

Assessing the cross-linguistic validity of phraseological complexity measures as indices of L2 proficiency

Rachel Rubin

Vrije Universiteit Brussel | Université catholique de Louvain

Co-authors:

Magali Paquot, Université catholique de Louvain

Alex Housen, Vrije Universiteit Brussel

Linguistic complexity

“...complexity refers to a property or quality of a phenomenon or entity in terms of

- (1) the number and the nature of the discrete components that the entity consists of, and
- (2) the number and the nature of the relationships between the constituent components”

(Bulté & Housen, 2012 p. 22)

"the range of forms that surface in language production and the degree of sophistication of such forms"

(Ortega, 2003 p. 492)

Complexity in L2 Dutch

- Focus of L2 complexity research to date:
 - L2 English
 - syntax (and to a lesser extent lexis)
- Dutch is therefore an under-researched language in terms of L2 complexity research
- Most research to date focuses on complexity as an outcome variable in L2 Dutch
 - Effects of task complexity on L2 Dutch CAF (Michel, Kuiken, & Vedder, 2007 & 2012)

L2 Dutch complexity research

- **Kuiken & Vedder (2019):** Syntactic complexity across L2s & proficiency
 - No significant correlation between syntactic complexity and proficiency (A2-B1)
 - No significant difference between L1 & L2
- **Bulon et al. (2017):** Impact of CLIL on L2 English & Dutch proficiency
 - Gains English CLIL students > gains Dutch CLIL students
 - Target language input

Phraseological complexity

- “Thus, phraseological complexity is defined as the range of phraseological units that surface in language production and the degree of sophistication of such phraseological units” (Paquot, 2019)
- Operationalization:
 - Phraseological units: relational or grammatical co-occurrences
 - in terms of: diversity and sophistication

Phraseological complexity in Dutch

Example: high-frequency verb *maken* (to make)

1. Dat heeft niks te **maken** met mij → *That has nothing to do with me*
2. Morgen **maken** de studenten hun eindexamens → *The students take their final exams tomorrow*
3. Pap heeft mijn fiets **gemaakt** → *Dad fixed my bike*
4. **Maak** er een mooie dag van → *Have a nice day*
5. Geld **maakt** niet gelukkig → *Money doesn't make you happy*
6. Ik **maak** me zorgen over zijn gedrag → *I'm worried about his behavior*
7. Ze **maken** vaak ruzie met elkaar → *They fight with each other often*
8. We **maken** goed gebruik van onze tuin → *We make good use of our garden*
9. Dat **maakt** mij niet uit → *It doesn't matter to me*

Measuring phraseological complexity

Phraseological units under investigation: grammatical dependencies

(verb + noun) in a direct object relation

(adjective/noun + noun) in an adjectival modifier head relation

(adjective/adverb + adjective/adverb/verb) in an adverbial modifier head relation

Diversity	Sophistication
Root type-token ratios of grammatical dependencies [T/VN]	Pointwise mutual information (MI) means for three grammatical dependencies
	Proportion of grammatical dependencies in collocational bands (MI-based)
	Ratio of academic to total dependencies

Phraseological measures of sophistication were better suited to predicting proficiency in intermediate to advanced (B2-C2) EFL speakers than conventional syntactic and lexical measures.

Measuring phraseological complexity

Phraseological units under investigation: grammatical dependencies

(verb + noun) in a direct object relation

(adjective/noun + noun) in an adjectival modifier head relation

(adjective/adverb + adjective/adverb/verb) in an adverbial modifier head relation

MI scores of adverbial modifier dependencies in L2 English

$0 < MI < 1$: *clearly negative; clearly described; important enough; measure + typically; represent + directly; very theoretical*

$6 < MI$: *mutually exclusive; fiercely debated; scarcely tenable; evenly distributed; firmly rooted*

Collocational bands

Below threshold	Dependency appears fewer than 5 times in the reference corpus
Non-collocational	$MI < 3$
Collocational: low	$3 \leq MI < 5$
Collocational: medium	$5 \leq MI < 7$
Collocational: high	$MI \geq 7$

The Present Study

An investigation of phraseological complexity as an index of L2 Dutch writing proficiency

Research questions

- **RQ 1:** How well do phraseological complexity measures perform as indices of L2 Dutch proficiency?
- **RQ 2:** How do measures of phraseological complexity compare to more traditional measures of lexical and syntactic complexity as indices of L2 Dutch proficiency?
- **RQ 3:** How do the results here compare to the results obtained for L2 English in Paquot (2018, 2019)?

Dutch learner data

CNaVT Corpus

Certificaat Nederlands als Vreemde Taal
Centrum voor Taal en Onderwijs (CTO) –
KU Leuven

(Thanks to Steven Verheyen & Goedele
Vandommele)

L2 Dutch

Multi-L1

- English, French, German, Dutch

1,595 texts used here

5 task variables, 14 learner variables

- e.g. writing topic, exam time, L1
background, education level

Binary pass/fail variable for each exam

Proficiency

- CEFR level of the CNaVT exam

Proficiency	Speakers	Tokens
B1	462	100,014
B2	1,059	305,008
C1	74	22,667



Syntactic measures

- Length
 - Words per sentence
 - Words per clause
- Embedding
 - Number of finite clauses per sentence
 - Number of clauses per sentence
 - Multiple finite embeddings
 - Multiple embeddings
- Coordination
 - Coordinate main clauses
 - Coordinate subordinate clauses
- Developmental
 - D-level (developmental level)
 - Proportion of sentences with D-level above 4
- Specific structures
 - Density of nominalizations
 - Density of passive forms
 - Density of adverbial modifiers
 - Density of adjectival modifiers
- Dependency distance (in words)
 - verb and accompanying subject
 - verb and accompanying object
 - noun and accompanying article
 - verb and verb from verbal complement
 - verb and separable part of this verb
 - average dependency distance

Lexical measures

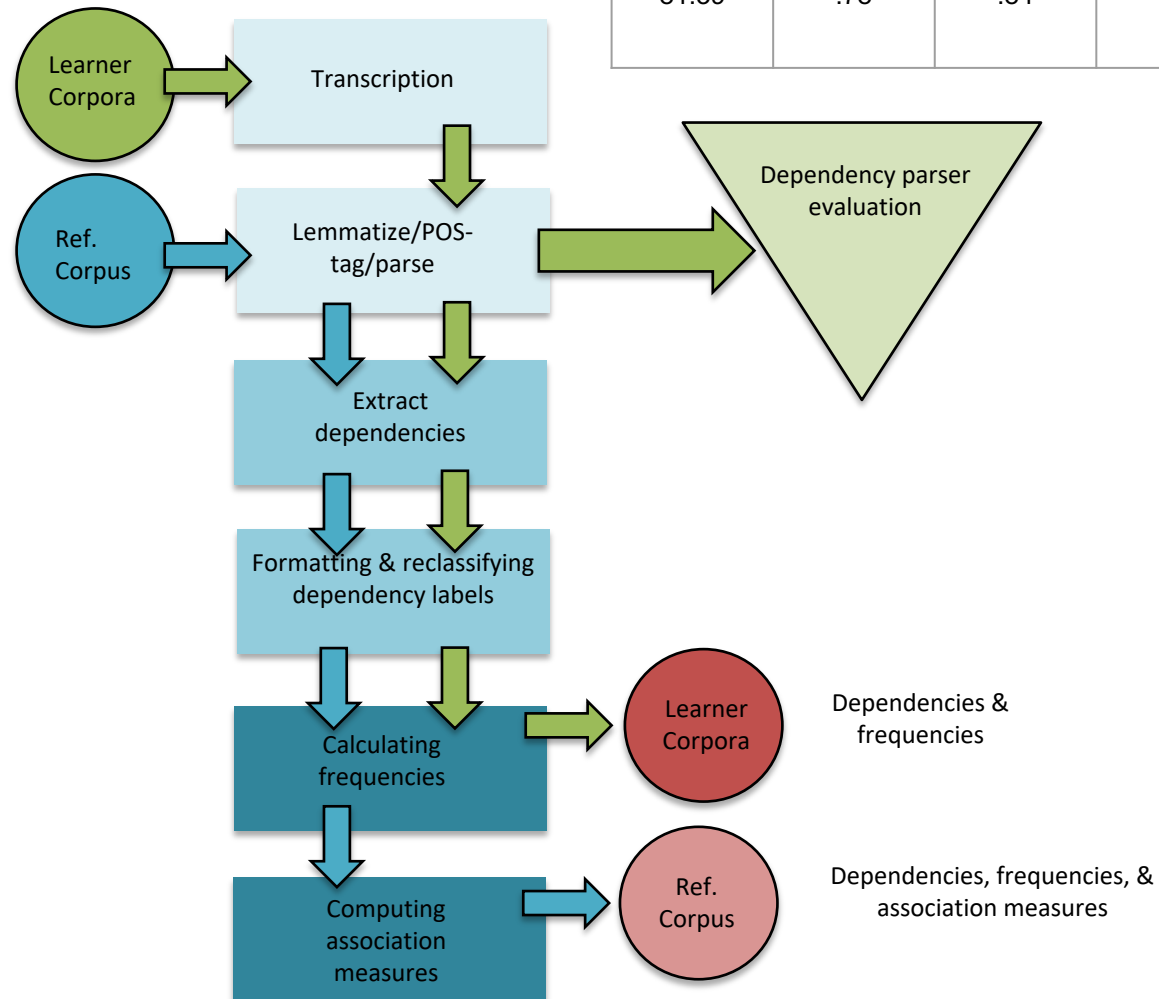
- *Morphological*
 - Morphemes per word (excluding names)
- Sophistication
 - lemma frequency logarithm (excluding names)
 - proportion of content words among 1000 most frequent words
 - proportion of content words among 5000 most frequent words
- Length
 - Corrected word length in letters for nouns
- Density
 - Density of content words
 - Density of compositional nominal compounds
- Diversity
 - TTR word
 - TTR lemma
 - MTLD lemma
 - TTR content words
 - MTLD content words

Phraseological measures

- Diversity
 - TTR direct object dependencies
 - TTR adjectival modifier dependencies
 - TTR adverbial modifier dependencies
 - Sophistication
 - dobj MI IQR
 - amod MI IQR
 - advmod MI IQR
 - dobj MI mean
 - amod MI mean
 - advmod MI mean
 - proportion dobj MI mean below threshold
 - prop. amod MI mean below threshold
 - prop. advmod MI mean below threshold
 - prop. dobj MI mean non-collocational
 - prop. amod MI mean non-collocational
 - prop. advmod MI mean non-collocational
- prop. amod mi mean collocational: low
 - prop. advmod mi mean collocational: low
 - prop. dobj mi mean collocational: med
 - prop. amod mi mean collocational: med
 - prop. advmod mi mean collocational: med
 - prop. dobj mi mean collocational: high
 - prop. amod mi mean collocational: high
 - prop. advmod mi mean collocational: high

From corpus pre-processing to complexity measures

LAS	F dobj	F amod	F advmod
81.39	.78	.84	.84



Dutch reference data

Lassy Groot Corpus version 4.0 (2016)

~700 million tokens (~350 million used here)

Lemmatized & POS-tagged (Frog)

Syntactically parsed (Alpino)

Genre variety:

- Books; brochures; legal texts; newspapers; policy documents; websites; wikipedia; etc.

Print & electronic media

Statistical analysis

- Statistical approach here assesses the role of complexity measures in the pass/fail assessment
 - unequal sample sizes
 - uncertainty in proficiency level
- 3 x binary logistic regression model
 - Outcome = binary pass/fail variable
 - Predictors determined by stepwise selection method (bidirectional)
 - Correlation table for multicollinearity (& VIF scores)

Results: binary logistic regression - B1

Model Fit

Model $\chi^2(11) = 98.57$ $p < .01$; Nagelkerke's $R^2: .50$

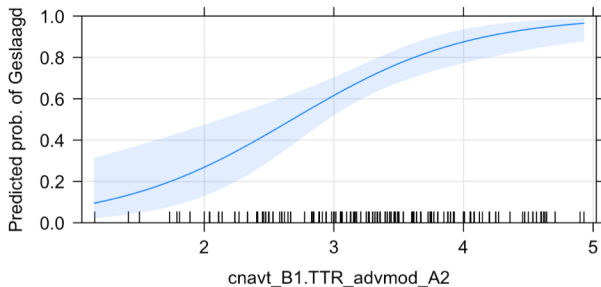
Predictors	B	S.E.	β	OR	z	P-value
Intercept	-19.23	4.09	0.87	2.39	-4.7	< .01
TTR_advmod_A2**	1.47	0.36	1.09	2.96	4.05	< .01
Lijdv_d_A2*	0.38	0.15	0.57	1.76	2.56	.01
AL_sub_ww_A2*	0.58	0.28	0.46	1.59	2.1	.04
amod_mi_coll_low_A2*	2.14	0.82	0.51	1.67	2.6	.01
advmod_mi_mean_A2_nobt**	1.77	0.49	1.17	3.23	3.59	< .01
advmod_mi_non_coll_A2*	3.13	1.4	0.6	1.83	2.23	.03
mi_iqr_advmod_A2	0.44	0.23	0.41	1.5	1.9	.06
dobj_mi_non_coll_A2	-2.28	1.2	-0.37	0.69	-1.9	.06
Wrd_per_zin_A2 *	2.27	0.84	0.82	2.28	2.69	.01
dobj_mi_coll_hi_A2	2.01	1.23	0.32	1.37	1.64	.10
D_level_A2	-0.59	0.32	-0.54	0.58	-1.85	.06

binary
pass/fail
outcome
variable

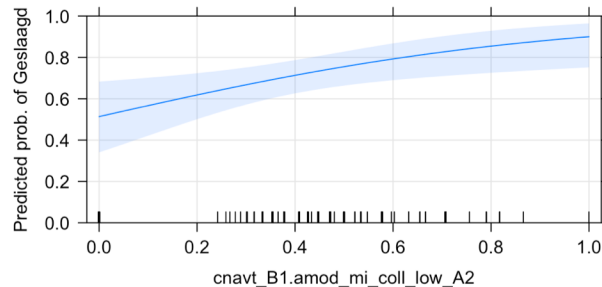
* significant at $p < .05$; ** significant at $p < .01$

B1 Effect Plots

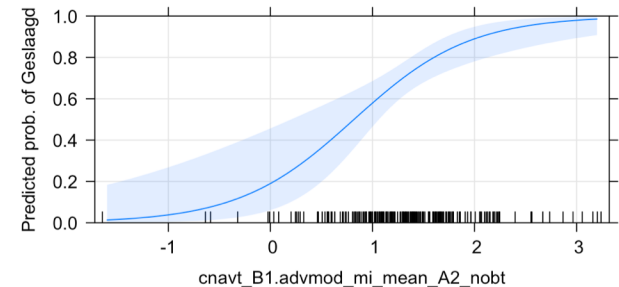
cnavt_B1.TTR_advmod_A2 predictor effect plot



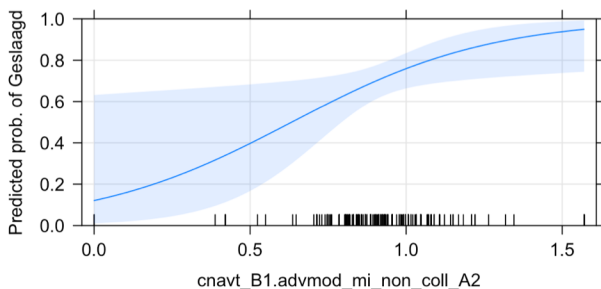
cnavt_B1.amod_mi_coll_low_A2 predictor effect plot



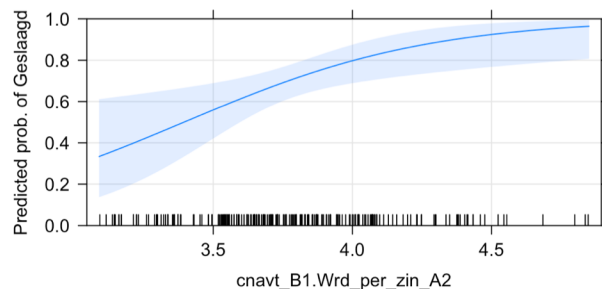
cnavt_B1.advmod_mi_mean_A2_nobt predictor effect plot



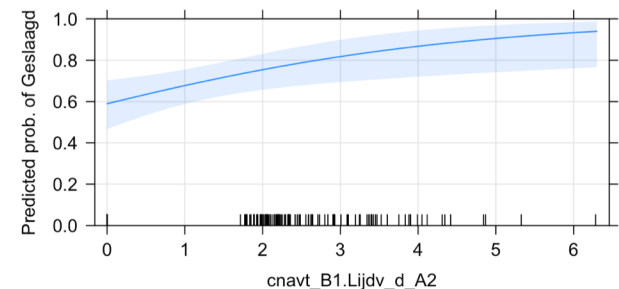
cnavt_B1.advmod_mi_non_coll_A2 predictor effect plot



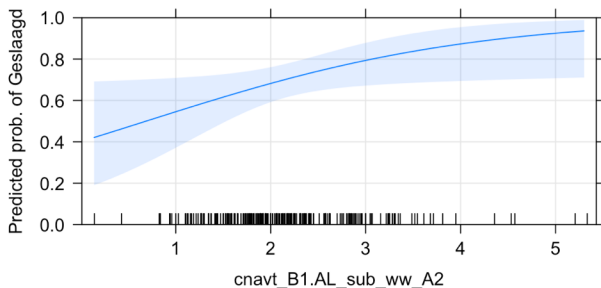
cnavt_B1.Wrd_per_zin_A2 predictor effect plot



cnavt_B1.Lijdv_d_A2 predictor effect plot



cnavt_B1.AL_sub_ww_A2 predictor effect plot



Results: binary logistic regression – B2

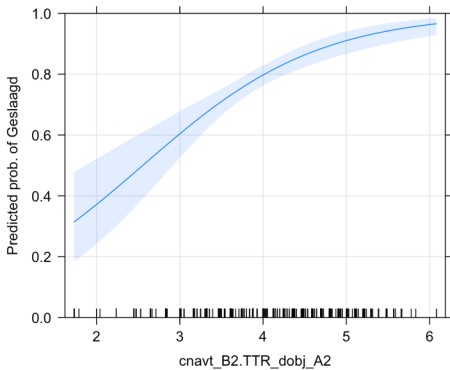
binary
pass/fail
outcome
variable

Model Fit						
Model $\chi^2(22) = 331.39$ $p < .01$; Nagelkerke's $R^2: .42$						
Predictors	B	S.E.	β	OR	z	P-value
Intercept	3.11	5.8	1.26	3.52	0.54	.59
TTR_dobj_A2**	0.95	0.17	.73	2.07	5.69	< .01
Bijw_bep_d_A2**	0.02	0.05	0.57	1.77	4.68	< .01
AL_sub_ww_A2**	1.64	0.43	0.50	1.65	3.82	< .01
dobj_mi_mean_A2_nobt**	0.38	0.12	0.39	1.48	3.32	< .01
Bijv_bep_d_A2**	0.01	0.01	0.39	1.48	2.58	0.01
Ljdv_d_A2**	0.44	0.12	0.32	1.38	3.78	< .01
dobj_bt_A2*	-1.62	0.69	-0.23	0.79	-2.35	.02
dobj_mi_coll_hi_A2**	2.55	0.78	0.38	1.47	3.26	< .01
Nom_d_A2**	-0.03	0.01	-0.85	0.43	-4.7	< .01
advmod_mi_mean_A2_nobt**	1.13	0.39	0.27	1.31	2.93	< .01
TTR_advmod_A2**	0.55	0.16	0.44	1.56	3.38	< .01
dobj_mi_coll_med_A2**	-2.07	0.72	-0.33	0.72	-2.86	< .01
Morf_per_wrd_zn_A2	4.06	2.15	0.23	1.26	1.89	.06
AL_gem_A2**	-1.78	0.54	-0.47	0.62	-3.27	< .01
Bijzin_conj_A2**	1.43	0.49	0.29	1.34	2.94	< .01
Lem_freq_zn_log_A2**	-5.1	1.11	-0.67	0.51	-4.62	< .01
Freq5000_inhwrld_A2**	10.61	2.46	0.76	2.14	4.32	< .01
MTLD_lem_A2	0.01	0.01	0.17	1.88	1.65	0.1
advmod_bt_A2	-0.89	0.58	-0.14	0.87	-1.54	0.12
AL_lidw_zrnw_A2	1.13	0.64	0.21	1.23	1.77	0.08
TTR_amod_A2	-0.33	0.19	-0.24	0.78	-1.72	0.08
Hzin_conj_A2	0.68	0.46	0.14	1.15	1.48	0.14

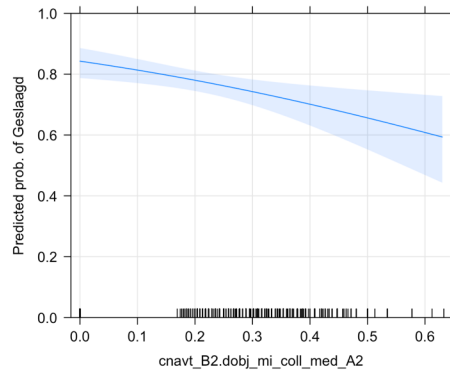
* significant at $p < .05$; ** significant at $p < .01$

B2 Effect Plots - Phraseological

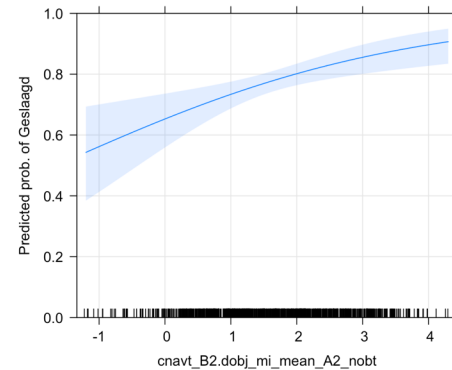
cnavt_B2.TTR_dobj_A2 predictor effect plot



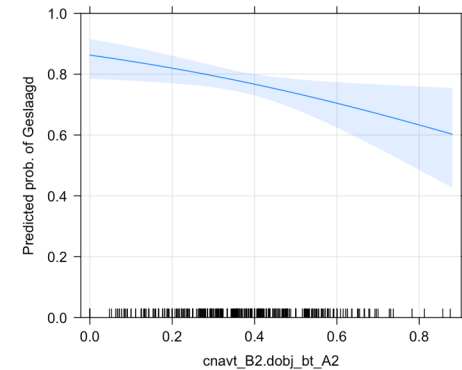
cnavt_B2.dobj_mi_coll_med_A2 predictor effect plot



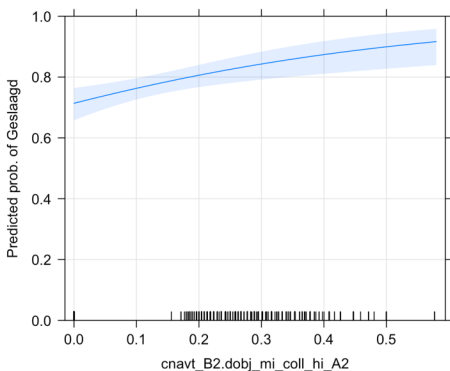
cnavt_B2.dobj_mi_mean_A2_nobt predictor effect plot



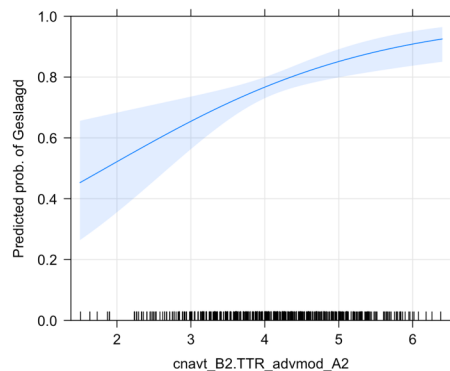
cnavt_B2.dobj_bt_A2 predictor effect plot



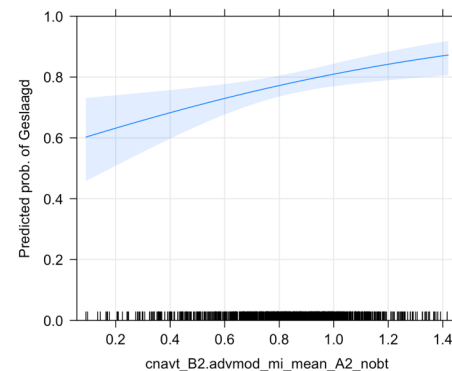
cnavt_B2.dobj_mi_coll_hi_A2 predictor effect plot



cnavt_B2.TTR_advmod_A2 predictor effect plot

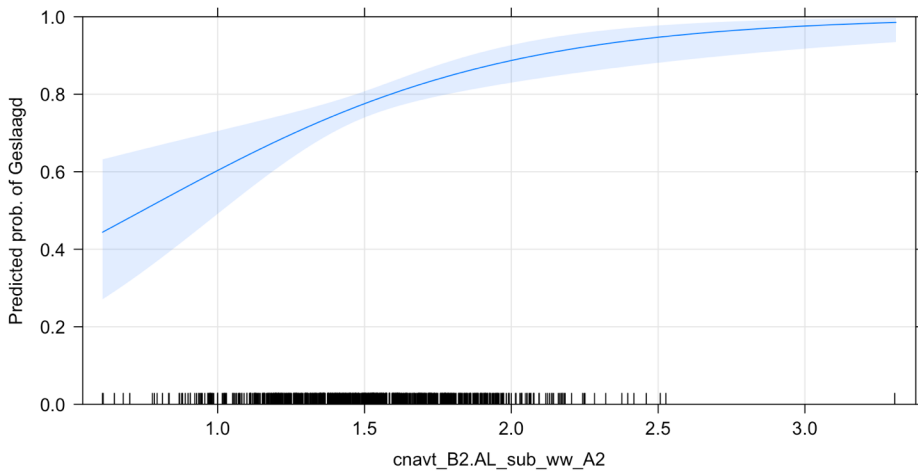


cnavt_B2.advmod_mi_mean_A2_nobt predictor effect plot

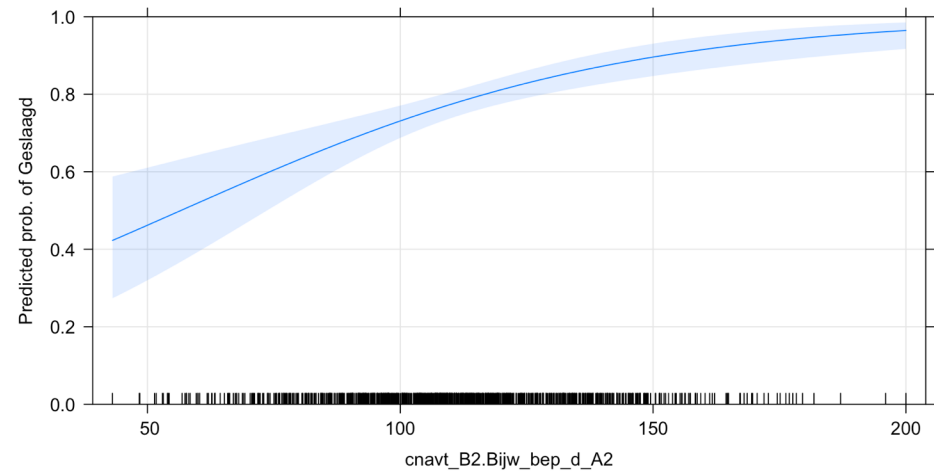


B2 Effect Plots – Syntactic & Lexical

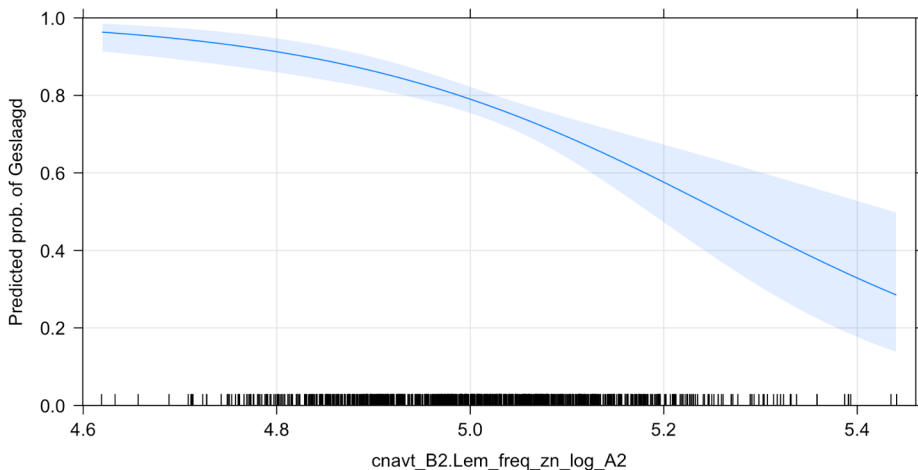
cnavt_B2.AL_sub_ww_A2 predictor effect plot



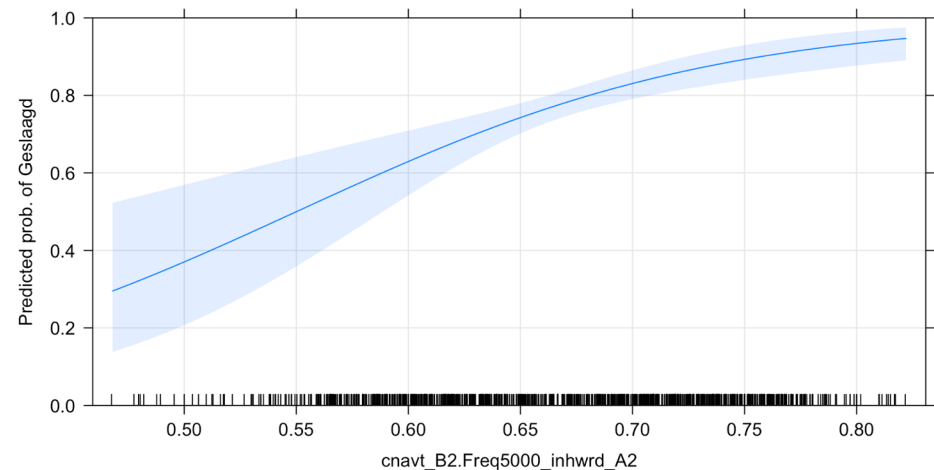
cnavt_B2.Bijw_bep_d_A2 predictor effect plot



cnavt_B2.Lem_freq_zn_log_A2 predictor effect plot



cnavt_B2.Freq5000_inhwrld_A2 predictor effect plot



Examples: B1 Fail

Er is problemen met Jacob straat. Tim Prince, de adviser heeft deze problemen gevonden. Eerst, de straat is [te slechte], en veel auto [komen] [te [snel]]. De oorzaak is dat auto om 50 km/H zijn, maar de straat heeft een 30 km/H reglementatie.

Een goede oplossing voor de probleem met [te snel] auto is om een permanent fliets te plaseren, we [kunnen ook] de begin van het straat zion hebben.

Duurtbewoner kunnen idee geven naar het E-mail van Jacob straat. Ik denk dat de permanent fliets de beste idee is, omdat een mogelijke solutie is.

Intercept = -19.2298			
variable	score	slope	total change
TTR_advmod:	1.788854	1.4732	+2.64
Lijdv_d:	0	0.3791	+0
AL_sub_ww:	1.3	0.5826	+0.76
amod_mi_coll_low:	0	2.1441	+0
advmod_mi_mean_nobt:	1.7931	1.7694	+3.17
advmod_mi_non_coll:	0.8860771	3.1334	+2.78
mi_iqr_advmod:	2.6931	0.4390	+1.18
dobj_mi_non_coll:	0.4	-2.2815	-0.91
Wrd_per_zin:	3.792553	2.2662	+8.59
dobj_mi_coll_hi	0	2.0122	+0
D_level	2	-0.5907	-1.18

-19.2298 + total change = -2.1998

advmod-v(kom+verb,snel+adj)	0.6544, mi_non_coll
advmod-adj(slecht+adj,te+adv)	1.2969, mi_non_coll
advmod-adj(snel+adj,te+adv)	3.3475, mi_coll_low
advmod-v(kan+verb,ook+adv)	0.3192, mi_non_coll
advmod-adj(snel+adj,te+adv)	3.3475, mi_coll_low

Examples: B1 Fail

Er **is** problemen met Jacob straat. Tim Prince, de adviser **heeft** deze problemen gevonden. Eerst, de **straat** **is** te slechte, en veel **auto** **komen** te snel. De **oorzaak** **is** dat **auto** om 50 km/H **zijn**, maar de **straat** **heeft** een 30 km/H reglementatie.

Een goede **oplossing** voor de probleem met te snel auto **is** om een permanent fliets te plaseren, **we** **kunnen** ook de begin van het straat zion hebben.

Duurtbewoner **kunnen** idee geven naar het E-mail van Jacob straat.

Ik **denk** dat de permanent **fliets** de beste idee **is**, omdat een mogelijke solutie is.

Intercept = -19.2298

variable	score	slope	total change
TTR_advmod:	1.788854	1.4732	+2.64
Lijdv_d:	0	0.3791	+0
AL_sub_ww:	1.3	0.5826	+0.76
amod_mi_coll_low:	0	2.1441	+0
advmod_mi_mean_nobt:	1.7931	1.7694	+3.17
advmod_mi_non_coll:	0.8860771	3.1334	+2.78
mi_iqr_advmod:	2.6931	0.4390	+1.18
dobj_mi_non_coll:	0.4	-2.2815	-0.91
Wrd_per_zin:	3.792553	2.2662	+8.59
dobj_mi_coll_hi	0	2.0122	+0
D_level	2	-0.5907	-1.18

-19.2298 + total change = -2.1998

Examples: B1 Pass

Tim Prinz, die voor “jaag op straat” werkt, wil de mening van de bewoners van Utrecht hebben omdat [er twee problemen daar zijn]. Het eerste probleem is de feit dat de buurtbewoners de anderen in hun straat [niet kennen] omdat een straat een klein stadje is. Daarom moeten buurtbewoners samen praten, [soms nieuws van de anderen vragen] en hen [beter kennen]. Als buurtbewoner zou ik proberen met de anderen te praten of [tenminste hen groeten] wanneer ik hen in de straat zien. Ik zou ook proberen geen problemen met hen te hebben, [bijvoorbeeld zou ik mijn auto niet voor hun huisje parkeren]. Ik vind belangrijk goede relaties met mijn burens hebben. Het tweede probleem betreft de snelheid van de auto’s in de straat: veel auto’s rijden te snel want ze komen te snel in de straat en ze respecteren de snelheidsbeperking niet. Er zijn twee mogelijkheden om dit probleem om te lossen... We kunnen kleine heuvels bouwen zodat de bestuurders verplicht zullen zijn hun snelheid te verminderen, of we kunnen een permanente flits inzetten. Dit apparaat zou fotos nemen wanneer een auto te snel rijdt dus kunnen we de snelheid beter controleren en de bestuurders die te snel reden moeten een boete betalen. Maar een flits zou een invloed op alle bewoners hebben daarom wil Tim Prinz hun mening hebben. Ze kunnen ook een email sturen als ze vragen of andere ideeën hebben. Als ik moeste kiezen zou ik liever een flits inzetten omdat ik geen probleem zal hebben als ik de snelheidsbeperking respecteer. Met de eerste oplossing, de kleine heuvels, zou het vervelend zijn zelfs als ik niet te snel rijd daarom vind ik deze idee niet goed. Als u uw mening wil ook geven kunt u een email sturen naar Tim Prinz met “jaagopstraat@gemeentutrecht.com”.

Intercept = -19.2298

variable	score	slope	total change
TTR_advmod:	4.199206	1.4732	+6.19
Lijdv_d:	0	0.3791	+0
AL_sub_ww:	2.85965	0.5826	+1.67
amod_mi_coll_low:	0.4472136	2.1441	+0.96
advmod_mi_mean_nobt:	1.064045	1.7694	+1.88
advmod_mi_non_coll:	0.9911566	3.1334	+3.11
mi_iqr_advmod:	3.2985	0.4390	+1.45
dobj_mi_non_coll:	0.4642857	-2.2815	-1.06
Wrd_per_zin:	4.336283	2.2662	+9.83
dobj_mi_coll_hi	0.3273268	2.0122	+0.66
D_level	4.46667	-0.5907	-2.64

-19.2298 + total change = 2.82

advmod-v(ben+verb,daar+adv)	-0.1768, mi_non_coll
advmod-v(ben+verb,er+adv)	1.4446, mi_non_coll
advmod-v(ken+verb,niet+adv)	0.9177, mi_non_coll
advmod-v(ken+verb,goed+adj)	2.1245, mi_non_coll
advmod-v(vraag+verb,soms+adv)	0.5964, mi_non_coll
advmod-v(groet+verb,tenminste+adv)	0.6507, bt
advmod-v(parkeer+verb,bijvoorbeeld+adv)	-1.1107, mi_non_coll
advmod-v(parkeer+verb,niet+adv)	-0.8867, mi_non_coll
advmod-v(probeer+verb,ook+adv)	0.0625, mi_non_coll

Examples: B1 Pass

Tim Prinz, die voor “jaag op straat” **werkt**, **wil** de mening van de bewoners van Utrecht hebben omdat er twee **problemen** daar **zijn**. Het eerste **probleem** **is** de feit dat de **buurtbewoners** de anderen in hun straat niet **kennen** omdat een **straat** een klein stadje **is**. Daarom **moeten** **buurtbewoners** samen praten, soms nieuws van de anderen **vragen** en hen beter **kennen**. Als buurtbewoner **zou** **ik** proberen met de anderen te praten of tenminste hen groeten wanneer **ik** hen in de straat **zien**. **Ik** **zou** ook proberen geen problemen met hen te hebben, bijvoorbeeld **zou** **ik** mijn auto niet voor hun huisje parkeren. **Ik** **vind** belangrijk goede relaties met mijn burens hebben. Het tweede **probleem** **betreft** de snelheid van de auto's in de straat: veel **auto's** **rijden** te snel want **ze** **komen** te snel in de straat en **ze** **respecteren** de snelheidsbeperking niet. Er **zijn** twee **mogelijkheden** om dit probleem om te lossen... **We** **kunnen** kleine heuvels bouwen zodat de **bestuurders** verplicht **zullen** zijn hun snelheid te verminderen, of we kunnen een permanente flits inzetten. Dit apparaat zou fotos nemen wanneer een auto te snel rijdt dus kunnen we de snelheid beter controleren en de bestuurders die te snel reden moeten een boete betalen. Maar een flits zou een invloed op alle bewoners hebben daarom wil Tim Prinz hun mening hebben. Ze kunnen ook een email sturen als ze vragen of andere ideeën hebben. Als ik moeste kiezen zou ik liever een flits inzetten omdat ik geen probleem zal hebben als ik de snelheidsbeperking respecteer. Met de eerste oplossing, de kleine heuvels, zou het vervelend zijn zelfs als ik niet te snel rijd daarom vind ik deze idee niet goed. Als u uw mening wil ook geven kunt u een email sturen naar Tim Prinz met “jaagopstraat@gemeentutrecht.com”.

Intercept = -19.2298

variable	score	slope	total change
TTR_advmod:	4.199206	1.4732	+6.19
Lijdv_d:	0	0.3791	+0
AL_sub_ww:	2.85965	0.5826	+1.67
amod_mi_coll_low:	0.4472136	2.1441	+0.96
advmod_mi_mean_nobt:	1.064045	1.7694	+1.88
advmod_mi_non_coll:	0.9911566	3.1334	+3.11
mi_iqr_advmod:	3.2985	0.4390	+1.45
dobj_mi_non_coll:	0.4642857	-2.2815	-1.06
Wrd_per_zin:	4.336283	2.2662	+9.83
dobj_mi_coll_hi	0.3273268	2.0122	+0.66
D_level	4.46667	-0.5907	-2.64

-19.2298 + total change = 2.82

Examples: B2 Fail

Dag iedereen, Vandaag zullen we over industrialisatie spreken. Industrialisatie is een proces waarop de machines het [werk doet]. We [gebruiken de stoommachines] voor goederen om efficiënter te werken. Deze revolutie is op korte termijn ontstaat maar de gevolgen/oorzaken worden op lange termijn. We zeggen dat industrialisatie in Engeland in de 18de eeuw ontstaat. Het is niet echt de waarheid want in de 3de eeuw voor Christus reeds tijdens de Hellenistische periode worden geen stoommachines, maar er was een markt economie, geld en vraag naar goederen (de verschillende maatstaf van industrialisatie) Maar waarom wordt industrialisatie in Engeland aanwezig? Er zijn 5 voorwaarden: Eerst de kapitaal accumulatie: Engeland had veel [geld te investeren] om nieuwe [zaken te bouwen]. Ten tweede is de voorkomende steenkool, ijzer en katoen en ijzer was in Engeland en ze [exporteren de katoen] van Amerika. Derde voorwaarden is de politiek klimaat: Engeland was stabiel en de rijke burgerijen [hadden hun eigen omgeving]. Daarna is omgeving die grote maat tolerantie aanwikkend gedrag. Er was een andere manier van denken. Ze [hadden meer vrijehdi dus better ideeën]. Ten laatste, ze [hadden technische intellectueel klimaat] ze [hadden een heftige zee] dus ze moeten goed varen en ze moeten [schepen hebben] die technologisch goed zijn.

variable	score	slope	total change
TTR_dobj_A2:	3.316625	0.947598	+3.14
Bijw_bep_d_A2:	88.2353	0.023048	+2.03
AL_sub_ww_A2:	0.9765411	1.642229	+1.6
dobj_mi_mean_A2_nobt:	0.2951857	0.382232	+0.11
Bijv_bep_d_A2:	147.059	0.013240	+1.95
Lijdv_d_A2:	0	0.444812	+0
dobj_bt_A2:	0.3636364	-1.615545	-0.59
dobj_mi_coll_hi_A2:	0	2.553140	+0
Nom_d_A2:	63.7255	-0.027931	-1.78
advmod_mi_mean_A2_nobt:	0.6371911	1.133330	+0.72
TTR_advmod_A2:	2.333333	0.552114	+1.29
dobj_mi_coll_med_A2:	0	-2.065298	+0
Morf_per_wrd_zn_A2:	1.44898	4.063862	+5.89
AL_gem_A2:	1.31541	-1.780949	-2.34
Bijzin_conj_A2:	0.2773501	1.434630	+0.4
Lem_freq_zn_log_A2:	5.05086	-5.101016	-25.76
Freq5000_inhwrz_A2:	0.440367	10.606658	+4.67
MTLD_lem_A2:	55.8928	0.014114	+0.79
advmod_bt_A2:	0.3333333	-0.893396	-0.3
AL_lidw_znw_A2:	0.5443202	1.128410	+0.61
TTR_amod_A2:	3.741657	-0.332088	-1.24
Hzin_conj_A2:	0.8770582	0.681915	+0.6

3.11 + total change = -5.01

dobj(doe+verb,werk+noun)	2.2924, mi_non_coll
dobj(gebruik+verb,stoommachine+noun)	4.1191, bt
dobj(bouw+verb,zaak+noun)	-2.6188, mi_non_coll
dobj(investeer+verb,geld+noun)	4.3845, mi_coll_low
dobj(exporteer+verb,katoen+noun)	8.5835, bt
dobj(heb+verb,omgeving+noun)	-1.4686, mi_non_coll
dobj(heb+verb,better_idee+noun)	NA, bt
dobj(heb+verb,vrijehdi+noun)	NA, bt
dobj(heb+verb,technisch+adj)	0.5462, mi_non_coll
dobj(heb+verb,zee+noun)	0.2121, mi_non_coll
dobj(heb+verb,schip+noun)	-1.2815, mi_non_coll

Examples: B2 Pass

Uitleg over industrialisatie

Allereerst is de [term industrialisatie uitgelegd]. Men spreekt soms van de industriële revolutie, van machines die uitgevonden zijn voor het maken van goederen. Het wordt vaak gezien als een revolutionair proces. Maar eigenlijk is de industrialisatie geen revolutie zoals men het vaak hoort. Een revolutie verloopt inderdaad op korte termijn en [heeft snel een effect] op de samenleving. Voor de industrialisatie is het niet het geval want het is een langzaam proces. Dan zeggen we vaak dat industrialisatie een typisch Engels fenomeen is. Het is niet helemaal juist. Zoals eerder gezegd is de industrialisatie heel langzaam en het is er al de vijf laatste honderd jaren aan begonnen. Ook in de Oude Griekenland (IIIde eeuw voor Christus) waren geld, vraag naar goederen, economische markt en kennis aanwezig maar we kunnen niet spreken van industrialisatie want [die werden niet ingezet] voor het produceren van goederen. Die vier [aspecten noemen] we de algemene vruchtbare omstandigheden voor industrialisatie. Dit vier omstandigheden zijn noodzakelijk voor industrialisatie maar er bestaan ook voorwaarden ermee. Ten eerste is de kapitaalaccumulatie in de loop van de XVIIde eeuw in Engeland aanwezig (ook in China). Immers heeft de Engels bevolking geld verdiend, ze worden nieuwe zaken ontwikkeling, geld was beschikbaar voor investeringen. Ten tweede is het voorkomen van kolen, ijzeren en katoen nodig. Het zijn grondstoffen voor de industriële bewerking. Alleen maar katoen vinden we in Engeland. De derde voorwaarde bestaat uit het politieke klimaat. Op het einde van de XVIIde eeuw heeft een groot deel van de bevolking invloed op de gang van zaken. Vooral de burgerij (rijke mensen) heeft een steeds belangrijkere positie. Op de vierde plaats vindt men de tolerantie voor afwijkend gedrag of afwijkende denkbeelden. Het kan op allerlei manieren gebeuren. We krijgen meer vrijheid om anders te denken. Dit leidt tot een vernieuwing. Eindelijk is het technische intellectuele klimaat ook een voorwaarde. Men bouwde meer en meer schepen die beter en beter uitgerust waren om goed te navigeren. Men hechtte meer belang aan bouwkwaliteit, aan technologie en de kennis was aanwezig in die technologie. Men had meer technische ruimte en men stelde zich de vraag: "Wat kunnen wij nog anders doen met die stoommachine?"

variable	score	slope	total change
Intercept = 3.11			
TTR_dobj_A2:	4.358899	0.947598	+4.13
Bijw_bep_d_A2:	112.948	0.023048	+2.6
AL_sub_ww_A2:	1.111031	1.642229	+1.82
dobj_mi_mean_A2_nobt:	2.588607	0.382232	+0.99
Bijv_bep_d_A2:	154.27	0.013240	+2.04
Lijdv_d_A2:	1.926486	0.444812	+0.86
dobj_bt_A2:	0.2105263	-1.615545	-0.34
dobj_mi_coll_hi_A2:	0.3973597	2.553140	+1.01
Nom_d_A2:	79.8898	-0.027931	-2.23
advmod_mi_mean_A2_nobt:	1.120558	1.133330	+1.27
TTR_advmod_A2:	5.112077	0.552114	+2.82
dobj_mi_coll_med_A2:	0.4588315	-2.065298	-0.95
Morf_per_wrd_zn_A2:	1.47059	4.063862	+5.98
AL_gem_A2:	1.32687	-1.780949	-2.36
Bijzin_conj_A2:	0	1.434630	+0
Lem_freq_zn_log_A2:	5.08581	-5.101016	-25.94
Freq5000_inhwrld_A2:	0.775401	10.606658	+8.22
MTLD_lem_A2:	61.5238	0.014114	+0.87
advmod_bt_A2:	0.2581989	-0.893396	-0.23
AL_lidw_znw_A2:	0.563901	1.128410	+0.64
TTR_amod_A2:	4.8	-0.332088	-1.59
Hzin_conj_A2:	0.7319249	0.681915	+0.5

3.11 + total change = 3.22

dobj(leg_uit+verb,term+noun)	2.6872, mi_non_coll
dobj(vind_uit+verb,die+pron)	0.6159, bt
dobj(zie+verb,het+det)	0.5951, mi_non_coll
dobj(hoor+verb,het+det)	1.462, mi_non_coll
dobj(heb+verb,effect+noun)	2.7424, mi_non_coll
dobj(zet_in+verb,die+det)	0.5442, mi_non_coll
dobj(noem+verb,aspect+noun)	0.8695, mi_non_coll
dobj(verdien+verb,geld+noun)	5.04, mi_coll_med
dobj(hecht+verb,belang+noun)	8.5055, mi_coll_hi

Research questions

- **RQ 1:** How well do phraseological complexity measures perform as indices of L2 Dutch proficiency?
- **RQ 2:** How do measures of phraseological complexity compare to more traditional measures of lexical and syntactic complexity as indices of L2 Dutch proficiency?
- **RQ 3:** How do the results here compare to the results obtained for L2 English in Paquot (2018, 2019)?

Linguistic complexity in the CNaVT corpus

- The results are interpreted in terms of the importance of the various complexity measures in the pass/fail assessment of learners at two CEFR levels — B1 & B2
- At B1, mostly phraseological (advmod) and a few syntactic measures contribute significantly to the prediction of the pass/fail outcome of the exam
 - **mostly phraseological (diversity & sophistication), some syntactic measures (dependency distance, sentence length)**
- At B2, many measures contribute significantly to the prediction of the pass/fail outcome variable, including syntactic, lexical, and phraseological (dobj)
 - **largest effects:** phraseological & lexical measures

Cross-linguistic perspective

- **L2 English** (Paquot, 2018)
 - syntactic complexity: no significant differences across proficiency levels
 - lexical complexity: significant difference in lexical density between B2/C1 & C2
 - phraseological complexity: significant increase in seven measures
- **Comparison**
 - Comparing different cross-sections of proficiency (B2-C2 vs. B1-C1)
 - Register/genre differences (academic vs. descriptive texts)
 - Research design differences (pass/fail vs. proficiency as outcome variable)
 - Difficulty of English syntax in comparison to Dutch syntax
 - Differences in target language exposure
 - Homogeneity of the learner group
 - Available complexity measures

Avenues for future research

- Reanalyzing once we address inequality in sample size between proficiency levels
- Considering the effect of topic on complexity
- Make the most of rich metadata set
 - task variables (text topic; oral/written prompt)
 - learner variables (L1 background, education level, motivation, etc.)
- Other phraseological measures
 - collocations (different phraseological units)
 - alternative association measures

Thanks!

Contact:

Rachel Rubin (rachel.rubin@vub.be)



References

- Banerjee, S. & Pederson, T. (2003). The Design, Implementation, and Use of the Ngram Statistics Package. Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics. Mexico City
- Bulon, A., Hendriks, I., Meunier, F., & Van Goethem, K. (2017). Using global complexity measures to assess second language proficiency: Comparing CLIL and non-CLIL learners of English and Dutch in French-speaking Belgium. *Papers of the Linguistic Society of Belgium*, 11(1), 1-25.
- Bulté, B. & Housen, A. (2012). Defining and operationalizing L2 complexity. In A. Housen, F. Kuiken, & I. Vedder (Eds.), *Dimensions of L2 Performance and Proficiency. Complexity, Accuracy and Fluency in SLA*. Amsterdam: John Benjamins, pp. 21-46.
- Gries, S.T. (2006). Some proposals towards more rigorous corpus linguistics. *Zeitschrift für Anglistik und Amerikanistik*, 54(2), 191-202.
- Gries, S.T. (2012). Frequencies, probabilities, association measures in usage-/exemplar-based linguistics: some necessary clarifications. *Studies in Language*, 36(3), 477-510
- Kuiken, F. & Vedder, I. (2019). Syntactic complexity across proficiency and languages: L2 and L1 writing in Dutch, Italian and Spanish. *International Journal of Applied Linguistics*, 1-19.
- Lassy Groot Corpus (Version 4.0) (2016) [Data set]. Available at the Dutch Language Institute: <http://hdl.handle.net/10032/tm-a2-f6>
- Michel, M., Kuiken, F., & Vedder, I. (2007). The influence of complexity in monologic versus dialogic tasks in Dutch L2. *International Review of Applied Linguistics in Language Teaching*, 45(3), 241-259.
- Michel, M., Kuiken, F., & Vedder, I. (2012). Task complexity and interaction: (combined) effects on task-based performance in Dutch as a second language. *EUROSLA Yearbook*, 12(1), 164-190.
- Ortega, L. (2003). Syntactic complexity measures and their relationship to L2 proficiency: a research synthesis of college-level L2 writing. *Applied Linguistics*, 24(4), 492-518.
- Ortega, L. (2012). Interlanguage complexity: A construct in search of theoretical renewal. In B. Kortmann & B. Szmrecsanyi (Eds.), *Linguistic Complexity: Second Language Acquisition, Indigenization, Contact*. Berlin, Boston: De Gruyter, pp. 127-155.
- Pallotti, G. (2015). A simple view of linguistic complexity. *Second Language Research*, 31(11), 117-134. <https://doi.org/10.1177/0267658314536435>
- 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.