A friend in need?

Language Technology for Second Language Learning

Elena Volodina

University of Gothenburg, Department of Swedish, Språkbanken

elena.volodina@svenska.gu.se

Language Technology is ...

- ...a cross-disciplinary research area that covers development of computer programs for analysis, interpretation and generation of natural languages, etc... [*Wikipedia*]

- Other names: Computational Linguistics, Natural Language Processing (NLP), etc...

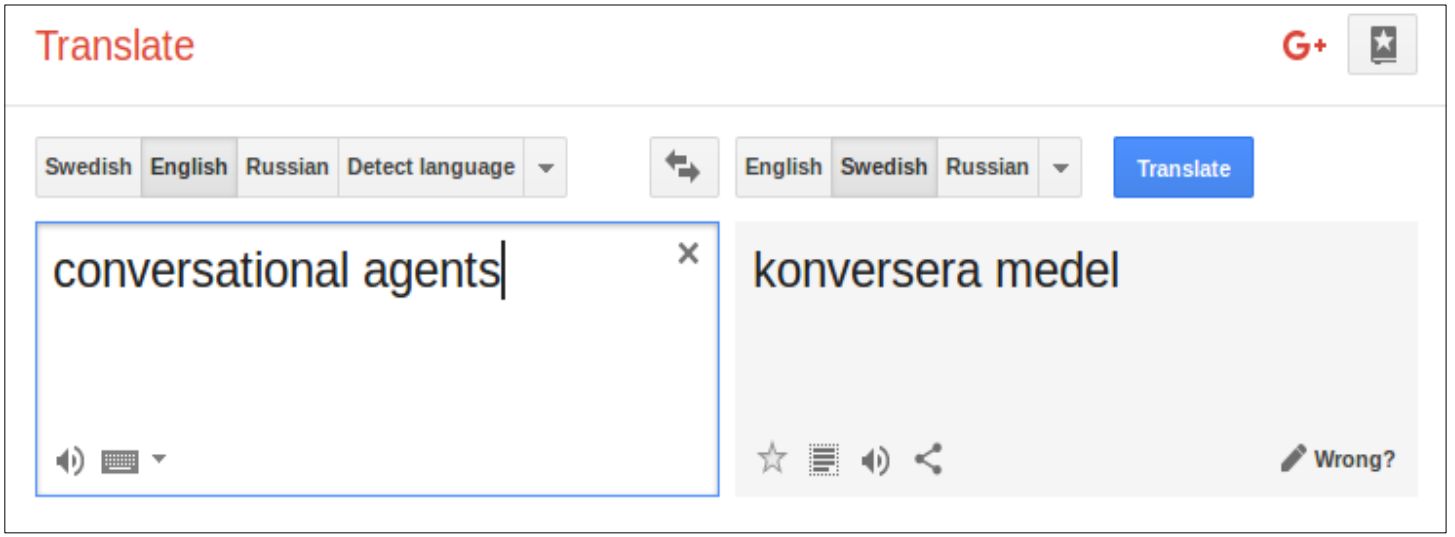
Origins

- 1950's: Machine translation (MT), Russian-English, first attempts in USA

Origins

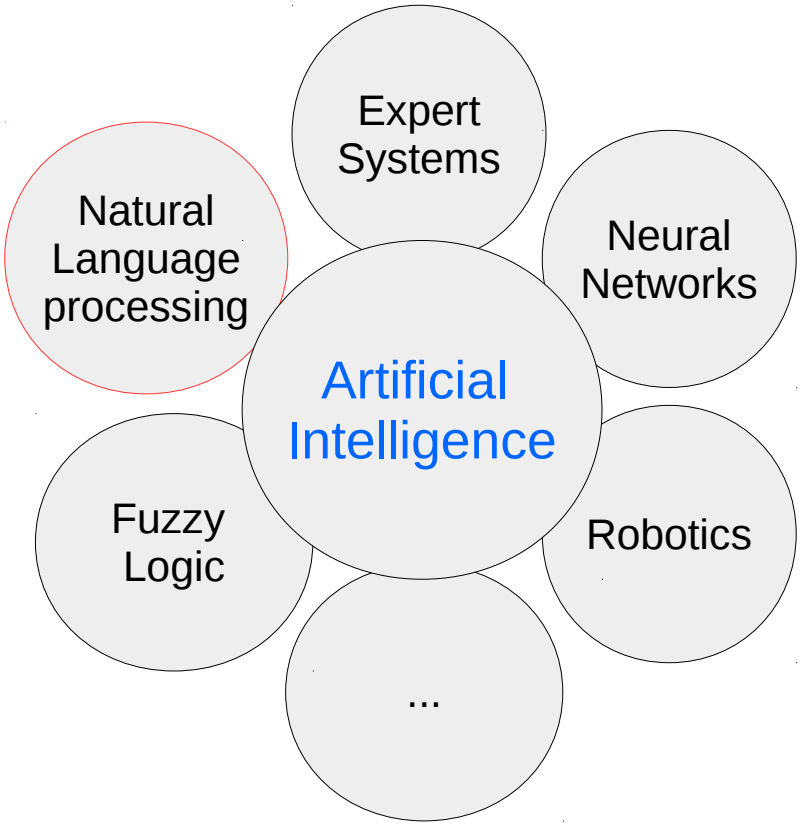
- 1960's: MT failure => Computational linguistics as a new research field

* grammar in both lan-ges
* morphology (inflections)
* syntax
* semantics
* lexicon
* pragmatics

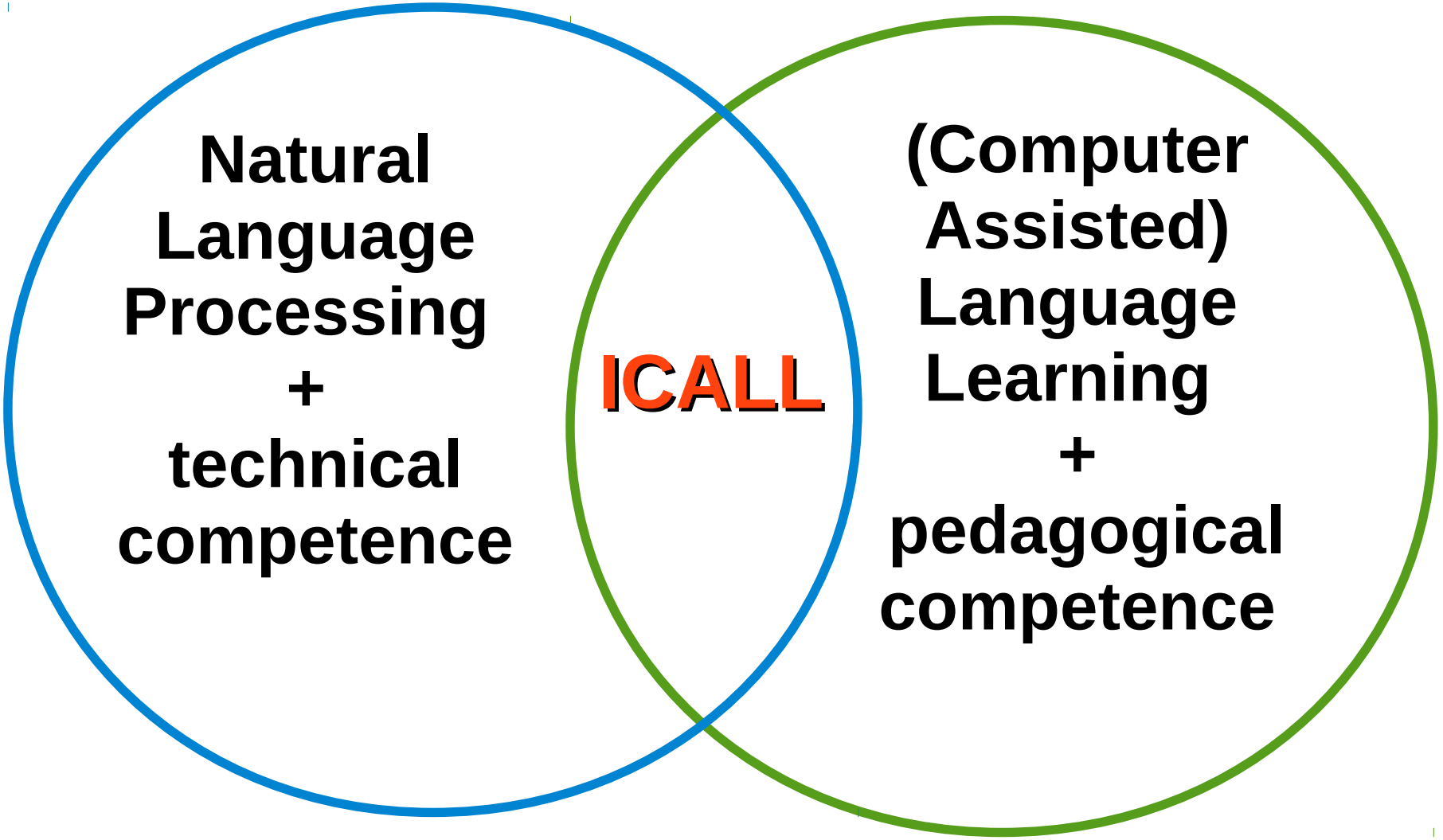


Current context

- Later: Language Technology/NLP is placed under *Artificial Intelligence*



NLP + CALL = ICALL

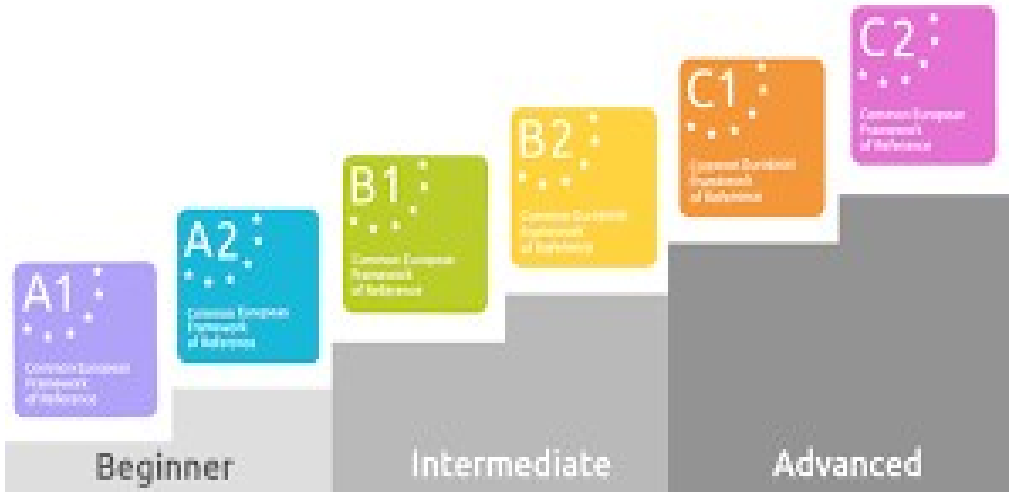


Focus on literacy

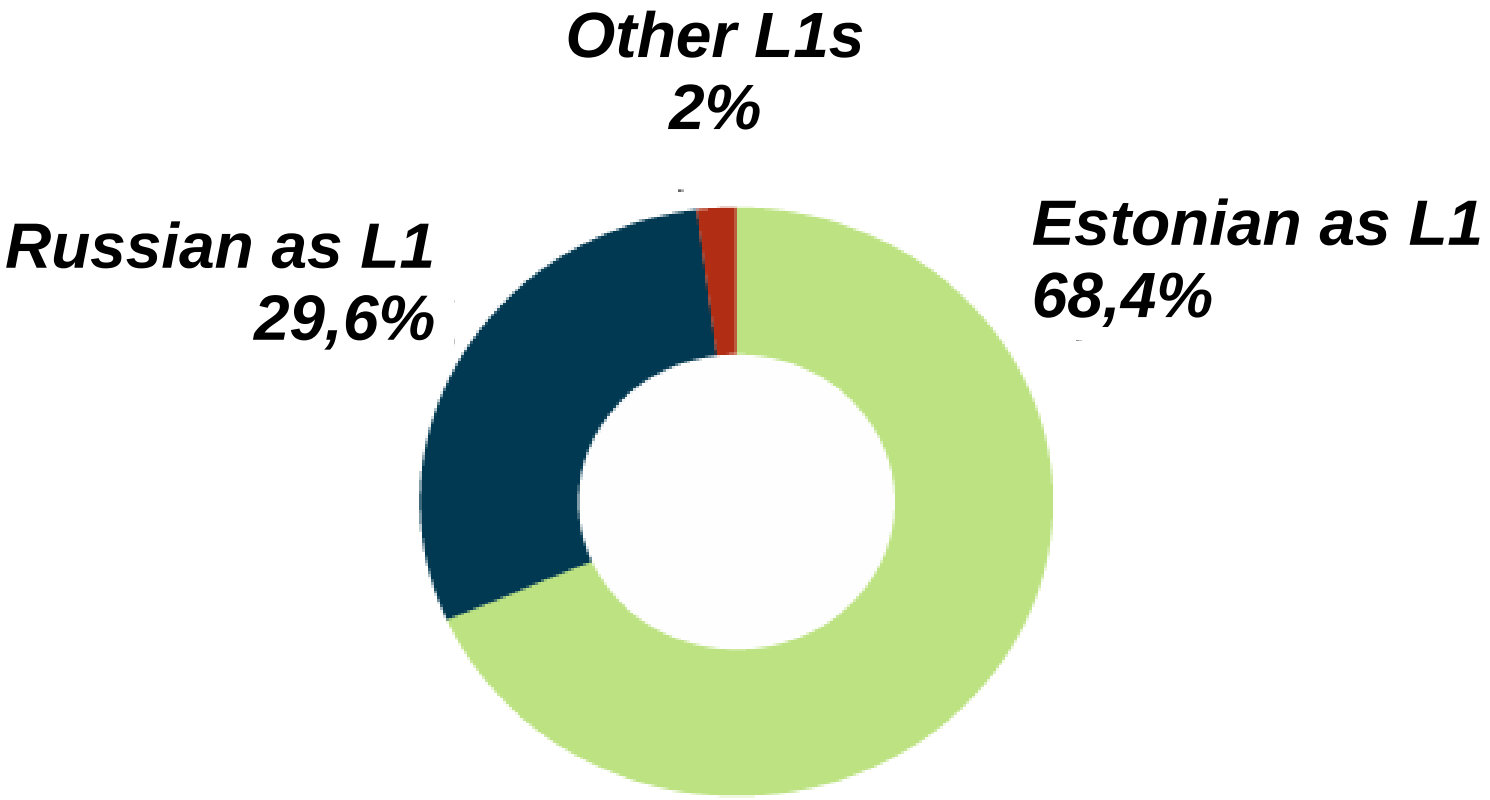
- Dutch study:

  → Average reading comprehension ~B1 level

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



Focus on Estonia, 2016

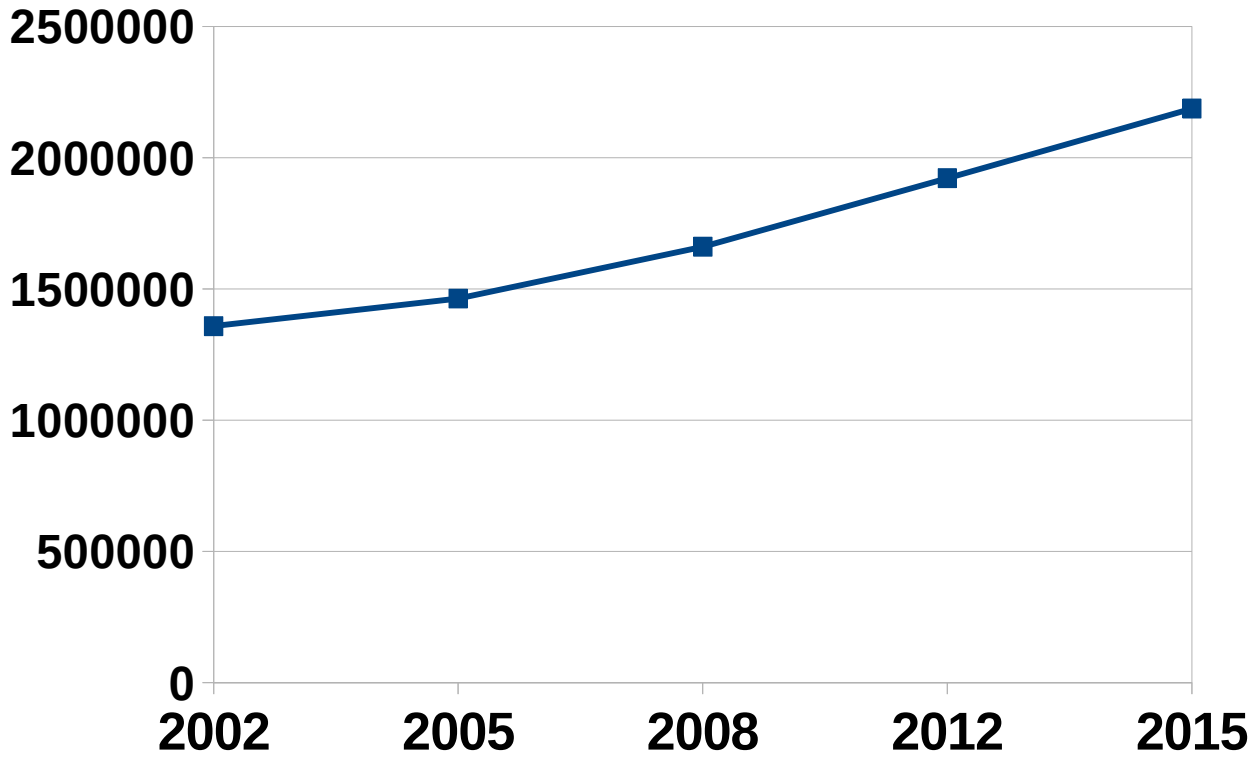


Out ot 1.3 mln citizens 0.4 mln (potentially) need training in Estonian

Source: Statistikaamet: http://www.stat.ee/main-indicators.

Swedish societal need

Citizens with foreign background, 2002-2015



2015: out of **9,9 mln** citizens, **2,2 mln** have foreign backgrund, dvs **22,2 %**
(Statistiska centralbyrån)

What can we do?

cause versus symptoms

ICALL tools for Second language (L2) learning

Language skill

Data & Resources

Tools & algorithms

Prototypes

Evaluation

Target group

ICALL tools for L2 learning

Target group

Grown-ups

Children

Analphabets

Special needs

etc

ICALL tools for L2 learning

Language focus

- Speaking
- Writing
- Listening
- Reading

Skills

- Pronunciation
- Vocabulary
- Grammar

Competences

ICALL tools for L2 learning

Data & Resources

Corpora

Essays

Word lists

Grammar rules

etc

Tools & Algorithms

POS-taggers

Lemmatizers

Sentence/text readability

Error detectors

Speech synthesis

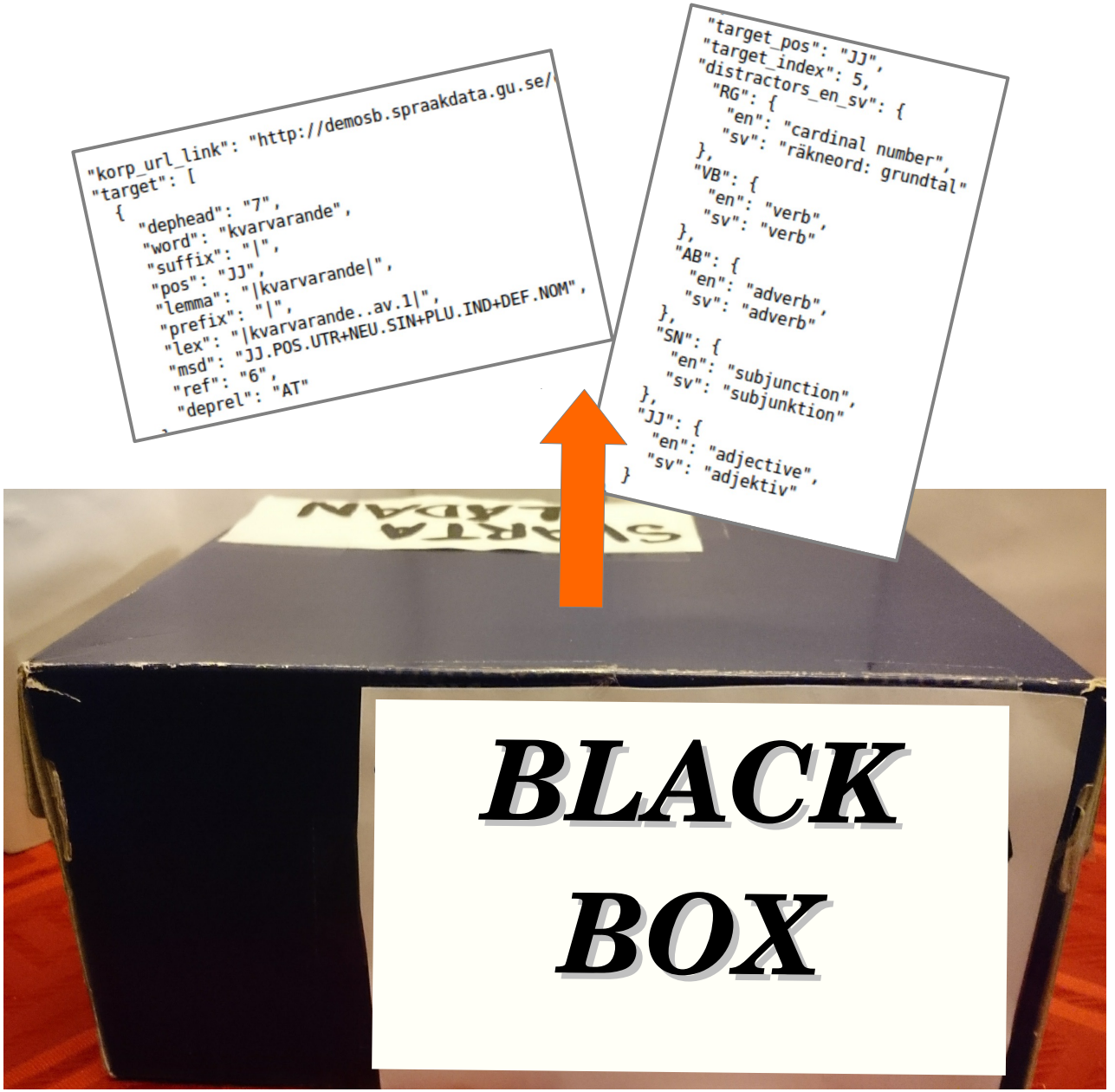
etc

ICALL tools for L2 learning

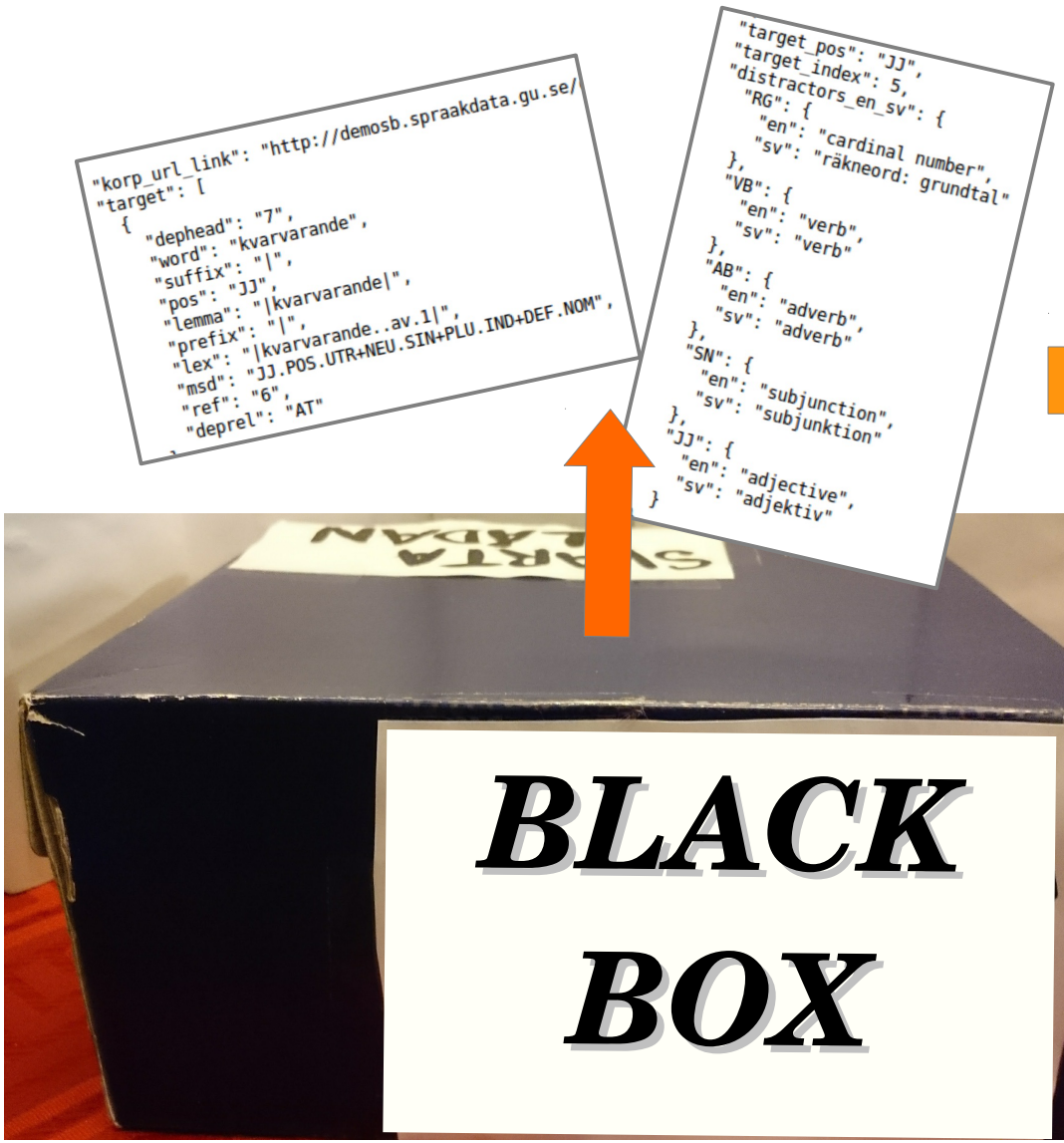
Resources

Algorithms

ICALL tools for L2 learning



ICALL tools for L2 learning



ICALL tools for L2 learning

Application-development and maintenance

versus

Prototype-development and evaluation (proof-of-concept)

Lark Trills for Language Drills

Text-to-speech technology for language learners

- Dictation and spelling exercise
- Focus on
  - evaluation of the quality of TTS
  - finding ways to give feedback on **spelling errors**

Elena Volodina, Dijana Pijetlovic. 2015. Lark Trills for Language Drills: Text-to-speech technology for language learners. Proceedings of the 10th workshop on Building Educational Applications (BEA10), NAACL 2015, Denver, USA

Fully automatic

self-study mode | test mode | timed test

word | inflected word | phrase | sentence | performance | Generate

repeat | slower | faster

Result Tracker

Exercise name	Correct/Total »
Learners/spelling-word, self-study	1/2

Train spelling, word level

Type the word you hear

TIPS

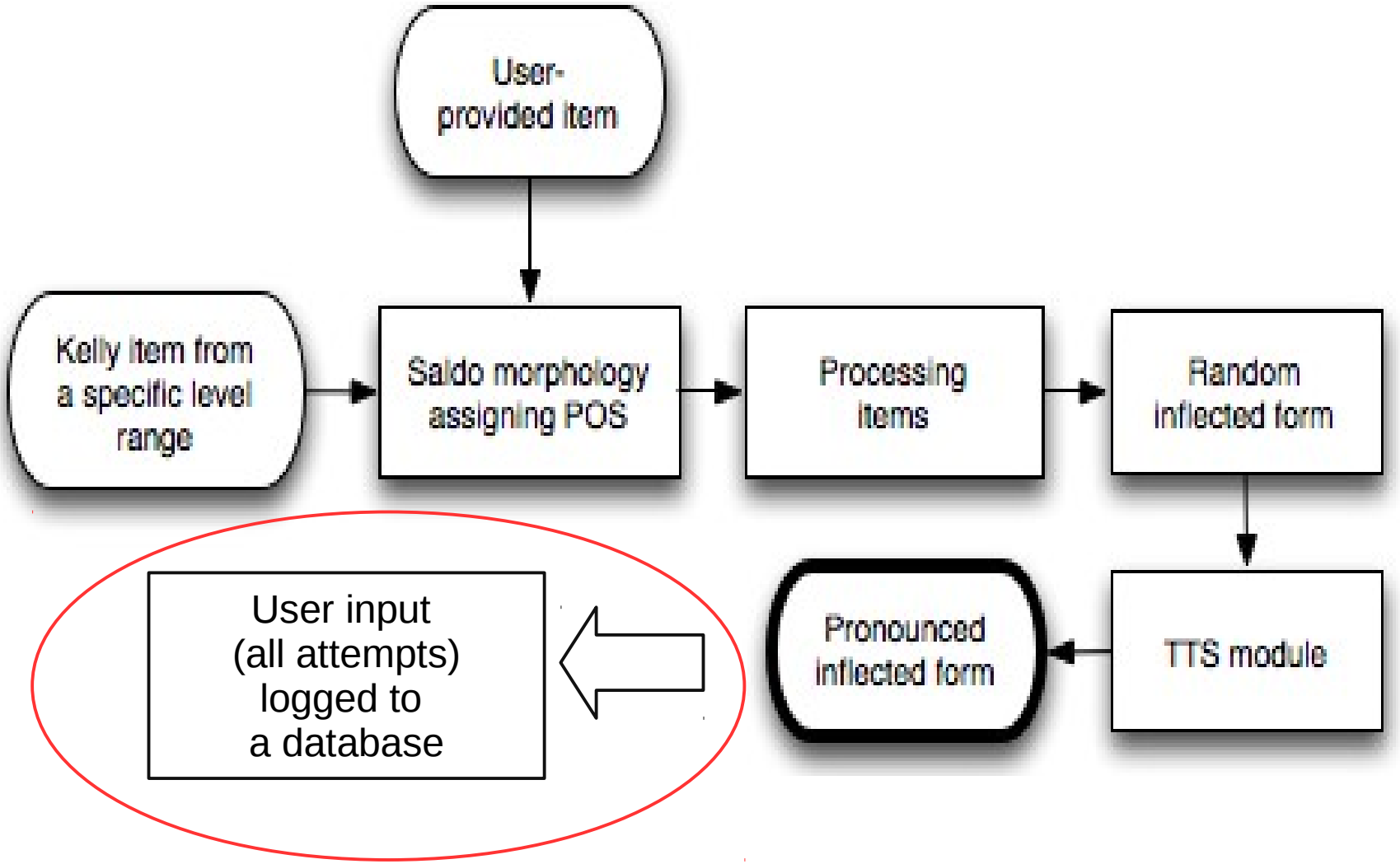
Nr	Word	Correct answer	Links
2	submit		JSON
1	matematik submit	matematik	

- Saldo morphology: *matematik*
- Wikipedia: *matematik*
- Wiktionary: *matematik*
- Monica: listen to pronunciation

matematik

Pipeline

for word & (inflected word) levels



SPEED

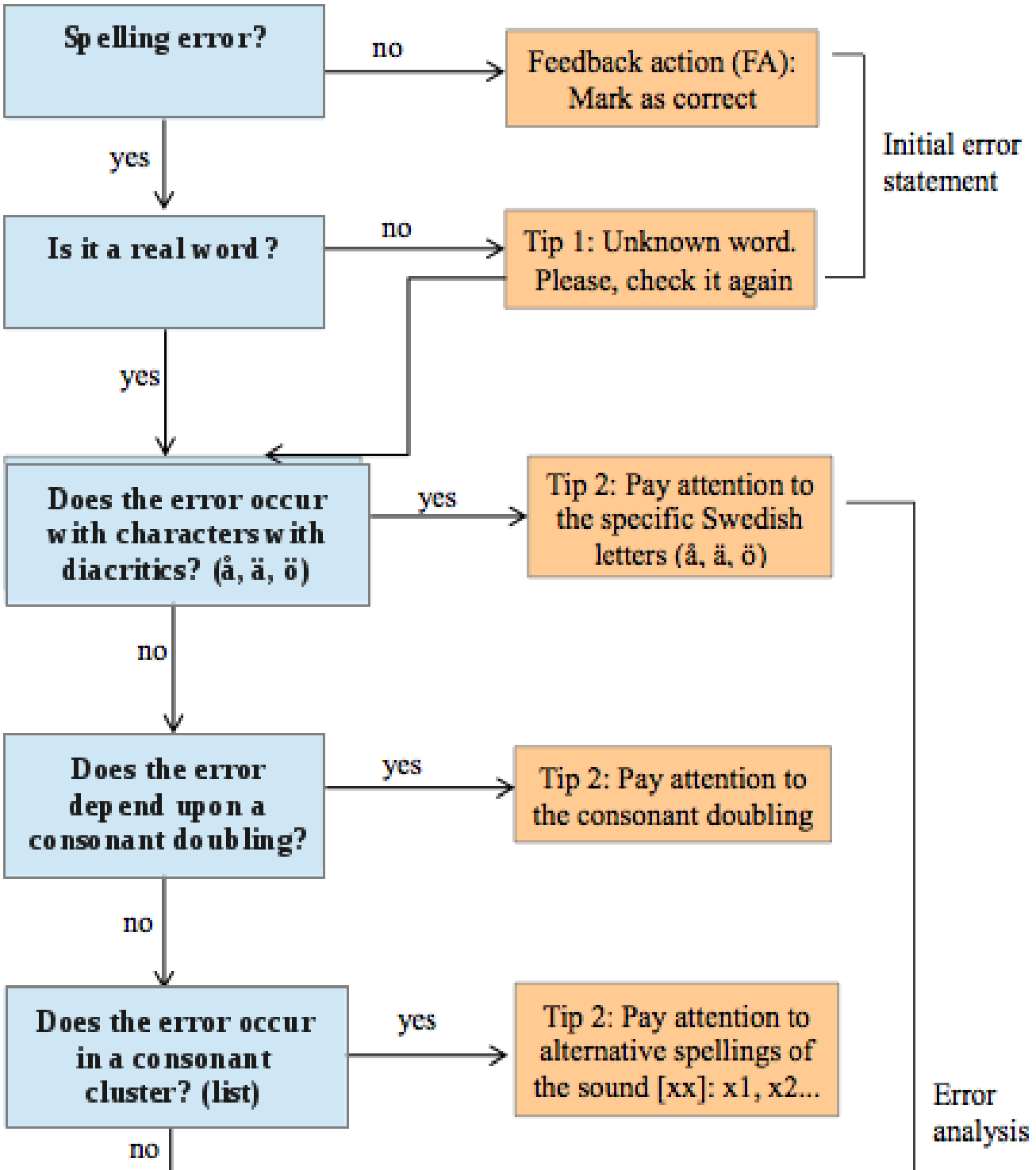
SPElling Error Database

- For each correct item (base form + word class) we store:
  - session ID (no personal data, such as L1)
  - incorrect spelling(s)

L2 spelling error database, SPEED

Error data

Error types	Nr,%	Example (correct →*incorrect*)
Competence-based errors	**55**	
Consonant doubling	28	sto**pp**a → *sto**p**a*
Diacritics (å, ä, ö)	23	h**ö**gre → *h**o**gre*
Phonetic errors (e.g. voiced vs voiceless)	25	relevan**s** --> *relevan**z***
Consonant clusters (phoneme-grapheme mappings, incl. cases of homonyms)	20	**sk**ön → ***sj**ön*
Other (unclassifiable)	4	**Israel** → ***visträv***
Performance-based errors	**17**	
Typos (neighbouring keys, addition, deletion, insertion, replacement)	17	förb**ä**ttra → *förb'ttra*
Across one word (phrases & sentences)	**28**	**se en** bild → ***sen** bild*


Spelling error?
no
Feedback action (FA): Mark as correct
yes
Is it a real word?
no
Tip 1: Unknown word. Please, check it again
Initial error statement
yes
Does the error occur with characters with diacritics? (å, ä, ö)
yes
Tip 2: Pay attention to the specific Swedish letters (å, ä, ö)
no
Does the error depend upon a consonant doubling?
yes
Tip 2: Pay attention to the consonant doubling
no
Does the error occur in a consonant cluster? (list)
yes
Tip 2: Pay attention to alternative spellings of the sound [xx]: x1, x2...
no
Error analysis

SPEED

SPElling Error Database

Advantages of collecting a corpus by applying this method: participants are quickly attracted, while cost, time and effort of collecting a corpus are reduced

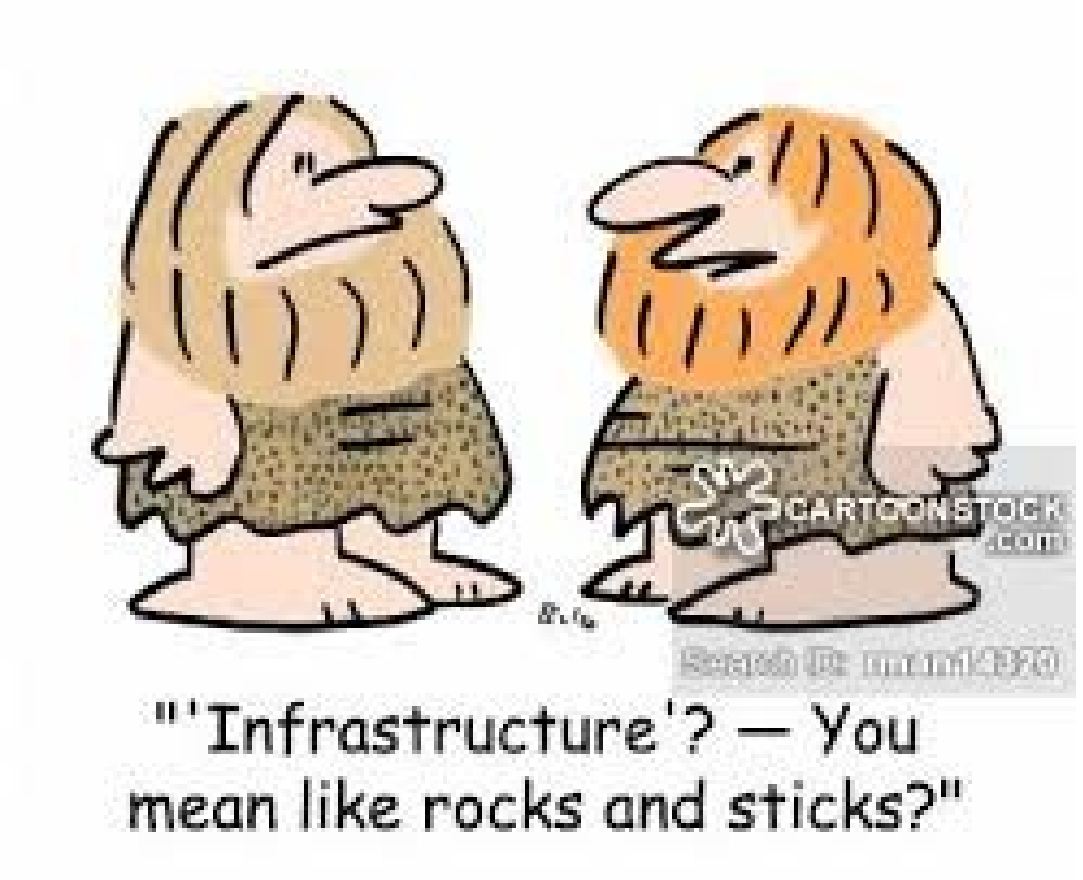
THIS is RESEARCH DATA!

And we need more of it!

L2 infrastructure –

a possible answer?

What is infrastructure?



An electronic research infrastructure

- (free accessible) data in electronic format
- technical platform for exploring data, including tools and algorithms for data analysis, and visualization
- a set of tools and technical solutions for new data collection and preparation, including data processing and annotation
- a network of experts in the relevant disciplines, incl. legal and ethical questions

How can it help?

- Collect and annotate data (L2 essays, error logs, course books ...)
- Develop tools for analyzing L2 data (e.g essays, reading comprehension texts)
- Set up and maintain applications/databases
- Gain expert knowledge

  ➔ to support research on L2 Swedish
  ➔ to support course book writers, L2 teachers, L2 assessors, L2 students
  ➔ to support instruction of future L2 teachers

Collecting data

Two types of data

- Produced **BY** L2 learners
  - → essays
  - → exercise logs
  - → errors
  - → (interviews)

- Produced by experts **FOR** L2 learners
  - → reading comprehension texts
  - → exercises
  - → (recordings of listening excerpts)

Challenges:
L2 learner-produced data

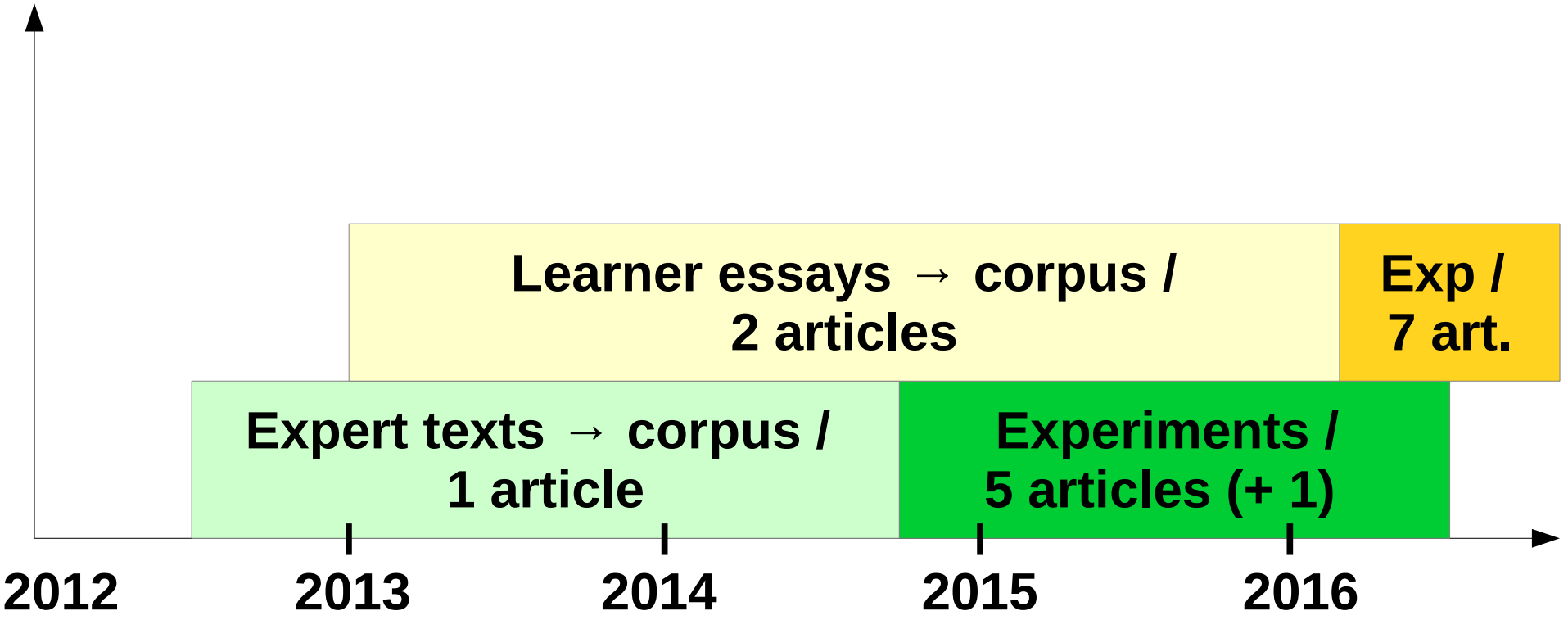
- Electronic L2 essays/logs are very difficult to collect
  - → NOT available online
  - → Need learner permits / copyright
  - → Need learner variables (gender, age, L1) / personal privacy act
  - → Sensitive in nature / anonymize
  - → Those who have it – don't want or CAN'T share
- We need an infrastructure/environment for storing and collecting L2 data
  - → same variables for comparison
  - → same student – same ID, etc

Challenges:
expert texts for L2 learners

- Electronic coursebooks
  - → NOT available online
  - → Aren't shared by publishing houses
  - → Need to be selected, bought, OCR-ed, proof-read, etc
  - → Can't be shared for copyright reasons
- We need extra information added (MANUALLY)
  - → text genres, topics, levels of difficulty
  - → exercise types, formats, target skills and competences

Curios “time & effort” fact:

Data vs experiments



L2 essay pre-processing

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

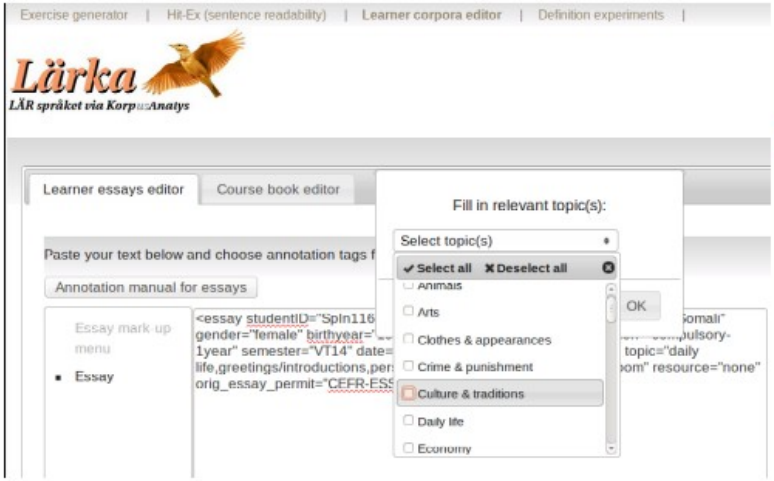
Essay variables
- Assigned CEFR level
- Essay setting (exam/home)
- Use of extra materials
- Academic term and date
- (Title, topic, genre, grade)

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 revel 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**
   * each illegible letter → replace with @
   * stricken text → leave out

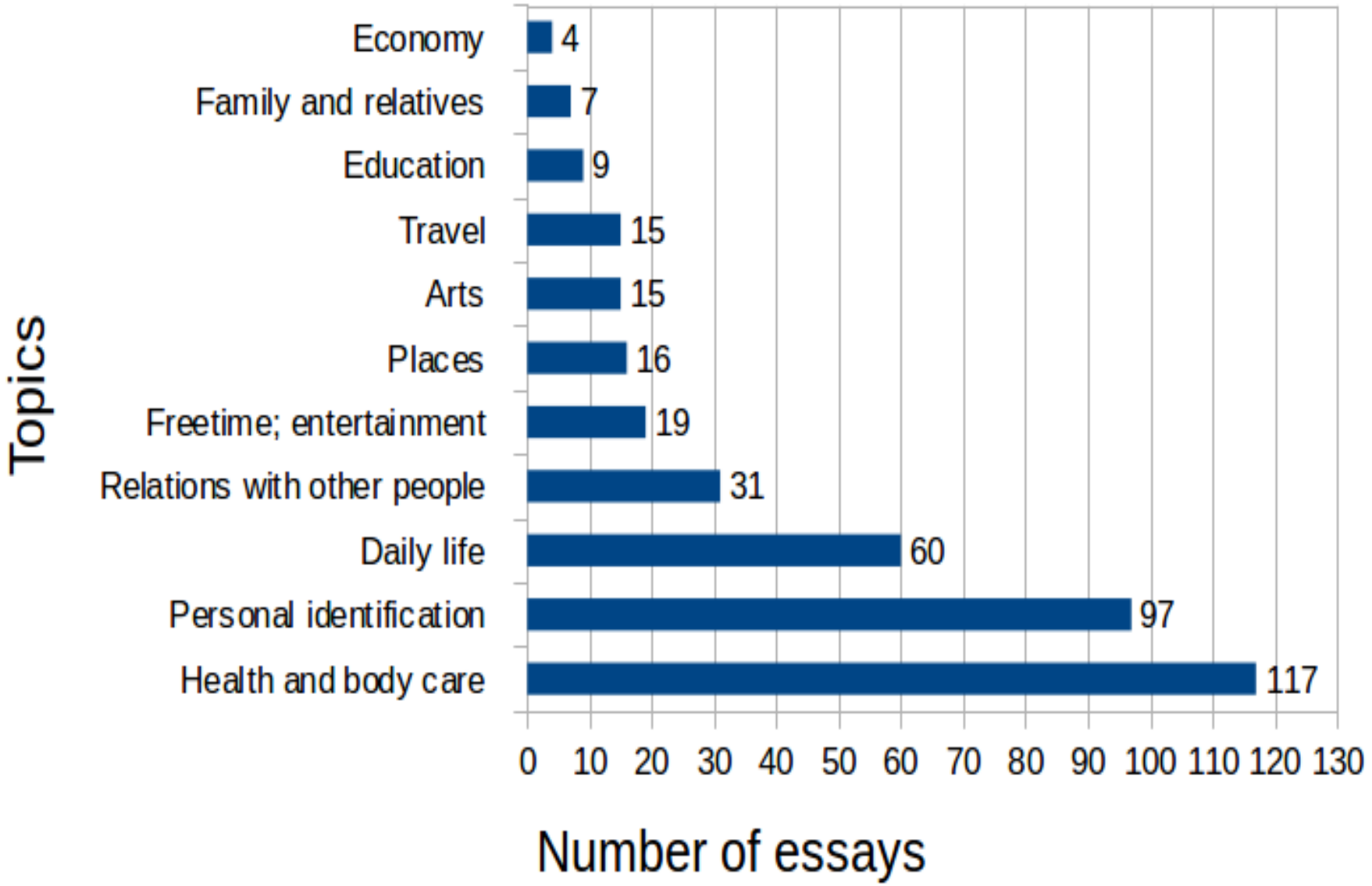


* *http://spraakbanken.gu.se/sites/spraakbanken.gu.se/files/tillstand eng-24042013 v03.pdf*

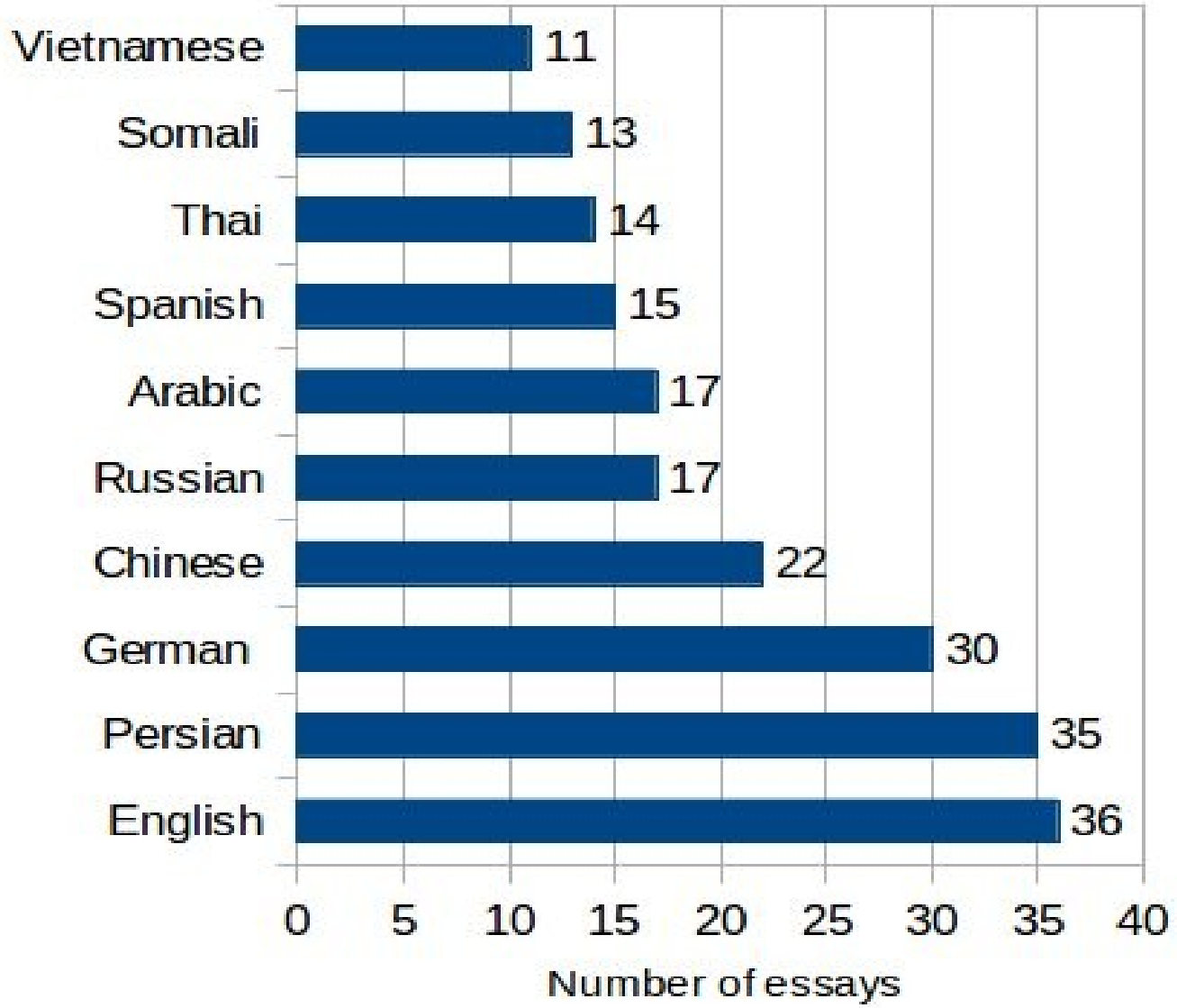
SweLL corpus

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**

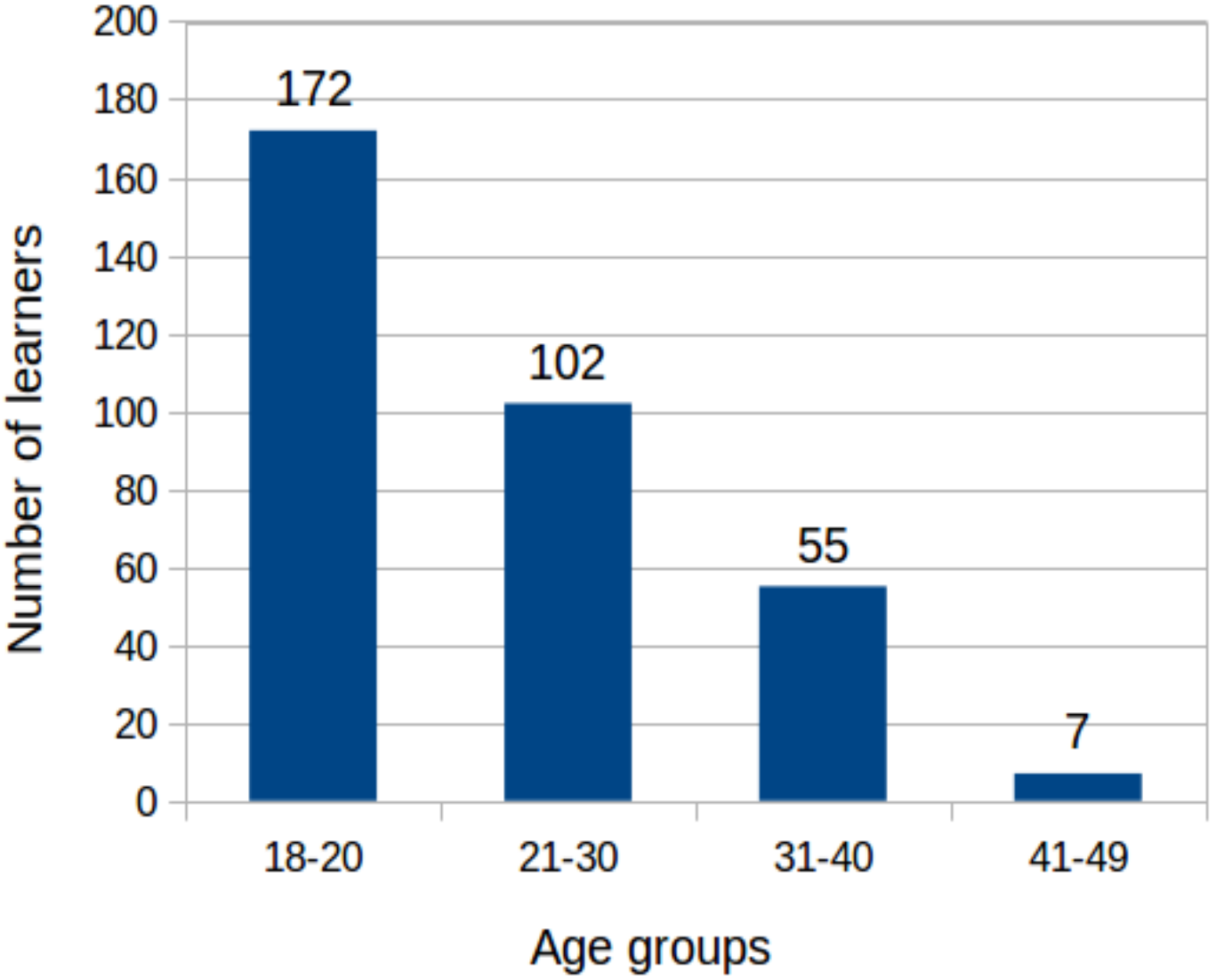
SweLL corpus: topics



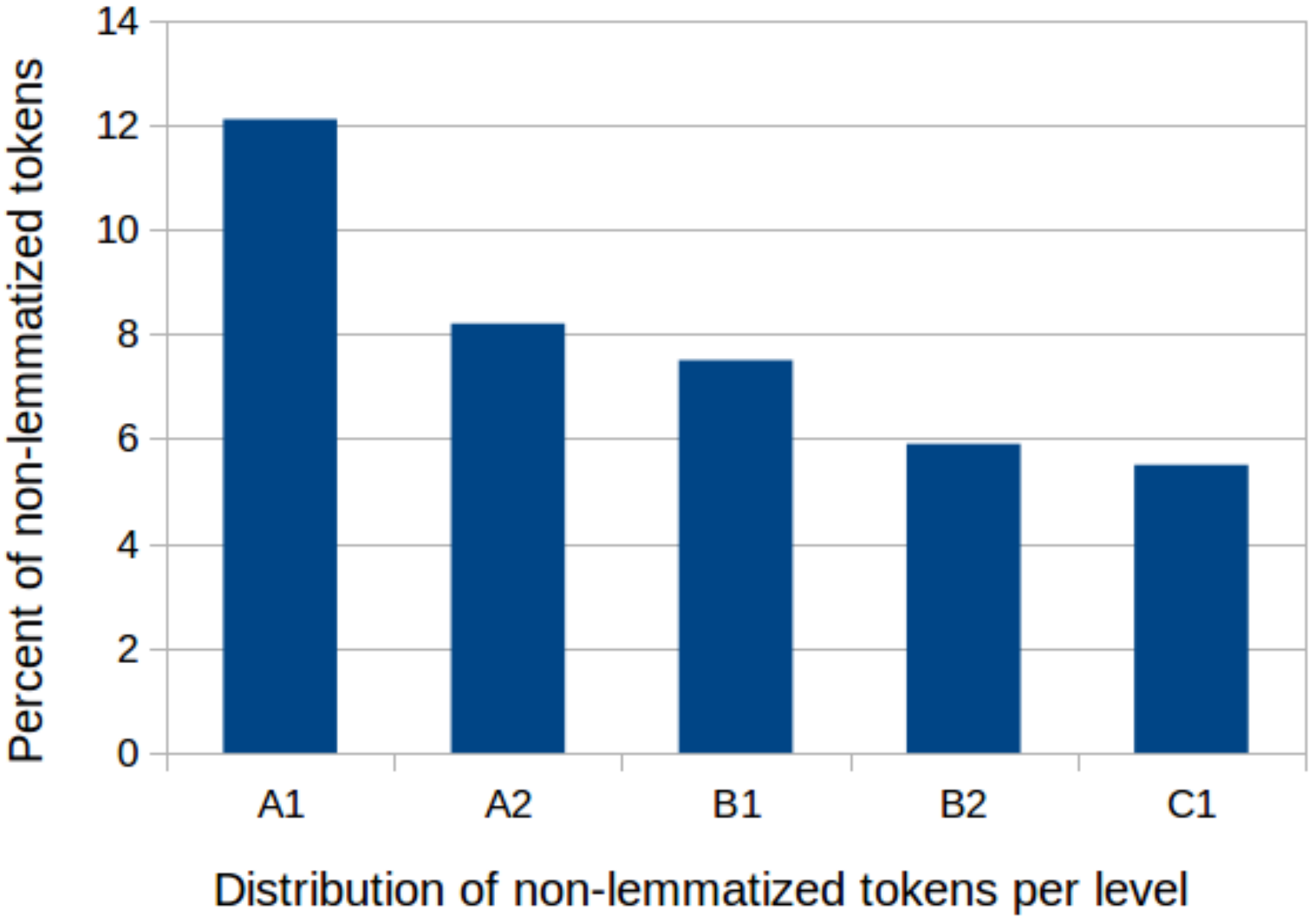
SweLL corpus: L1s



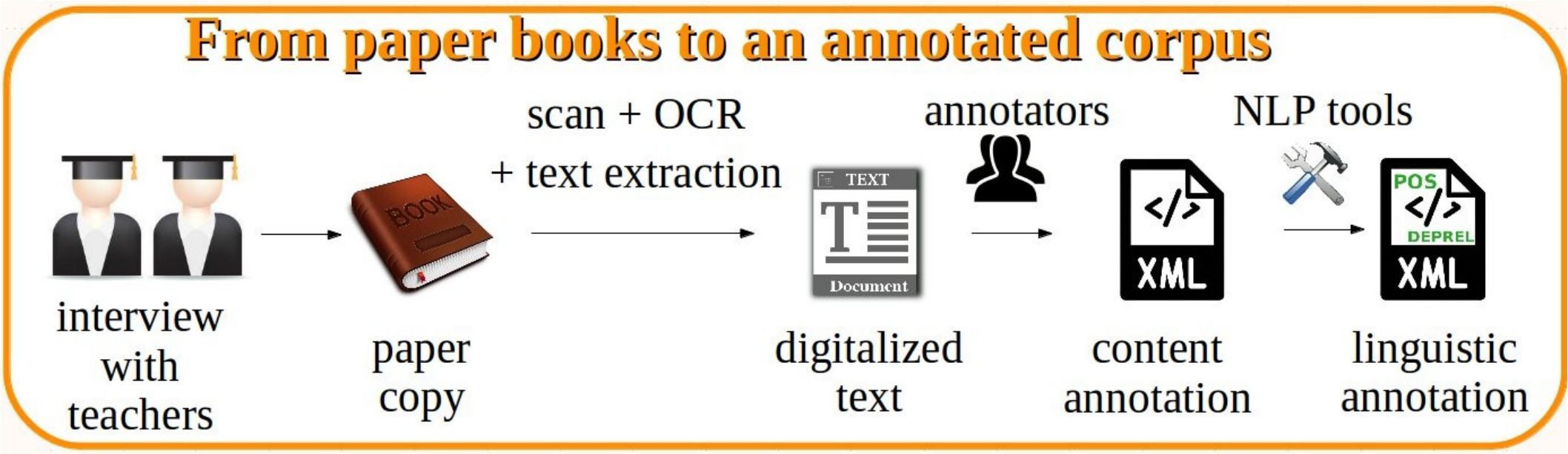
SweLL corpus: age

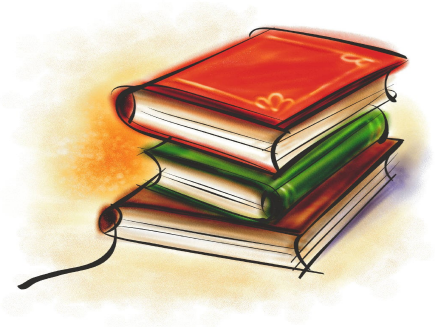


SweLL corpus: non-lemmatized items



COCTAILL corpus

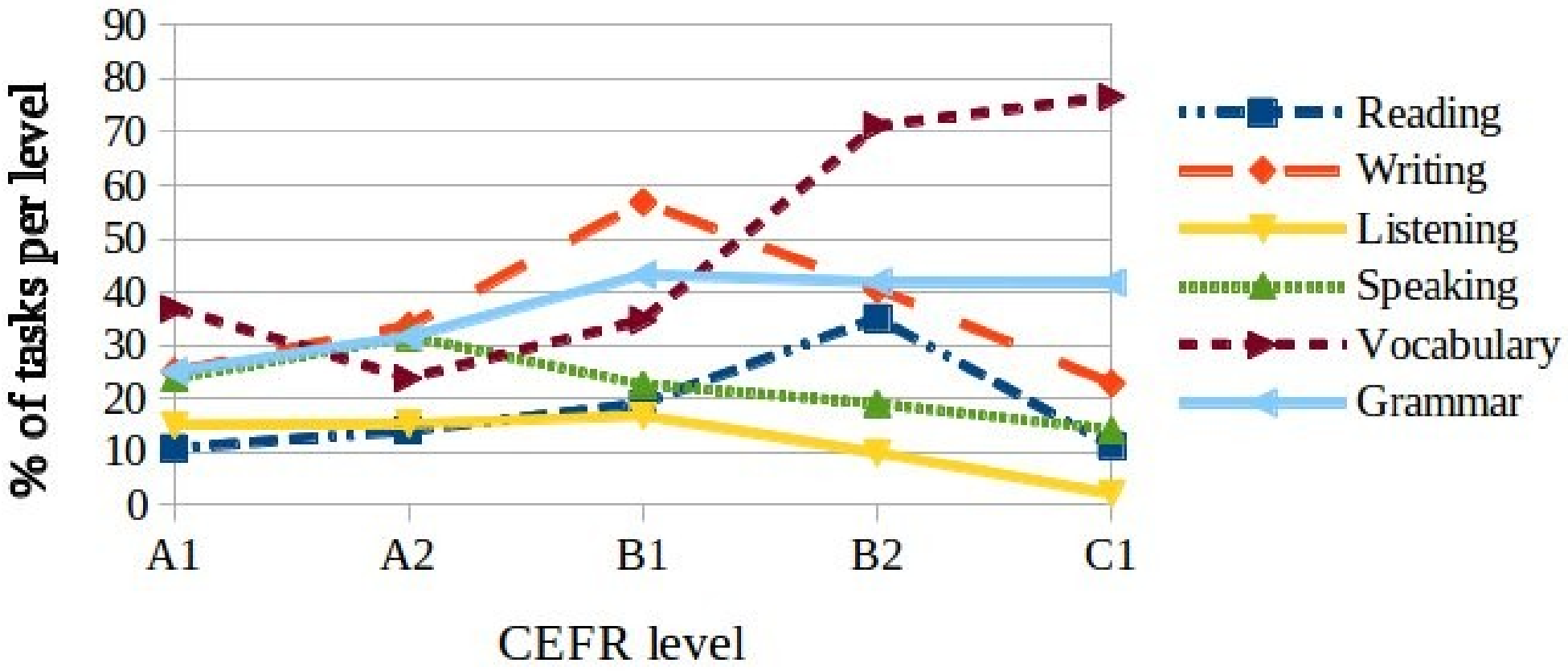




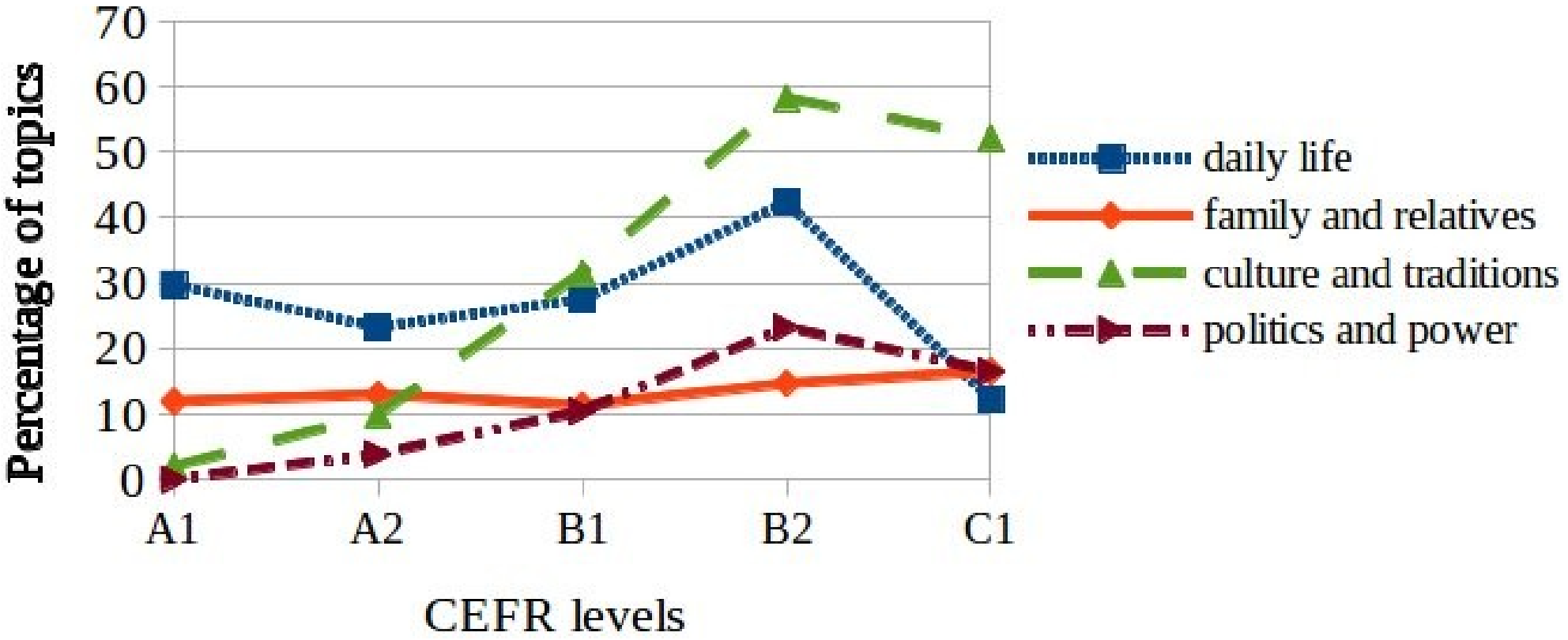
COCTAILL “ingredients”

ELEMENTS	ATTRIBUTES
Coursebook	title, author etc.
└ Extras	
└ Lesson	CEFR level, (title)
└ Text	topic (e.g. arts), (title)
└ Genre	genre family (e.g. facts), macrogenre (e.g. dialogue)
└ Subheadings	
└ Activity instruction	skills/competences, activity type/format
└ Task	
└ List	skills/competences, linguistic unit
└ Language Example	
Common attributes:	id, ref

COCTAILL quantitative explorations:
target skills across levels

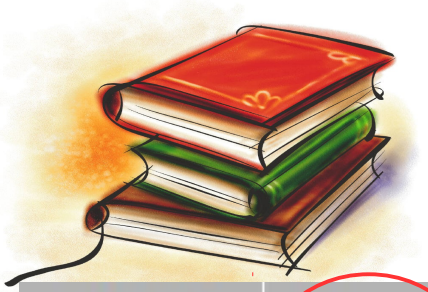


COCTAILL quantitative explorations:
topics across levels



GÖTEBORGS UNIVERSITET

CLT



COCTAILL

CEFR level	Nr. of books	Nr. of authors	Nr. of lessons	Nr. of texts	Nr. of tasks	Nr. of sentences (texts)	Nr. of tokens (texts)
A1	4	10	37	101	160	1581	11132
A2	4	10	105	232	244	4217	37259
B1	4	12	83	345	389	6510	79402
B2	4	8	31	314	368	8527	101583
C1	2	2	22	115	333	5085	71991
Total	18 (12 titles)	42 (26 different names)	278	1106	1494	25920	301367

Exploiting data

GÖTEBORGS UNIVERSITET

SVALex
L2 receptive vocabulary

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

user: englishprofile
password: vocabulary

GÖTEBORGS UNIVERSITET

SweLLex
L2 productive vocabulary

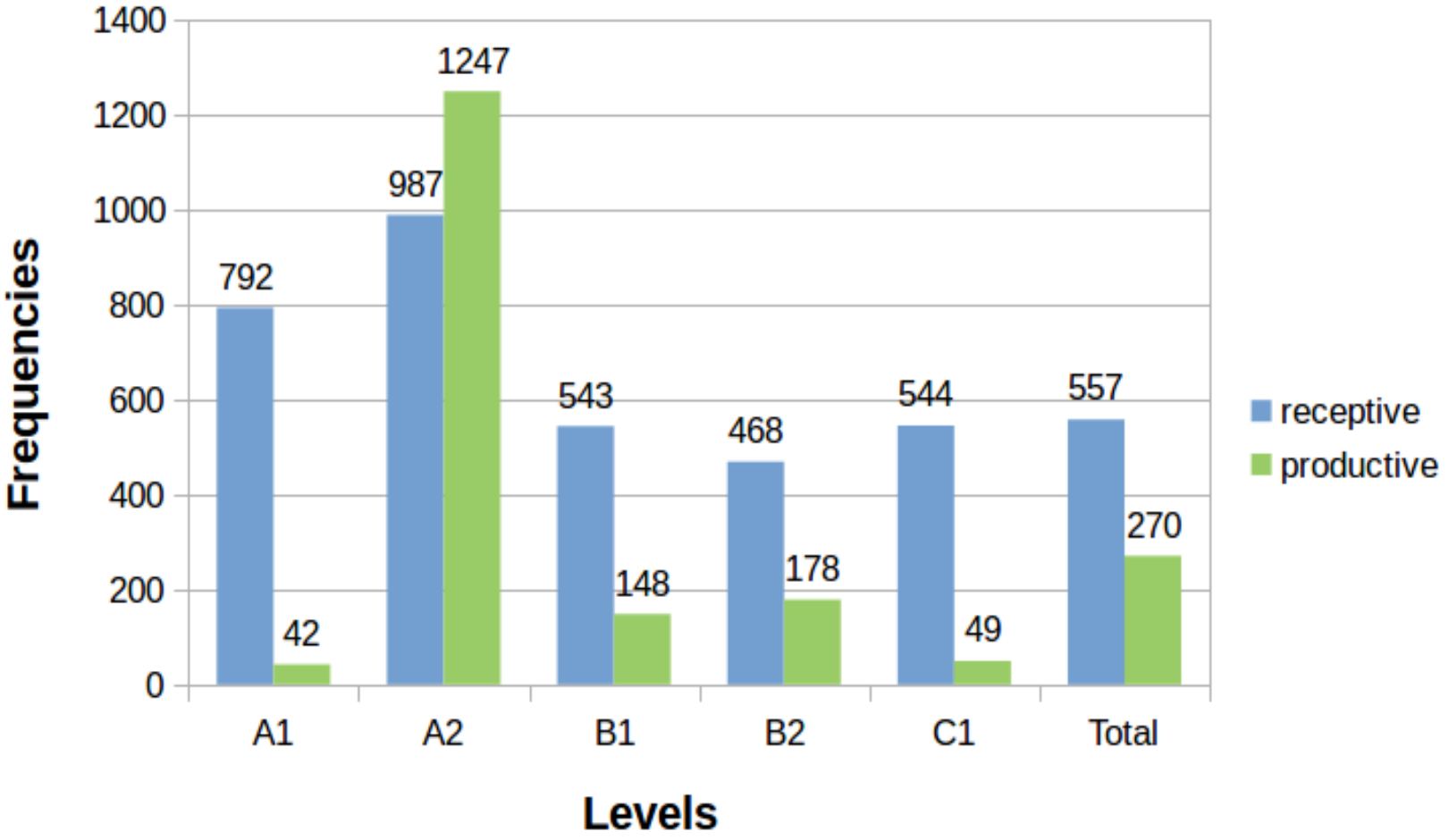
Lev	#items	#new	#MWE	#hapax	Doc.hapax examples	#SVALex	#EVP
A1	398	398	15	0	-	1,157	601
A2	1,327	1,038	82	12	*i kväll* "tonight"	2,432	925
B1	2,380	1,542	206	36	*fylla år* "have birthday"	4,332	1,429
B2	2,396	959	264	58	*fatta beslut* "make a decision"	4,553	1,711
C1	3,566	1,545	430	152	*sätta fingret*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

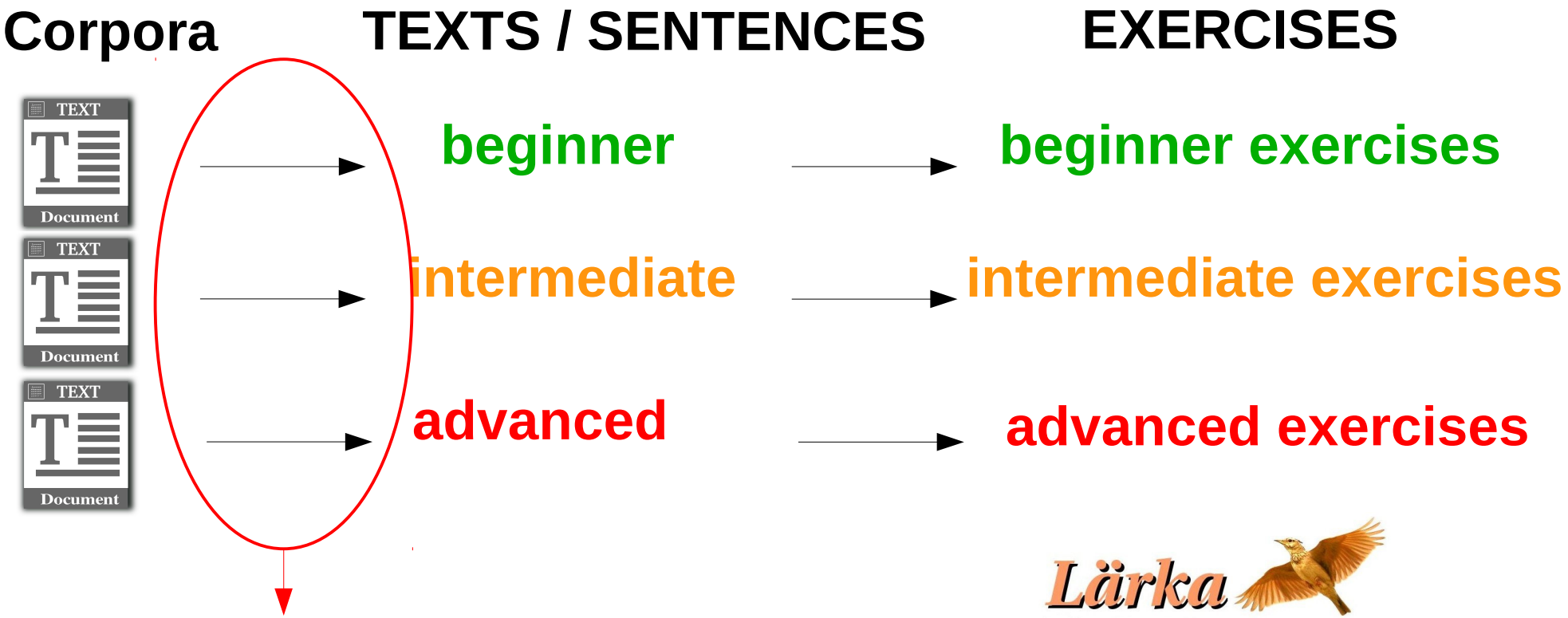
Studera (Eng. “study”)



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

Material generation for L2 learners

- **Aim:** reuse corpora to support language learning



Lärka

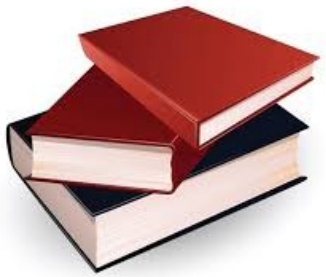
How can we automatically assess linguistic *complexity* (readability) on text or sentence level?

Readability experiments

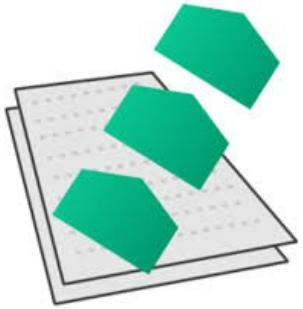
- Machine learning methods for automatic classification (WEKA, scikit-learn)
- Text- and sentence level
- Based on COCTAILL corpus
- 5 CEFR levels (A1-C1)
- 61 different linguistic features

Ildikó Pilán

From course books to automatic CEFR level assessment



Course books
(training data)

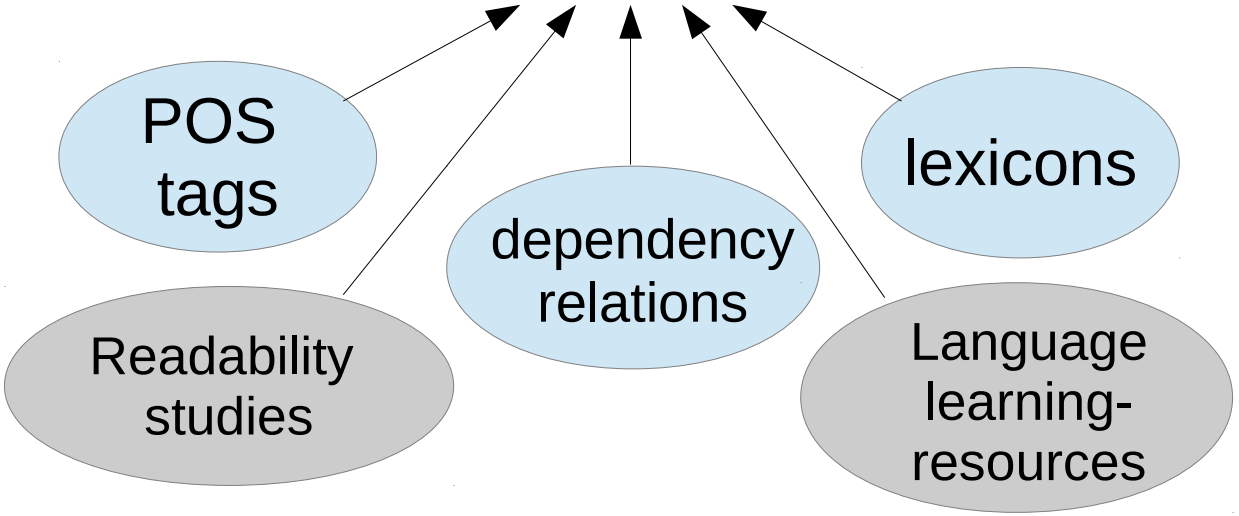


features

Machine learning
(supervised
training)

trained
classifier

80% correct
(texts)
63% correct
(sentences)



Features

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 tool for text evaluation

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

Exercise Generator | Hit-Ex | Learner Corpora Editor | Text evaluation | Svenska | English

Language Acquisition Reusing Korp

Filmen hanlar om en pojke . Han heter NN och han gilla dansar ballet så mycket . När han har idrott leklion , brukor 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 stejjkar för att alla person måste jobba mycket hårt men lön inte mycket pengar för familj , mat och deras liv . De jätte 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

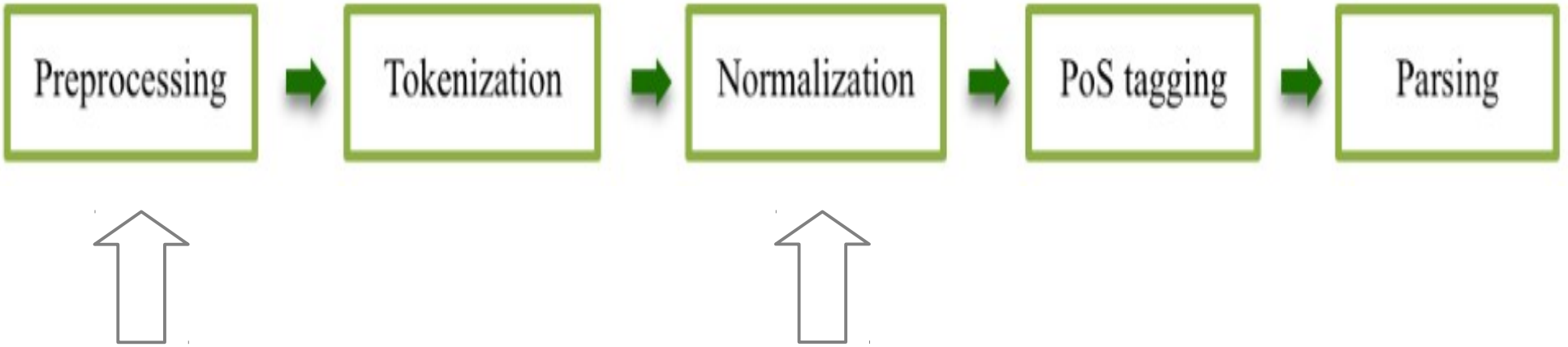
Detailed evaluation

Number of sentences	16
Number of tokens	189

David Alfter

Ildikó Pilán

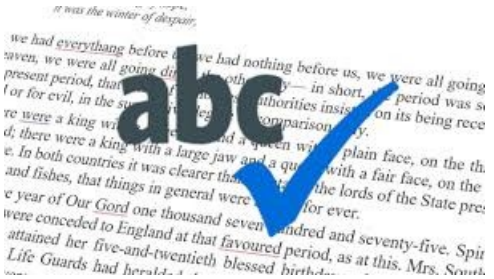
Automatic !?!? annotation

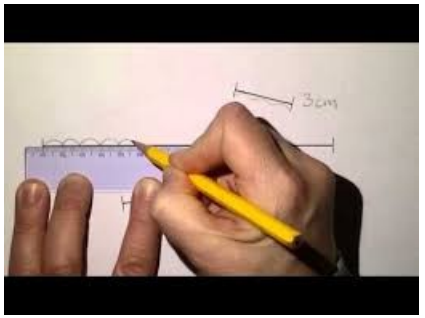


→ Segmentation

→ Single-word errors
(non-words vs real words)
→ Phrase-level errors
(grammar, combinability)
→ Word order errors

L2 word-level normalization

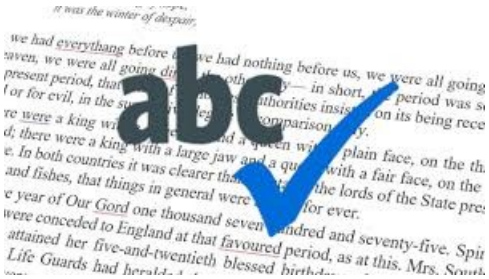




- **Levenstein distance (as is)**
  - Good for advanced levels (edit distance of 1)
  - Fails at lower levels (with multiple edits)

- **LanguageTool + candidate ranking**
  - 73% correct variant selection
  - Failed to identify 30% of spelling errors

Levenstein distance



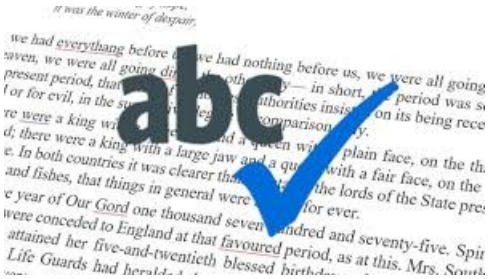
Level	Correct/total
A1	7/20
A2	13/20
B1	13/20
B2	15/20
C1	16/20

(1) substitution of one misspelled letter, e.g.: ursprang* → ursprung (origin);

(2) deletion of an extra letter, e.g.: sekriva* →skriva (to write), naman* → namn (name);

(3) insertion of one missing letter, i.e. sammanfata* → sammanfatta (summarize).

LanguageTool + ranking



	# tokens	% tokens
A1	204	9.7
A2	1118	6.0
B1	1650	5.5
B2	3526	10.8
C1	7511	12.4

Number corrected tokens per level

scores for noun-verb and noun-adjective combinations included with a threshold of LMI ≥ 50

L2 “alternative” data

- Logs – acc. to a defined research interest
- Steps:
    - Implement an activity for learners
    - Prepare database for storing (structured) data
    - Implement a way to browse logs, visualize statistics etc
    - If necessary – add extra annotation steps (manual, automatic)

Pilot 1 on L2 “alternative” data

- Identifying most predictive features for a language proficiency level (for diagnostic purposes)
    - Multi-word expressions
    - Syntactic properties (e.g. word order)
    - Knowledge of word morphology (e.g. inflections)

David Alfter

L2 “alternative” data (logs)

LÄR språket via Korpus Analys

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

Pilot 2 on L2 “alternative” data

- Automatic assigning new words to a proficiency level
    - We predict the level automatically
    - Learners (of a known level) get the word in an exercise (or a series of exercises)
    - We see whether learners can cope with it

David Alfter

Ildikó Pilán

L2 “alternative” data (logs)

LÄR språket via Korpus Analys

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

Word guess

Tries: 0/7

Score: 0

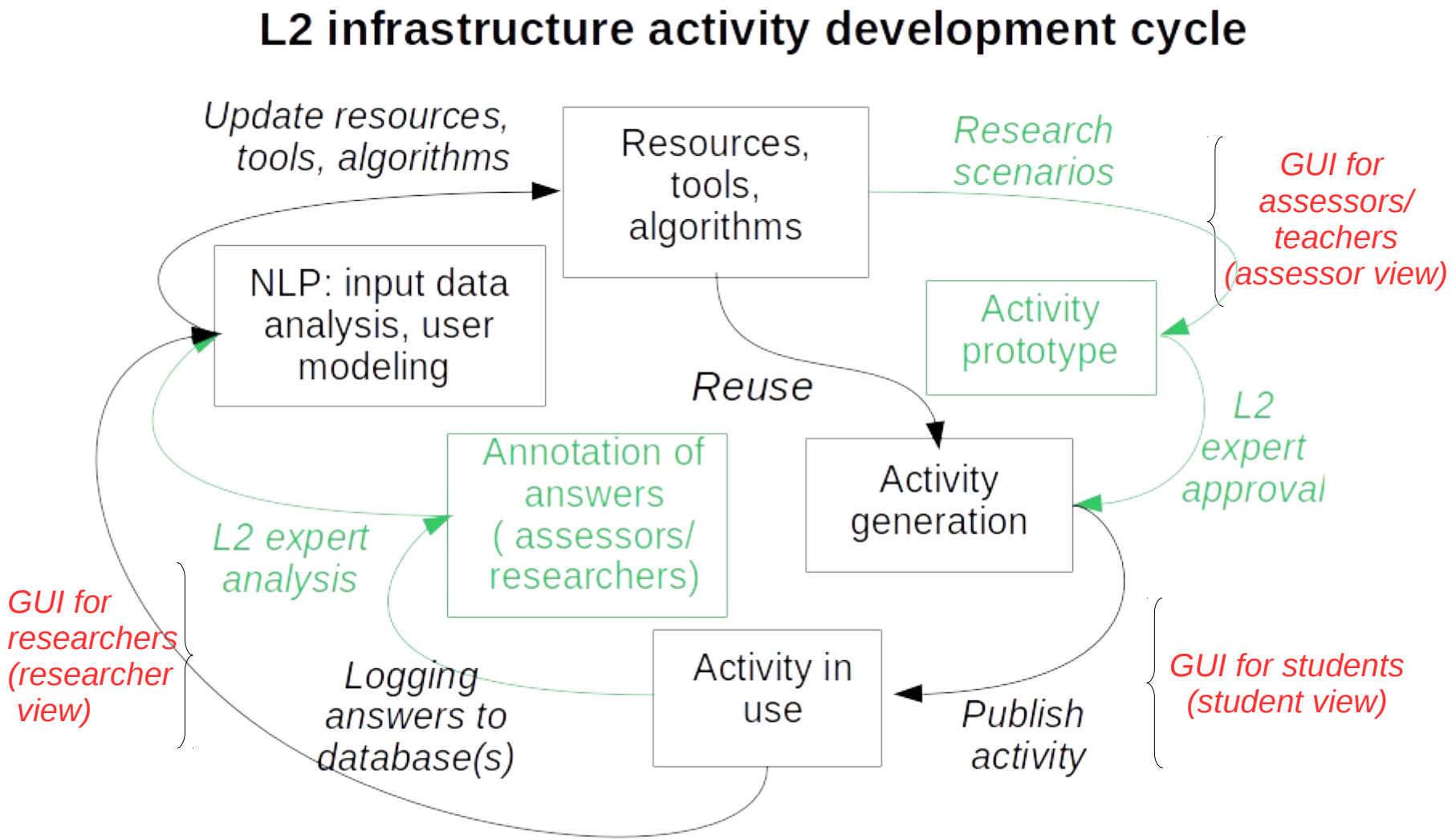
Definition:
rör sig tyst och försiktigt för att inte bli upptäckt, tassar (pres ind aktiv)

Hjälp:
slip; steal, sneak, creep

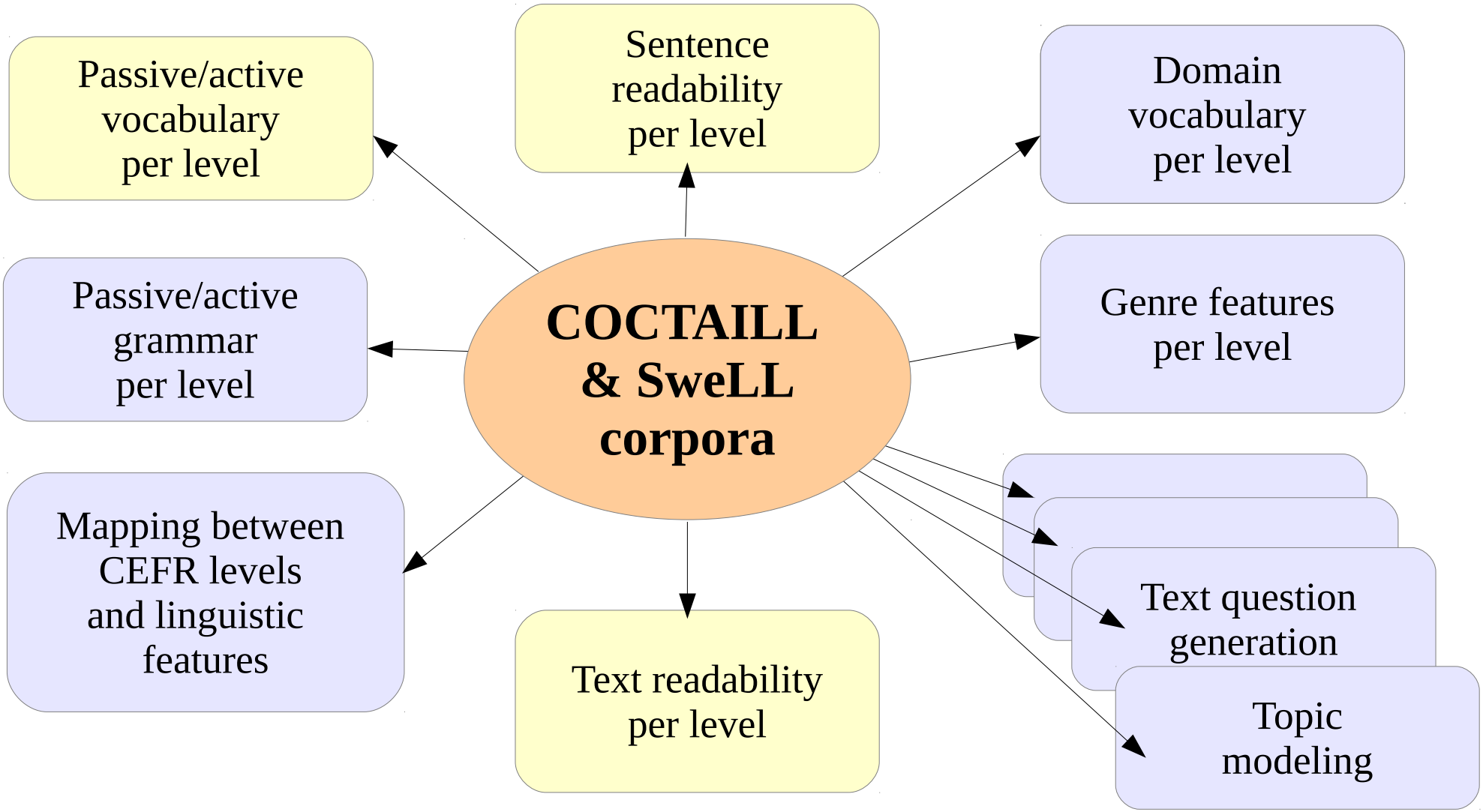
A B C D E F G H I J K L M N O P Q R S T U V W X
Y Z Ö Ä Å

Evaluation
Reliability of tools

The ultimate goal



L2 data for Lärka's research agenda



EESTI
RAKENDUSLINGVISTIKA
ÜHING ESTONIAN ASSOCIATION OF APPLIED LINGUISTICS

SWE-CLARIN

Språk-
BANKEN

Thank you!