ASU-korpusen

Dess syfte, uppbyggnad och särart

Björn Hammarberg
Stockholms universitet, Institutionen för lingvistik

Swe-Clarin på turné, Stockholms universitet 2018-05-16

Vad är ASU-korpusen tänkt för?

- Dataresurs för forskning och utbildning
  (senior forskning, avhandlingar, uppsatser, kursuppgifter)
- Utvecklingsmönster i inlärarspråket, framväxten av ett utvecklat svenskt andraspråk hos vuxna
- I någon mån yttrandeplanerings- och tillägnandeprocesser
- Grammatiska och lexikala undersökningar
  (även fonologi, diskurs/text, samtalsanalys)
- Språkproduktion på individnivå
- Sökning, analys och dokumentation av språkliga detaljdata
- Fånga upp relativt frekventa företeelser i språket och karakteristiska inlärarfenomen i svenskan

Historik

Korpusen byggdes upp 1990-93 (inlärardelen) och 1998 (infödda delen) i projektet ***Andraspråkets strukturutveckling (ASU)*** vid Institutionen för lingvistik, SU, av en grupp under ledning av Björn Hammarberg.

Ursprunglig programvara: Benny Broddas *PC Beta* och *PC Tagger* (Brodda 1982, 1991). ASCII-format och DOS-kommandon.

På 2000-talet modernisering i samarbete med Språkbanken (Lars Borin, Leif-Jöran Olsson, Elena Volodina). XML och koppling till *ITG-gränssnittet*.

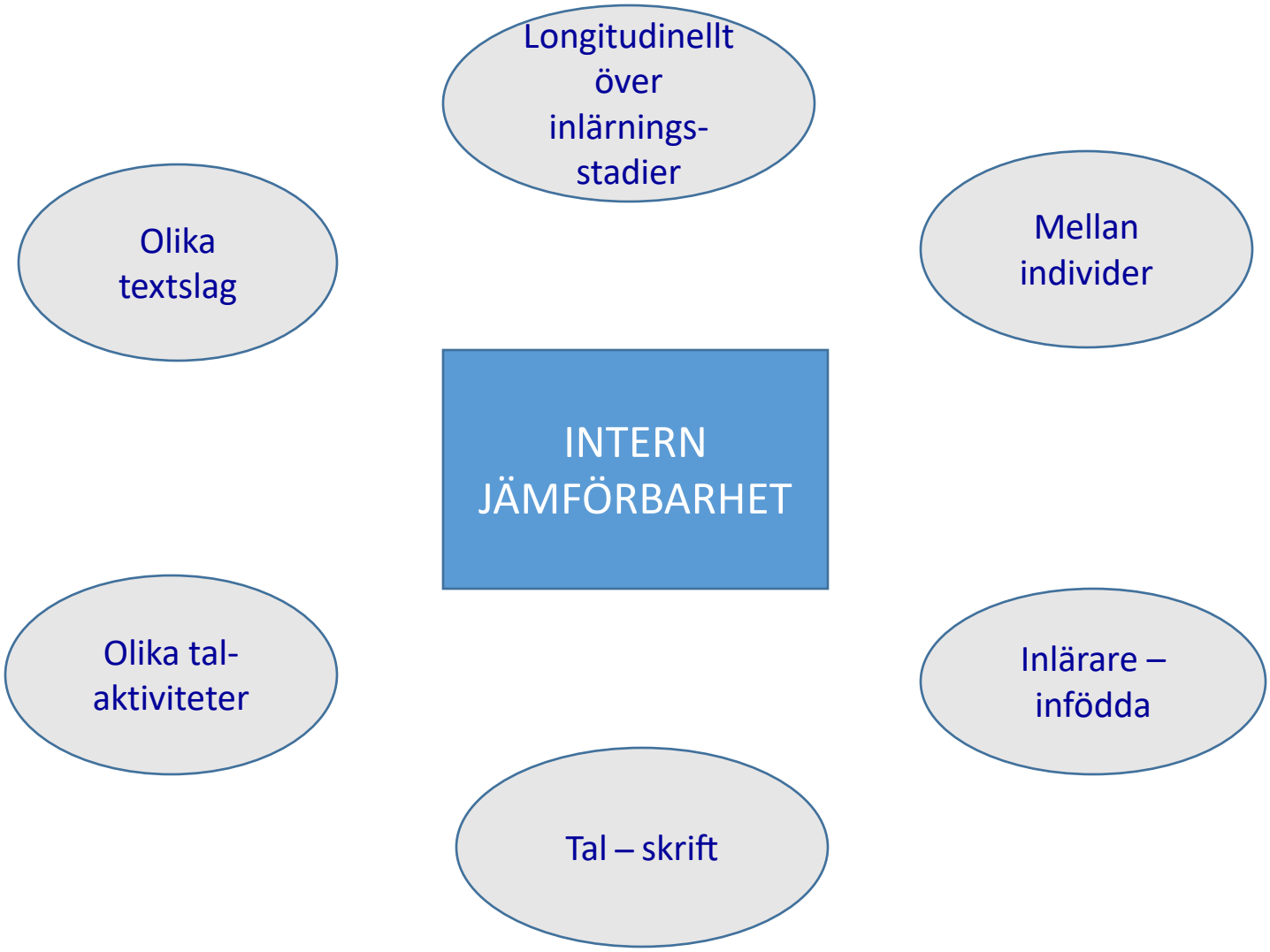
Extern finansiering från HSFR, Magn. Bergvalls Stiftelse, Birgit & Gad Rausings Stiftelse, Henrik Granholms Stiftelse.

Riktlinjer för uppbyggnaden

Grundläggande syfte: dokumentera inlärarspråkets dynamik, utveckling och förhållande till målspråket; vuxna personers språk.

Kriterier:

- Individinriktning - följa och jämföra individer; ”mycket från få”
- Inlärartyp - inlärare som strävar till utvecklat språk
- Utveckling - påtaglig förändring över tid, observera ofta
- Stadieomfång - noll till avancerat stadium hos samma personer
- Tal och skrift - tal o skrift parallellt över tid hos samma personer
- Inföddas språkbruk - jämförbar kontrollkorpus från infödda (interlanguage-modell – ej felanalysmodell)
- Intern jämförbarhet - inlärningsstadier; individer; inlärare-infödda; tal-skrift; olika aktiviteter i tal, slag av text i skrift


Longitudinellt över inlärnings-stadier
Olika textslag
Mellan individer
INTERN JÄMFÖRBARHET
Olika tal-aktiviteter
Inlärare – infödda
Tal – skrift

Indelning och omfång

Fyra huvuddelar, antal löpord:

	Inlärare	Infödda	Summa Inl+Inf
Muntligt*	269 000 / 147 000	149 000 / 98 000	418 000
Skriftligt	50 000	25 000	75 000
ASU totalt			493 000

*Uppgiften avser: hela dialogen / informanternas yttranden

Texterna indelas efter huvuddel > person > kronologi

Korpusens inlärardel

Informanter:

- 10 studenter 19-28 år, 3K 7M
- Förstaspråk: kin (3), gre (2), por (2), spa (1), pol (1), ty+eng (1); L2-kunskaper i engelska + ytterligare språk
- Rekryterades från preparandkursen i svenska för utländska studenter vid SU, lå 1990-91
- Nybörjare i svenska i starten, läste vid svenska högskolor i slutet, efter rikstestet (TISUS)
- Allmän karakteristik: *'semi-formella'*, *'kvalificerade'*, *'snabba'* inlärare

Material:

- 10 inspelade samtal individuellt i studio varvade med 11 uppsats-tillfällen, allt insamlat utanför kursen

Korpusens kontrolldel

- 7 svenska studenter 20-29 år, 4K 3M, centralsvensk standardvarietet
- 5 inspelningar + 5 uppsatstillfällen, vt 1998
- Informanter, metod och innehåll jämförbart med inlärardelen, så långt möjligt

Texternas innehåll

Muntliga delen

- Berättelse av bildserier utan text
- Beskrivning av föremål och foton
- Intervjusamtal och diskussioner
- Diskussion utifrån lästa tidningsartiklar

Skriftliga delen

- Berättelse av bildserier
- Beskrivande texter
- Diskuterande och redogörande texter

Textens form i korpusen

Muntliga delen
(Kvasi-)ortografisk transkription som återger talspråksformer; tillagda tecken för talaridentifikation, meningsgränser, pauser, turstruktur mm. Även inlagda mellanrubriker.

Skriftliga delen
Transkription efter handskrivna originalet med viss normalisering av skiljetecken.

Taggning
Morfologisk taggning av inlärarnas text.
Halvautomatisk procedur (PC Tagger, Brodda 1982).

Fast radbrytning och radnumrering

Typ av korpus (1)

Nomotetisk och *idiografisk* forskningsinriktning

(Windelband 1894, Geschichte und Naturwissenschaft)

NOMOTETISKT FOKUS
- söka generell kunskap
- ofta kvantitativa metoder
- makro-perspektiv
- abstraherande, reduktionistisk
- typiskt för: naturvetenskap
- psykologi: klasser, kohorter, populationer
- språkvetenskap

IDIOGRAFISKT FOKUS
- belysa enskilda händelser
- ofta kvalitativa metoder
- mikro-perspektiv
- konkretiserande
- typiskt för: historia, humaniora
- psykologi: individer
- språkvetenskap

Hur placerar sig språkkorpusar?

Typ av korpus (2)

Korpusar, schematiskt betraktade:

- Övervägande **nomotetiskt** perspektiv; språket i samhället

  Ex.: BNC, COCA, Språkbankens storkorpusar

- Framträdande **idiografiskt** perspektiv; språket hos individen

  Ex.: ASU

- Utpräglat **idiografiskt** perspektiv

  Ex.: Strindbergs brev (i Korp)

ASU ska kunna ”zooma in”:

- Involverar detaljarbete ner på person- och beläggnivå, tidpunkter, händelser
- Sökning → bearbetning av sökresultatet

Krav på ett gränssnitt

- Lagring i ordning efter korpusens indelning
- Flexibla korpusurval
- Flexibla sökmöjligheter
- Frekvensuppgifter, statistik
- Beläggställen i konkordanser och textvisning
- Textvisning: kontext till sökträffar; fulltext
- Bearbeta sökresultat, spara/återhämta versioner
- Exportera

Lagring i ordning efter korpusens indelning

Huvuddel (muntligt/skriftligt, inlärare/infödda)

> person (informant)

> textenhet (inspelningstillfälle, uppsats)

> tidsföljd i text

Visa konkordanser och text i samma ordning

Frekvensuppgifter, statistik

Räkningar på text

Textlängd (valda textdelar; antal löpord), talarval (informanten resp. andra talare), typ/tecken-relationer, ordfrekvenser, taggfrekvenser …

Skilja mellan *token* / (egentliga) *ord* / *skiljetecken*

Räkningar på sökträffar

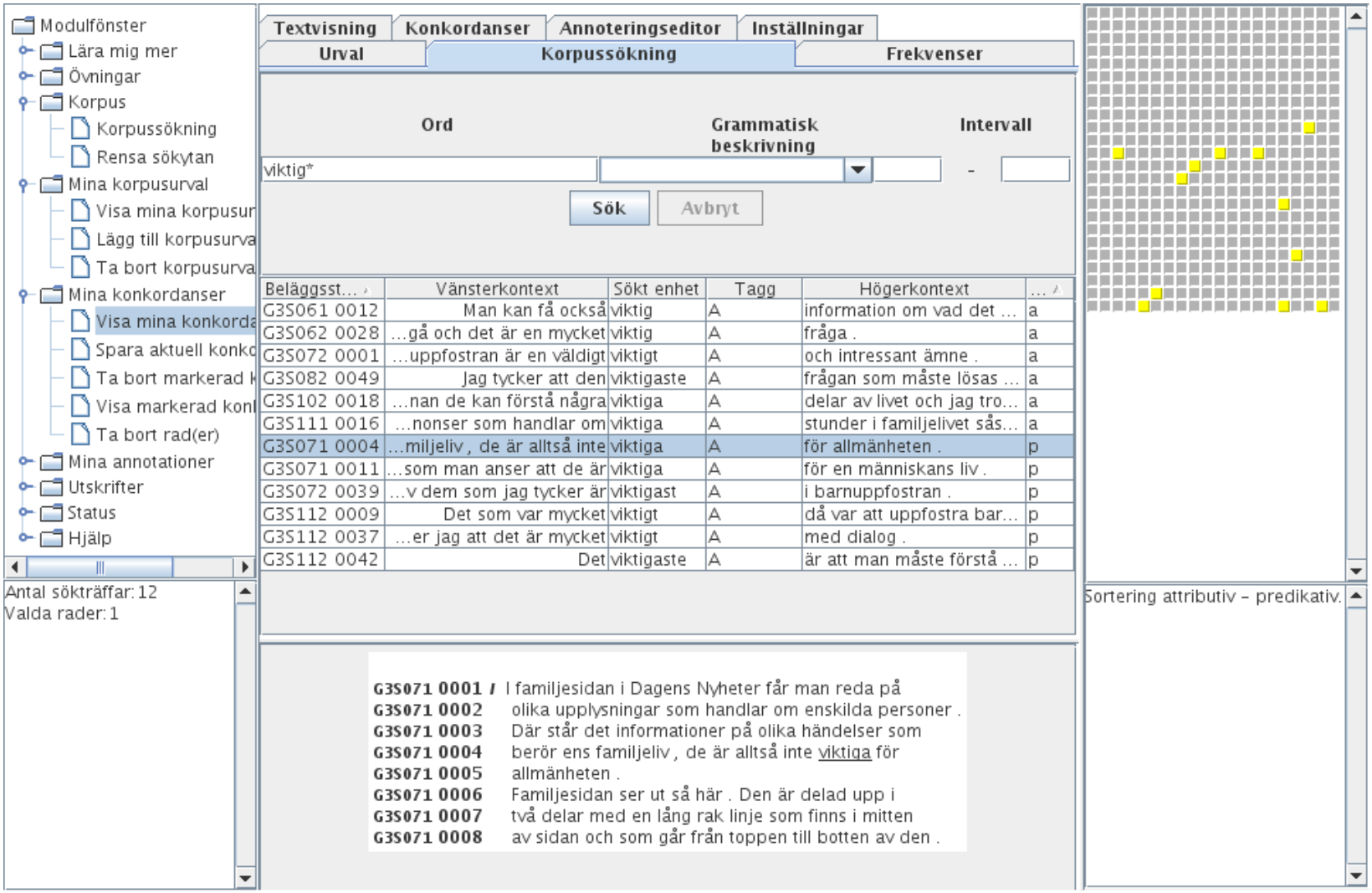
Antal träffar i konkordans, antal träffar av visst slag (angivna rader).

Räkningar på bearbetade konkordanser.

Bearbetning av konkordans

- Annotera (tilläggskategorisera) träffar
- Omsortera (även komplexa omsorteringar)
- Stryka överflödiga träffar (irrelevanta items, iterationer, ekon)
- Ny räkning efter bearbetning
- Benämna och spara/återhämta arbetsversioner
- Skriva ut

Konkordansegenskaper



Exempel på en egenuppmärkt och omsorterad ASU-konkordans i ITG-gränssnittet.

Exportera

Exportera frekvenser i text

- - -

Exportera konkordanser

t.ex. för vidare bearbetning i Excel:

- Hela konkordansen
- Intakt struktur

Forskningsfrågor: Ett detaljexempel

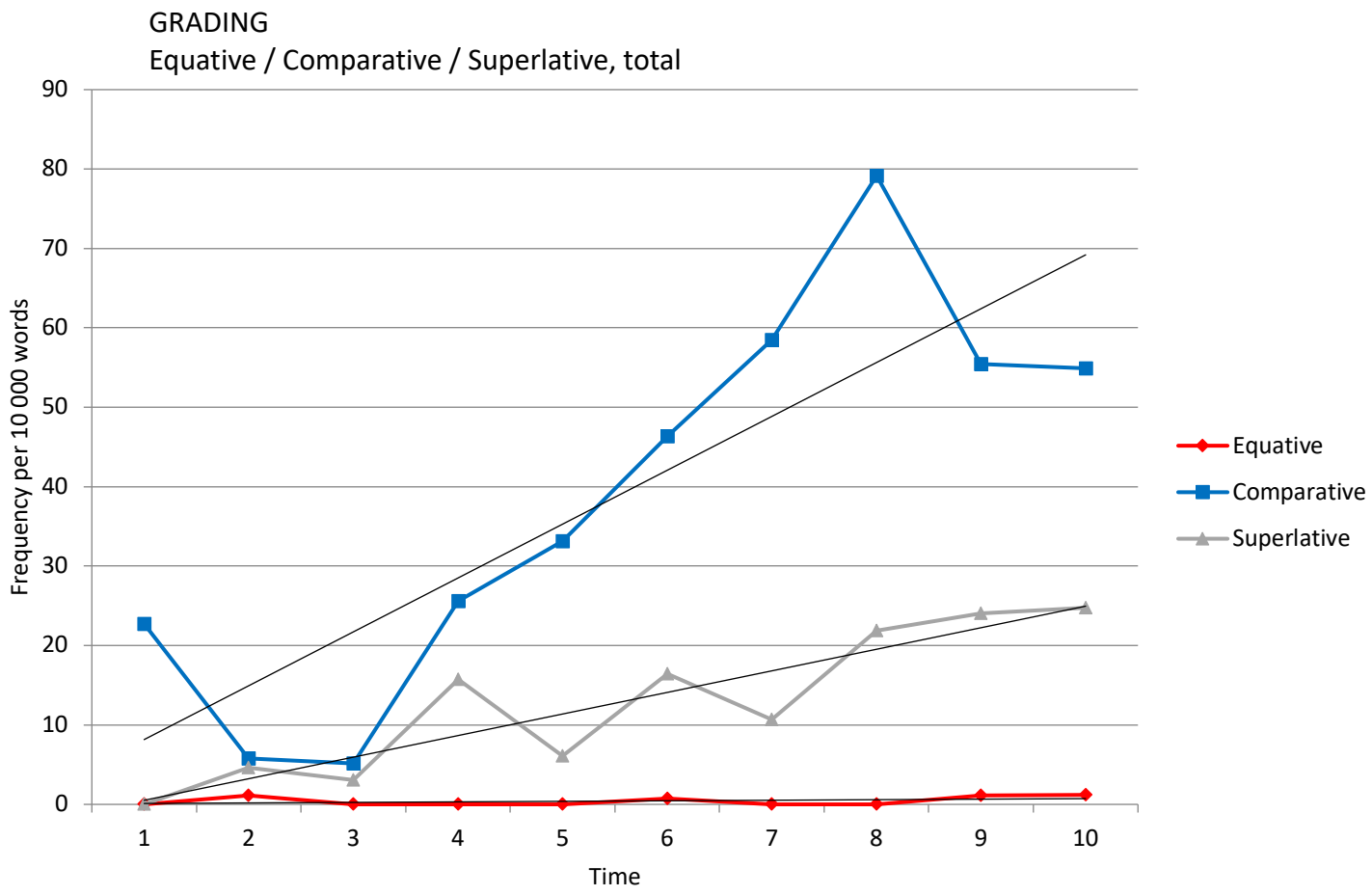


Figure 3. Developmental profile in the GRADING dimension:
equative vs. *comparative* (*inequative-exclusive*) vs. *superlative* (*inequative-inclusive*), total frequencies.
(Hammarberg, B. 2014. Constructions of comparison in Swedish: Quantitative dominance patterns in acquisition and use. *Constructions* 1-5/2014.)