



UPPSALA
UNIVERSITET

Trädgårdens storlek och dess växter

Beáta B. Megyesi

Institutionen för lingvistik och filologi
Uppsala universitet



Trädgårdens funktioner

- Trädbankens kärna skall utgöra gold standard annotation för parsningsstandarder för svenska
- Kunna användas för att
 - inducera grammatik
 - utvärdera parsrar



Trädgårdens sammansättning

Ett representativt urval av

- redan existerande
- morfo-syntaktiskt annoterade och
- manuellt kontrollerade korpusar
- för svenska
 - Talbanken (Einarsson, 1976; Nivre et al, 2006)
 - SUC (Ejerhed et al, 1992)



Innehåll: Talbanken

Ca 300 k tokens

Skriftspråksmaterialet

- Bruksprosa: 86 929 ord
- Gymnasistuppsatser: 88 528 ord

Talspråksmaterialet

- Intervjuer Borås: 67 025 ord
- Samtal: 21 031 ord
- Debatt: 26 706 ord



Innehåll: SUC

Kategori där varje kategori innehåller [#] texter á 2000 ord

A: reportage [44] (8,8%)

B: ledare, debattartiklar [17] (3,4%)

C: recensioner [27] (5,4%)

E: arbete & fritid [58] (11,6%)

F: populärvetenskap [48] (9,6%)

G: biografier, memoarer, essäer [26] (5,2%)

H: myndighetstexter [70] (14%)

J: lärda och vetenskapliga texter [83] (16,6%)

K: skönlitteratur [127] (25,4%)



Trädgårdens storlek och urval

- Storlek:
 - Minst 100 k tokens med tanke på state-of-the-art för trädbankers storlek idag
 - Högst 200 k tokens med tanke på tid och resurser för manuell rättning av annotering
- Trädgården bör vara en delmängd av Talbanken och SUC
- Sampling måste ske för urval av data



Samplingsmetoder 1

- Utan slump
 - Lättillgänglighetssampling (fråga släkt och vänner)
 - Styrd sampling (väljer element som passar bra)
 - Snöbollssampling (valet av ett element bidrar till att andra element väljs)
 - Kvotsampling (populationen delas in i olika kategorier och man letar reda på element för varje kategori)



Samplingsmetoder 2

Med slump

- Enkel slumpmässig sampling (alla element har lika stor chans att dras)
- Systematisk slumpmässig sampling (börjar på ett slumpmässigt ställe och drar var i :nde element)
- Stratifierad sampling (populationen delas in i kategorier (strata) från vilka slumpmässiga sampel dras)



Samplingsmetod och urval för trädgården

Stratifierad sampling:

– Representation:

- Vilka kategorier bör representeras?
 - Alla eller vissa, vilka i så fall?

– Urval av texter

- Proportionerlig fördelning över kategorier i trädgården som motsvarar SUCs och (ev.) Talbankens representation av kategorier

eller

- Ej proportionerligt: Ta de första N texterna i varje kategori



Trädgårdsdata - diskussion

- Talbanken - Bruksprosa: 86 929 ord
 - Ta hela eller sampla 50 k ord?
- SUC
 - Sampla 87 k ord eller 50 k ord?
 - Samplingsmetod:
 - alla kategorier eller vissa utvalda?
 - proportionerligt enligt SUC eller första N texter i varje kategori?



SUC – kategorier av intresse?

- Dagspress: reportage, ledare och recensioner
 - Kategori: A, B, C [176 k ord]
- Skönlitteratur: allmän, deckare, scifi, trivial, humor
 - Kategori: K [254 k ord]
- Eventuellt:
 - Populärvetenskap: F [96 k]
 - Biografi, memoarer: Ga [14 k]
 - Myndighetstexter: H [140 k]
 - Yrkes- och fackföreningspress: Ec [40 k]
 - Religiös veckopress: Ed [6 k]