



# Towards a gold standard for Swedish CEFR-based ICALL



*Elena Volodina, Dijana Pijetlovic, Ildikó Pilán, Sofie Johansson Kokkinakis  
Språkbanken (Swedish Language Bank), University of Gothenburg*

## Some useful terminology

- **Lärka** – **Lär** språket via **KorpusAnalys**, a platform for learning Swedish as a Second Language (L2)
- **CEFR** – Common European Framework of Reference for Languages, a document providing guidelines and standards for language learning, teaching and testing including the scale of proficiency levels



# Presentation plan

- Pedagogical framework (CEFR)
- Lärka: exercise generation (sub-projects) + Demo
- Immediate research agenda
- **A Gold Standard**: *CEFR-corpus* (sub-project) + Demo
- A “*taste*” of initial sentence readability tests
- Planned uses for the corpus

# Pedagogical framework 1

- **CEFR** - Common European Framework of Reference for Languages (CEFR, Council of Europe, 2001)
  - international initiative, accepted inside and outside of Europe
  - guidelines about teaching and assessing languages

The six proficiency levels are named as follows:

C2	Mastery	}	Proficient user
C1	Effective Operational Proficiency		
B2	Vantage	}	Independent user
B1	Threshold		
A2	Waystage	}	Basic user
A1	Breakthrough		

# Pedagogical framework 2

## “can-do” statements

- **CEFR** “can-do” statements for each competence or skill and each level of proficiency.

*Can collate short pieces of information from several sources and summarise them for somebody else. Can paraphrase short written passages in a simple fashion, using the original text wording and ordering.*

CEFR descriptor for **B1**, for **ability to process text**

*Can understand familiar names, words and very simple sentences for example in notices, posters or in catalogues.*

CEFR descriptor for **A1**, **overall reading skills**

# Pedagogical framework 3

## CEFR weaknesses

- non-specific, expressed in terms of competences rather than linguistic constituents
- given to subjective interpretations
- performance outweighs competence
- efforts to interpret CEFR guidelines:
  - **top-down** approaches: start with CEFR guidelines, e.g. Reference Level Descriptions
  - **bottom-up**: start with interpretations that have been made up-to-date (e.g. course materials and graded essays)

# Introducing Lärka

- **Lärka** (Eng. Lark) – **LÄR** språket via **KorpusAnalys**:
  - web-service based ICALL platform
  - at the moment consisting of an exercise generator, and two supportive modules for rating corpus hits and annotating learner-oriented corpora
  - eventually other learner-related activities, e.g. performing readability analysis, selecting texts for language learners from the web, etc.

<http://spraakbanken.gu.se/eng/Research/icall/architecture>



# Zooming into the exercise generator

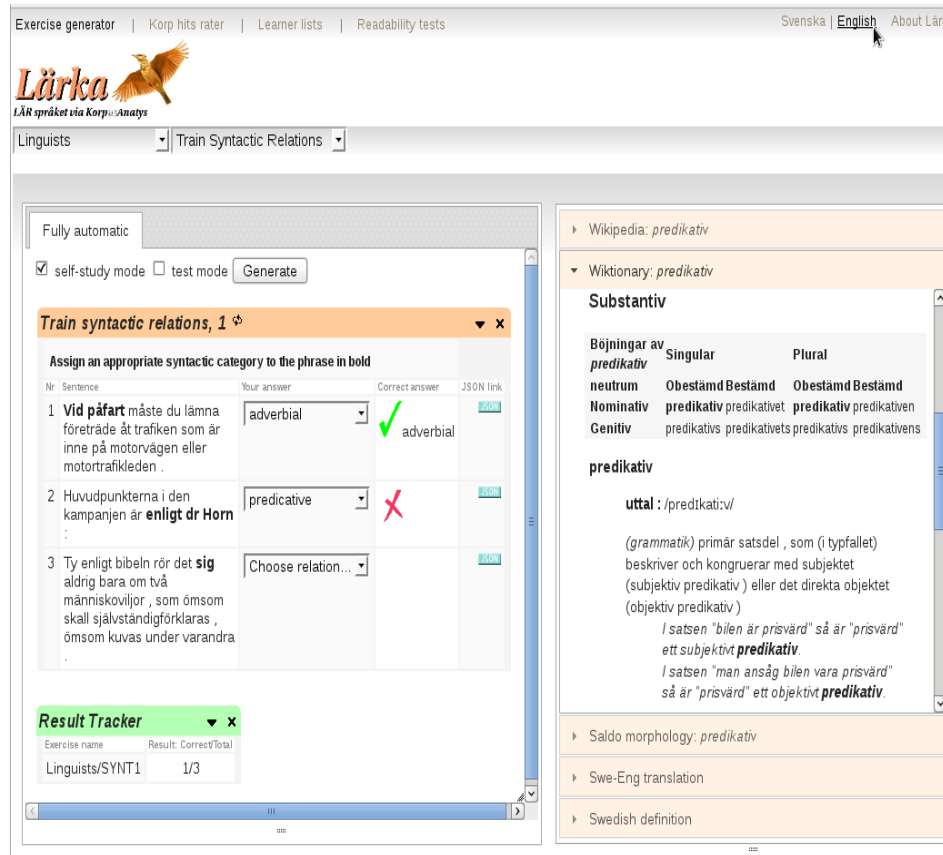
## Target groups

- Students of Linguistics
  - ✓ Items for training parts of speech
  - ✓ Items for training syntactic relations
  - ✓ (coming) Items for training semantic roles
- Learners of Swedish
  - ✓ Multiple-choice items for vocabulary training
  - ✓ Dictation&spelling items
  - ✓ (coming) Wordbox items for training vocabulary and morphology



# The user interface in Lärka

<http://spraakbanken.gu.se/larka>



The screenshot displays the Lärka web application interface. At the top, there is a navigation bar with links for 'Exercise generator', 'Korp hits rater', 'Learner lists', and 'Readability tests'. The language is set to 'English'. The main header features the 'Lärka' logo and the tagline 'LÄR språket via Korp & Anatys'. Below this, a dropdown menu shows 'Linguists' and 'Train Syntactic Relations'.

The central area is titled 'Train syntactic relations, 1'. It contains a table with three columns: 'Nr', 'Sentence', and 'Your answer'. The first row shows a sentence about leaving a car and the user's answer 'adverbial', which is marked as correct. The second row shows a sentence about a campaign and the user's answer 'predicative', which is marked as incorrect. The third row shows a sentence about a sign and the user's answer 'Choose relation...'. A 'Generate' button is visible above the table.

On the right side, there is a panel titled 'Wikipedia: predikativ' and 'Wiktionary: predikativ'. It contains a table of inflections for the word 'predikativ' in Swedish, showing singular and plural forms for neutrum, nominativ, and genitiv. Below the table, there is a section for 'predikativ' with a definition and examples.

At the bottom left, there is a 'Result Tracker' section showing the exercise name 'Linguists/SYNT1' and the result '1/3'.

# Vocabulary items for language learners

- Based on SUC3.0; eventually more corpora
- Multiple-choice principle
- Builds on vocabulary from the Kelly list
- As soon as one item is answered, a new one is generated

# Vocabulary items 2

## simplified **present** version\*

1. Randomly select an item for training from the Kelly list

2. Randomly select a sentence from SUC containing the target item

3. Select distractors

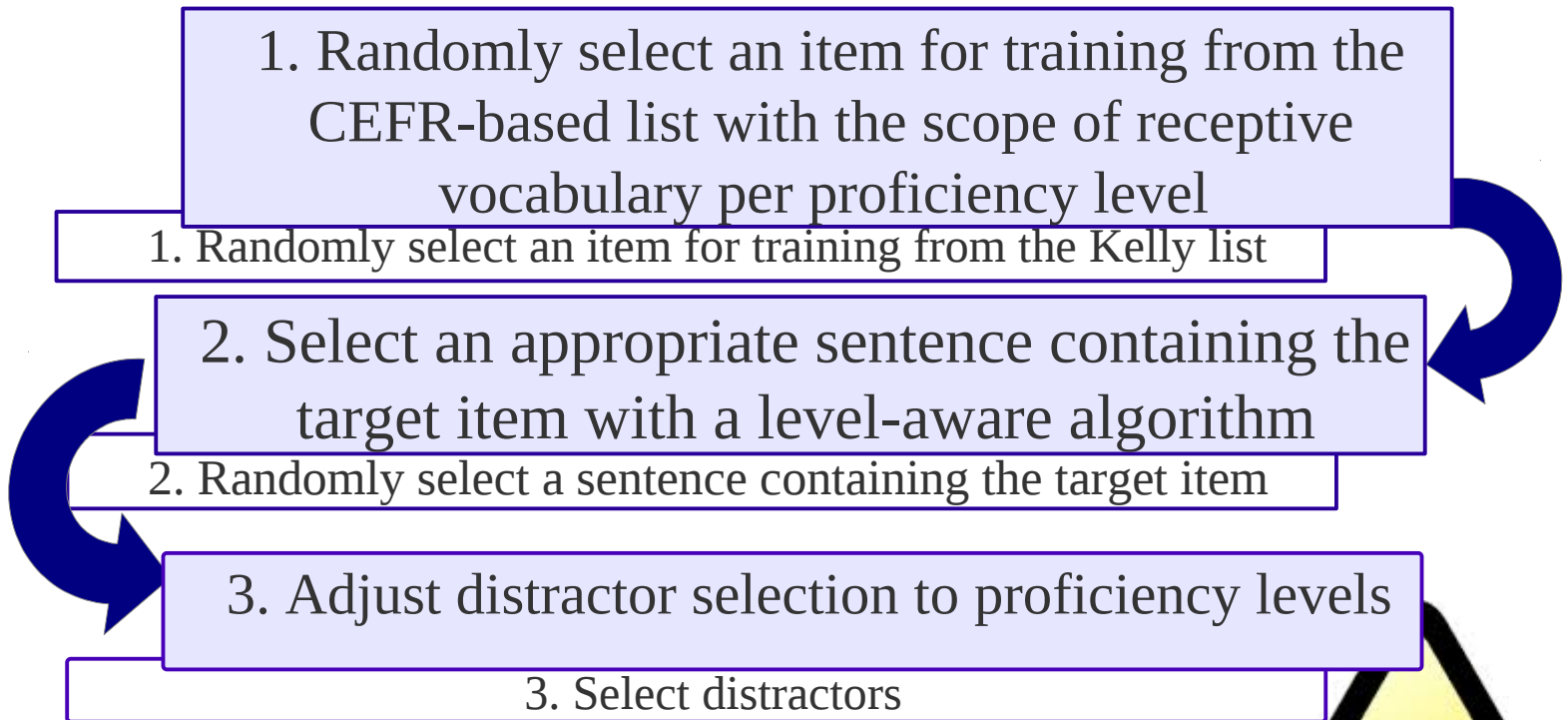
*\* Same refers to the dictation exercises.*

*Soon: item and sentence selection – depending on proficiency level, defined wordclass and/or domain*



# Vocabulary items 3

## future version\*



*\*Same refers to the spelling and dictation items*



# Lärka's research agenda

- Dictated by the practical needs of development
- Automatic generation of learning materials
  - ✓ *for L2 vocabulary training & for students of Linguistics (at the moment)*
  - ✓ *in sentence-long context (at the moment)*
- Practical needs:
  - ✓ *receptive vocabulary scope per level*
  - ✓ *sentence readability measure per level*
- How?
  - ✓ *e.g. study texts used for teaching CEFR-based courses, per level?*
  - ✓ *crowdsourcing?*
  - ✓ *any other ways?*

# CEFR-corpus

project financed by the Department of Swedish

- Gold standard for CEFR-based research
- Text types: normative (input) and learner-produced (output)



- Focus in this project: normative texts

# CEFR-corpus 2

## identifying relevant sources

- Interviews with teachers on relevant course books & novels used in CEFR-based teaching
  - ✓ *resulted in a list of 15+ titles*
  - ✓ *that contain 3187+ pages;*
  - ✓ *with an estimated corpus size of approx. 3 mln tokens*
- Contacts with publishers
  - ✓ *Folkuniversitets förlag, Studentlitteratur, Natur och Kultur, Svenska institutet – negative to sharing electronic materials*
  - ✓ *Liber – positive to collaboration; provided e-texts for research*



# CEFR-corpus 3

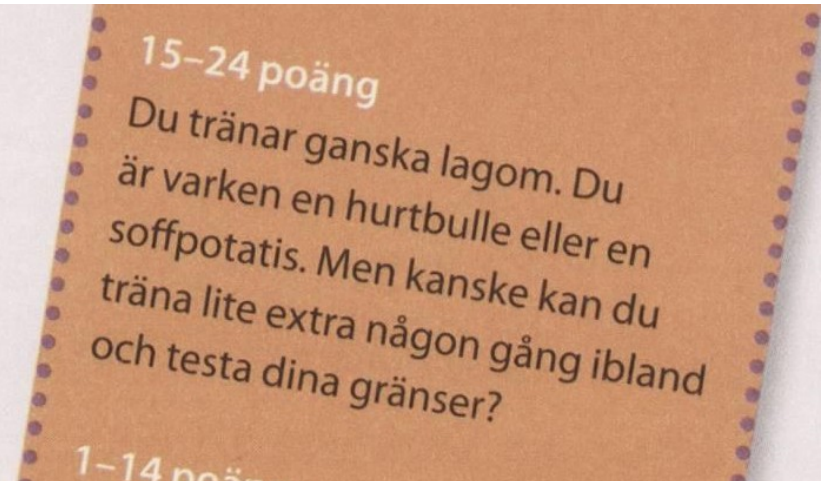
## optical scanning approach

armhävningar efter det.

### 9 Vad är idrott för dig?

- a En stor öl på en sportbar med en gigantisk teveskärm.
- b Idrott är jobbigt men nödvändigt.
- c Idrott är fantastiskt roligt.

### 10 Vad är en bra bantningskur enligt dig?



9 Vad är idrott för dig?  
a En stor öl på en sportbar med en gigantisk teveskärm.  
b Idrott är jobbigt men nödvändigt.  
c Idrott är fantastiskt roligt.  
braät"öTn!hU"bu"e!Det  
ar  
^ attrÖraP åsigoch mer> kanske\* CL. . a  
tänka på att  
danske ska r,  
: Osanna . Uanfa  
; Kr°PPenbehöv ""ani"9iliVet ; 'bland. Också vila  
:  
ar varken enh5ka 'd90m Du SOffP«afe.Men?U"ee"eren 'rän\*!\*eeZ  
.kansk^ndu  
• Och^«nagZ2fn9ibl\*TM  
10 Vad är en bra bantningskur enligt dig?

# CEFR-corpus 4

## text-level annotation

- Lärka-based editor helps to insert text variables:
  - ✓ *proficiency level for the “lesson” (i.e. chapter in a course book)*
  - ✓ *texts, genres, topics*
  - ✓ *other types of language: activity instructions, tasks themselves (e.g. gaps), lists, grammar/vocabulary focus, etc.*

# Taxonomy of text variables

Text parameters: Genre	Text parameters: Topic	Other types of text in lessons
<p>Genre</p> <ul style="list-style-type: none"> <li>• Narration <ul style="list-style-type: none"> <li>• Personal story</li> <li>• Fiction</li> <li>• Description</li> <li>• News article</li> </ul> </li> <li>• Facts <ul style="list-style-type: none"> <li>• Historical facts</li> <li>• Biography</li> <li>• Autobiography</li> <li>• Explanation</li> <li>• Instruction</li> <li>• Rules</li> <li>• Procedures</li> <li>• Report</li> <li>• Demonstration</li> </ul> </li> <li>• Evaluation <ul style="list-style-type: none"> <li>• Argumentation</li> <li>• Exposition</li> <li>• Discussion</li> <li>• Personal reflection</li> <li>• Review</li> <li>• Interpretation, exegesis</li> <li>• Persuasion</li> </ul> </li> <li>• Other <ul style="list-style-type: none"> <li>• Dialogue</li> </ul> </li> </ul>	<p>Topic</p> <ul style="list-style-type: none"> <li>• Personal identification</li> <li>• House and home, environment</li> <li>• Daily life</li> <li>• Free time, entertainment</li> <li>• Travel</li> <li>• Relations with other people</li> <li>• Health and body care</li> <li>• Education</li> <li>• Shopping</li> <li>• Food and drink</li> <li>• Services</li> <li>• Places</li> <li>• Languages</li> <li>• Weather</li> </ul>	<p>Activity instruction</p> <ul style="list-style-type: none"> <li>• Listening</li> <li>• Reading</li> <li>• Writing</li> <li>• Speaking</li> <li>• Discussion</li> <li>• Grammar exercise</li> <li>• Vocabulary exercise</li> <li>• Text question</li> </ul> <p>Task</p> <ul style="list-style-type: none"> <li>• Listening</li> <li>• Reading</li> <li>• Writing</li> <li>• Speaking</li> <li>• Discussion</li> <li>• Grammar exercise</li> <li>• Vocabulary exercise</li> <li>• Text question</li> <li>• Gaps</li> </ul> <p>List</p> <ul style="list-style-type: none"> <li>• Vocabulary</li> <li>• Grammar</li> <li>• Sentences</li> </ul> <p>Language example</p> <ul style="list-style-type: none"> <li>• Vocabulary</li> <li>• Grammar</li> <li>• Pronunciation</li> </ul>

Course book editor

Learner essays editor

Paste your text below and choose annotation tags from the menu to the left

Update tags and IDs

Last update 2013/03/06 7:52:32

Annotation menu

- Coursebook
- Extras
- Lesson
- Text
- Genre
- Topic
- Activity instruction
- Task
- List
- Language example

envisnet  
fantasi  
taktkänsla  
snabbhet  
taktik  
uthållighet  
</list>  
<activity\_instruction id="ai\_1\_8" type="listening">  
A Lyssna på folk som pratar om sport. Vilka tre sporter talar de om?  
</activity\_instruction>  
B Arbeta i par och läs frågorna. Lyssna igen och svara på frågorna.)  
<activity\_instruction id="ai\_1\_9" type="listening">  
<task id="task\_1\_1" ref="#a1\_1\_9" type="listening">  
1 a Vad gör Ulf innan han börjar springa?  
b Vad kallar man hans träningsmetod?  
2 a Vad blev resultatet i matchen mellan Sjöbergs IF och Lindbergs IF?  
b Hur gammal var Marika när hon började spela?  
3 a Vad tränar Jan på lördagar?  
b Varför började han träna?

IDs &amp; Entities

ai\_1\_8  
list\_1\_2  
ai\_1\_7  
list\_1\_1  
langex\_1\_3  
ai\_1\_6  
langex\_1\_2  
ai\_1\_5  
langex\_1\_1  
ai\_1\_4  
text\_1\_1  
ai\_1\_3  
ai\_1\_2  
ai\_1\_1  
1

Download edited text as file



## LASSE-MAJA

På 1800-talet var biografen *Lasse-Majas äventyr* den mest lästa boken i Sverige vid sidan av Bibeln. I boken berättar tjuven och transvestiten Lars Molin själv om sina öden och äventyr. Lars Molin föddes som Lars Larsson, men bytte efter en tid efternamn. Namnet Lasse-Maja kommer av att Lars oftast gick klädd i kvinnokläder.



Lars Larsson föddes 1785 i närheten av Arboga, i södra Bergslagen. Han var en busig pojke som tyckte mer om att snatta pengar, spela kort och lata sig än att arbeta. En bit ifrån Lasse bodde Maja som blev hans första fästmö. Lasse var en smal och söt pojke och en dag provade han sin fästmöns kläder. Maja tyckte att Lasse var mycket fin i kvinnokläder och även hennes föräldrar blev imponerade.

Från den dagen gick Lasse oftast klädd i kvinnokläder och han övade sig på allt som en kvinna förväntades kunna. Han arbetade klädd som bondpiga på gårdar, ibland tillsammans med Maja, och lärde sig att mjölka, laga mat, baka, städa, tvätta och annat. Det verkade som om kvinnojobb passade honom bättre än mansjobb, som lätt tråkade ut honom.

Lasse-Maja blev en duktig kock och han lärde sig att sy och dansa. Många trodde att han var en fin fröken och han fick enkelt tjänst på olika fina gårdar med hjälp av förfälskade betyg. På gårdarna blev han

## CEFR-corpus teaser1

- What is the genre?
  - ✓ narration / description?
  - ✓ facts / biography?

- What is the topic?
  - ✓ famous people?
  - ✓ crime & punishment?

# CEFR-corpus

## teaser 2

- What is the genre?
  - ✓ facts / instruction?
  - ✓ evaluation / personal reflection?

Ulf Frövi är psykolog och arbetar som konsult med personer som ska flytta utomlands. Här ger han några råd som kan underlätta processen:

- När det känns jobbigt, tänk på att livet till stor del består av små och stora problem som ska lösas, även i ditt hemland.
- Försök lära dig så mycket som möjligt om den nya kulturens historia, politik, geografi, konsthistoria osv. Ju mer du vet desto mer förstår du.
- Glöm inte bort dina intressen och hobbyer. Om du förut var medlem i någon klubb, försök då att hitta motsvarande klubb i ditt nya land. Det kan vara svårt att få nya vänner, men ett bra sätt är garanterat att försöka hitta personer som har samma intressen som du.
- Värdera inte och jämför inte olika länder. Konstatera bara att man gör på olika sätt i olika länder.
- Om du är frustrerad över hur dina nya landsmän beter sig, tänk då på att du inte kan ändra på ett helt folk. De är ganska nöjda med sakernas tillstånd. Du kan bara ändra på din egen attityd och ditt eget beteende.
- Om du har en partner från ett annat land, försök då att vara flexibel utan att glömma bort värderingar och principer som är viktiga för dig. Välj dina krig. Ta bara upp diskussioner om sådant som du tycker är extra viktigt. Det är kanske viktigare t ex att vara överens om hur barnen ska uppfostras, än exakt vilka jultraditioner ni ska ha.





# CEFR-corpus 5

## linguistic annotation

- Linguistic annotation (standard Korp pipeline):
  - parts of speech (*pos*), morpho-syntactic information (*msd*),
  - syntactic relations (*ref*, *dephead*, *deprel*), *lemmas*,
  - and linking to morphology lexicon (*lex*, *saldo*).

```
<w pos="HP" msd="HP.NEU.SIN.IND" lemma="|vad|"
lex="|vad..pn.1|" saldo="|vad..1|" prefix="|" suffix="|"
ref="1" dephead="4" deprel="+F">vad</w>
```

```
<w pos="VB" msd="VB.PRS.AKT" lemma="|vilja|"
lex="|vilja..vb.1|" saldo="|vilja..1|" prefix="|" suffix="|"
ref="2" dephead="4" deprel="MS">vill</w>
```

# CEFR-corpus

## present-day status

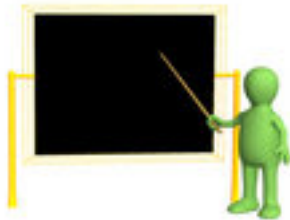
- Two course books for B1 (intermediate) level
  - ✓ *scanned*
  - ✓ *annotated*
  - ✓ *uploaded into Korp (demo?)*
- Tests on sentence readability for B1 level
  - ✓ *master thesis project by Ildikó Pilán*
  - ✓ *to be presented at EuroCALL 2013*





## MT on sentence readability: Purpose

- Automatically **select** and **rank** sentences from Swedish native language texts.
- Sentences should be:
  - **understandable** by students of Swedish as a second language (L2), especially at B1 level
  - suitable **exercise item**
  - appropriate **examples** to illustrate a new lexical item.
- Target users:



Teachers of  
L2 Swedish



Students of  
L2 Swedish



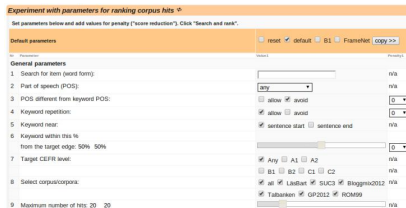
Lexicographers

Most informative features	Logistic Reg. (RFE)	Decision Tree
Sentence length	✓	✓
Average token length	✓	✓
Percentage of words longer than 6 characters		✓
Modifiers	✓	
Average dependency depth		✓
Average number of senses per word	✓	✓
Nominal Ratio	✓	✓
Average frequency in the Wikipedia list	✓	
Average frequency in Kelly list	✓	✓
Percentage of difficult words	✓	✓
Number of difficult words	✓	
Adverb variation	✓	
Noun / Verb ratio	✓	
Model Verb / Verb ratio	✓	

# The readability module



preferences



Experiment with parameters for ranking corpus hits

Set parameters below and add values for penalty ("score reduction"). Click "Search and rank".

Default parameters:

General parameters

Parameter	Value	Unit
1. Search for item (word form)	<input type="text" value="any"/>	ms
2. Part of speech (POS)	<input type="text" value="any"/>	ms
3. POS different from keyword POS	<input type="checkbox"/> allow <input checked="" type="checkbox"/> avoid	ms
4. Keyword repetition	<input type="checkbox"/> allow <input checked="" type="checkbox"/> avoid	ms
5. Keyword near	<input checked="" type="checkbox"/> sentence start <input type="checkbox"/> sentence end	ms
6. Keyword within this % from the target edge: 50% 50%	<input type="text" value="any"/>	ms
7. Target CEFR level	<input type="checkbox"/> Any <input type="checkbox"/> A1 <input type="checkbox"/> A2 <input type="checkbox"/> B1 <input type="checkbox"/> B2 <input type="checkbox"/> C1 <input type="checkbox"/> C2	ms
8. Select corpora/corpora	<input type="checkbox"/> all <input checked="" type="checkbox"/> Laidback <input type="checkbox"/> SLIC3 <input type="checkbox"/> Biogram2012 <input type="checkbox"/> Tashkent <input type="checkbox"/> UP2012 <input type="checkbox"/> HCM99	ms
9. Maximum number of hits	<input type="text" value="20"/>	ms

**FRONTEND**  
(user interface)

keyword



sentences

parameters




**BACKEND**  
(web service)

**filtered and  
ranked  
sentences**

# The user interface in Lärka

[http://spraakbanken.gu.se/larka/larka\\_hitex\\_index.html](http://spraakbanken.gu.se/larka/larka_hitex_index.html)

**Experiment with parameters for ranking corpus hits** 

Set parameters below and add values for penalty ("score reduction"). Click "Search and rank".

**Default parameters** ☐ reset ☒ default ☐ B1 ☐ FrameNet

No	Parameter	Value1	Penalty1
<b>General parameters</b>			
1	Search for item (word form):	<input type="text"/>	n/a
2	Part of speech (POS):	<input type="text" value="any"/>	n/a
3	POS different from keyword POS:	<input type="checkbox"/> allow <input checked="" type="checkbox"/> avoid	<input type="text" value="0"/>
4	Keyword repetition:	<input checked="" type="checkbox"/> allow <input type="checkbox"/> avoid	<input type="text" value="0"/>
5	Keyword near:	<input checked="" type="checkbox"/> sentence start <input type="checkbox"/> sentence end	n/a
6	Keyword within this % from the target edge: 50% 50%	<input type="text" value="50"/>	<input type="text" value="0"/>
7	Target CEFR level:	<input checked="" type="checkbox"/> Any <input type="checkbox"/> A1 <input type="checkbox"/> A2 <input type="checkbox"/> B1 <input type="checkbox"/> B2 <input type="checkbox"/> C1 <input type="checkbox"/> C2	n/a
8	Select corpus/corpora:	<input checked="" type="checkbox"/> all <input checked="" type="checkbox"/> LäsBart <input checked="" type="checkbox"/> SUC3 <input checked="" type="checkbox"/> Bloggmix2012 <input checked="" type="checkbox"/> Talbanken <input checked="" type="checkbox"/> GP2012 <input checked="" type="checkbox"/> ROM99	n/a
9	Maximum number of hits: 20 20	<input type="text" value="20"/>	n/a

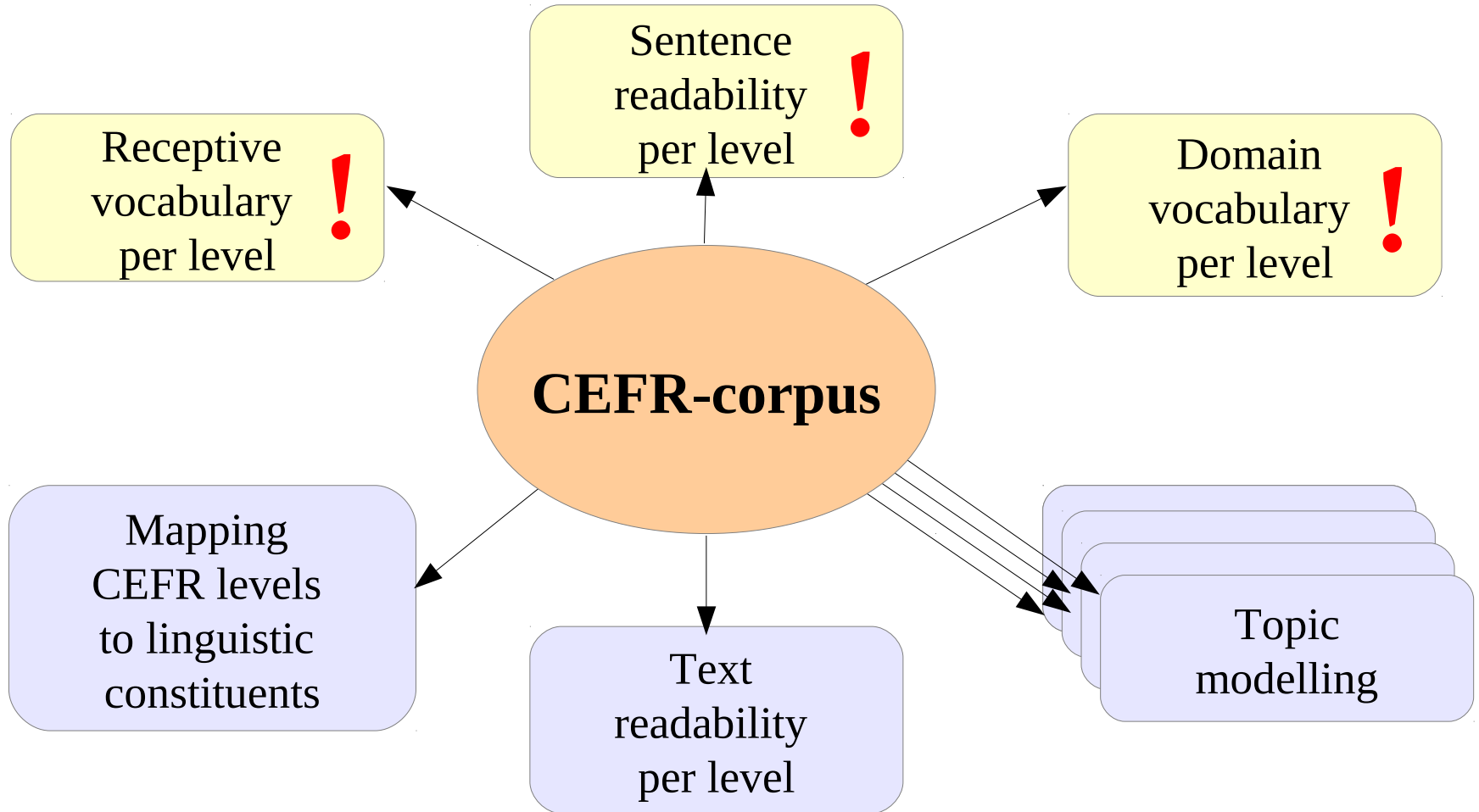
# CEFR-corpus

## intended use

- Identification of receptive vocabulary per proficiency level
- Test on *sentence* readability per proficiency level
- Tests on *text* readability per proficiency level
- Topic modeling
- Question generation
- Mapping CEFR “can-do” statements to linguistic constituents
- etc.



# Lärka's research agenda





**Thank you!**

**Questions?**