# From Language Learning Platform to Infrastructure for Research on Language Learning

David Alfter*, Lars Borin*, Ildikó Pilán*, Therese Lindström Tiedemann° and Elena Volodina*

* Språkbanken, University of Gothenburg, **Sweden**

° Department of Finnish, Finno-Ugrian and Scandinavian Studies, University of Helsinki, **Finland**

CLARIN conference

10 October 2018

Pisa, Italy

# Presentation overview

- Lärka Overview
- Lärka for learning and teaching
- Lärka as research infrastructure
- Ongoing and future work
- Conclusion

# Lärka Overview

- ICALL platform developed at CLARIN B Center Språkbanken

https://spraakbanken.gu.se/larka

# Lärka Overview

- ITG (Borin and Saxena, 2004)
- A system architecture for ICALL (Volodina et al., 2012)
- Lärka

# Lärka Overview

- Openly available web-based tool
- Builds on a variety of existing SWE-CLARIN language resources such as
  - Korp for querying corpora (Borin et al., 2012)
  - Karp for querying lexical resources (Borin et al., 2013)
  - language technology tools
- Service-oriented architecture

# Presentation overview

- Lärka Overview
- Lärka for learning and teaching
- Lärka as research infrastructure
- Ongoing and future work
- Conclusion

# Lärka for learning and teaching

- Two target learner groups:
  - students of (Swedish) linguistics
  - second/foreign language learners

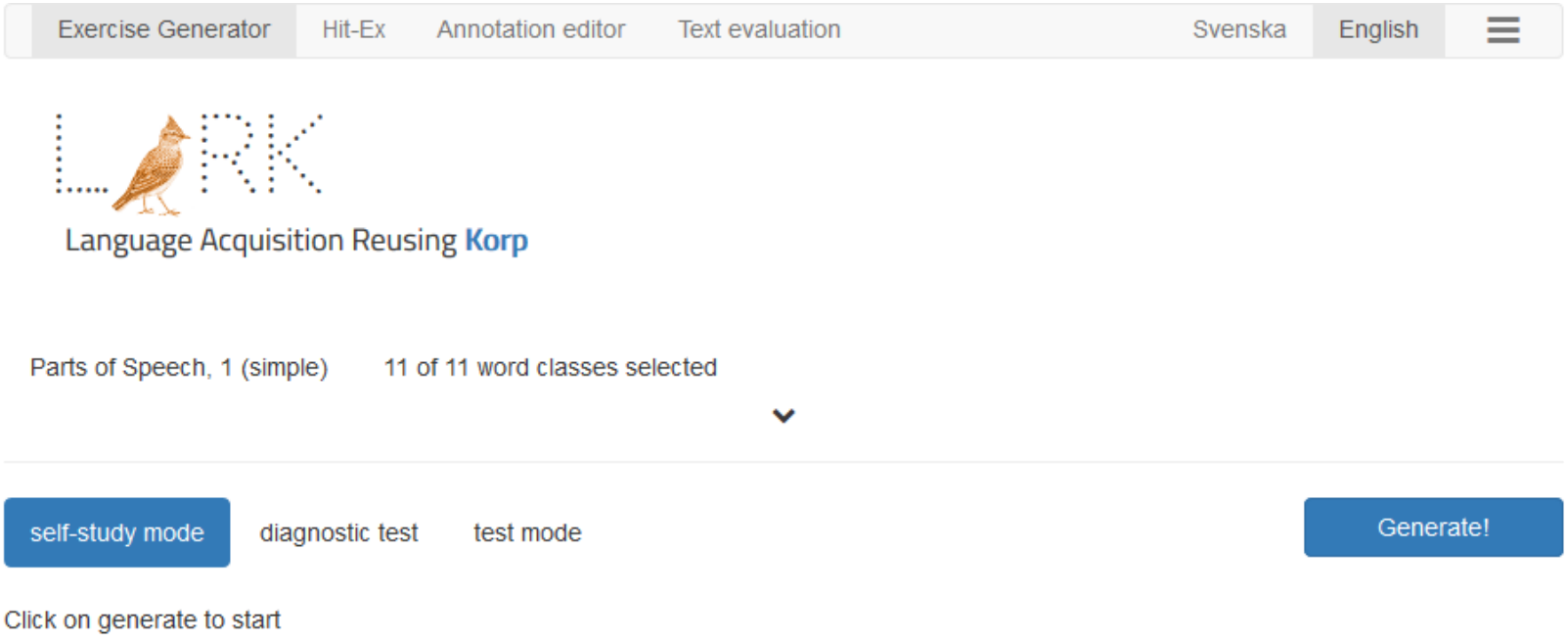
- Automatically generated exercises based on real-life authentic language examples from corpora

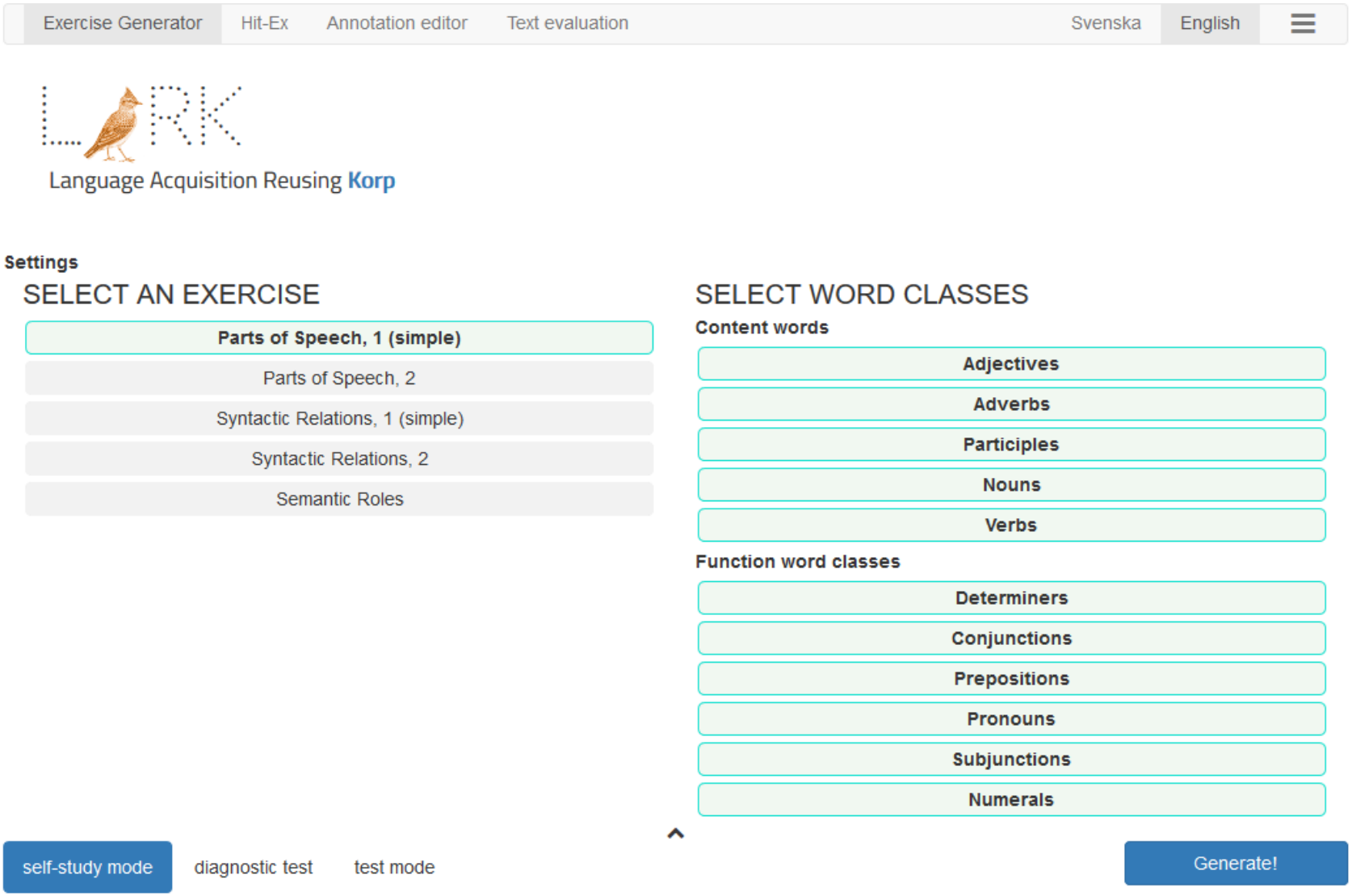
# Exercises for linguists

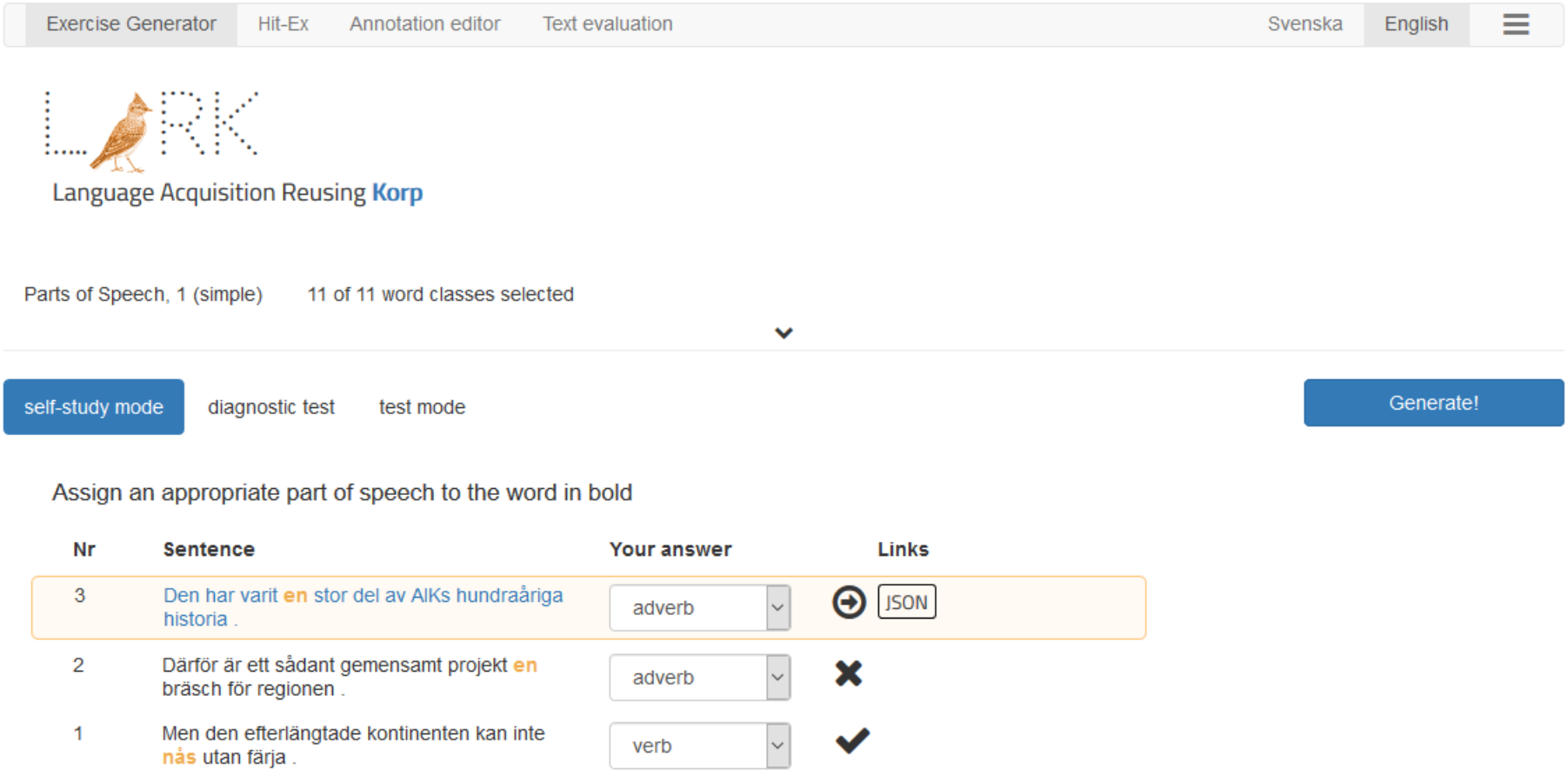
- Exercises for linguistic analysis of
  - word classes
  - syntactic relations
  - semantic roles
- Immediate automatic feedback
- Option to send a log to a teacher and students

# Exercises for linguists

- Two levels (beginner and intermediate), semantic roles one level
- Three modes:
  - self-study
  - diagnostic test
  - test


Exercise Generator
Hit-Ex
Annotation editor
Text evaluation
Svenska
English
Language Acquisition Reusing Korp
Parts of Speech, 1 (simple)
11 of 11 word classes selected
self-study mode
diagnostic test
test mode
Generate!
Click on generate to start


Exercise Generator
Hit-Ex
Annotation editor
Text evaluation
Svenska
English
Language Acquisition Reusing Korp
Settings
SELECT AN EXERCISE
Parts of Speech, 1 (simple)
Parts of Speech, 2
Syntactic Relations, 1 (simple)
Syntactic Relations, 2
Semantic Roles
SELECT WORD CLASSES
Content words
Adjectives
Adverbs
Participles
Nouns
Verbs
Function word classes
Determiners
Conjunctions
Prepositions
Pronouns
Subjunctions
Numerals
self-study mode
diagnostic test
test mode
Generate!
Click on generate to start


Exercise Generator
Hit-Ex
Annotation editor
Text evaluation
Svenska
English
Language Acquisition Reusing Korp
Parts of Speech, 1 (simple)
11 of 11 word classes selected
self-study mode
diagnostic test
test mode
Generate!
Assign an appropriate part of speech to the word in bold
Nr
Sentence
Your answer
Links
3
Den har varit en stor del av AIKs hundraåriga historia .
adverb
JSON
2
Därför är ett sådant gemensamt projekt en bräsch för regionen .
adverb
1
Men den efterlängtade kontinenten kan inte nås utan färja .
verb

# Lärka in practice

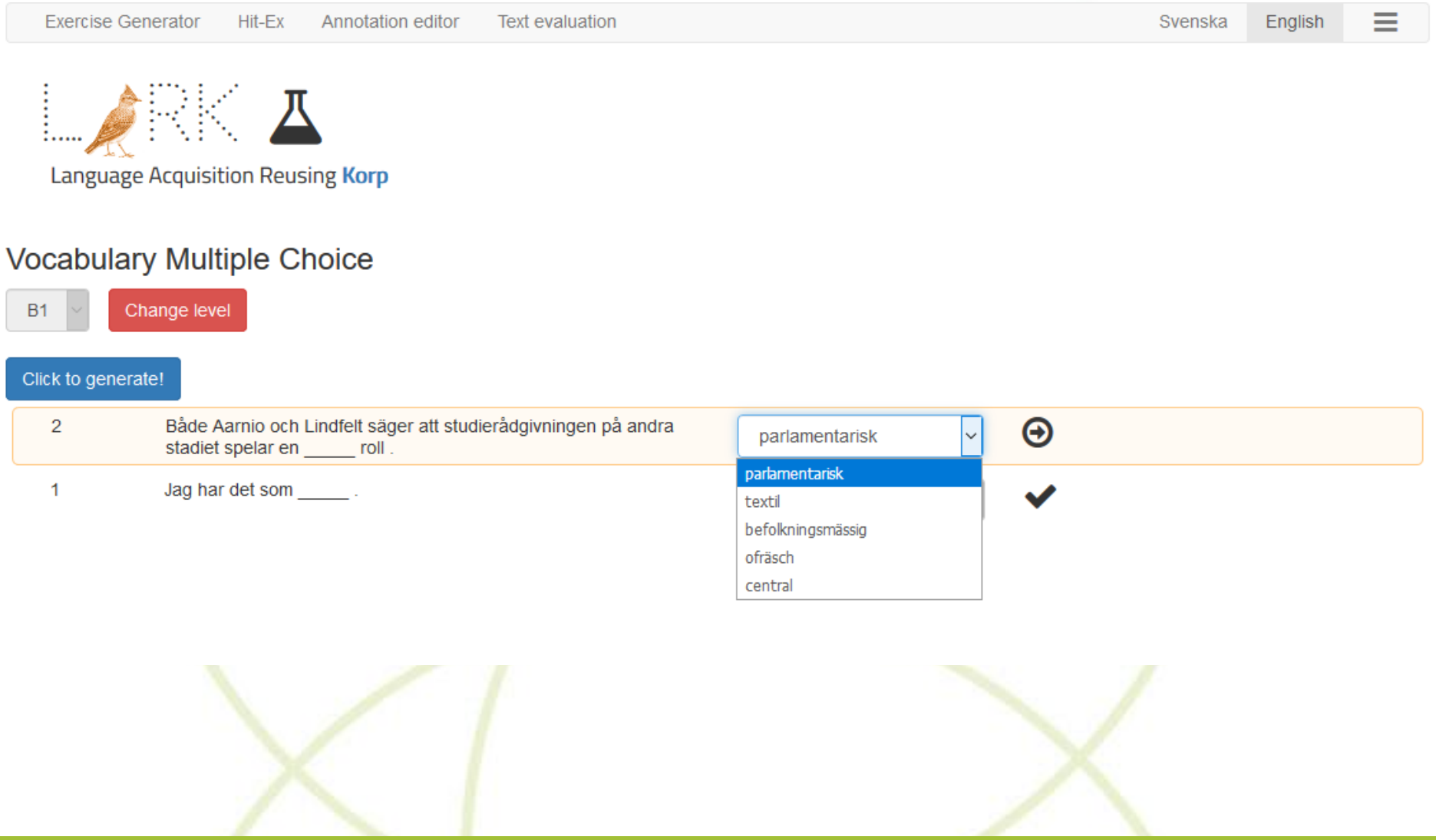
- Used in Introductions to grammar and linguistics in Sweden and Finland (Volodina et al., 2014; Lindström Tiedemann et al., 2016)

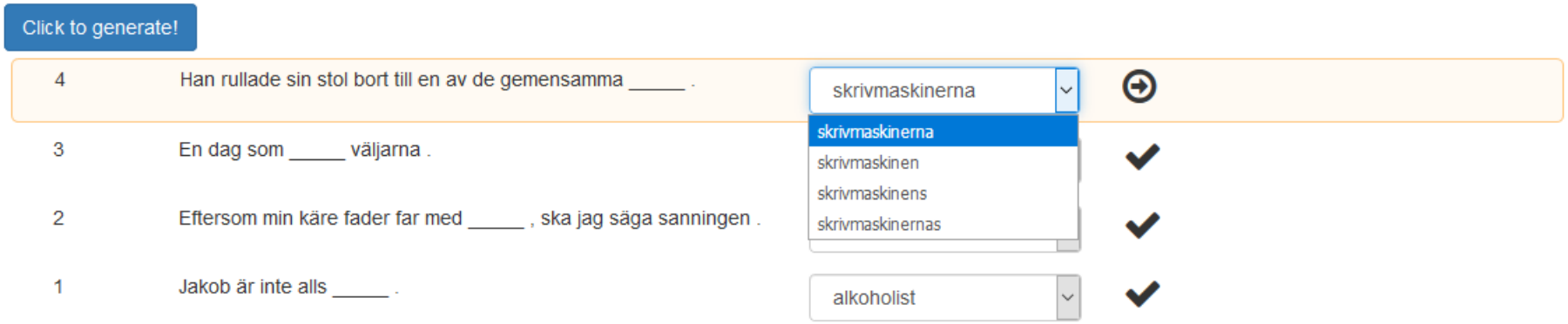
- Labb sessions with teacher
- Discussion of results with fellow students

- Perceived value of platform
- Use in the future

# Exercises for language learners

- Vocabulary multiple-choice
- Inflection multiple-choice
- WordGuess
- Liwrix


Exercise Generator
Hit-Ex
Annotation editor
Text evaluation
Svenska
English
Language Acquisition Reusing Korp
Vocabulary Multiple Choice
B1
Change level
Click to generate!
2
Både Aarnio och Lindfelt säger att studierådgivningen på andra stadiet spelar en _____ roll .
parlamentarisk
parlamentarisk
textil
befolkningsmässig
ofräsch
central
1
Jag har det som _____ .


Click to generate!
4
Han rullade sin stol bort till en av de gemensamma _____ .
skrivmaskinerna
skrivmaskinerna
skrivmaskinen
skrivmaskinens
skrivmaskinernas
3
En dag som _____ väljarna .
2
Eftersom min käre fader far med _____ , ska jag säga sanningen .
1
Jakob är inte alls _____ .
alkoholist

Language Acquisition Reusing **Korp**

# Word guess

**Tries**: 1/7

**Score**: 0

**Definition:**
växt, ört

**Help:**
Show translation

L A

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

LiWriX - Listen and Write Exercise

**Correct:** 2/3 **Score:** 2

Check Show sentences

| | | | |
|---|---|---|---|
| ▶ | | | |
| ▶ | hon | ☹ | han |
| ▶ | faster | ☺ | faster |
| ▶ | molnig | ☺ | molnig |

# Presentation overview

- Lärka Overview
- Lärka for learning and teaching
- Lärka as research infrastructure
- Ongoing and future work
- Conclusion

# Lärka as research infrastructure

- Collection of data from learners
- Annotation of textbooks
- Experimentation and visualization of ongoing research

# Corpus example selection

- HitEx (Pilán et al., 2017)

https://spraakbanken.gu.se/larka/hitex

- Find suitable sentences for learners from corpora

HitEx sentence selection tool

**Search for:**

Lemma ▾ | hund

**Select part-of-speech (optional)**

noun (NN)

Search!

☐ Use default parameters

Advanced options

**1. General**

**2. Search term criteria**

**3. Well-formedness criteria**

| | | |
|---|---|---|
| No dependency root | Filter | Ranker |
| Elliptic | Filter | Ranker |
| Incomplete sentence | Filter | Ranker |
| Non-lemmatized tokens over | % | |
| | Filter | Ranker |
| Non-alphabetical tokens over | % | |
| | Filter | Ranker |

**4. Context independence criteria**

**5. Readability**

**CEFR level:** A1 | A2 | B1 | B2 | C1

**6. Additional structural criteria**

**7. Additional lexical criteria**

HitEx sentence selection tool

**Search for:**

Lemma ▾ | hund

**Select part-of-speech (optional)**

noun (NN)

Search!

☑ Use default parameters

Results

| Rank | Score | Sentence |
|---|---|---|
| 1 | 5 | Jag har alltid älskat **hundar** . |
| 2 | 4 | **Hunden** får sin egen sida och kan ha vänner , både bland människor på Facebook och hundar på Dogbook . |
| 3 | 3 | Till slut började hans två **hundar** äta av kroppen . |

Results with violations

| Rank | Score | Sentence |
|---|---|---|
| 4 | -1 | Han fick sy fyra stygn på knäet efter att ha ramlat i samband med att han bar hem **hunden** . |
| 5 | -1 | Han gav Rex mat , och medan **hunden** åt satt han hopsjunken vid köksbordet med huvudet på armen . |
| 6 | -1 | – Att få **hunden** att lägga leksaker i en låda är inga problem . |
| 7 | -1 | De är två snälla och livliga **hundar** som jag ska ta hand om i en månad . |
| 8 | -1 | Att **hundarna** lär sig sitta still . |
| 9 | -1 | – Jag hade en **hund** som hette Pepe och som blev dödad . |

**Contains proper names:** Pepe

**Contains participles:** dödad

**Sensitive vocabulary:** dödad

**Typicality:** 463.066109242

| | | |
|---|---|---|
| 10 | -1 | Nu sprids efterlysningen av **hunden** Wilja i rekordfart på internet . |

# Text complexity prediction

- Texteval

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

- Assess texts according to CEFR
  - Learner productions (essays)
  - Expert-written texts (reading material)

Vad är egentligen laktosintolerans ? Att vara laktosintolerant betyder att man är överkänslig mot laktos ( mjölksocker ) . Laktos är en kolhydrat som finns naturligt i mjölk och andra mejeriprodukter , till exempel grädde och yoghurt . Laktosintolerans orsakas egentligen av laktasbrist , det vill säga brist enzymet laktas som bryter ner laktos i tunntarmen.Utan laktasenzym förblir laktosen ospjälkad i tunntarmen och går vidare till tjocktarmen där sen mjölksockret bryts ner av bakterierna som finns där och gaser bildas . Detta gör att man kan få magknip , gasbildning , diarré och / eller en känsla av uppblåsthet . Symptomen är individuella och kan variera . En del får väldigt ont medan andra får lindriga besvär . Man kan uppleva att man tål mycket laktos ena dagen och bara lite en annan . Tolerans av laktos kan också till exempel bero på måltidens sammansättning .

**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

**Suggested overall level:** C1

*Given the limited amount of underlying data, this CEFR level should be considered as a suggestion and its use as a basis for decisions in high-stakes assessment is discouraged.*

**Detailed evaluation**

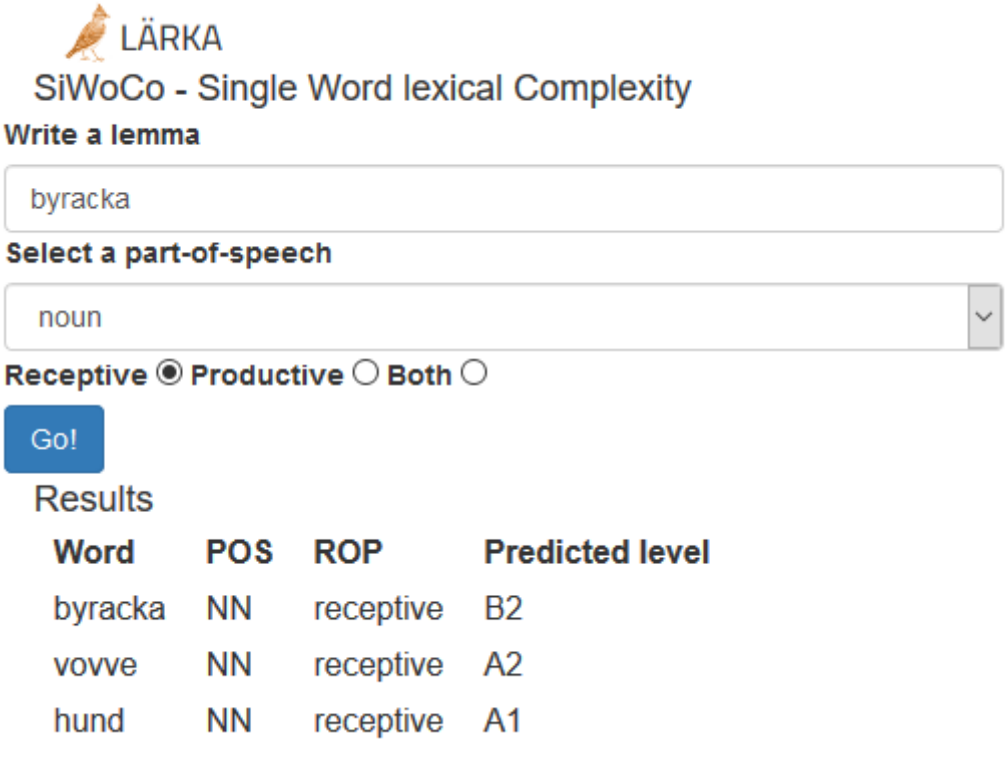
| | |
|---|---|
| Number of sentences | 9 |
| Number of tokens | 146 |
| Non-lemmatized forms | 2 |
| Average sentence length | 16.22 |
| Average token length | 4.99 |
| Average dependency length | 2.52 |
| LIX score | 42 (normal) |
| Nominal ratio | 1.09 |
| Pronoun-to-noun ratio | 0.35 |

# Lexical complexity prediction

- Siwoco (Alfter and Volodina, 2018)

https://spraakbanken.gu.se/larka/siwoco

- Lexical complexity prediction of single words


LÄRKA
SiWoCo - Single Word lexical Complexity
Write a lemma
byracka
Select a part-of-speech
noun
Receptive
Productive
Both
Go!
Results
Word POS ROP Predicted level
byracka NN receptive B2
vovve NN receptive A2
hund NN receptive A1

# Annotation editor

- XML markup of textbooks

https://spraakbanken.gu.se/larka/editor

Language Acquisition Reusing Korp

# Cefrit annotation editor

Annotation manual

**Session key:** 11121 Restore session

**Textbook content**

**List of tags**

- Coursebook
- Extras
- Lesson
- Text
- Genre
- Subheading
- Activity/Task
- Activity instruction
- List
- Language example
- Finalize

Download edited text | Save progress

# Presentation overview

- Lärka Overview
- Lärka for learning and teaching
- Lärka as research infrastructure
- Ongoing and future work
- Conclusion

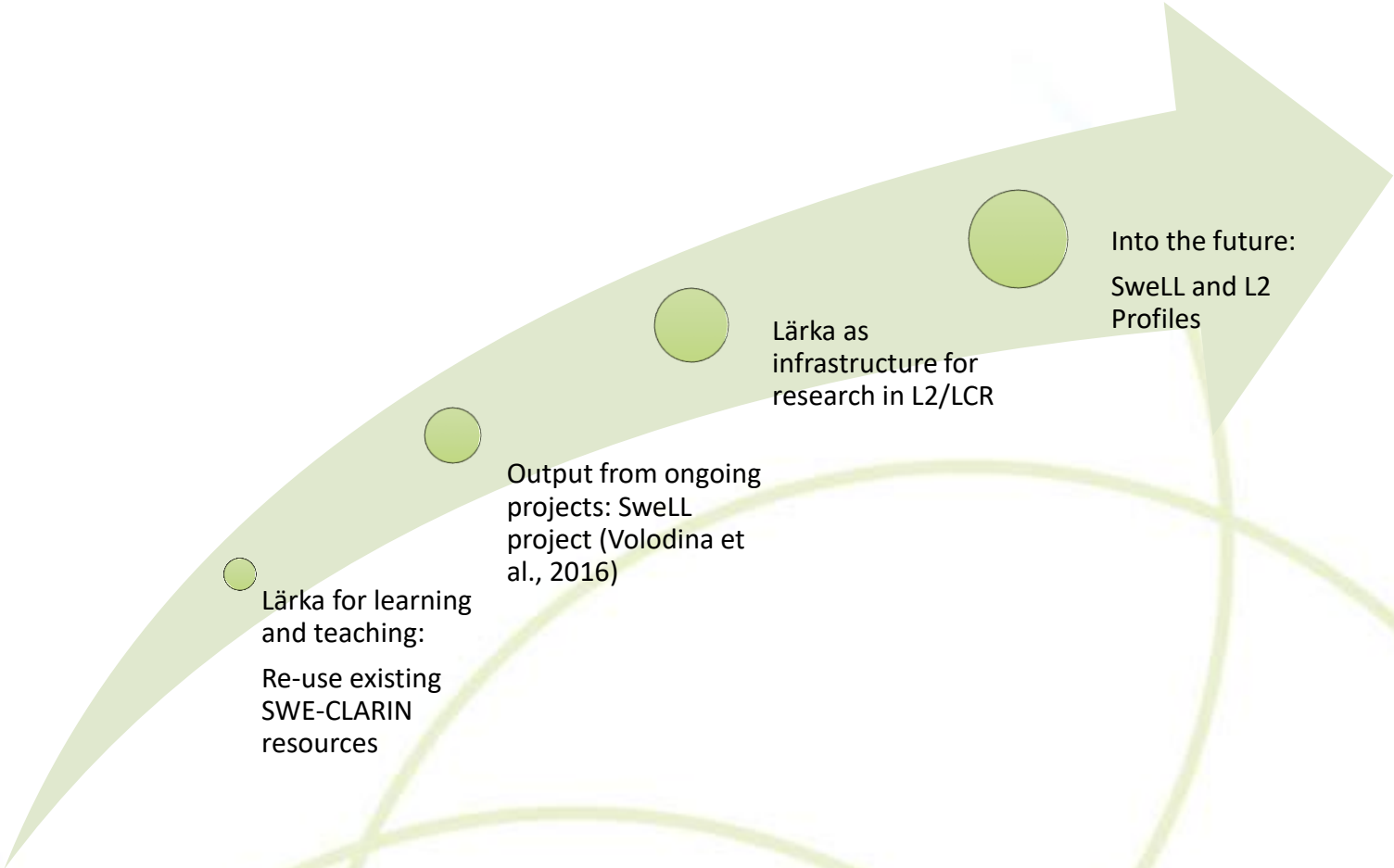
# Ongoing and future work

- Diagnostic placement test
- New exercise formats
- Login and logging of more specific user data
  - L1-specific errors
  - Learner development over time
  - Adaptive exercises

# Presentation overview

- Lärka Overview
- Lärka for learning and teaching
- Lärka as research infrastructure
- Ongoing and future work
- Conclusion

# Conclusion


Lärka for learning and teaching:
Re-use existing SWE-CLARIN resources
Output from ongoing projects: SweLL project (Volodina et al., 2016)
Lärka as infrastructure for research in L2/LCR
Into the future:
SweLL and L2 Profiles

# References I


- David Alfter and Elena Volodina. 2018. Towards Single Word Lexical Complexity Prediction. In Proceedings of the 13th Workshop on Innovative Use of NLP for Building Educational Applications, pages 79–88.
- Lars Borin and Anju Saxena. 2004. Grammar, incorporated. Copenhagen studies in language, 30:125.
- Lars Borin, Markus Forsberg, and Johan Roxendal. 2012. Korp-the corpus infrastructure of Språkbanken. In LREC, pages 474–478.
- Lars Borin, Markus Forsberg, Leif-Jöran Olsson, Olof Olsson, and Jonatan Uppström. 2013. The lexical editing system of Karp. In Proceedings of the eLex 2013 conference, pages 503–516.

# References II


- Therese Lindström Tiedemann, Elena Volodina, and Håkan Jansson. 2016. Lärka: ett verktyg för träning av språkterminologi och grammatik. LexicoNordica, 23:161–181.
- Ildikó Pilán, Elena Volodina, and Lars Borin. 2017. Candidate sentence selection for language learning exercises: from a comprehensive framework to an empirical evaluation. TAL, 57(3/2016):67–91.
- Elena Volodina, Lars Borin, Hrafn Lofsson, Birna Arnbjörnsdóttir, and Guðmundur Örn Leifsson. 2012. Waste not; want not: Towards a system architecture for icall based on nlp component re-use. In Proceedings of the SLTC 2012 workshop on NLP for CALL; Lund; 25th October; 2012, number 080, pages 47–58. Linköping University Electronic Press.
- Elena Volodina, Ildikó Pilán, Lars Borin, and Therese Lindström Tiedemann. 2014. A flexible language learning platform based on language resources and web services. In Proceedings of LREC 2014, Reykjavik, Iceland, pages 3973–3978.
- Elena Volodina, Ildikó Pilán, Ingegerd Enström, Lorena Llozhi, Peter Lundkvist, Gunlög Sundberg, and Monica Sandell. 2016. SweLL on the rise: Swedish learner language corpus for european reference level studies. arXiv preprint arXiv:1604.06583.

# Thank you for your attention!