FACULTY OF ARTS
AND PHILOSOPHY

CENTRE FOR DIVERSITY AND LEARNING, MULTIPLES, LANGUAGE & TRANSLATION TECHNOLOGY TEAM (LT3)

Amaury Van Parys, Vanessa De Wilde, Lieve Macken, Maribel Montero Perez

LEXPRO: A PLURILINGUAL LEXICAL PROFILING TOOL FOR RESEARCH AND MATERIALS DEVELOPMENT

Theoretical background

Vocabulary

- Essential predictor of L2 reading comprehension (Jeon & Yamashita, 2022)
- More words known = better comprehension (Schmitt et al., 2011)
- Need for objective method to assess how demanding a text's vocabulary will be for L2 learners

Lexical profiling

- Method for determining **vocabulary demands** of L2 input
- Often used in previous research: TV series (Webb & Rodgers, 2009), novels (Nation, 2006), L2 textbooks (Van Parys et al., 2024), etc.
- Categorising vocabulary across **word frequency** levels: Higher-frequency words have higher odds of being known by learner (Nation, 2013)
- Allows to estimate **vocabulary loads**, i.e., estimates of required vocabulary size for achieving crucial points of vocabulary coverage (Webb, 2020):
  - **95%** coverage: needed for basic comprehension
  - **98%** coverage: needed for detailed comprehension
- Example: according to Webb & Rodgers (2009), the 3,000 most frequent word families in English need to be known for 95% coverage of TV series and thus basic comprehension

Gaps in prior profiling methods

Word counting unit

- Typical counting unit in profiling is the **word family**, which covers a headword (e.g., 'act') with all its inflections (e.g., 'acts', 'acting') and derivations (e.g., 'actor')
- However: increasing criticism (e.g., McLean, 2018; Stoeckel et al., 2024)
- Potentially more appropriate counting unit: the **flemma,** which covers a headword with all inflections (across different parts of speech), but not derivations

Word frequency as proxy for learner knowledge

- Typically used frequency lists in profiling are based on **broad corpora** covering a wide range of topics (e.g., British National Corpus)
- However: these lists do not reflect learner knowledge as closely as once presumed (Pinchbeck et al., 2022)
- Lists derived from **subtitle corpora** (e.g., SubtLex-UK) appear to align more with learner knowledge (Pinchbeck et al., 2022; van Heuven et al., 2014)

Lack of focus on non-English L2s

- Most existing tools (e.g., LexTutor; Cobb, n.d.) mainly target English
- In line with overall focus on English in Second Language Acquisition (Brezina & Pallotti, 2019)

Goals of LexPro

LexPro aims to set itself apart from existing profiling tools by:

- Using the **flemma** as main word counting unit
- Using **subtitle-based frequency lists**
- Targeting **English** in addition to multiple other L2s (currently **French**, **Spanish**, and **Dutch**)

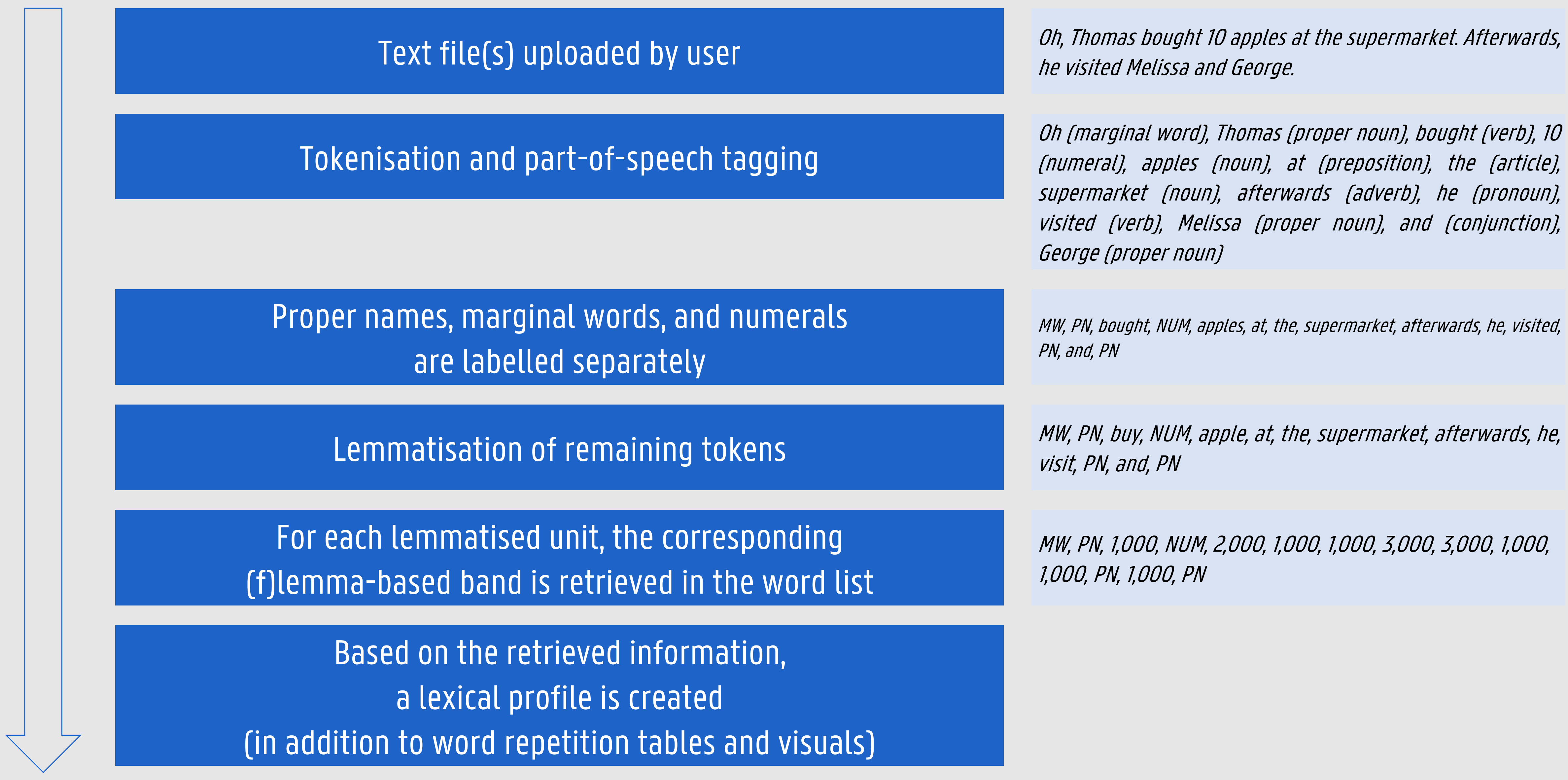
Output

Example corpus: French Netflix series 'Lupin' (Kay & Uzan, 2021-present)

- 3 seasons
- 17 episodes
- 54,502 running tokens
- On average 3,206 tokens per episode

How does LexPro work?

Programmed in **Python** and relying on the **spaCy** NLP library (Honnibal & Montani, 2017)



Text file(s) uploaded by user

Oh, Thomas bought 10 apples at the supermarket. Afterwards, he visited Melissa and George.

Tokenisation and part-of-speech tagging

Oh (marginal word), Thomas (proper noun), bought (verb), 10 (numeral), apples (noun), at (preposition), the (article), supermarket (noun), afterwards (adverb), he (pronoun), visited (verb), Melissa (proper noun), and (conjunction), George (proper noun)

Proper names, marginal words, and numerals are labelled separately

MW, PN, bought, NUM, apples, at, the, supermarket, afterwards, he, visited, PN, and, PN

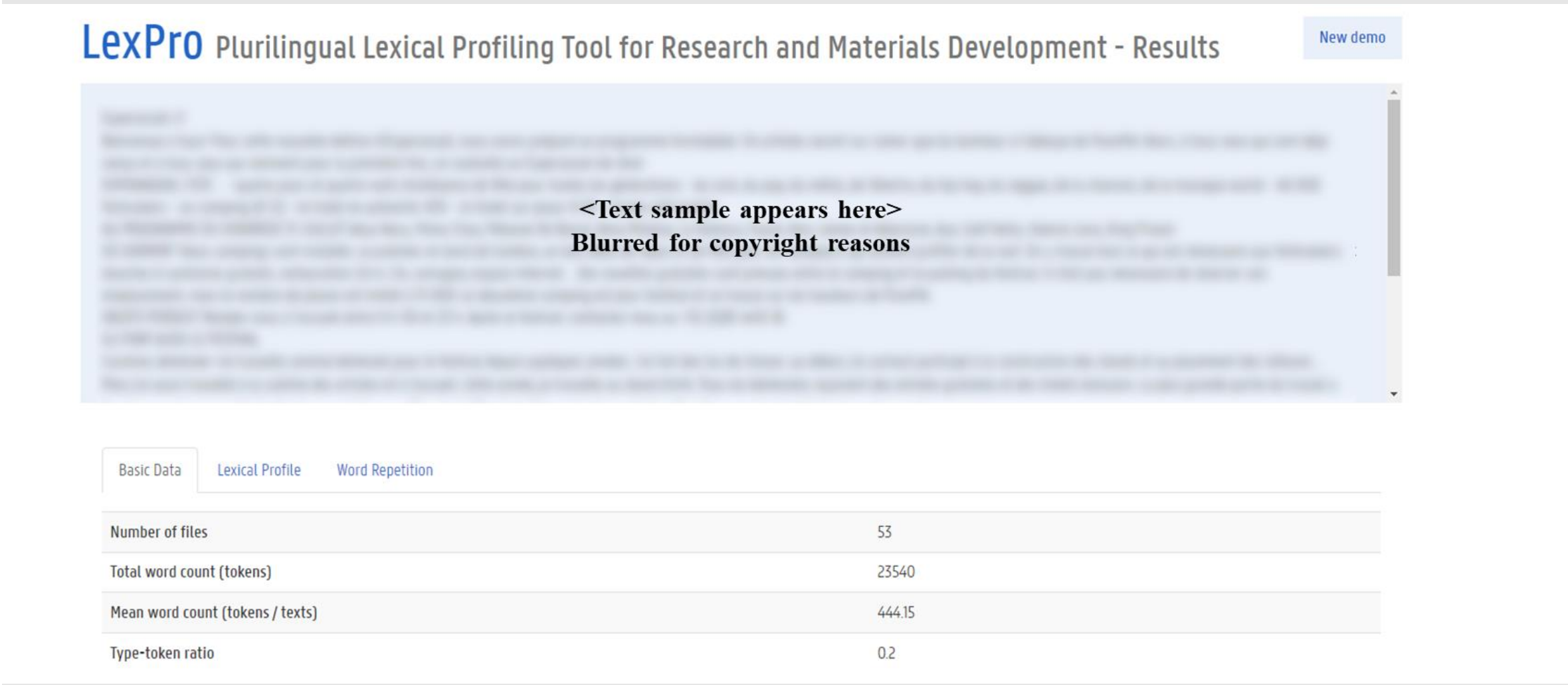
Lemmatisation of remaining tokens

MW, PN, buy, NUM, apple, at, the, supermarket, afterwards, he, visit, PN, and, PN

For each lemmatised unit, the corresponding (f)lemma-based band is retrieved in the word list

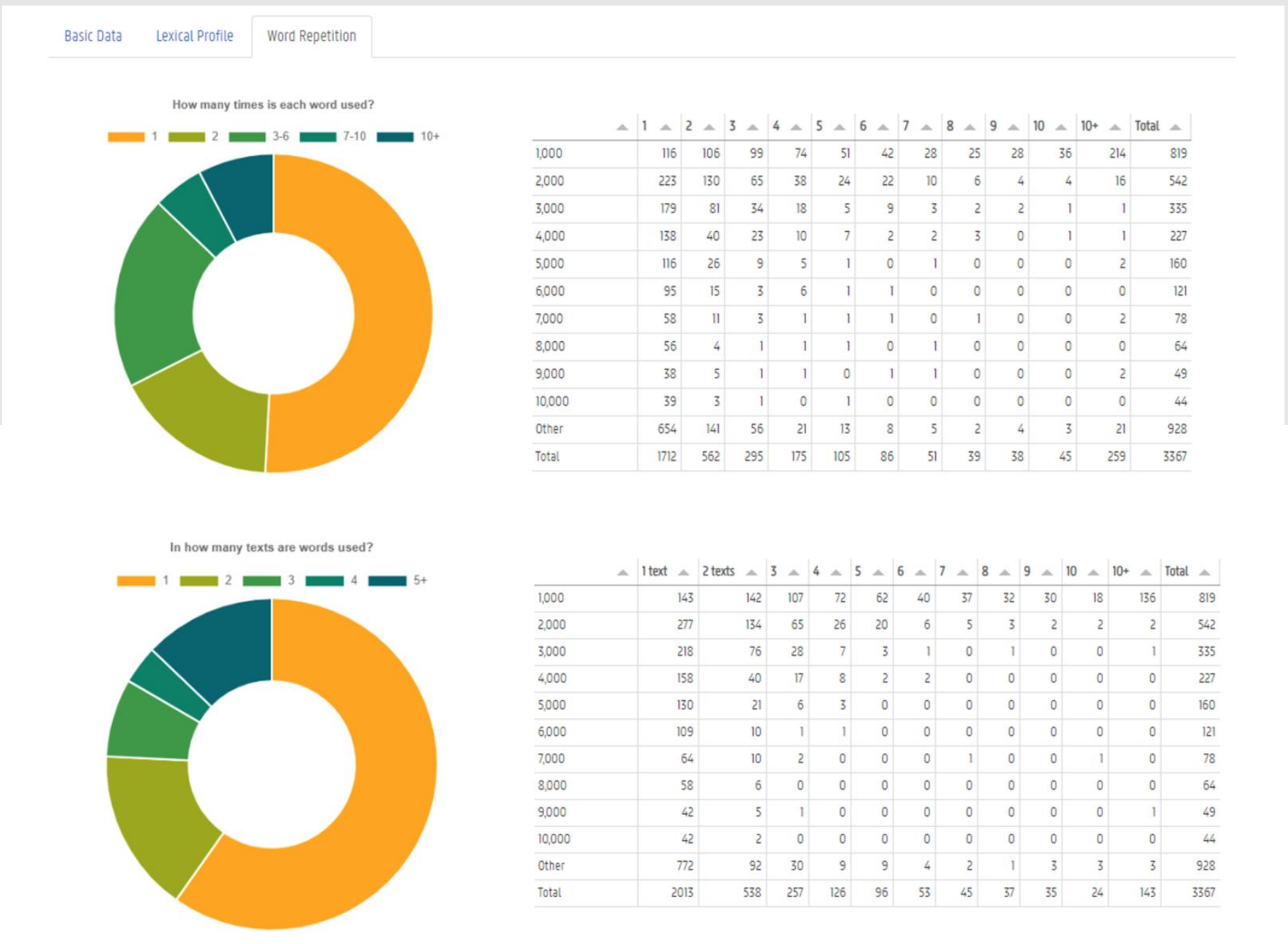
MW, PN, 1,000, NUM, 2,000, 1,000, 1,000, 3,000, 3,000, 1,000, 1,000, PN, 1,000, PN

Based on the retrieved information, a lexical profile is created (in addition to word repetition tables and visuals)



LexPro Plurilingual Lexical Profiling Tool for Research and Materials Development - Results

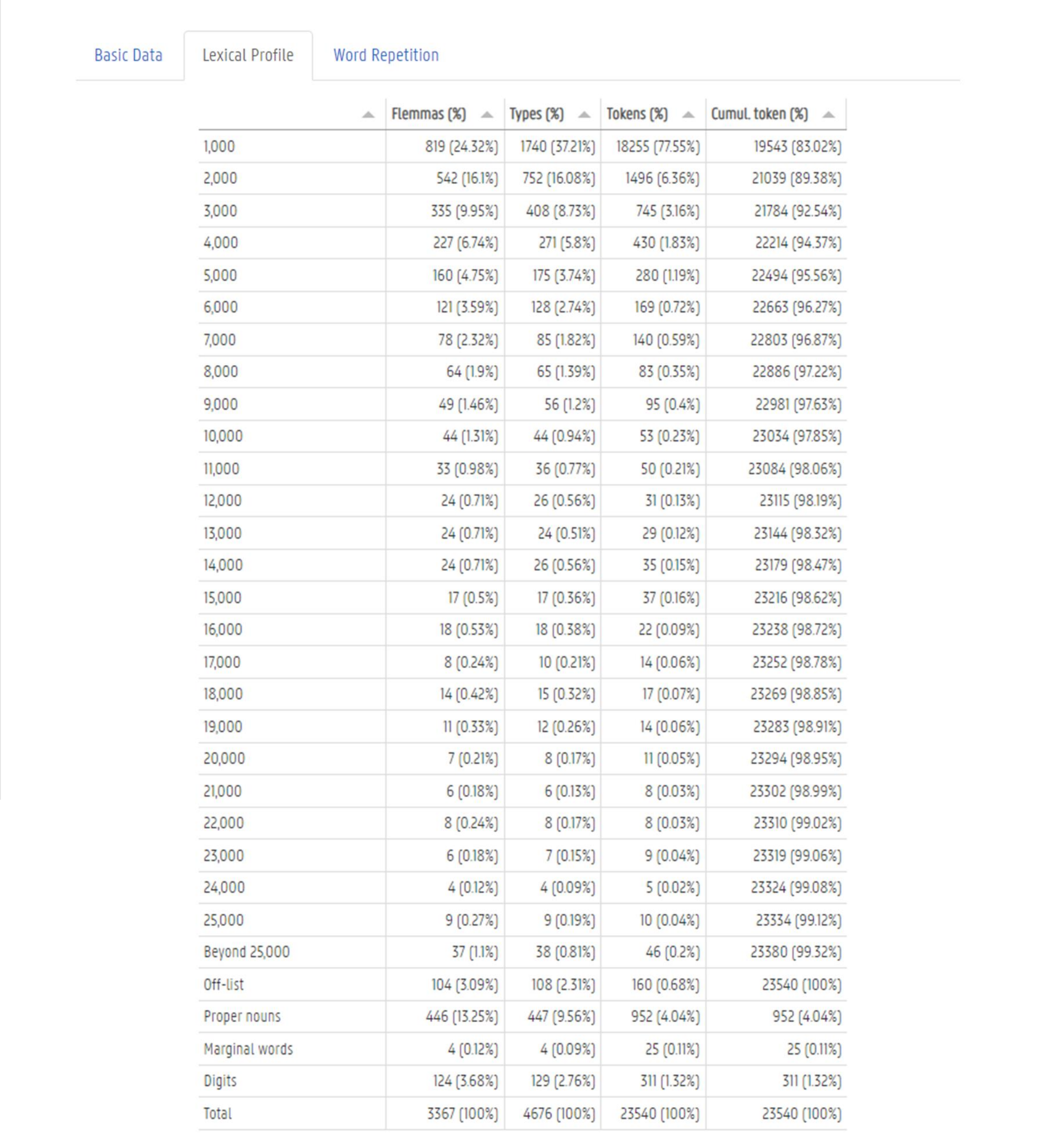
Number of files	53
Total word count (tokens)	23540
Mean word count (tokens / texts)	444.15
Type-token ratio	0.2



Basic Data
Lexical Profile
Word Repetition
How many times is each word used?
1
2
3-6
7-10
10+
In how many texts are words used?
1
2
3
4
5+

	1	2	3	4	5	6	7	8	9	10	10+	Total
1,000	116	106	99	74	51	42	28	25	28	36	214	819
2,000	223	130	65	38	24	22	10	6	4	4	16	542
3,000	179	81	34	18	5	9	3	2	2	1	1	335
4,000	138	40	23	10	7	2	2	3	0	1	1	227
5,000	116	26	9	5	1	0	1	0	0	0	2	160
6,000	95	15	3	6	1	1	0	0	0	0	0	121
7,000	58	11	3	1	1	1	0	1	0	0	2	78
8,000	56	4	1	1	1	0	1	0	0	0	0	64
9,000	38	5	1	1	0	1	1	0	0	0	2	49
10,000	39	3	1	0	1	0	0	0	0	0	0	44
Other	654	141	56	21	13	8	5	2	4	3	21	928
Total	1712	562	295	175	105	86	51	39	38	45	259	3367

	1 text	2 texts	3	4	5	6	7	8	9	10	10+	Total
1,000	143	142	107	72	62	40	37	32	30	18	136	819
2,000	277	134	65	26	20	6	5	3	2	2	2	542
3,000	218	76	28	7	3	1	0	1	0	0	1	335
4,000	158	40	17	8	2	2	0	0	0	0	0	227
5,000	130	21	6	3	0	0	0	0	0	0	0	160
6,000	109	10	1	1	0	0	0	0	0	0	0	121
7,000	64	10	2	0	0	0	1	0	0	1	0	78
8,000	58	6	0	0	0	0	0	0	0	0	0	64
9,000	42	5	1	0	0	0	0	0	0	0	1	49
10,000	42	2	0	0	0	0	0	0	0	0	0	44
Other	772	92	30	9	9	4	2	1	3	3	3	928
Total	2013	538	257	126	96	53	45	37	35	24	143	3367



	Flemmas (%)	Types (%)	Tokens (%)	Cumul. token (%)
1,000	819 (24.32%)	1740 (37.21%)	18255 (77.55%)	19543 (83.02%)
2,000	542 (16.1%)	752 (16.08%)	1496 (6.36%)	21039 (89.38%)
3,000	335 (9.95%)	408 (8.73%)	745 (3.16%)	21784 (92.54%)
4,000	227 (6.74%)	271 (5.8%)	430 (1.83%)	22214 (94.37%)
5,000	160 (4.75%)	175 (3.74%)	280 (1.19%)	22494 (95.56%)
6,000	121 (3.59%)	128 (2.74%)	169 (0.72%)	22663 (96.27%)
7,000	78 (2.32%)	85 (1.82%)	140 (0.59%)	22803 (96.87%)
8,000	64 (1.9%)	65 (1.39%)	83 (0.35%)	22886 (97.22%)
9,000	49 (1.46%)	56 (1.2%)	95 (0.4%)	22981 (97.63%)
10,000	44 (1.31%)	44 (0.94%)	53 (0.23%)	23034 (97.85%)
11,000	33 (0.98%)	36 (0.77%)	50 (0.21%)	23084 (98.06%)
12,000	24 (0.71%)	26 (0.56%)	31 (0.13%)	23115 (98.19%)
13,000	24 (0.71%)	24 (0.51%)	29 (0.12%)	23144 (98.32%)
14,000	24 (0.71%)	26 (0.56%)	35 (0.15%)	23179 (98.47%)
15,000	17 (0.5%)	17 (0.36%)	37 (0.16%)	23216 (98.62%)
16,000	18 (0.53%)	18 (0.38%)	22 (0.09%)	23238 (98.72%)
17,000	8 (0.24%)	10 (0.21%)	14 (0.06%)	23252 (98.78%)
18,000	14 (0.42%)	15 (0.32%)	17 (0.07%)	23269 (98.85%)
19,000	11 (0.33%)	12 (0.26%)	14 (0.06%)	23283 (98.91%)
20,000	7 (0.21%)	8 (0.17%)	11 (0.05%)	23294 (98.95%)
21,000	6 (0.18%)	6 (0.13%)	8 (0.03%)	23302 (98.99%)
22,000	8 (0.24%)	8 (0.17%)	8 (0.03%)	23310 (99.02%)
23,000	6 (0.18%)	7 (0.15%)	9 (0.04%)	23319 (99.06%)
24,000	4 (0.12%)	4 (0.09%)	5 (0.02%)	23324 (99.08%)
25,000	9 (0.27%)	9 (0.19%)	10 (0.04%)	23334 (99.12%)
Beyond 25,000	37 (1.1%)	38 (0.81%)	46 (0.2%)	23380 (99.32%)
Off-list	104 (3.09%)	108 (2.31%)	160 (0.68%)	23540 (100%)
Proper nouns	446 (13.25%)	447 (9.56%)	952 (4.04%)	952 (4.04%)
Marginal words	4 (0.12%)	4 (0.09%)	25 (0.11%)	25 (0.11%)
Digits	124 (3.68%)	129 (2.76%)	311 (1.32%)	311 (1.32%)
Total	3367 (100%)	4676 (100%)	23540 (100%)	23540 (100%)

Contact

amaury.vanparys@ugent.be
https://research.flw.ugent.be/en/amaury.vanparys

Universiteit Gent
@ugent
Ghent University

GHENT
UNIVERSITY