# Kort och gott – Svenskt basordförråd

Katarina Mühlenbock, DART

VÄSTRA GÖTALANDSREGIONEN
SAHLGRENSKA UNIVERSITETSSJUKHUSET

# Vad är ett ord?

- Vi kan göra pauser då vi uttalar ett ord
- Ett ord kan oftast bytas ut mot ett annat med liknande funktion och betydelse
- Kan (oftast) stå som ensamt yttrande
- I svensk skrift markeras gränsen genom ett tomrum

# Ordbildning i olika språk

- Flekterande där orden böjs för att uttrycka grammatiska fenomen (alla indoeuropeiska): ex. *katt –er –na  jama –r -de*
- Agglutinerande där enskilda ord ändrar form för att kunna användas i olika grammatiska sammanhang, ex. finska: *talossani* (”i mitt hus”)

# Lite terminologi

- Korpus = en text eller samling texter, insamlade för ett visst ändamål
- Lemman = grundord
  - Verb: infinitivformen (hitta)
  - Substantiv: 1 pers sing (stol)
  - Adjektiv: positiv singular (liten)
- Lexem = betydelseenhet i det semantiska systemet (ex. ”krona”)

**krona** substantiv ~*n kronor* **1** en huvudprydnad för kungliga personer **2** *Kronan* statsmakten **3** översta del av ngt: *trädets krona* **4** myntenheten i bl.a. Sverige: *reformerna är finansierade krona för krona (SAOL)*

# För avgränsa ett ordförråd måste orden kategoriseras

- Efter ordklass
- Efter användning (talspråk/skriftspråk)
- Efter ”age-of-acquisition” – i vilken ordning uppträder vanligen orden i ett barns ordförråd?
- Efter domän (medicinskt, slang, låneord)
- Efter betydelse (semantisk kategori)

# Kategorisering efter ordklasser

- Böjliga (öppna) ordklasser: Verb, substantiv, adjektiv, pronomen och räkneord
- Oböjliga (slutna) ordklasser: Prepositioner, konjunktioner, subjunktioner, interjektioner och adverb

# Hur stort är ett ”normalt” ordförråd?

- En treåring kan mellan 1 000 och 3 000 ord
- När man börjar skolan kan man i snitt 7 000 ord
  - kan skilja mycket mellan de språksvaga som kan ca 5 000 ord och språkstarka som kan upp till 20 000 ord
- För att kunna läsa en artikel i en dagstidning eller en för en 15-16 åring avpassad bok behöver vi ”kunna” 50-70 000 ord

# Basordförråd

- På 1930-talet extraherade Ogden 850 basord för engelska
- De Mauro gjorde på 1980-talet en indelning av 8 000 ord som kan vara fundamentala för vardaglig kommunikation:
  - Vardagliga ord som vid testning har visat sig vara begripliga för grundskoleelever
  - En delmängd av ovanstående som är högfrekventa
  - En uppsättning ord som inte förekommer särskilt ofta i vare sig skrift eller tal, men som är starkt knutna till vardagliga situationer och föremål, ex. *tandborste*

# Hur skapar man ett basordförråd?

- Tittar på hur språket används i olika uttrycksformer
  - Skrift
  - Tal
- Räknar frekvenser av olika slag
  - Ordförekomster
  - Lemmaförekomster

# Ordförrådets bredd och djup

- Ett ordförråds **bredd** är hur många ord man kan
  - hur många ord man kan ge en definition till eller matcha till en bild.
- Ett ordförråds **djup** hänger samman med hur bra man kan orden i sitt ordförråd.
  - Kan man använda orden i flera olika kontexter?
  - Kan man definiera orden i detalj?
  - Vet man hur orden relaterar till andra ord?
  - Kan man förstå orden utan kontext?

# De 100 vanligaste orden i svenska språket

i
och
att
det
som
en
på
är
av
för
med
till
den
har
de
inte
om
ett
han
men

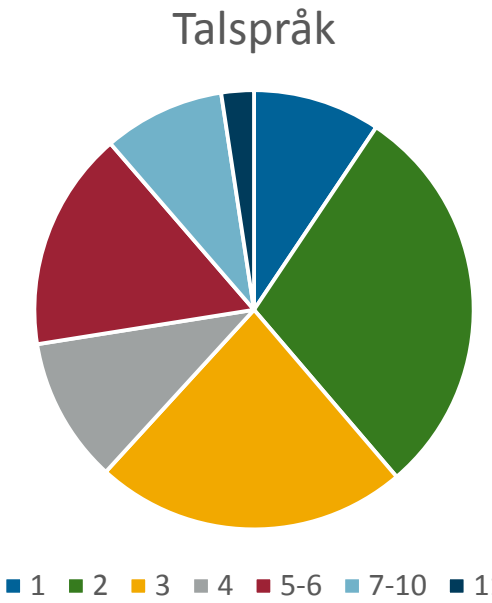
var
jag
sig
från
vi
så
kan
man
när
år
säger
hon
under
också
efter
eller
nu
sin
där
vid

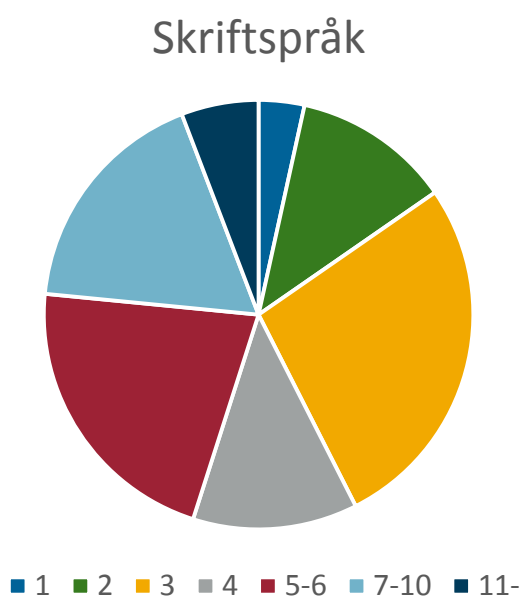
mot
ska
skulle
kommer
ut
får
finns
vara
hade
alla
andra
mycket
än
här
då
sedan
över
bara
in
blir

upp
även
vad
få
två
vill
ha
många
hur
mer
går
sverige
kronor
detta
nya
procent
skall
hans
utan
sina

något
svenska
allt
första
fick
måste
mellan
blev
bli
dag
någon
några
sitt
stora
varit
dem
bland
kl
bra
tre

# Ordlängd i tal och skrift




Från Talspråksfrekvenser. **Frekvenser för ord och kollokationer i svenskt tal- och skriftspråk**. Gothenburg Papers in Theoretical Linguistics, April 1999. Red. J. Allwood

# Exempel från korpusmaterial

- LäSBarT är en korpus med 1 114 305 löpord
- Består av
  - Lättläst skönlitteratur för vuxna
  - Barnlitteratur
  - Lättläst nyhetstext
  - Lättläst samhällsinformation
- ~ 43 000 ordförekomster (tokens)
- ~ 20 000 lemman (grundformer)

# Ordfrekvenser i LäSBarT

Barnbokstext (138 000 ord)


Regeringskansliets årsbok (5 000 ord)


8 SIDOR (156 000 ord)


# Ordfrekvenser i LäSBarT, 2-bokstavsord

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| en | 13 721 | ge | 593 | yr | 21 | it | 3 |
| du | 8 668 | ur | 564 | is | 15 | mc | 2 |
| de | 7 096 | ny | 430 | aj | 14 | km | 2 |
| vi | 5 647 | bo | 220 | kr | 13 | hy | 2 |
| om | 5 206 | er | 217 | ko | 11 | cd | 2 |
| av | 5 146 | jo | 198 | by | 11 | an | 2 |
| sa | 3 927 | va | 122 | hm | 9 | ah | 2 |
| nu | 3 336 | be | 113 | oh | 8 | tu | 1 |
| ut | 2 854 | la | 77 | mm | 7 | my | 1 |
| ha | 2 169 | oj | 60 | kg | 6 | mu | 1 |
| in | 1 900 | ro | 45 | sy | 5 | il | 1 |
| ju | 1 315 | el | 42 | kl | 5 | hu | 1 |
| ta | 1 301 | tv | 41 | go | 4 | ek | 1 |
| se | 1 205 | fy | 33 | di | 4 | ej | 1 |
| ni | 868 | le | 29 | al | 4 | dy | 1 |
| ja | 847 | te | 28 | nr | 3 | bi | 1 |

# Ordfrekvenser i LäSBarT, 3-bokstavsord

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| och | 29 041 | vad | 3 027 | dom | 1 101 | hos | 467 |
| att | 25 702 | hur | 2 934 | ner | 1 032 | mej | 465 |
| det | 24 188 | mig | 2 673 | oss | 966 | tag | 445 |
| jag | 15 265 | man | 2 582 | tog | 900 | hus | 436 |
| han | 13 456 | mot | 2 430 | min | 875 | mat | 393 |
| har | 11 076 | upp | 2 104 | tar | 747 | bor | 379 |
| som | 10 704 | bli | 1 853 | nya | 744 | ger | 348 |
| med | 10 162 | bra | 1 728 | nej | 721 | dej | 342 |
| hon | 8 938 | kom | 1 700 | nog | 657 | ens | 323 |
| var | 8 887 | dem | 1 460 | dag | 648 | sej | 312 |
| ska | 7 311 | sin | 1 457 | hem | 607 | dog | 307 |
| men | 6 587 | mer | 1 355 | tre | 585 | din | 301 |
| kan | 6 349 | ser | 1 343 | del | 518 | dit | 285 |
| sig | 5 236 | vet | 1 283 | vem | 510 | tio | 267 |
| ett | 5 186 | vid | 1 187 | sen | 497 | bil | 261 |
| den | 4 297 | dig | 1 164 | tid | 473 | liv | 260 |

# SweVoc

- Sammanslagning av SBVP (SUC-corpus), Kelly-ordlistan (läroboksord), ICF (ord för kroppsfunktioner, ord ur individ- och samhällsperspektiv)
- 8 000 lemman uppdelade i:
  - Basvokabulär (2 201)
  - Vardagliga föremål och händelser (1 019)
  - Högfrekventa ord i skriven text (1 518)
  - Läromedelsord (288)
  - Supplementära ord (3 442)

# Verb

- **Huvudverb**: betecknar en *aktion*
- **Hjälpverb** används som *formord*:
  - Temporala: ha, ska
  - Passivbildande: bli, vara
  - Modala: bör, måste, kunna, vilja, låta
- **Lös** och **fast** sammansättning *Ex. tillsätta grädde / sätta till grädde*
  - En fast sammansättning har ofta en mer abstrakt betydelse, ex. avbryta/bryta av, uppgå/gå upp

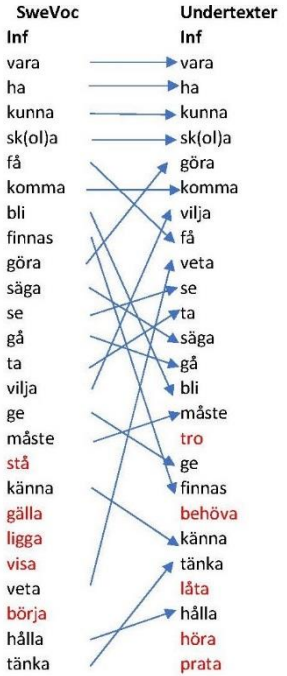
# Verbens betydelsefält

- Fysiska fältet
  - Befintlighet (bo, ligga, stå)
  - Rörelse (flytta, lägga)
  - Produktion (baka, bygga)
  - Konsumtion (dricka, äta)
  - Fysisk hantering (måla, skära)
- Psykiska och sociala fältet
  - Varseblivning (höra, se)
  - Kommunikation (be, skriva)
  - Tänkande och minne (förstå, tro)
  - Avsikt (försöka, planera)
  - Önskan (hoppas, ångra)
  - Förmåga (kunna, orka)
  - Socialt handlande o relationer (lyda, gifta sig)
- Logiska fältet
  - Orsak, verkan (bero på)
  - Likhet (jämföra)
  - Beteckning (betyda)
  - Existens (finnas, försvinna)
  - Kvantitet (minska, väga)
  - Innehav (få, ge)
  - Del/helhet (bestå av)

# Högfrekventa Verb i SweVoc och talat språk

Jämförelse av verbfrekvenser i skrivet och talat språk

- de fyra i topp är identiska
- 67 % av samtliga finns både i tal och skrift
- i talat språk fler verb som tillhör det psykiska och sociala fältet



# Referenser


- De Mauro, T. 1980. *Guida all'uso delle parole*. Roma: Editori Riuniti.
- Forsbom, E. 2006: A Swedish Base Vocabulary Pool. *Proc. of SLTC*. Dept. of Linguistics and Philology. Uppsala University.
- Heimann Mühlenbock, K. 2013. *I see what you mean*. Assessing readability for specific targets group. Avhandling, inst. för svenska språket, Göteborgs universitet.
- Heimann Mühlenbock, K. & Johansson Kokkinakis, S. 2012. SweVoc – A Swedish vocabulary resource for CALL. *Proc of SLTC Workshop on NLP for CALL*. Lund: Linköping University Electronic Press.
- Johansson Kokkinakis, S. & Volodina, E. 2011. Corpus-based approaches for the creation of a frequency based vocabulary list in the EU project Kelly. *eLex Conference*. Slovenia.
- Ogden, C.K. 1930. *Basic English: A general introduction with rules and grammar*. London: Paul Treber & Co. Ltd.
- WHO 2008. International Classification of functioning, disability and health (ICF).


VÄSTRA
GÖTALANDSREGIONEN
SAHLGRENSKA UNIVERSITETSSJUKHUSET