FACULTY OF ARTS AND PHILOSOPHY

DEPARTMENT OF TRANSLATION, INTERPRETING AND COMMUNICATION
RESEARCH GROUPS LT³ AND MULTIPLES

LexComSpaL2: A Lexical Complexity Corpus for Spanish as a Foreign Language

Jasper Degraeuwe & Patrick Goethals – LREC-COLING 2024

GHENT UNIVERSITY

CORE CONCEPTS

- Foreign/second language acquisition (SLA) → L2 Spanish
- Lexical complexity prediction

GHENT UNIVERSITY

SLA

GHENT
UNIVERSITY

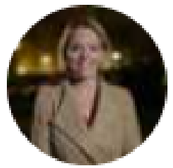
IDENTIFICATION OF DIFFICULT WORDS

- Text comprehension and vocabulary knowledge positively correlated (Schmitt et al., 2011)
- 95 to 98% of words in text should be known for optimal comprehension (Laufer & Ravenhorst-Kalovski, 2010)

GHENT UNIVERSITY

22:12

Sunak kicks off campaign with familiar speech



Hannah Miller
Political correspondent

The first Conservative campaign event at the Excel centre was essentially a photo opportunity - a repetition of many of the lines the prime minister tested out during his Downing Street pitch to the nation.

The party members who surrounded their leader with placards were quickly whisked away afterwards, with little opportunity to find out what any of them make of the timing of this election or their party's prospects.

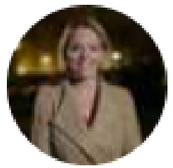
Rishi Sunak will be grateful that at least this time it was indoors.

Share

Source: BBC

22:12

Sunak kicks off campaign with familiar speech



Hannah Miller
Political correspondent

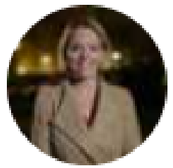
The first Conservative campaign event at the Excel centre was essentially a photo opportunity - a repetition of many of the lines the prime minister tested out during his Downing Street pitch to the nation.

The party members who surrounded their leader with placards were quickly whisked away afterwards, with little opportunity to find out what any of them make of the timing of this election or their party's prospects.

Rishi Sunak will be grateful that at least this time it was indoors.

22:12

Sunak kicks off campaign with familiar speech



Hannah Miller
Political correspondent

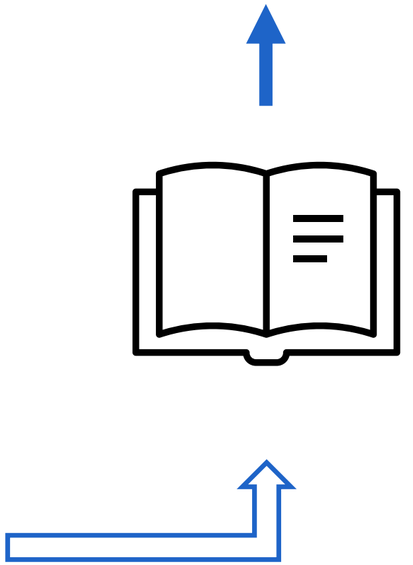
The first Conservative campaign event at the Excel centre was essentially a photo opportunity - a repetition of many of the lines the prime minister tested out during his Downing Street pitch to the nation.

The party members who surrounded their leader with placards were quickly whisked away afterwards, with little opportunity to find out what any of them make of the timing of this election or their party's prospects

Rishi Sunak will be grateful that at least this time it was indoors.

Share

- Placard = […]
- To be whisked away = […]
- Prospect = […]



KWIC | Plot | File View | Cluster | N-Gram | Collocate | Word | Keyword | Wordcloud

Keyword Types 163/5558 **Keyword Tokens** 8595/34246 **Page Size** 100 hits 1 to 100 of 163 hits

	Type	Rank	Freq_Tar	Freq_Ref	Range_Tar	Range_Ref	Keyness (Likelihood)	Keyness (Effect)
1	god	1	188	4	16	2	403.840	0.011
2	of	2	1677	1946	17	38	377.452	0.089
3	christian	3	89	2	13	2	190.266	0.005
4	doctrine	4	75	0	6	0	175.264	0.004
5	religion	5	79	1	8	1	174.611	0.005
6	church	6	77	2	13	2	162.775	0.004
7	divine	7	57	1	9	1	123.821	0.003
8	social	8	62	4	10	3	117.666	0.004
9	sacred	9	59	3	8	3	116.065	0.003
10	theology	10	49	0	7	0	114.480	0.003
11	scripture	11	47	0	5	0	109.806	0.003
12	science	12	57	4	3	3	106.629	0.003
13	theological	13	44	0	9	0	102.794	0.003
14	faith	13	44	0	12	0	102.794	0.003
15	justification	15	41	0	2	0	95.783	0.002

Source: AntConc

KWIC | Plot | File View | Cluster | N-Gram | Collocate | Word | Keyword | Wordcloud

Keyword Types 163/5558 **Keyword Tokens** 8595/34246 **Page Size** 100 hits 1 to 100 of 163 hits

	Type	Rank	Freq_Tar	Freq_Ref	Range_Tar	Range_Ref	Keyness (Likelihood)	Keyness (Effect)
1	god	1	188	4	16	2	403.840	0.011
2	of	2	1677	1946	17	38	377.452	0.089
3	christian	3	89	2	13	2	190.266	0.005
4	[redacted]							
5	religion	5	79	1	8	1	174.611	0.005
6	church	6	77	2	13	2	162.775	0.004
7	divine	7	57	1	9	1	123.821	0.003
8	social	8	62	4	10	3	117.666	0.004
9	sacred	9	59	3	8	3	116.065	0.003
10	theology	10	49	0	7	0	114.480	0.003
11	[redacted]							
12	science	12	57	4	3	3	106.629	0.003
13	[redacted]							
14	faith	13	44	0	12	0	102.794	0.003
15	justification	15	41	0	2	0	95.783	0.002

Source: AntConc

LEXICAL COMPLEXITY PREDICTION (LCP)

GHENT UNIVERSITY

LCP: SCALE

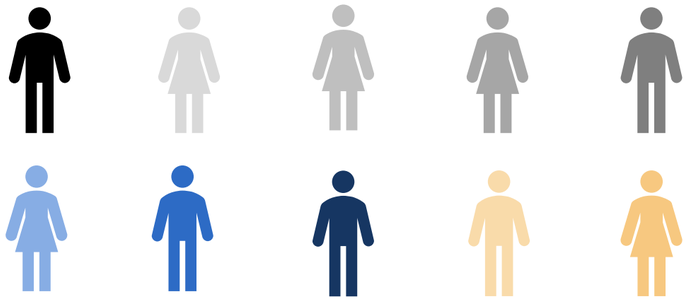
LCP label	Description
1	Very easy: this word is very familiar to me
2	Easy: I am aware of the meaning of this word
3	Neutral: this word is neither difficult nor easy
4	Difficult: the meaning of this word is unclear to me, but I may be able to infer it from the sentence
5	Very difficult: I have never seen this word before / this word is very unclear to me

LCP: EXAMPLE

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

Content word	LCP label
paediatric	5
waiting	1
room	1
filled	2
children	1
sniffling	4
coughing	3

LCP → MACHINE LEARNING CLASSIFIER



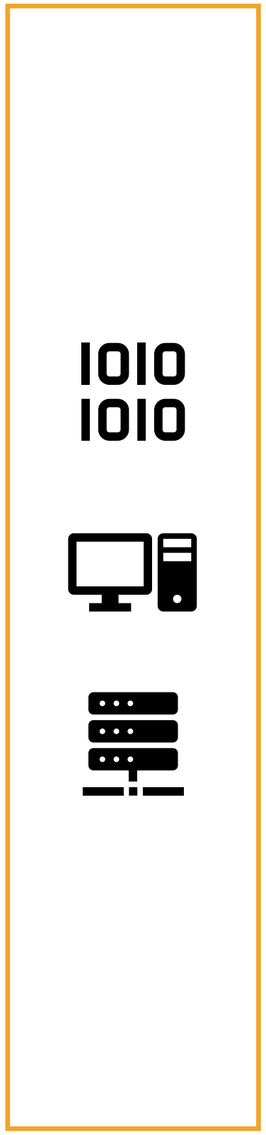
S1: The paediatric waiting room […]

S2: The manager has assured that […]

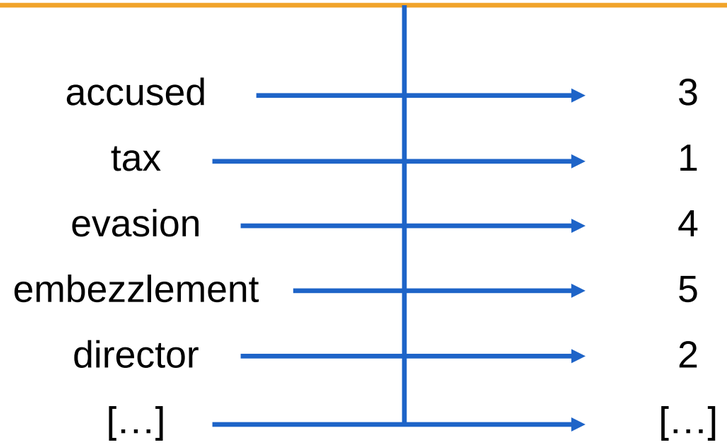
S3: Research into primary care, neurology, […]

[…]

1	2	3	4	5



S_{new}: Being accused of tax evasion and embezzlement, the director […]



LexComSpaL2

GHENT
UNIVERSITY

DATA COLLECTION

- Representative dataset of 200 sentences
  - **4 domains** (economics, health, law, and migration): specialised vocabulary knowledge is crucial to learning a particular topic (Webb and Nation, 2017)
  - **Pedagogically suitable corpus sentences** selected according to specific framework (Pilán et al., 2016)
- Target words = all nouns, verbs, and adjectives

GHENT UNIVERSITY

DATA LABELLING

- Participants: 26 L2 Spanish students (L1 = Dutch)
- Different proficiency levels (PLs)
  - PL1: 2nd year L2 Spanish career at university (≈ B1)
  - PL2: 3rd year (≈ B2)
  - PL3: 4th year (≈ C1)
- LCP descriptions adapted to vocabulary knowledge continuum (Schmitt, 2019): no knowledge → receptive knowledge → productive knowledge

LCP: ADAPTED SCALE

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.

DATA LABELLING

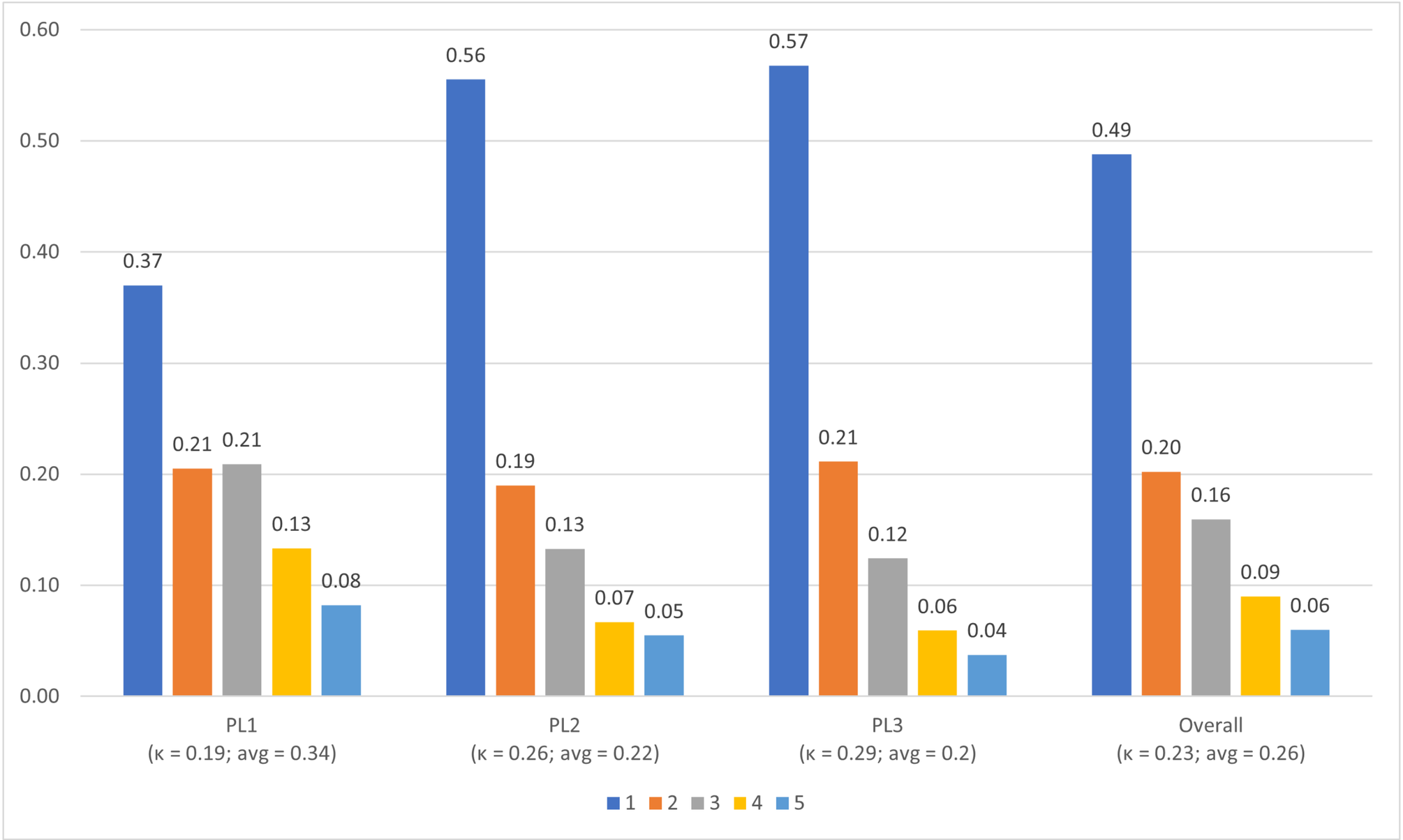
Sentence: The paediatric waiting room is filled with children sniffling and coughing.

Content word	PARTP1 (PL1)	PARTP2 (PL1)	[…]	PARTP26 (PL3)
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

DATASET STATISTICS

Sentences		Target words		Frequency target words	
Total (per domain)	Average length (SD)	Total (unique)	Average per sentence (SD)	Frequency range	Percentage
200 (50)	28.85 (2.98)	2,240 (1,863)	11.2 (2.14)	1 - 1,000 1,001 - 2,000 2,001 - 3,000 3,001 - 4,000 4,001 - 5,000 >5,000	0.24 0.14 0.09 0.07 0.05 0.41

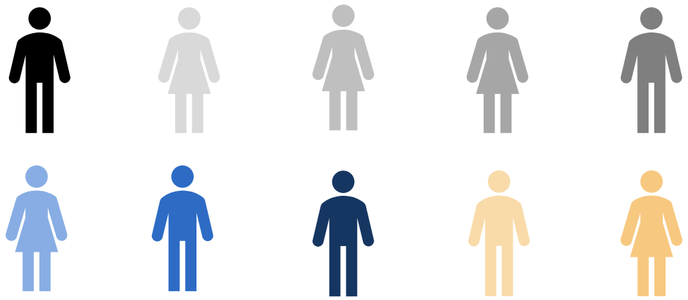
DATASET STATISTICS



DATASET SAMPLE

Sentence ID	Sentence text	Target word	Average judgement	Individual judgements
1_1	El directivo, que ha celebrado un almuerzo de Navidad con la prensa, ha asegurado que [...] ('The manager, who has held a Christmas lunch with the press, has assured that [...]')	directivo	{PL1: 0.3, PL2: 0.34, PL3: 0.22, overall: 0.29}	{PARTP1: 3, ..., PARTP26: 1}
		celebrado	{PL1: 0.13, PL2: 0, PL3: 0.06, overall: 0.07}	{PARTP1: 2, ..., PARTP26: 1}
		...		
...				
4_50	Las investigaciones sobre atención primaria, neurología, oncología médica y microbiología van después, [...] ('Research into primary care, neurology, medical oncology and microbiology comes after, [...]')	investigaciones	{PL1: 0.28, PL2: 0.03, PL3: 0.06, overall: 0.13}	{PARTP1: 1, ..., PARTP26: 1}
		atención	{PL1: 0.2, PL2: 0.03, PL3: 0.03, overall: 0.1}	{PARTP1: 2, ..., PARTP26: 1}
		...		

LCP → MACHINE LEARNING CLASSIFIER



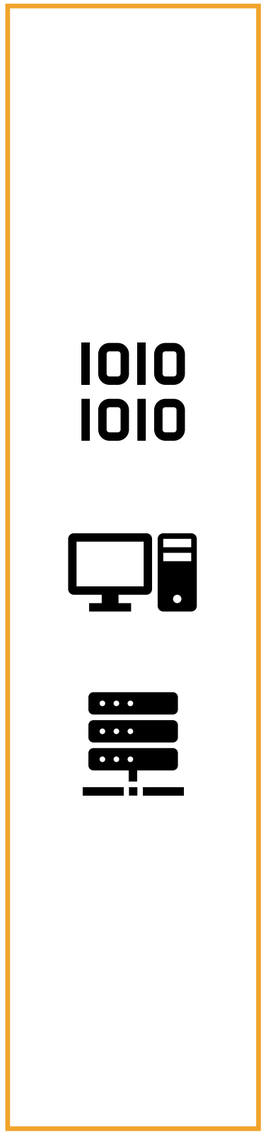
S1: The paediatric waiting room […]

S2: The manager has assured that […]

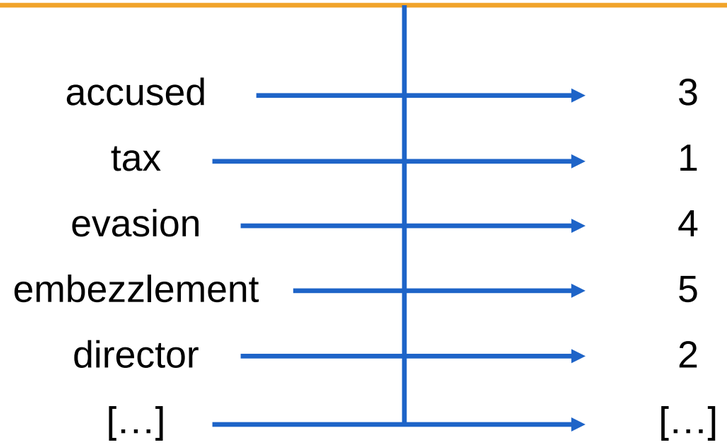
S3: Research into primary care, neurology, […]

[…]

1	2	3	4	5



S_{new}: Being accused of tax evasion and embezzlement, the director […]



CONCLUSION

GHENT
UNIVERSITY

CONTRIBUTIONS

- First of its kind: LCP corpus for L2 Spanish
- Continuous predictions (←→ traditionally binary)
- Representativeness
- Individual annotations

GHENT UNIVERSITY

LIMITATIONS

- Single words
- No information on word senses yet
- Participants = L1 Dutch

GHENT UNIVERSITY

FUTURE WORK

- Expand L1s
- Release baseline ML classifier

GHENT UNIVERSITY

LINKS

- Dataset repository: https://github.com/JasperD-UGent/LexComSpaL2

GHENT UNIVERSITY

REFERENCES

- Batia Laufer and Geke C. Ravenhorst-Kalovski. 2010. Lexical threshold revisited: Lexical text coverage, learners' vocabulary size and reading comprehension. *Reading in a Foreign Language*, 22(1):15–30.
- Kai North, Marcos Zampieri, and Matthew Shardlow. 2023. Lexical Complexity Prediction: An Overview. *ACM Computing Surveys*, 55(9):1–42.
- Ildikó Pilán, Elena Volodina, and Lars Borin. 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.
- Norbert Schmitt. 2019. Understanding vocabulary acquisition, instruction, and assessment: A research agenda. *Language Teaching*, 52(02):261–274.
- Norbert Schmitt, Xiangying Jiang, and William Grabe. 2011. The Percentage of Words Known in a Text and Reading Comprehension. *The Modern Language Journal*, 95(1):26–43.
- Matthew Shardlow, Michael Cooper, and Marcos Zampieri. 2020. CompLex — A New Corpus for Lexical Complexity Prediction from Likert Scale Data. In *Proceedings of the 1st Workshop on Tools and Resources to Empower People with REAding DIfficulties (READI)*, pages 57–62, Marseille, France. European Language Resources Association.
- Stuart Webb and Paul Nation. 2017. *How vocabulary is learned*. Oxford Handbooks for Language teachers. Oxford University Press, Oxford.

FACULTY OF ARTS AND PHILOSOPHY

Jasper Degraeuwe

PhD researcher

DEPARTMENT OF TRANSLATION,
INTERPRETING AND COMMUNICATION

E Jasper.Degraeuwe@UGent.be

www.ugent.be

Universiteit Gent
@ugent
@ugent
Ghent University

GHENT UNIVERSITY