

Läsbarhetsstudier för svenska som andraspråk

Ildikó Pilán, Elena Volodina

**Språkbanken
Göteborgs Universitet, Sweden**

Nov 30, 2016

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)

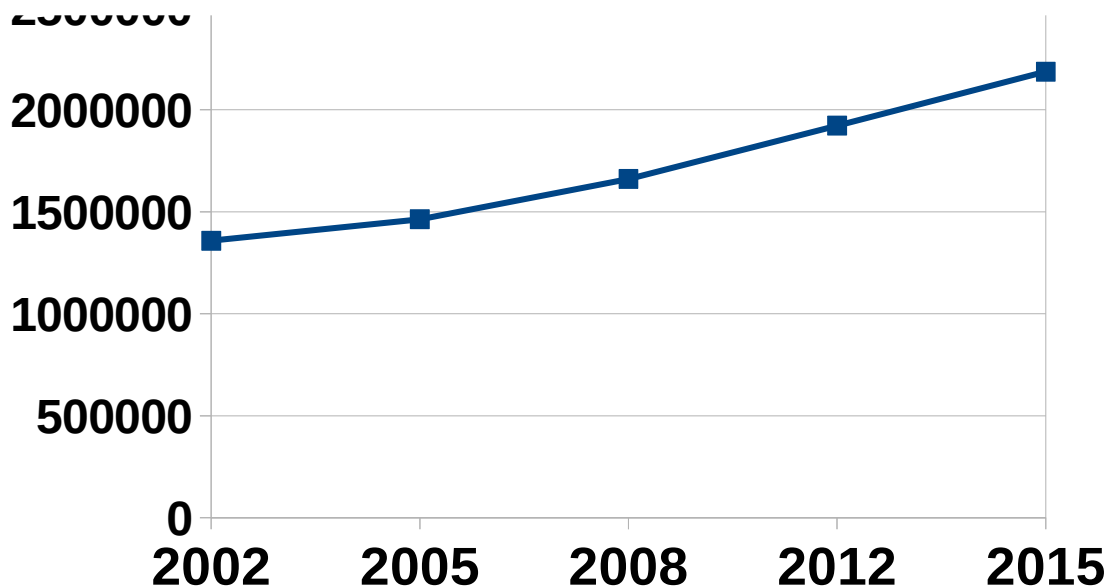


Läskunnighet: Sverige

- PIAAC studie 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

Samhällsbehov

Medborgare med utländsk bakgrund, Sverige, 2002-2015

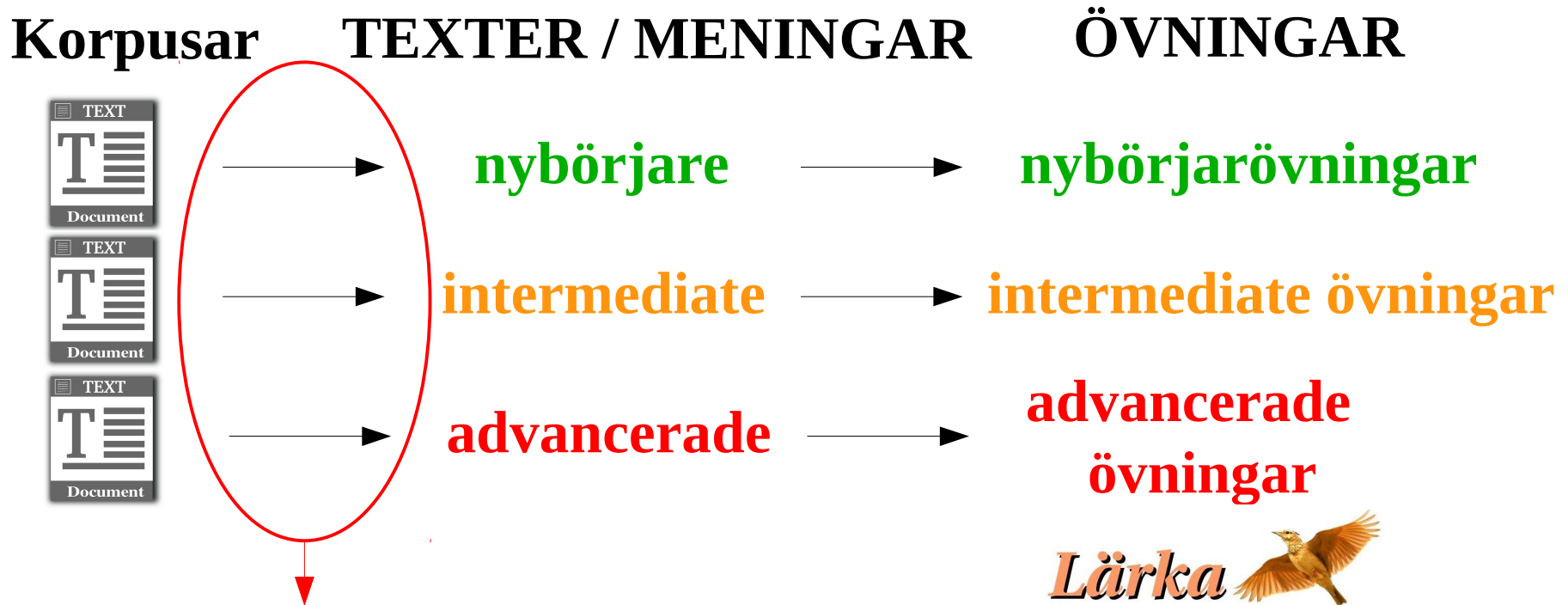


2015: av **9,9 mln** invånare, **2,2 mln** har utländsk bakgrund, dvs **22,2 %**
(Statistiska centralbyrån)

Vad kan vi göra?



- **Mål:** återanvända korpusar för att stödja språkinlärning



Hur kan vi automatisk bedömma *textsvårighet*
(läsbarhet) på text eller meningsnivå?

Common European Framework of Reference (CEFR)

Kan inte ...



Kan det mesta



Data





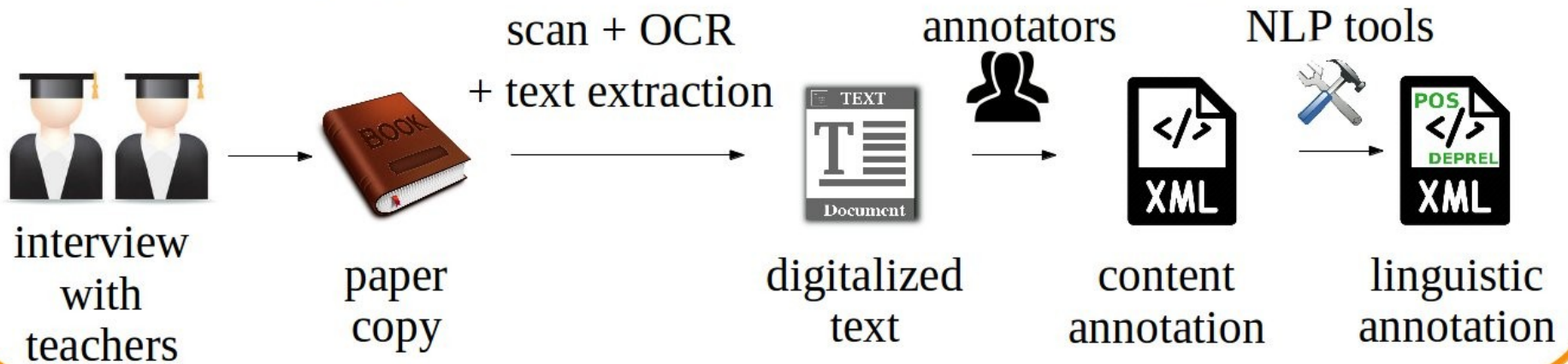
COCTAILL

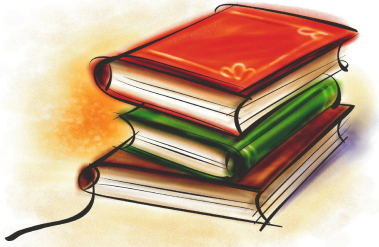
- COCTAILL - **C**orpus **o**f **C**EFR **T**extbooks **a**s **I**nput for **L**earner **L**evel modeling
- Januari 2013-oktober 2014
- CLT-finansierat (Center for Language Technology, GU)
- A1-C1 nivåer (ej C2)

*Elena Volodina, Ildikó Pilán, Stian Rødven Eide and Hannes Heidarsson 2014.
You get what you annotate: a pedagogically annotated corpus of coursebooks
for Swedish as a Second Language. NEALT Proceedings Series 22*

COCTAILL korpus

From paper books to an annotated corpus





COCTAILL “ingredienser”

ELEMENTS

Coursebook →

└ Extras

└ Lesson →

└ Text →

└ Genre →

└ Subheadings

└ Activity instruction

└ Task

└ List

└ Language Example

ATTRIBUTES

title, author etc.

CEFR level, (title)

topic (e.g. arts), (title)

{ genre family (e.g. facts)

{ macrogenre (e.g. dialogue)

} skills/competences,

} activity type/format

} skills/competences,

} linguistic unit

Common attributes: → id, ref

Resurser





SVALex - receptivt ordförråd

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

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

user: englishprofile
password: vocabulary



Browsing SVALex

Enter a word

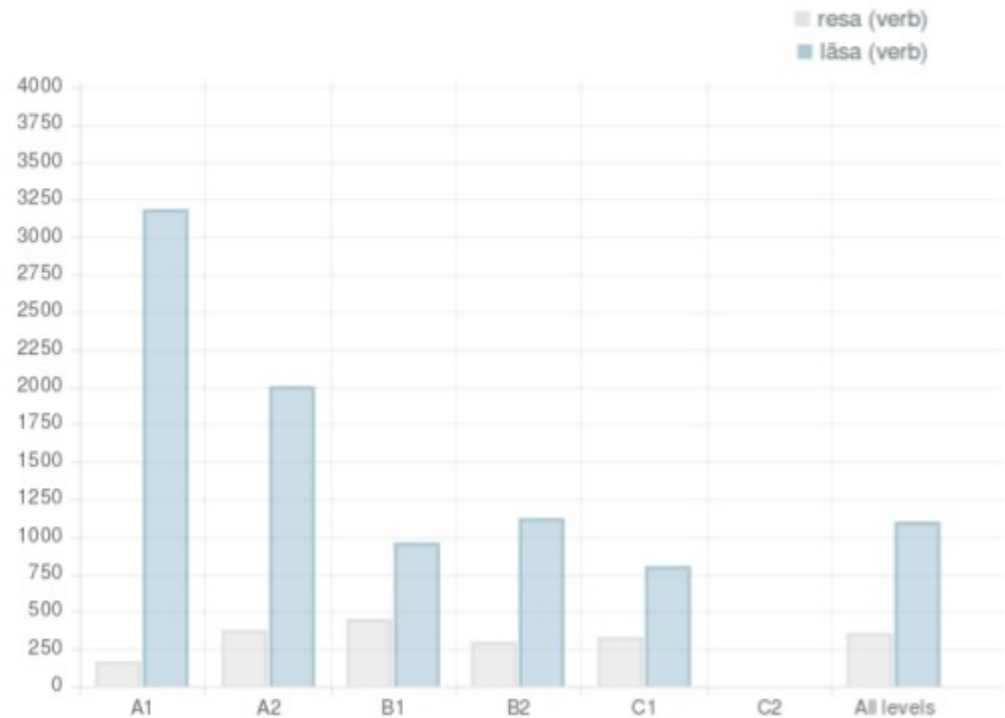
resa

läsa

-

Search

Frequencies by CEFR levels for the words resa and läsa.



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

Kelly listan

- Moderna språk (9 st)
- Mest frekventa ord
- Från webb-korpusar (min 100 mln löpord)
- Objektivt urval kontra subjektivt bedömning
- Indelning in i CEFR-nivåer

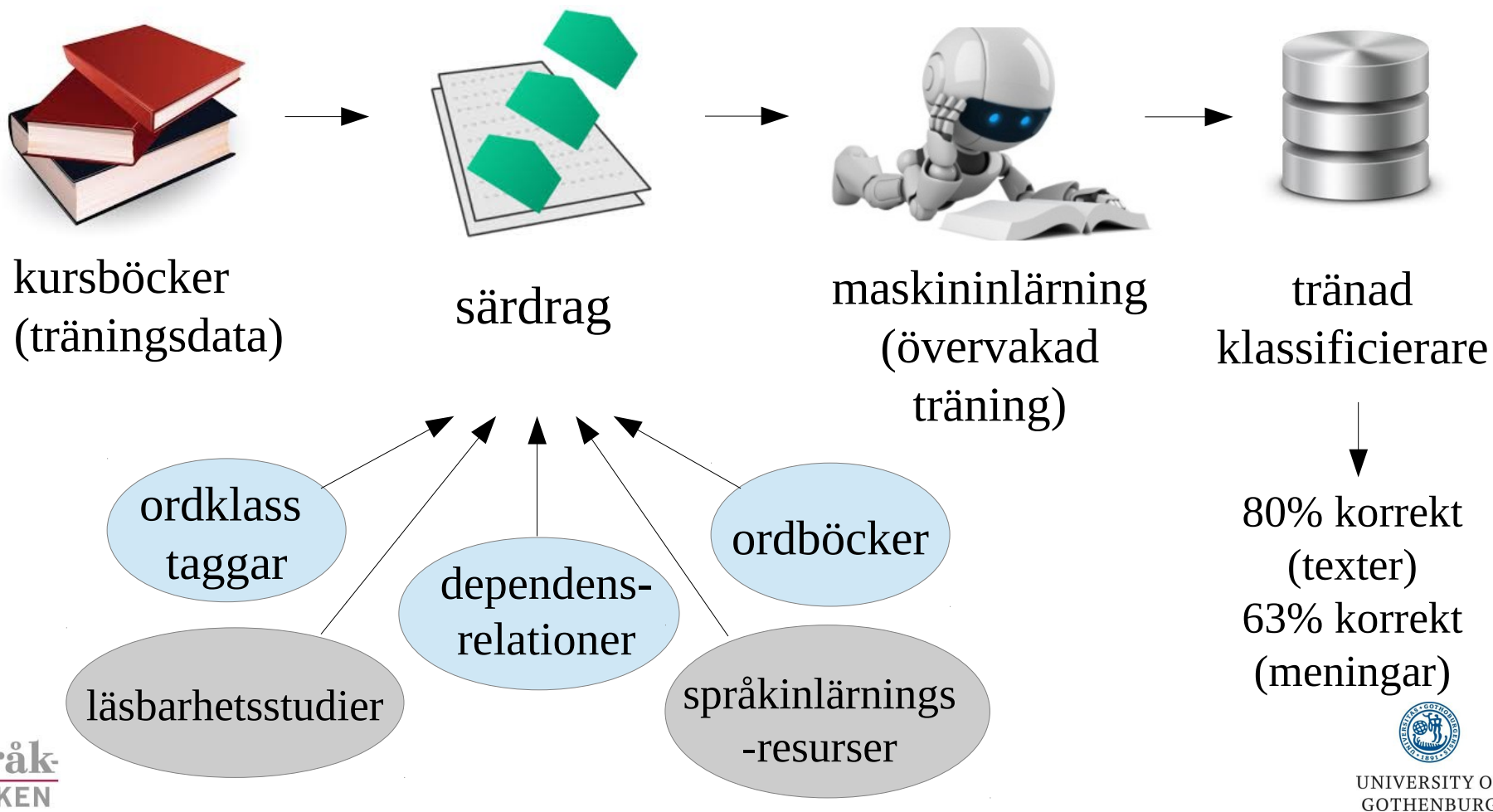
ID	Raw freq	WPM	CEFR levels	Source	Grammar	Swedish items for translation	Word classes	Example
1319	7920	69.47	A1	SweWaC	en	häst	noun-en	

Experimenter

Läsbarhetsexperiment

- Maskininlärningsmetoder för automatisk klassificering (WEKA, scikit-learn)
- Text- och meningsnivå
- Baserad på COCTAILL korpusen
- 5 CEFR nivåer (A1-C1)
- 61 olika lingvistiska särdrag

Från kursböcker till automatisk CEFR-nivå bedömning



Särdrag

Count	Lexical	Syntactic	Morphological	
Sentence length	A1 lemma IS	Avg DepArc length	Modal V to V	Verb IS
Avg token length	A2 lemma IS	DepArc Len > 5	Particle IS	V variation
Extra-long token	B1 lemma IS	Max length DepArc	3SG pronoun IS	Function W IS
Nr characters	B2 lemma IS	Right DepArc Ratio	Punctuation IS	Lex tkn to non-lex tkn
LIX	C1 lemma IS	Left DepArc Ratio	Subjunction IS	Lex tkn to Nr tkn
Bilog TTR	C2 lemma IS	Modifier variation	PR to N	Neuter N IS
Square root TTR	Difficult W IS	Pre-modifier IS	PR to PP	CJ + SJ IS
Semantic	Difficult N&V IS	Post-modifier IS	S-VB IS	Past PC to V
Avg senses per token	OOV IS	Subordinate IS	S-V to V	Present PC to V
N senses per N	No lemma IS	Relative clause IS	ADJ IS	Past V to V
	Avg. KELLY log freq	PP complement IS	ADJ variation	Present V to V
			ADV IS	Supine V to V
			ADV variation	Relative structure IS
			N IS	Nominal ratio
			N variation	N to V

Online verktyg för textbedömning

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

[Exercise Generator](#)[Hit-Ex](#)[Learner Corpora Editor](#)[Text evaluation](#)[Svenska](#)[English](#)

Language Acquisition Reusing **Korpus**

Filmen hanlar om en pojke . Han heter NN och han gilla dansar ballet så mycket . När han har idrott lektion , brukar han inte träna boxning så att träna ballet med många tjejer i nästa klassrummet . Han tränar dansa mycket , var och när han kan . Hans lärare ger för han ett par skor av ballet och han gömmer den mellan för två medrasser . Den mest intressanta person i filmen är Billy . Han är en snäll pojke . Hans mamma dog , han bor med hans pappa , bror och hans mormor . Han älskar hans mormor så mycket . Hon är gammal och hon brukar göra konstiga saker . Billy sörjer när han saknar hans mamma . Hans pappa är en gruva och han stöjkar för att alla person måste jobba mycket hårt men lön inte mycket pengar för familj , mat och deras liv . De lätte arg filmen utspelar @ ag i England på 1984-talet . Liv av människor är fattiga . Man syns på kläderna i filmen . Man trots att det måste vara på 1984-talet .

What do you want to assess? ?

[Learner essay](#)[Text readability](#)

Show all words of the following CEFR level(s) ?

☒ A1☒ A2☒ B1☒ B2☒ C1

Additional options ?

☒ Mark all potentially incorrect words☐ Use Spellchecker[Edit text](#)[Reset](#)

Evaluation

Overall level: B1

Detailed evaluation

Number of sentences

16

Number of tokens

189

Framtida planer och behov

- “Meningsnivå” korpus (upptäckt av svåra meningar i texter)
 - Utvärdering 2015 - en mindre mängd av data samlats in
- Analys av ordkomplexitet
- Läsförståelse → stöd (länkning till lexikala resurser, mm.)
- Logga elevernas interaktion med texterna och hjälpmedel
- Ytterligare utvärdering med lärare och elever
- Samarbete (Språkrådet mm.)