More than you ever wanted to know about: Lexical diversity

Day 3

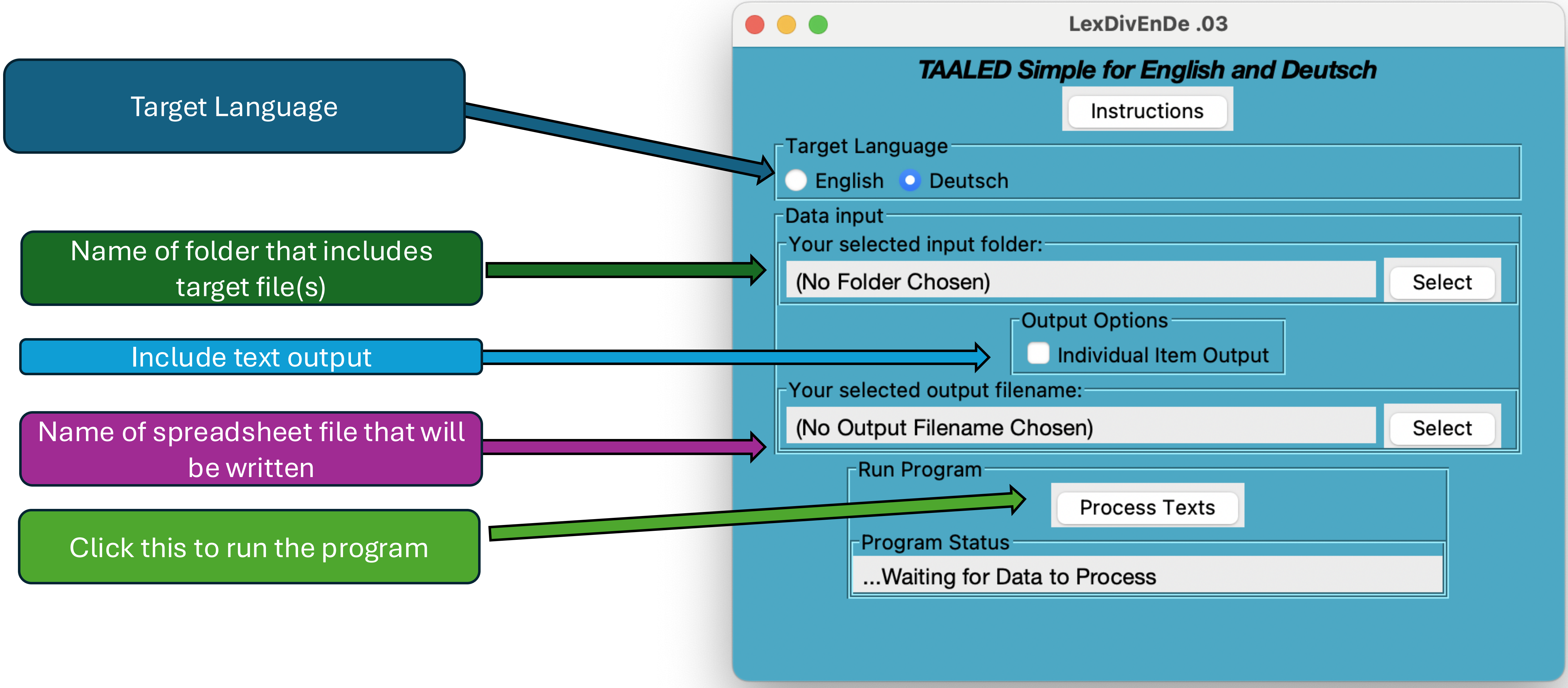
Kristopher Kyle, University of Oregon

https://github.com/LCR-ADS-Lab/

UNIVERSITY OF OREGON

TAALED_EnDe

- Can process English or German texts
- Currently works on Mac (not Windows – make friends with someone who has a Mac ☺)
- Takes plain text (.txt) files (with UTF-8 encoding) as input
- Files must be placed in a folder (TAALED processes all .txt files in a folder)
- Output includes:
  - Spreadsheet with lexical diversity scores for each text
  - A processed version of each text (to check how TAALED processed each file)


Target Language
Name of folder that includes target file(s)
Include text output
Name of spreadsheet file that will be written
Click this to run the program
LexDivEnDe .03
TAALED Simple for English and Deutsch
Instructions
Target Language
English
Deutsch
Data input
Your selected input folder:
(No Folder Chosen)
Select
Output Options
Individual Item Output
Your selected output filename:
(No Output Filename Chosen)
Select
Run Program
Process Texts
Program Status
...Waiting for Data to Process

Your texts + Your scores + LD

- How are things going?

An Empirical Evaluation of Lexical Diversity Indices in L2 Korean Writing Assessment

Sung, Cho, & Kyle (2024; Language Assessment Quarterly)

UNIVERSITY OF OREGON

Introduction

- Lexical diversity (LD) indices have been used as measures of lexical proficiency/development in L2 assessment studies (Crossley, Salsbury, McNamara, & Jarvis, 2011, Vidal & Jarvis, 2020, inter alia)
- Breadth of lexical knowledge refers to the number of lexical items that L2 learners know (Nation, 2001; Nassaji, 2004)
- More proficient L2 learners would use a wider variety of lexical items to complete a given production task

Research Background: 1. LD indices & Reliability

- Reliability: Developing more text-length stable indices
- Learners of different proficiency will produce different lengths of a text
  - L2 English writing assessment (Zenker & Kyle, 2021)

More reliable (revised)	Less reliable (classic)
HD-D (McCarthy & Jarvis, 2007)	Type-token ratio (TTR) (Johnson, 1944)
MATTR (Covington & McFall, 2010)	Log TTR (Carroll, 1938; Chotlos, 1944)
MTLD (McCarthy & Jarvis, 2010)	Root TTR (Guiraud, 1960)

Research Background: 2. LD indices & Validity

- Validity: Measuring correlations with proficiency scores / human judgments
- Valid LD index should measure lexical diversity itself
  - L2 English

Holistic proficiency / writing scores (indirect)	Human judgments on LD (direct)
Holistic judgments on essays (L2) (Engber, 1995)	Human ratings (L1, L2) (Jarvis, 2017)
CEFR level, vocab / writing scores (L2) (Treffers-Daller et al., 2018)	Human ratings (L1, L2) (Kyle et al., 2021)

Research Background: 3. Lexical item in Korean

- LD is an index of productive lexical breadth, measured using the diversity of lexical items in a text.
- [English LD] One lexical item refers to one (functional/lexical) word,
  a written unit based on space (Biber et al., 2021)
- The notion word cannot be defined consistently across languages (...) A written unit separated by spaces may not reflect an important grammatical unit in some languages (Haspelmath & Michaelis, 2017)

Research Background: 3. Lexical item in Korean

- The index of synthesis (the number of morphemes per word) (Payne, 2006)

one ⟷ lots

analytic / isolating languages — synthetic languages — polysynthetic languages

English — Korean

- *Eojeol*: Korean syntactic word separated by space
- Korean POS tagging has evolved from separating an eojeol into morphemes (Lee, Cha, & Lee, 2002)

Research Background: 3. Lexical item in Korean

sentence	나는		한국에		와서		태권도를		배웠다			
	na-nun		*han-kwuk-ey*		*wa-se*		*Taekwondo-lul*		*pay-wess-ta*			
translation	'I came to Korea and learned Taekwondo'											*n*token
Stanza	나는		한국에		와서		태권도를		배웠다			5
Okt	나	-는	한국	-에	와서		태권도	-를	배웠다			8
Mecab	나	-는	한국	-에	와서		태권도	-를	배웠 (stem+tense)		-다	9
Kkma	나	-는	한국	-에	오-	-아서	태권도	-를	배우-	-었-	-다	11

Tokenization type

eojeol

morpheme

josa

stem

prefinal-ending (tense)

final-ending

- Each Korean-specific tokenizer shows different parsing styles, which would influence LD calculation.

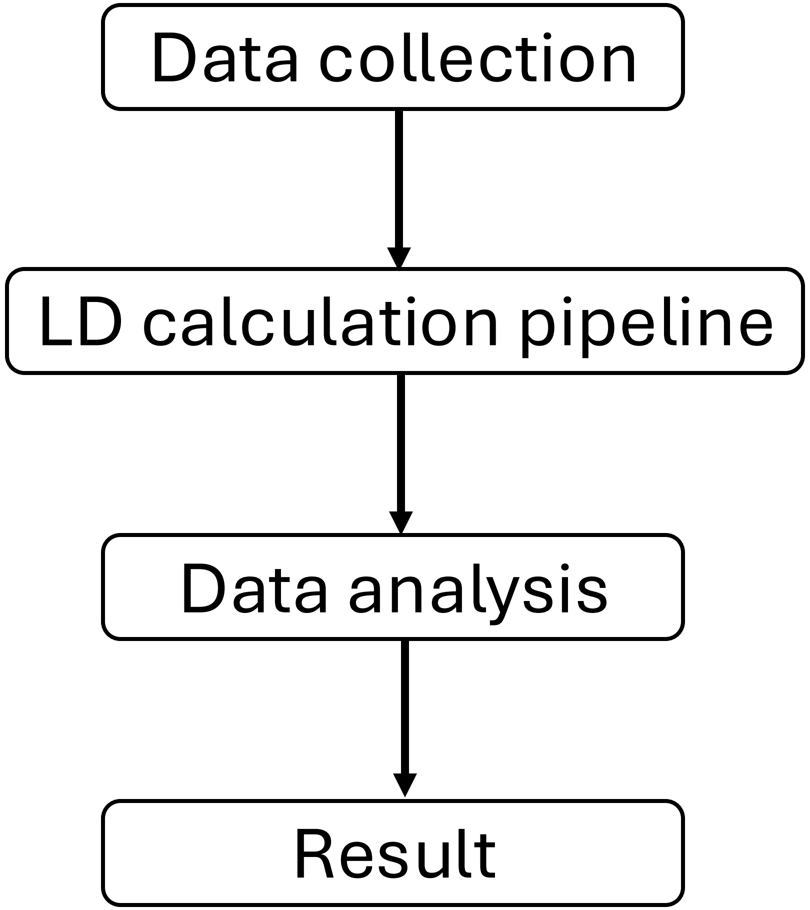
Research Object & Question

- Object: Evaluating LD of L2 Korean writing
  - Include various LD indices (both classic & revised indices)
  - Use Korean-specific tokenizers (Okt, Mecab, Kkma) (Stanza: reference)
- Questions

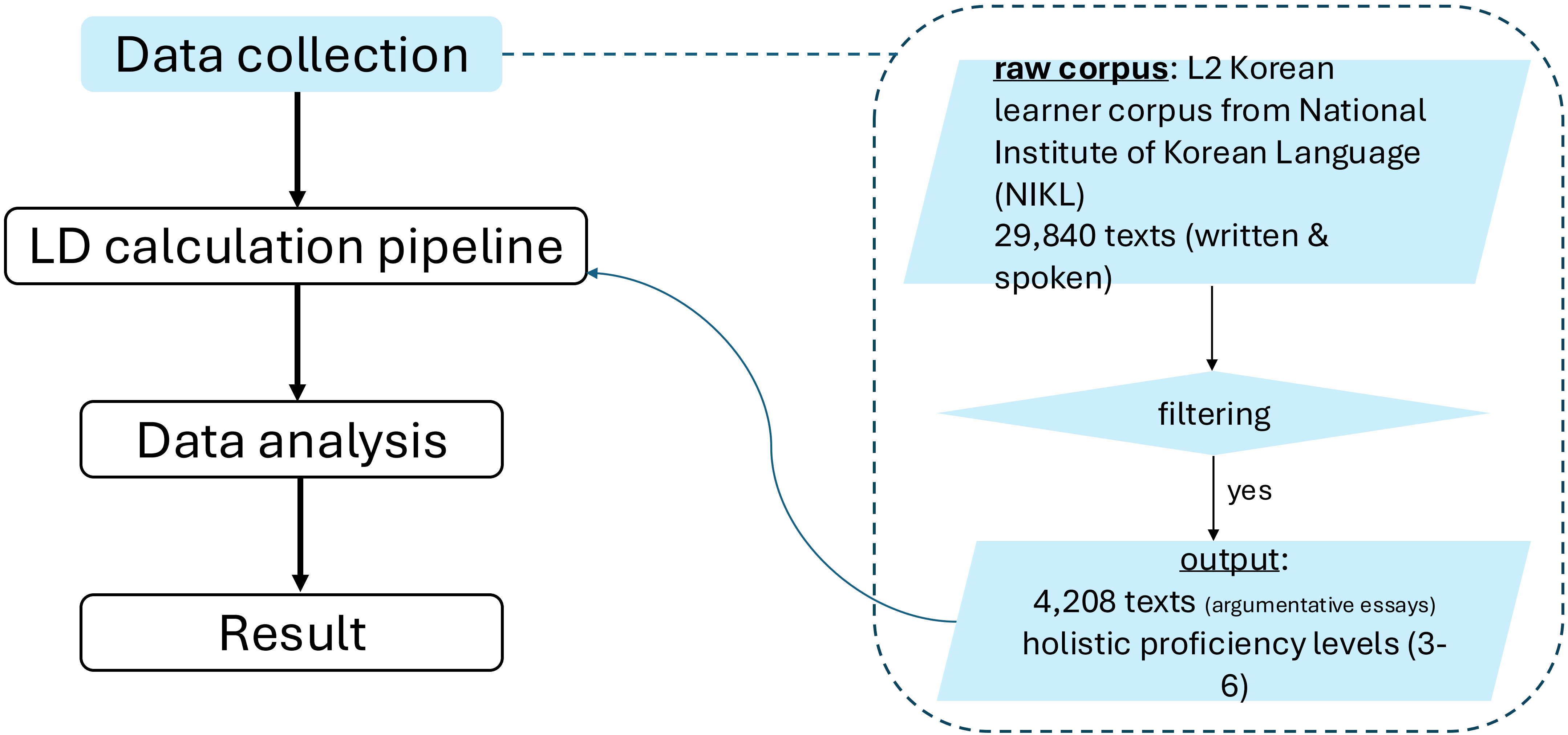
  [1] What is the relationship between LD indices and text length? (reliability)

  [2] What is the relationship between LD indices and holistic proficiency level? (validity)

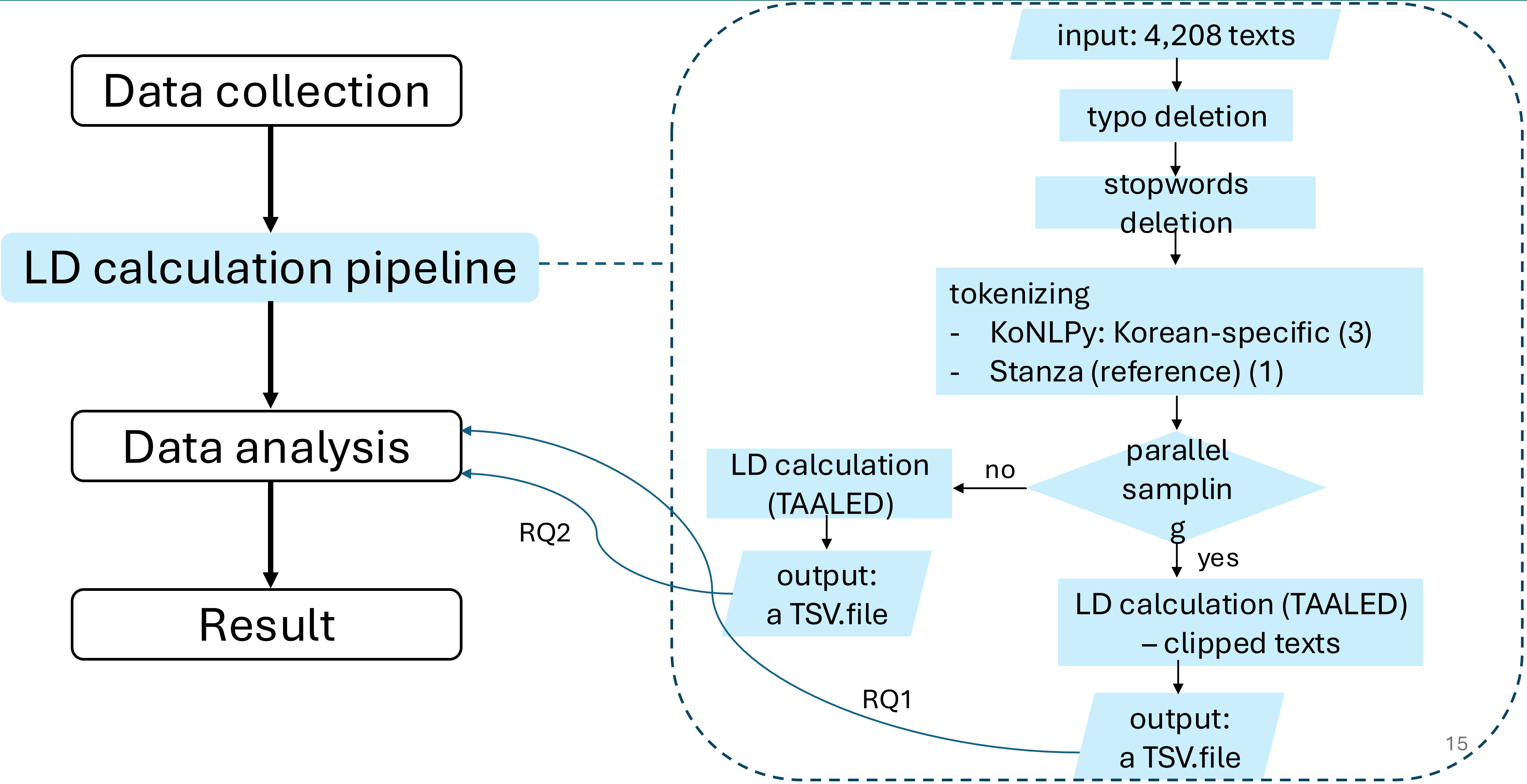
Research Process: Framework



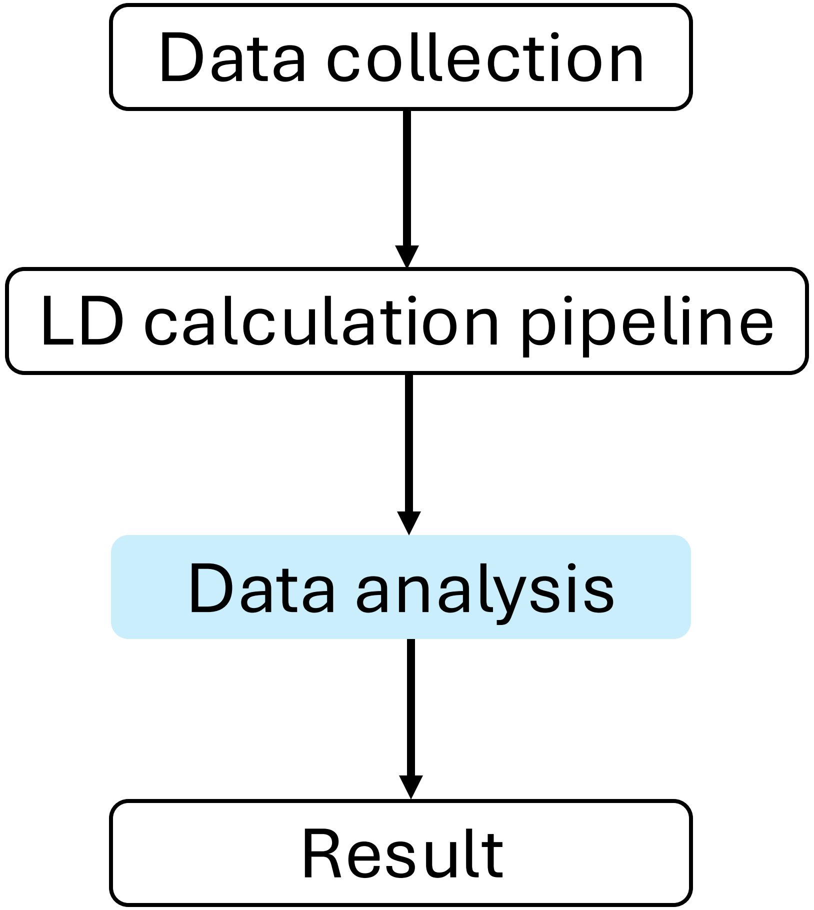
Research Process: Framework



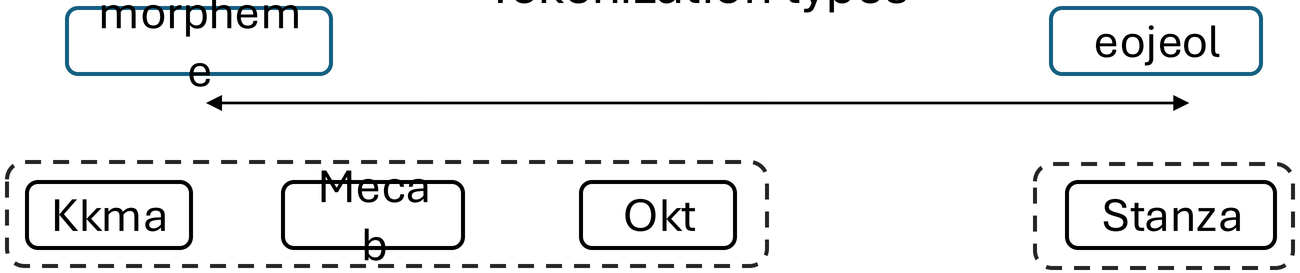
Research Process: Framework



Research Process: Framework



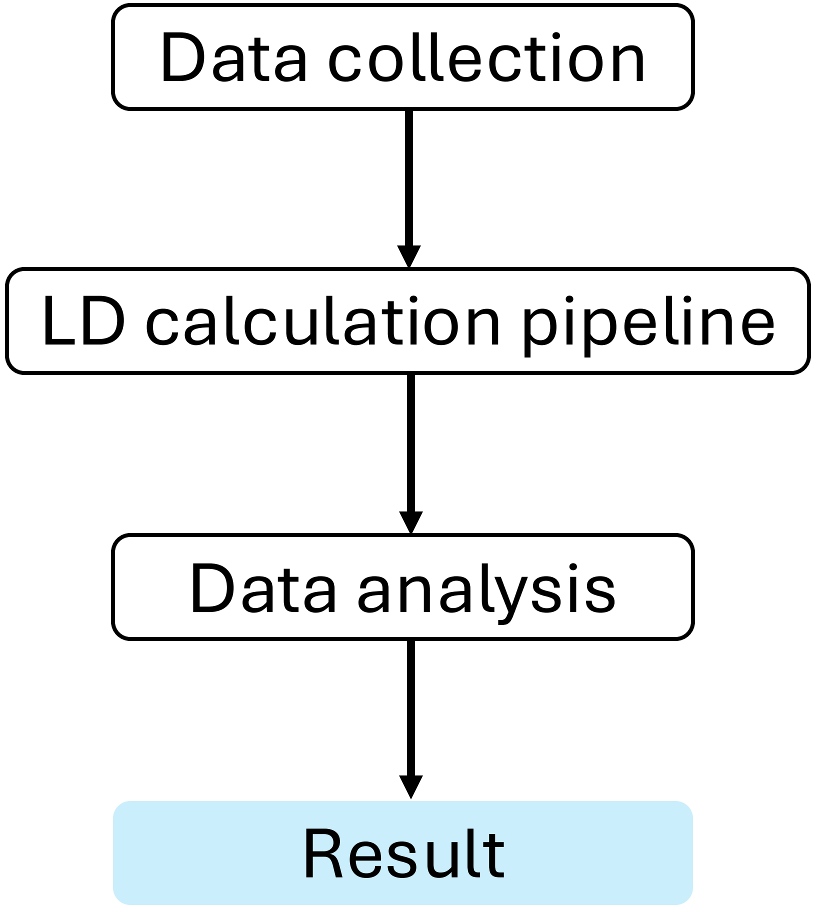
Tokenization types



LD indices types

Classic	TTR, Log TTR (LTTR), Root TTR (RTTR)
Revised	HD-D, MATTR, MTLD

Research Process: Framework

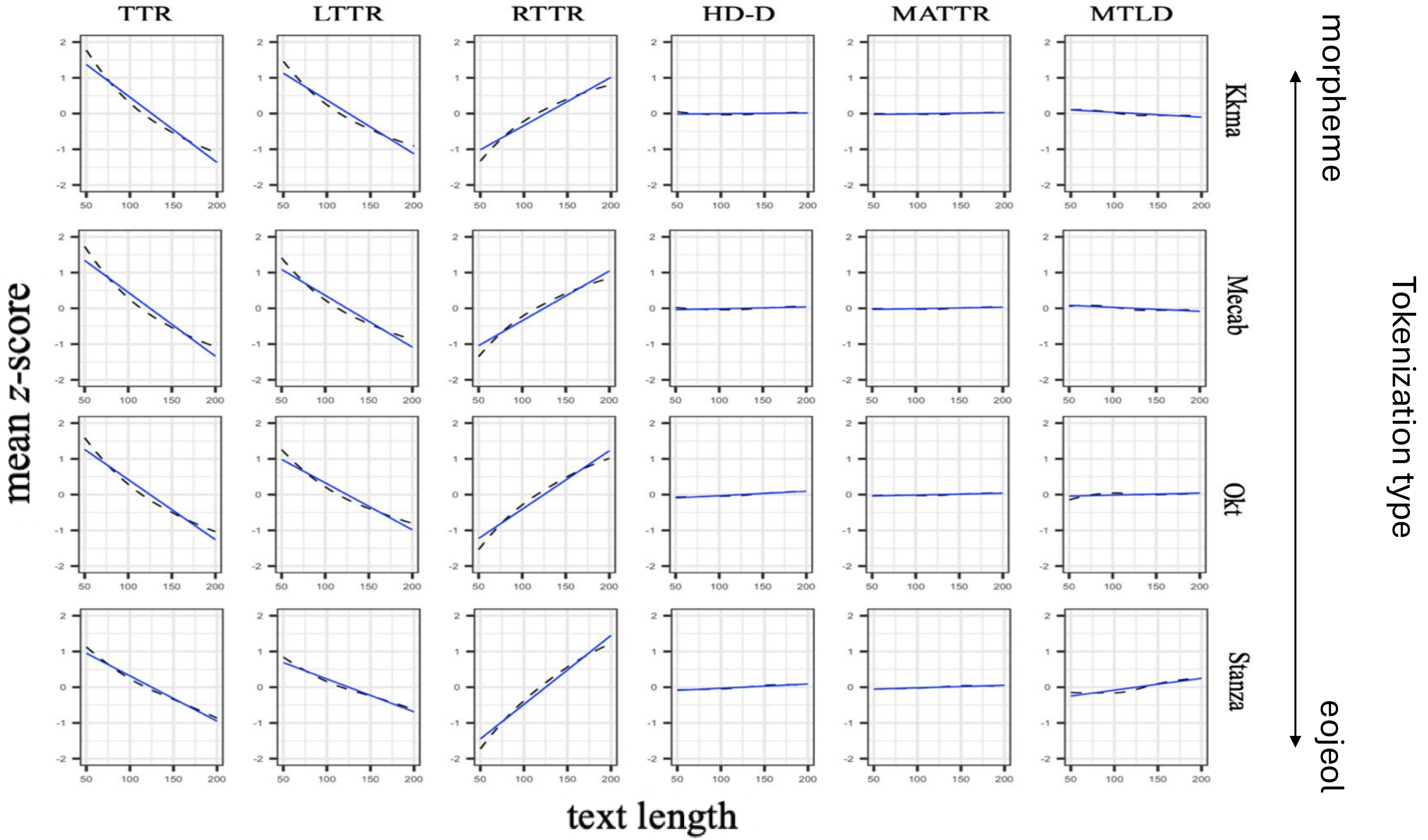


Result: RQ1

RQ1. What is the relationship between LD indices and text length?

- Mean *z*-score (text length – LD value)
- Correlation (text length – LD value)

Result: RQ1 | Mean *z*-score



Result: RQ1 | Correlations (text length – LD value)

Tokenization type

morpheme

eojeol

	TTR	LTTR	RTTR	HD-D	MATTR	MTLD
Kkma	-.818	-.673	.605	.009	.016	-.061
Mecab	-.780	-.648	.622	.023	.019	-.050
Okt	-.752	-.587	.732	.055	.022	.024
Stanza	-.566	-.411	.864	.053	.032	.148

Note. Colored boxes indicate absolute *r*-values below .100 (i.e., negligible effect size, Cohen, 1988)

Result: RQ1 | Discussion

- The mean *z*-scores and correlations indicate that HD-D, MATTR, and MTLD are stable across the text lengths (Zenker & Kyle, 2021)
- There was no noticeable difference across tokenizers
- MATTR remained the most stable among three most reliable indices, followed by HD-D
- MTLD were mostly stable with Kkma and Mecab, but it fluctuated with Okt and Stanza

Result: RQ2

RQ2. To what degree are LD indices predictive of proficiency levels?

- Descriptive statistics
- Correlation (proficiency level – LD value)
- Included *n*tokens & *n*types (Jarvis, 2017; Kyle et al., 2021)

Result: RQ2 | Descriptive statistics

Tokenization type: morpheme ↔ eojeol

		*n*tokens	*n*types	TTR	LTTR	RTTR	HD-D	MATTR	MTLD
Kkma	mean (SD)	270.58 (60.67)	113.28 (21.45)	0.42 (0.05)	0.85 (0.02)	6.89 (0.78)	0.76 (0.03)	0.74 (0.04)	53.14 (10.57)
	min	94	51	0.22	0.75	4.11	0.63	0.58	25.50
	max	585	216	0.64	0.92	9.74	0.85	0.84	107.53
Mecab	mean (SD)	250.88 (56.79)	112.55 (21.94)	0.45 (0.06)	0.86 (0.02)	7.11 (0.83)	0.78 (0.03)	0.75 (0.04)	58.01 (12.98)
	min	91	52	0.24	0.76	4.18	0.64	0.58	25.59
	max	551	220	0.69	0.93	10.24	0.87	0.86	136.55
Okt	mean (SD)	205.81 (47.51)	113.50 (22.98)	0.56 (0.06)	0.89 (0.02)	7.91 (0.92)	0.82 (0.04)	0.79 (0.04)	76.70 (22.74)
	min	69	48	0.30	0.79	4.68	0.67	0.61	28.80
	max	441	230	0.78	0.95	11.42	0.91	0.91	196.16
Stanza	mean (SD)	138.57 (30.97)	105.58 (23.24)	0.77 (0.07)	0.94 (0.02)	8.94 (1.12)	0.90 (0.04)	0.88 (0.05)	171.79 (70.68)
	min	54	43	0.44	0.85	5.29	0.67	0.64	28.99
	max	295	230	0.96	0.99	13.39	0.98	1.00	700.00

Result: RQ2 | Correlations (LD value, proficiency level)

Tokenization type

morpheme

eojeol

		*n*tokens	*n*types	TTR	LTTR	RTTR	HD-D	MATTR	MTLD
Kkma	level	0.206***	**0.293*****	0.076***	0.156***	0.289***	0.252***	0.263***	0.264***
Mecab		0.205***	**0.294*****	0.087***	0.162***	0.292***	0.272***	0.274***	0.274***
Okt		0.245***	**0.280*****	0.006***	0.069***	0.248***	0.180***	0.191***	0.179***
Stanza		-0.037*	-0.029.	0.014	0.010	-0.022	-0.003	0.002	-0.009

Note. *** indicates that *p*-value was less than .001 ** indicates that *p*-value was less than .01
* indicates that *p*-value was less than .05, . indicates that *p*-values was less than .1

Result: R2 | Discussion

- Overall, the *n*types demonstrates the largest correlation with holistic level (Mecab, r = .294) (Jarvis, 2017; Kyle el al., 2021: abundance (*n*types))
- *n*types is strongly correlated to *n*tokens (Mecab, r = .830)
- Among the indices that provided reliable values across different text-length (in RQ1; HD-D, MATTR, MTLD), MATTR (r = .274) and MTLD (r = .274) showed the largest correlations when they were calculated by Mecab
- This is still a small correlation (r < .300), but the result shows that Korean-specific tokenizers give more valid LD scores compared to Stanza.

Summary

[RQ1]

- The revised LD indices, HD-D, MATTR, and MTLD, were more reliable than TTR, LTTR, and RTTR regardless of the type of tokenizers

[RQ2]

- The text-length stable indices (HD-D, MATTR, MTLD) demonstrated small correlations to proficiency level
- Among various pairs of tokenizers and LD indices, [Mecab – MATTR/MTLD] represented the best options
- Morpheme-based tokenizers would be more valid way to calculate LD indices than eojeol-based tokenizers

Limitation and future direction

- In terms of tokenizer, other characteristics of Korean-specific tokenizers are underresearched (e.g., accuracy)
- In terms of validity, we need to compare the relationship between human judgment and different ways of tokenizing (w/wo lemmatization)

References 1

Bai, D-Y. (2012). A study on the lexical variation and lexical density shown in writing of KFL learners. *Journal of Language Sciences, 19*(1), 99-117.

Biber, D., Johansson, S., Leech, G. N., Conrad, S., & Finegan, E. (2021). *Grammar of spoken and written English*. John Benjamins.

Carroll, J. B. (1964). Language and thought. *Reading Improvement*, *2*(1), 80.

Covington, M. A., & McFall, J. D. (2010). Cutting the Gordian Knot: The Moving-Average Type–Token Ratio (MATTR). *Journal of Quantitative Linguistics*, *17*(2), 94–100. https://doi.org/10.1080/09296171003643098

Crossley, S. A., Salsbury, T., McNamara, D. S., & Jarvis, S. (2011). Predicting lexical proficiency in language learner texts using computational indices. *Language Testing*, *28*(4), 561-580.

Engber, C. A. (1995). The relationship of lexical proficiency to the quality of ESL compositions. *Journal of Second Language Writing*, *4*(2), 139–155. https://doi.org/10.1016/1060-3743(95)90004-7

Guiraud, P. (1960). *Problèmes et méthodes de la statistique linguistique*. Presses universitaires de France.

Haspelmath, M., & Michaelis, S. M. (2017, July). Analytic and synthetic. In *Language Variation-European Perspectives VI: Selected papers from the Eighth International Conference on Language Variation in Europe (ICLaVE 8), Leipzig 2015*. John Benjamins Publishing Company.

Jarvis, S. (2017). Grounding lexical diversity in human judgments. *Language Testing*, *34*(4), 537–553. https://doi.org/10.1177/0265532217710632

Johnson, W. (1944). I. A program of research. *Psychological Monographs*, *56*(2), 1–15. https://doi.org/10.1037/h0093508

Kang, J-H. (2018). A study on the vocabulary development and the lexicon use aspect of Korean learners – Focusing on the lexicon development appearing in the writing test. *Bilingual Research, 71,* 31-64.

Koizumi, R. (2012). Relationships between text length and lexical diversity measures: Can we use short texts of less than 100 tokens? *Vocabulary Learning and Instruction*, *01*(1), 60–69. https://doi.org/10.7820/vli.v01.1.koizumi

References 2

Koizumi, R., & In’nami, Y. (2012). Effects of text length on lexical diversity measures: Using short texts with less than 200 tokens. *System*, *40*(4), 554–564. https://doi.org/10.1016/j.system.2012.10.012

Kyle, K., Crossley, S. A., & Jarvis, S. (2021). Assessing the Validity of Lexical Diversity Indices Using Direct Judgements. *Language Assessment Quarterly*, *18*(2), 154–170. https://doi.org/10.1080/15434303.2020.1844205

Lee, G. G., Cha, J., & Lee, J.-H. (2002). Syllable-Pattern-Based Unknown-Morpheme Segmentation and Estimation for Hybrid Part-of-Speech Tagging of Korean. *Computational Linguistics*, *28*(1), 53–70. https://doi.org/10.1162/089120102317341774

McCarthy, P. M., & Jarvis, S. (2007). vocd: A theoretical and empirical evaluation. *Language Testing*, *24*(4), 459–488. https://doi.org/10.1177/0265532207080767

McCarthy, P. M., & Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, *42*(2), 381–392. https://doi.org/10.3758/BRM.42.2.381

Nation, P. (2001). *Learning vocabulary in another language*. Cambridge University Press.

Nassaji, H. (2006). The Relationship Between Depth of Vocabulary Knowledge and L2 Learners’ Lexical Inferencing Strategy Use and Success. *The Modern Language Journal*, *90*(3), 387–401.

Payne, T. (2006). *Exploring language structure: A student's guide*. Cambridge University Press.

Treffers-Daller, J., Parslow, P., & Williams, S. (2018). Back to basics: How measures of lexical diversity can help discriminate between CEFR levels. *Applied Linguistics*, *39*(3), 302-327.

Vidal, K., & Jarvis, S. (2020). Effects of English-medium instruction on Spanish students’ proficiency and lexical diversity in English. *Language Teaching Research*, *24*(5), 568-587.

Zenker, F., & Kyle, K. (2021). Investigating minimum text lengths for lexical diversity indices. *Assessing Writing*, *47*, 100505. https://doi.org/10.1016/j.asw.2020.100505

UD-KSL-treebank: Annotations

- No funding for annotations, but a good colleague
- v1.1 (# sents = 7,530) (2023, 2024)
  - Started small, only with the morpheme annotations
  - Expanded to dependency relations
- v1.2 (# sents = 12,984) (2025a)
  - Able to recruit 2-3 annotators
- v1.3 (# sents = 15,982) (2025b)
  - Bigger research group

Gyu-Ho Shin

UIC UNIVERSITY OF ILLINOIS CHICAGO

Sung, H., Shin, G-H., Sung, Y., Lee, C., Jung, B-K., (*Accepted*). In *Proceedings of 19th Linguistic Annotation Workshop*

```
# sent_id = B200018-9-3968
# text = 그래서 나는 그 산 앞에서 옷을 입고 사진을 찍어 기념으로 남겼다.
1	그래서	그래서	ADV	MAJ	_	11	cc	_	_
2	나는	나+는	PRON	NP+JX	_	11	nsubj	_	_
3	그	그	DET	MM	_	4	det	_	_
4	산	산	NOUN	NNG	_	7	obl	_	_
5	앞에서	앞+에서	ADP	NNG+JKB	_	4	case	_	_
6	옷을	옷+을	NOUN	NNG+JKO	_	7	obj	_	_
7	입고	입+고	VERB	VV+EC	_	9	advcl	_	_
8	사진을	사진+을	NOUN	NNG+JKO	_	9	obj	_	_
9	찍어	찍+어	VERB	VV+EC	_	11	advcl	_	_
10	기념으로	기념+으로	ADV	NNG+JKB	_	11	obl	_	_
11	남겼다	남기+었+다	VERB	VV+EP+EF	_	0	root	_	SpaceAfter=No
12	.	.	PUNCT	SF	_	11	punct	_	_
```

Tag	Description	Tag	Description
NNG	Noun, common	EP	Ending, prefinal
NNP	Noun, proper	EF	Ending, closing
NNB	Noun, bound	EC	Ending, connecting
NR	Numeral	ETN	Ending, nounal
NP	Pronoun	ETM	Ending, determinative
VV	Verb, main	XPN	Prefix, nounal
VA	Adjective	XSN	Suffix, noun derivative
VX	Verb, auxiliary	XSV	Suffix, verb derivative
VCP	Copular, positive	XSA	Suffix, adjective derivative
VCN	Copular, negative	XR	Root
MM	Determiner	NF	Undecided (considered as a noun)
MAG	Adverb, common	NV	Undecided (considered as a predicate)
MAJ	Adverb, conjunctive	NA	Undecided
IC	Exclamation	SF	Period, Question, Exclamation
JKS	Case particle, nominative	SE	Ellipsis
JKG	Case particle, prenominal	SP	Comma, Colon, Slash
JKO	Case particle, objectival	SO	Hyphen, Swung Dash
JKB	Case particle, adverbial	SW	Symbol
JKC	Case particle, complement	SS	Quotation, Bracket, Dash
JKV	Case particle, vocative	SH	Chinese characters
JKQ	Case particle, conjunctive	SL	Foreign characters
JX	Case particle, auxiliary	SN	Number

UD-KSL-treebank: Fine-tuning

- **Evaluated/trained** models with the annotated dataset

			BiLSTM	*tok2vec*	*transformer*
Dataset	Metric	Baseline	Stanza	SpaCy	Trankit
L2K-UD-test (in-domain)	XPOS	82.44	89.72	83.15	**91.81**
	LEMMA	89.61	**95.64**	87.97	88.84
	UAS	76.72	85.53	82.21	**92.28**
	LAS	60.69	80.36	75.21	**89.13**
KoLLA (out-of-domain)	XPOS	77.79	81.87	71.21	**84.51**
	LEMMA	88.03	**91.01**	79.64	86.90
	UAS	72.30	81.17	74.48	**88.93**
	LAS	58.53	75.14	63.56	**85.45**

Table 2: Evaluation metrics