INDIVIDUALISED PREDICTION OF PERCEIVED VOCABULARY DIFFICULTY: FROM DATASET TO CLASSIFIER

Jasper Degraeuwe – 20 March 2025 – Tübingen

LATILL workshop on "Foreign language learning and proficiency-rated reading materials"

GHENT UNIVERSITY

ABOUT ME

- Ghent University (Belgium)

- PhD on **Intelligent Computer-Assisted Language Learning** (ICALL)

  ICALL ≈ CALL + AI and NLP

- Doctor-assistant on **educational technologies and AI for language learning**

GHENT
UNIVERSITY

WHAT BROUGHT ME TO TÜBINGEN?

WHAT BROUGHT ME TO TÜBINGEN?

WHAT BROUGHT ME TO TÜBINGEN?

Individualised prediction of perceived vocabulary difficulty: from dataset to classifier

CORE CONCEPTS

Individualised prediction of perceived vocabulary difficulty: from dataset to classifier

CORE CONCEPTS

Individualised prediction of perceived vocabulary difficulty: from dataset to classifier

GHENT
UNIVERSITY

CORE CONCEPTS

Individualised prediction of perceived vocabulary difficulty: from dataset to classifier

GHENT
UNIVERSITY

Individualised prediction of perceived vocabulary difficulty: from dataset to classifier

Individualised prediction of perceived vocabulary difficulty: from dataset to classifier

CORE CONCEPTS

Individualised prediction of perceived vocabulary difficulty: from dataset to classifier

1. Vocabulary learning < foreign/second language acquisition (SLA)

1. SLA

GHENT
UNIVERSITY

BACKGROUND

– Text comprehension and vocabulary knowledge are positively correlated (Schmitt et al., 2011)

– 95 to 98% of words in text should be known for optimal comprehension (Laufer & Ravenhorst-Kalovski, 2010)

BACKGROUND

– Text comprehension and vocabulary knowledge are positively correlated

– 95 to 98% of words known for optimal comprehension

→ **Lack of vocabulary knowledge = obstacle**

BACKGROUND

— Text comprehension and vocabulary knowledge are positively correlated

— 95 to 98% of words known for optimal comprehension

— Lack of vocabulary knowledge = obstacle

→ **Develop tools and resources**

BACKGROUND

– Text comprehension and vocabulary knowledge are positively correlated

– 95 to 98% of words known for optimal comprehension

– Lack of vocabulary knowledge = obstacle

– Develop tools and resources

→ **Identification of difficult words = essential**

GHENT
UNIVERSITY

EXAMPLE: READING ASSISTANT

BBC

Home   News   Sport   Business   Innovation   Culture   Arts   Travel   Earth   Audio   Video   Live

The organisers of an arts market in Leeds have amended the application process after visitors complained about the amount of AI-generated art on sale at a recent trading event.

It is the first time the Alternative Market, which has been running since 2017 and receives hundreds of applications from potential vendors, has faced complaints about AI, say organisers.

After more than 100 comments appeared on Reddit after the event on Saturday, organisers at the Leeds Festival of Gothica have promised to engage with the community about the issue of AI-generated art.

GHENT UNIVERSITY

Source: BBC

EXAMPLE: READING ASSISTANT

BBC

Home   News   Sport   Business   Innovation   Culture   Arts   Travel   Earth   Audio   Video   Live

The organisers of an arts market in Leeds have amended the application process after visitors complained about the amount of AI-generated art on sale at a recent trading event.

It is the first time the Alternative Market, which has been running since 2017 and receives hundreds of applications from potential vendors, has faced complaints about AI, say organisers.

After more than 100 comments appeared on Reddit after the event on Saturday, organisers at the Leeds Festival of Gothica have promised to engage with the community about the issue of AI-generated art.

GHENT UNIVERSITY

Source: BBC
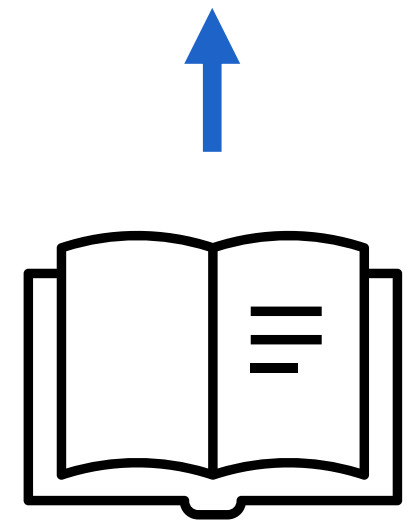
EXAMPLE: READING ASSISTANT

- To amend = […]
- Complaint = […]
- To engage with = […]

The organisers of an arts market in Leeds have amended the application process after visitors complained about the amount of AI-generated art on sale at a recent trading event.

It is the first time the Alternative Market, which has been running since 2017 and receives hundreds of applications from potential vendors, has faced complaints about AI, say organisers.

After more than 100 comments appeared on Reddit after the event on Saturday, organisers at the Leeds Festival of Gothica have promised to engage with the community about the issue of AI-generated art.

GHENT UNIVERSITY

Source: BBC

EXAMPLE: VOCABULARY LISTS

	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   22

	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	79	0	0	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	108.886	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 23

EXISTING TOOLS

EXISTING TOOLS: MultilingProfiler

UNIVERSITY of York

| Home | **MultilingProfiler** | Word Families | FAQ | About | Dissemination and Use | Contact |

MultilingProfiler

Select the *list type* and the related options (if any) you want to use to profile your text.

List type	Language	Level	Remove Inflected Forms ❓	Add Derived Forms ❓
Frequency list ⌄	German ⌄	Top 1000 words ⌄	All forms selected ⌄	No forms selected ⌄

For accurate results, split compounds in your texts by adding a space between the words of which they consist (e.g., *Sommerferien → Sommer Ferien*).

Profile window ❓

James Bond kann viel. Das zeigt er ab dem 1. November wieder einmal in einem neuen 007-Kinofilm, Skyfall. Als Kosmopolit spricht der Agent in den Filmen manchmal auch ein bisschen Deutsch - dafür gibt es einen guten Grund: Bond ist am 11. November 1920 in Wattenscheid (Nordrhein-Westfalen) geboren. So steht es wenigstens in einem Buch über die 007-Figur, das der Engländer John Pearson geschrieben hat. Pearson ist ein Freund des britischen Autors Ian Fleming, der die Idee für die Agenten-Ikone hatte. In Flemings Bond-Büchern steht auch, dass 007 der Sohn eines Schotten und einer Schweizerin ist.

EXISTING TOOLS: LATILL

007 Spricht Deutsch

ⓘ Text info   **🔍 Text level**   **📂 Collections**

👆 Select a CEFR level to highlight its corresponding words in the text.

A1   A2   B1   B2   C1   C2

James Bond kann viel. Das zeigt er ab dem 1. November wieder einmal in einem neuen 007-Kinofilm, Skyfall. Als Kosmopolit spricht der Agent in den Filmen manchmal auch ein bisschen Deutsch - dafür gibt es einen guten Grund: Bond ist am 11. November 1920 in Wattenscheid (Nordrhein-Westfalen) geboren. So steht es wenigstens in einem Buch über die 007-Figur, das der Engländer John Pearson geschrieben hat. Pearson ist ein Freund des britischen Autors Ian Fleming, der die Idee für die Agenten-Ikone hatte. In Flemings Bond-Büchern steht auch, dass 007 der Sohn eines Schotten und einer Schweizerin ist.

CORE CONCEPTS

Individualised prediction of perceived vocabulary difficulty: from dataset to classifier

1. SLA

2. Personalisation

GHENT
UNIVERSITY

READING ASSISTANT FOR LEARNER A

BBC

Home   News   Sport   Business   Innovation   Culture   Arts   Travel   Earth   Audio   Video   Live

The organisers of an arts market in Leeds have amended the application process after visitors complained about the amount of AI-generated art on sale at a recent trading event.

It is the first time the Alternative Market, which has been running since 2017 and receives hundreds of applications from potential vendors, has faced complaints about AI, say organisers.

After more than 100 comments appeared on Reddit after the event on Saturday, organisers at the Leeds Festival of Gothica have promised to engage with the community about the issue of AI-generated art.

GHENT
UNIVERSITY

Source: BBC

READING ASSISTANT FOR LEARNER B

BBC

Home   News   Sport   Business   Innovation   Culture   Arts   Travel   Earth   Audio   Video   Live

The organisers of an arts market in Leeds have amended the application process after visitors complained about the amount of AI-generated art on sale at a recent trading event.

It is the first time the Alternative Market, which has been running since 2017 and receives hundreds of applications from potential vendors, has faced complaints about AI, say organisers.

After more than 100 comments appeared on Reddit after the event on Saturday, organisers at the Leeds Festival of Gothica have promised to engage with the community about the issue of AI-generated art.

GHENT
UNIVERSITY

Source: BBC

2. PERSONALISATION

GHENT
UNIVERSITY

WHY PERSONALISE PREDICTIONS?

– Influence of L1 (e.g., cognates)

– Cultural factors (e.g., dubbing versus subtitling)

– Personal interests

GHENT
UNIVERSITY

WHY PERSONALISE PREDICTIONS?

– One Size Does Not Fit All: The Case for Personalised Word Complexity Models (Gooding & Tragut, 2022)

– *"The difficulty of a word is a highly idiosyncratic notion that depends on a reader's first language"*

– *"Models are best when predicting word complexity for individual readers"*

CORE CONCEPTS

Individualised prediction of perceived vocabulary difficulty: from dataset to classifier

1 - 2   SLA | Personalisation

3.   Lexical complexity prediction

3. 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
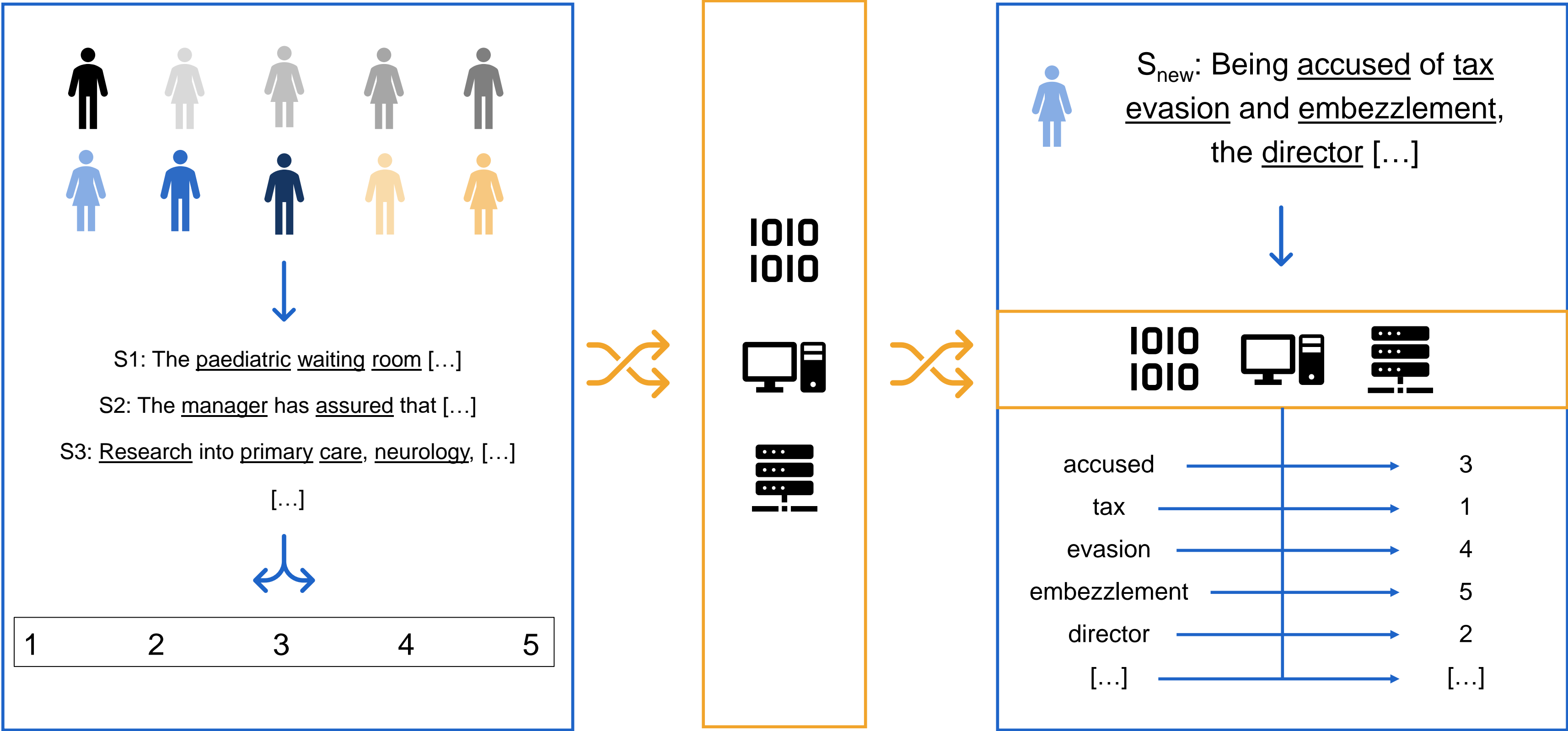
CORE CONCEPTS

Individualised prediction of perceived vocabulary difficulty: from dataset to classifier

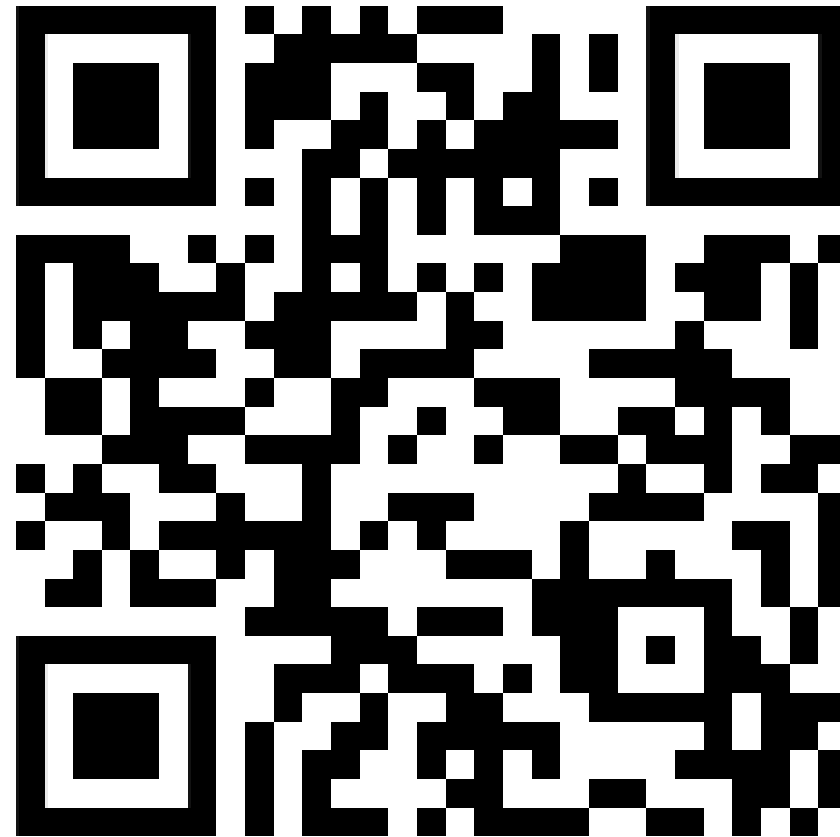1 - 3   SLA | Personalisation | LCP

4.   Dataset

GHENT
UNIVERSITY

4. DATASET

GHENT
UNIVERSITY

LexComSpaL2

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

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 & 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

GHENT UNIVERSITY

LCP: ADAPTED SCALE

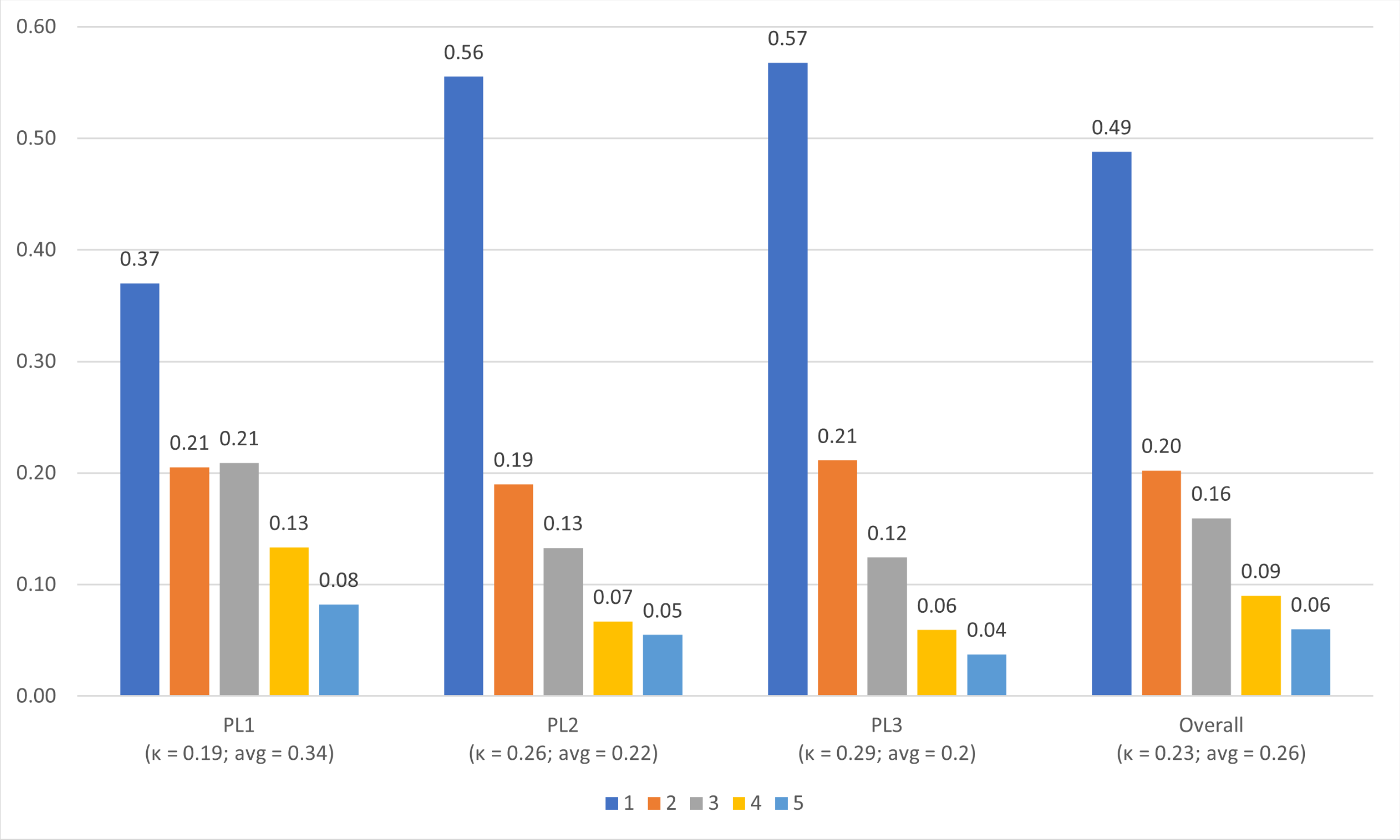| LCP label	Description
1	I know this word and its meaning, and I also use it actively in speaking/writing.
2	I know this word and its meaning, but I might not be able to use it on the top of my head in an oral/written conversation. When I have some time to think, however, I do think I would use it naturally.
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	0.24
				1,001 - 2,000	0.14
				2,001 - 3,000	0.09
				3,001 - 4,000	0.07
				4,001 - 5,000	0.05
				>5,000	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
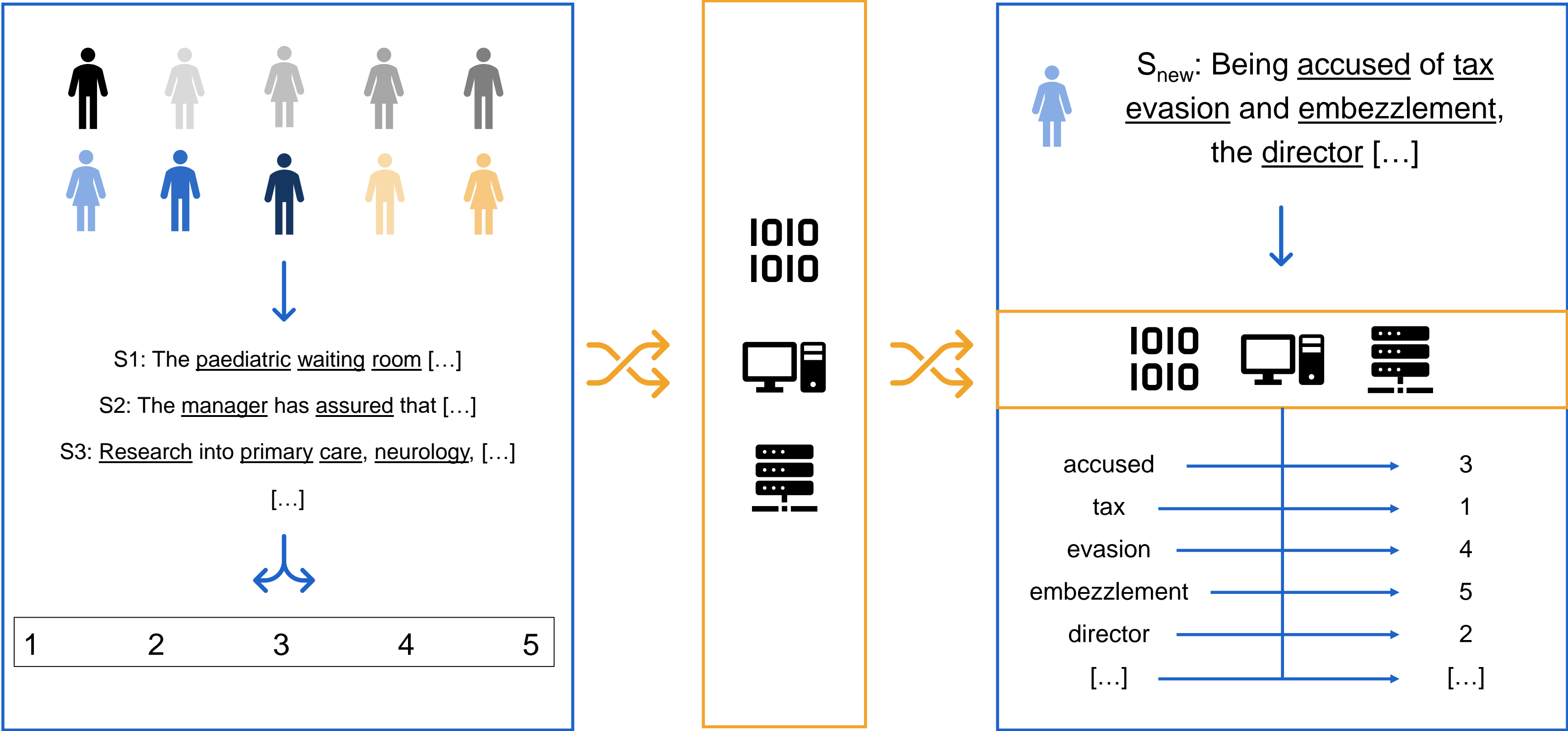
CORE CONCEPTS

Individualised prediction of perceived vocabulary difficulty: from dataset to classifier

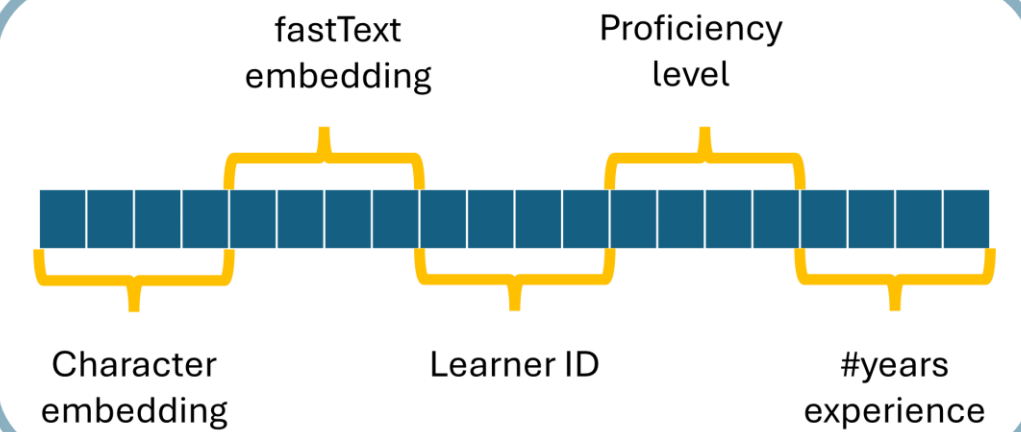1 - 4   SLA | Personalisation | LCP | Dataset

5.   Classifier

GHENT
UNIVERSITY

5. CLASSIFIER

DETAILS

– Architecture based on previous research (Tack, 2018)

– BiLSTM neural classifier

– Goal: baseline model for future research

GHENT
UNIVERSITY

This is a sentence with a **difficult** word.

word → vector

fastText embedding   Proficiency level

Character embedding   Learner ID   #years experience

vector → model

BiLSTM classifier

model → prediction

This is a sentence with a **difficult** word.

softmax activation

1	0.05
2	0.11
3	**0.51**
4	0.26
5	0.07

GHENT UNIVERSITY

52

PERFORMANCE

Fold	#observations (target words | annotations)			D'	MCC	F1
	Training	Validation	Test			
1	1,796 | 46,696	216 | 5,616	228 | 5,928	0.18	0.32	0.53
2	1,797 | 46,722	227 | 5,902	216 | 5,616	0.16	0.3	0.52
3	1,787 | 46,462	226 | 5,876	227 | 5,902	0.19	0.36	0.58
4	1,775 | 46,150	226 | 5,876	226 | 5,876	0.18	0.33	0.53
5	1,799 | 46,774	202 | 5,252	239 | 6,214	0.18	0.32	0.51
6	1,818 | 47,268	220 | 5,720	202 | 5,252	0.19	0.34	0.51
7	1,803 | 46,878	217 | 5,642	220 | 5,720	0.17	0.34	0.52
8	1,804 | 46,904	219 | 5,694	217 | 5,642	0.18	0.32	0.54
9	1,775 | 46,150	246 | 6,396	219 | 5,694	0.17	0.31	0.52
10	1,766 | 45,916	228 | 5,928	246 | 6,396	0.17	0.31	0.5
Mean ± SD				0.18 ± 0.01	0.32 ± 0.02	0.53 ± 0.02
Median				0.18	0.32	0.53

GHENT
UNIVERSITY

NEXT STEPS

– Adding more features and train other machine learning models (e.g., XGBoost)

– Can LLMs be employed for this purpose?

– ?

GHENT
UNIVERSITY

LIMITATIONS

GHENT
UNIVERSITY

LIMITATIONS

— Single words

— No information on word senses yet

— Participants = L1 Dutch

— Too many sentences to apply in real life → identify most important ones with item analysis?

— No classifier yet that is ready to be implemented in real-life applications

GHENT
UNIVERSITY

LINKS

– Dataset repository and slides

REFERENCES

Gooding, S., & Tragut, M. (2022). One Size Does Not Fit All: The Case for Personalised Word Complexity Models. *Findings of the Association for Computational Linguistics: NAACL 2022*, 353–365. https://doi.org/10.18653/v1/2022.findings-naacl.27

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.

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.

Schmitt, N. (2019). Understanding vocabulary acquisition, instruction, and assessment: A research agenda. Language Teaching, 52 (02), 261–274. https://doi.org/10.1017/S0261444819000053

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. https://doi.org/10.1111/j.1540-4781.2011.01146.x

Tack, A. (2021). *Mark My Words! On the Automated Prediction of Lexical Difficulty for Foreign Language Readers* [PhD thesis]. UCLouvain & KU Leuven.

Webb, S., & Nation, I. S. P. (2017). *How vocabulary is learned*. Oxford University Press.