser och exempel. Nedanstående villkor gäller.

Forskare som önskar tillgång till den transkriberade fulltexten kan meddela sig per e-post till Björn Hammarberg, ham@ling.su.se.
För förfrågningar om ITG-gränssnittet, kontakta Språkbanken per e-post sb@svenska.gu.se.

Villkor för användning av ASU-korpusen

För att utnyttja korpusen gäller följande villkor:

- Generellt: ASU-korpusen är avsedd för forskning och utbildning. Den omfattas av upphovsrätt. De medverkande informanternas integritet ska respekteras.
- Användare får inte sprida korpusens texter i kommersiellt syfte eller sprida texter så att de kan komma att utnyttjas kommersiellt. Dock får exempel och kontextutdrag återges enligt vanlig vetenskaplig citeringspraxis.
- När material ur korpusen återges, i publikationer eller på annat sätt, ska ASU-korpusen, Institutionen för lingvistik, Stockholms universitet anges som källa. (För orientering och detaljinformation om korpusen kan den här skriften, *Introduktion till ASU-korpusen*, anges och citeras.)
- Informanternas identitet ska skyddas. De är anonymiserade i korpusen, men skulle ledtrådar till deras identitet ändå förekomma, får dessa inte röjas. Informanternas personliga värdighet ska alltid iakttas.

Litteraturreferenser

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. I *English Computer Corpora. Selected Papers and Research Guide*, red. 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.