

Forskning om textförenkling och textkomplexitetsmätning

Forskning om läsbarhet och textförenkling

- Verktyg för att göra texter mer tillgängliga
 - Automatisk sammanfattning
 - Omskrivning till lätt svenska
 - Synonymer
 - Automatisk textkomplexitetsmätning
- Resultat
 - 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

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

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

Vår utmaning

- Känn till vem läsaren är!
 - Figurativt språk?
 - Förklara kulturella fenomen?
 - Förklara svåra ord?
 - Dela upp långa meningar?
 - Förenkla grammatiken?
- 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
- Används för att förstå textförenkling
 - Språkliga analyser
 - Bygga modell som förenklar
 - Guldstandard
 - Mäta textkomplexitet

	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%

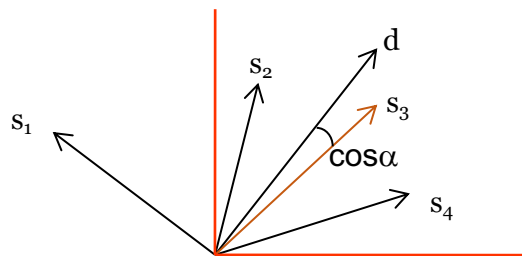
Teknik

- Regler
- Korpusinsamling och träning
 - Egen korpus
 - Alla vanliga och lättlästa texter från offentliga förvaltningar i Sverige, myndigheter, kommuner, landsting m.fl
 - Parallellställda
- Utvärderingar
 - Guldstandarder
 - Mänsklig utvärdering
- Vektorrums- och grafmodeller
 - Maskininlärning

	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

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 meningar och dokument mäts som närhet i vektorrummet



Distribuerad semantik

- Ett ords betydelse beror på hur det används i kontext
- Varje ord representeras som en vektor
 - Några hundra dimensioner
 - Träningskorpus, några miljoner ord
- Exempel, random indexing

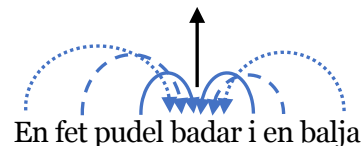
Han bet hunden.

1. Tilldela en gles slumpvektor till varje ord

<i>Han</i>	$(0 -1 0 1 0 -1 0 0 0 1)$
<i>bet</i>	$(0 0 0 0 0 1 1 -1 -1 0)$
<i>hunden</i>	$(-1 0 1 0 0 1 0 -1 0 0)$

2. Varje gång ett ord förekommer i texten, addera kontextvektor i ett fönster runt ordet
bet: $(0 0 0 0 0 1 1 -1 -1 0) + (0 -1 0 1 0 -1 0 0 0 1) + (-1 0 1 0 0 1 0 -1 0 0) = (-1 -1 1 1 0 1 1 -2 -1 1)$

- Träningskorpusar på miljontals ord, och mer



Exempel

- Automatisk sammanfattning
 - Extraktiv – plocka ut de viktigaste meningarna
 - Vektorrumslighet
 - Abstraktiv – skriver om
 - Modell tränad på sammanfattningar

Original

Tillsammans med Kanadas nye stjärna Sidney Crosby duellerade 20-åriga Alexander Ovetjkin om vem som skulle bli årets rookie i NHL. Poängmässigt vann den tuffe och spelskicklige ryssen i Capitals, laget från USA:s huvudstad. På 81 matcher under sin första säsong i NHL gjorde han 52 mål och passade till 54. 106 poäng en rookiesäsong är bland det absolut bästa en nykomling presterat i ligan. Han debuterade redan som 16-åring i den ryska ligan och trots locktoner om feta dollarbuntar från andra sidan av Atlanten stannade han kvar hemma i Moskva fram till och med förra säsongen. Det gjorde också att Ovetjkin var en redan klar elitspelare när han landade i Washington, där han för övrigt gjorde mål redan i sin debutmatch. Washington som lag räckte dock inte alls till för att ta sig till årets Stanley Cup-slutspel. Det är därför Alexander Ovetjkin nu kan komma och visa upp sig i Globen redan i kväll och därefter förstärka det ryska VM-laget i Riga. Det är alla fall var den ryska lagledningen hoppas på. Ovetjkin skulle egentligen ha anslutit till det ryska laget som på torsdagseftermiddagen kom från Helsingfors. Ryssland spelade sin första match i Hockey Games mot Finland i Helsingfors och vann i onsdags kväll. Nu blev Ovetjkin försenad från USA och landar inte i Stockholm förrän på fredagsmorgonen. Alexander Ovetjkin är bara en i raden av unga framgångsrika ryska hockeyspelare som kommer fram just nu. Tillsammans med bland andra Ilja Kovaltjuk, Atlanta och Jevgenij Malkin, Magnitogorsk i ryska ligan, har rysk hockey fått fram stjärnor som kan ta tillbaka landets hockeylandslag till en nivå vi inte sett sedan början av 1990-talet. Med Ovetjkin i laget blir det ännu svårare för Sverige att dels vinna dagens match i Globen och även att ta hem slutsegern i Euro Hockey Tour, som avgörs på måndag.

Extraktiv sammanfattning

- Det är därför Alexander Ovetjkin nu kan komma och visa upp sig i Globen redan i kväll och därefter förstärka det ryska VM-laget i Riga. Alexander Ovetjkin är bara en i raden av unga framgångsrika ryska hockeyspelare som kommer fram just nu. Tillsammans med Kanadas nye stjärna Sidney Crosby duellerade 20-åriga Alexander Ovetjkin om vem som skulle bli årets rookie i NHL. Med Ovetjkin i laget blir det ännu svårare för Sverige att dels vinna dagens match i Globen och även att ta hem slutsegern i Euro Hockey Tour, som avgörs på måndag.

Abstraktiv sammanfattning

- Alexander Ovetjkin är klar för spel i det ryska VM-laget i ishockey. Det innebär att han kommer att förstärka det ryska laget i kvällens match mot Sverige i Globen. Det blir Ovetjkins första match i Moskva och den blir tuffare än vad han hade tänkt sig innan han lämnade Washington Capitals för att spela i ryska ligan. Det är nu ryska hockeyligan, NHL, är på väg att ta upp sporten på den nivå man inte sett sedan början av 1990-talet. I dag möts Sverige och Ryssland i Globen i en match som kan bli avgörande för om Sverige ska vinna Euro Hockey Tour.

Textförenklingar

- Regelbaserat verktyg
 - Dependensgrammatik
- Modell för textförenkling
 - Träna modell

Inom kort kommer bostadsbolaget Bohososs AB troligtvis att höja omkostnader rörande lokalhyra och serviceavgifter. Beslutet kommer tas av bolagsstyrelsen under bolagets årliga sammanträde i maj. En markant stegring av avgifter kan emotes av hyresgästerna om förslaget bifalles. "Hyreshöjningen motsvarar den höjda kvalitén på bostadsservicen", kommenterar en kontaktperson för Bohososs.

Förenklingsgrad

☐ Låg

☐ Medel

☐ Hög

Förenkla

Bostadsbolaget Bohososs AB kommer troligtvis att höja omkostnader rörande lokalhyra och kort.

MOD: [Straight word order]

Beslutet kommer tas av bolagsstyrelsen under bolagets årliga sammanträde i maj.

En markant stegring av avgifter kan emotes av hyresgästerna om förslaget bifalles.

En kontaktperson för Bohososs kommenterar: " Hyreshöjningen motsvarar den höjda kvalitén på bostadsservicen ".
MOD: [Quote inverted]

☒ Anpassad

☒ Passiv till aktiv

☒ Rak ordföljd

☒ Meningsuppdelning

☐ Synonymutbyte

☒ Citatomvändning

☐ Decker set#1

☐ Decker set#2

Textkomplexitetsmätning

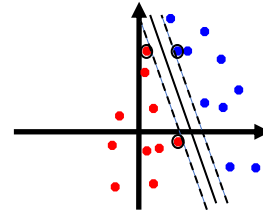
- Ordnivå
 - Svåra, tvetydliga, långa ord och begrepp
 - Andra språk
 - Förkortningar
- Meningsnivå
 - Långa/Svåra meningar
 - Andel bisatser
- Övergripande
 - Längd, innehåll, variation

Textkomplexitetsmått

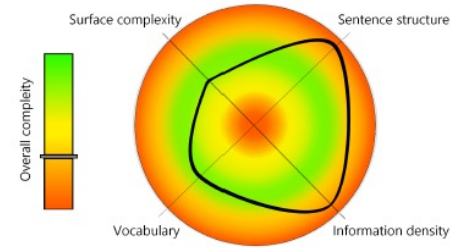
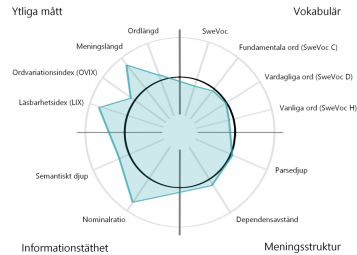
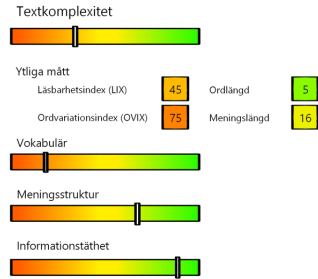
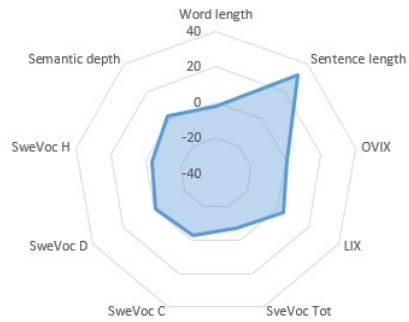
- Ytliga mått
 - Räknar ord och bokstäver, t.ex. antal ord/mening
- Lexikala mått
 - Baserade på ordfrekvenser och grundläggande svensk vokablar, t.ex. vardagliga ord (SweVocD)
- Morfosyntaktiska mått
 - Bygger på en morfologisk analys av texten, t.ex. andel innehållsord
- Syntaktiska mått
 - Egenskaper beräknade efter en syntaktisk analys av texten, t.ex. meningsdjup, dependenslängd
- Textkvalitetsmått
 - Mått som vanligtvis används för att mäta läsbarhet, t.ex. LIX, NR

Läsbarhet

- Avgöra om en text är lätt att läsa eller ej
- Använder textkomplexitetsmått, 117 totalt
- Tränar en modell med de olika måtten
 - SVM (Weka)
 - 7-fold korsvalidering
- Korpus med lättlästa texter och vanliga texter. 700 dokument i varje korpus.
- Testa modellen
 - 98,9% (Stdev 1,0)

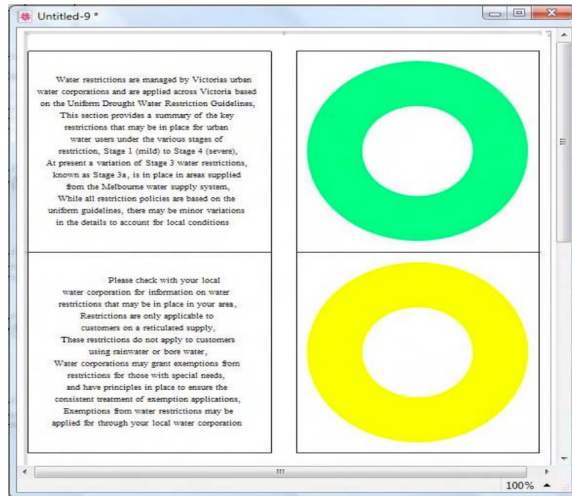


Visualisering

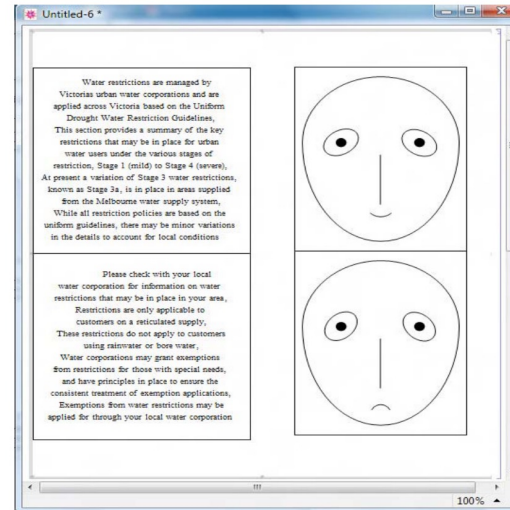


Visualisera flera olika läsbarhetsindex

Fem meningslängdsmått (Flesch-Kincaid, Gunning Fox, Coleman-Liau, SMOG, Automatic readability index)



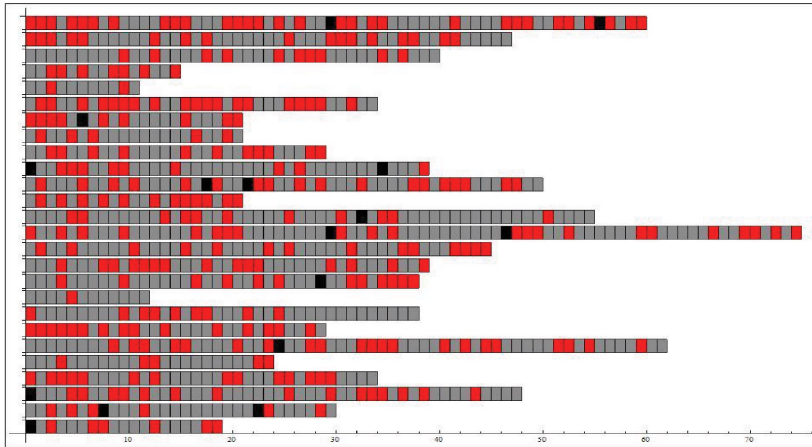
... Viktade och färgkodade



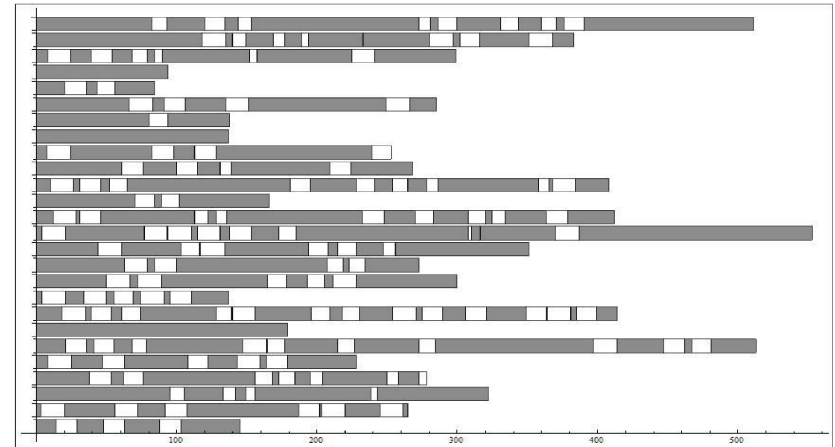
... mappade till ansiktformer

Visualisera hela texter

Varje stapel representerar en mening. Varje del representerar ett ord (vänster) eller en fras (höger).



Lexikal komplexitet från ordlista (Dale-Chall) i grått. Rött betyder att det saknas och svart är icke-ord.



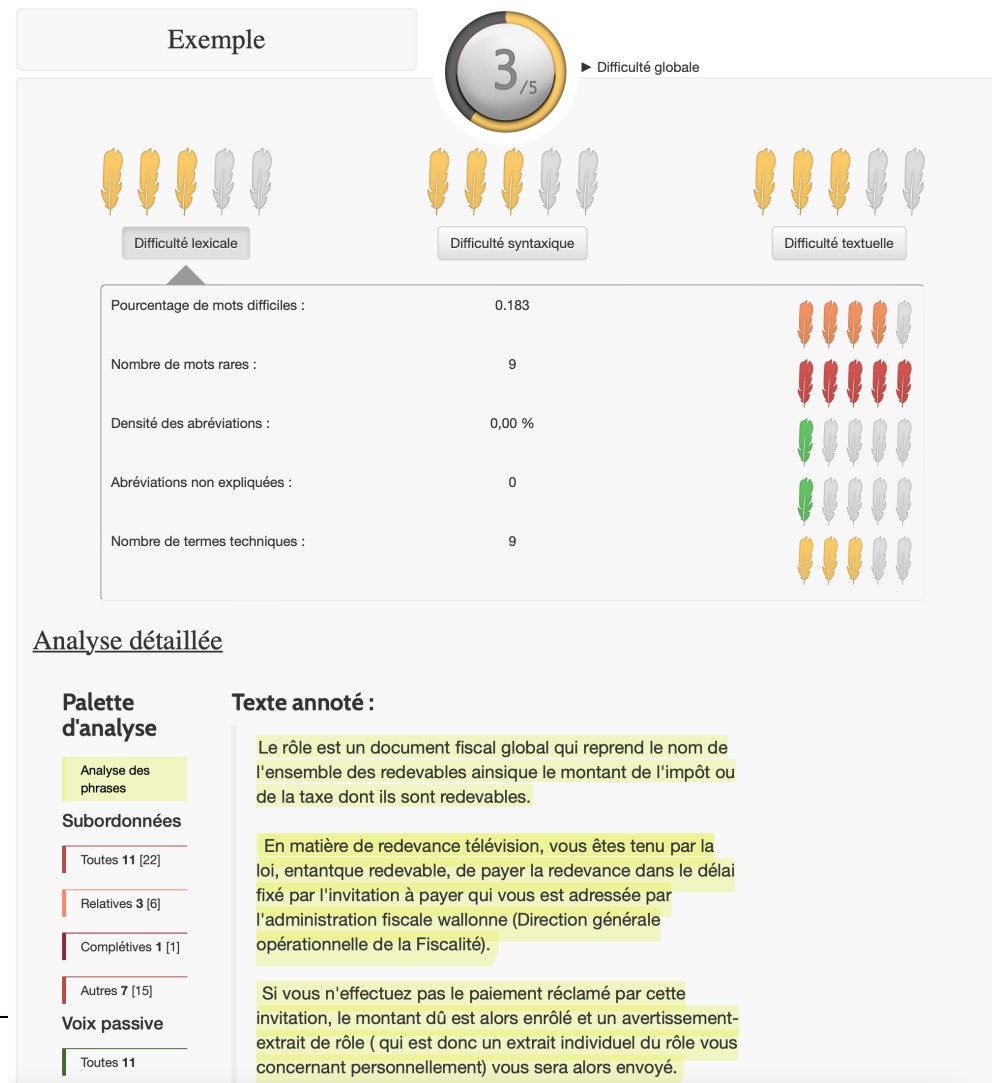
Fraser i grått. Vitt indikerar parsedjup.

Färgkoderna representerar olika läsbarhetsnivåer

liU LINKÖPINGS
UNIVERSITET

AMesure

- Skrivverktyg
- Franska



LX CEFR

- Flesch Reading Ease Index
- Lexikal densitet i proportion till antal substantiv
- Medelordlängd som antal stavelser per ord
- Medelmeningslängd



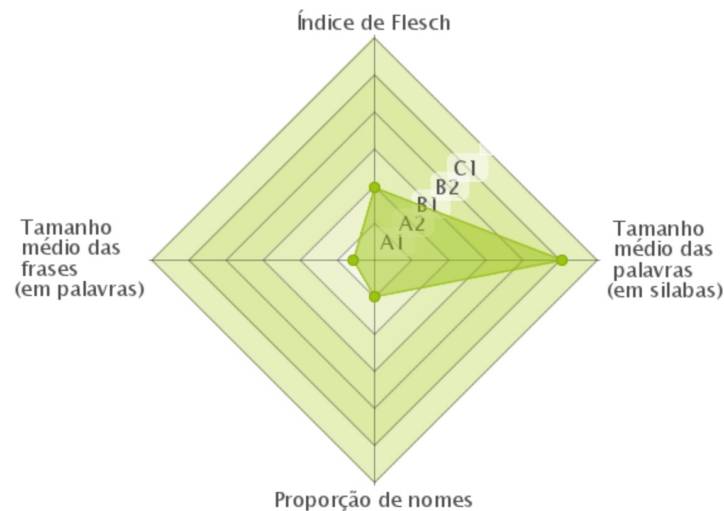
Desenvolvido na [Universidade de Lisboa](#), Departamento de Informática, pelo NLX-Grupo de Fala e Linguagem Natural.

[ver um exemplo](#) | [características](#)

Introduza uma frase:

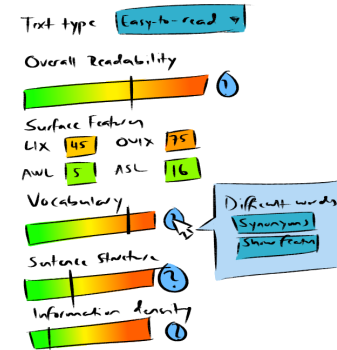
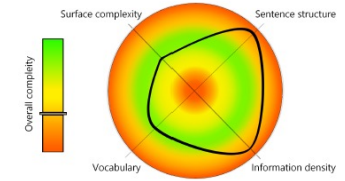
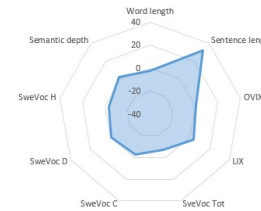
A família de Luciano
Luciano tem oito anos e possui uma família que admira muito. Ele gosta de
desenhar seus familiares quando está de férias.

Sua mãe chama-se Olívia e tem cabelos castanhos e longos. Ela gosta de
assistir novelas e cultivar um jardim de rosas brancas.



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



Vilka mått skall användas

- Workshop med experter
 - 12 stycken bland annat från Språkrådet, Företag, akademien
- Fyra grupper:
 - Ord
 - Meningar
 - Stycken
 - Dokument



Vilka parametrar i varje grupp

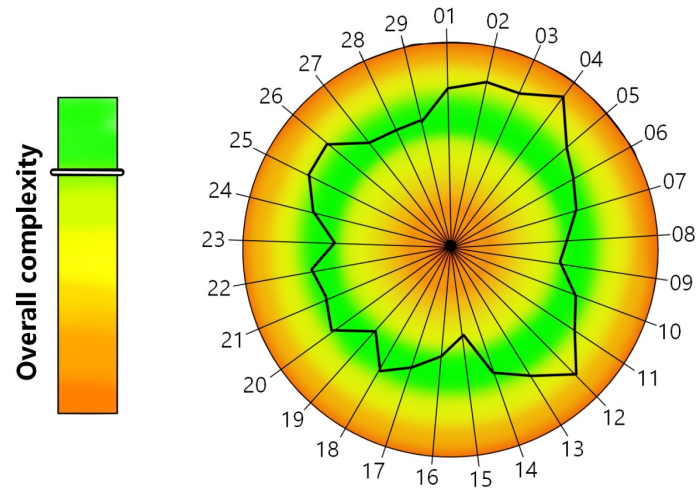
- 117 att välja på
- Datadriven analys
 - LäsBart och SUC
- Förmåga att mäta läsbarhet, prediktion
- Parametrar korrelerar

Val av parametrar

- Faktoranalys (PCA)
 - 29 komponenter
 - Mer än 60% av variansen (61,47%)
 - 85 parametrar utnyttjas
 - Namnge komponenterna?
 - Varje komponent består av flera olika parametrar med olika vikt

Visualisering av komponenterna

- Stapel
 - Övergripande komplexitet
- Radardiagram
 - Mönster



Utvärdering av textvisualiseringarna

- Tre semistrukturerade fokusgrupper med speciallärare
- Öppna frågor om textkomplexitet och två radardiagram
- Resultat
 - Vet inte alltid vad studenterna har svårt med
 - Intuition gällande textkomplexitet
 - Utmaning att förklara vilka aspekter som gör en text komplex
 - Visualisering
 - Meningslängd, ordlängd, antal långa ord ganska enkelt
 - Meningsdjup, pre | post-modifierare, anaforer, SweVoc-ordlistor svårare att förstå

