FACULTY OF ARTS
AND PHILOSOPHY

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

YOU SHALL KNOW A WORD'S DIFFICULTY BY THE FAMILY IT KEEPS

Jasper Degraeuwe – 31 July 2025 – ACL@Vienna

Workshop on Innovative Use of NLP for Building Educational Applications (BEA)

GHENT
UNIVERSITY

ACL 2025
VIENNA
JULY 27 - AUGUST 1

ABOUT ME

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

  ICALL ≈ CALL + AI and NLP

- Postdoctoral researcher on **educational technologies and AI for language learning**

SNEAK PEEK

GHENT
UNIVERSITY

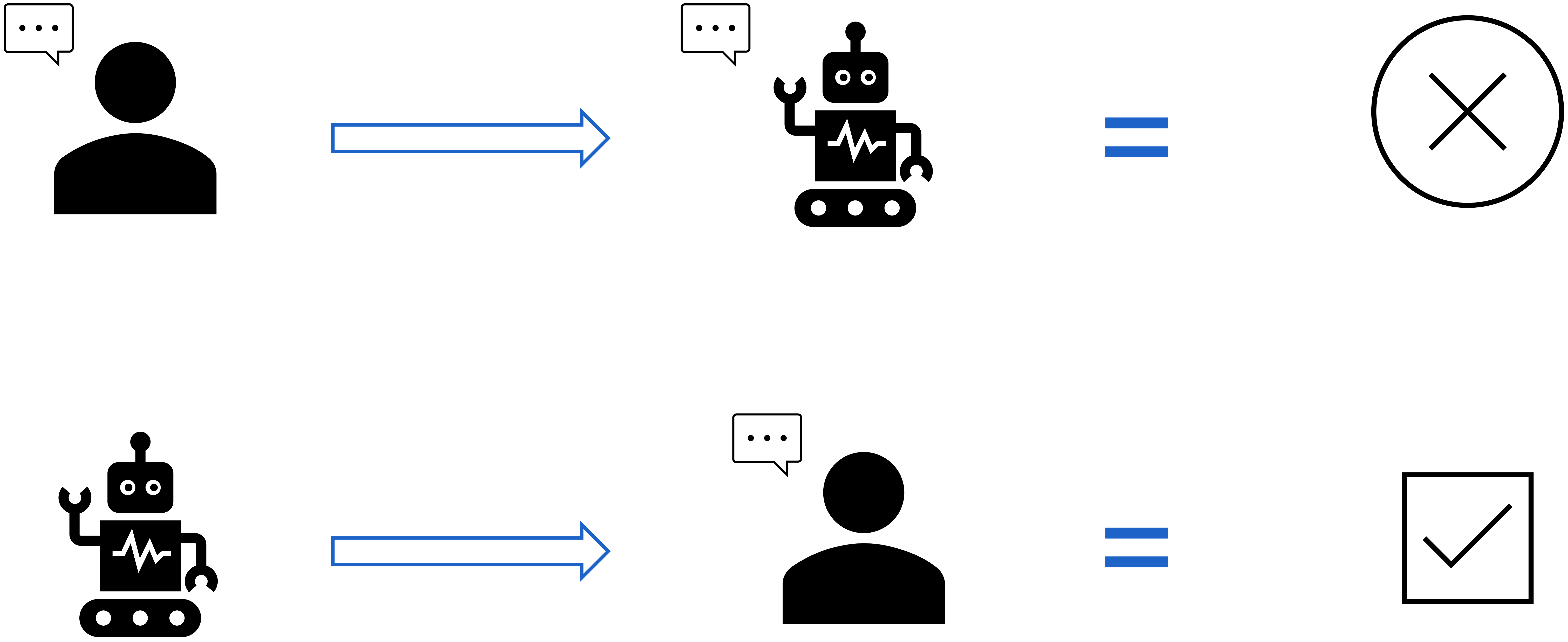
SNEAK PEEK

- Word difficulty as perceived by individual language learners
- Automated prediction on 1 – 5 scale
- Added value of word family knowledge as feature

TARGET SETTING

Second language acquisition (SLA)

?



RESEARCH QUESTION

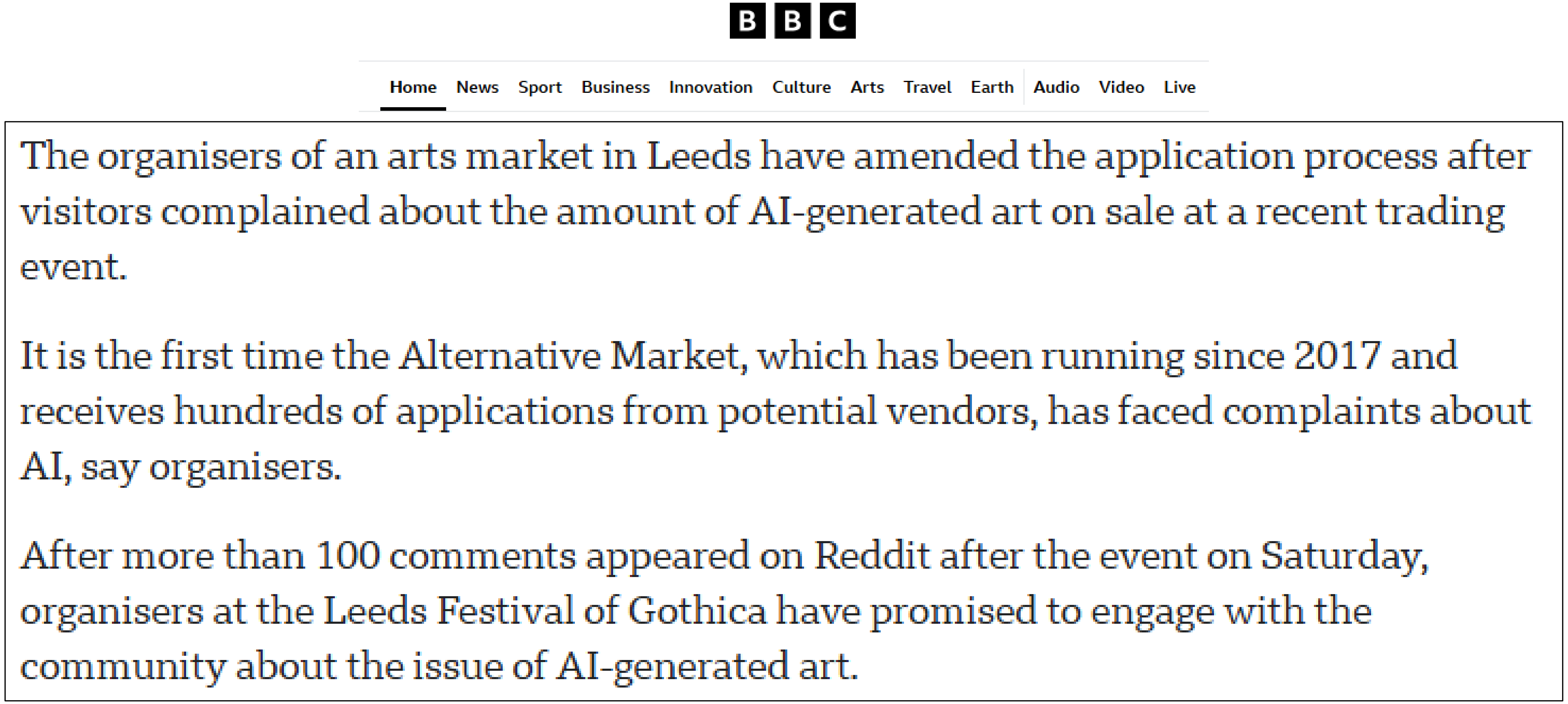
GHENT
UNIVERSITY

RESEARCH QUESTION

Can machine learning systems accurately predict how easy/difficult individual words are for individual language learners?

REAL-LIFE UTILITY

(1) READING ASSISTANT



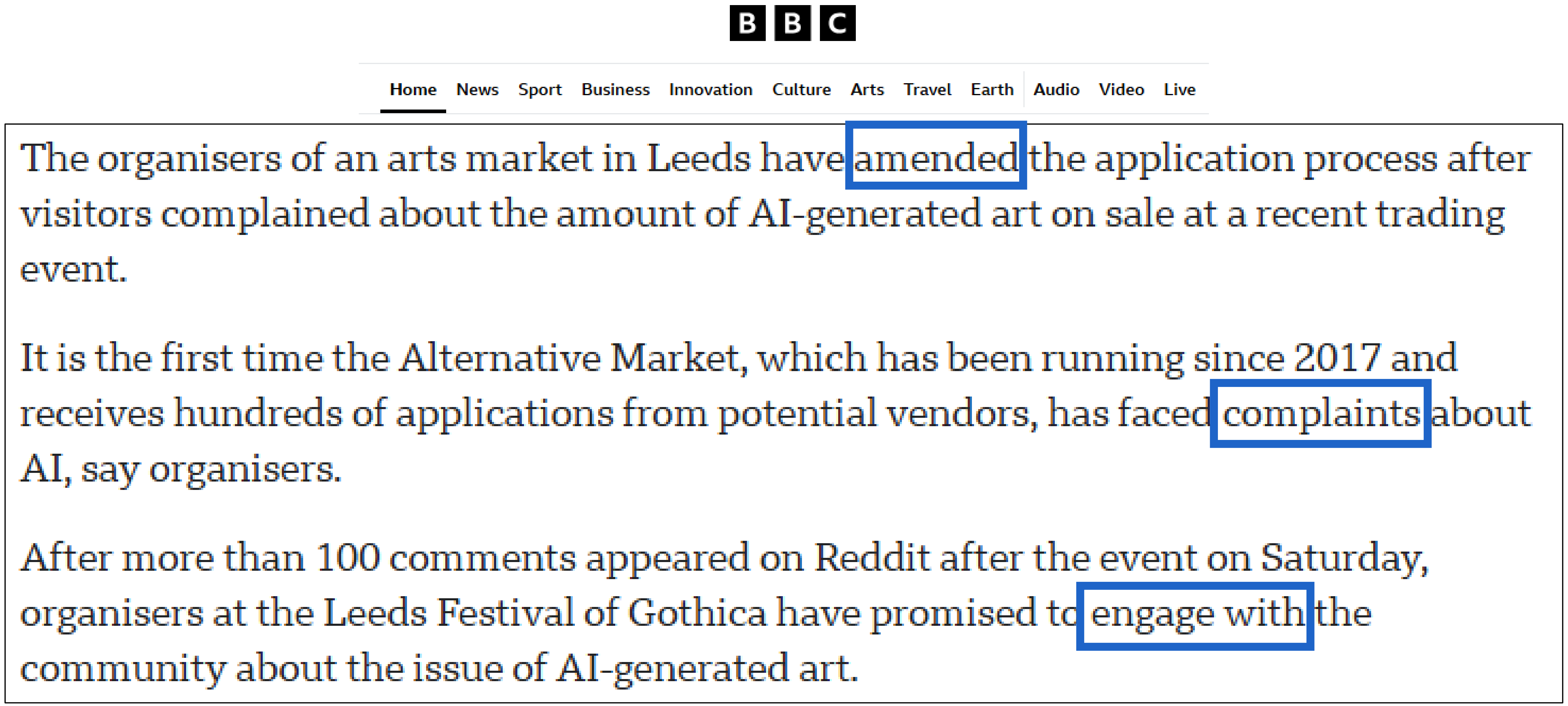
The organisers of an arts market in Leeds have amended the application process after visitors complained about the amount of AI-generated art on sale at a recent trading event.

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.

Source: BBC

(1) READING ASSISTANT



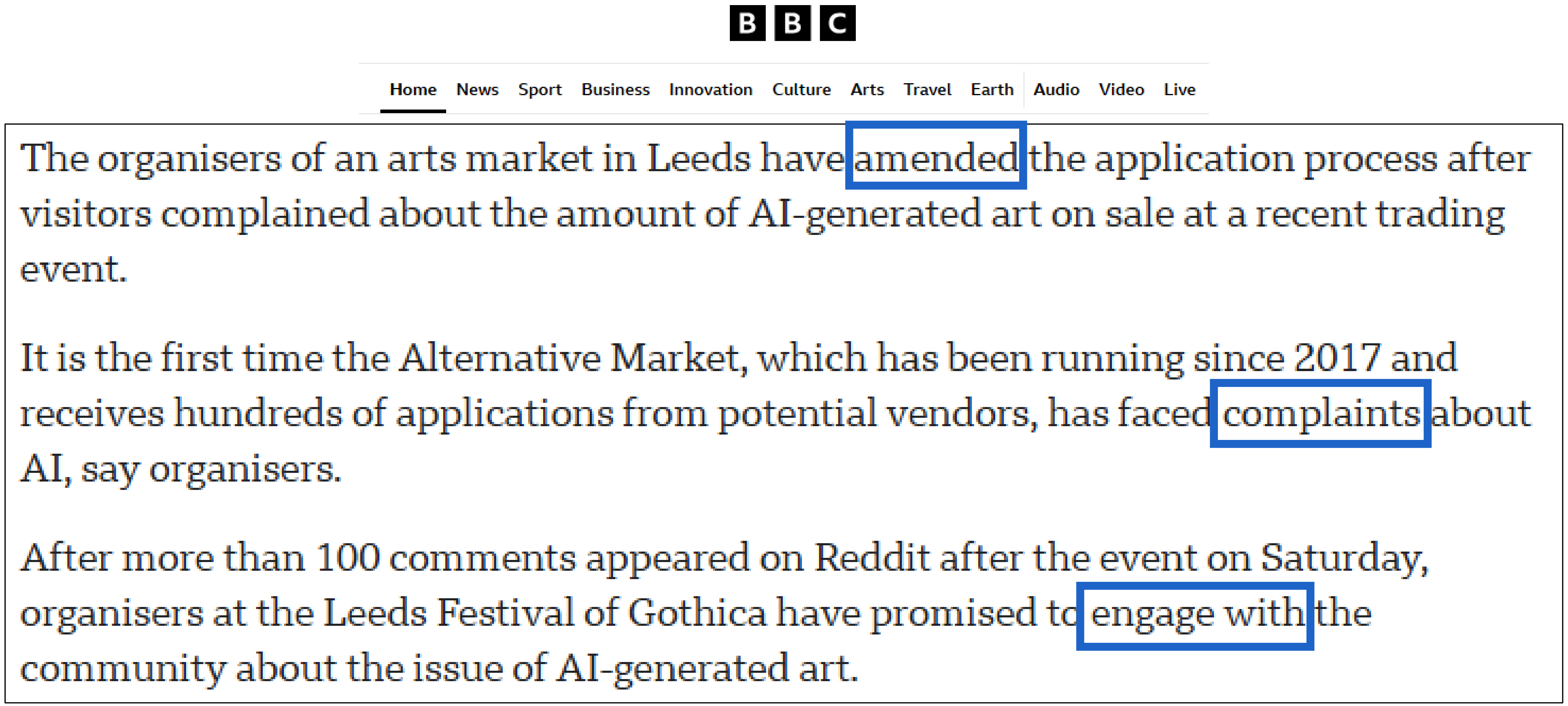
The organisers of an arts market in Leeds have amended the application process after visitors complained about the amount of AI-generated art on sale at a recent trading event.

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.

Source: BBC

(1) READING ASSISTANT



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.

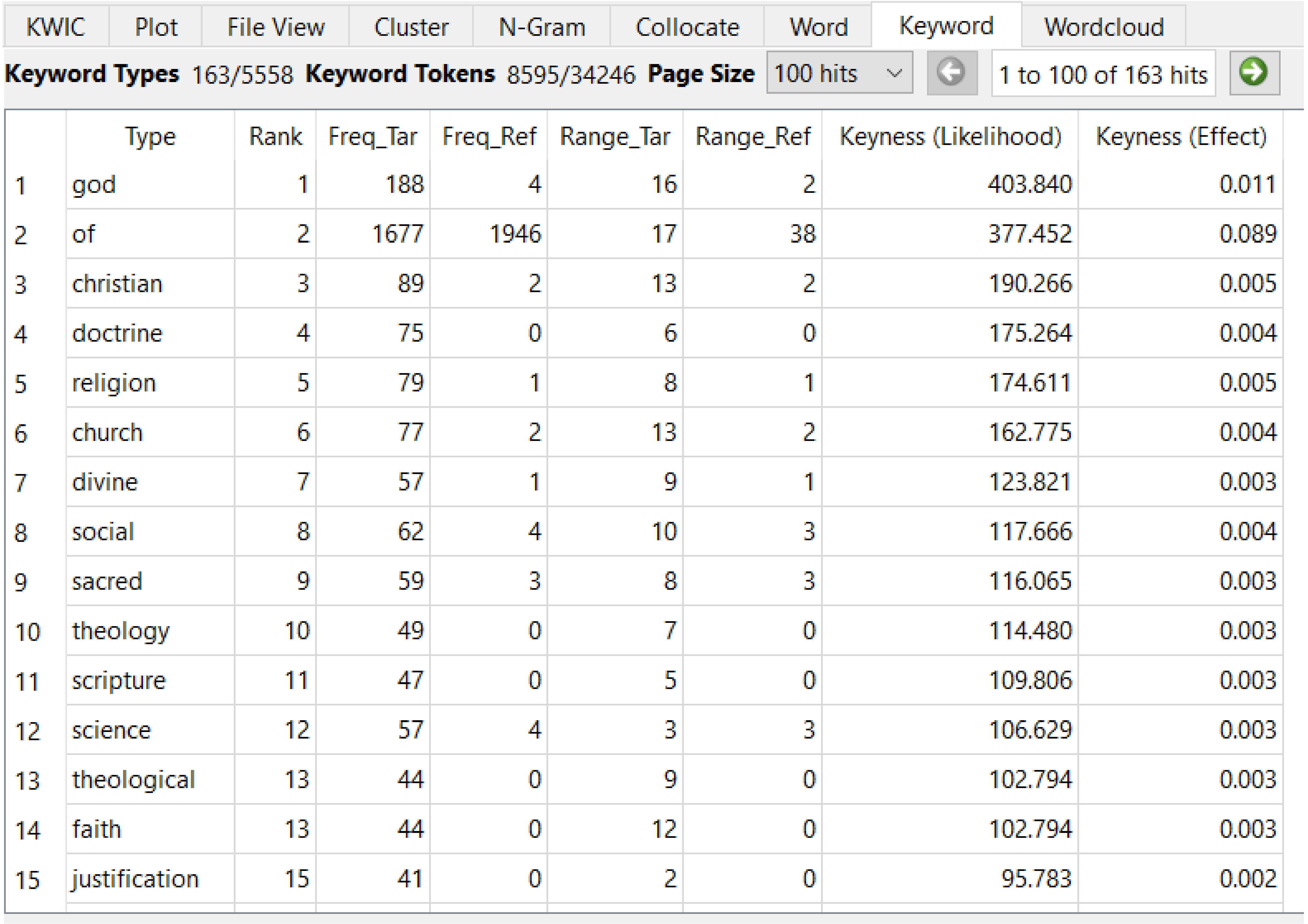
It is the first time the Alternative Market, which has been running since 2017 and receives hundreds of applications from potential vendors, has faced complaints about AI, say organisers.

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.

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

Source: BBC

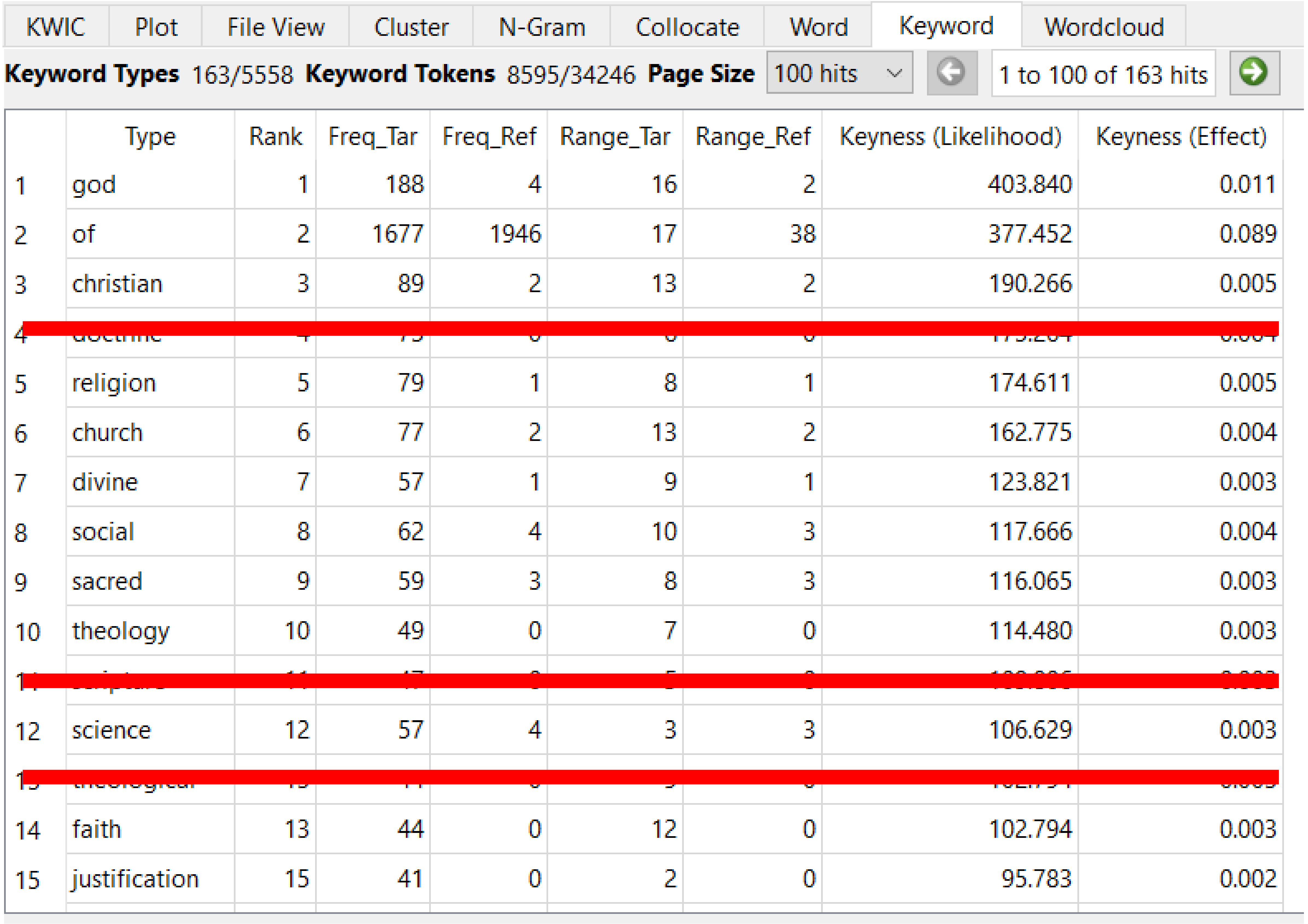
(2) VOCABULARY LISTS



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



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~~	[illegible]	[illegible]
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~~	[illegible]	~~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

GHENT UNIVERSITY

Source: AntConc

METHODOLOGY

GHENT UNIVERSITY

DATA

– LexComSpaL2 corpus (Degraeuwe & Goethals, 2024)

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

LABELLING: LEXICAL COMPLEXITY PREDICTION

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

GHENT UNIVERSITY

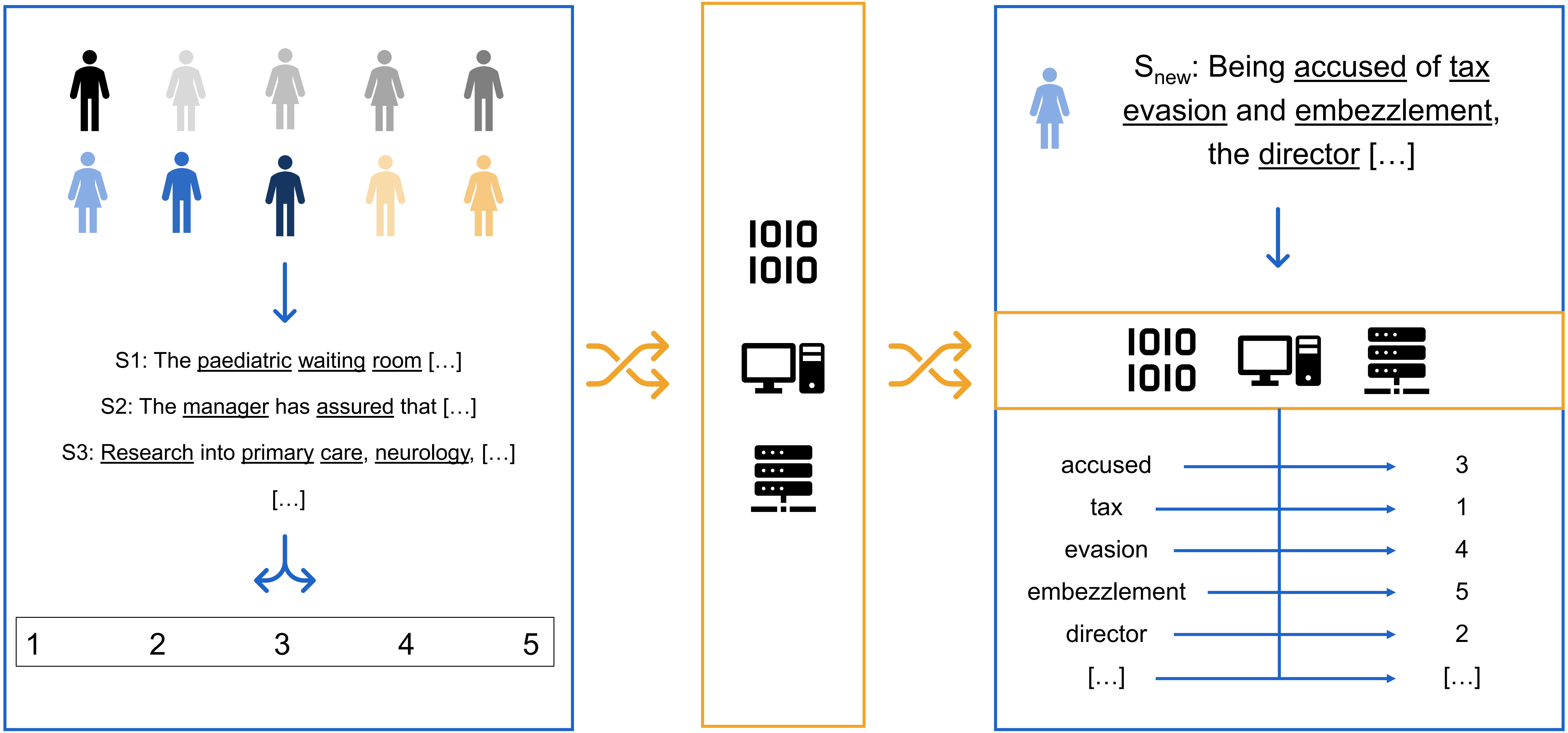
Shardlow et al. (2020)

LABELLING: ADAPTED LCP 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.

BASE CLASSIFIER

CONCEPTUAL OVERVIEW



S1: The paediatric waiting room […]

S2: The manager has assured that […]

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

[…]

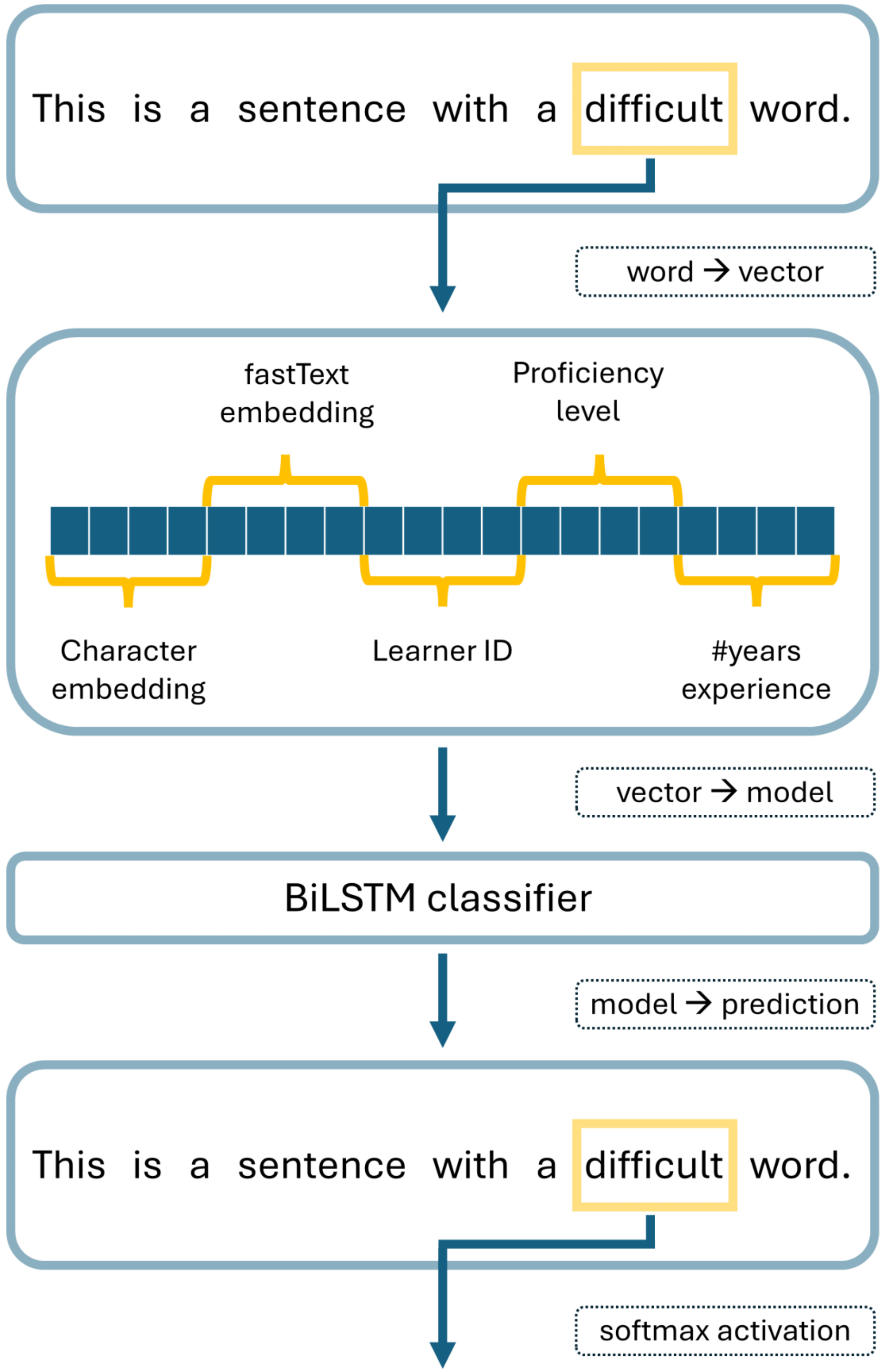
1	2	3	4	5

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

accused	3
tax	1
evasion	4
embezzlement	5
director	2
[…]	[…]

DETAILS

- Architecture based on previous research (Tack, 2021)
- BiLSTM neural classifier
- Input features
  - Character embedding
  - fastText static word embedding
  - Learner ID
  - Proficiency level learner
  - Number of years of experience learner



This is a sentence with a difficult word.

word → vector

fastText embedding

Proficiency level

Character embedding

Learner ID

#years experience

vector → model

BiLSTM classifier

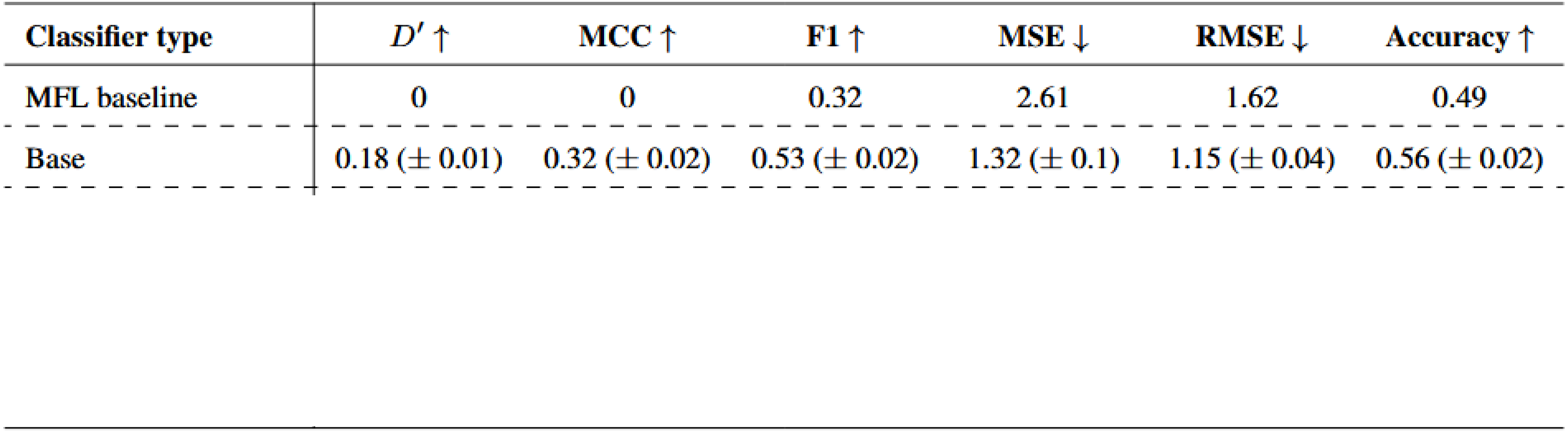
model → prediction

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

PERFORMANCE



Classifier type	D' ↑	MCC ↑	F1 ↑	MSE ↓	RMSE ↓	Accuracy ↑
MFL baseline	0	0	0.32	2.61	1.62	0.49
Base	0.18 (± 0.01)	0.32 (± 0.02)	0.53 (± 0.02)	1.32 (± 0.1)	1.15 (± 0.04)	0.56 (± 0.02)

WORD FAMILY-ENRICHED CLASSIFIER

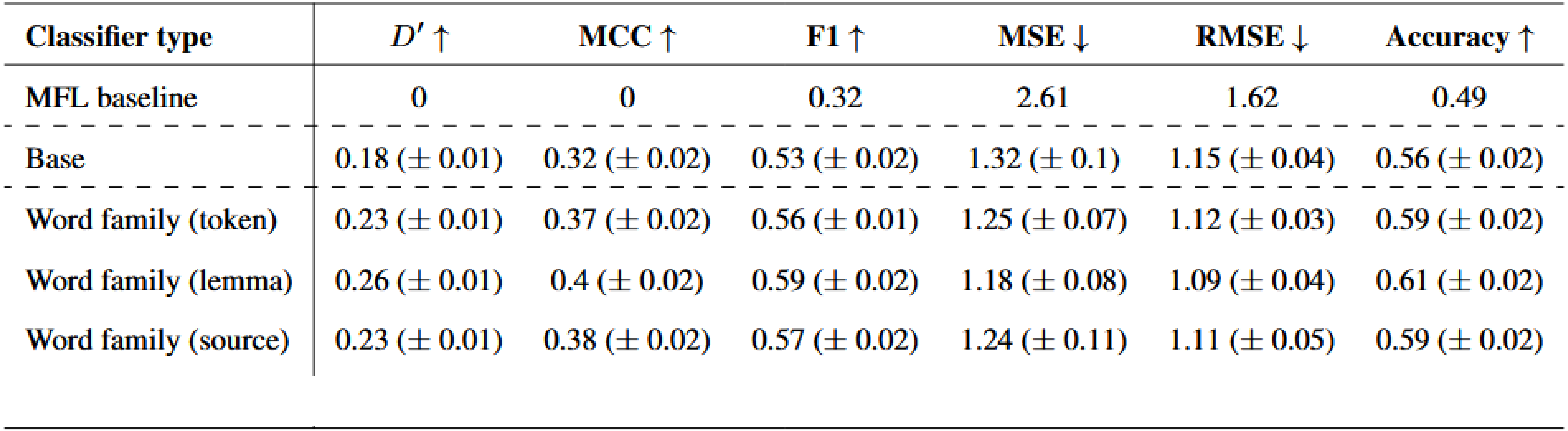
WORD FAMILY LEVELS

Level	Description	Example	
Token	Family = all occurrences of the exact word form in the dataset	disappears	disappears
Lemma	Family = “base form” of word + all its inflected forms		**disappear** disappears disappeared disappearing
Source	Family = “parent” of word (i.e. the lemma the word is derived from) + all inflected forms of this parent		**appear** appears appeared appearing

EXAMPLE (SOURCE LEVEL)

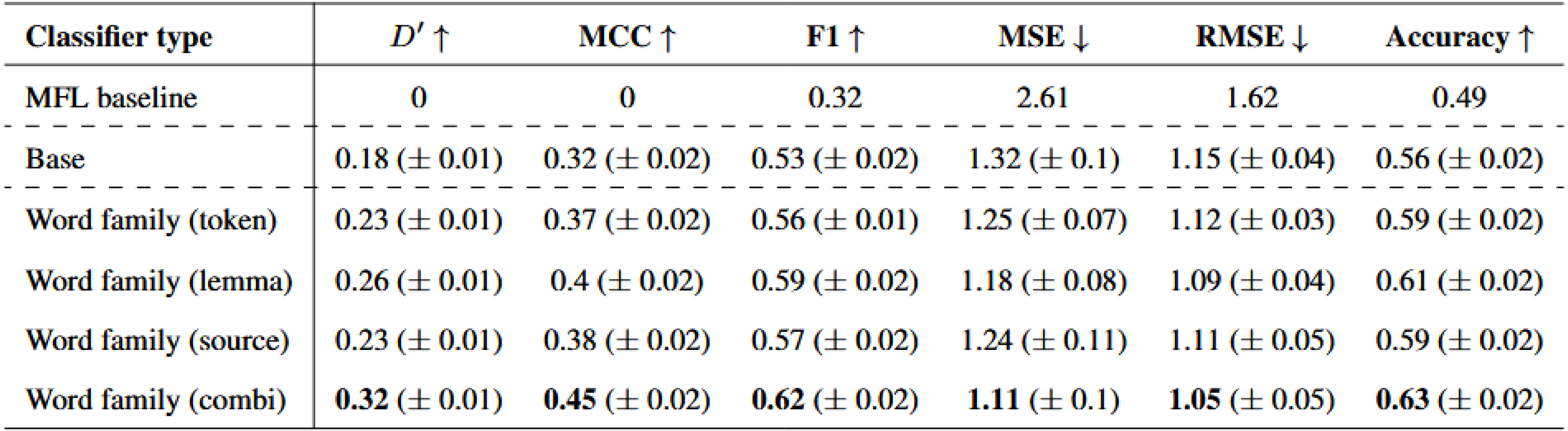
- Participant ID: 11
- Annotations
  - *appears*: 1 (“very easy”)
  - *appearing*: 2 (“easy”)
- Word to be predicted: *disappeared*

PERFORMANCE



Classifier type	D' ↑	MCC ↑	F1 ↑	MSE ↓	RMSE ↓	Accuracy ↑
MFL baseline	0	0	0.32	2.61	1.62	0.49
Base	0.18 (± 0.01)	0.32 (± 0.02)	0.53 (± 0.02)	1.32 (± 0.1)	1.15 (± 0.04)	0.56 (± 0.02)
Word family (token)	0.23 (± 0.01)	0.37 (± 0.02)	0.56 (± 0.01)	1.25 (± 0.07)	1.12 (± 0.03)	0.59 (± 0.02)
Word family (lemma)	0.26 (± 0.01)	0.4 (± 0.02)	0.59 (± 0.02)	1.18 (± 0.08)	1.09 (± 0.04)	0.61 (± 0.02)
Word family (source)	0.23 (± 0.01)	0.38 (± 0.02)	0.57 (± 0.02)	1.24 (± 0.11)	1.11 (± 0.05)	0.59 (± 0.02)

PERFORMANCE



Classifier type	D' ↑	MCC ↑	F1 ↑	MSE ↓	RMSE ↓	Accuracy ↑
MFL baseline	0	0	0.32	2.61	1.62	0.49
Base	0.18 (± 0.01)	0.32 (± 0.02)	0.53 (± 0.02)	1.32 (± 0.1)	1.15 (± 0.04)	0.56 (± 0.02)
Word family (token)	0.23 (± 0.01)	0.37 (± 0.02)	0.56 (± 0.01)	1.25 (± 0.07)	1.12 (± 0.03)	0.59 (± 0.02)
Word family (lemma)	0.26 (± 0.01)	0.4 (± 0.02)	0.59 (± 0.02)	1.18 (± 0.08)	1.09 (± 0.04)	0.61 (± 0.02)
Word family (source)	0.23 (± 0.01)	0.38 (± 0.02)	0.57 (± 0.02)	1.24 (± 0.11)	1.11 (± 0.05)	0.59 (± 0.02)
Word family (combi)	**0.32** (± 0.01)	**0.45** (± 0.02)	**0.62** (± 0.02)	**1.11** (± 0.1)	**1.05** (± 0.05)	**0.63** (± 0.02)

CONCLUSION

CONTRIBUTIONS

- Adapted LCP scale → **predictions are tailored to SLA** target setting
- Research gap filled: **multi-label** & **personalised** word difficulty prediction for SLA purposes (←→ binary & personalised prediction; Tack, 2021)
- Analysis of adding **word family knowledge** as input feature

LIMITATIONS (DATASET)

- Only for L1 speakers of Dutch → same results for other L1s?
- Words in dataset are not disambiguated for different meanings (e.g., *bat* as a baseball instrument and *bat* as a night animal)

LIMITATIONS (MODEL)

- Real-life applicability?
  - Is RMSE of 1.05 good enough for pedagogical setting?
  - Using model for new learners = learners first have to annotate the 200 LexComSpaL2 sentences

REFERENCES

Degraeuwe, J., & Goethals, P. (2024). LexComSpaL2: A Lexical Complexity Corpus for Spanish as a Foreign Language. In N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, & N. Xue (Eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* (pp. 10432–10447). ELRA and ICCL. https://aclanthology.org/2024.lrec-main.912

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

Shardlow, M., Cooper, M., & Zampieri, M. (2020). CompLex—A New Corpus for Lexical Complexity Prediction from Likert Scale Data. *Proceedings of the 1st Workshop on Tools and Resources to Empower People with REAding DIfficulties (READI)*, 57–62. https://aclanthology.org/2020.readi-1.9

FACULTY OF ARTS
AND PHILOSOPHY

Jasper Degraeuwe
Postdoctoral researcher

DEPARTMENT OF TRANSLATION,
INTERPRETING AND COMMUNICATION

Jasper.Degraeuwe@UGent.be

www.ugent.be

Universiteit Gent
@ugent
@ugent
Ghent University

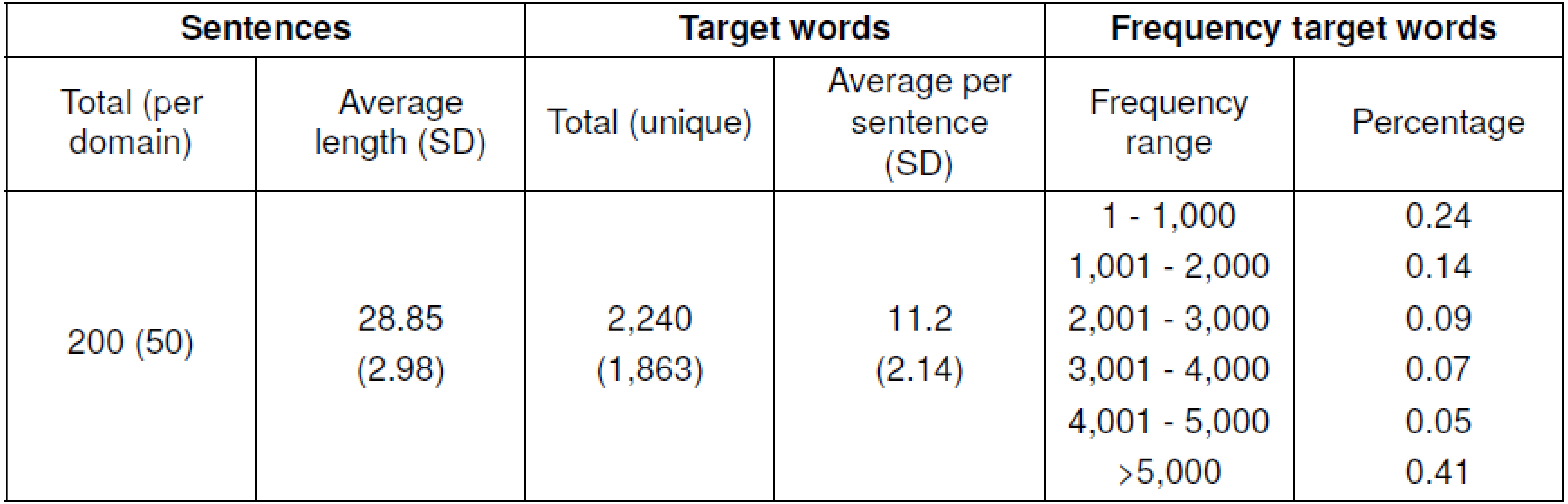
Dataset →

Paper →

GHENT
UNIVERSITY

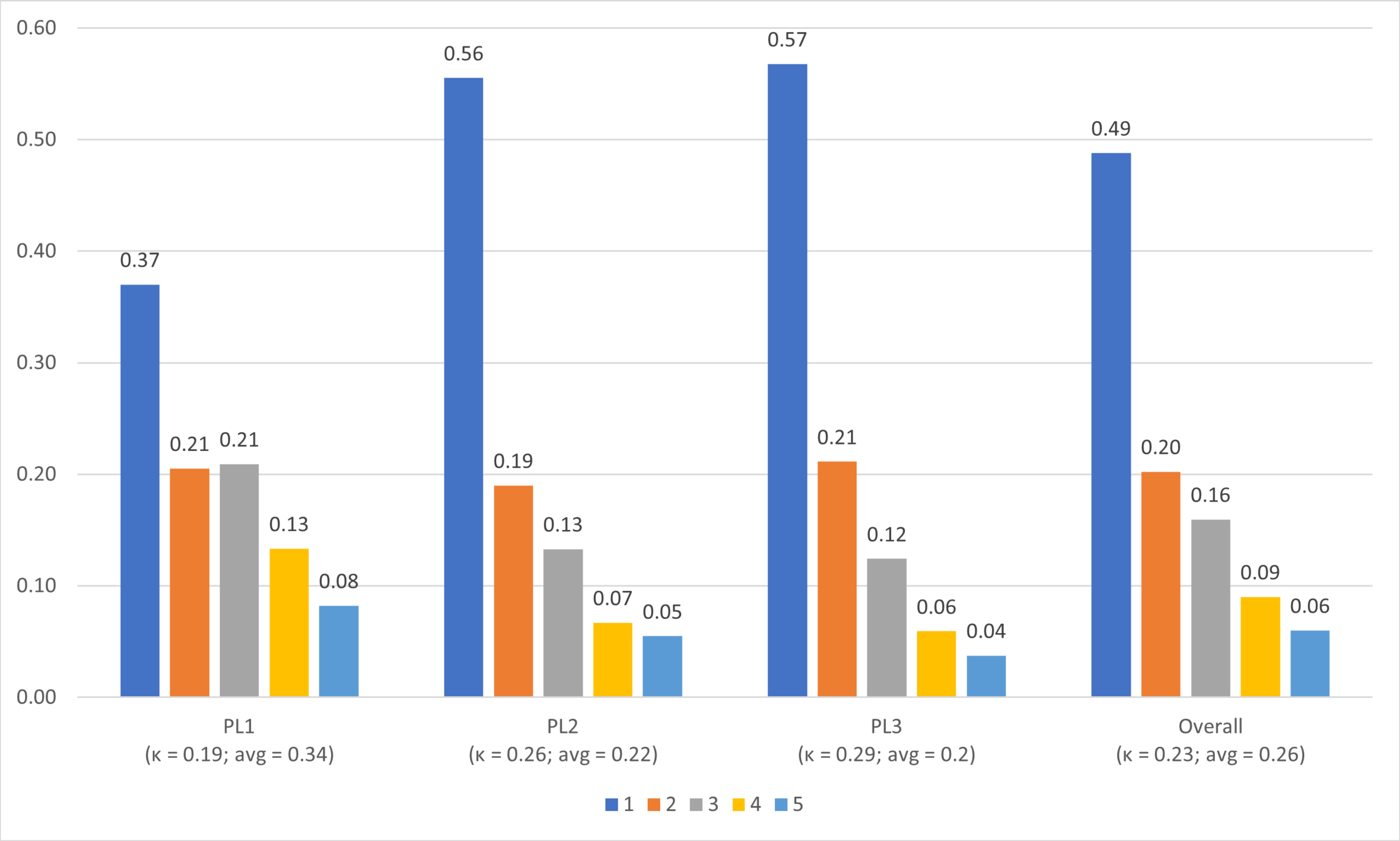
SUPPLEMENTARY SLIDES

LexComSpaL2: STATISTICS



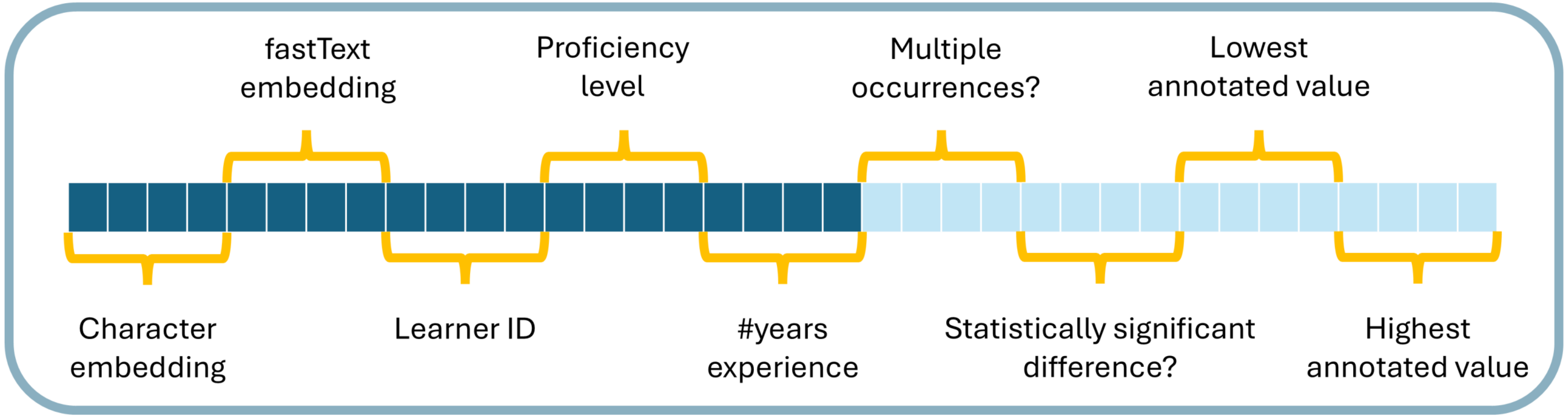
Sentences		Target words		Frequency target words	
Total (per domain)	Average length (SD)	Total (unique)	Average per sentence (SD)	Frequency range	Percentage
200 (50)	28.85 (2.98)	2,240 (1,863)	11.2 (2.14)	1 - 1,000	0.24
				1,001 - 2,000	0.14
				2,001 - 3,000	0.09
				3,001 - 4,000	0.07
				4,001 - 5,000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

LexComSpaL2: STATISTICS



WORD FAMILY KNOWLEDGE: EXAMPLE

	Multiple occurrences?	Statistically significant difference?	Lowest annotated value by participant	Highest annotated value by participant
Token	No	N/A	N/A	N/A
Lemma	Yes	No	1	2
Source	Yes	Yes	1	4


fastText embedding
Proficiency level
Multiple occurrences?
Lowest annotated value
Character embedding
Learner ID
#years experience
Statistically significant difference?
Highest annotated value