Lexical features in adolescents' writing: Insights from the trilingual parallel corpus SWIKO

Nina Selina HICKS

Institute of Multilingualism
University of Fribourg, Switzerland

Workshop on Profiling second language vocabulary and grammar
University of Gothenburg, April 2023

SWIKO – Context & aim

- Shift in FL teaching in Switzerland
  (Council of Europe 2001, 2020; EDK 2011, 2017; Bertschy et al. 2015; Lenz & Wiedenkeller 2019; Peyer et al. 2016)
  - Competence based; action, content & tasks oriented
  - HarmoS: 2 FL mandatory (~ ages 8 + 10), one national language + English, goal = A2.2 overall, A2.1 writing
- Empirically document & analyze learners' interlanguage at the end of mandatory schooling
  - Lexical & grammatical features

Theoretical background & previous work

- Task effects on CAF measures
  (Housen et al. 2012, 2019, Norris & Ortega 2009, Wolfe-Quintero et al. 1998)

- "There is currently no theory of how the myriad design variables interact to affect task complexity."
  (Ellis et al. 2020: 348, emphasis added)

- Previous studies on lexical features
  (e.g., Abdi Tabari et al. 2021, Alexopoulou et al. 2017, Berman 2008, Bi 2020, Eckstein & Ferris 2018, Olinghouse & Wilson 2012, Ong 2014, Qin & Uccelli 2016, 2020, Révész et al. 2017, Vyatkina 2012, Yoon 2017, Yu 2010, Yoon & Polio 2017)
  - Large-scale exams or few (advanced college-level) classes
  - One task characteristic
  - One language and acquisitional context

Research question

How do different task characteristics – rhetorical type, topic familiarity and structure – influence lexical features in Swiss adolescents' low-level written productions…

…in three languages (German, English, French)?

…across two acquisitional contexts (language of schooling L1 vs. foreign language FL)?

SWIKO – Task variation

Task characteristics	Variation	SWI01	SWI02	SWI03	SWI04	SWI05	SWI06	SWI07	SWI08
Rhetorical type	descriptive	x	x			x	x		
	argumentative			x	x			x	x
Topic / familiarity	personal	x		x		x		x	
	academic		x		x		x		x
Structure	more	x	x	x	x				
	less					x	x	x	x

cf. *tasks* in Ellis et al. 2020

Task production conditions:

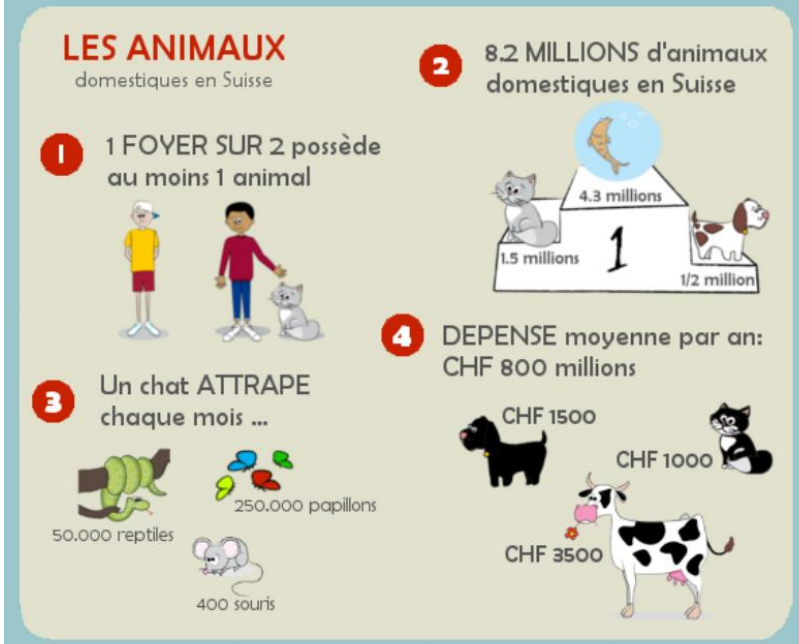
– 3 languages (language of schooling "L1" & 2 foreign languages "FL")
– 2 modes: writing and speaking (Karges, Studer & Hicks 2022)
– 2 types of medium: paper and computer (Karges, Studer & Wiedenkeller 2017, 2020)

SWIKO - Example task SWI02

Sur Internet, tu as trouvé un graphique sur les animaux domestiques en Suisse.

Regarde le graphique et réfléchis. Quelles informations dois-tu donner pour décrire le graphique à un(e) collègue qui ne peut pas voir le graphique ?

Puis, clique sur « Suivant ». Sur la prochaine page, tu enregistreras ton texte.



- Language of schooling (L1): French
- Target language: German FL
- Mode: writing
- Medium: computer-based
- Rhetorical type: descriptive
- Topic: academic
- Structure: more

Suivant

SWIKO - Example task SWI07

SWI07_e

Des politiciens proposent que les cours à l'école secondaire ne commencent qu'à 10 heures du matin. En revanche, les pauses de midi seraient raccourcies et l'école ne finirait qu'à 17h30.

Write in English!

Écris en anglais !

Qu'est-ce que tu penses de cette idée ? Donne des arguments pour et contre.

Écris environ 60-80 mots.

- Language of schooling: French
- Target language: English FL
- Mode: speaking
- Medium: paper-based
- Rhetoric type: argumentative
- Topic: personal
- Structure: less

Example: SWI02 vs. SWI07 (DaF)

SWI02: describe a graph (descriptive, academic, more structured)

Ein katz essen vier Hunden Mause in einen Monate. In die Schweiz, Es gibt 1,5 millionen Katze und ein Katze cost 1000 für zwölf monat.

[A cat eats four hundred mice in a month. In Switzerland, there are 1,5 million cats and a cat costs 1000 for twelve months.]

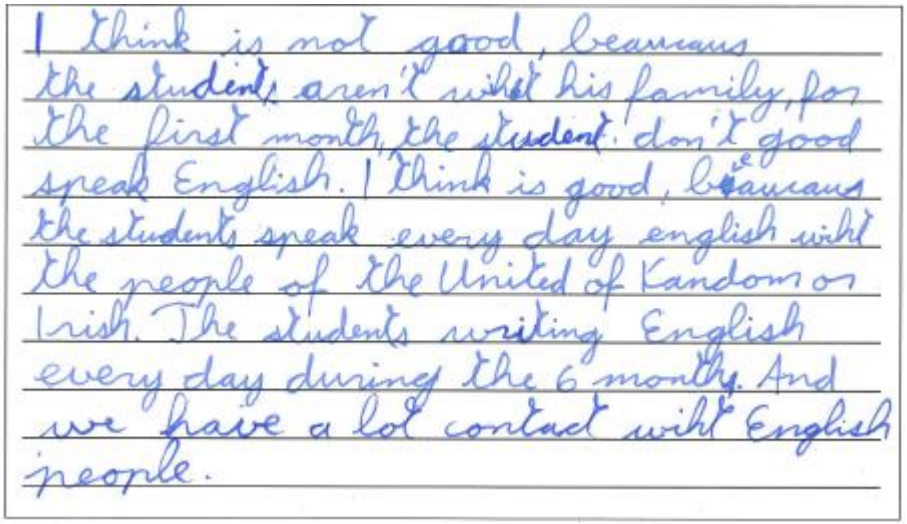
SWI07: discuss later school hours (argumentative, personal, less structured)

Ja, Das ist sehr gut weil, wir habe + schlafen weil, Wir sind sehr müde in die schule. Auch, ich habe 30 minuten minus die Schule normal. Aber, Wir haben finir plus tard in 5.30

[Yes, this is very good because we get more sleep because we are very tired at school. Also, I have 30 minutes less than usually at school. But we have to finish later at 5:30pm.]

SWIKO – Transcription & annotation

The original learner texts…


I think is not good, beaucaus
the students aren't with his family, for
the first month, the student don't good
speak English. I think is good, beaucaus
the students speak every day english with
the people of the United of Kandom or
Irish. The students writing English
every day during the 6 monthy. And
we have a lot contact with English
people.

SWIKO – Transcription & annotation

The original learner texts…

↓

transcribed and normalized manually…

Copyright 2012 project Merlin, http://merlin-platform.eu; adapted for
Transcriber: JBE
Checked by: NMU
Author ID: Mo121
Task ID: SWI08_fE
Medium: p
Original Text:
I think is not good, beaucaus the students aren't whit his family, for the first month, the student don't good speak English. I think is good beaucaus the students speak every day english wiht the people of the United of Kandom or Irish. The students writing English every day during the #6# month. And we have a lot contact wiht English people.

Tagged Text:
I think is not good, [beaucaus because] the students aren't [wiht with] his family, for the first month, the student don't good speak English. I think is good, [beaucaus because] the students speak every day [english English] [wiht with] the people of the United [of Kandom Kingdom]or Irish. The students writing English every day during the #6# months. And we have a lot contact [wiht with] English people.

*In XMLmind with the help of an xml-specification kindly provided by the **MERLIN project**: "MERLIN-Multilingual Platform for European Reference Levels: Interlanguage Exploration in Context" (project number: 518989-LLP-1-2011-1-DE-KA2-KA2MP).*

SWIKO – Transcription & annotation

The original learner texts…

↓

transcribed and annotated structurally…

↓

POS annotated…

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	original	doc_id	token	common.t	en.POS.ta	lemma	lttr	wclass	desc	stop	stem	idx	sntc	tag	markup	comment
2	I	SWI08_fE	I	PRO:PER	PP	I	1	pronoun	NA	FALSE	NA	1	1	NA	NA	NA
3	think	SWI08_fE	think	VER:PRE	VVP	think	5	verb	NA	FALSE	NA	2	1	NA	NA	NA
4	is	SWI08_fE	is	VER:PRE	VBZ	be	2	verb	NA	FALSE	NA	3	1	NA	NA	NA
5	not	SWI08_fE	not	ADV	RB	not	3	adverb	NA	FALSE	NA	4	1	NA	NA	NA
6	good	SWI08_fE	good	ADJ	JJ	good	4	adjective	NA	FALSE	NA	5	1	NA	NA	NA
7	,	SWI08_fE	,	$,	,	,	1	comma	NA	FALSE	NA	6	1	NA	NA	NA
8	beaucaus	SWI08_fE	because	PRP	IN	because	7	preposition	NA	FALSE	NA	7	1	error	NA	NA
9	the	SWI08_fE	the	DET	DT	the	3	determine	NA	FALSE	NA	8	1	NA	NA	NA
10	students	SWI08_fE	students	NN	NNS	student	8	noun	NA	FALSE	NA	9	1	NA	NA	NA
11	aren't	SWI08_fE	are	VER:PRE	VBP	be	3	verb	NA	FALSE	NA	10	1	NA	NA	NA
12	aren't((2))	SWI08_fE	n't	ADV	RB	n't	3	adverb	NA	FALSE	NA	11	1	NA	NA	NA
13	wiht	SWI08_fE	whit	NN	NN	whit	4	noun	NA	FALSE	NA	12	1	NA	NA	NA
14	his	SWI08_fE	his	PRO:POS	PP$	his	3	pronoun	NA	FALSE	NA	13	1	NA	NA	NA
15	family	SWI08_fE	family	NN	NN	family	6	noun	NA	FALSE	NA	14	1	NA	NA	NA
16	,	SWI08_fE	,	$,	,	,	1	comma	NA	FALSE	NA	15	1	NA	NA	NA
17	for	SWI08_fE	for	PRP	IN	for	3	preposition	NA	FALSE	NA	16	1	NA	NA	NA
18	the	SWI08_fE	the	DET	DT	the	3	determine	NA	FALSE	NA	17	1	NA	NA	NA
19	first	SWI08_fE	first	ADJ	JJ	first	5	adjective	NA	FALSE	NA	18	1	NA	NA	NA
20	month	SWI08 fE	month	NN	NN	month	5	noun	NA	FALSE	NA	19	1	NA	NA	NA

by TreeTagger (Schmid 2013), wrapped in the koRpus package (Michalke 2017) in R (R Core Team 2022).

SWIKO – Transcription & annotation

The original learner texts…

↓

transcribed and annotated structurally…

↓

POS annotated…

↓

and imported to EXMARaLDA for further analyses.

	0	1	2	3	4	5	6	7	8	9	10
Mol21 [tok]	I	think	is	not	good	,	beaucaus	the	students	aren't	
Mol21 [ctok]	I	think	is	not	good	,	because	the	students	are	n't
Mol21 [lemma]	I	think	be	not	good	,	because	the	student	be	n't
Mol21 [clemma]											
Mol21 [commonPOS]	PRO:PER	VER:PRE	VER:PRE	ADV	ADJ	$,	PRP	DET	NN	VER:PRE	ADV
Mol21 [lg-specific POS]	PP	VVP	VBZ	RB	JJ	,	IN	DT	NNS	VBP	RB
Mol21 [cpos]											
Mol21 [tag]							error				
Mol21 [markup]											
[comment]											

Schmidt & Wörner 2009

SWIKO – Scope (as of 31.03.2022)

Language	German		French		English		TOTAL
	FL	L1	FL	L1	FL	L1	
Class level	10 & 11	11 & 12	11 & 12	10 & 11	10 - 12	10	
Written							
Originals	524	355	396	426	770	103	**2'574**
Transcripts	499	344	*299*	*149*	475	101	**1'867**
Tokens	*23121*	*23648*	*17208*	*10425*	*33239*	8490	**116'310**
Spoken							
Originals	49	72	57	64	140	28	**410**
Transcripts	42	72	7	0	106	8	**235**
Tokens	2157	13533	174	0	15009	3028	**33'901**

Lexical features

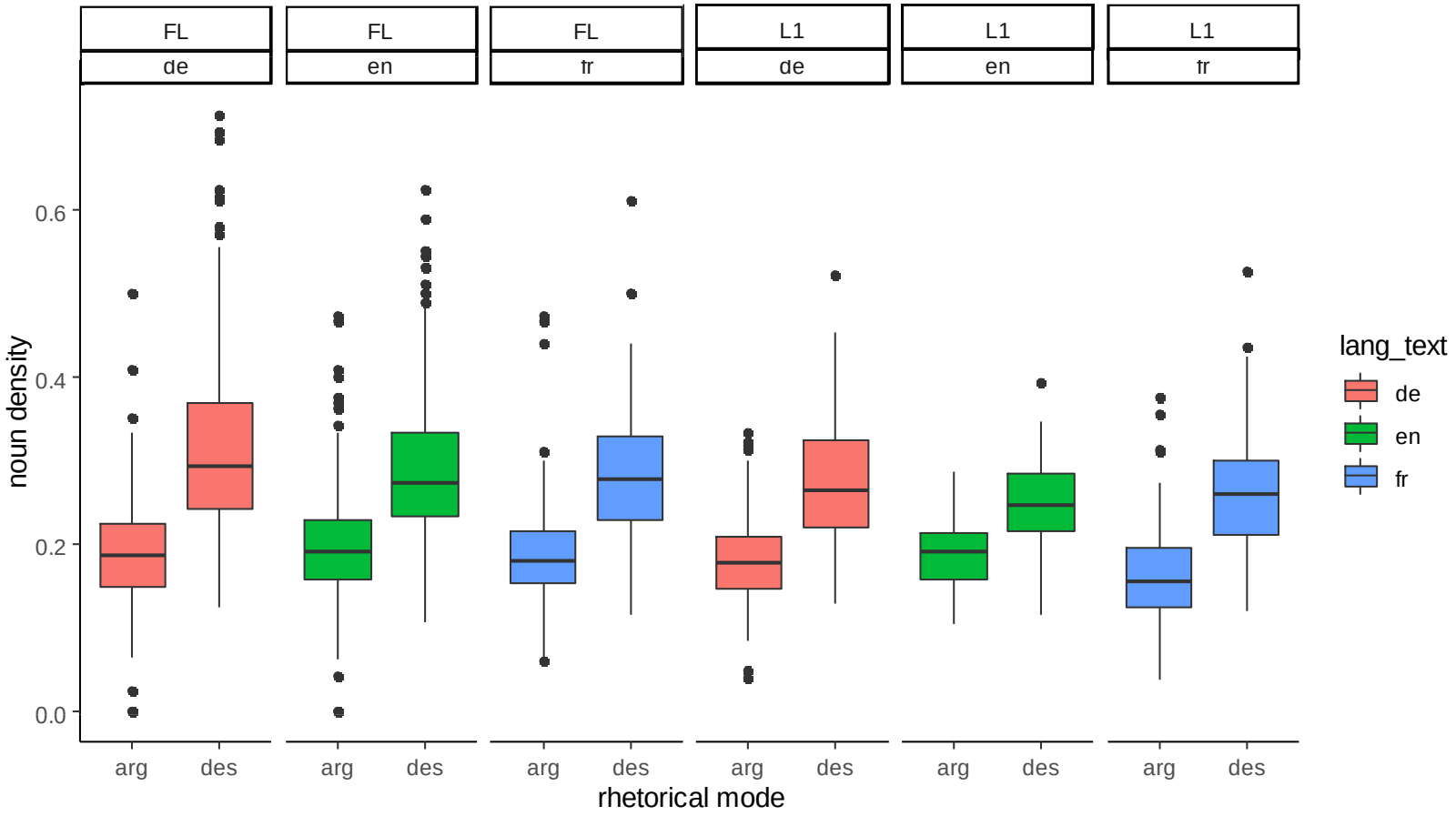
Lexical Richness in Read 2000, *Lexical Complexity* in Bulté & Housen 2012

- Lexical *density* *(Ure 1971)*
- Lexical *diversity* *(Treffers-Daller et al., 2018; McCarthy & Jarvis 2010)*
- Lexical *sophistication* *(Kyle & Crossley 2015)*

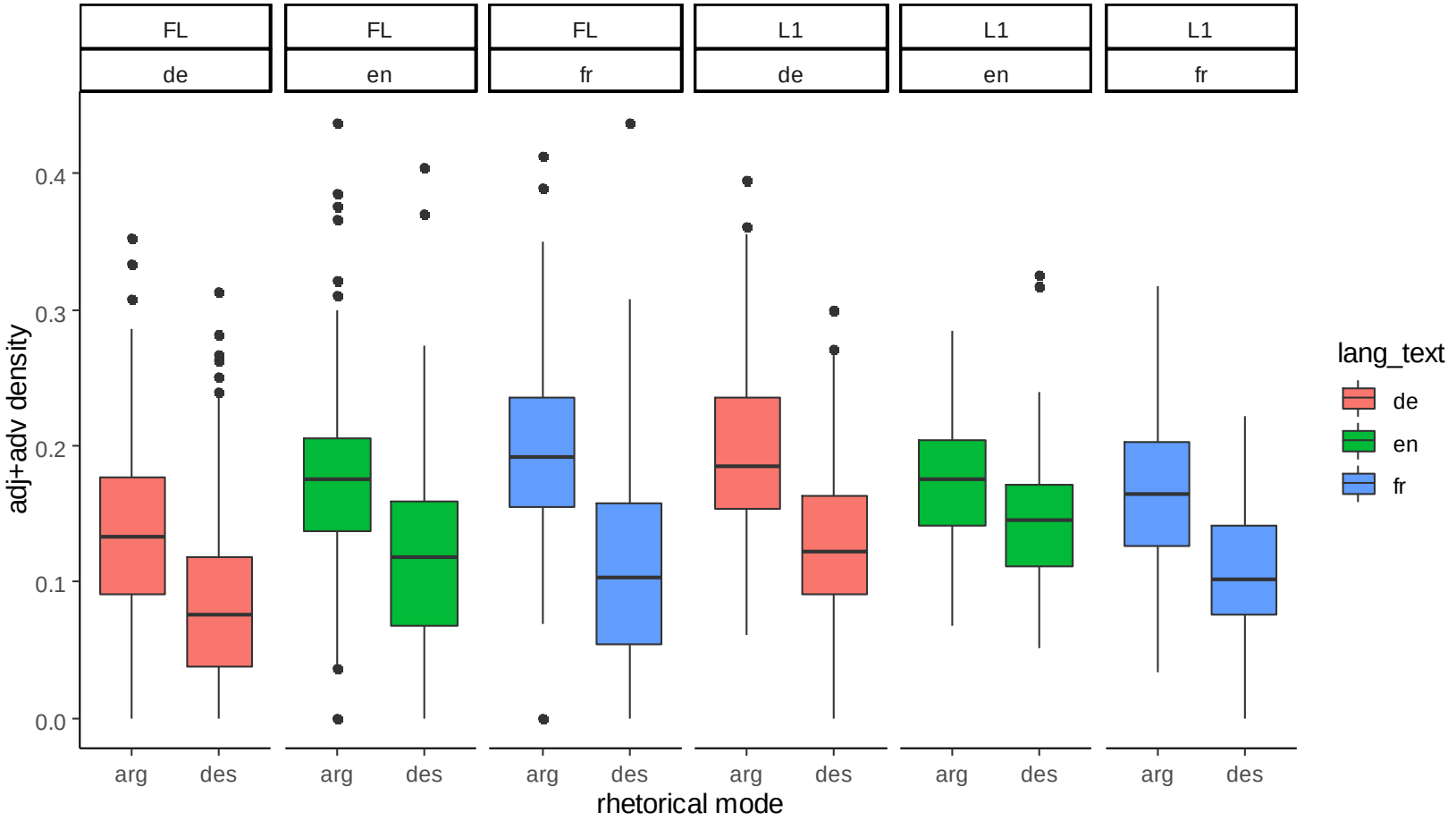
- *Lexical errors*

Analyses using the koRpus package (Michalke 2017) in R (R Core Team 2022)

Noun density by rhetorical type (argumentative vs. descriptive)



Adj + adv density by rhetorical type (argumentative vs. descriptive)



Lexical features

Lexical Richness in Read 2000, cf. *Lexical Complexity* in Bulté & Housen 2012

- Lexical *density* *(Ure 1971)*
  (ratio of content words, i.e., ***nouns***, verbs, adjectives, adverbs)
- Lexical *diversity* *(Treffers-Daller et al., 2018; McCarthy & Jarvis 2010)*
  (Guiraud, HD-D)
- Lexical *sophistication* *(Kyle & Crossley 2015)*
  (frequency based: ratio of lemmas beyond the top 1000 frequent lemmas; reference corpora: CoCa, Lexique, DeReKo)

- *Lexical errors*

Analyses using the koRpus package (Michalke 2017) in R (R Core Team 2022)

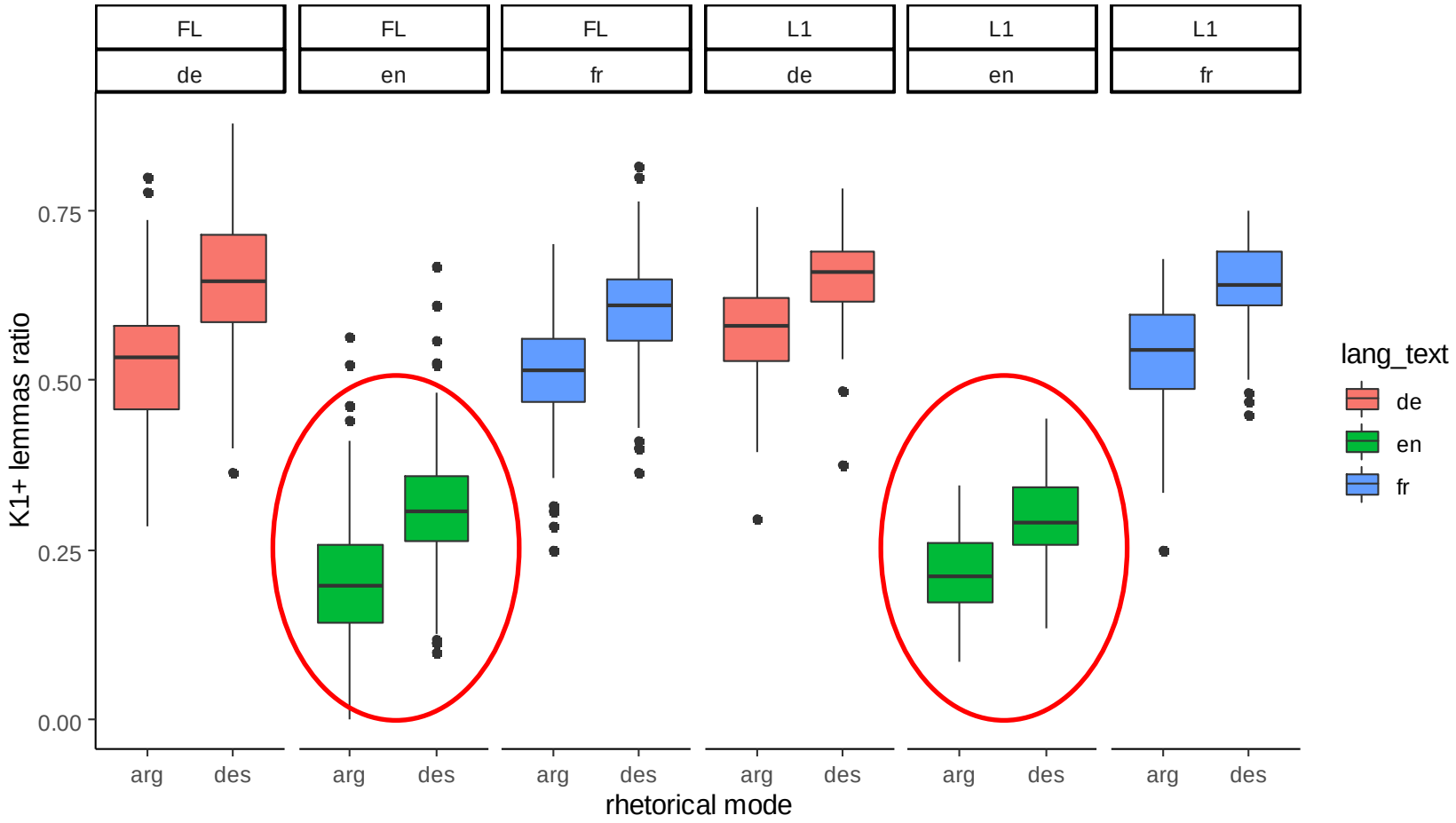
Rhetorical mode (descriptive vs. argumentative)

		density (noun ratio)	diversity (Guiraud/HDD)	sophistication (K1+ ratio)
FL	de	d > a	d > a	d > a
	en			
	fr			
L1	de			
	en			
	fr			

→ Across contexts and languages:
descriptive = more dense & more sophisticated

(e.g., Berman & Nir-Sagiv 2007, Olinghouse & Wilson 2012 for L1; Alexopoulou et al. 2017, Qin & Uccelli 2016, Bi 2020 for L2; Weiss et al. 2022, Yoon & Polio 2017 for both L1 and L2)

Lexical sophistication by rhetorical type



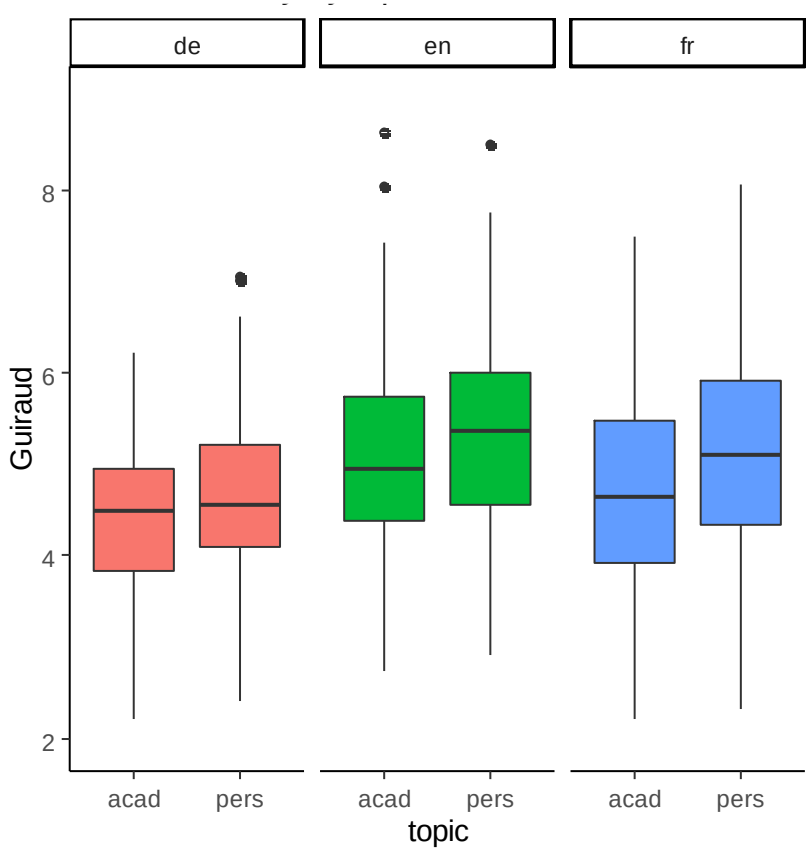
Topic / familiarity (personal vs. academic)

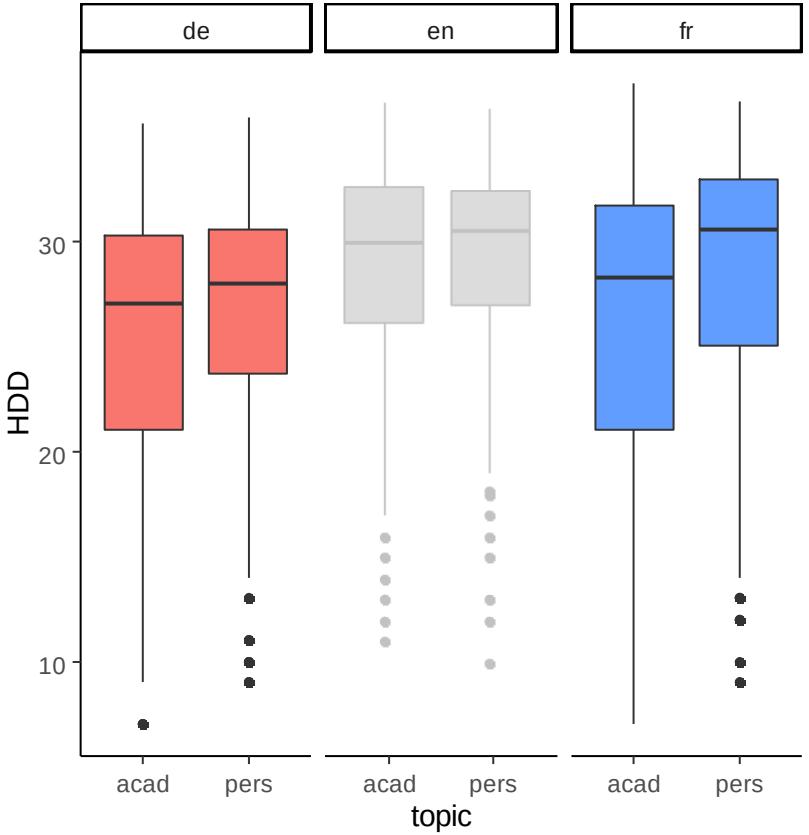
		density (noun ratio)	diversity (Guiraud)	sophistication (K1+ ratio)
FL	de	p < a	p > a	p > a
	en			
	fr			
L1	de	p < a		
	en			
	fr			

→ Foreign languages only:
personal = slightly less dense, more diverse & elaborate

(Abdi Tabari et al. 2021, Qin & Uccelli 2020, Yoon 2017, Yu 2010 for L2)

Lexical diversity by topic familiarity (FL)





Structure (more vs. less)

		density (noun ratio)	diversity (Guiraud)	sophistication (K1+ ratio)
FL	de	m > l		m > l
	en		m < l	
	fr			
L1	de			
	en			
	fr			m > l

→ Across contexts and languages:
more structured = more dense & sophisticated

(«provision of ideas & macro-structure» in Ong 2014, Révész et al. 2017, Yoon 2021 for L2)

Example: SWI02 vs. SWI07 (DaF)

SWI02, Es148: describe a graph (descriptive, academic, more structure)

Ein katz essen vier Hunden Mause in einen Monate. In die Schweiz, Es gibt 1,5 millionen Katze und ein Katze cost 1000 für zwölf monat.

noun ratio 36% Guiraud 3.8 K1+ ratio 39%

SWI07, Es148: discuss later school hours (argumentative, personal, less structure)

Ja, Das ist sehr gut weil, wir habe + schlafen weil, Wir sind sehr müde in die schule. Auch, ich habe 30 minuten minus die Schule normal. Aber, Wir haben finir plus tard in 5.30

noun ratio 9% Guiraud 4.2 K1+ ratio 46%

Summary

- Effects of task characteristics on lexical features
  - Rhetorical type → density & sophistication
  - Topic → density, diversity & sophistication, FL only
  - Structure → density & sophistication
- More effects in FL than L1
- Most differences in German, least in French

Limitations & further directions

- Limitations
  - Short texts → reliability of measures?
  - Task variables: fuzzy boundaries, combinations
  - Not necessarily reflective of text quality
    (cf. Studer & Hicks 2022)

- Outlook
  - Further CAF annotation & analyses (ratings!)
  - Transfer back to school context: SWIKOweb, DDL

Conclusion

- Complex effects of task variables on lexical features in adolescents' writing (SWIKO)
  → relevant for education & research

- Many similar across all three languages, but some observations language-specific
  → widen English focus

References

Abdi Tabari, M., Bui, G., & Wang, Y. (2021). The effects of topic familiarity on emotionality and linguistic complexity in EAP writing. Language Teaching Research, 13621688211033564.

Alexopoulou, T., Michel, M., Murakami, A., & Meurers, D. (2017). Task Effects on Linguistic Complexity and Accuracy: A Large-Scale Learner Corpus Analysis Employing Natural Language Processing Techniques. Language Learning, 67(S1), 180–208.

Berman, R. A. (2008). The psycholinguistics of developing text construction. Journal of Child Language, 35(4), 735–771.

Berman, R. A., & Nir-Sagiv, B. (2007). Comparing narrative and expository text construction across adolescence: A developmental paradox. Discourse Processes, 43, 79–120.

Bertschy, I., Cuenat, M. E., & Stotz, D. (2015). Lehrplan Französisch und Englisch. Passepartout - Fremdsprachen an der Volksschule. https://be.lehrplan.ch/passepartout/Lehrplan_Passepartout.pdf

Bi, P. (2020). Revisiting genre effects on linguistic features of L2 writing: A usage-based perspective. International Journal of Applied Linguistics, 30(3), 429–444.

Bulté, B., & Housen, A. (2012). Defining and operationalising L2 complexity. In A. Housen, F. Kuiken, & I. Vedder (Hrsg.), Dimensions of L2 Performance and Proficiency: Complexity, Accuracy and Fluency in SLA (S. 21–46). John Benjamins Publishing.

Council of Europe. (2001). Common European framework of reference for languages: Learning, teaching, assessment. Cambridge University Press.

Council of Europe. (2020). Common European framework of reference for languages: Learning, teaching, assessment: companion volume. Council of Europe Publishing.

Eckstein, G., & Ferris, D. (2018). Comparing L1 and L2 Texts and Writers in First-Year Composition. TESOL Quarterly, 52(1), 137-162.

EDK. (2011). Grundkompetenzen für die Fremdsprachen: Nationale Bildungsstandards.

EDK. (2017). Empfehlungen zum Fremdsprachenunterricht (Landessprachen und Englisch) in der obligatorischen Schule.

Ellis, R., Skehan, P., Li, S., Shintani, N., & Lambert, C. (2020). Task-based language teaching: Theory and practice. Cambridge University Press.

Housen, A., De Clercq, B., Kuiken, F., & Vedder, I. (2019). Multiple approaches to complexity in second language research. Second Language Research, 35(1), 3–21.

Housen, A., Kuiken, F., & Vedder, I. (Hrsg.). (2012). Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA. John Benjamins.

Karges, K., Studer, T., & Hicks, N. S. (2022). Lernersprache, Aufgabe und Modalität: Beobachtungen zu Texten aus dem Schweizer Lernerkorpus SWIKO. Zeitschrift für germanistische Linguistik, 50(1), 104–130.

Karges, K., Studer, T., & Wiedenkeller, E. (2019). On the way to a new multilingual learner corpus of foreign language learning in school: Observations about task variation. In A. Abel, A. Glaznieks, V. Lyding, & L. Nicolas (eds.), Widening the Scope of Learner Corpus Research. Selected papers from the 4th Learner Corpus Research Conference (S. 137-165). Presses universitaires de Louvain.

Karges, K., Studer, T., & Wiedenkeller, E. (2020). Textmerkmale als Indikatoren von Schreibkompetenz. Bulletin suisse de linguistique appliquée, No spécial Printemps 2020, 117–140.

Kyle, K., & Crossley, S. A. (2015). Automatically Assessing Lexical Sophistication: Indices, Tools, Findings, and Application. TESOL Quarterly, 49(4), 757–786.

Lenz, P., & Wiedenkeller, E. (2019). Kurzbericht zum Projekt «Ergebnisbezogene Evaluation des Französischunterrichts in der 6. Klasse (HarmoS 8) in den sechs Passepartout-Kantonen». Institut für Mehrsprachigkeit.

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

Michalke, M. (2017). koRpus: An R Package for Text Analysis (0.10-2). reaktanz.de. http://reaktanz.de/?c=hacking&s=koRpus

Norris, J. M., & Ortega, L. (2009). Towards an Organic Approach to Investigating CAF in Instructed SLA: The Case of Complexity. Applied Linguistics, 30(4), 555–578.

Olinghouse, N. G., & Wilson, J. (2013). The relationship between vocabulary and writing quality in three genres. Reading and Writing, 26(1), 45–65.

Ong, J. (2014). How do Planning Time and Task Conditions Affect Metacognitive Processes of L2 Writers? Journal of Second Language Writing, 23, 17–30.

Peyer, E., Andexlinger, M., Kofler, K., & Lenz, P. (2016). Projekt Fremdsprachenevaluation BKZ: Schlussbericht zu den Sprachkompetenztests. Institut für Mehrsprachigkeit.

Qin, W., & Uccelli, P. (2016). Same language, different functions: A cross-genre analysis of Chinese EFL learners' writing performance. Journal of Second Language Writing, 33, 3–17.

Qin, W., & Uccelli, P. (2020). Beyond linguistic complexity: Assessing register flexibility in EFL writing across contexts. Assessing Writing, 45, 100465.

R Core Team. (2022). R: A Language and Environment for Statistical Computing (4.0.2). R Foundation for Statistical Computing. http://www.R-project.org

Read, J. (2000). Assessing Vocabulary. Cambridge University Press.

Révész, A., Kourtali, N.-E., & Mazgutova, D. (2017). Effects of Task Complexity on L2 Writing Behaviors and Linguistic Complexity. Language Learning, 67(1), 208–241.

Schmid, H. (2013). TreeTagger—A Language Independent Part-of-speech Tagger (3.2). http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

Schmidt, T., & Wörner, K. (2009). EXMARaLDA – creating, analysing and sharing spoken language corpora for pragmatic research. Pragmatics, 19(4), 565–582.

Studer, T., & Hicks, N. S. (2022). The interplay of task variables, linguistic measures, and human ratings: Insights from the multilingual learner corpus SWIKO. European Second Language Acquisition Conference, Fribourg.

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.

Ure, J. (1971). Lexical density and register differentiation. In G. E. Perren & J. L. M. Trim (Hrsg.), Applications of linguistics (S. 443–452). Cambridge University Press.

Vyatkina, N. (2012). The development of second language writing complexity in groups and individuals: A longitudinal learner corpus study. The Modern Language Journal, 96(4), 576–598.

Weiss, Z., Hicks, N. S., Meurers, D., & Studer, T. (2022). Using linguistic complexity to probe into genre differences? Insights from the multilingual SWIKO learner corpus. Learner Corpus Research Conference, Padua.

Wolfe-Quintero, K., Inagaki, S., & Kim, H.-Y. (1998). Second language development in writing: Measures of fluency, accuracy, & complexity. University of Hawaii Press.

Yoon, H.-J. (2017). Linguistic complexity in L2 writing revisited: Issues of topic, proficiency, and construct multidimensionality. System, 66, 130–141.

Yoon, H.-J. (2021). Challenging the connection between task perceptions and language use in L2 writing: Genre, cognitive task complexity, and linguistic complexity. Journal of Second Language Writing, 54, 100857.

Yoon, H.-J., & Polio, C. (2017). The Linguistic Development of Students of English as a Second Language in Two Written Genres. TESOL Quarterly, 51(2), 275–301.

Yu, G. (2010). Lexical Diversity in Writing and Speaking Task Performances. Applied Linguistics, 31(2), 236–259.

Questions & comments?

Thank you for your attention!

Research Centre on Multilingualism
Institute of Multilingualism
Murtengasse 24
CH-1700 Freiburg

Mail: nina.hicks@unifr.ch

Web: www.institut-plurilinguisme.ch/en