Lärka: an online platform where language learning meets natural language processing

cost
EUROPEAN COOPERATION IN SCIENCE & TECHNOLOGY

Funded by the Horizon 2020 Framework Programme of the European Union

Ildikó Pilán, David Alfter and Elena Volodina
Swedish Language Bank, University of Gothenburg, Sweden

GÖTEBORGS UNIVERSITET

Språk-BANKEN

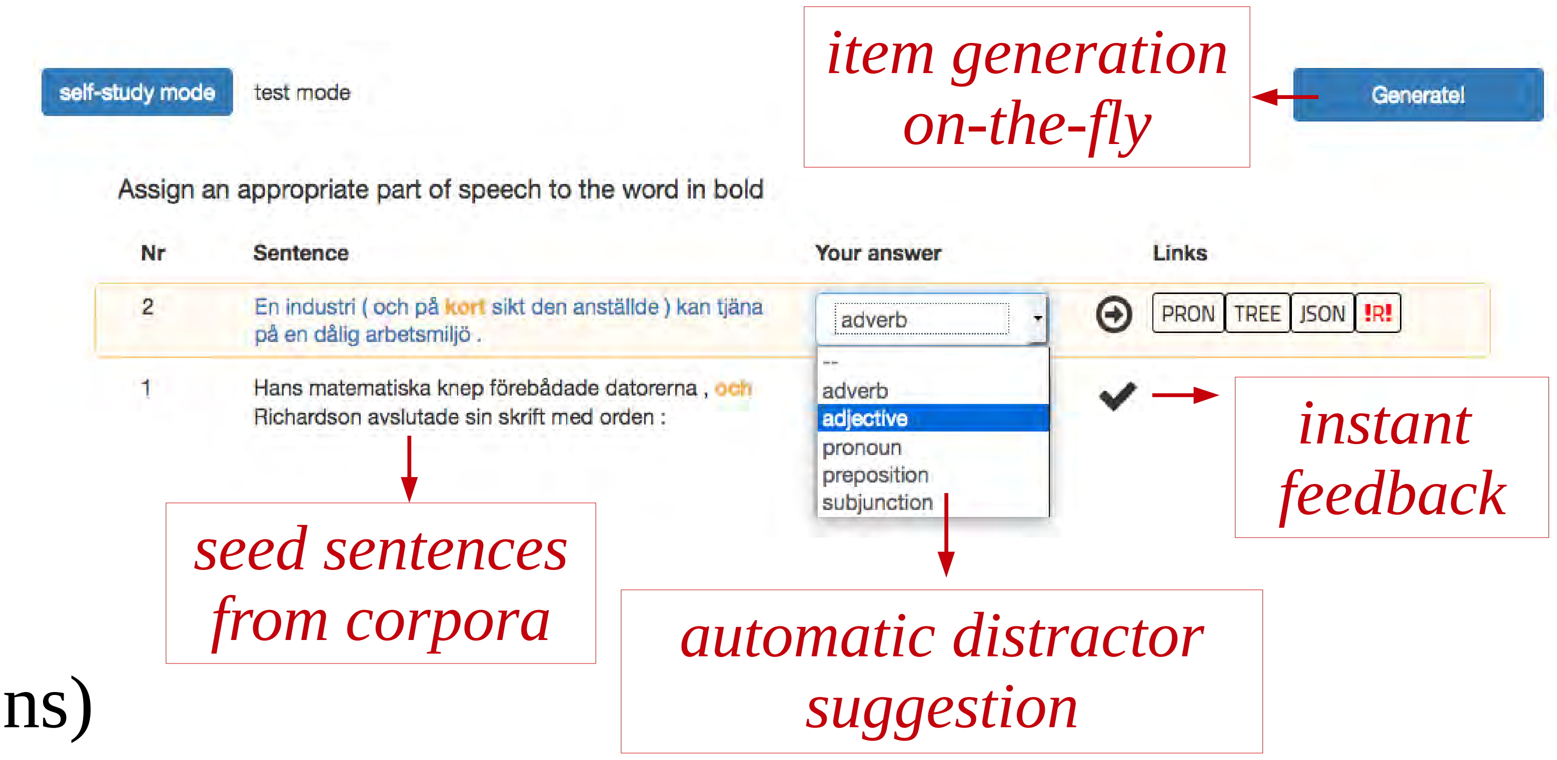
Lärka: introduction

LÄRKA
LÄR språket via Korpus Analys

- A **free**ly available Intelligent Computer Assisted Language Learning (**ICALL**) **platform** developed at Språkbanken
- **Builds on** a variety of language technology **resources** (corpora, lexical resources, NLP tools)
- **Aim**: **support** language **teachers**, **learners** and **researchers**
- **Web-based,** functionalities also available as **web services**

Multiple-choice exercises

- Exercises for **learners** of **Swedish** and students of **linguistics**
- 1 correct + **distractors** (incorrect options)



item generation on-the-fly

Assign an appropriate part of speech to the word in bold

Nr	Sentence	Your answer	Links
2	En industri (och på **kort** sikt den anställde) kan tjäna på en dålig arbetsmiljö .	adverb	PRON TREE JSON !R!
1	Hans matematiska knep förebådade datorerna , **och** Richardson avslutade sin skrift med orden :	✓	

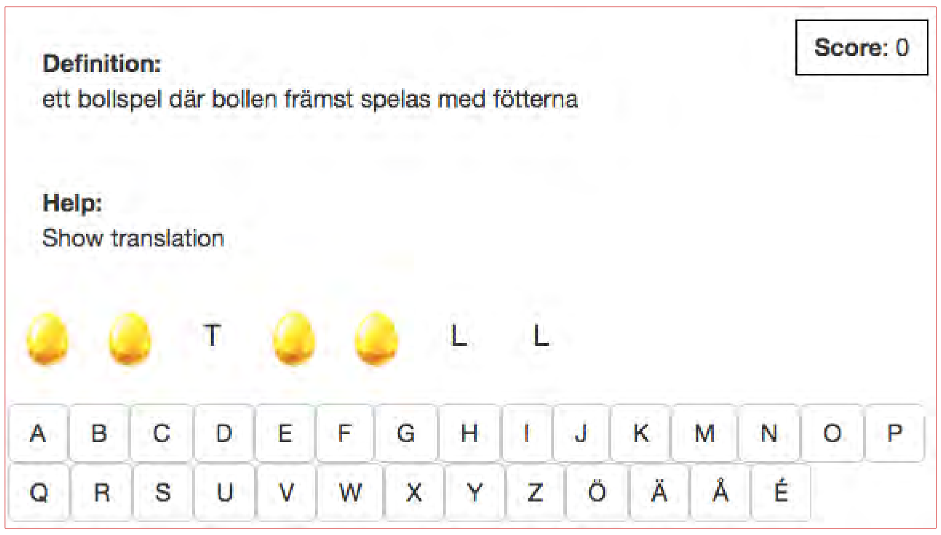
instant feedback

seed sentences from corpora

automatic distractor suggestion

Word Guess

- **Gamified** lexical resourse re-use
- Hangman game mechanism
- **Clues**: definitions, translations


Score: 0

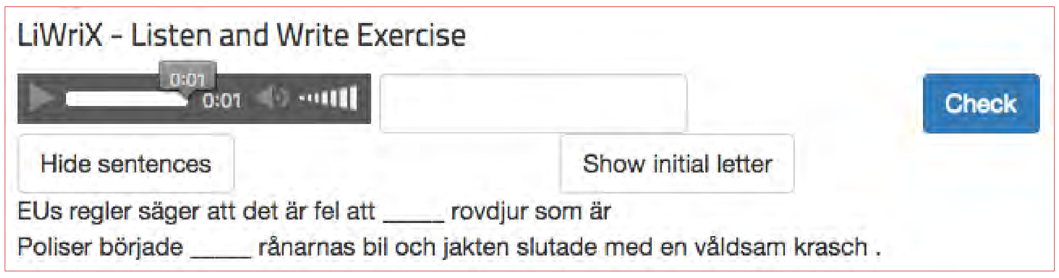
Definition:
ett bollspel där bollen främst spelas med fötterna

Help:
Show translation

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

Listening exercise

- **Listen** to a word and **type** it
- **Clues**: sentences or first letter
- **Nuance** text-to-speech technology
- **Machine-learning** for the automatic classification of target **word complexity**

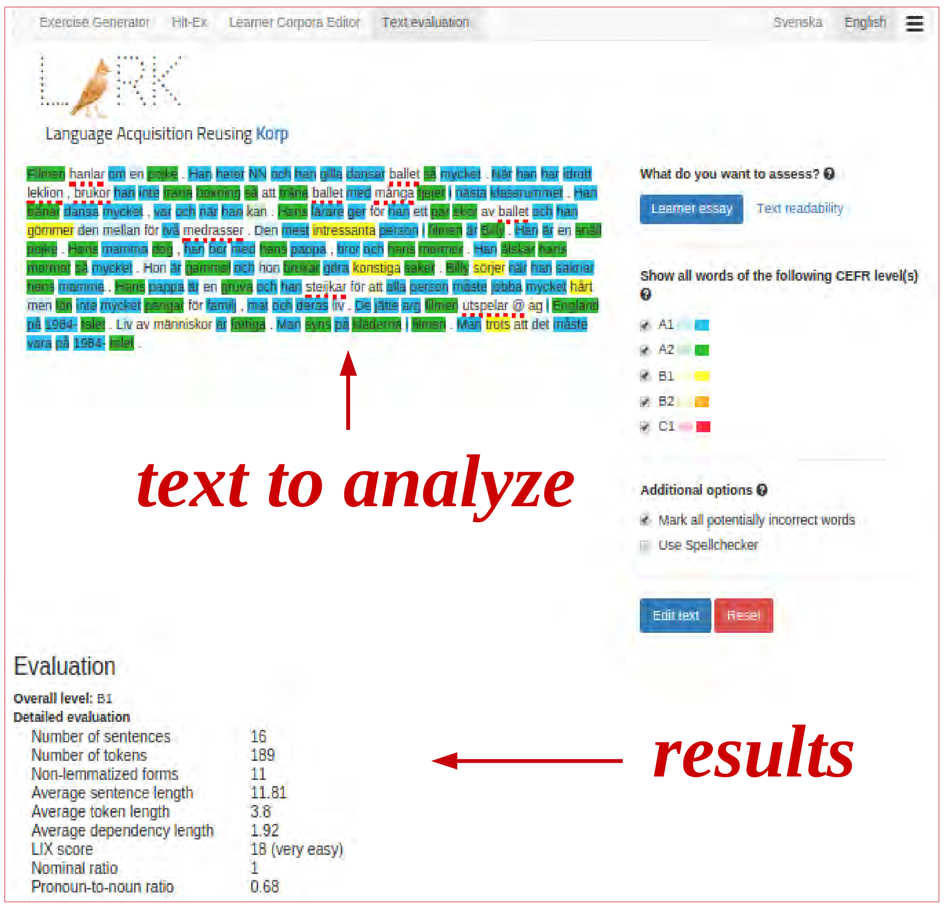

LiWriX - Listen and Write Exercise

EUs regler säger att det är fel att ____ rovdjur som är
Poliser började ____ rånarnas bil och jakten slutade med en våldsam krasch .

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

TextEval: text complexity analysis

- **Aim**: classification of second and foreign language (L2) texts into **<u>proficiency (CEFR) levels</u>**
- Texts written **by (*productive*)** or **for (*receptive*)** L2 learners can be analyzed for complexity
- **Machine-learning** based CEFR level assessment and **statistics**
- Per-**token color**-coded CEFR levels
  - **darker** shades: receptive vocabulary
  - **lighter** shades: productive vocabulary



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

LÄRK

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 steijkar 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
Non-lemmatized forms	11
Average sentence length	11.81
Average token length	3.8
Average dependency length	1.92
LIX score	18 (very easy)
Nominal ratio	1
Pronoun-to-noun ratio	0.68

text to analyze

results

Feature set

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

- **Supervised** machine learning
- **Training data**:
  - receptive: L2 coursebook corpus (COCTAILL)
  - productive: L2 essay corpus (SweLL)
- **Accuracy**: 81% (receptive)

  72% (productive)

- From KELLY, an L2 frequency word **lists** with **mappings** to **CEFR** levels based on frequency bands
- **CEFR levels** of tokens, not words as **lexical** features (data sparsity)

Incidence score

(normalized category count)

$$\boxed{\frac{1000}{N^{token}} \times N^{category}}$$

$$\boxed{\frac{N^{category}}{N^{noun} + N^{verb} + N^{adj} + N^{adv}}}$$

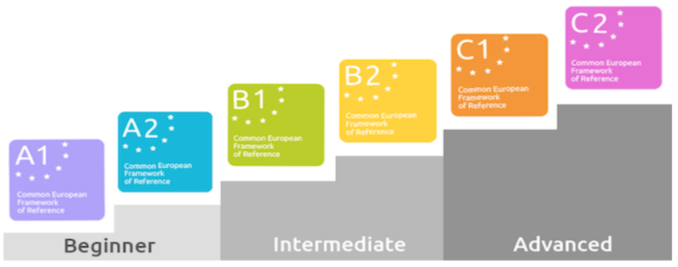
http://spraakbanken.gu.se/larka/texteval

HitEx: exercise item selection

Overview

- **Aim**: identify sentences from corpora to be used as L2 exercise items
- **Why**: a. **authentic** language use
  b. available in **large quantities**
- **Challenges:**
  a. sentences might be too **difficult** or
  b. sometimes **out of context**.
- **HitEx**: a **hybrid** system for sentence selection with **25 criteria**
  - **machine learning** → L2 complexity
  - **rules** → all other criteria
- Criteria set as **filters** or **rankers**

Common European Framework of Reference for Languages (CEFR)



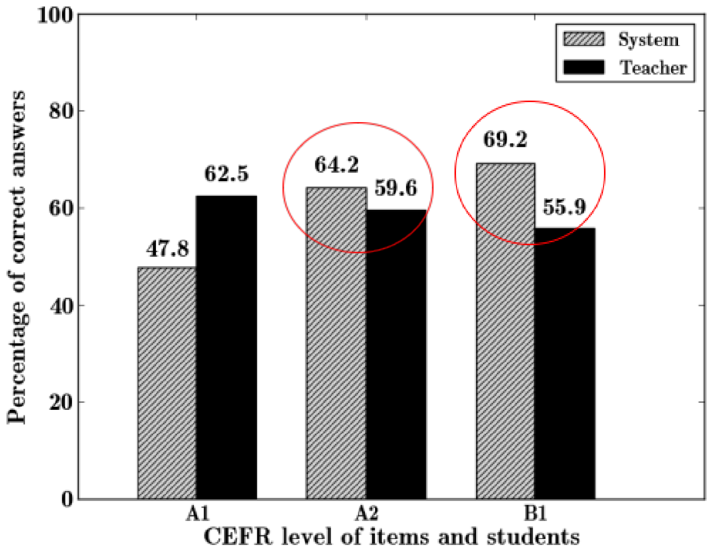
http://gostudylink.net/en/blog/cefr-levels-explained

How well does it work?

Ideal item difficulty
(IID)

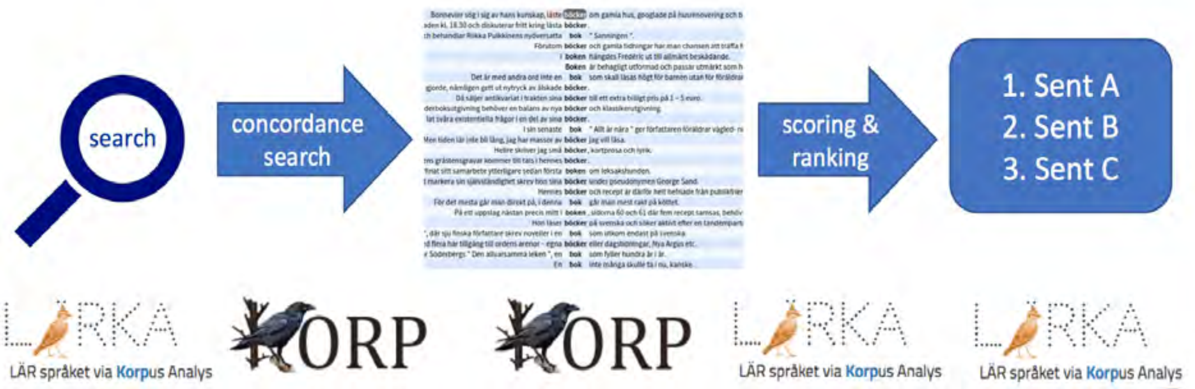
$IID = P_c + (100-P_c) / 2$ → *correct %*

IID = **0.645**
(= 64,5 % students answering correctly)

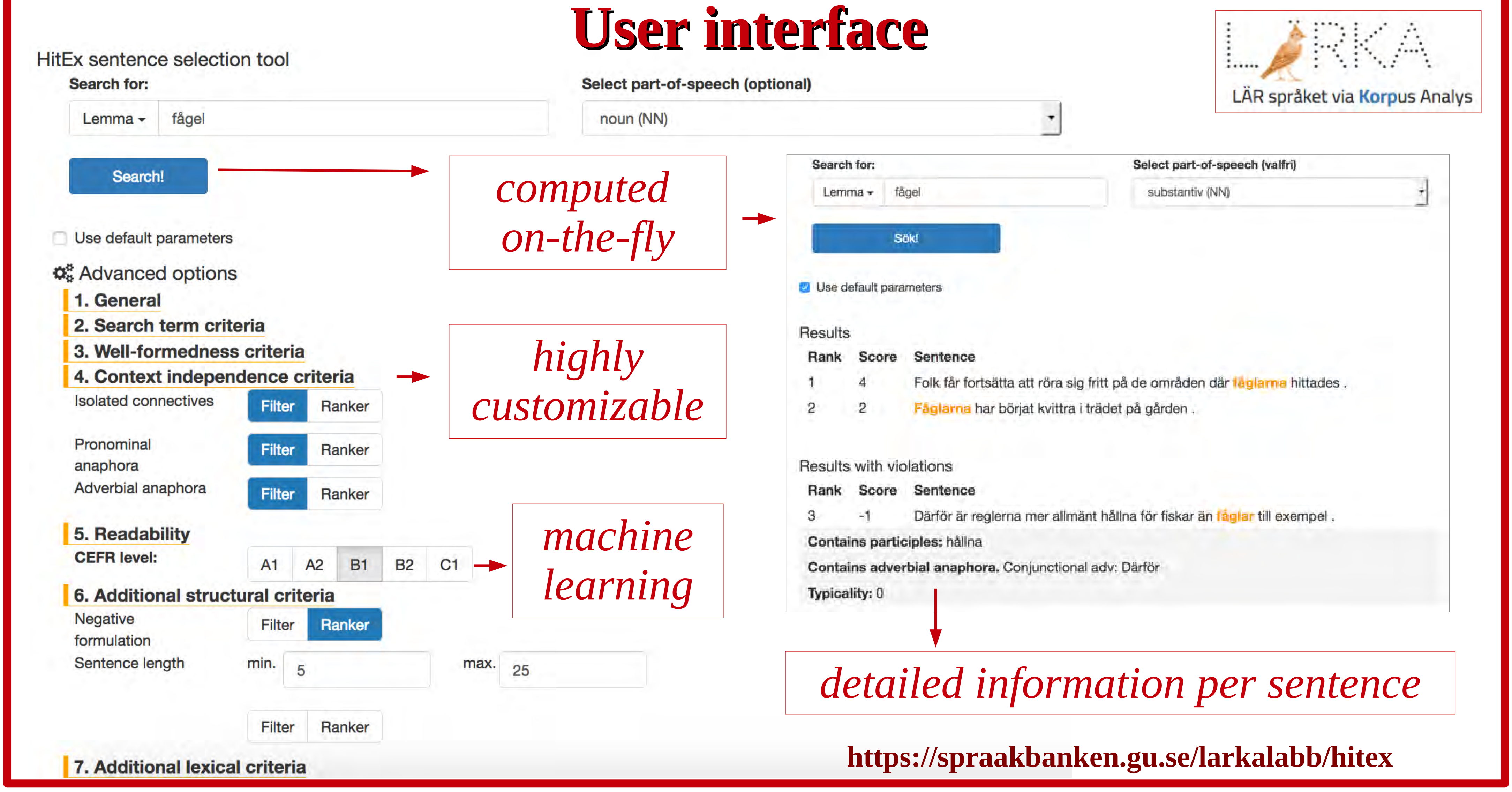


A1: harder than it should be
A2, B1: appropriate difficulty

Architecture



User interface



HitEx sentence selection tool

Search for: Lemma ▾ fågel

Select part-of-speech (optional): noun (NN)

computed on-the-fly

□ Use default parameters

Advanced options

1. General
2. Search term criteria
3. Well-formedness criteria
4. Context independence criteria

Isolated connectives: [✓ Filter] Ranker

Pronominal anaphora: [✓ Filter] Ranker

Adverbial anaphora: [✓ Filter] Ranker

highly customizable

5. Readability

CEFR level: A1 A2 [✓ B1] B2 C1

machine learning

6. Additional structural criteria

Negative formulation: Filter [✓ Ranker]

Sentence length: min. 5 max. 25

Filter Ranker

7. Additional lexical criteria

Search for: Lemma ▾ fågel

Select part-of-speech (valfri): substantiv (NN)

☑ Use default parameters

Results

Rank	Score	Sentence
1	4	Folk får fortsätta att röra sig fritt på de områden där fåglarna hittades .
2	2	Fåglarna har börjat kvittra i trädet på gården .

Results with violations

Rank	Score	Sentence
3	-1	Därför är reglerna mer allmänt hållna för fiskar än fåglar till exempel .

Contains participles: hållna

Contains adverbial anaphora. Conjunctional adv: Därför

Typicality: 0

detailed information per sentence

LÄRKA
LÄR språket via Korpus Analys

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