A Friend in Need?

Research agenda for electronic Second Language infrastructure

Elena Volodina, Beata Megyesi, Mats Wirén,

Lena Granstedt, Julia Prentice, Monica Reichenberg,
Gunlög Sundberg

SLTC 2016, Umeå

What is infrastructure?

"'Infrastructure'? — You mean like rocks and sticks?"

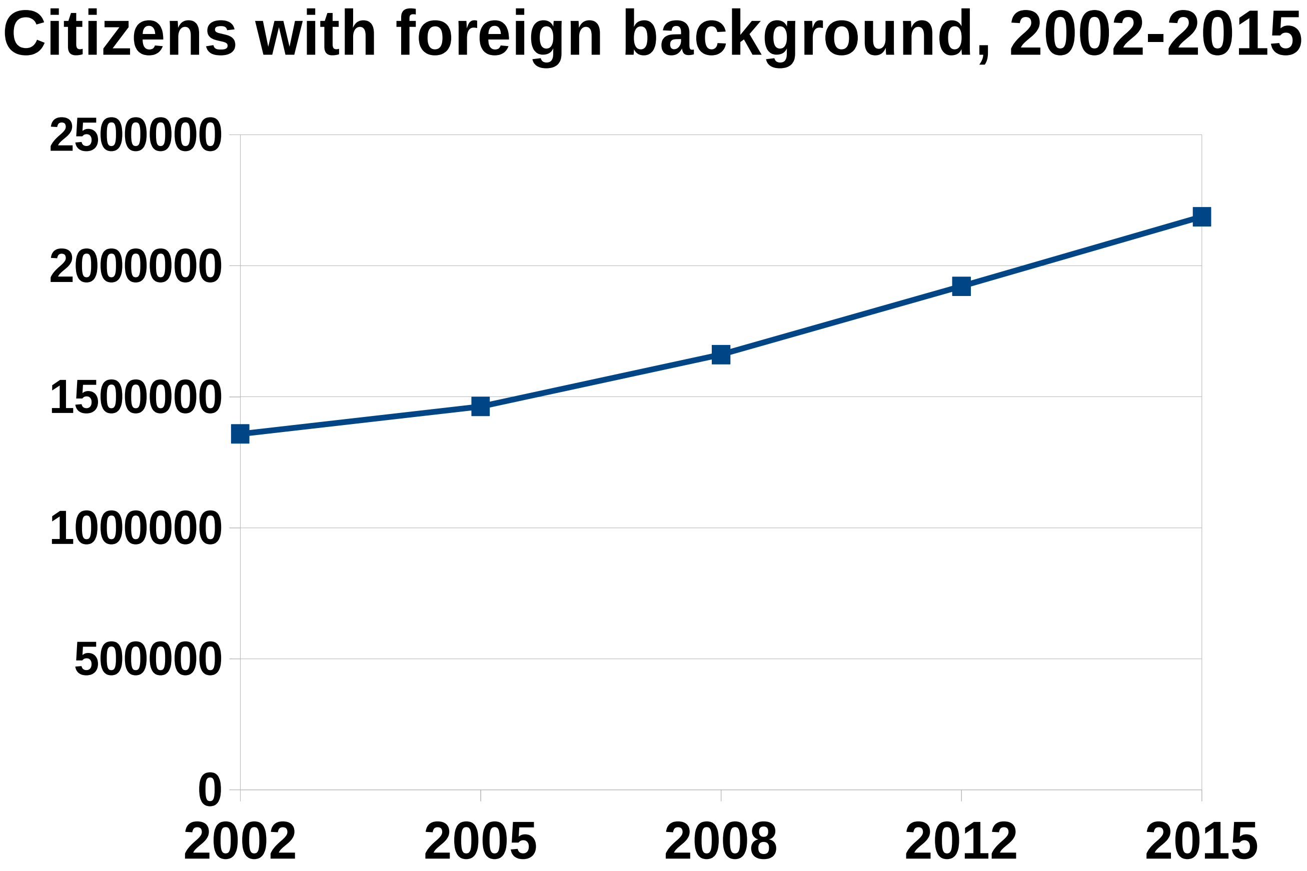
An electronic research infrastructure

- (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

Key terminology

SweLL	**Swe**dish **L**earner **L**anguage
L2	Second (and foreign) language

Societal need



2015: out of **9,9 mln** citizens, **2,2 mln** have foreign background. i.e. **22,2 %**
(Statistiska centralbyrån)

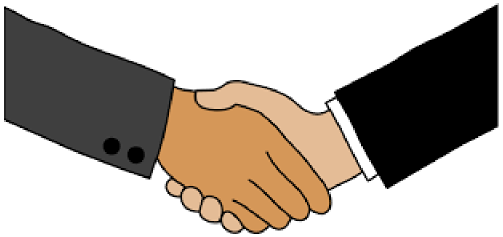
How can we help?

- Collect and annotate data (L2 essays, error logs, ...)
- Develop tools for analyzing L2 data (e.g essays)
- 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

Partners

- University of Gothenburg: NLP, L2, assessment
- Stockholm university: NLP, L2
- Uppsala university: NLP
- Umeå university: L2/assessment



Guess what?

© Enache Dumitru Bogdan • www.free-cartoon-clipart.blogspot.com

- Riksbankens Jubileumsfond, infrastructure project IN16-0464:1

- 2017-2019

Our focus is on...

- L2 essays (writing)
- exercise logs (reading and listening comprehension, vocabulary and grammar training)

- NO speech data – yet

- target group: adult learners

Problem 1: lack of L2 data

- Electronic L2 production is very difficult to collect
  - ➔ NOT available online,
  - ➔ Need learner permits
  - ➔ Need learner variables (gender, age, native language, etc)
  - ➔ Sensitive in nature
- We need an infrastructure/environment for storing and collecting L2 data

Problem 2: lack of coordination

- There is a national need to coordinate various (individual and bigger-scale) efforts aimed at collecting L2 production (e.g. which permits, learner variables, formats etc so that the data could be comparable and usable between projects)

- There is a need to digitize and process hand-written L2 language samples (e.g. National tests in Swedish and L2 Swedish) in an organized nation-wide effort

Problems 3: lack of L2 tools and models

- Existing NLP tools are not capable to analyze L2 learner language due to numerous infelicities (normative language analysis versus error analysis)
  - ➔ Adaptation of existing NLP tools required
  - ➔ Adaptation of tools targeting ”deviating” forms of language: historical texts or social media
- Development of new tools require specific, often hand-annotated data
  - ➔ Error-tagged corpora
  - ➔ Learner profiles (grammar, vocabulary, etc. per level of proficiency)
- ...

Natural Language Processing for Language Learning

New activities for learners and teachers

SUPER-new resources and algorithms

Support of language skills for 21 century

Logging results and indidualizing learning

Writing support tools

Essay grading/classification

Error detection

Feedback generation

You name it....

Initial steps and pilot studies

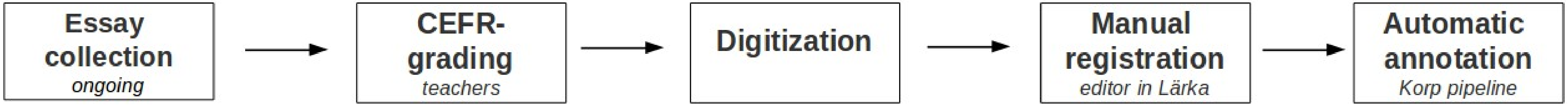
- **Data** collection and digitiation
  - SweLL corpus
  - The Uppsala Corpus of Student Writings
- **Resource creation** (e.g. SweLLex – L2 productive vocabulary)
- **Algorithm development**: L2 error normalization
- **User-oriented tools**:
  - L2 annotatoion pipeline: SweGRAM
  - L2 essay classification (Lärka-based online tool)

Data

SweLL corpus

core data

SweLL workflow



Learner variables

Collected through permits*

Student variables:

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

Essay variables

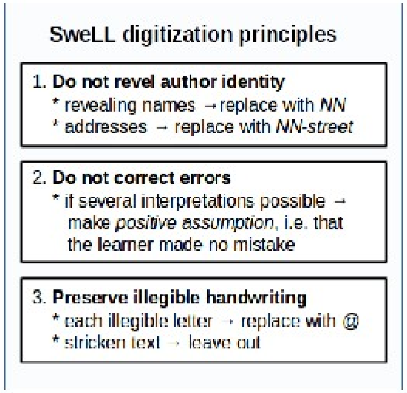
- Assigned CEFR level
- Essay setting (exam/home)
- Use of extra materials
- Academic term and date
- (Title, topic, genre, grade)

Assessors

Minimum of two trained assessors

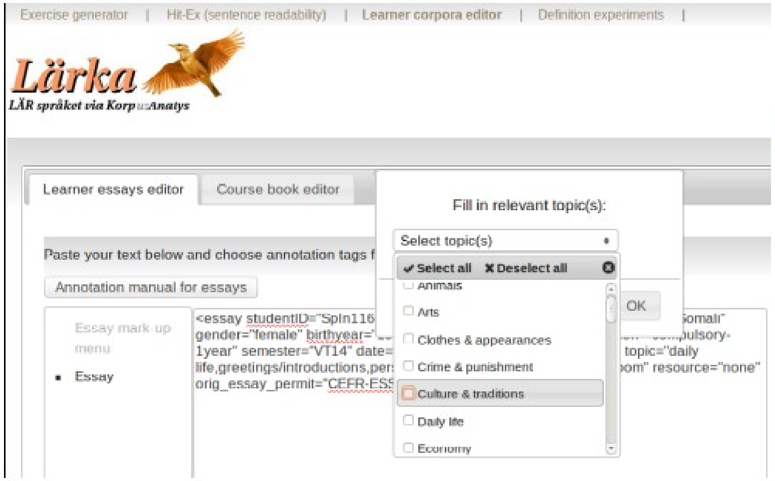
Inter-annotator agreement

- 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

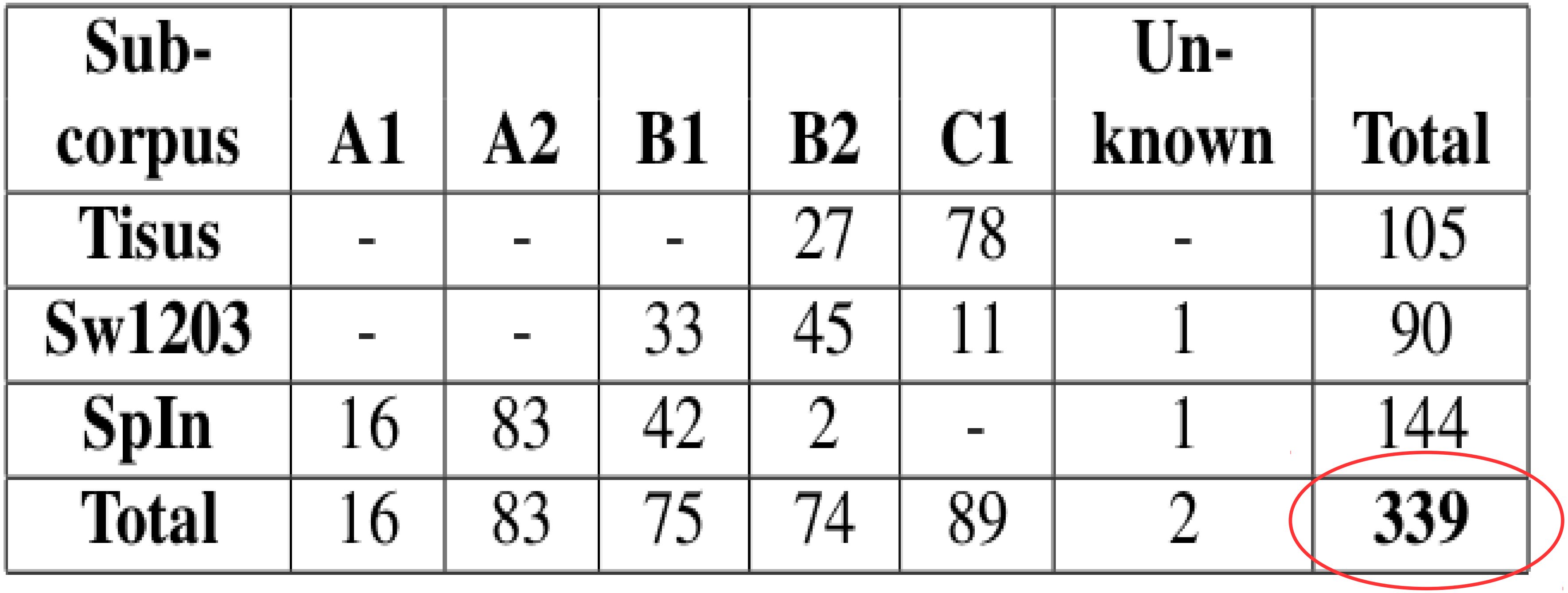

Exercise generator | Hit-Ex (sentence readability) | Learner corpora editor | Definition experiments |
Lärka
LÄR språket via KorpusAnalys
Learner essays editor
Course book editor
Paste your text below and choose annotation tags f
Annotation manual for essays
Essay mark-up menu
Essay
Fill in relevant topic(s):
Select topic(s)
✔ Select all ✘ Deselect all
Animals
Arts
Clothes & appearances
Crime & punishment
Culture & traditions
Daily life
Economy
OK

```
<essay studentID="SpIn116                    Somali"
gender="female" birthyear="                  compulsory-
1year" semester="VT14" date=                 topic="daily
life,greetings/introductions,per             om" resource="none"
orig_essay_permit="CEFR-ESS
```

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

SweLL corpus

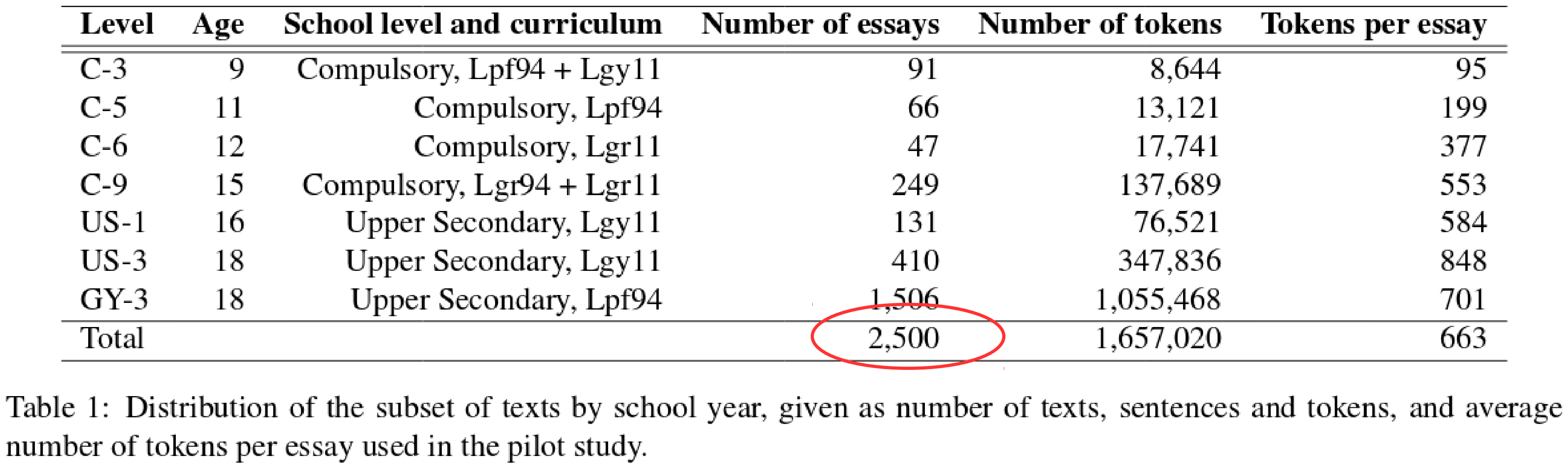
core data



Sub-corpus	A1	A2	B1	B2	C1	Unknown	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**

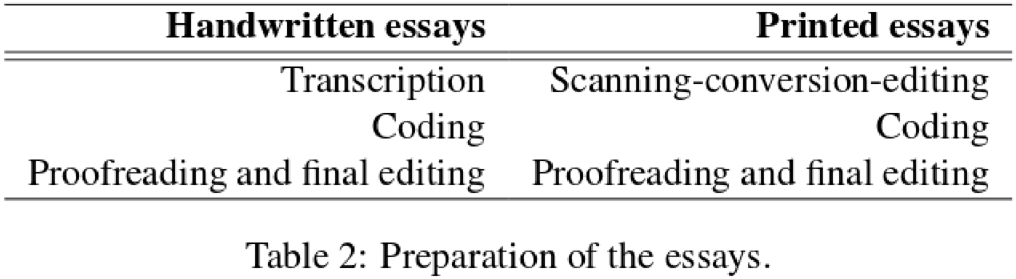
The Uppsala Corpus of Student Writings

reference 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.

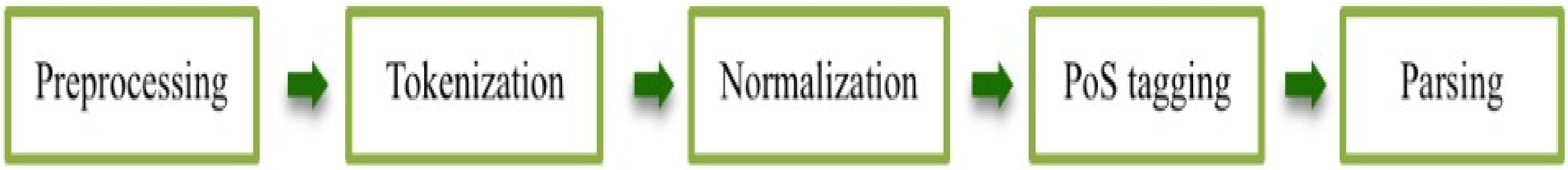


Handwritten essays	Printed essays
Transcription	Scanning-conversion-editing
Coding	Coding
Proofreading and final editing	Proofreading and final editing

Table 2: Preparation of the essays.

The Uppsala Corpus of Student Writings

reference corpus



Preprocessing → Tokenization → Normalization → PoS tagging → Parsing

Next steps for the core corpus

- Error annotation

- Normalized version(s), i.e. hypotheses what a learner has meant

Resources

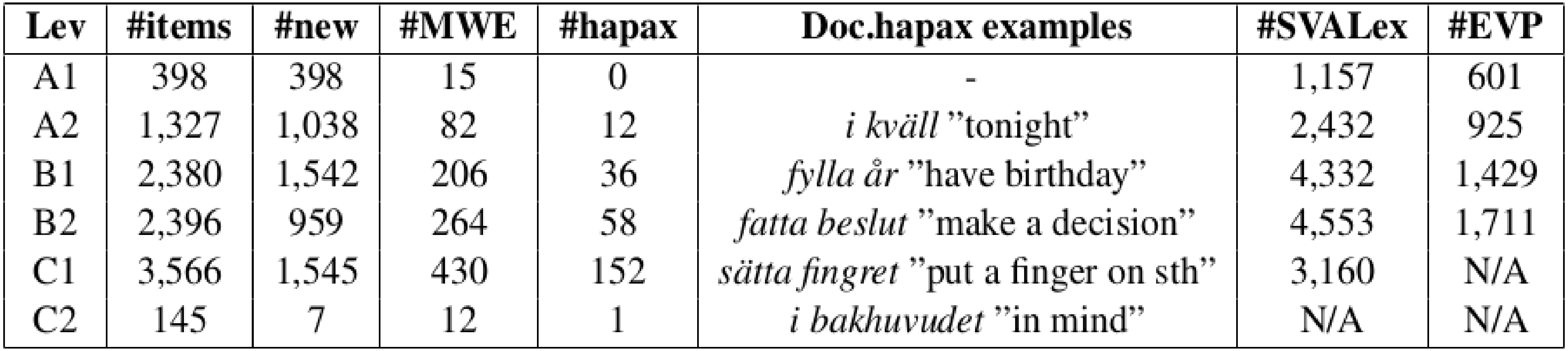
re·sound·ing
loud: resounding
tory. —re·sound·
re·source
n. 1. Something
library is a
drawn upon
difficult or
means of
total means
cluding
ments.

GÖTEBORGS UNIVERSITET

CLT

Språkbanken

SweLLex
productive L2 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

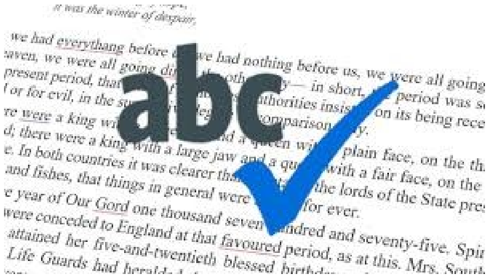
Total 5,475

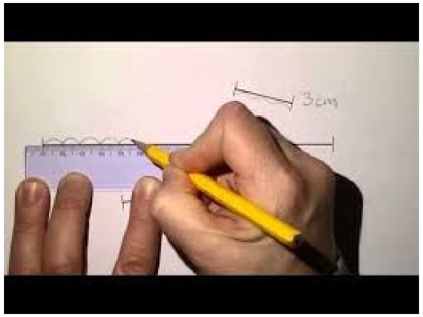
http://cental.uclouvain.be/svalex/

Algorithms

```
inp_array[a], c) &&
inp_array[a], use_wystepuje
[b.length - 1].b, inp_array));
a.reverse(); b = m(a, " "); -1 <
m(a, void 0); -1 < b && a.splice(b,
< b && a.splice(b, 1); return a; }
rn a.replace(RegExp(",", "g"), " "); }
b) { for (var c = 0, d = 0;d < b.length
a && c++; } return c; } function m(
-1, d = 0;d < a.length;d++) { if
break; } } return c; } function
rie", b = 1; "-" === a[0] &&
```

L2 word-level normalization





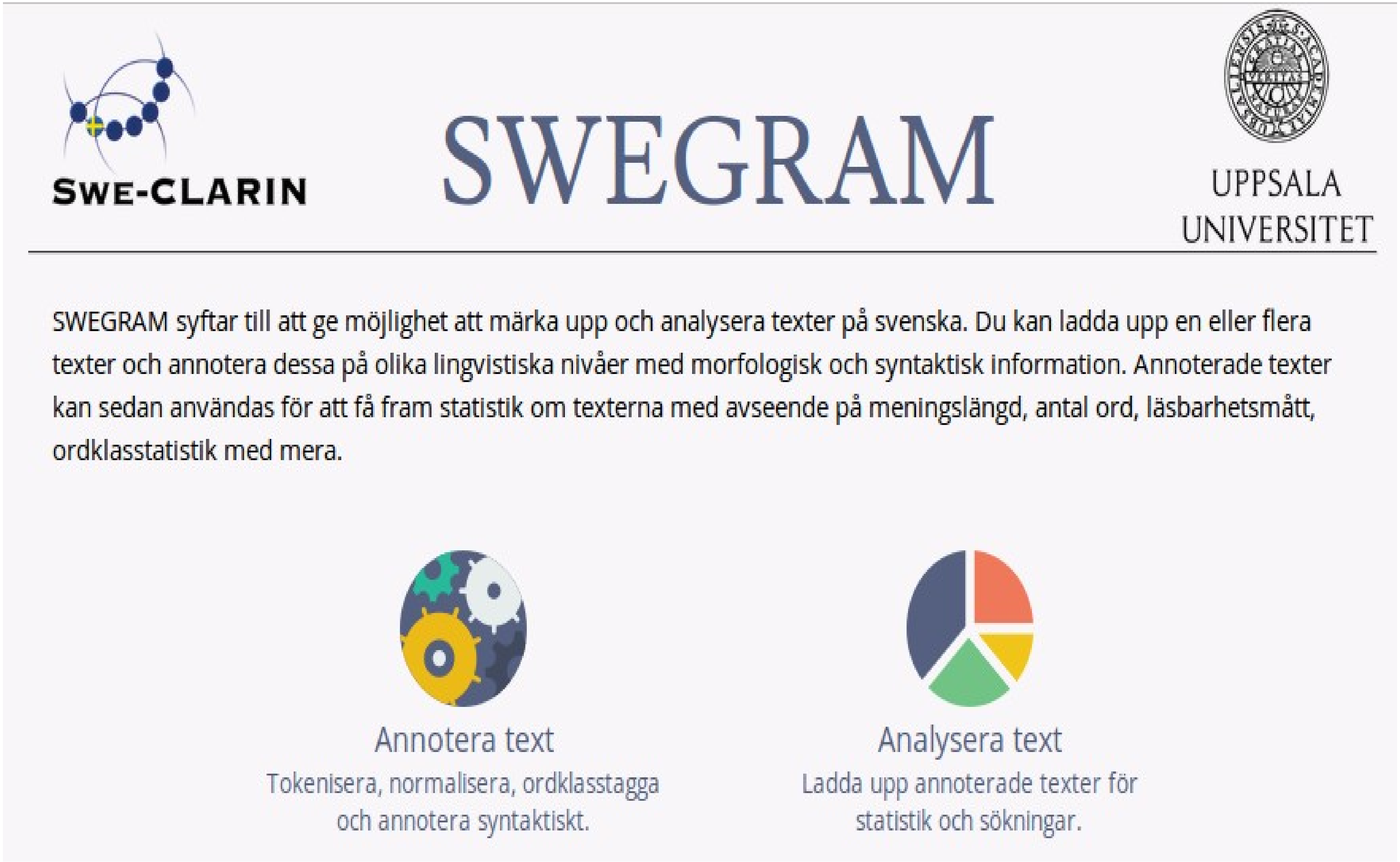
- **Levenstein distance (as is)**
  - Good for advanced levels (edit distance of 1)
  - Fails at lower levels (with multiple edits)
- **Levenstein distance (for historical texts)**

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

Next steps for tool development

- Normalization on phrase level, etc
- Error detection (need to identify which types to target first)
- ...

User-oriented tools



SWEGRAM

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

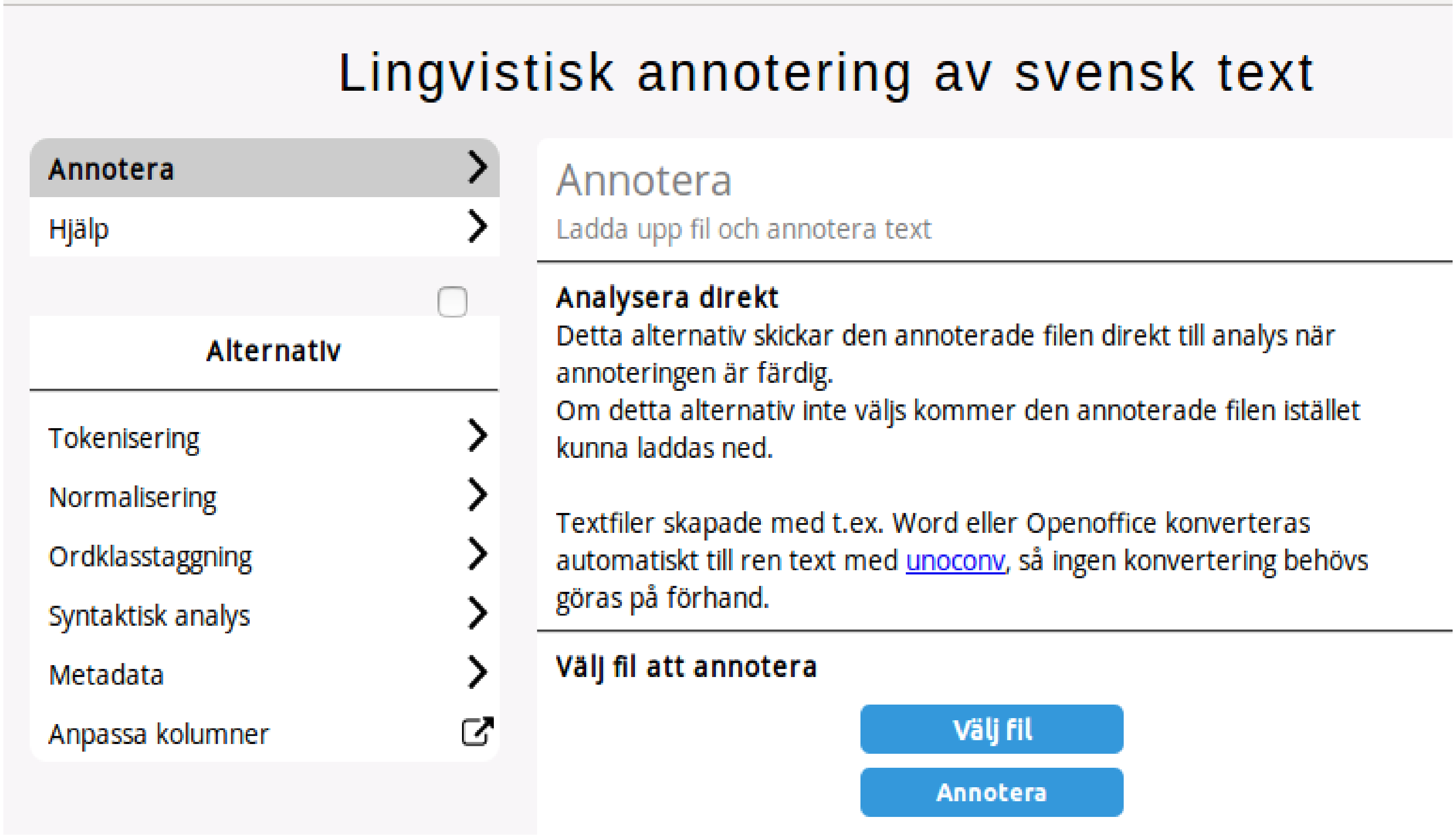
Tokenisera, normalisera, ordklasstagga och annotera syntaktiskt.

Analysera text

Ladda upp annoterade texter för statistik och sökningar.

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

SweGRAM annotation pipeline



Annotera

Ladda upp fil och annotera text

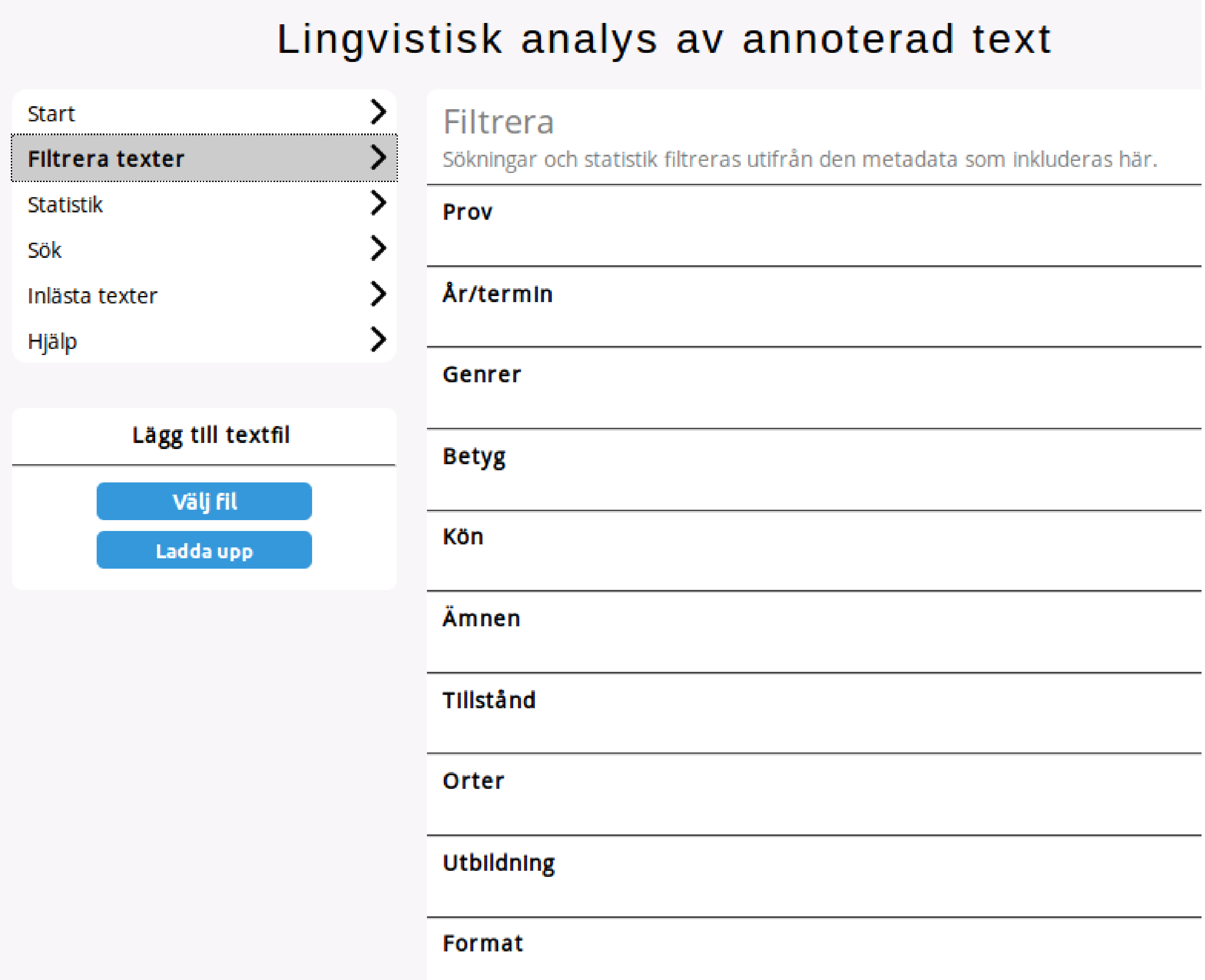
Analysera direkt

Detta alternativ skickar den annoterade filen direkt till analys när annoteringen är färdig.
Om detta alternativ inte väljs kommer den annoterade filen istället kunna laddas ned.

Textfiler skapade med t.ex. Word eller Openoffice konverteras automatiskt till ren text med unoconv, så ingen konvertering behövs göras på förhand.

Välj fil att annotera

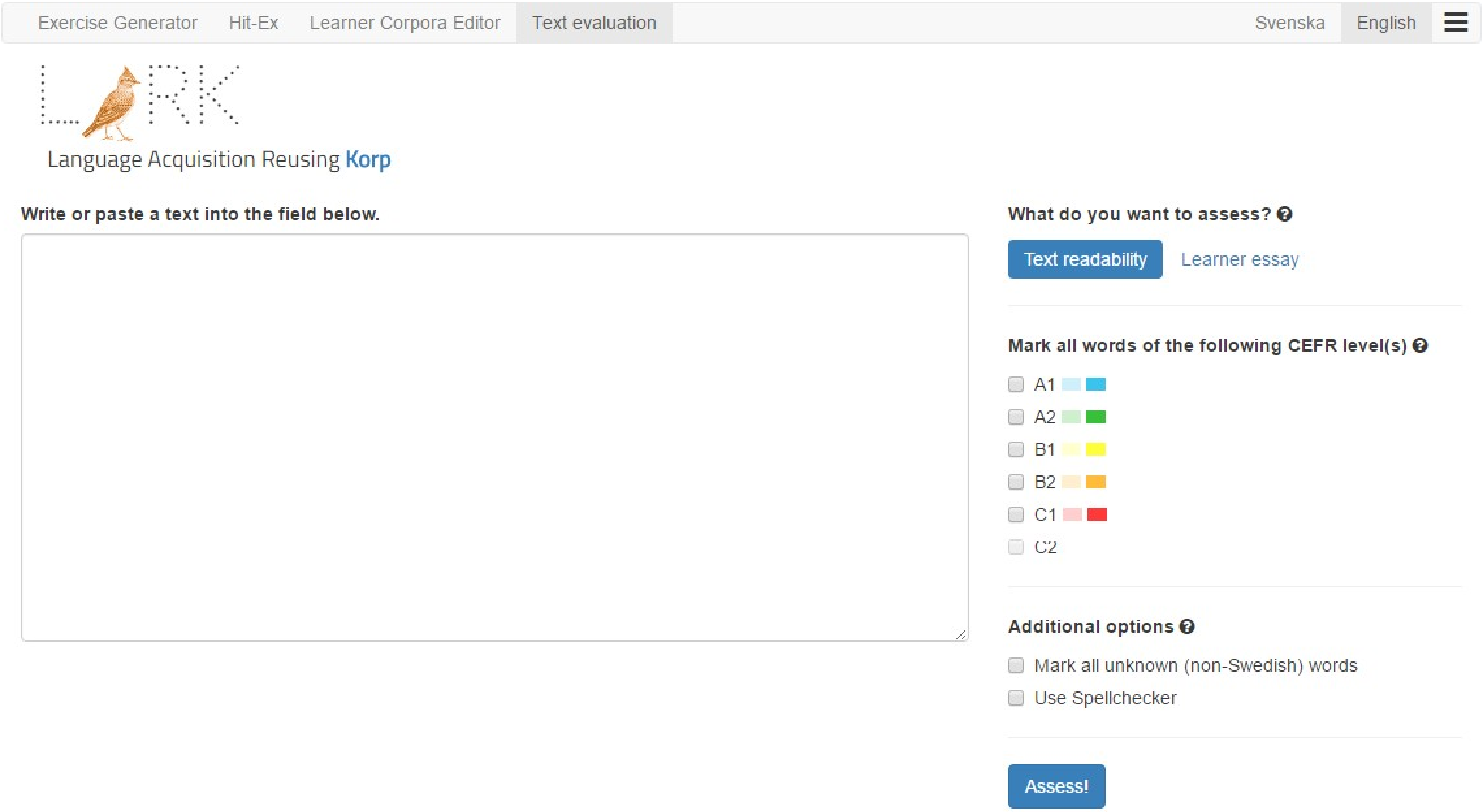
SweGRAM exploration tool



Filtrera

Sökningar och statistik filtreras utifrån den metadata som inkluderas här.

L2 text assessment in CEFR terms



Write or paste a text into the field below.

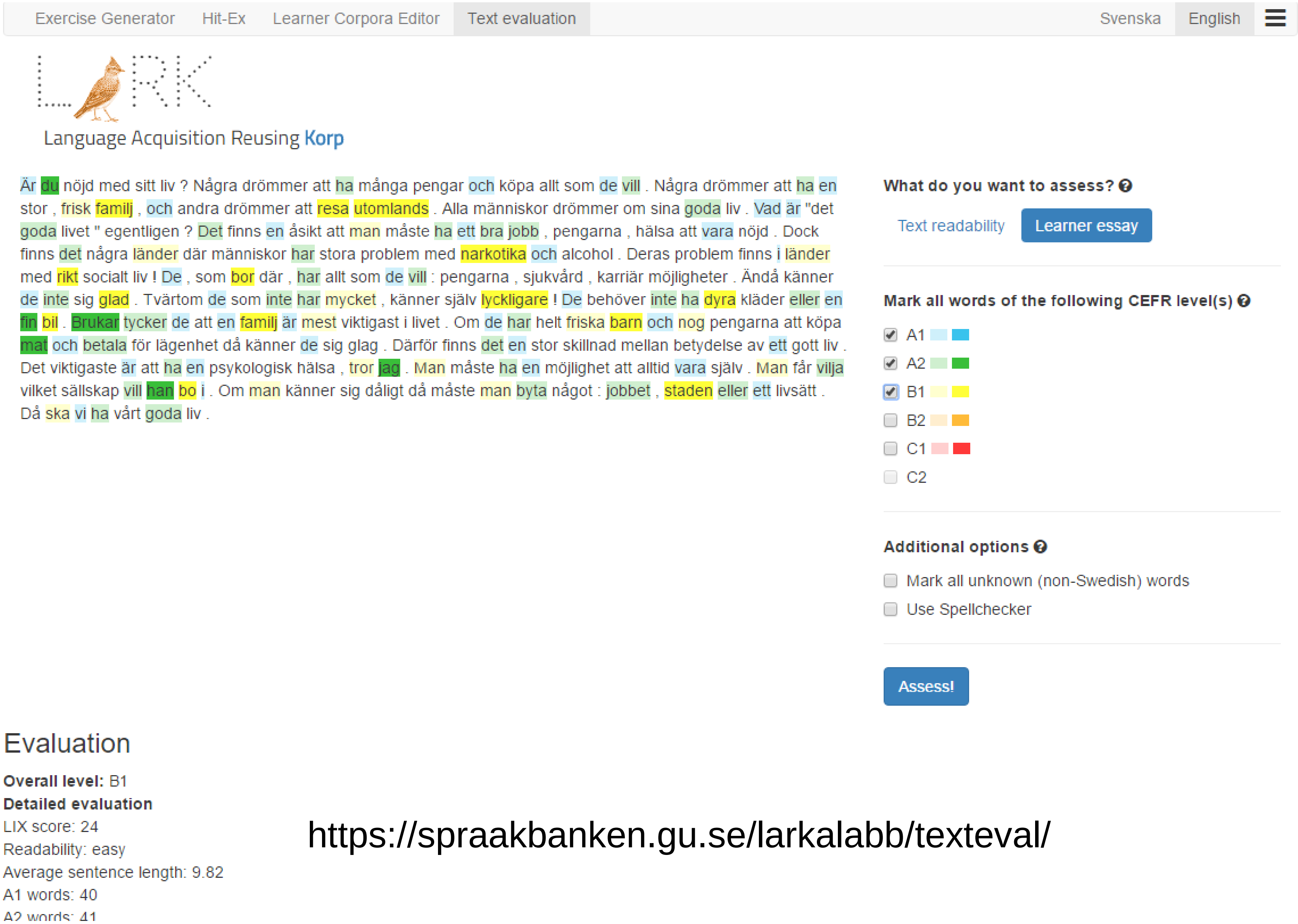
What do you want to assess?

Mark all words of the following CEFR level(s)

- □ A1 🟦🟦
- □ A2 🟩🟩
- □ B1 🟨🟨
- □ B2 🟧🟧
- □ C1 🟥🟥
- □ C2

Additional options

- □ Mark all unknown (non-Swedish) words
- □ Use Spellchecker



Är du nöjd med sitt liv ? Några drömmer att ha många pengar och köpa allt som de vill . Några drömmer att ha en stor , frisk familj , och andra drömmer att resa utomlands . Alla människor drömmer om sina goda liv . Vad är "det goda livet " egentligen ? Det finns en åsikt att man måste ha ett bra jobb , pengarna , hälsa att vara nöjd . Dock finns det några länder där människor har stora problem med narkotika och alcohol . Deras problem finns i länder med rikt socialt liv ! De , som bor där , har allt som de vill : pengarna , sjukvård , karriär möjligheter . Ändå känner de inte sig glad . Tvärtom de som inte har mycket , känner själv lyckligare ! De behöver inte ha dyra kläder eller en fin bil . Brukar tycker de att en familj är mest viktigast i livet . Om de har helt friska barn och nog pengarna att köpa mat och betala för lägenhet då känner de sig glag . Därför finns det en stor skillnad mellan betydelse av ett gott liv . Det viktigaste är att ha en psykologisk hälsa , tror jag . Man måste ha en möjlighet att alltid vara själv . Man får vilja vilket sällskap vill han bo i . Om man känner sig dåligt då måste man byta något : jobbet , staden eller ett livsätt . Då ska vi ha vårt goda liv .

What do you want to assess?

Text readability | **Learner essay**

Mark all words of the following CEFR level(s)

- ☑ A1 🟦
- ☑ A2 🟩
- ☑ B1 🟨
- ☐ B2 🟧
- ☐ C1 🟥
- ☐ C2

Additional options

- ☐ Mark all unknown (non-Swedish) words
- ☐ Use Spellchecker

Assess!

Evaluation

Overall level: B1
Detailed evaluation
LIX score: 24
Readability: easy
Average sentence length: 9.82
A1 words: 40
A2 words: 41

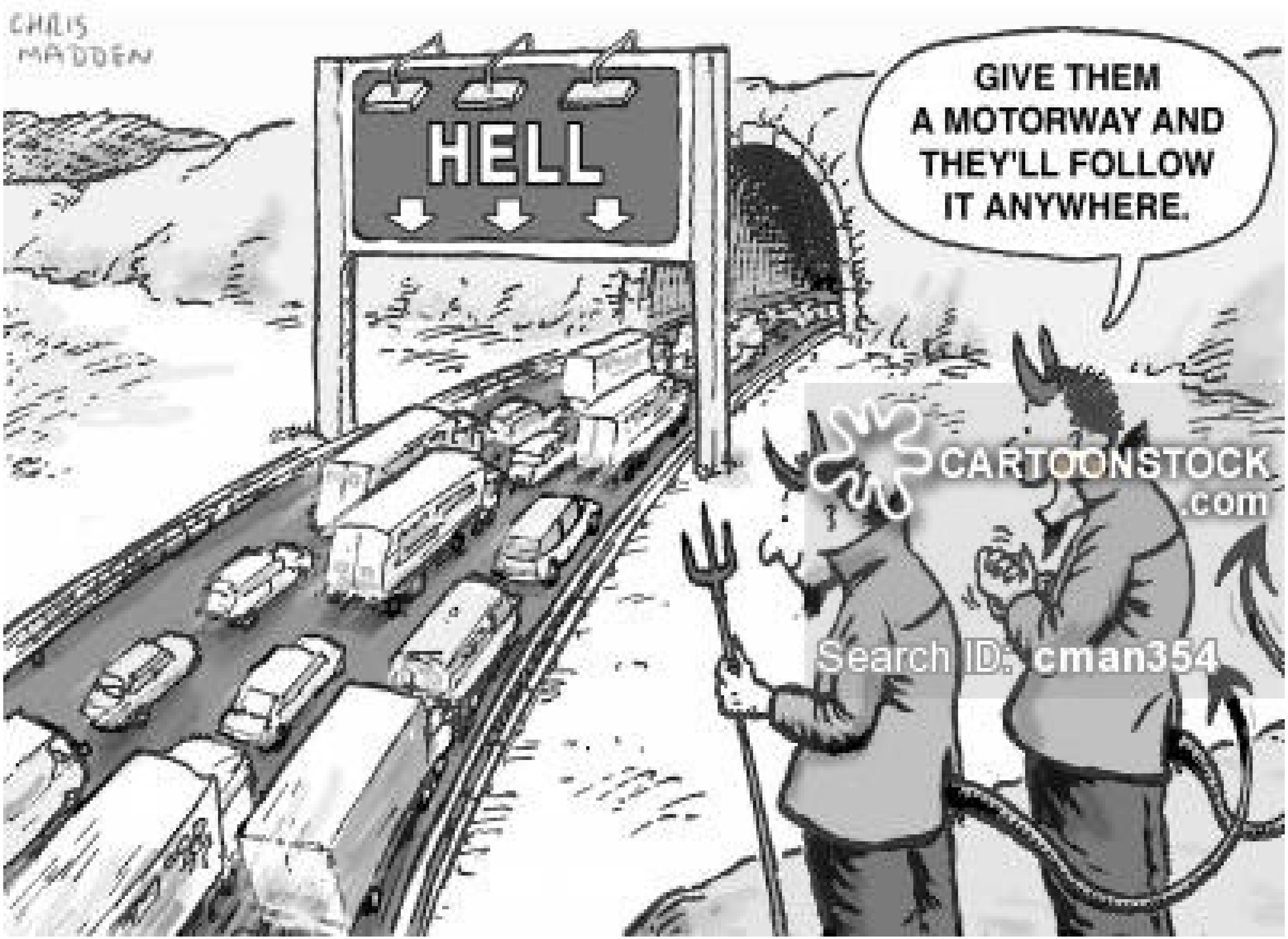
https://spraakbanken.gu.se/larkalabb/texteval/

Next step - reliability of tools

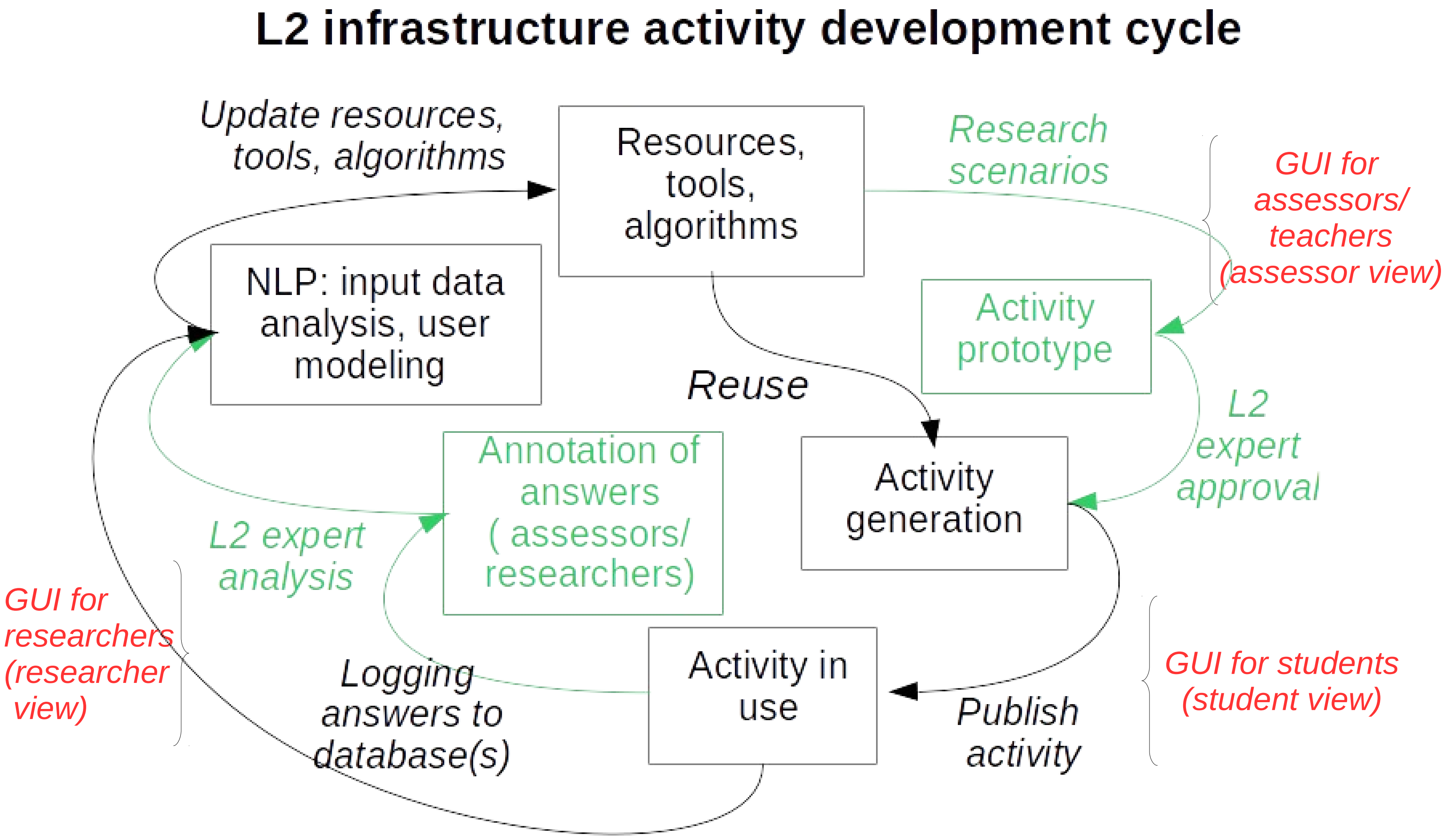
SweLL:
Lärka-based L2 infrastructure

- … as a unit under Språkbanken's infrastructure
- … in the context of CLARIN

Where will this lead?



The ultimate goal


L2 infrastructure activity development cycle
Update resources, tools, algorithms
Resources, tools, algorithms
Research scenarios
GUI for assessors/ teachers (assessor view)
NLP: input data analysis, user modeling
Activity prototype
Reuse
L2 expert approval
Annotation of answers (assessors/ researchers)
Activity generation
L2 expert analysis
GUI for researchers (researcher view)
Logging answers to database(s)
Activity in use
Publish activity
GUI for students (student view)

Thank you!

Questions?