SweLL -

Forskningsinfrastruktur för svenska som andraspråk

Elena Volodina
Göteborgs universitet, Språkbanken
elena.volodina@svenska.gu.se

Själv-presentation: Elena Volodina

https://spraakbanken.gu.se/eng/personal/elena

- 1998 PhD i Lingvistik (Moskva, Ryssland)
- 2008 MA i Språkteknologi (Göteborg, Sverige)



"Teachers never tell you their first names because they don't want you to *Google* them"

- 2010 - ... Forskningsingenjör Språkbanken, UGOT
  - Lärka utveckling
  - ICALL forskning
  - Svenska som andraspråk (L2)-resurser och algoritmer
  - Infrastruktur för L2 svenska
  - ...

Några förkortningar

SweLL	**Swe**dish **L**earner **L**anguage
L2	Andraspråk, samt främmande språk
PIAAC	Programme for the International Assessment of Adult Competencies
CEFR aka **GERS**	Common European Framework of Reference for Languages Gemensam Europeiska Referensramen för Språk

Fokus på läskunnighet

- Nederländsk studie:

  → Genomsnittlig läskompetens ~B1 nivå

Velleman, E., van der Geest, T.: Online test tool to determine the CEFR reading comprehension level of text. Procedia Computer Science 27 (2014)



Kan inte ... *Kan det mesta*



Läskunnighet: Sverige

- PIAAC studie (2013) med fokus på läskunnighet
- Sverige bland 5 “**bästa**” av 23 länder i genomsnitt
- **Störst** klyfta mellan infödda och icke-infödda invånare

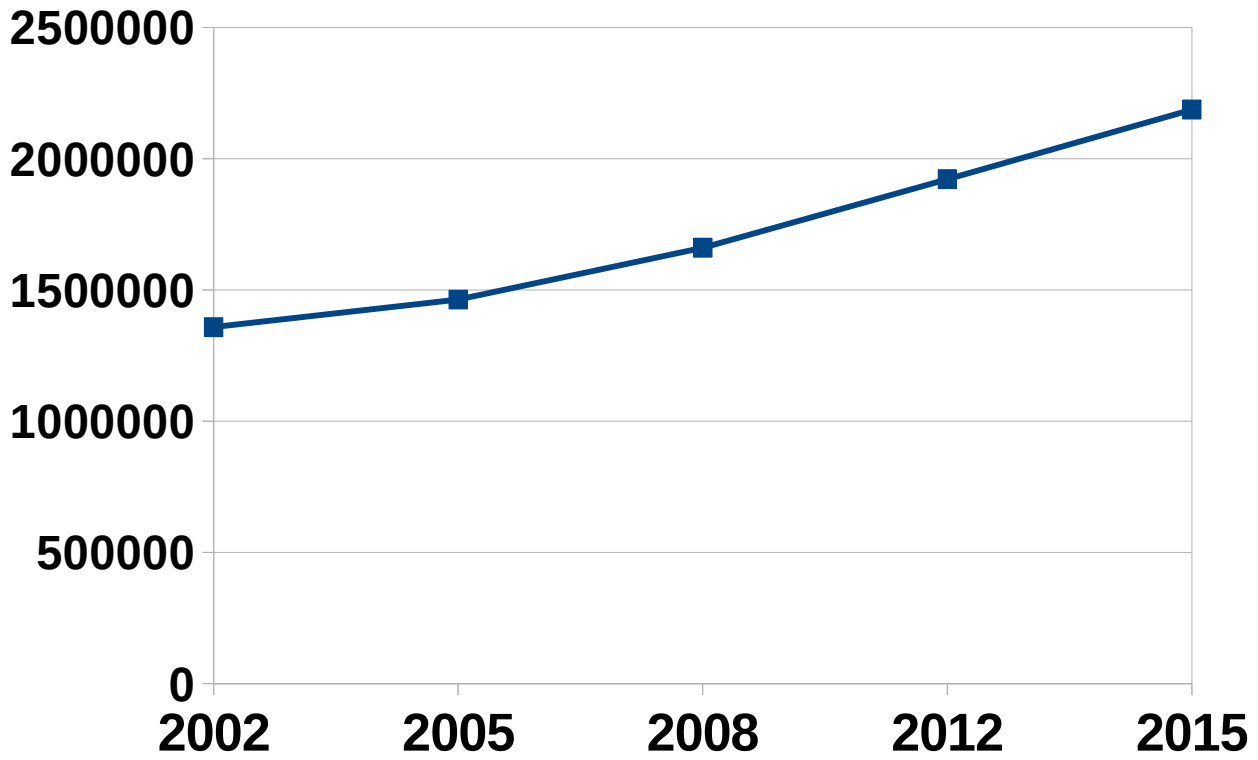
  → låg anställningsbarhet

  → större risk för dålig hälsa

OECD. 2013. OECD Skills Outlook 2013. First Results from the Survey of Adult Skills.
PIAAC. 2013. Survey of Adult Skills (PIAAC).
SCB. 2013. Tema utbildning, rapport 2013:2, Den internationella undersökningen av vuxnas färdigheter. Statistiska centralbyrån.

Samhällsbehov

Medborgare med utländsk bakgrund, Sverige, 2002-20015



2015: av **9,9 mln** invånare, **2,2 mln** har utländsk bakgrund, dvs **22,2 %**
(Statistiska centralbyrån)

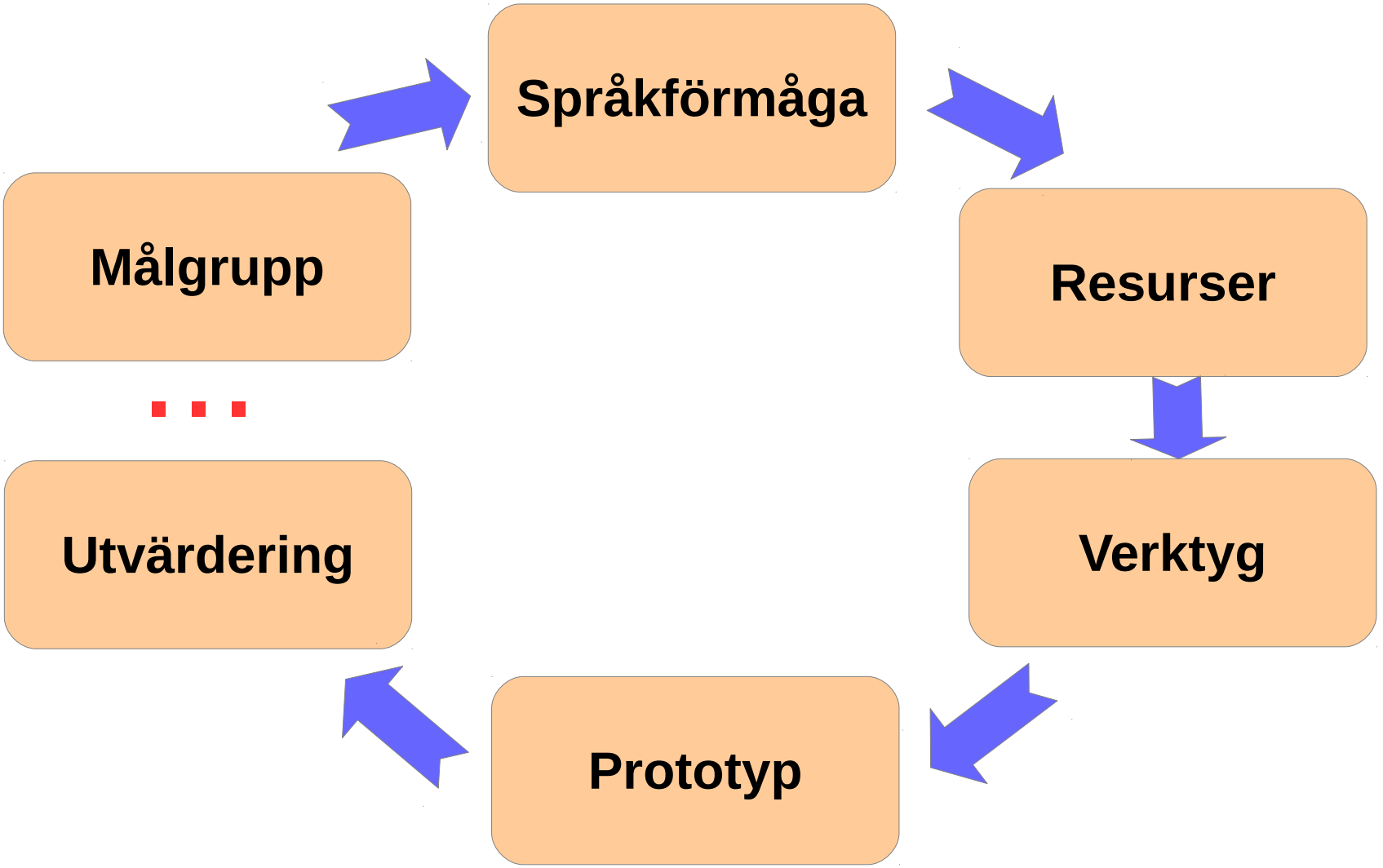
Vad kan vi göra (med NLP)?

orsak- och symtombehandling





Datorverktyg för L2 språkinlärning



Datorverktyg för L2 språkinlärning



Målgrupp

Vuxna

Barn

Analfabeter

Dyslektiker

osv



Datorverktyg för L2 språkinlärning



Språkförmåga

Tala

Skriva

Höra

Läsa

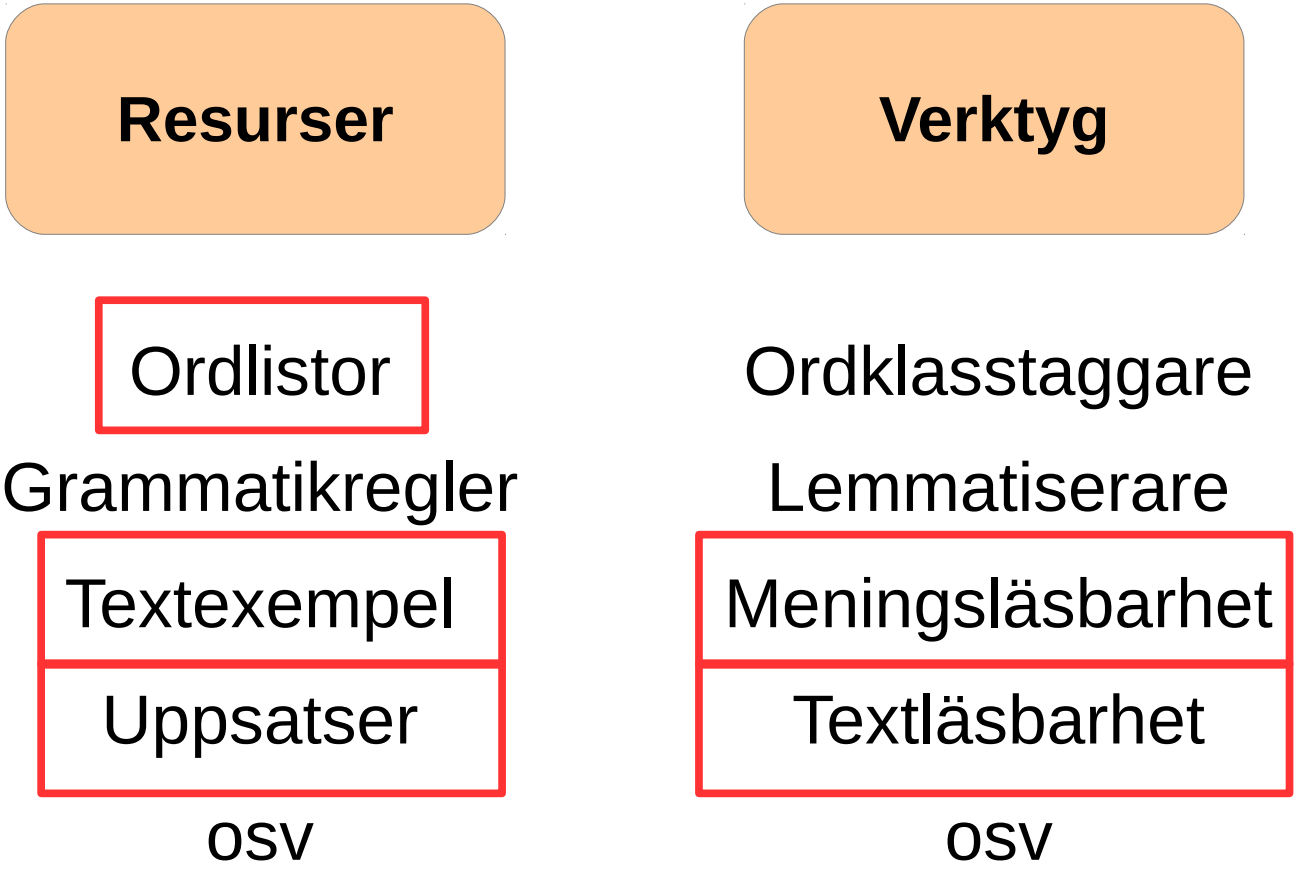
Uttala

Ordförråd

Grammatik



Datorverktyg för L2 språkinlärning



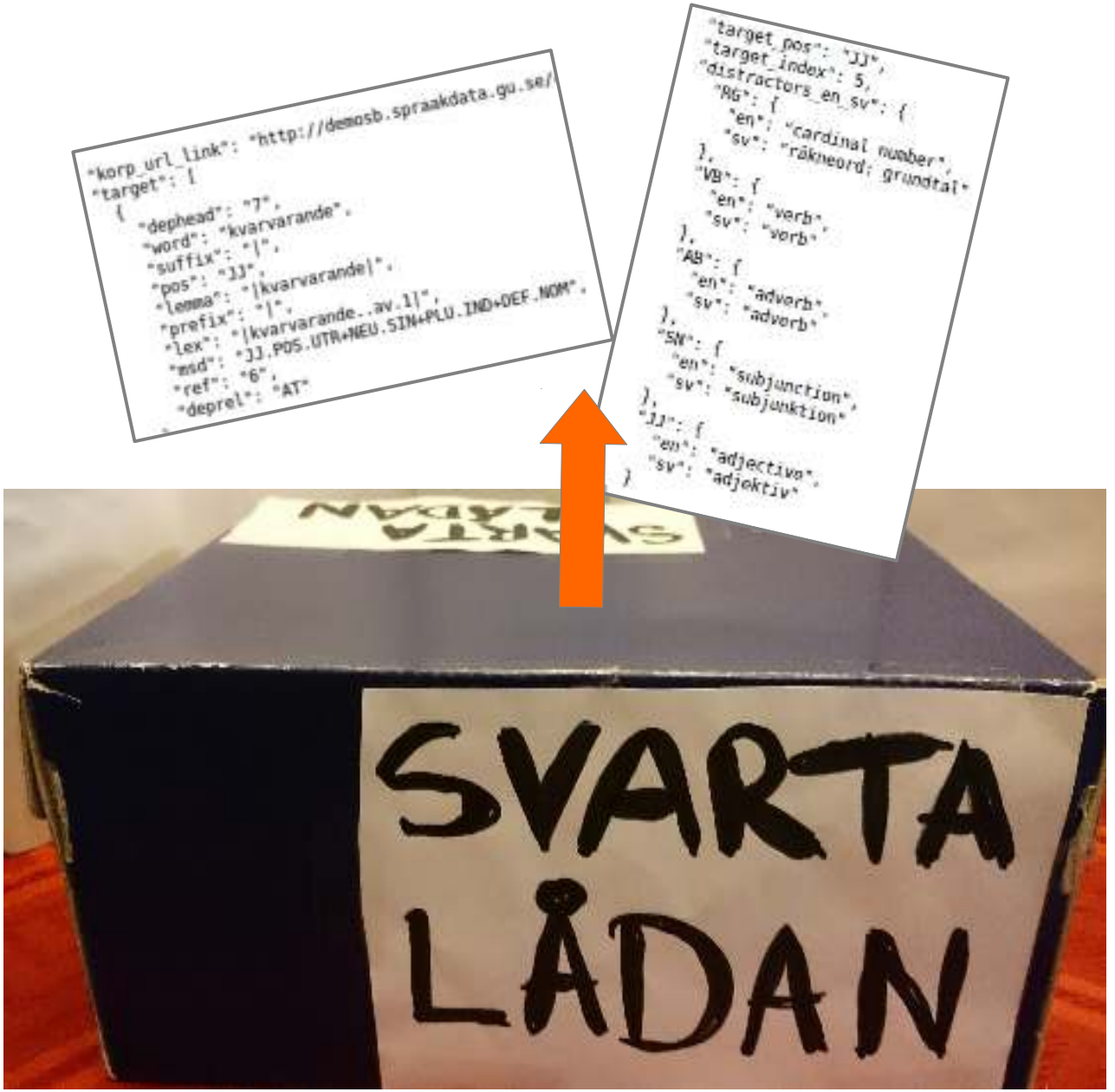
Datorverktyg för L2 språkinlärning

Resurser

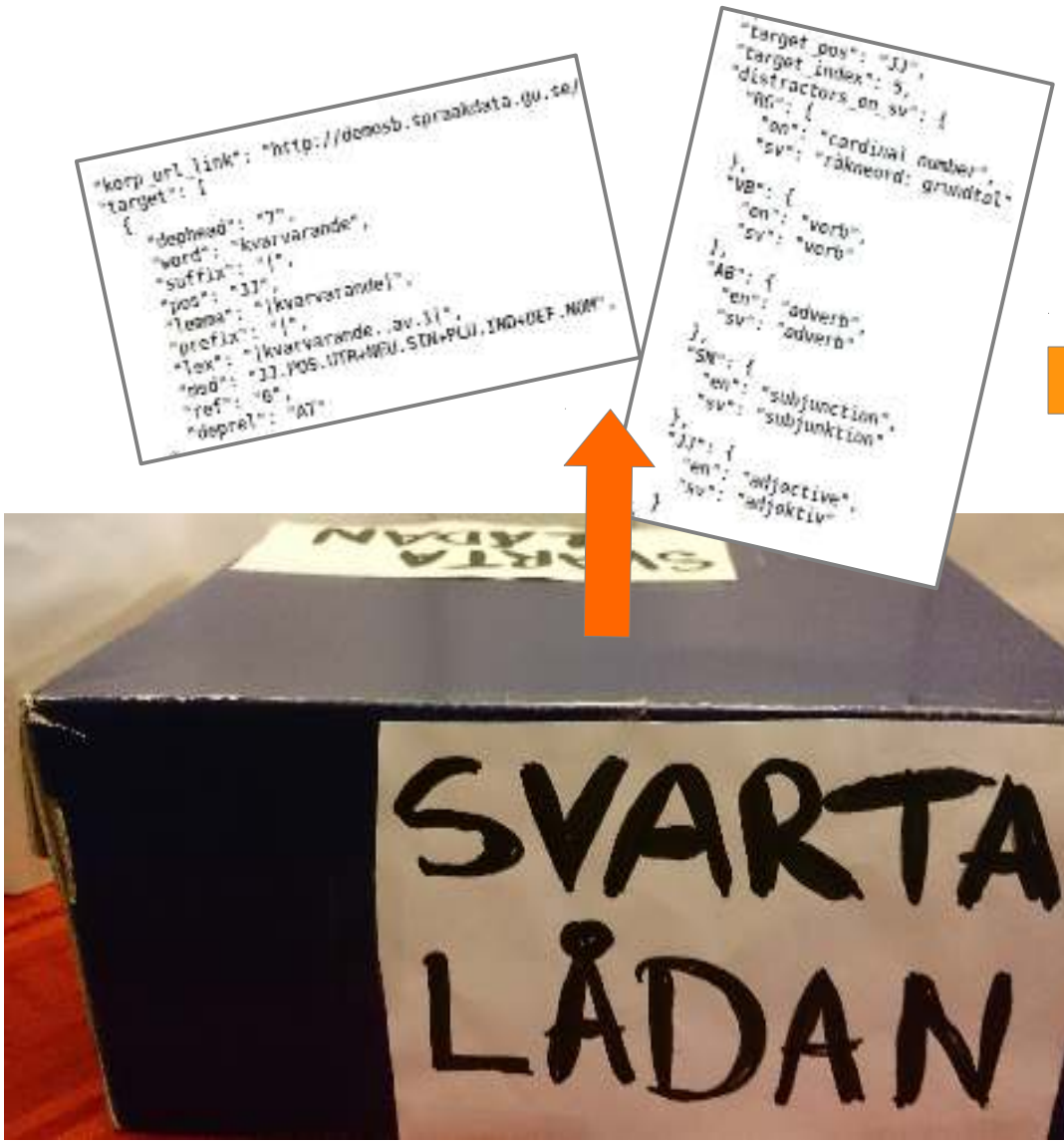
Verktyg



Datorverktyg för L2 språkinlärning



Datorverktyg för L2 språkinlärning



Datorverktyg för L2 språkinlärning

Applikation-
utveckling och underhåll

versus

Prototyp-
utveckling och utvärdering



Datorverktyg för L2 språkinlärning

Målgrupp

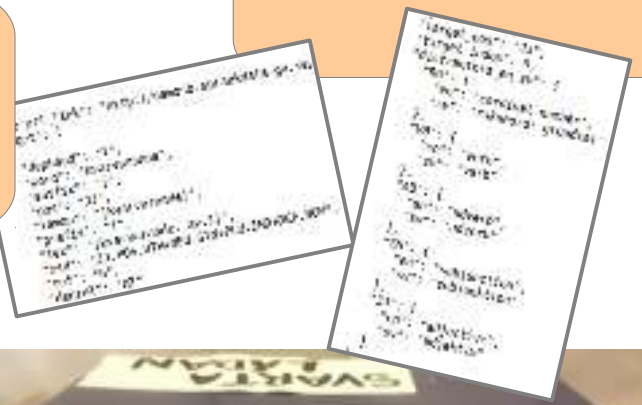
Språkförmågor



Prototyp
+ utvärdering

Resurser

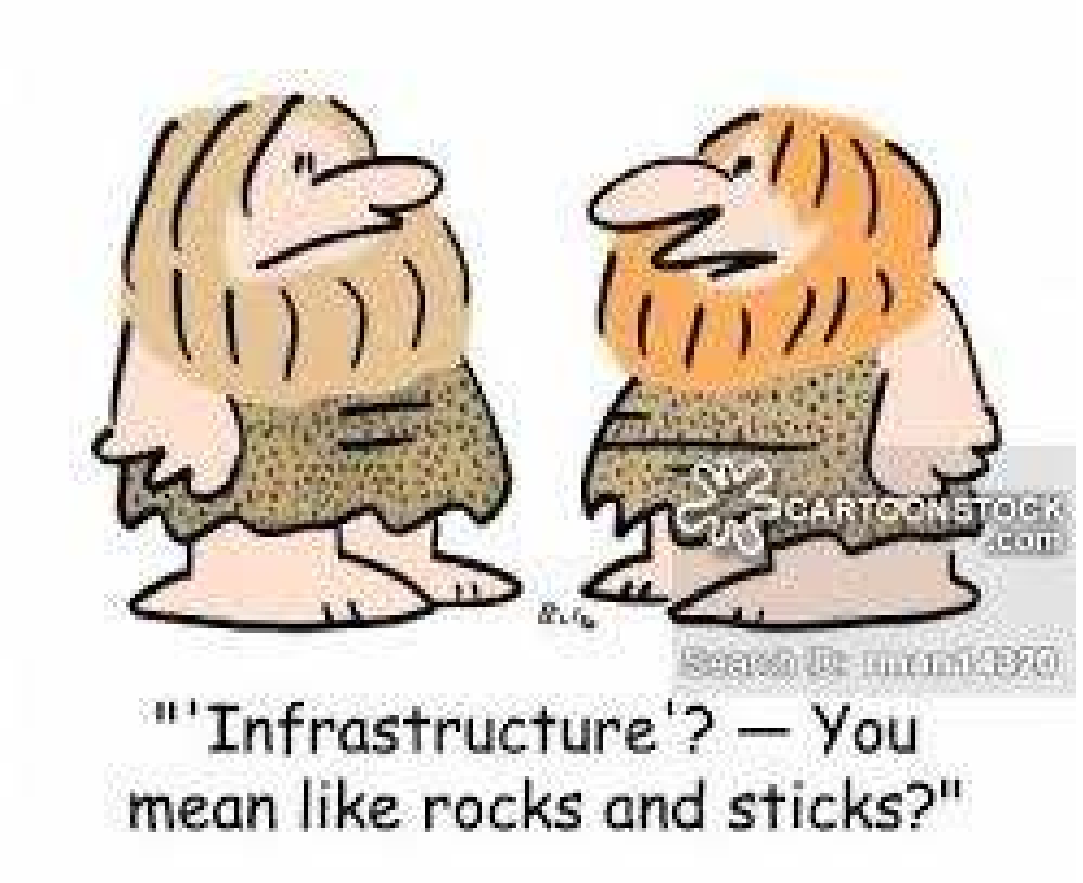
Verktyg







Vad är infrastruktur?



En elektronisk forskningsinfrastruktur

- (fri tillgänglig) data i elektronisk format
- teknisk plattform för att utforska data, inkl. verktyg och algoritmer för analysen och visualisering av datan
- en verktygslåda med tekniska lösningar för insamling och bearbetning av ny data (inkl bl.a. automatisk annotering)
- ett nätverk med experter inom relevanta områden, inkl. bl.a. juridiska och etiska frågor



Hur kan vi hjälpa till?

- Samla och annotera data (L2 uppsatser, felloggar, ...)
- Utveckla verktyg för L2 dataanalys (e.g uppsatser)
- Utveckla expertkunskap för att kunna

  - ➔ stödja forskning inom L2 svenska
  - ➔ stödja L2 kursboksförfattare, L2 lärare, L2 bedömmare
  - ➔ stödja undervisning av blivande L2 lärare
  - ➔...och genom detta stödja undervisning av L2 studenter



Partners

- Göteborgs universitet: NLP, L2, språkbedömning
  Elena Volodina, Julia Prentice, Monica Reichenberg
- Stockholm universitet: NLP, L2
  Mats Wirén, Gunlög Sundberg
- Uppsala universitet: NLP
  Beata Megyesi
- Umeå universitet: L2/bedömning
  Lena Granstedt



Guess what?



- Riksbankens Jubileumsfond, infrastruktur projektet IN16-0464:1
- 7 mln



- 2017-2019

Vår primära fokus

- L2 uppsatser (skriftspråk)
- Loggar från online övningar (läsförsåelse, hörförståelse, vokabulär och grammatik)

- INGEN taldata – än så länge

- målgrupp: vuxna L2 inlärare

Våra löfte

För att tillgodose (bl.a. dessa) behov syftar SweLL till att skapa en infrastruktur som består av:

- en portal för datainsamling, både för import från filer och via online övningar
- verktyg för analys av inlärarspråk
- en L2-korpus på ca 600 texter annoterade med bl.a. fel
- sökmöjligheter för L2-material, såsom sökningar på behärskningsnivåer och grammatiska strukturer

- Materialet och verktyg kommer att tillgängliggöras genom Språkbanken.

Problem 1: avsaknad av L2 data

- Digitaliserad L2 produktion är svårt att samla
  - kan INTE laddas från webben
  - behov av medföljande tillstånd
  - behov att ta reda på inlärarvariabler (kön, ålder, modersmål, osv.)
- Vi behöver en infrastruktur för insamling av
  - L2 produktion
  - loggar över fel
  - osv

Problem 2:
avsaknad av koordinering av insatser

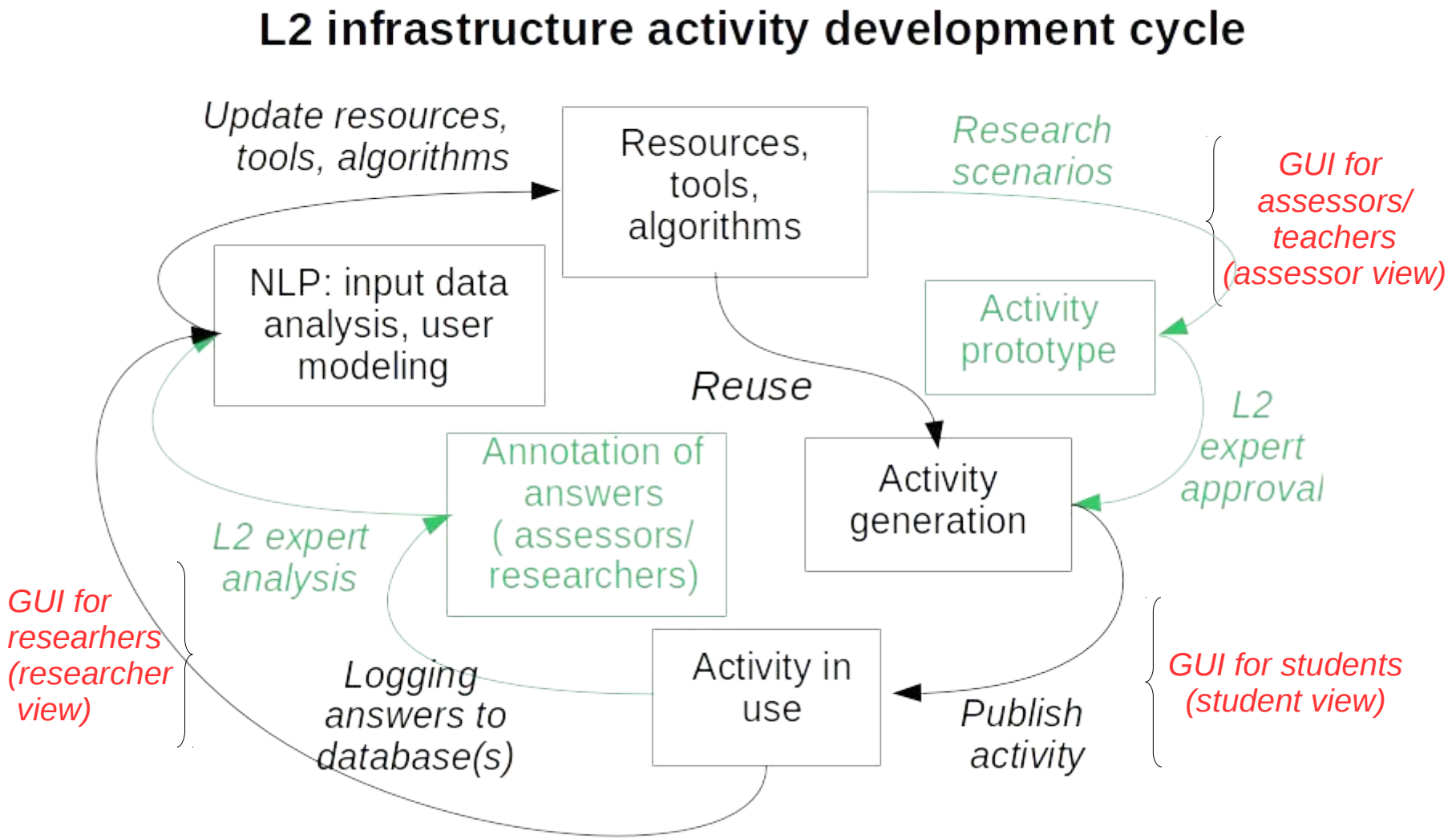
- Typ av variabler, tillstånd, dataformat bör koordineras mellan flera intresserade parter

- Digitalisering samt annotering av stora arkiv med elevproduktion (t.ex. Nationella prov i svenka och svenska som andraspråk)

Problem 3:
avsaknad av metoder/algoritmer för bearbetning av inlärardata

- NLP metoder/algoritmer ej anpassade att hantera avvikelser i L2 språk
- Felannotering av L2 uppsatser

Det ultimata målet



Natural Language Processing för språkinlärning

Nya aktiviteter/övningar för studenter och lärare

SUPER-nya resurser och algoritmer

Stöd till utveckling av språkfärdigheter för den nya millenium

Loggning av resultat och individanpassning



Skrivstöd (verktyg)

Upsatsbedömning

Feldetektering

Automatisk återkoppling

You name it....

Första steg och pilotstudier

- **Data:** insamling och digitalisering
  - SweLL korpus
  - The Uppsala Corpus of Student Writings
- **Resurser:** t.ex. SweLLex – L2 produktivt lexikon
- **Algoritmer**: t.ex. L2 felnormalizering
- **Användaranpassade tools**:
  - L2 annoterings pipeline: SweGRAM
  - Klassificering av L2 uppsatser (Lärka-baserad online verktyg)



Data



SweLL – en korpus med L2 uppsatser

- Från A1 till C1, totalt 339 digitaliserade
- Lingvistikt annoterade texter

- Pågående studier:

✔ skapa ordlista med aktiv vokabulär (SweLLex)

✔ skapa en metod att länka nya uppsatser till uppnådd CEFR nivå

- Framöver:

✔ Annotera fel

✔ Aktiv grammatik per nivå

SweLL korpus

core data

SweLL workflow

Essay collection *ongoing* → **CEFR-grading** *teachers* → **Digitization** → **Manual registration** *editor in Lärka* → **Automatic annotation** *Korp pipeline*

Learner variables
Collected through permits*

Student variables:
- Age/birthyear
- Gender
- Mother tongue(s)
- Residence time in Sweden
- Education level

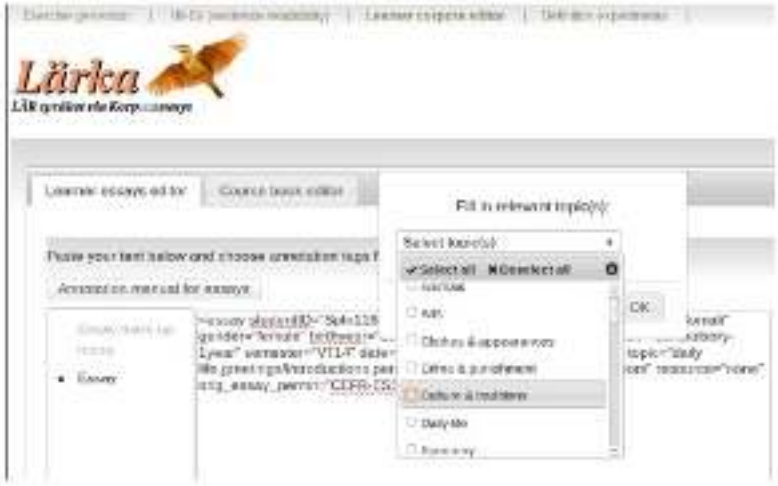
Essay variables
- Assigned CEFR level
- Essay setting (exam/home)
- Use of extra materials
- Academic term and date
- (Title, topic, genre, grade)

Assessors
Minimum of two trained assessors

Inter-annotator agreement
- A degree to which several annotators agree about assigning attributes
- Reported for SW1203 subcorpus
- Krippendorff's alpha for pairwise agreement = 0.80
- **0.80 = good annotation quality** (Artstein & Poesio 2008)

SweLL digitization principles

1. Do not reveal author identity
   * revealing names → replace with NN
   * addresses → replace with NN-street
2. Do not correct errors
   * if several interpretations possible – make positive assumption i.e. that the learner made no mistake
3. Preserve illegible handwriting
   * [illegible] → replace with [illegible]
   * [illegible text] → [illegible text]



* http://spraakbanken.gu.se/sites/spraakbanken.gu.se/files/tillstand eng-24042013 v03.pdf

SweLL korpus

core data

Sub-corpus	A1	A2	B1	B2	C1	Un-known	Total
Tisus	-	-	-	27	78	-	105
Sw1203	-	-	33	45	11	1	90
SpIn	16	83	42	2	-	1	144
Total	16	83	75	74	89	2	**339**

SweLL korpus

core data

Språk-
BANKEN



GÖTEBORGS
UNIVERSITET

Tillstånd

till användning av inläraruppsatser för forskningsändamål

Jag, (namn, efternamn) _____

tillåter Göteborgs universitet, Språkbanken, att använda mina uppsatser för forskningsändamål enligt följande.

Jag tillåter
- ☐ begränsad användning av godkända användargrupper med lösenordsskydd
- ☐ obegränsad användning förutsatt att min identitet förblir anonym

SweLL korpus

core data

Personuppgifter:

Kön: ☐ Kvinna ☐ Man Ålder_____

Modersmål (ett eller flera)_____

Vistelsetid i Sverige: _____år _____ månader

Utbildningsnivå: ☐ Grundskola antal år _____

☐ Högstadiet antal år _____

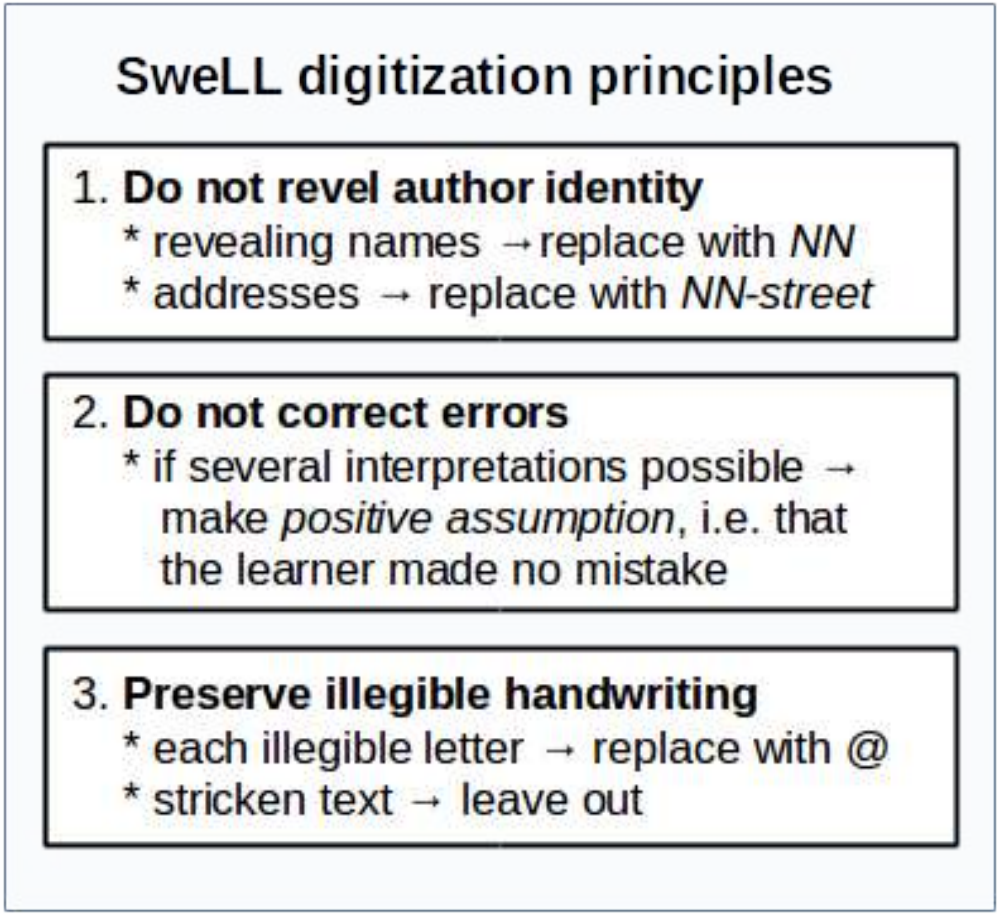
☐ Gymnasiet antal år _____

☐ Högskola/unversitet antal år _____

☐ Forskarutbildning antal år _____

Ort och datum Namnunderskrift:

Digitalisering



The Uppsala Corpus of Student Writings

referenskorpus (monitor corpus)

Level	Age	School level and curriculum	Number of essays	Number of tokens	Tokens per essay
C-3	9	Compulsory, Lpf94 + Lgy11	91	8,644	95
C-5	11	Compulsory, Lpf94	66	13,121	199
C-6	12	Compulsory, Lgr11	47	17,741	377
C-9	15	Compulsory, Lgr94 + Lgr11	249	137,689	553
US-1	16	Upper Secondary, Lgy11	131	76,521	584
US-3	18	Upper Secondary, Lgy11	410	347,836	848
GY-3	18	Upper Secondary, Lpf94	1,506	1,055,468	701
Total			2,500	1,657,020	663

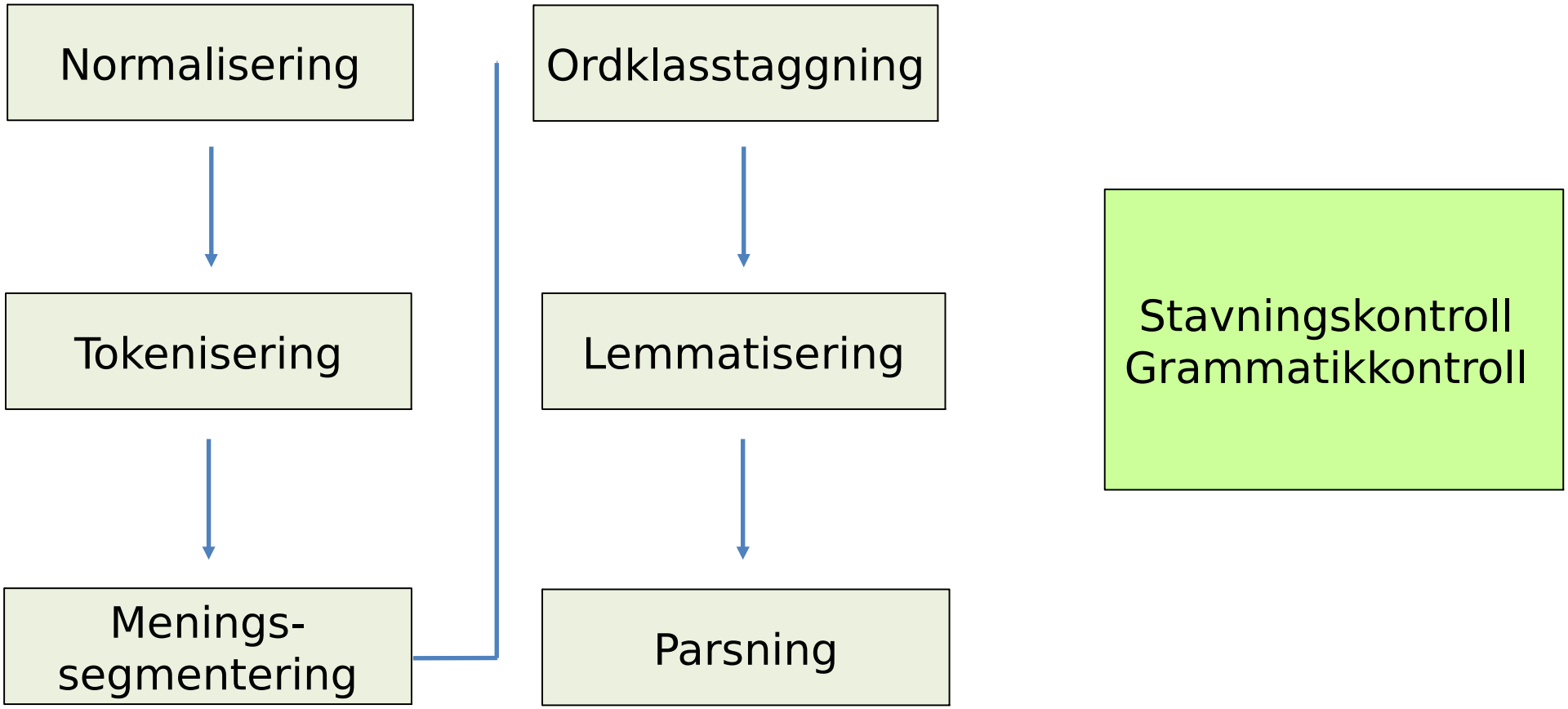
Table 1: Distribution of the subset of texts by school year, given as number of texts, sentences and tokens, and average number of tokens per essay used in the pilot study.

Handwritten essays	Printed essays
Transcription	Scanning-conversion-editing
Coding	Coding
Proofreading and final editing	Proofreading and final editing

Table 2: Preparation of the essays.

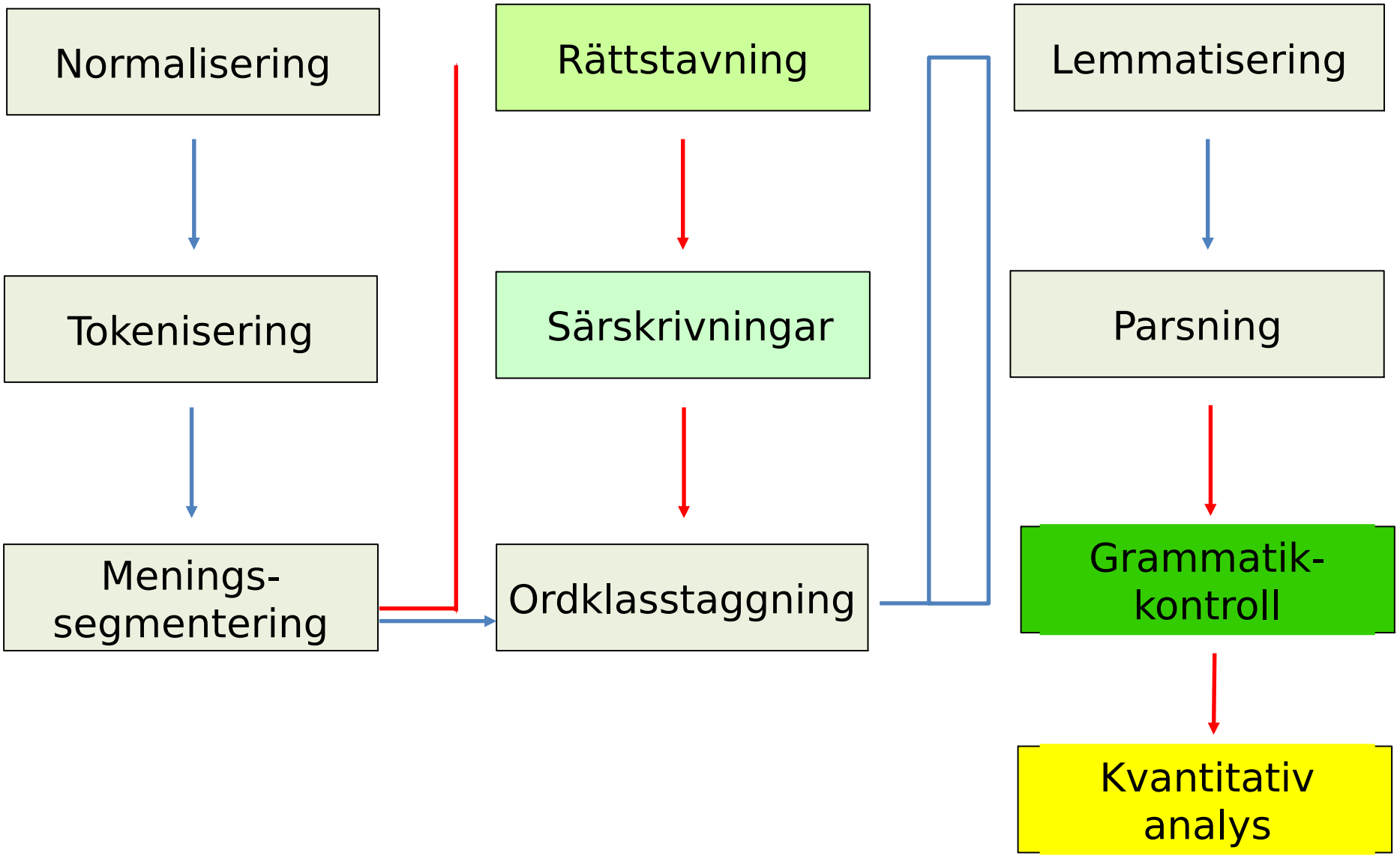
The Uppsala Corpus of Student Writings

referenskorpus



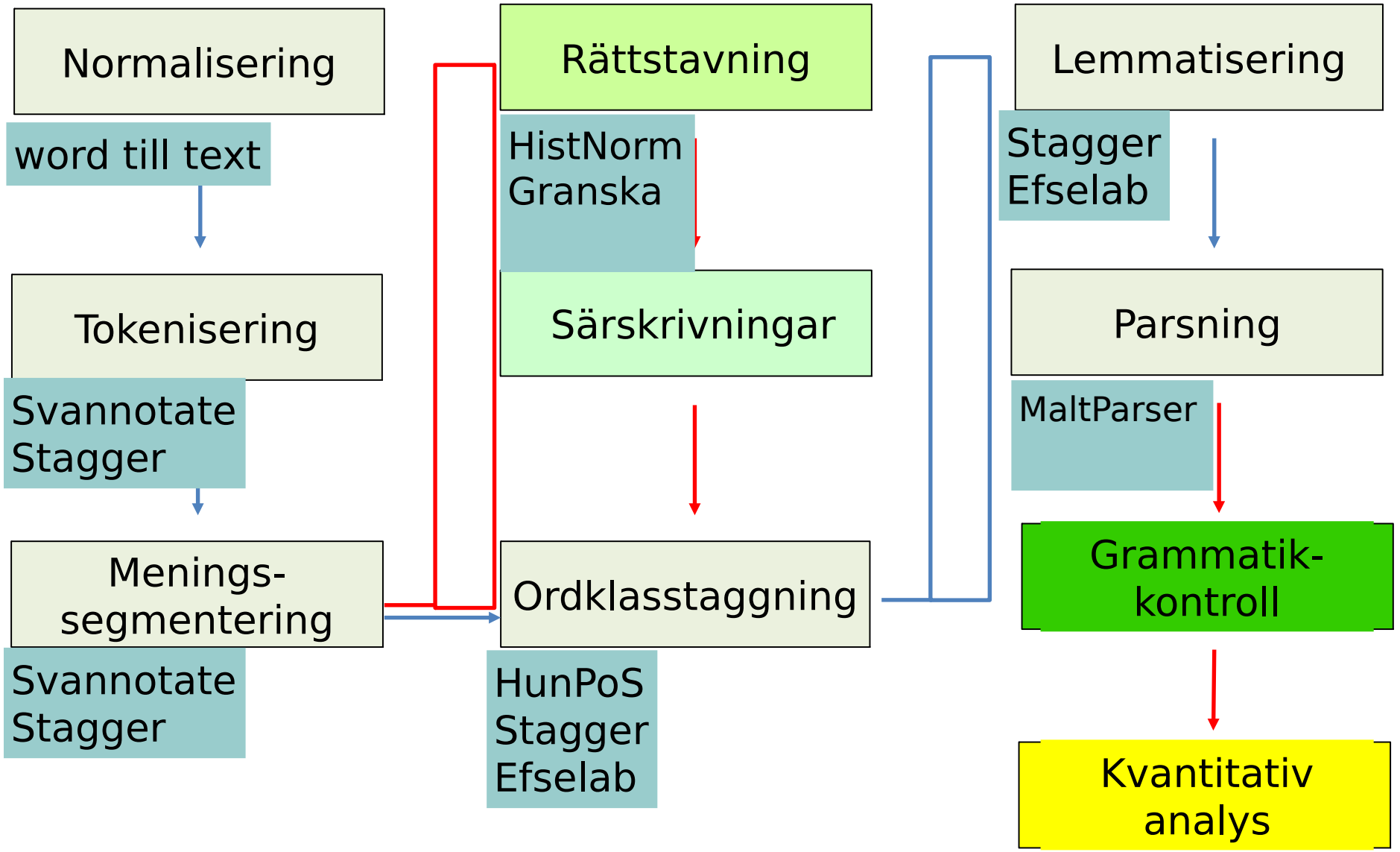
The Uppsala Corpus of Student Writings

referenskorpus



The Uppsala Corpus of Student Writings

referenskorpus



Nästa steg för vår *core* korpus

- Manuell felannotering

- Normalisering, dvs omskrivning till en “förståelig” version, vilket medför bl.a. en del av hypotisering av vad en inlärare menar

Resurser





SweLLex
L2 produktivt ordförråd

Lev	#items	#new	#MWE	#hapax	Doc.hapax examples	#SVALex	#EVP
A1	398	398	15	0	-	1,157	601
A2	1,327	1,038	82	12	*i kväll* "tonight"	2,432	925
B1	2,380	1,542	206	36	*fylla år* "have birthday"	4,332	1,429
B2	2,396	959	264	58	*fatta beslut* "make a decision"	4,553	1,711
C1	3,566	1,545	430	152	*sätta fingret* "put a finger on sth"	3,160	N/A
C2	145	7	12	1	*i bakhuvudet* "in mind"	N/A	N/A

Total 5,475

http://vocabulary.englishprofile.org/staticfiles/about.html

user: englishprofile
password: vocabulary

GÖTEBORGS UNIVERSITET

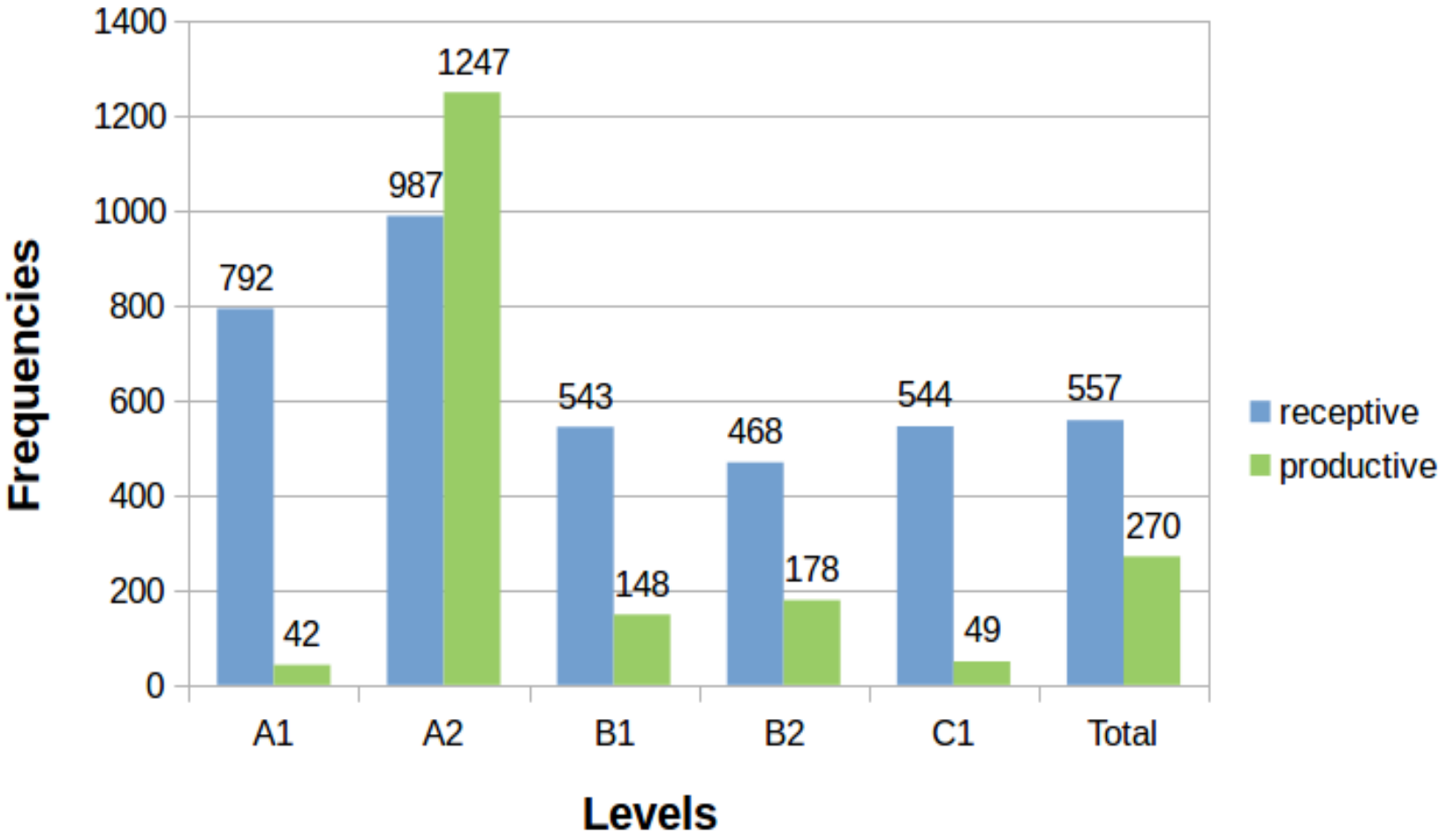


SVALex - receptivt ordförråd

Level	# items	# new items	# MWE	# doc.hapax	Doc.hapax examples	# EVP
A1	1,157	1,157	92	99	*postnummer* "zip code"	601
A2	3,327	2,432	300	635	*jurist* "lawyer"	925
B1	6,554	4,332	617	1,868	*öga mot öga* "face to face"	1,429
B2	8,728	4,553	880	3,051	*snigelfart* "snail speed"	1,711
C1	7,564	3,160	783	2,709	*inom synhåll* "within eyesight"	N/A
Total	15,681	15,681	1,426	8,362		

http://cental.uclouvain.be/svalex/

SVALex vs SweLLex, *att studera*



http://cental.uclouvain.be/svalex/

Algoritmer


a.reverse(); b = m(a, " "); -1 < b
m(a, void 0); -1 < b && a.splice(b, 1); return a;
a.replace(RegExp(",", "g"), " ");
&& c++; } return c; } function
-1, d = 0;d < a.length;d++)
break; } } return c; } function
"-" === a[0]

L2 normalisering på ordnivå



- **Levenstein distance (as is)**
  - Funkar bra på avancerade nivåer (editeringsavstånd 1)
  - Misslyckas på lägre nivåer (vid flera editeringar)
- **Levenstein distance (för historiska texter)**



- **LanguageTool + kandidatrankning**
  - 73% korrekt vid urvalet av passande alternativ
  - Misslyckas att identifiera approx 30% stavfel

HitEx – *“hitta exempel”*

- En algoritm för urval av meningar från korpusar/texter som kan användas för övningar (samling av L2 övningsloggar)
- Viktiga drag:
  - 5 CEFR nivåer
  - Kontextoberoende
  - Möjlighet att filtrera meningar med “känsliga” ord
  - Pedagogisk filtrering för ordförråd (baserad på SVALex)
  - Representativitet / Typiskhet (baserad på MI värden from Korp) osv.

Analys av L2 texter

- **Nivåbedömning (CEFR)**
  - för L2 uppsatser & läsförståelsetexter (lästexter)
  - CEFR nivåer (A1-C1) + extra indikatorer (e.g. LIX)



Nästa steg för utveckling av verktyg

- Normalisering av uppsatser på frasnivån, osv.
- Feldetektering (behöver att identifiera vilka fel vi ska satsa på, samt felannotera ett urval av datan)
- ...

User-oriented tools



SWE-CLARIN

SWEGRAM

UPPSALA UNIVERSITET

SWEGRAM syftar till att ge möjlighet att märka upp och analysera texter på svenska. Du kan ladda upp en eller flera texter och annotera dessa på olika lingvistiska nivåer med morfologisk och syntaktisk information. Annoterade texter kan sedan användas för att få fram statistik om texterna med avseende på meningslängd, antal ord, läsbarhetsmått, ordklasstatistik med mera.



Annotera text

Tokenisera, normalisera, ordklasstagga och annotera syntaktiskt.

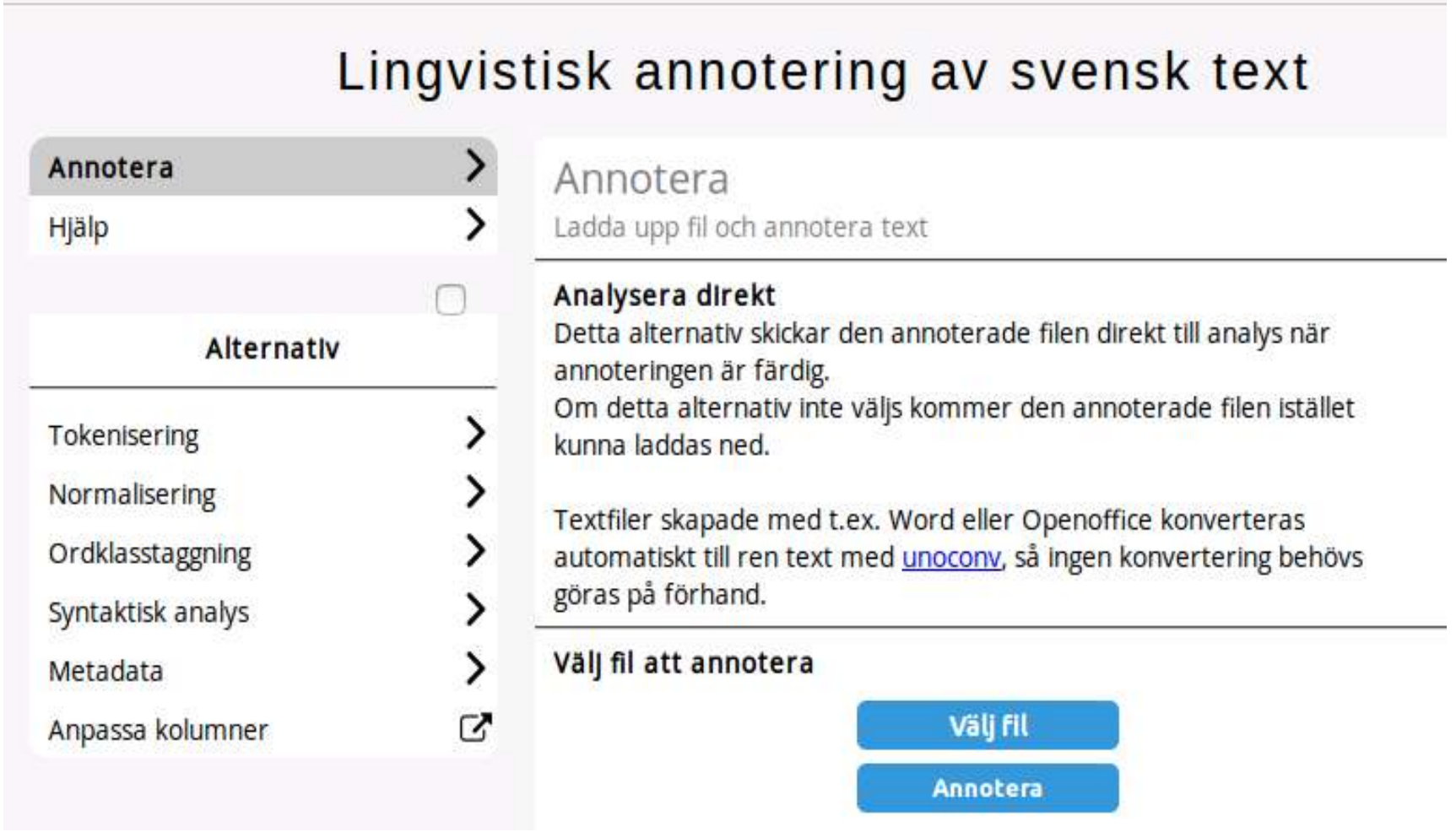


Analysera text

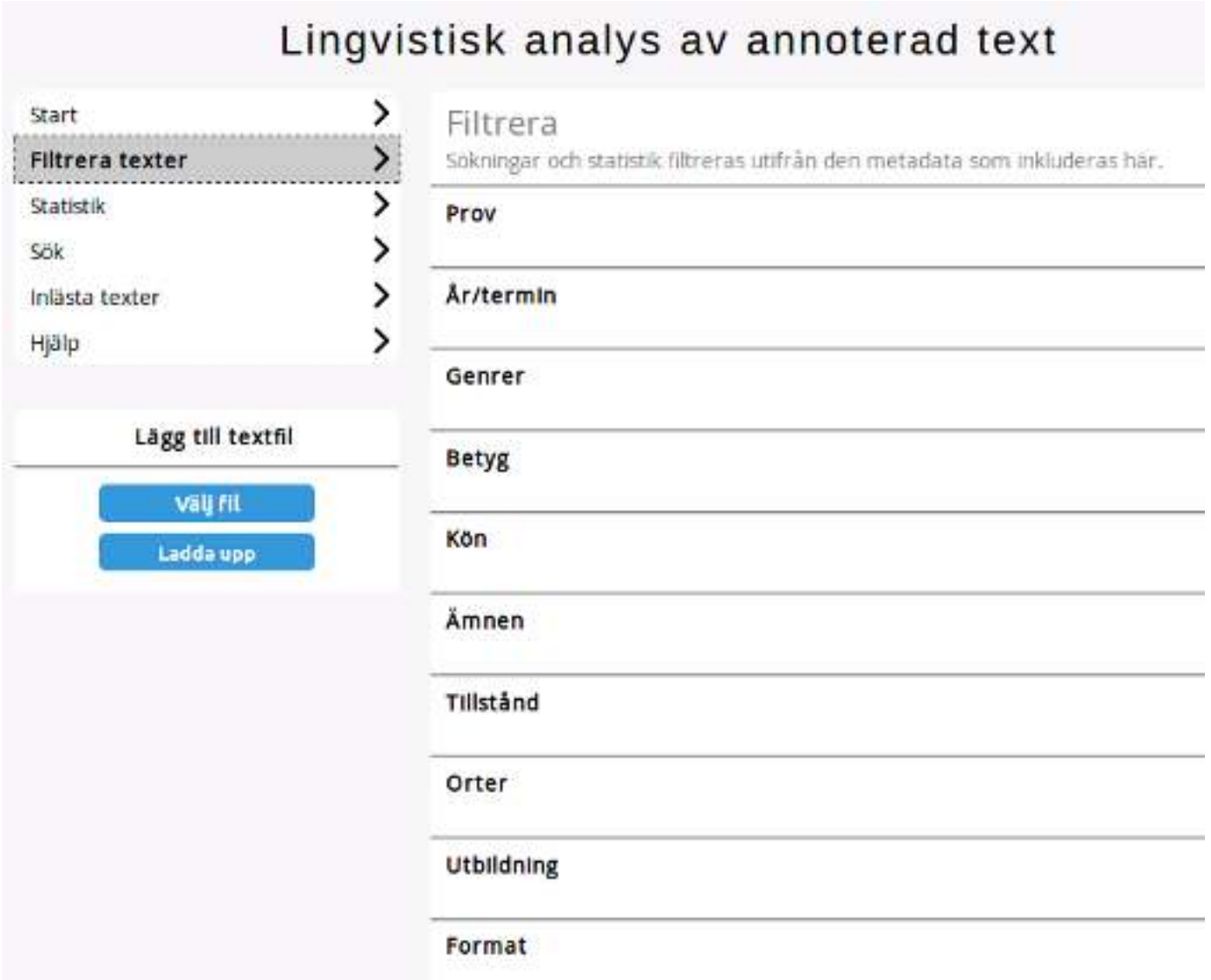
Ladda upp annoterade texter för statistik och sökningar.

http://stp.lingfil.uu.se/swegram

SweGRAM annoteringspipeline



SweGRAM visualiseringar



Kvantitativ analys av elevtexter

Statistisk sammanställning:

- Antal tokens, meningar och hela texter och genomsnittlig längd för dessa
- Fördelning av ordklasser med eller utan morfologiska särdrag
- Ordföljd baserat på syntax
- Läsbarhetsmått: LIX, OVIX, nominalkvot
- Jämförelser mellan olika grupper definierade av användaren

Kvantitativ analys av elevtexter

Metadata

Prov: GY
År/terminer: HT08, HT09, VT02, VT03, VT04, VT08, VT09, HT10, VT10
Genrer: BESK, _, UTR, ARG
Betyg: G, VG, IG, MVG
Kön: _, M, K
Ämnen: SVAS, SV
Tillstånd: ET
Utbildningar_
Format_
Orter_

Texter: 209
Meningar: 7516
Ord: 127331
Tokens: 140616
Särskrivningar: 12
Felstavningar: 859

Genomsnittlig ordlängd: 4.87 tecken
Genomsnittlig meningslängd: 18.71 ord
Genomsnittlig textlängd (tokens): 672.80 tokens
Genomsnittligt antal felstavningar per text: 4.11

Ord med fler än 5 tecken: 41364
Ord med färre än 5 tecken: 72185
Ord med exakt 5 tecken: 13782

Kvantitativ analys av elevtexter: felstavningar

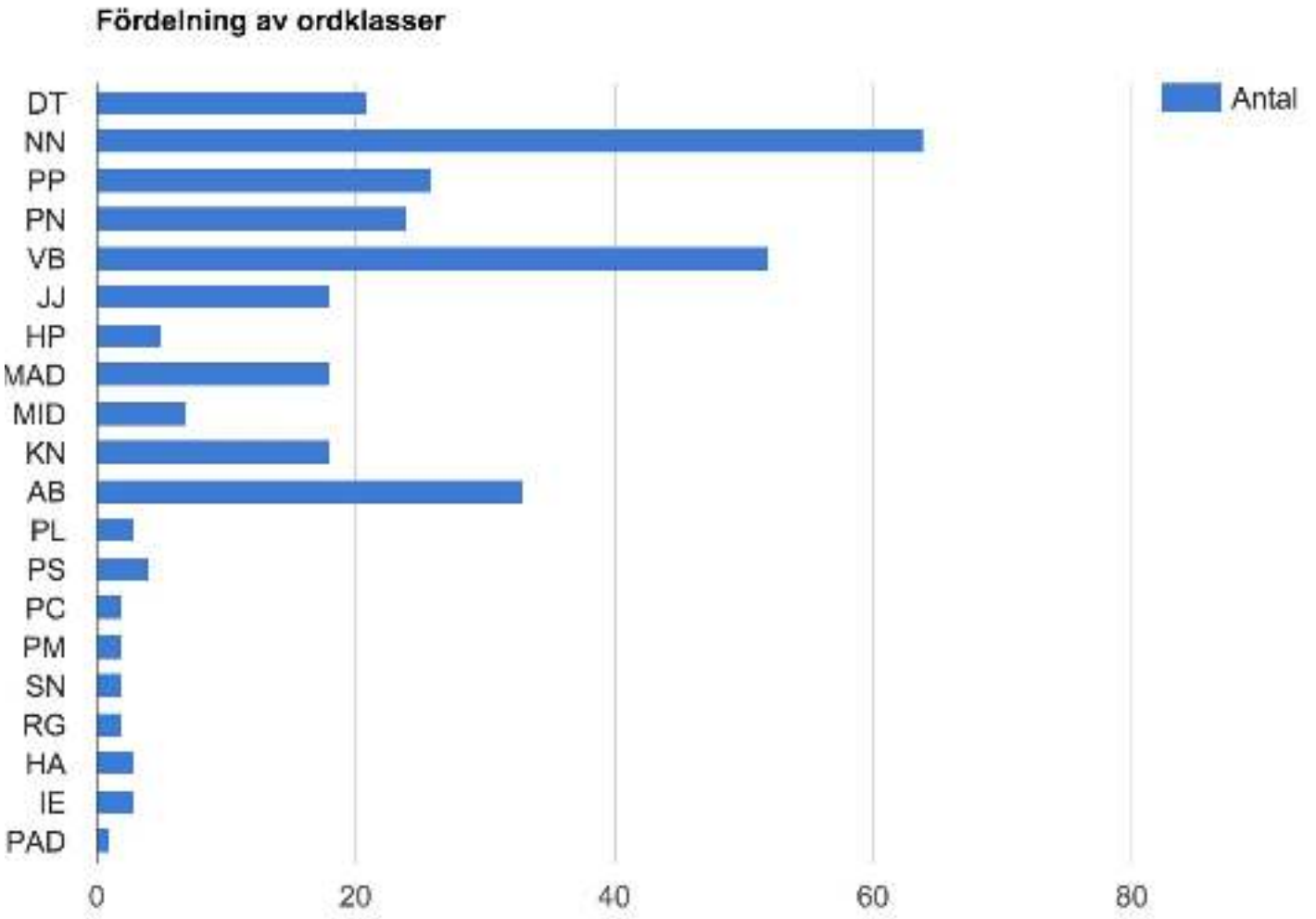
Fel	Rätt
tv	TV
imorrn	imorgon
kakibyxor	khakibyxor
Dokysåpor	Dokusåpor
skillda	skilda
Farmoden	Farmodern
svällter	svälter
Ihopskrynklade	hopskrynklade
cafét	caféet
idéen	idén
års-åldern	årsåldern
hjälva	hjälpa
massajen	massagen
your	jour
oengagerande	engagerande
imorrn	imorgon
andledning	anledning

Kvantitativ analys av elevtexter

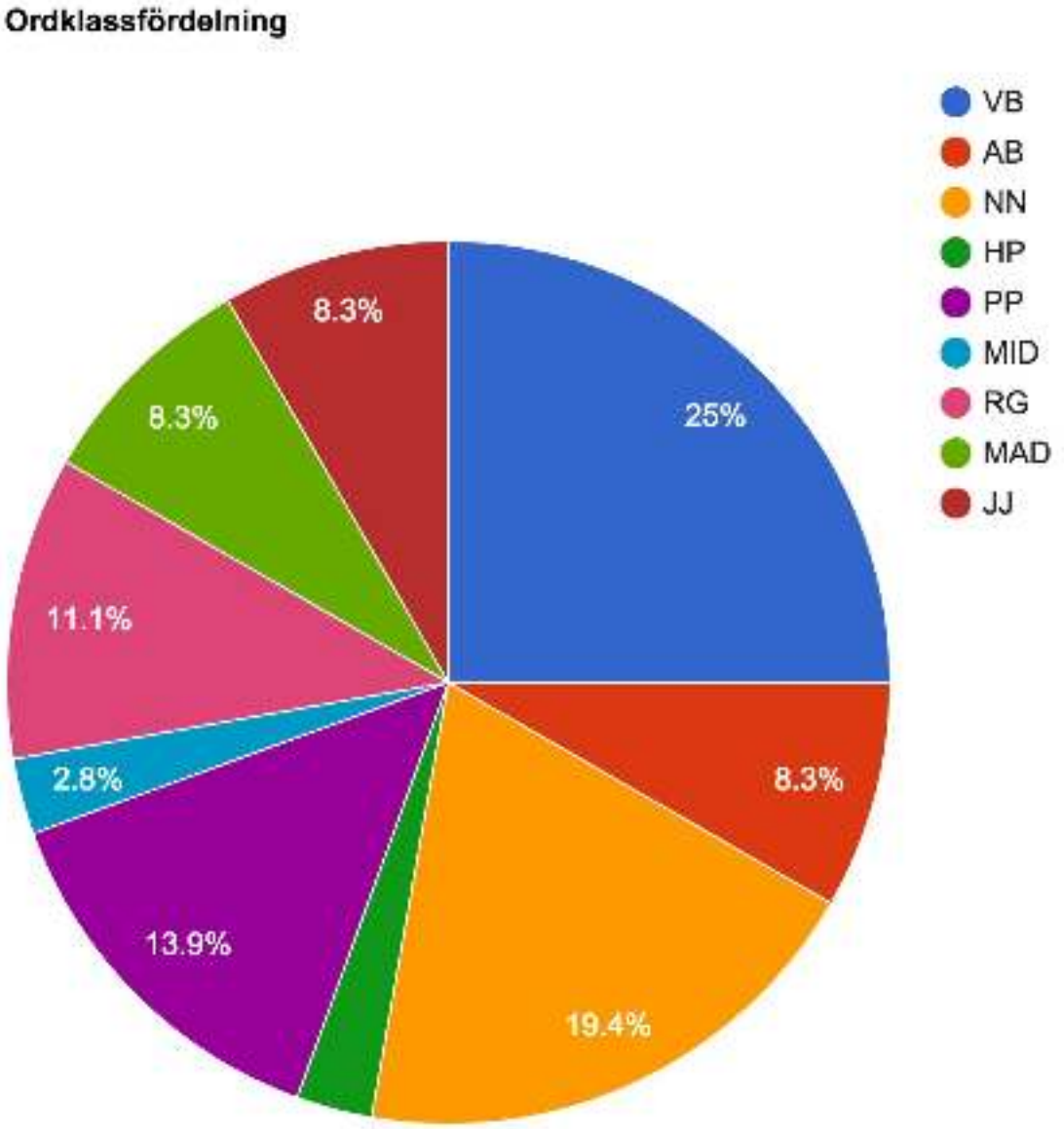
Ordklass (Tagg)	Förekomster	Procentandel
Verb (VB)	6794	17.54%
Substantiv (NN)	6650	17.17%
Pronomen (PN)	3559	9.19%
Preposition (PP)	3378	8.72%
Adverb (AB)	3205	8.27%
Adjektiv (JJ)	2497	6.45%
Skiljetecken (MAD)	2221	5.73%
Konjunktion (KN)	2004	5.17%
Determinerare (DT)	1700	4.39%

...

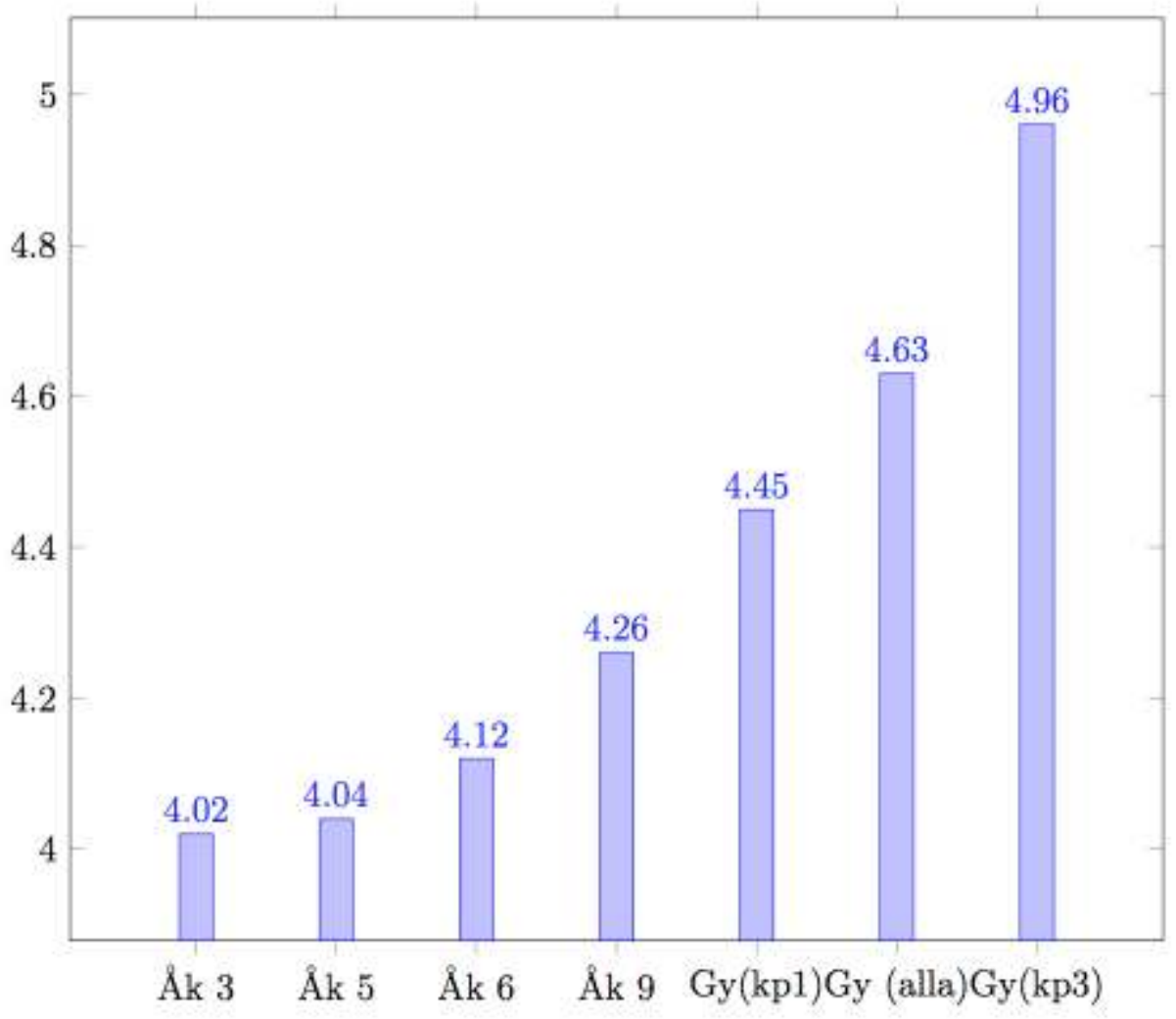
Visualisering: Fördelning av ordklasser



Visualisering



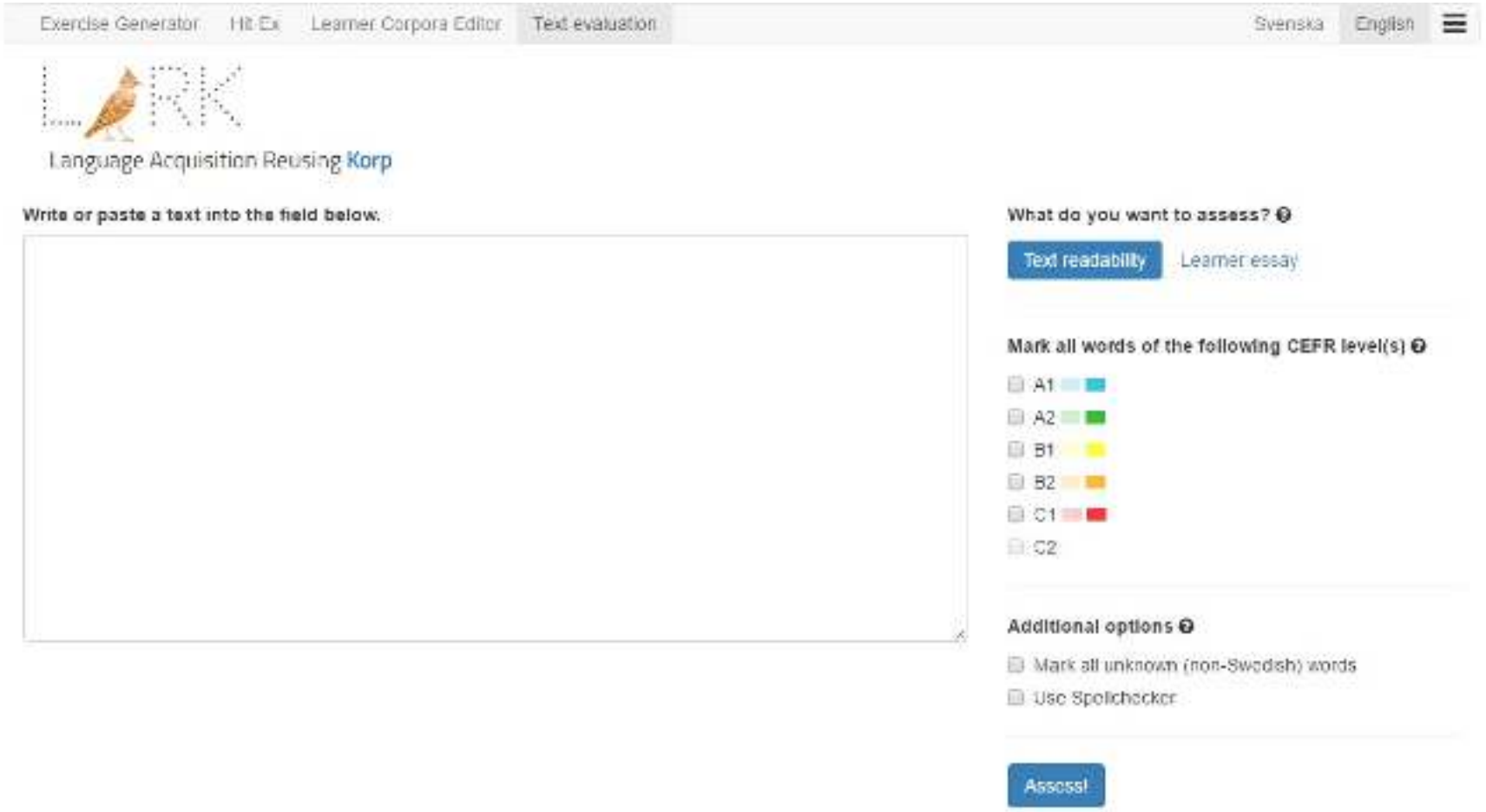
Visualisering:
Genomsnittlig tokenlängd per årskurs

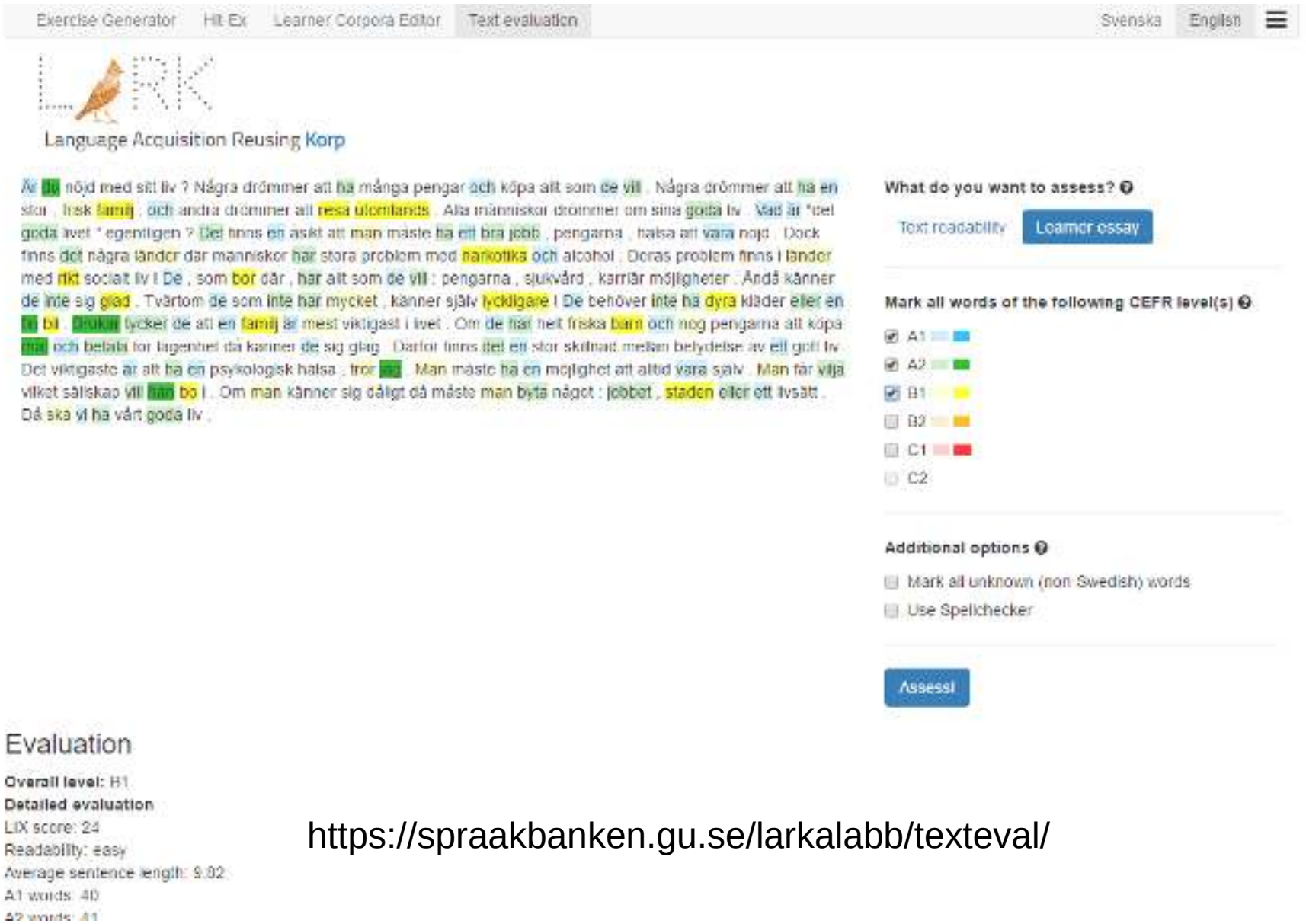


Framtiden

- Förfina analysverktygen för att anpassa till L2-texter
  - Vidareutveckling av rättning av stavning och grammatik;
- Skapa en verktygskedja som möjliggör rättning av automatisk lingvistisk analys av L2-texter;
- Utveckla visualiseringsverktyg för kvantitativ analys
  - Idag kan data läses in i Excel
- Genomföra preliminära studier baserat på kvantitativ analys

L2 textbedömning (CEFR nivåer)





https://spraakbanken.gu.se/larkalabb/texteval/

Exempel på en L2 uppsats

Berätta om dig själv! – B1

Jag heter NN some är 18 år.Jag kom i Sverige 2012 sedan började jag skolan efter några månader.Just nu bor jag ensam på en egen lägenhet som ligger i NN.

Jag bara studera 4 ämne i skolan och på frididen träna jag på gym och ibland jag går på bibiliotek eftersom vill jag lära mig mer svenska och engelska.

Jag har sökt sommarjobb också, hoppas att kan hitta jag nån jobb under sommern annars det bli jätte tråkig.

jag har två kompisar som hete NN och NN de är från NN-land också jag älskar de för att när jag behöver hjälp de hjälpar gärna mig och jag också hjälpa de.

Jag vill gärna bli sjukskörtöske i framm tiden för att jag vill hjälpar andra människor som behöver hjälp.

Min drömar är en dag kan jag tala perfekt svenska och engelska sen vill jag resa i andra länder och besök de andra människor med andra kultor och tala som de har. Jag vill gärna känna mig många nya saker.

Mina intressant är spela fotboll, resa, träna på gym, och hitta en yrke som passar mig bra. Det är allt.

L2 “alternativ” data (loggar)



Exercise type evaluation

Bundled gaps (variant 1)

Which word fits into these gaps? Each gap contains the same word. Write the word.

Hennes _____ var på hans lår , gned in värme i hans kalla ben .

En annan taxi tar _____ om skolbarnen .

Novelty hade flera trumf på _____ .

I första _____ har hon spelat dragspel och fiol .

Evaluation

For which levels is this exercise type relevant?

☐ A1 ☐ A2 ☐ B1 ☐ B2 ☐ C1

Comments

Page 1 of 7

Continue

https://spraakbanken.gu.se/larkalabb/exeval

L2 “alternativ” data (loggar)

https://spraakbanken.gu.se/larkalabb/wordguess

LÄR språket via Korpus Analys

Word guess

Tries: 0/7

Score: 0

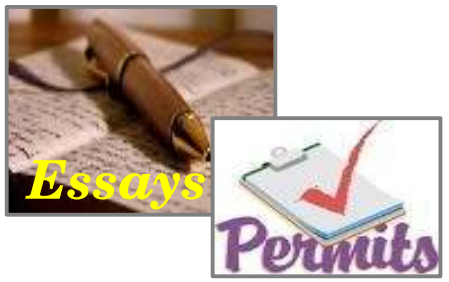
Definition:
rör sig tyst och försiktigt för att inte bli upptäckt, tassar (pres ind aktiv)

Hjälp:
slip; steal, sneak, creep



A B C D E F G H I J K L M N O P Q R S T U V W X
Y Z Ö Ä Å

Arbetsflöde (när allt är klart)



Filförberedelse
Fil-upload
Metadata registrering
<Digitalisering (M)>
Export till textformat (A)

Initiell annotering
Tokenisering (A)
Lemmatisering (A)
NER (A)?
Anonymisering (A)?

Normalisering
- stavfel (A)
- frasfel/kongruens (A)
- ordföljd (SA)
- ordval (SA)
- ...

Felannotering
- enl feltaxonomi,
delvis automatisk

Lingvistisk annotering
- tokenisering
- lemmatisering
- syntaktisk annotation

Analys
Nivåbedömning (A)
Topik klassificering (A)?
Genre klassificering (A)?
L2 vs L1
(eg overuse, underuse)?

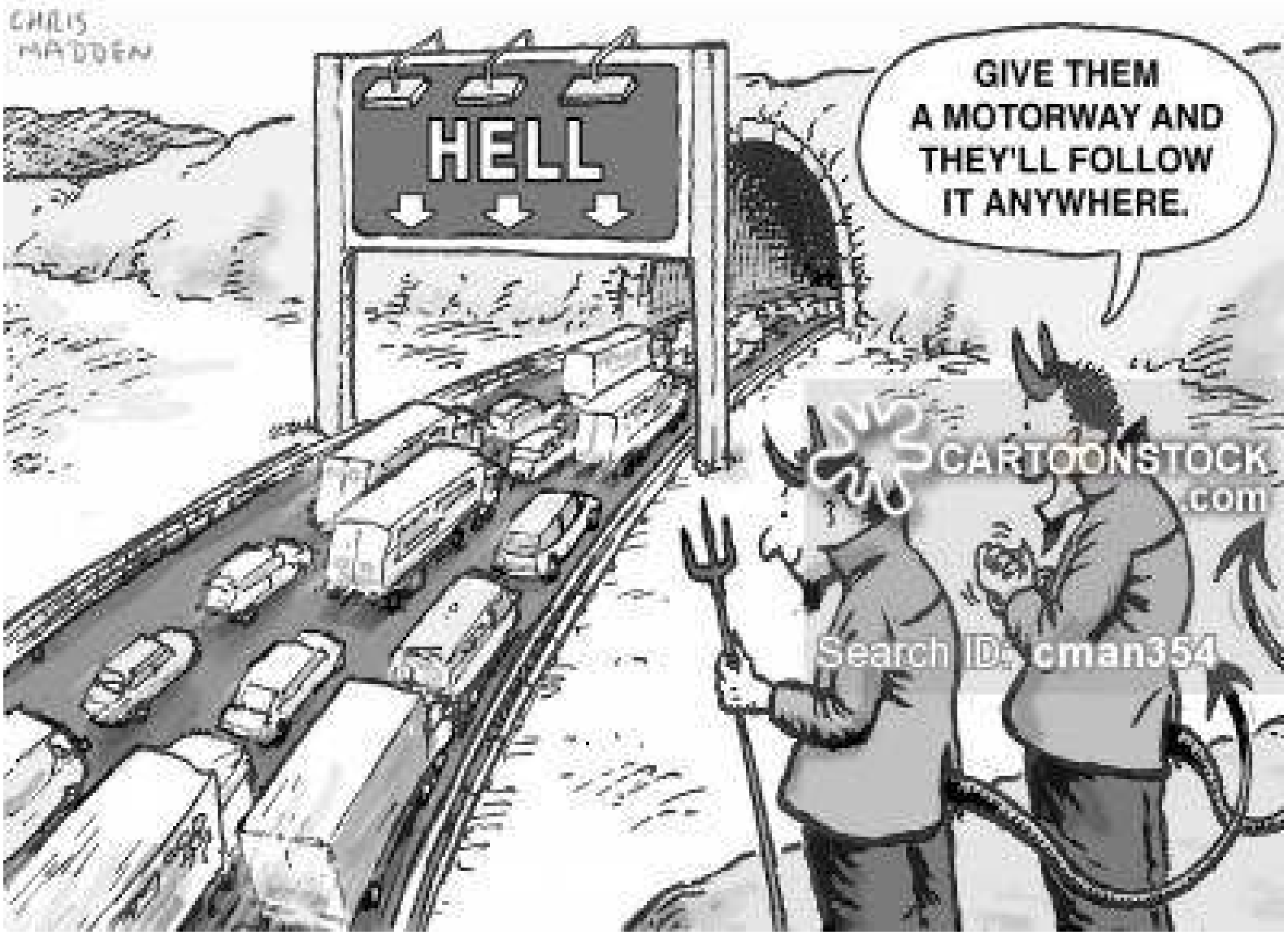
Sökmiljö (STRIX)
- enl specifikationer

Annoteringsmiljö
manuella annoteringar
korrigeringar

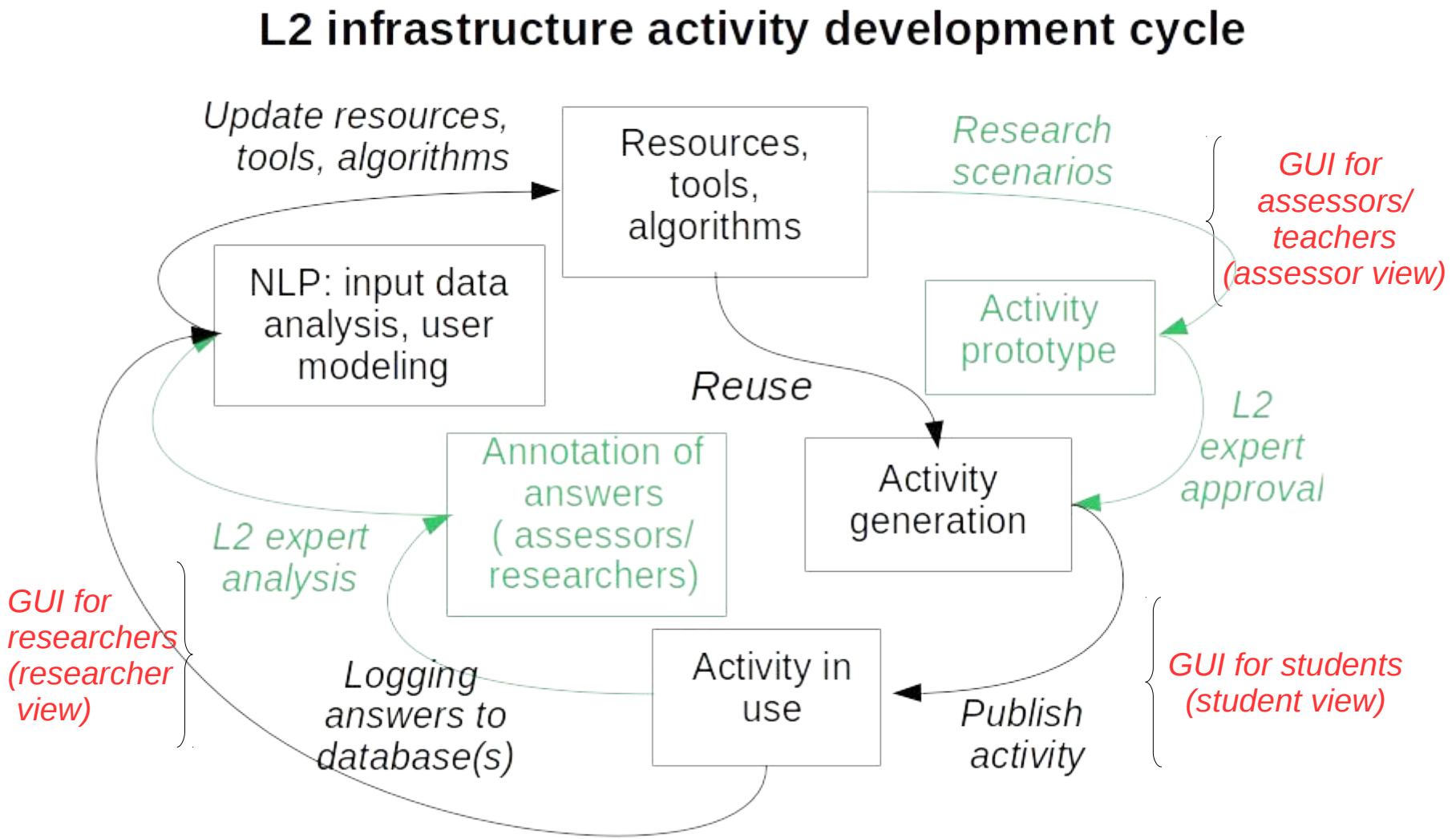
Next step - reliability of tools



Vart ska det leda?



Det ultimata målet



Tack!

Frågor?