Fördjupningsuppgift

Kickoff

LINKÖPINGS UNIVERSITET

Lärandemål

- Redogöra för grundläggande begrepp inom språkvetenskap, särskilt morfologi, syntax och semantik.
- Utföra grundläggande språkvetenskapliga analyser såsom ordklassbestämning och dependensanalys.
- Använda datorbaserade verktyg och programmering för att samla in, analysera och validera språkliga datamängder (korpusar).
- **Planera** och **utföra** enklare korpusundersökningar, samt **redovisa** och **värdera** resultaten.

Läsbarhet

- Begreppet läsbarhet syftar på ”summan av sådana språkliga egenskaper hos en text, vilka gör den mer eller mindre svårtillgänglig för läsaren”. (Björnsson, 1968).
- Hur läsbar en text är beror bl.a. på dess mängd, dess uppbyggnad, samt på komplexiteten hos olika morfologiska och syntaktiska formationer som förekommer i den.
- Kan man mäta läsbarheten hos en text?

Kommunen har ansvar

Kommunen har ansvar för att det finns ledningsnät för vatten och avlopp.

Det går att koppla sin bostad

till kommunalt vatten och avlopp

i de flesta delar av Linköping där det är tätort,

alltså där många bostäder är samlade.

Kommunen ansvarar för att det finns ledningsnät för vatten och avlopp. Det är möjligt att ansluta till kommunalt vatten och avlopp i de flesta av Linköpings tätortsområden.

Komplexitetsmått

- Ta fram numeriska värden för att beskriva en texts komplexitet.

- Men:
  - Många faktorer påverkar komplexitet (mängd, ordval, syntax, struktur,…). Ska vi kombinera dem till ett mått, ha ett urval med olika mått, eller finns det ett mått som är representativ för alla egenskaper?
  - Värden kan vara svårt att tolka.

Ett vanligt mått för svenska

- Måttet LIX (Läsbarhetsindex) viktar antalet långa ord och antalet meningar i texten mot antalet ord i hela texten.

$$\text{LIX} = \frac{\text{antal ord i texten}}{\text{antal meningar i texten}} + \frac{\text{antal långa ord} \cdot 100}{\text{antal ord i texten}}$$

- Ett ord räknas som långt om det innehåller fler än 6 bokstäver.

LIX-tal för olika slag av texter

LIX-värde	Texttyp
under 25	barnböcker
25–30	enkla texter
30–40	normaltext/skönlitteratur
40–50	sakinformation, t.ex. Wikipedia
50–60	facktexter
över 60	svåra facktexter, forskning, avhandlingar

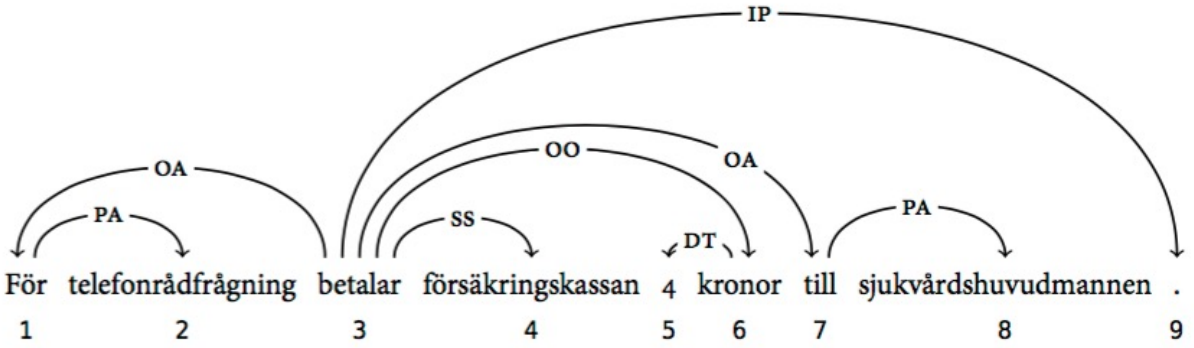
Komplexitetsmått i denna studie

Nominalkvot

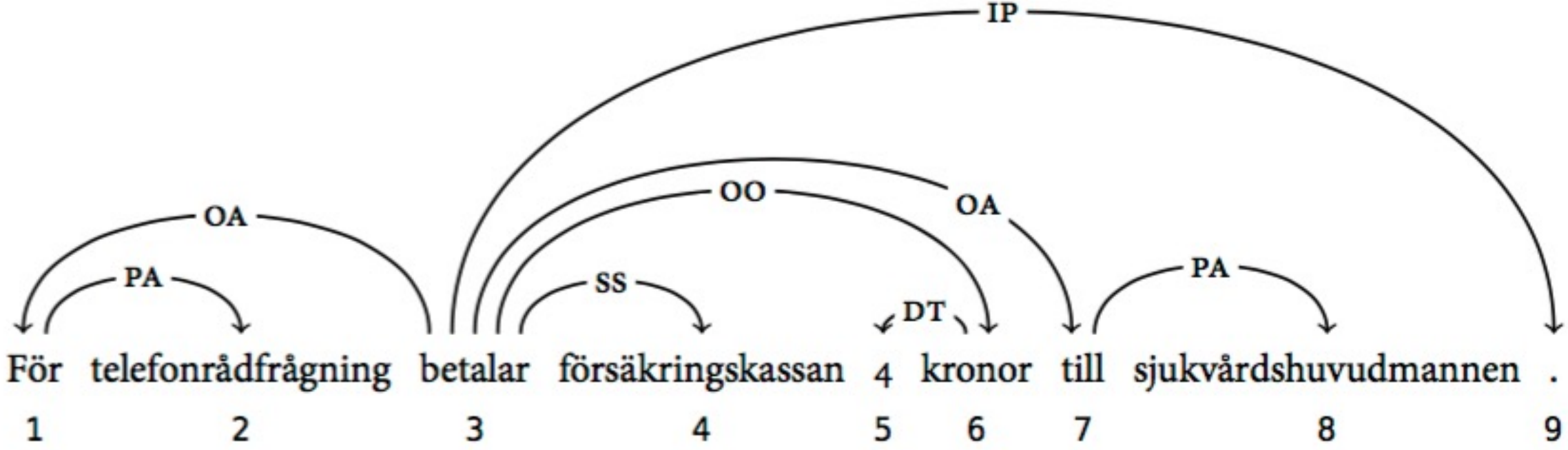
- Kvoten mellan ord med nominal karaktär (substantiv, prepositioner, particip) och ord med verbal karaktär (verb, pronomen, adverb).

Genomsnittlig dependenslängd

- Det genomsnittliga avståndet mellan ett ord och dess huvudord i en dependensanalys av texten.
  - Hur många ord måste man passera för att komma till huvudet?



Genomsnittlig dependenslängd



Bedömningskriterier

För att bli **godkänd** på denna uppgift ska du

- implementera ett skript som beräknar nominalkvot och genomsnittlig dependenslängd för en given korpus
- använda ditt skript för att jämföra syntaktisk komplexitet hos två olika korpusar (Göteborgsposten, 8 sidor)
- redovisa din metod, dina resultat, din diskussion och dina slutsatser i en skriftlig rapport

Dimensioner för bedömning

- Hur mycket material är rapporten baserad på?
- Hur komplexa är argumenten i rapporten?
- I vilken utsträckning problematiserar rapporten?
- I vilken utsträckning används konkreta exempel?
- I vilken utsträckning innehåller rapporten relevanta slutsatser?
- Hur väl genomarbetad är rapporten (struktur, språk)?

Tilltänkt läsare

- När jag och Daniel bedömer din rapport kommer vi läsa den som om vi vore personer som inte alls vet vad uppgiften handlar om.
- Detta innebär att du behöver förklara utförligt **vad** du har gjort, **varför** du har gjort det och **hur** du har gjort.
  - Skriv för den som du var första dagen på programmet!
- Detta innebär t.ex. att du inte ska förutsätta att läsaren kan något om korpusar eller programmering.

Kriterier för Väl Godkänd

- För betyget VG krävs en **omfattande** diskussion med **välutvecklade** omdömen. En sådan diskussion kräver i normalfallet en utökning av din studie bortom de ramar som presenteras nedan. Exempel:
  - undersöka andra komplexitetsmått
  - inkludera fler data än de givna korpusarna eller
  - utföra fler analyser

- Även en välskriven **metodkritik** som mynnar ut i konkreta förslag på hur studien kan förbättras kan motivera det högre betyget.

Formella kriterier

- mellan 2 000 och 4 000 ord, motsvarande ca. 4–8 sidors text (exkl. titel, bilder, tabeller, formler, referenser)

- akademiskt språk, väl genomarbetad och korrekturläst
  - referenser enligt vedertaget format
  - t.ex. det som du använt på introduktionskursen

Struktur

- **Inledning**. Förklara syftet med studien med egna ord.
- **Teori**. Definiera läsbarhetsmåtten med egna ord.
- **Data**. Beskriv datamaterialet med egna ord.
- **Metod**. Beskriv hur du genomfört ditt arbete.
- **Resultat**. Presentera dina resultat på ett objektivt sätt.
- **Diskussion**. Ge din tolkning av resultaten gentemot frågeställningen.
- **Slutsatser**. Dra tydliga slutsatser. Vad blev svaret på frågan?

Inledning

Beskriv studien översiktligt och sätt den i ett sammanhang. Försök att göra läsaren intresserad av arbetet.

- Vad är syftet med studien?
- Varför är studien intressant och viktig?
- Hur ska syftet uppnås?

Teori

- Definiera läsbarhetsmåtten med egna ord.
- Referera till arbeten som länkas till från kurshemsidan och till annat material du hittar om läsbarhetsmåtten.
- Tips: Illustrera läsbarhetsmåtten med exempel!

Data

- Beskriv datamaterialet med egna ord.
- Tips: Inkludera deskriptiv statistik, t.ex. antal meningar.
- Ju mer du vet om hur datan kom till, desto bättre!

Metod

- Beskriva hur du faktiskt genomfört din korpusstudie.
- **Replikerbarhet:** Någon som har läst din metodbeskrivning ska kunna göra om samma studie och få samma resultat.
- Tips: Fundera över vilka metodbeslut du fattade.
- Kod hör **inte** hemma i metoddelen. Den som läser din metoddel ska inte behöva kunna Python.

Metod – för många detaljer

I funktionen för att beräkna textkorpusarnas nominalkvot skapades först två listor med de ordklasstaggar som var relevanta. Den ena innehöll taggarna för ord av nominal karaktär och den andra innehöll de taggar som klassificerar ord av verbal karaktär. Sedan öppnades korpusfilen och placerades i en lista för att kunna formateras så att den blev lätthanterlig. Därefter itererades denna lista igenom efter ordklasstaggarna på index 3 och jämfördes med de två listorna med relevanta taggar. För varje tagg i korpusfilen som matchade någon av taggarna i någon av de två listorna ökades antingen nominal_count-variabelns eller verbal_count-variabelns värde med ett, beroende på vilken ordklasstagg som hade hittats i korpusfilen. Då listan med korpusfilen loopats igenom färdigt någon av listorna dividerades den variabel som räknade antalet ord med nominal karaktär med den variabel som räknat de ord med verbal karaktär. Kvoten utav detta var nominalkvoten.

Resultat

- Presentera dina resultat på ett objektivt sätt. Tolka inte!
- Dina resultat är underlaget till din diskussion.
- Tips: Använd tabeller eller grafer!

Resultat – använd tabeller!

Tabell 1. *Nominalkvot och genomsnittlig dependenslängd hos korpusarna.*

	Nominalkvot	Genomsnittlig dependenslängd
Göteborgsposten		
8 sidor		

Presentera med fördel dina resultat i en tabell eller en graf
men glöm inte att även sammanfatta resultaten i en kort text.

Diskussion

- **Resultatdiskussion**
  Vad blev resultaten? Hur tolkar du dem?
- **Metoddiskussion**
  Var din metod lämplig? Vilka begränsningar har den?
- **Framtida forskning**
  Vilka fortsättningar eller utvidgningar finns det för ditt arbete?

Diskussion: exempel

- Ger måtten de förväntade resultaten?
- Mäter måtten det de ska (validitet, reliabilitet)?
- Vilka resurser krävs det för att göra dessa mätningar?
- Hur skulle man kunna använda mätningarna?
- Kan man egentligen mäta läsbarhet?
- Vad är läsbarhet, egentligen?

Talspråk versus formellt språk

Talspråk	Formellt språk
våran, vårat	vår, vårt
dom	de, dem
mej, dej, sej	mig, dig, sig
medans	medan
kolla	kontrollera, undersöka, se, …
nåt	något

Talspråk versus formellt språk

Talspråk	Formellt språk
våran, vårat	vår, vårt
dom	de, dem
mej, dej, sej	
	medan
kolla	kontrollera, undersöka, se, …
nåt	något

Men formellt språk behöver inte innebära krångligt språk!

Några enkla språkregler

- Använd fullständiga meningar (subjekt, predikat)
- Undvik personliga pronomen (du, vi, ni, etc.)
- Undvik ”visa på att”
- Inled inte meningar med ”Detta för att”
- **Kontrollera stavfel, särskilt särskrivningar**

Skrivhjälp av Generativ AI

- *"För fördjupningsuppgiften (UPG1) får dessa verktyg endast användas för att hjälpa dig att förbättra eller korrigera text som du redan har skrivit; de får inte användas för att generera helt ny text till din rapport, oavsett om du bara kopierar texten eller omformulerar den."*

Skrivhjälp av Generativ AI

Metod

För att utföra vår studie och beräkna nominalkvot och dependenslängd för texterna från Göteborgsposten och 8 sidor använde vi oss av en strukturerad process. Nedan beskriver vi stegen som vi följde för att samla in och analysera data samt beräkna de syntaktiska komplexitetsmåtten:

1. **Insamling av data**: Vi samlade in datamaterialet från två källor: Göteborgsposten och 8 sidor. Dessa källor tillhandahöll nyhetsartiklar som representerar olika nivåer av språklig komplexitet. Datamaterialet var särskilt utvalt för att omfatta en bredd av texter för att säkerställa en representativ analys.
2. **Preprocessering av data**: Innan vi kunde beräkna nominalkvot och dependenslängd för texterna, utförde vi preprocessering av datamaterialet. Detta inkluderade att omvandla texterna till ett format som var lämpligt för analys, inklusive tokenisering och annotering av ordklasser och huvudord.

???

Kod

- Skulle vara läsbar.
- Enkla kommentarer vilket mått en funktion/del tillhör.
- Använd gärna funktioner och kodskelett från labbarna.

- Även här kan ni ta hjälp av AI-verktyg för att avbugga koden.
    - Kopiera dock aldrig större sekvenser och installera inga nya paket; ni måste förstå koden ni har skrivit helt och hållet!
- Att kopiera andras kod är plagiat.

Återkoppling, resultatseminarium

- Frivilligt
- Återkoppling på resultat
- 14 maj 10-12 (inlämning 7 maj)
- Ingen annan förberedelse, ta med resultat
- Få och ge feedback

Återkoppling, rapportseminarium

- Frivilligt
  - Inte betygsgrundande, återkoppling på utkast från oss lärare och övriga i seminariegrupp.
- Lämna in rapport som ett PDF-dokument i Lisam/Inlämningar/Slutrapport
- Döp filen till ditt LiU-ID som filnamn, exempelvis **kunje68.pdf**
- Inlämningsdatum 21 maj
- Seminarier 27 maj i mindre grupper
  - Tiden för seminariegrupperna kommer efter inlämningen.
  - Varje seminarium tar ca. 45 minuter.
  - Ta emot och ge konstruktiv återkoppling på rapportutkasten, samt att diskutera uppgiften i sin helhet.

Slutgiltig inlämning

- Lämna in rapport som ett PDF-dokument i Lisam/Inlämningar/Slutrapport
- Döp filen till ditt LiU-ID som filnamn, exempelvis **kunje68.pdf**
- Bifoga kod som en Python-fil (.py eller .ipynb).
- Inlämningsdatum: 2024-05-31
- Bedömning: Rapporten bedöms med ett av betygen U, G eller VG enligt betygskriterierna. Detta betyg blir betyg på examinationsmomentet UPG1.

Frågor?

Läsbarhet

Mer Bakgrund & Forskning @ LiU

LINKÖPINGS UNIVERSITET

Förbättra läsbarhet

- Automatisk sammanfattning
- Omskrivning till lätt svenska
  - Använda lättare synonymer
  - Förklara ord
  - Enklare och kortare meningar
  - Struktur och rubriker
  - Skriva om metafor, negationer, förkortningar

https://www.mtm.se/vagledning-och-lektioner/att-skriva-lattlast/

Forskning på LiU

- Evelina Rennes, Arne Jönsson m.m.
- Tjänster för ökad läsbarhet på webben
- Rangordning av texter efter läsbarhet samt klustring av texter med samma innehåll
- Nya läsbarhetsmodeller
- Verktyg för att välja ut texter anpassade efter läsförmåga

Evelinas avhandling:
https://www.diva-portal.org/smash/record.jsf?pid=diva2%3A1647431&dswid=3379

Läsning: en mänsklig rättighet

- ”…utan dröjsmål och extra kostnader förse personer med funktionsnedsättning med information som är avsedd för allmänheten i tillgängligt format och teknologi anpassad för olika funktionsnedsättningar”

Exempel på lässvårigheter

Målgrupp	Exempel på upplevda svårigheter
Dyslexi	Långa och ovanliga ord, homofoner, ord som är ortografiskt lika, nya ord, icke-ord
Afasi	Hög informationsdensitet, långa meningar, långa sekvenser av adjektiv, passiv form, sammansatta ord
Andraspråksinlärning	För litet vokabulär, kulturella företeelser, textstruktur
Hörselskada	Komplexa grammatiska konstruktioner, generalisering av ord till en vidare kontext, textstruktur
Intellektuell funktionsnedsättning	Svårigheter relaterade till arbetsminnet, motivation till läsning

Riktlinjer för lättläst

https://www.mtm.se/var-verksamhet/lattlast/att-skriva-lattlast/

One size fits all… eller?

- Dagens lösningar anpassar ofta en text efter generella riktlinjer
- Med dagens teknik borde vi kunna anpassa teknikerna efter folk med olika typer av läsbeteende
  - Kanske till och med individuell anpassning?
- Inte säkert att personer **inom** en grupp har samma behov

Utmaningar

Känn till vem läsaren är!

- Figurativt språk?
- Förklara kulturella fenomen?
- Förklara svåra ord?
- Dela upp långa meningar?
- Förenkla grammatiken?

Utmaningar

Överförenkla inte!

- Utvecklingsaspekt
- Om texten är *för* enkel kan det störa läsupplevelsen

Datainsamling

Korpusar

- LäSBarT, SUC.
- Alla vanliga och lättlästa texter från offentliga förvaltningar i Sverige, myndigheter, kommuner, landsting m.fl.
- Parallellställning, alignment.

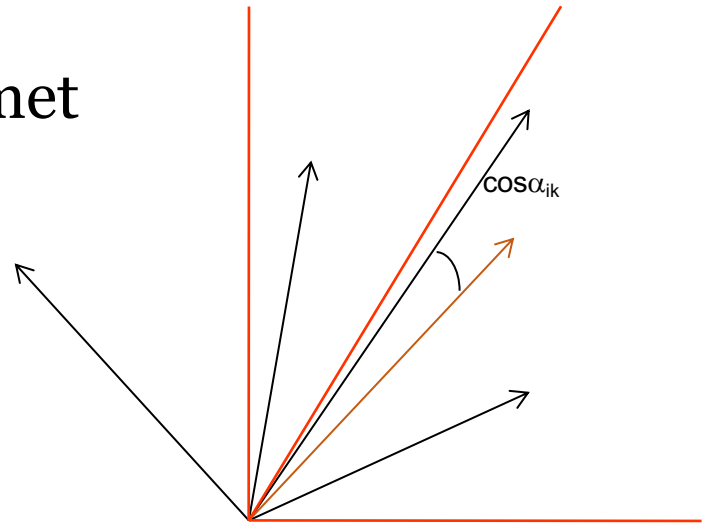
	Vanlig svenska	Lätt svenska
Antal dokument	115 027	2247
Antal meningar	1 333 474	26 461
Antal tecken	20 649 700	338 977
Meningslängd, medel	15	12
LIX	52	44
SweVoc	49%	50%

Teknik

- Vektorrums- och grafmodeller
- Maskininlärning
- Regler
- Korpusinsamling och träning
- Användarcentrerad utveckling
- Utvärderingar
  - Guldstandarder
  - Ögonrörelsemätningar, frågeformulär

Vektorrumsmodeller

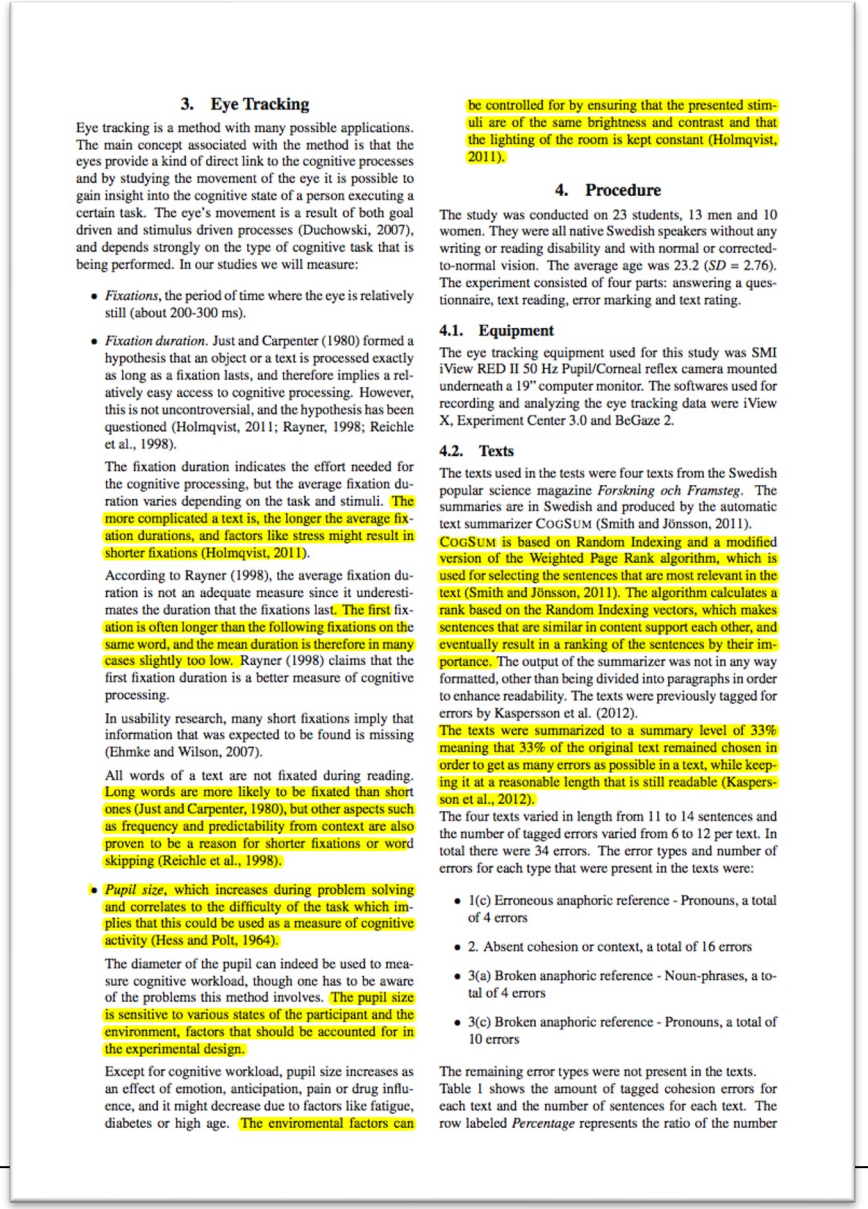
- Tekniker som reducerar den lingvistiska variationen och fångar semantiskt relaterade begrepp
- Ord representeras som vektorer (word embeddings)
- Meningar fås genom att addera ordvektorer, dokument genom att addera meningsvektorer
- Likhet mellan dokument mäts som närhet i vektorrummet



Exempel: Textsammanfattning

- **Extraktiv**: Hitta bara de semantiskt mest relevanta meningarna.
- **Abstraktiv**: Stora språkmodeller genererar en kortare text.

- Kommer i Språkteknolog-kursen (729G17)!



3. Eye Tracking

Eye tracking is a method with many possible applications. The main concept associated with the method is that the eyes provide a kind of direct link to the cognitive processes and by studying the movement of the eye it is possible to gain insight into the cognitive state of a person executing a certain task. The eye's movement is a result of both goal driven and stimulus driven processes (Duchowski, 2007), and depends strongly on the type of cognitive task that is being performed. In our studies we will measure:

- *Fixations*, the period of time where the eye is relatively still (about 200-300 ms).
- *Fixation duration*. Just and Carpenter (1980) formed a hypothesis that an object or a text is processed exactly as long as a fixation lasts, and therefore implies a relatively easy access to cognitive processing. However, this is not uncontroversial, and the hypothesis has been questioned (Holmqvist, 2011; Rayner, 1998; Reichle et al., 1998).

  The fixation duration indicates the effort needed for the cognitive processing, but the average fixation duration varies depending on the task and stimuli. The more complicated a text is, the longer the average fixation durations, and factors like stress might result in shorter fixations (Holmqvist, 2011).

  According to Rayner (1998), the average fixation duration is not an adequate measure since it underestimates the duration that the fixations last. The first fixation is often longer than the following fixations on the same word, and the mean duration is therefore in many cases slightly too low. Rayner (1998) claims that the first fixation duration is a better measure of cognitive processing.

  In usability research, many short fixations imply that information that was expected to be found is missing (Ehmke and Wilson, 2007).

  All words of a text are not fixated during reading. Long words are more likely to be fixated than short ones (Just and Carpenter, 1980), but other aspects such as frequency and predictability from context are also proven to be a reason for shorter fixations or word skipping (Reichle et al., 1998).

- *Pupil size*, which increases during problem solving and correlates to the difficulty of the task which implies that this could be used as a measure of cognitive activity (Hess and Polt, 1964).

  The diameter of the pupil can indeed be used to measure cognitive workload, though one has to be aware of the problems this method involves. The pupil size is sensitive to various states of the participant and the environment, factors that should be accounted for in the experimental design.

  Except for cognitive workload, pupil size increases as an effect of emotion, anticipation, pain or drug influence, and it might decrease due to factors like fatigue, diabetes or high age. The enviromental factors can be controlled for by ensuring that the presented stimuli are of the same brightness and contrast and that the lighting of the room is kept constant (Holmqvist, 2011).

4. Procedure

The study was conducted on 23 students, 13 men and 10 women. They were all native Swedish speakers without any writing or reading disability and with normal or corrected-to-normal vision. The average age was 23.2 (SD = 2.76). The experiment consisted of four parts: answering a questionnaire, text reading, error marking and text rating.

4.1. Equipment

The eye tracking equipment used for this study was SMI iView RED II 50 Hz Pupil/Corneal reflex camera mounted underneath a 19" computer monitor. The softwares used for recording and analyzing the eye tracking data were iView X, Experiment Center 3.0 and BeGaze 2.

4.2. Texts

The texts used in the tests were four texts from the Swedish popular science magazine *Forskning och Framsteg*. The summaries are in Swedish and produced by the automatic text summarizer CogSum (Smith and Jönsson, 2011).
CogSum is based on Random Indexing and a modified version of the Weighted Page Rank algorithm, which is used for selecting the sentences that are most relevant in the text (Smith and Jönsson, 2011). The algorithm calculates a rank based on the Random Indexing vectors, which makes sentences that are similar in content support each other, and eventually result in a ranking of the sentences by their importance. The output of the summarizer was not in any way formatted, other than being divided into paragraphs in order to enhance readability. The texts were previously tagged for errors by Kaspersson et al. (2012).
The texts were summarized to a summary level of 33% meaning that 33% of the original text remained chosen in order to get as many errors as possible in a text, while keeping it at a reasonable length that is still readable (Kaspersson et al., 2012).
The four texts varied in length from 11 to 14 sentences and the number of tagged errors varied from 6 to 12 per text. In total there were 34 errors. The error types and number of errors for each type that were present in the texts were:

- 1(c) Erroneous anaphoric reference - Pronouns, a total of 4 errors
- 2. Absent cohesion or context, a total of 16 errors
- 3(a) Broken anaphoric reference - Noun-phrases, a total of 4 errors
- 3(c) Broken anaphoric reference - Pronouns, a total of 10 errors

The remaining error types were not present in the texts. Table 1 shows the amount of tagged cohesion errors for each text and the number of sentences for each text. The row labeled *Percentage* represents the ratio of the number

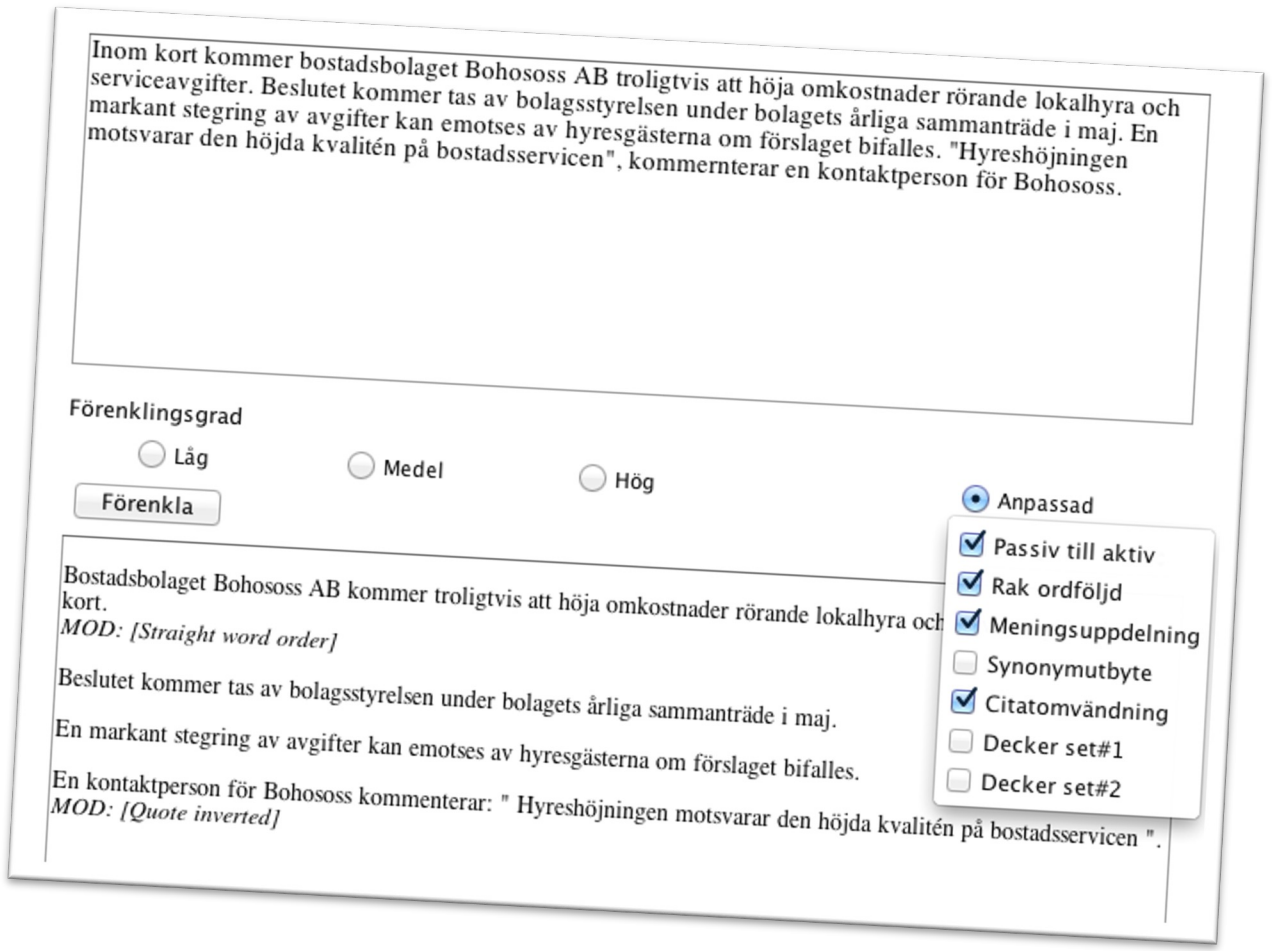
Textförenklingar

Regelbaserat verktyg

- Dependensgrammatik

Modell för textförenkling

- Träna modell
  - LSTM, encoder-decoder



Textkomplexitetsmått

- Ytliga mått
  - Räknar ord och bokstäver, t.ex. antal ord/mening
- Lexikala mått
  - Baserade på ordfrekvenser och grundläggande svensk vokabulär, t.ex. vardagliga ord (SweVocD)
- Morfosyntaktiska mått
  - Bygger på en morfologisk analys av texten, t.ex. andel innehållsord
- Syntaktiska mått
  - Egenskaper beräknade efter en syntaktisk analys av texten, t.ex. meningsdjup, dependenslängd

Textkomplexitetsmätning

Ordnivå

- Svåra, tvetydliga, långa ord och begrepp
- Andra språk
- Förkortningar

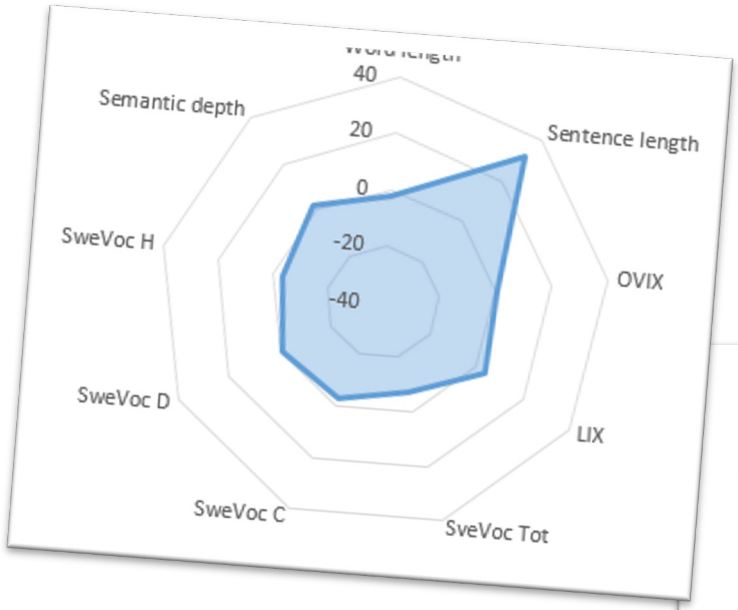
Meningsnivå

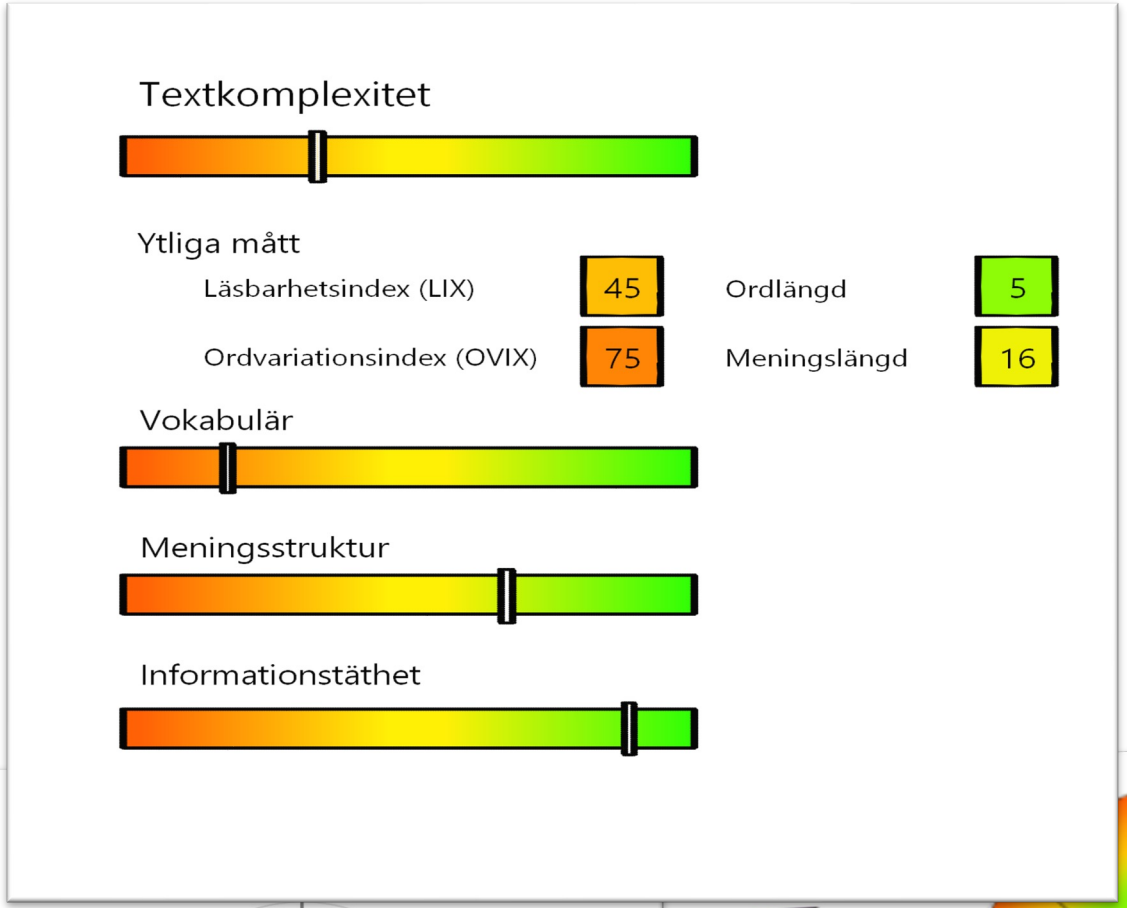
- Långa/Svåra meningar
- Andel bisatser

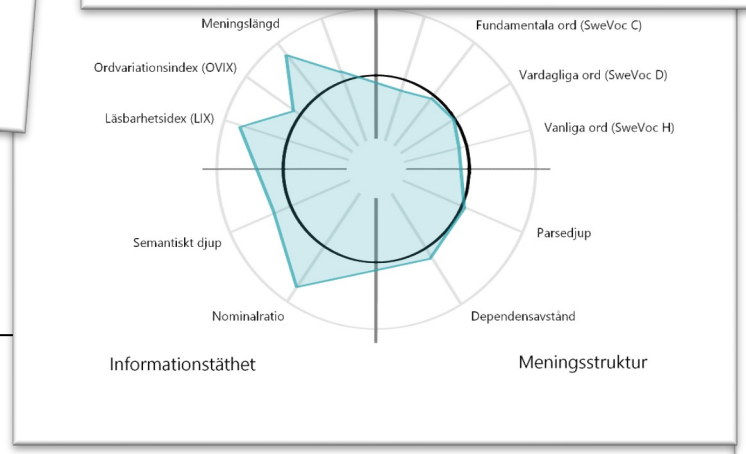
Övergripande

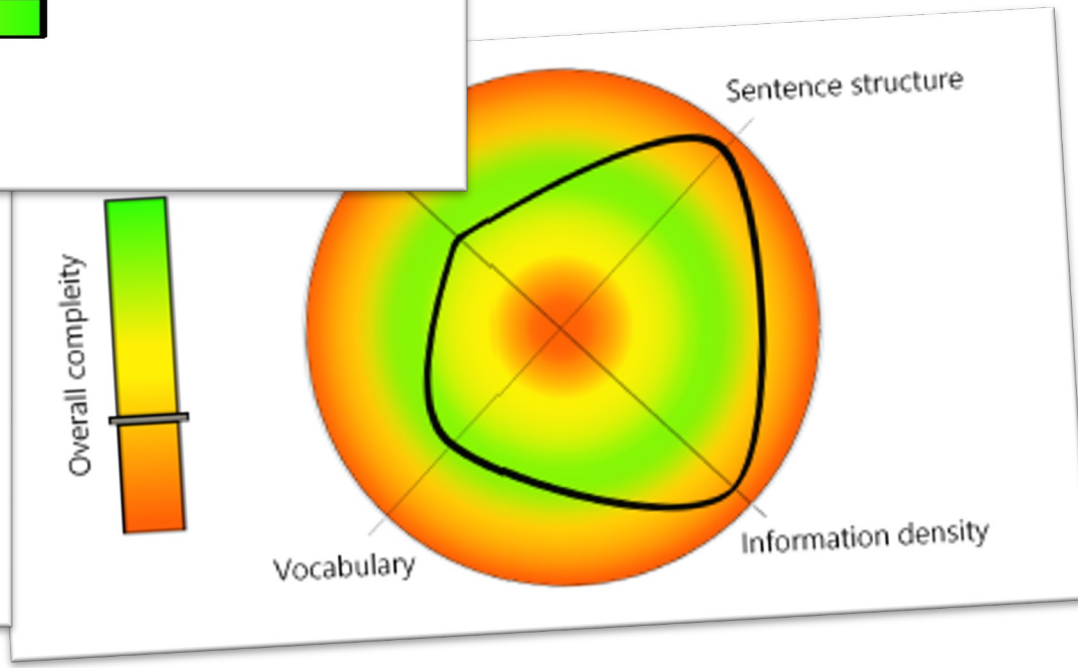
- Längd, innehåll, variation

Visualisering









Utvärdering av visualiseringar

Webbenkät

- Jämförde staplar och radar
- 11 av 26 webbredaktörer svarade

Resultat

- Föredrog stapel för enkelheten
- Radardiagrammet mer informativt
- Kombinera visualiseringarna
  - Stapel för dess användning av färg och färre explicita parametrar
  - Radardiagram ger en mer nyanserad bild, är kompakt och informativt

Tack för idag!

LINKÖPINGS UNIVERSITET