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

Språk-
BANKEN

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)

Språk-BANKEN

UNIVERSITY OF GOTHENBURG

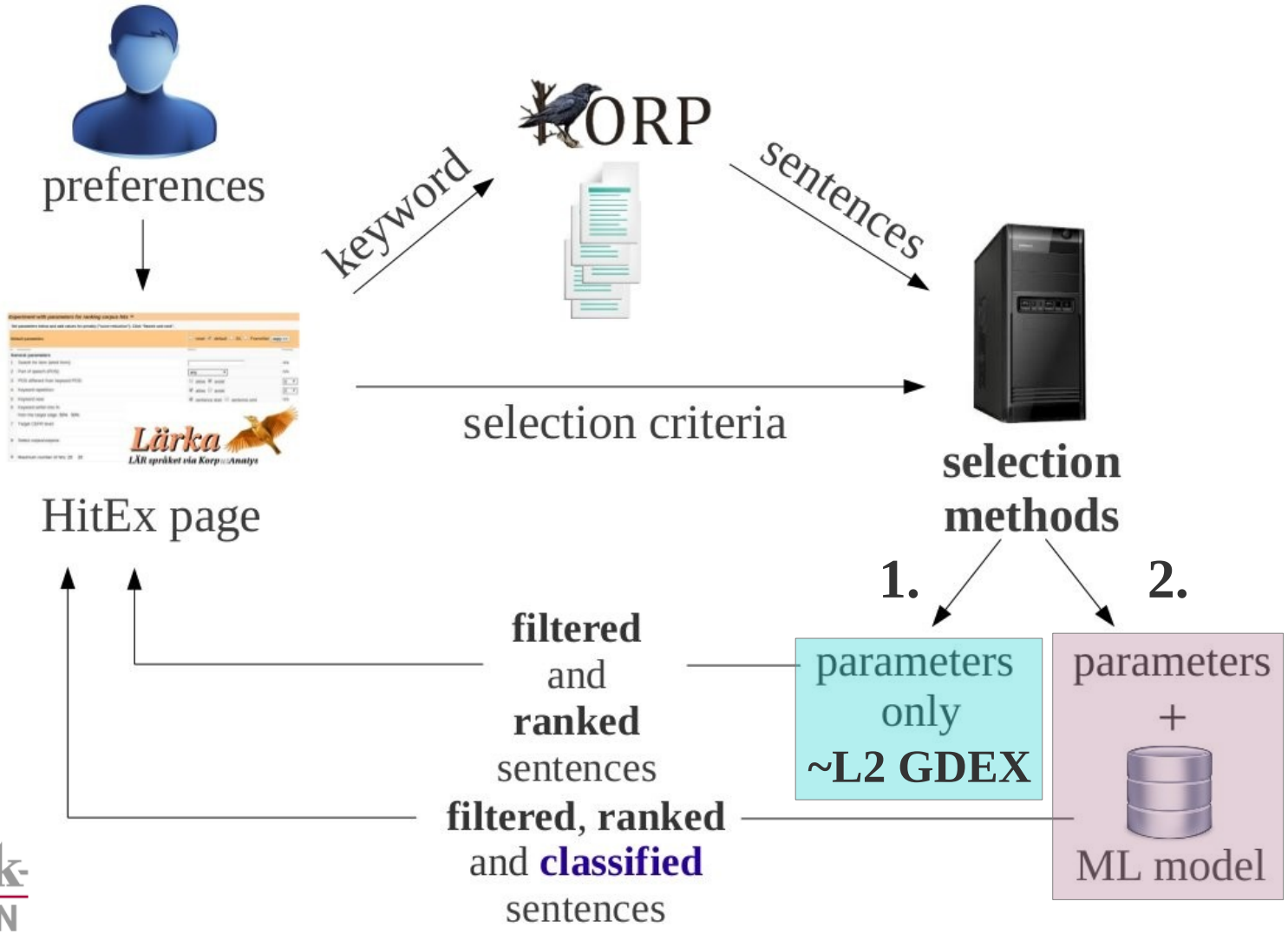
HitEx: Resources used

- Swedish corpora of a variety of genres (Korp)
- **COCTAILL corpus**:
  - **C**orpus **o**f **C**EFR **T**extbooks **a**s **I**nput for **L**earner **L**evel 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

Språk-BANKEN

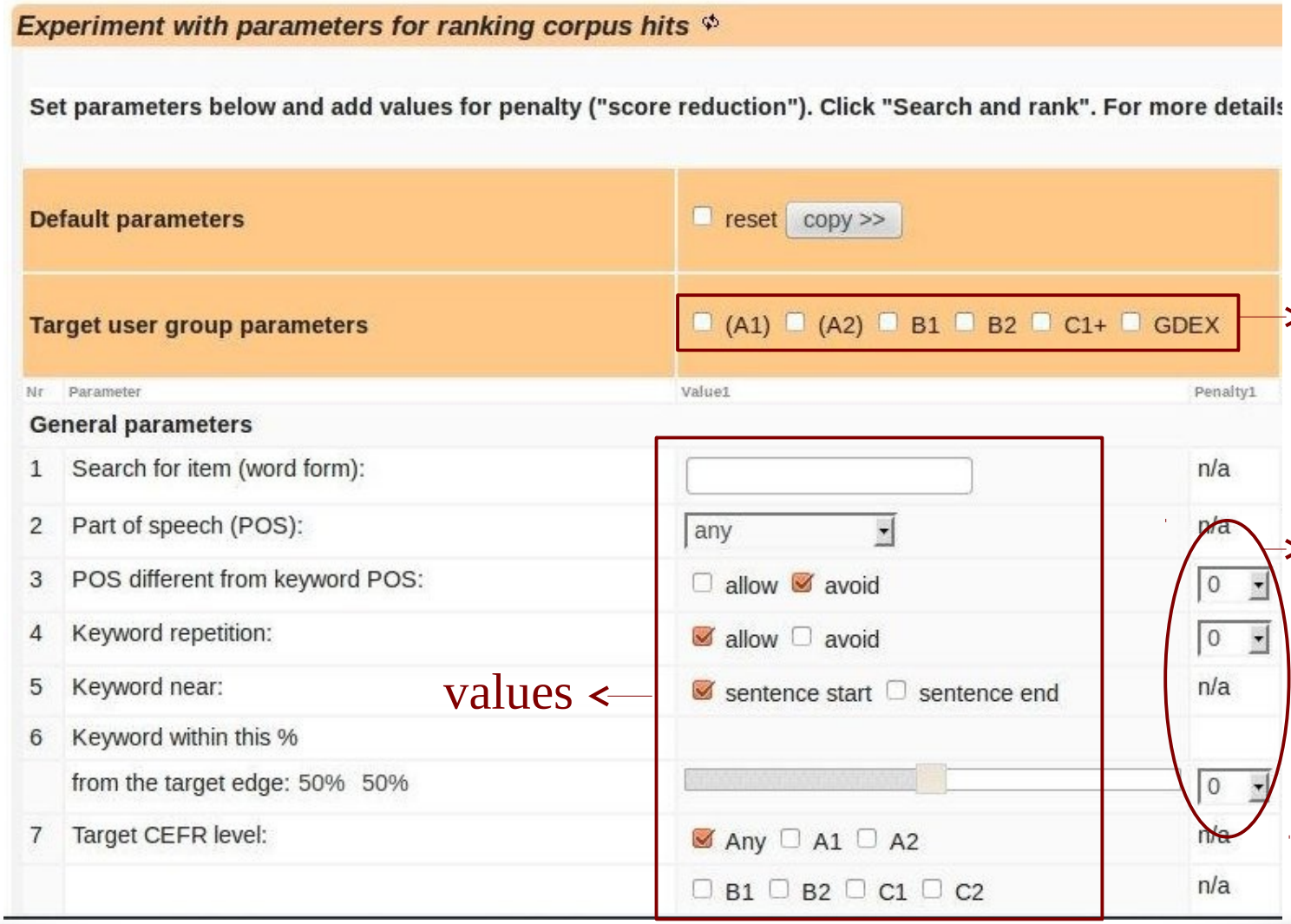
UNIVERSITY OF GOTHENBURG

Overview of HitEx



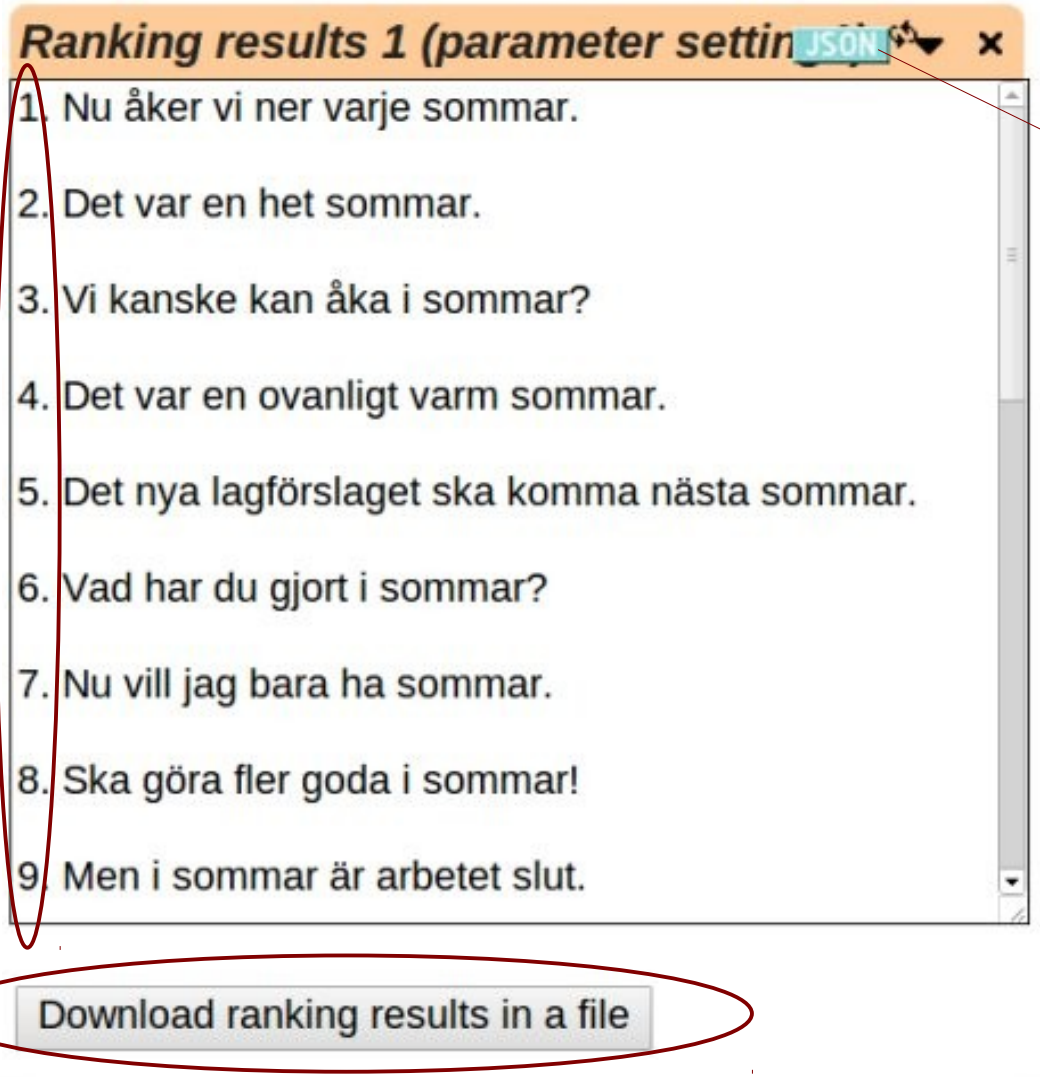
Språk-BANKEN

HitEx: user interface



UNIVERSITY OF GOTHENBURG

HitEx: results



ranks

additional information

Språk-BANKEN

UNIVERSITY OF GOTHENBURG

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

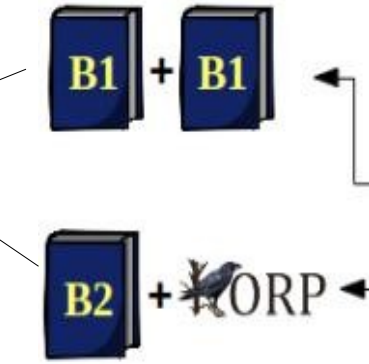
Språk-BANKEN

UNIVERSITY OF GOTHENBURG

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)
- Data:

COCTAILL corpus



	Number of sentences (bw. 5-30 tokens)	Percentage of sentences
B1 level	2358	50%
Above B1	2323 (795 +1528)	50% (17% + 33%)
Total size	**4681**	100%

Språk-BANKEN

UNIVERSITY OF GOTHENBURG

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

Mostly lexical and morphological features

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

Språk-BANKEN

UNIVERSITY OF GOTHENBURG

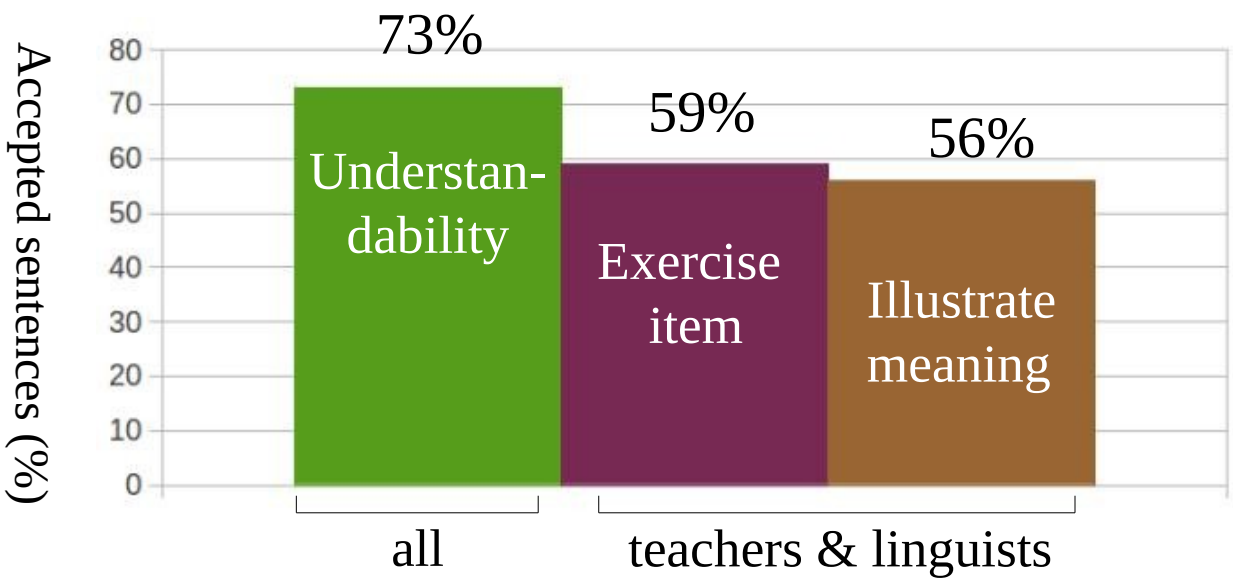
Evaluation

- **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)

Språk-BANKEN

UNIVERSITY OF GOTHENBURG

Evaluation: results



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

Språk-BANKEN

UNIVERSITY OF GOTHENBURG

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

Språk-BANKEN

UNIVERSITY OF GOTHENBURG

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.

Språk-BANKEN

UNIVERSITY OF GOTHENBURG

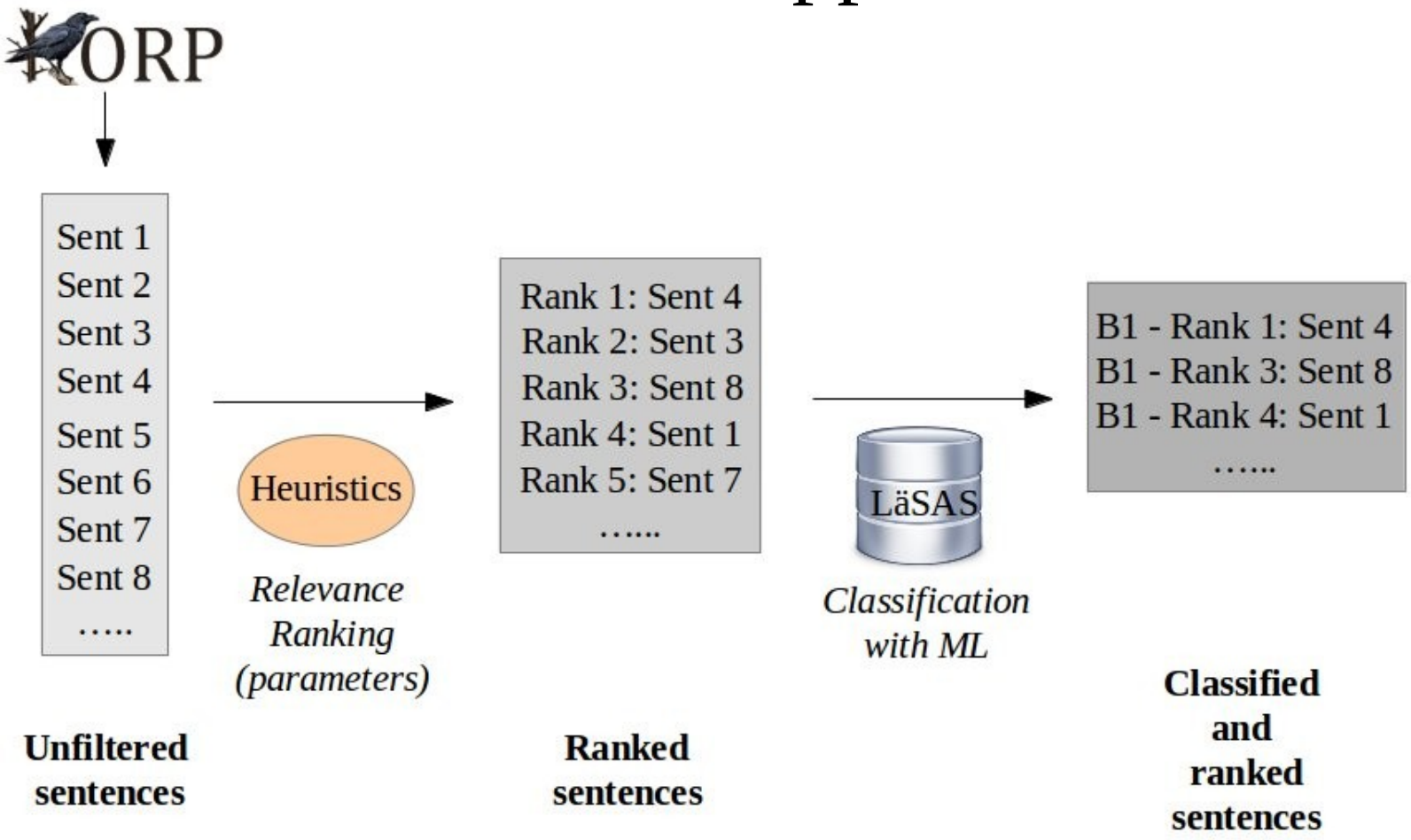
Demo

Thank you!

Språk-
BANKEN

UNIVERSITY OF
GOTHENBURG

Combined approach



Språk-BANKEN

UNIVERSITY OF GOTHENBURG