

Issues in Applying Phraseological Complexity Indices to L2 French

Nate Vandeweerd

Broadening the Scope of L2 Complexity Research
Vrije Universiteit Brussel, Nov. 14, 2019

Phraseological complexity in L2 English (Paquot, 2019)

Data: Vespa Corpus

- L2 English
- 98 linguistics essays
- holistic CEFR grades

Method:

ADJ+NOUN

ADV+VERB

VERB+DOBJ

- Phraseological complexity measures
 - diversity: RTTR
 - sophistication: Mean MI score, Proportion ACL
- Lexical measures → Lexical Complexity Analyzer (Lu, 2012)
- Syntactic measures → L2 Syntactic Complexity Analyzer (Lu, 2010)

CEFR Level
Predictions

Phraseological complexity in L2 English (Paquot, 2019)

	Diversity			Sophistication (Mean PMI)		
	advmod	amod	dobj	advmod	amod	dobj
L2 English (Paquot, 2019)	n.s.	n.s.	n.s.	* (B2->C1) (B2->C2)	* (B2->C2)	* (B2->C2) (C1->C2)
Regression Model - Mean PMI advmod (14%) - Mean PMI amod (11%) - Mean PMI dobj (15%)				Sophistication (ACL)		
				n.s.	n.s.	n.s.

Phraseological complexity in L2 English (Paquot, 2018)

	Sophistication (Mean PMI)			Sophistication (Coll. Bands)		
	advmod	amod	dobj	advmod	amod	dobj
L2 English (Paquot, 2018)	*	*	*	n.s.	* (non-coll, low, med)	* (med)

Mixed effects regression model

- Fixed effects: Mean PMI dobj, Mean PMI amod (25% variance explained)
- Random effects: learner
- $R^2 = 0.3158$

Complexity in L2 French

- Significant increase in **lexical sophistication** (LFP) from low-advanced to high-advanced) (Lindqvist, Bardel Gudmundson, 2011)
- Significant increase in **lexical diversity, sophistication and density** as well as **collocational density** (num. V+(P)+N collocations per text) with proficiency (De Clercq, 2015)
- Significant increase in **syntactic complexity** from A->B level (MLT, MLSC, MLC) (Gyllstad et al., 2014; Bernardini & Granfeldt, 2019)
- Significant increase in **syntactic elaboration** (length), **syntactic diversity** with proficiency (De Clercq & Housen, 2017)
- Increase in **morphological diversity** with proficiency (De Clercq & Housen, 2019)

L2 French Phraseology

- increase in **quantity of formulaic language** (Lexical FS) from A2->B1(*), B1->B2(n.s.), B2->C2(*) (Forsberg & Bartning, 2010)
- Advanced long-term residency L2 French speakers have less **diversity of multi-word units** compared to NS (Erman et al. 2015)
- Increase in **productive collocation knowledge** of VERB+DOBJ collocations from B2->C1 levels (Forsberg Lundell et al., 2018)

The Gap

Phraseological **complexity** (diversity and sophistication) in L2 French.

Current Study

Partial replication of Paquot (2018; 2019)
for L2 French

Research Questions

RQ1. How does phraseological complexity compare in written L2 French at different proficiency levels?

RQ2. To what extent does phraseological complexity relate to other aspects of linguistic complexity as tapped by the current repertoire of measures of lexical, syntactic and morphological complexity in L2 written French?

Methodology

Data: Leerdercorpus Frans

- L2 French (L1 Dutch)
- 168 argumentative essays
- 6 topics: youth delinquency, euthanasia, foreigners, freedom, language, nuclear proliferation, state reform
- Texts assessed by two trained CEFR Raters ($\kappa = 0.87$, $p < 0.001$)
 - B2 (n=22)
 - C1 (n=111)
 - C2 (n=35)

*NB: contested texts removed from further analysis

Methodology

Lexical Measures:

- Lexical Diversity Measures (koRpus package, Michalke 2019 and in-house R scripts)
 - Root type-token ratio
 - Type-token ratio (lexical words, verbs, nouns, adj, adv)
 - Corrected type-token ratio (overall, verbs, nouns, adj, adv)
 - Lexical word variation (verbs, nouns, modifiers)
- Lexical Sophistication Measures
 - Prop. lemmas (lexical, verbs, nouns, adj, adv) absent from top 2k SUBTLEX-FR (Lexique3, New et. al, 2001)
 - Prop. of nouns, adj, verbs on the 'lexique transdisciplinaire' list (Jacques & Tutin, 2018)

Methodology

Syntactic Measures:

(calculated with in-house R scripts)

- Average sentence length
- Conjunction ratio: num. conjunctions/num. tokens

Morphological Diversity Measures:

(not present in Paquot 2018, 2019)

- Mean number of derivational morphemes/text (Lexique3, New et. al, 2001)
- Inflectional Diversity: token RTTR - lemma RTTR (in-house R scripts)
- Morphological Complexity Index of verbs, nouns, adj (in-house R scripts)

Methodology

Phraseological Measures:

- Phraseological Diversity Measures
 - RTTR (advmod_v, amod, dobj)
- Phraseological Sophistication Measures (Hubert Naets, CENTAL)
 - > Reference corpus: FRCOW16 (Schäfer & Bildhauer, 2012; Schäfer, 2015)
 - Mean PMI (advmod_v, amod, dobj)
 - For each dependency type
 - prop. non-collocational (PMI < 3)
 - prop. low collocational (PMI 3 - 5)
 - prop. med. collocational (PMI 5 - 7)
 - prop. high collocational (PMI > 7)

Methodology

Analysis:

RQ1. How does phraseological complexity compare in written L2 French at different proficiency levels?

CEFR Level ~ Phraseological complexity measures (Kruskal-Wallis rank sum test)

- phraseological sophistication (mean PMI, proportion in collocation bands)
- phraseological diversity (RTTR)

→ Pairwise comparisons (Wilcoxon rank sum test)

RQ2. To what extent does phraseological complexity relate to other aspects of linguistic complexity as tapped by the current repertoire of measures of lexical, syntactic and morphological complexity in L2 written French?

CEFR Level ~ Complexity measures (party::cforest)

- Model 1: lexical, syntactic, morphological measures
- Model 2: lexical, syntactic, morphological and phraseological measures
- Model 3: phraseological measures

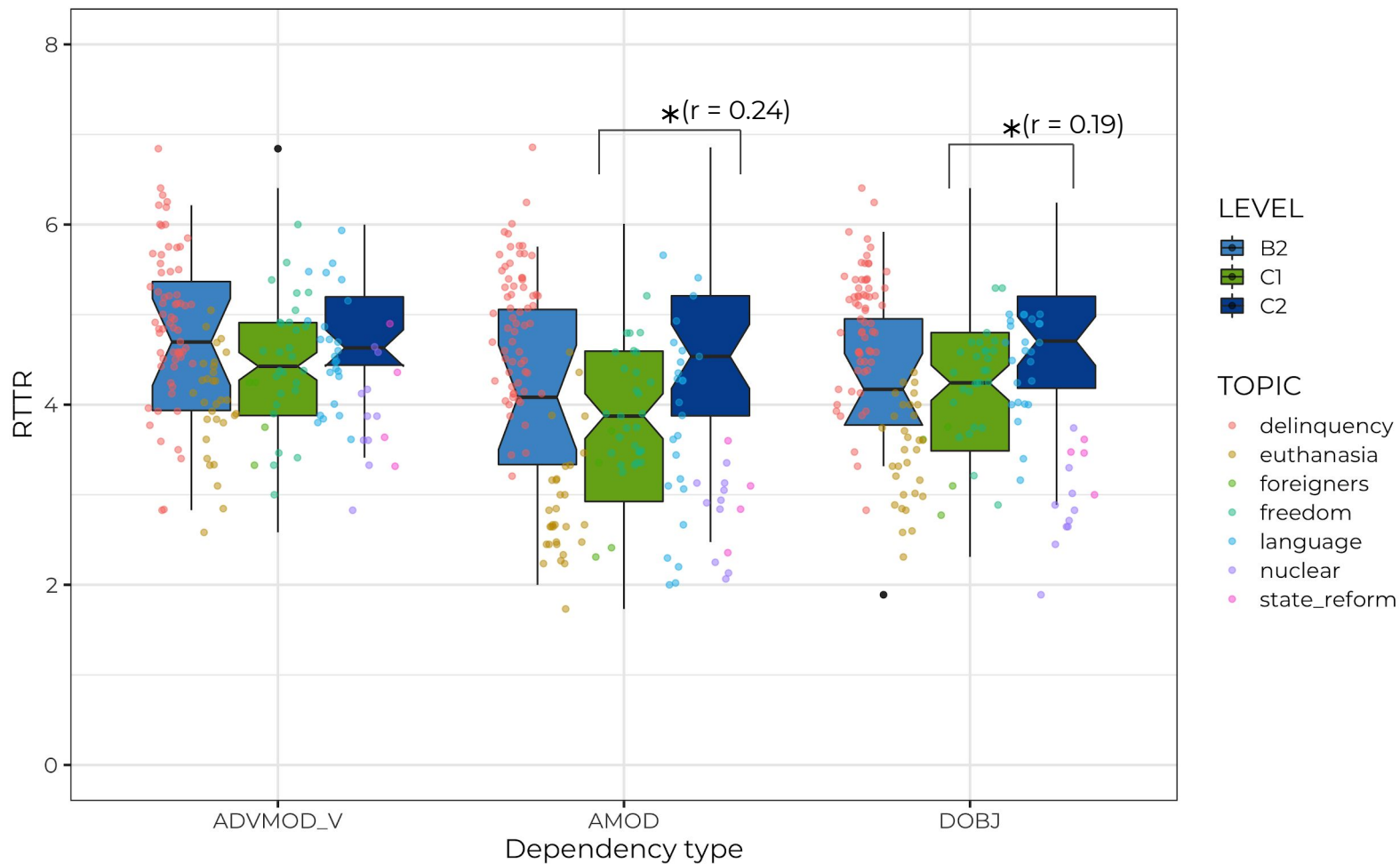
→ Difference in prediction accuracy between models (two-tailed binomial test)

RQ1:

Proficiency ~ Phraseological measures

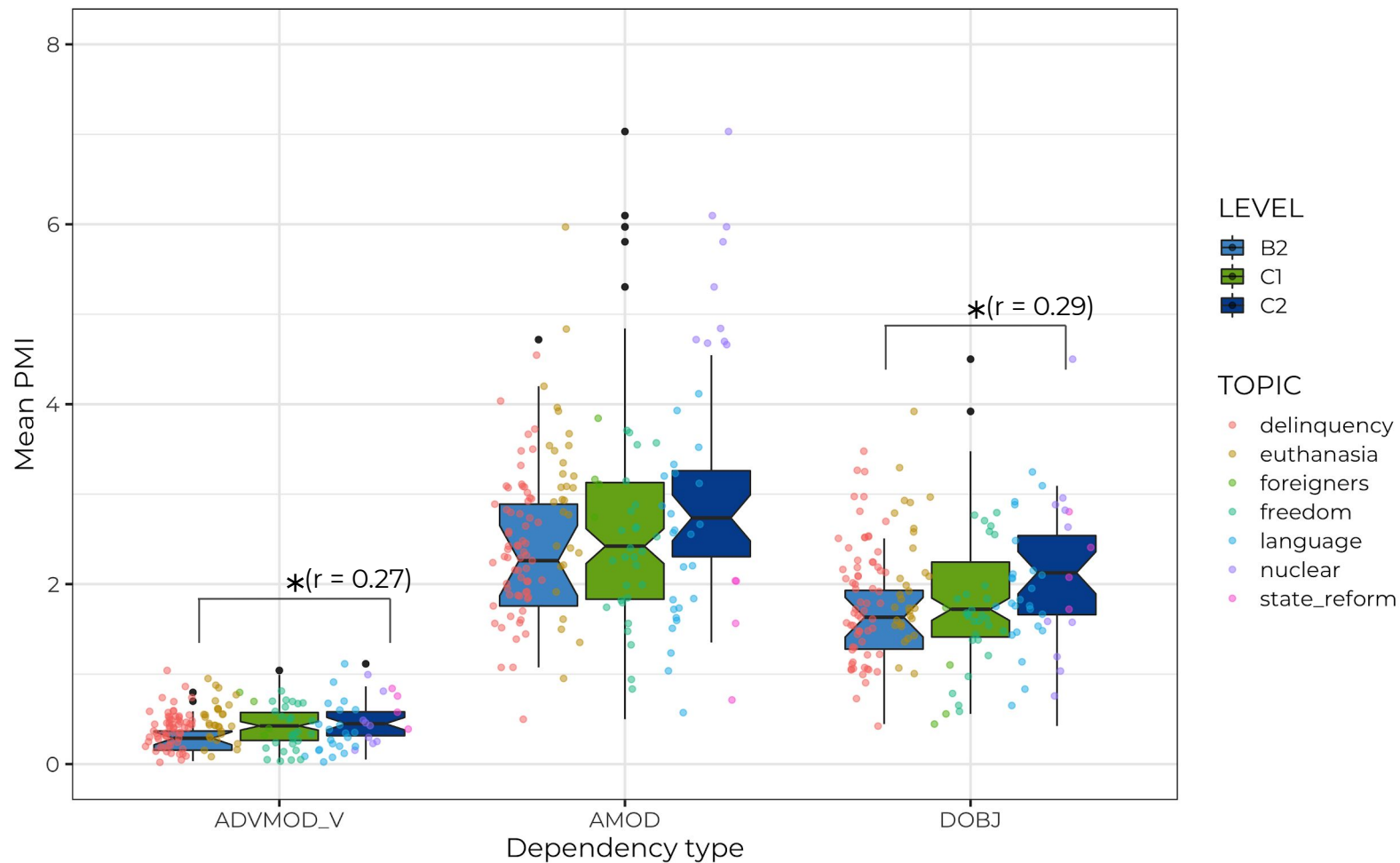
RTTR of dependency relations across CEFR levels

*p<0.05



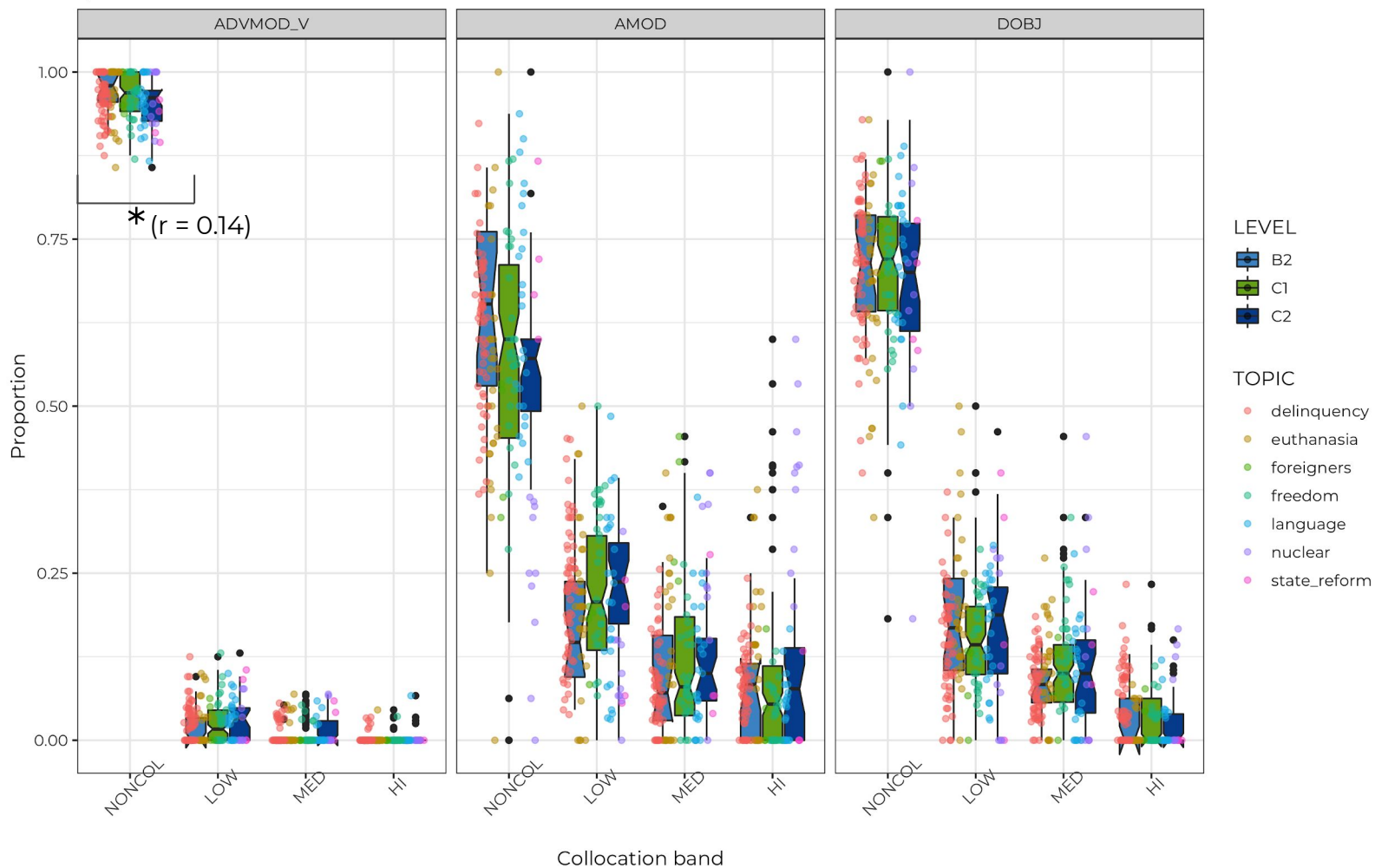
Mean PMI of dependency relations across CEFR levels

* $p < 0.05$



Dependency relation collocation bands across CEFR levels

*p<0.05



RQ1. How does phraseological complexity compare in written L2 French at different proficiency levels?

	Diversity			Sophistication (Mean PMI)			Sophistication (Coll. bands)		
	advmod	amod	dobj	advmod	amod	dobj	advmod	amod	dobj
L2 French (current study)	n.s.	* (C1->C2)	* (C1->C2)	* (B2->C2)	n.s.	* (B2->C2)	* (non-col: B2->C2)	n.s.	n.s.

	Diversity			Sophistication (Mean PMI)			Sophistication (Coll. bands)		
	advmod	amod	dobj	advmod	amod	dobj	advmod	amod	dobj
L2 French (current study)	n.s.	* (C1->C2)	* (C1->C2)	* (B2->C2)	n.s.	* (B2->C2)	* (non-col: B2->C2)	n.s.	n.s.
L2 English (Paquot, 2019)	n.s.	n.s.	n.s.	* (B2->C1) (B2->C2)	* (B2->C2)	* (B2->C2) (C1->C2)			
L2 English (Paquot, 2018)				*	*	*	n.s.	* (non-col, low, med)	* (med)

RQ2:

Phraseological vs. other measures

Model Comparison

Recall (precision)

	Overall	B2	C1	C2
Model 1 (lex, syn, morph)	80%	27% (100%)	100% (42%)	51% (100%)

Model Comparison

Recall (precision)

	Overall	B2	C1	C2
Model 1 (lex, syn, morph)	80%	27% (100%)	100% (42%)	51% (100%)
Model 2 (lex, syn, morph, phraseological)	82%	31% (100%)	100% (47%)	57% (100%)

Model Comparison

Recall (precision)

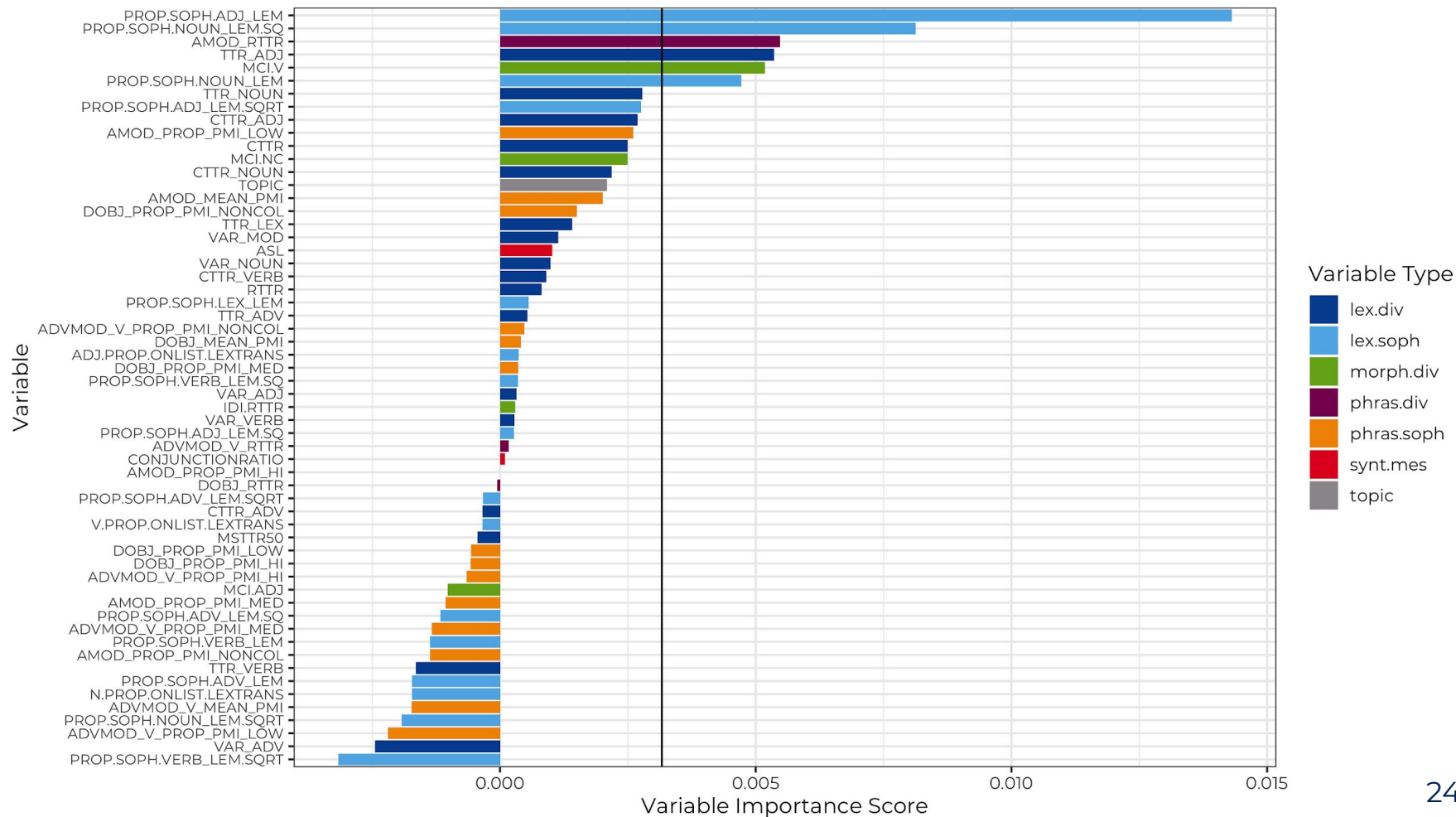
	Overall	B2	C1	C2
Model 1 (lex, syn, morph)	80%	27% (100%)	100% (42%)	51% (100%)
Model 2 (lex, syn, morph, phraseological)	82%	31% (100%)	100% (47%)	57% (100%)
Model 3 (phraseological)	78%	5% (100%)	100% (35%)	54% (100%)

Caveats

- based on small samples B2 (n=22), C1 (n=111), C2 (n=35)
- categorical predictions were not significantly different between models

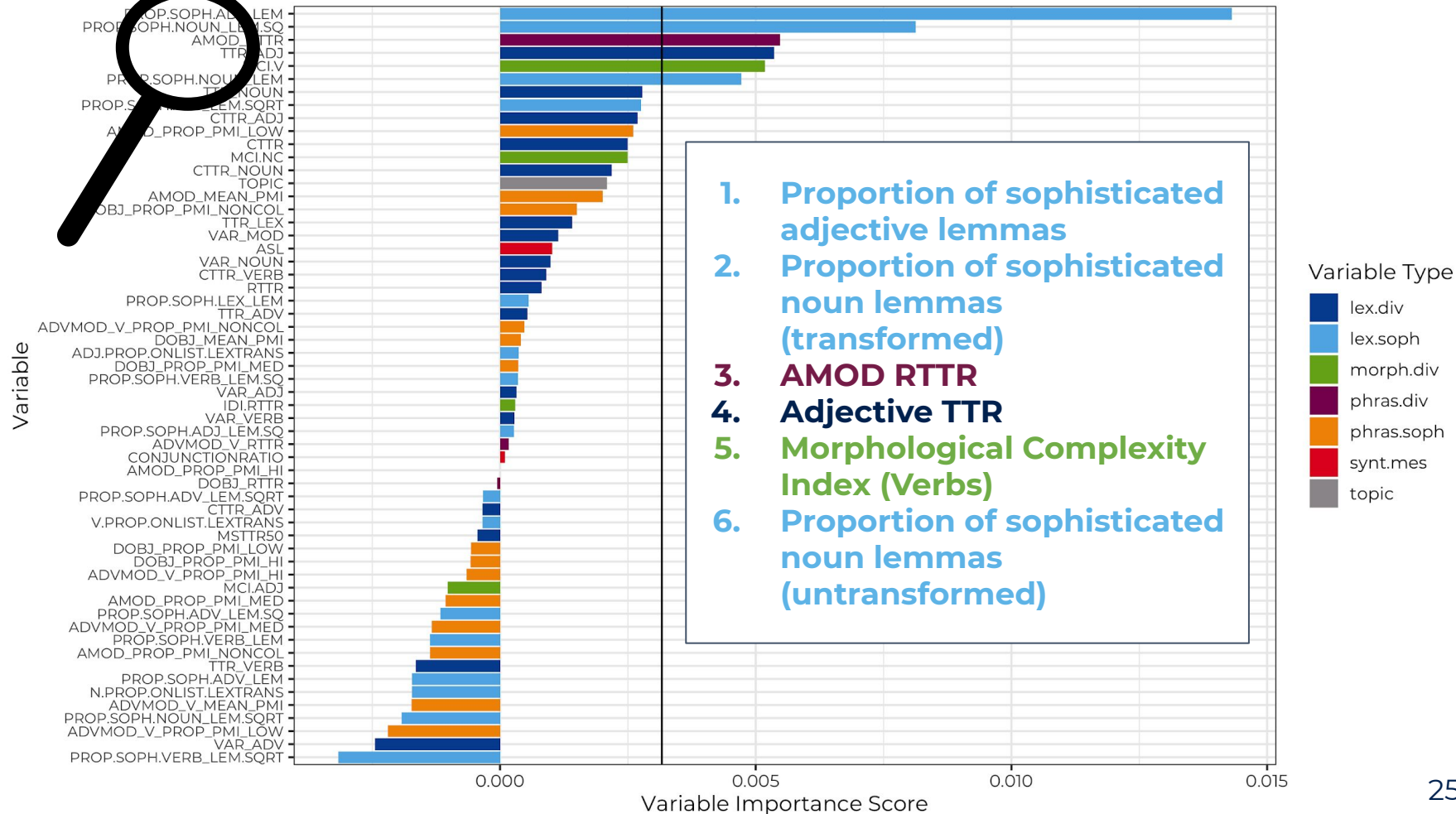
Variable Importance Scores

(vertical line indicates the absolute value of the minimum VIS)



Variable Importance Scores

(vertical line indicates the absolute value of the minimum VIS)



RQ2. To what extent does phraseological complexity relate to other aspects of linguistic complexity as tapped by the current repertoire of measures of lexical, syntactic and morphological complexity in L2 written French?

Most important variables:

- | | |
|---|---------------------------|
| 1. proportion of sophisticated adjectives | (lexical sophistication) |
| 2. proportion of sophisticated nouns | (lexical sophistication) |
| 3. diversity of amod dependencies | (phras. diversity) |
| 4. diversity of adjectives | (lexical diversity) |
| 5. diversity of verbal inflections | (morph. diversity) |

Two Example Texts

C4023-02.txt

Une **langue** sans **professeurs**, c'est comme une **justice** sans **juges**, un **contrat** sans **notaire**, une **amitié** sans des **amis**? L'**homme**, **est-il** **capable** de **développer** une **maîtrise** de la **langue parfaite** sans l'aide des profs? Dans une **première partie**, nous **jetterons** un **regard** sur l'**acquisition naturelle** d'une **langue**. Ensuite, nous **traiterons** l'**importance** des **professeurs** pour l'**orthographe**. Nous **terminerons** avec une **appréciation** de la **thèse** de Claudel...

Level: B2

Proportion of Sophisticated **Adjective** Lemmas: 8%

Proportion of Sophisticated **Noun** Lemmas: 57%

AMOD RTTR: 2

Morphological Complexity Index (**Verbs**): 5.10

Two Example Texts

C4020-02.txt

« Très faible croissance » ou « croissance négative », voilà deux **euphémismes** par lesquels les hommes politiques décrivent la crise économique actuelle. Et qu'est-ce que vous **pensez** de la **périphrase** « épuration ethnique » au lieu de **génocide** ? Le monde politique grouille des périphrases atténuants qui **modèrent** l'opinion publique. Et pas seulement le monde politique ! Les **hommes savent** très bien **manipuler** les **paroles** en leur **faveur**. Comment nous **mettre** en garde contre les **mystifications** ? Faut-il alors se **méfier** de la **parole**, la considérer comme un outil perfide ?...

Level: C2

Proportion of Sophisticated **Adjective** Lemmas: 70%

Proportion of Sophisticated **Noun** Lemmas: 65%

AMOD RTTR: 4.53

Morphological Complexity Index (**Verbs**): 3.68

Why does phraseological sophistication seem to play less of a role in this data?

- topic effect (Paquot, Naets Gries, forthcoming)
- accounted for by lexical sophistication
- stylistic preferences of French (Vinay & Darbelnet, 1995, p. 101)
 - French is a noun-centered language
 - less rich in adjectives and adverbs compared to English

Adjectival phrases:

a hopeless undertaking → une entreprise *sans espoir*

Adverbial phrases:

answered angrily → répondu *avec colère*

Prepositional phrases

C4020-02.txt

« Très faible croissance » ou « croissance négative », voilà deux euphémismes par lesquels les hommes politiques décrivent la crise économique actuelle. Et qu'est-ce que vous pensez de la périphrase « épuration ethnique » au lieu de génocide ? Le monde politique grouille des périphrases atténuants qui modèrent l'opinion publique. Et pas seulement le monde politique ! Les hommes savent très bien **manipuler** les paroles **en leur faveur**. Comment nous **mettre en garde** contre les mystifications ? Faut-il alors **se méfier de la parole**, la considérer comme un outil perfide ?

...

c'est **sans doute** un art de parler

Next Steps

- replication with a larger and more balanced dataset (exam corpus, in preparation)
- integrate more syntactic measures (may improve model performance at B2 level)
 - length of clause, t-unit, noun phrase, verb phrase
 - subordination ratio, coordination ratio
- explore other phraseological measures which may be more representative of French
 - including other types of dependencies (ex: prepositional modifiers)
 - expanding the scope beyond dependencies (ex: lexical bundles)
 - suggestions?

Q &

eh?

References

- Candito, M., Nivre, J., Denis, P., & Henestroza, A. E. (n.d.). Benchmarking of Statistical Dependency Parsers for French. In COLING'2010. Beijing, China.
- De Clercq, B., & Housen, A. (2017). A Cross-Linguistic Perspective on Syntactic Complexity in L2 Development: Syntactic Elaboration and Diversity. *Modern Language Journal*, 101(2), 315–334. <https://doi.org/10.1111/modl.12396>
- De Clercq, B., & Housen, A. (2019). The development of morphological complexity: A cross-linguistic study of L2 French and English, 35(1), 71–97.
- De Clercq, B. (2015). The development of lexical complexity in second language acquisition: A cross-linguistic study of L2 French and English. *EUROSLA Yearbook*, 15(2015), 69–94. <https://doi.org/10.1075/eurosla.15.03dec>
- Erman, B., Denke, A., Fant, L., & Forsberg Lundell, F. (2015). Nativelike expression in the speech of long-residency L2 users: A study of multiword structures in L2 English, French and Spanish. *International Journal of Applied Linguistics (United Kingdom)*, 25(2), 160–182. <https://doi.org/10.1111/ijal.12061>
- Forsberg, F., & Bartning, I. (2010). Can linguistic features discriminate between the communicative CEFR-levels? : A pilot study of written L2 French. In I. Bartning, M. Martin, & I. Vedder (Eds.), *Communicative proficiency and linguistic development : Intersections between SLA and language testing* (Vol. 1, pp. 133–157). European Second Language Association.
- Forsberg Lundell, F., Lindqvist, C., & Edmonds, A. (2018). Productive Collocation Knowledge at Advanced CEFR Levels: Evidence from the Development of a Test for Advanced L2 French. *Canadian Modern Language Review*, 74(4), 627–649. <https://doi.org/10.3138/cmlr.2017-0093>
- Gries, S. T. (2019). 15 Years of Collostructions: Some Long Overdue Additions/Corrections (To/of Actually All Sorts of Corpus-Linguistics Measures). *International Journal of Corpus Linguistics*, 24(3), 387–414.
- Gyllstad, H., Granfeldt, J., Bernardini, P., & Källkvist, M. (2014). Linguistic correlates to communicative proficiency levels of the CEFR: The case of syntactic complexity in written L2 English, L3 French and L4 Italian. *EUROSLA Yearbook*, 14, 1–30. <https://doi.org/10.1075/eurosla.14.01gyl>
- Jacques, M.-P., & Tutin, A. (2018). Les expressions polylexicales transdisciplinaires dans les articles de recherche en sciences humaines : retour d'expérience. In M.-P. Jacques & A. Tutin (Eds.), *Lexique transversal et formules discursives des sciences humaines* (pp. 75–89).
- Jarvis, S. (2013). Capturing the Diversity in Lexical Diversity. *Language Learning*, 63(SUPPL. 1), 87–106. <https://doi.org/10.1111/j.1467-9922.2012.00739.x>

References

- Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics*, 15(4), 474–496.
<https://doi.org/10.1075/ijcl.15.4.02lu>
- Lu, X. (2012). The Relationship of Lexical Richness to the Quality of ESL Learners' Oral Narratives. *The Modern Language Journal*, 96(2), 190–206.
- Lindqvist, C., Bardel, C., & Gudmundson, A. (2011). Lexical richness in the advanced learner's oral production of French and Italian L2. *IRAL - International Review of Applied Linguistics in Language Teaching*, 49(3), 221–240. <https://doi.org/10.1515/iral.2011.013>
- Michalke, M. (2019). koRpus: An R Package for Text Analysis (Version 0.12-1).
Available from <https://reaktanz.de/?c=hacking&s=koRpus>
- Paquot, M. (2018). Phraseological competence: a missing component in university entrance language tests? Insights from a study of EFL learners' use of statistical collocations. *Language Assessment Quarterly*, 15(1), 29–43.
- Paquot, M. (2019). The phraseological dimension in interlanguage complexity research. *Second Language Research* (special issue on linguistic complexity), 1–25.
- R Core Team. (2019). R: A Language and Environment for Statistical Computing. Vienna, Austria. Retrieved from <https://www.r-project.org/>
- Schäfer, R. (2015). Processing and querying large web corpora with the {COW14} architecture. In P. Bański, H. Biber, E. Breiteneder, M. Kupietz, H. Lungen, & A. Witt (Eds.), *Proceedings of Challenges in the Management of Large Corpora 3 (CMLC-3)*. Lancaster: IDS. Retrieved from <http://rolandschaefer.net/?p=749>
- Schäfer, R., & Bildhauer, F. (2012). Building Large Corpora from the Web Using a New Efficient Tool Chain. In N. C. (Conference Chair), K. Choukri, T. Declerck, M. U. Doğan, B. Maegaard, J. Mariani, ... S. Piperidis (Eds.), *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)* (pp. 486–493). Istanbul, Turkey: European Language Resources Association (ELRA). Retrieved from <http://rolandschaefer.net/?p=70>
- Schmid, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of International Conference on New Methods in Language Processing*. Manchester, UK.
- Vinay, J.-P., & Darbelnet, J. (1995). *Comparative Stylistics of French and English*. (J. Sager & M.-J. Hamel, Trans.). Amsterdam: John Benjamins Pub. Co.

Table 1: CEFR descriptor scales for linguistic competence

Linguistic competence	B1	B2	C1	C2
Grammatical accuracy	[B1] Uses reasonably accurately a repertoire of frequently used 'routines' and patterns associated with more predictable situations.	[B2] Shows a relatively high degree of grammatical control. Does not make mistakes which lead to misunderstanding.	Consistently maintains a high degree of grammatical accuracy; errors are rare and difficult to spot.	Maintains consistent grammatical control of complex language.
	[B1+] Communicates with reasonable accuracy in familiar contexts; generally good control though with noticeable mother tongue influence. Errors occur, but it is clear what he/ she is trying to express.	[B2+] Good grammatical control; occasional 'slips' or non-systematic errors and minor flaws in sentence structure may still occur, but they are rare.		
Vocabulary control	Shows good control of elementary vocabulary but major errors still occur when expressing more complex thoughts or handling unfamiliar topics and situations.	Lexical accuracy is generally high, though some confusion and incorrect word choice does occur without hindering communication.	Occasional minor slips, but no significant vocabulary errors.	Consistently correct and appropriate use of vocabulary.
Vocabulary range	Has sufficient vocabulary to express him/herself with some circumlocution on most topics pertinent to his/ her everyday life such as family, hobbies and interests, work, travel and current events.	Has a good range of vocabulary for matters connected to his/ her field and most general topics. Can vary formulation to avoid frequent repetition but lexical gaps can still cause circumlocution.	Has a good command of a broad lexical repertoire allowing gaps to be readily overcome with circumlocution, little obvious searching for expressions or avoidance strategies. Good command of idiomatic expressions and colloquialisms.	Has a good command of a very broad linguistic repertoire including idiomatic expressions and colloquialisms; shows awareness of connotative levels of meaning.
Orthographic control (punctuation and spelling)	Can produce continuous writing which is generally intelligible throughout. Spelling, punctuation and layout are accurate enough to be followed most of the time.	Can produce clearly intelligible continuous writing which follows standard layout and paragraphing conventions. Spelling and punctuation are reasonably accurate but may show signs of mother tongue influence.	Layout, paragraphing and punctuation are consistent and helpful. Spelling is accurate, apart from occasional typographic errors.	Writing is free of error.
Cohesion and coherence	Can link a series of shorter, discrete simple elements into a connected, linear sequence of points	[B2] Can use a limited number of cohesive devices to link his/ her utterances into clear, coherent discourse, though there may be some 'jumpiness' in a long contribution.	Can produce clear, smoothly flowing, well-structured texts, showing controlled use of organisational patterns, connectors and cohesive devices.	Can create coherent and cohesive text making full and appropriate use of a variety of organisational patterns and a wide range of cohesive devices.
		[B2+] Can use a variety of linking words efficiently to mark clearly the relationship between ideas.		