language and translation technology team

LexComSpaL2: A LEXICAL COMPLEXITY CORPUS FOR SPANISH AS A FOREIGN LANGUAGE

JASPER DEGRAEUWE & PATRICK GOETHALS

Sentence: The paediatric waiting room is filled with children sniffling and coughing.

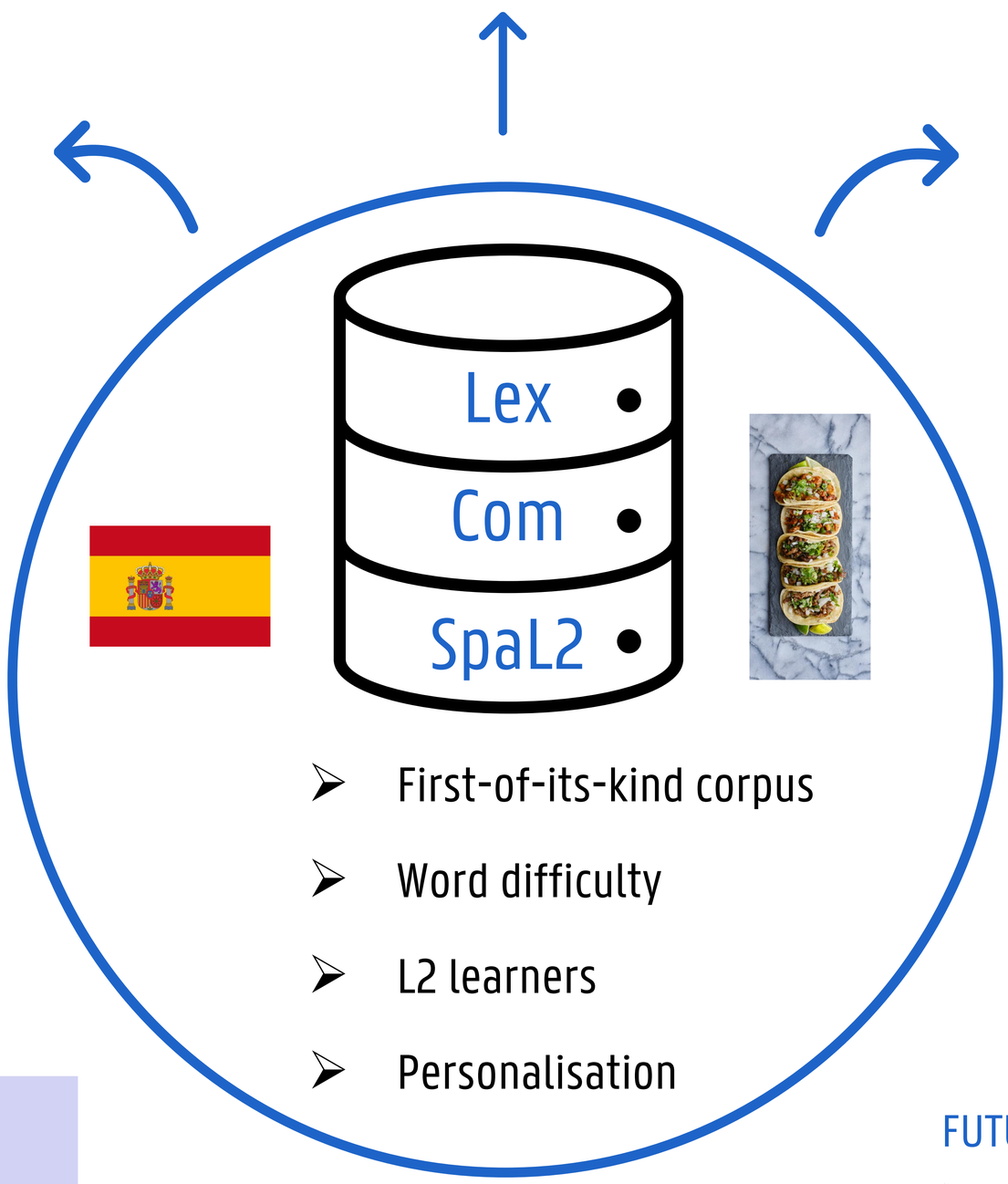
Content word	PARTP1	PARTP2	…	PARTP26
paediatric	5	3	⋯	4
waiting	1	2	⋯	1
room	1	1	⋯	1
filled	1	2	⋯	1
children	1	1	⋯	1
sniffling	3	4	⋯	4
coughing	3	4	⋯	3

- Set of **200 representative sentences**
  - 4 domains (≈ learning particular topics; Webb & Nation, 2017)
  - Pedagogically suitable (specific selection method; Pilán et al., 2016)
- **Tailor-made lexical complexity prediction scale** based on vocabulary knowledge continuum (no knowledge → passive mastery → active mastery)

LCP label	Description
1	I know this word and its meaning, and I also use it actively in speaking/writing.
2	I know this word and its meaning, but I might not be able to use it on the top of my head in an oral/written conversation. When I have some time to think, however, I do think I would use it naturally.
3	I have heard/seen this word before and given the context I think that I more or less know what it means, but I do not see myself using this word actively.
4	This word sounds vaguely familiar and based on the context I could make an educated guess about its meaning, but I would still need a dictionary to be able to understand its exact meaning.
5	This word does not sound familiar at all to me, and even based on the context I do not know what it means, so I would definitely need a dictionary to get to know its meaning.

CONTEXT

- **Second language acquisition** (SLA)
  - Correlation between text comprehension and vocabulary knowledge (Schmitt et al., 2011)
  - 95% to 98% of words in running text should be known (Laufer & Ravenhorst-Kalovski, 2010)
- **Identify difficult words to improve tools and resources** used in SLA
  - Tools: reading assistants, exercise generators, etc.
  - Resources: graded readers, graded word lists, etc.
  - Long tradition of manual identification



RESULTS

- Research gaps addressed
  - Continuous instead of binary labels (North et al., 2023)
  - L2 learners as target audience
  - Spanish as target language
  - Personalised predictions
- Methodology
  - 200 **representative** sentences
  - **5-point scale** of lexical complexity prediction (LCP)
  - **26 L2 Spanish students** as annotators
- End goal: train **machine learning classifiers** that predict word difficulty for individual L2 learners

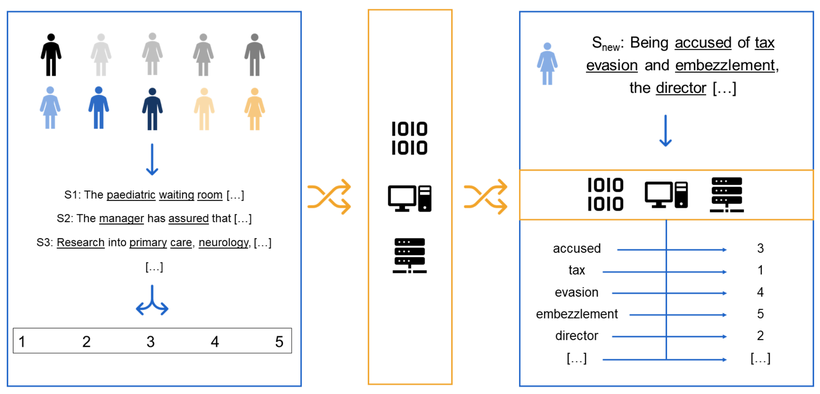
STATISTICS

- Students of 3 proficiency levels: B1, B2, C1
- L1 students = Dutch
- 200 sentences (50 / domain)
- 2,240 target words (11.2 / sentence)
- 58,240 observations
- Low inter-annotator agreement (κ = 0.23) → need for personalised predictions

FUTURE RESEARCH

- Extend L1s
- Features
  - Current: years of experience and proficiency level study year
  - Potential additions: results on proficiency tests (e.g., cloze tests)
- Identify most "valuable" sentences
- Release baseline classifier (LLM-based?)

LCP CLASSIFIER



GHENT UNIVERSITY

REFERENCES
[1] Laufer, B., & Ravenhorst-Kalovski, G. C. (2010). Lexical threshold revisited: Lexical text coverage, learners' vocabulary size and reading comprehension. *Reading in a Foreign Language*, *22*(1), 15–30.
[2] North, K., Zampieri, M., & Shardlow, M. (2023). Lexical Complexity Prediction: An Overview. *ACM Computing Surveys*, *55*(9), 1–42.
[3] Pilán, I., Volodina, E., & Borin, L. (2016). Candidate sentence selection for language learning exercises: From a comprehensive framework to an empirical evaluation. *Revue Traitement Automatique Des Langues*, *57*(3), 67–91.
[4] Schmitt, N., Jiang, X., & Grabe, W. (2011). The Percentage of Words Known in a Text and Reading Comprehension. *The Modern Language Journal*, *95*(1), 26–43.
[5] Webb, S., & Nation, I. S. P. (2017). *How vocabulary is learned*. Oxford University Press.