

Selection of Suitable Sentences for Language Learning Exercises: extending the initial algorithm

Ildikó Pilán, Elena Volodina, Richard Johansson

**Språkbanken (Swedish Language Bank)
University of Gothenburg, Sweden**

**February, 12 2015
Vienna, Austria**



**UNIVERSITY OF
GOTHENBURG**

Aims

- **Extend** the initial selection algorithm:
 - Increase the number of aspects taken into consideration
 - Use machine learning methods besides heuristic rules
- Create a **module** for experimenting with sentence selection within our free online language learning platform, **Lärka**
- HitEx: **HIT**ta **EX**empel [Find examples] or **HIT EX**amples
 - AIM: **select sentences** based on their **readability**

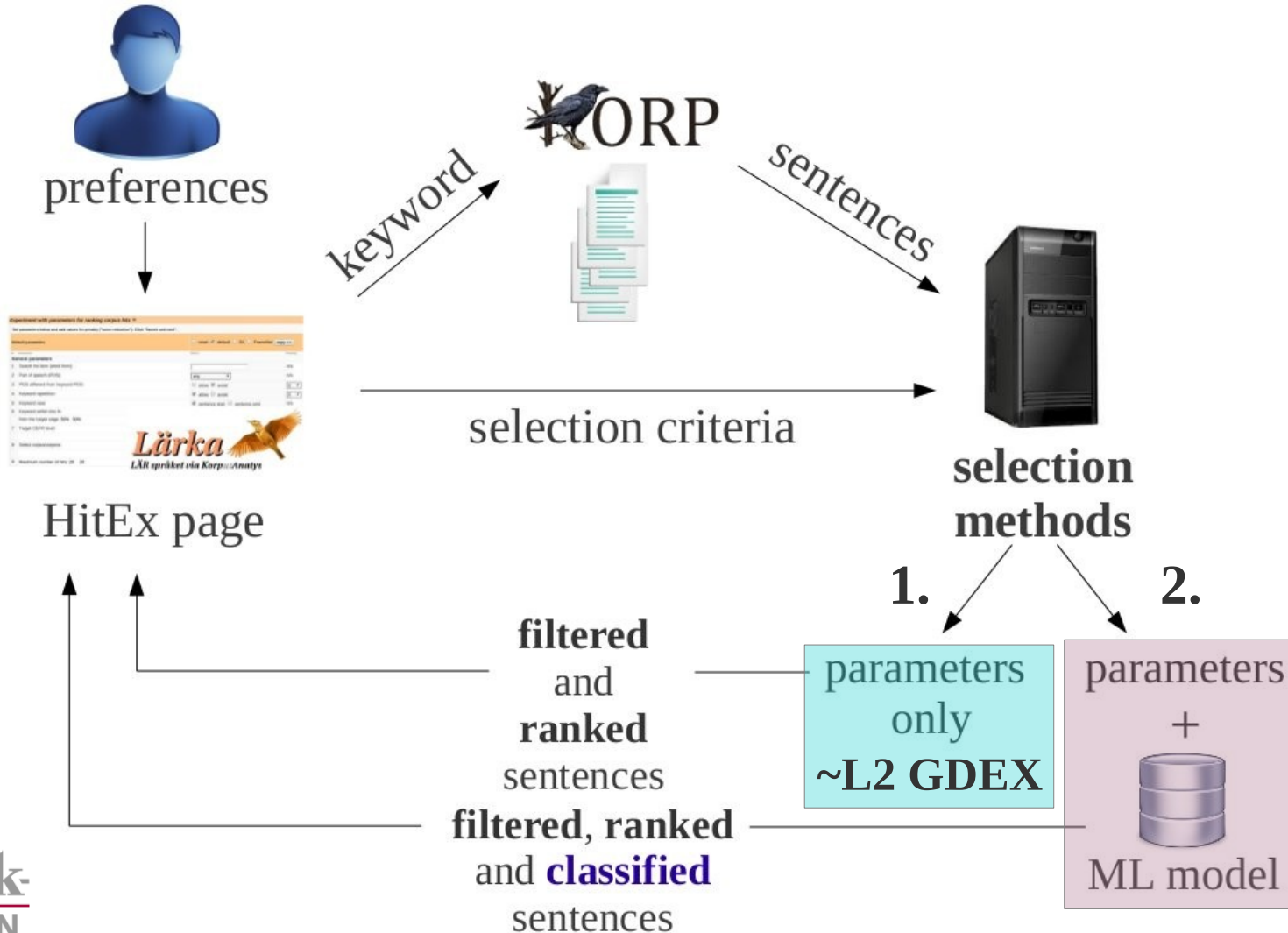
↓

≈ how difficult a text
is for the reader
(**CEFR** levles: A1-C2)

HitEx: Resources used

- Swedish corpora of a variety of genres (Korp)
- **COCTAILL corpus:**
 - Corpus of CEFR Textbooks as Input for Learner Level modeling
 - Collection of course book texts for L2 Swedish
 - 5 proficiency levels: A1 - C1
- **Kelly list:** a frequency-based vocabulary list with CEFR levels for each item

Overview of HitEx



HitEx: user interface

Experiment with parameters for ranking corpus hits

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

Default parameters

reset

Target user group parameters

(A1) (A2) B1 B2 C1+ GDEX

default settings

Nr	Parameter	Value1	Penalty1
----	-----------	--------	----------

General parameters

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="range" 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 n/a

values ←

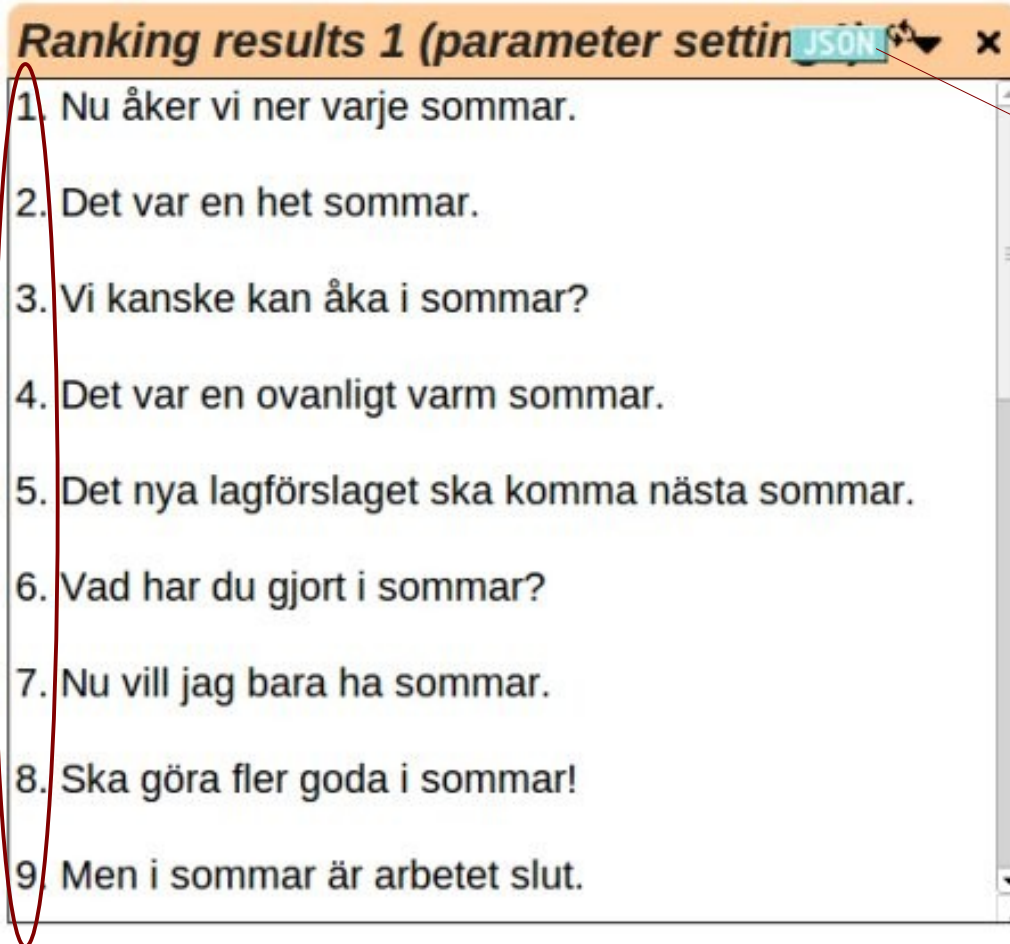
penalty points

list of parameters



UNIVERSITY OF
GOTHENBURG

HitEx: results



The screenshot shows a web interface titled "Ranking results 1 (parameter settin JSON)". It displays a list of 9 ranked Swedish sentences. A red oval highlights the list, and a red arrow points from the word "ranks" to it. Another red arrow points from the "JSON" dropdown menu to the text "additional information". At the bottom, a button labeled "Download ranking results in a file" is circled in red.

Ranking results 1 (parameter settin JSON)

1. Nu åker vi ner varje sommar.
2. Det var en het sommar.
3. Vi kanske kan åka i sommar?
4. Det var en ovanligt varm sommar.
5. Det nya lagförslaget ska komma nästa sommar.
6. Vad har du gjort i sommar?
7. Nu vill jag bara ha sommar.
8. Ska göra fler goda i sommar!
9. Men i sommar är arbetet slut.

Download ranking results in a file

ranks

additional
information

Rule-based approach: parameters

Structural parameters

allow / avoid

#10	Sentence length	#17	Pronoun / noun ratio
#11	Average word length	#18	Relative pronoun %
#12	Elliptic sentences	#19	Adverb %
#13	Negative formulations	#20	Preposition %
#14	Modal verbs	#21	Con- and subjunction %
#15	Participles	#22	Average dependency depth
#16	S-verbs		

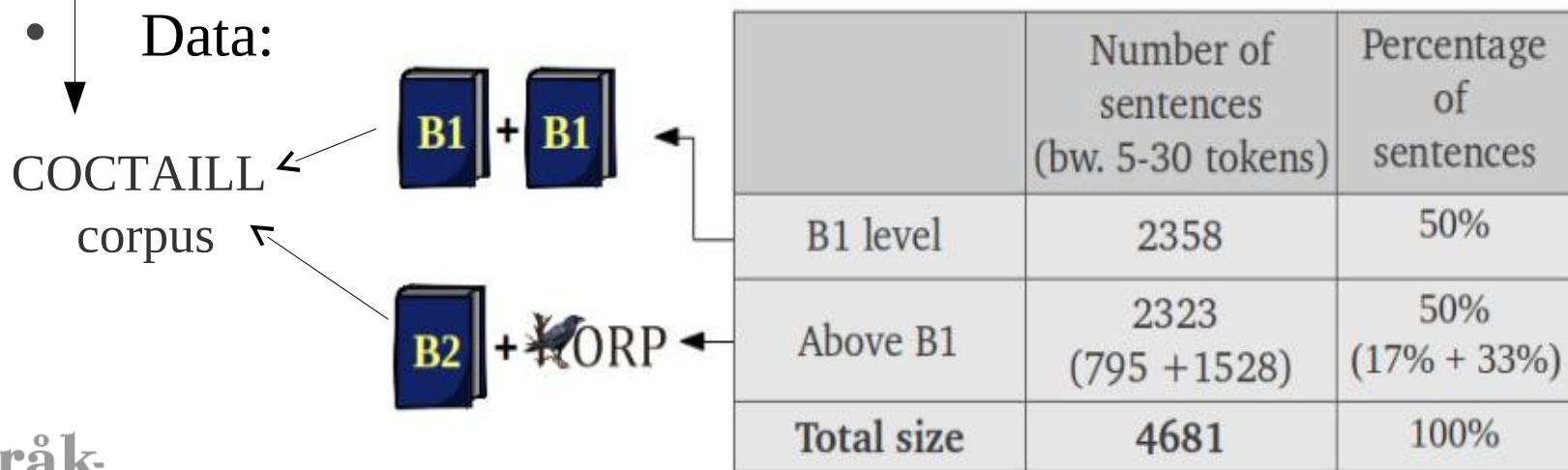
Lexical parameters

#23	Penalize words below a freq. limit	#25	Proper Names
#24	% of words above target CEFR level	#26	Abbreviations



Machine Learning for CEFR level classification

- **Supervised** classification methods – (SVM algorithm)
- **28 features** (mostly based on linguistic information e.g. parts of speech, dependency relations)

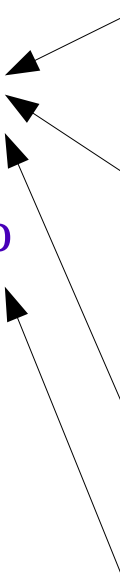


Classification results and top features

Classifier	Accuracy	F1	within B1 Precision	within B1 Recall
Baseline	0.50	0.66	0.50	1.00
All	0.71	0.70	0.73	0.68

Rank	Feature ID	SVM weight
1	Percentage of difficult words	0.576
2	Average number of senses per word	0.438
3	Nr of difficult words	0.422
4	Sentence length (nr of tokens)	0.258
5	Nr of modifiers	0.223
6	Average frequency in Kelly word list	0.215
7	Nominal Ratio (Nominal to verbal cat.s)	0.132
8	Adverb variation	0.114

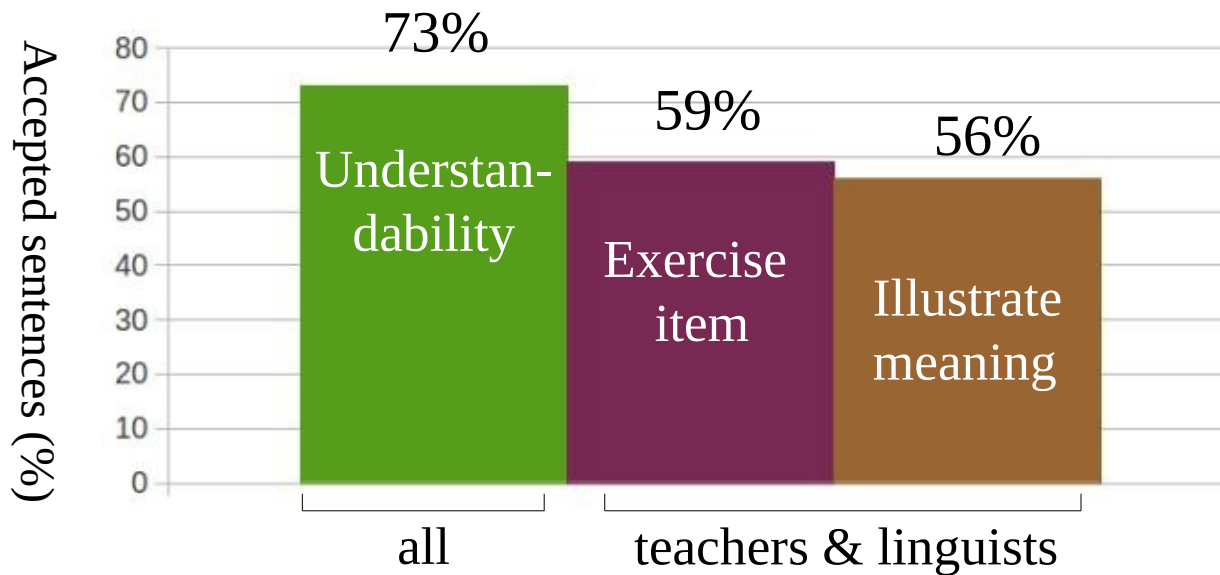
Mostly lexical and morphological features



Evaluation

- **Purpose**: evaluate whether **sentences** selected with our system from generic corpora are **suitable** for B1-level students
- 200 sentences selected with both heuristics-only and the combined approach
- **Participants (34)**:
 - 26 Students at B1 level
 - 3 Teachers of Swedish as L2
 - 5 Linguists (+ one lexicographer)

Evaluation: results



- Overall 7 out of 10 sentences rated as understandable
- 5% more sentences selected with the heuristics-only approach "accepted" by raters

Current work

- Machine learning model extended to 5 **CEFR levels (A1-C1)**
- Additional features (additional morpho-syntactic info etc.)
- Experiments repeated with data **annotated** at **sentence level**:
 - **63%** accuracy for distinguishing 5 CEFR levels
 - **92%** adjacent accuracy (=errors within 1 class distance)
- **text-level** experiments: 81% for 5-level classification

Conclusion and future work

- An approach for the **selection of readable sentences** for **language learning** purposes (7 out of 10 understandable)
- Sentences used in **automatically generated exercises**

Future work:

- Re-evaluate new models with users
- Optimization, increasing user-friendliness etc.

Demo

Thank you!

Combined approach

