

Bridging the gap between syntactic complexity and lexical diversity research: Phraseology – the keystone?

Magali Paquot

FNRS

Centre for English Corpus Linguistics

Université catholique de Louvain

Introduction

- Lexis and grammar are closely intertwined
- Language is essentially made up of word combinations that constitute single choices
- Words acquire meaning from context
- Word combinations play crucial roles in language acquisition, proficiency, fluency, idiomaticity and change

Sinclair (1991), Ellis (1996), Biber et al. (1999), Wray (2002), Stefanowitsch & Gries (2003), Schmitt (2004), Goldberg (2006), Ellis & Cadierno (2009), Römer (2009), Bybee & Beckner (2012)

Second language (L2) research

- Phraseology / formulaic language now at the forefront of debates in foreign language learning and teaching (Meunier & Granger, 2008; Polio, 2012)
- Not all domains of L2 research have navigated the transition.
 - Lexis « has been strikingly absent » in task-based learning research (Skehan, 2009)
 - « Phraseological competence is a relatively new perspective in the study of second language writing » (Connor-Linton & Polio, 2014)

Interlanguage complexity research

- « one of the major research variables in applied linguistic research » (Housen & Kuiken, 2009)
 - « At the core of the construct is the claim that the ability to produce more linguistically complex oral or written texts reflects increasingly more developed and mature capacities to use the second language. Thus, at stake is the relationship of linguistic complexity of L2 oral and written production to other constructs that are central in the field, such as **linguistic proficiency** and **linguistic development**. » (Ortega, 2012: 127)
- Narrowed down to **syntactic complexity**

Lexical complexity

- Lexical richness / diversity
 - Word length
 - Lexical density
 - Lexical sophistication
 - Lexical variation
 - Number of errors

Read (2000), Housen et al. (2008), Deane (2012),
Jarvis (2013), Bulté & Housen (2012, 2014)

Single word-based measures

- Lexical variation
 - Type-token ratio and its variants
 - *vocd-D* (Malvern & Richards, 1997)
 - MTLD (McCarthy & Jarvis, 2010)
- Lexical sophistication
 - Word frequency lists
- Jarvis, S. & Daller, M. (2013) *Vocabulary Knowledge. Human ratings and automated measures*. John Benjamins.

Automatic text analysis

- CHILDES CLAN programs (Mac Whinney, 2000)
- VocabProfile (Cobb, 2004)
- Coh-Metrix (Graesser et al, 2004)
- Linguistic Inquiry and Word Count (Pennebaker et al., 2007)
- L2 Syntactic Complexity Analyzer (Lu, 2010)
- Lexical Complexity Analyzer (Lu, 2012)
- R package: koRpus
- Tool for the Automatic Analysis of LEXical Sophistication (TAALES, Kyle & Crossley, 2014)
 - Bigram and trigram frequencies and proportions

Theoretical renewal

- “I argue that the construct (*complexity*) begs for long-ranging theoretical renewal if we are to see improvement in definitions, operationalizations, and measurement practices in the future.”
(Ortega, 2012: 128)
- Measures of lexical diversity « are not grounded in an adequate or clearly articulated theoretical account of the nature of the construct of lexical diversity » (Jarvis, 2013: 13)

Construct reductionism

- « The reduction of complexity to only one of the dimensions is an extended practice in L2 research » (Norris & Ortega, 2009: 566)
- « Syntactic complexity must be measured multidimensionally » (Norris & Ortega, 2009: 562; Ortega, 2012; Bulté & Housen, 2012)
- « Existing measures of lexical diversity take as their input far too little information to account for the diversity of word use in a text. » (Jarvis, 2013: 18)

Syntactic complexity

- Three measurable subconstructs tapping different sources of syntactic complexification
 - Overall syntactic complexity (e.g. mean length of T-unit)
 - Complexity by subordination (e.g. clauses per T-unit)
 - Complexity by subclausal or phrasal elaboration (e.g. mean length of clause, complex nominals per clause)

Lexical diversity (Jarvis, 2013)

- Variability
- Volume
- Evenness
- Rarity
- Dispersion
- Disparity

Word combinations?

MEET

MEET up

MEET your ends

MEET and greet (n.)

make ends MEET

Nice to MEET you

It's a pleasure to MEET you

Position statement

- A successful renewal of both interlanguage complexity research and lexical diversity research will require a better appreciation of the interdependence of lexis and grammar and the fundamental role of word combinations.

Objective

- Empirically demonstrate the theoretical and practical relevance of the **phraseological dimension** of language for L2 (syntactic/lexical) complexity research

Hanks's (2014) Theories of Norms and Exploitations

- Words of minimal complexity
 - *rhododendron, exasperate*
- Words of medium complexity
 - *eat, drink, swallow, chew, bite*

Medium semantic and syntagmatic complexity

bite: *Animals and indeed humans use their teeth to bite other animals (often one of the same species, as an act of aggression) or a physical object – a human or animal may **bite** something **off** – snakes bite their victims – insects bite humans and other animals too – people bite their nails when worried or as a nervous habit. The verb is also used in a number of idiomatic phrases: a process may begin to bite – a person may be bitten by a bug (for some activity) – an angry or upset person may bite another person's head off – a person may **bite back** something that he or she was about to say – an inanimate object may **bite into** another object by physical contact – an ungrateful person bites the hand that feeds him (or her) – sometimes a person must bite the bullet and get on with something – a person or a project may bite the dust and so die or stop.*

Hanks's (2914) Theories of Norms and Exploitations

- Words of minimal complexity
 - *rhododendron, exasperate*
- Words of medium complexity
 - *eat, drink, swallow, chew, bite*
- Words of maximum complexity
 - *up, at, on*
 - *take, make, come*
 - *hand, nose, head*

Usage-based approaches to second language acquisition

- « Language learning is primarily based on learners' exposure to their second language in use, that is, the linguistic input they receive. »
- « Language learning is the learning of constructions, pairings of form and meaning or function. Constructions range from simple morphemes like *-ing* to complex and abstract syntactic frames such as Subject-Verb-Object-Object (as in *Nick made Steffi a sandwich*). »
- « Ultimately constructions blur the traditional distinction between lexicon and grammar. »

Ellis & Wulff (2015)

Complexity measures in L2 research

- **Describe L2 performance**
- Assess L2 proficiency
- Trace L2 development

Wolfe-Quintero et al. (1998), Norris & Ortega (2009)
Ortega (2012), Housen et al. (2012)

Research questions

1. To what extent phraseological units constitute valid and reliable indices that can be used to objectively discriminate between learner texts at different proficiency levels in the target language?
2. How do phraseological indices compare with measures of syntactic and lexical complexity?

‘Advancedness’ in academic settings

- *Varieties of English for Specific Purposes Database (VESPA)*
 - L1s: Dutch, French, German, Italian, Norwegian, Spanish, Swedish
 - Disciplines: linguistics, business, engineering, ...
 - Genres: research papers, reports
 - Levels: BA + MA

<http://www.uclouvain.be/en-cecl-vespa.html>

VESPA-FR-LING

Per institutional level	Number of files	Total number of words	Means
B2	25	86,472	3,588
C1	62	216,283	3,488
C2	11	33,994	3,090
Total	98	336,749	3,436

To what extent phraseological units constitute valid and reliable indices that can be used to objectively discriminate between learner texts at different proficiency levels in the target language?

STUDY 1

Phraseological units – statistical definition

- “The co-occurrence of a form or a lemma of a lexical item and one or more additional linguistic elements of various kinds which functions as one semantic unit in a clause or sentence and whose frequency of co-occurrence is larger than expected on the basis of chance.” (Gries, 2008: 6)
- Pointwise Mutual Information scores

Statistically defined PUs in learner production

- Durrant & Schmitt (2009)
 - Compared with native speakers, EFL learners underuse premodifier-noun pairs with high PMI scores (e.g. *bated breath*, *preconceived notions*)
- Granger & Bestgen (2014)
 - Intermediate learner texts are characterized by a smaller proportion of low-frequency, but strongly associated bigrams as attested by higher PMI scores than advanced

RECURRENCE

Lexis-grammar interface

- Relational collocations (Evert, 2005)
 - Adjective + noun, adverb + adjective, adverb + adverb, adverb + verb, verb + direct object
- Stanford parser
 - Typed dependencies: triples of a relation between pairs of words
 - *dobj(analyzed, cells)*
 - « the object of *analyzed* is *cells* »

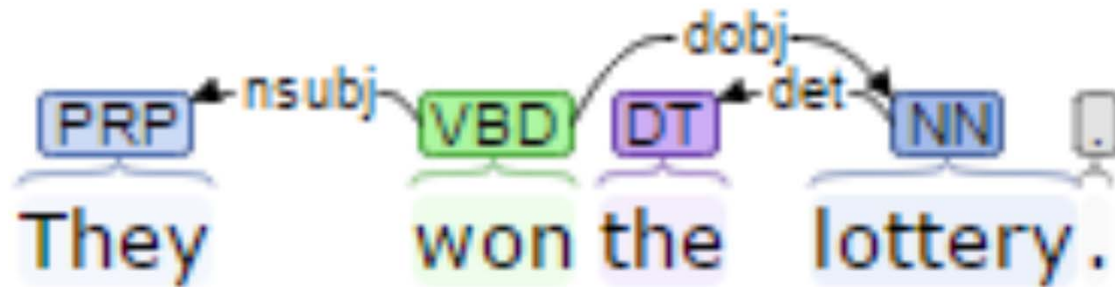
Corpus processing

Ref. corpus + VESPA	1. Lemmatisation and part-of-speech tagging	Stanford CoreNLP: a suite of core NLP tools (40 days 10 hours 52 min 51 sec!)
	2. Parsing and extraction of dependencies	
	3. Simplification of POS tags, computing frequencies, etc.	In-house Perl programs (several days as well ...)
Ref. corpus	4. Calculation of association measures between a pair of words in a particular Stanford typed dependency	Ngram Statistics Package (NSP)

1. Lemmatisation and part-of-speech tagging

- *They won the lottery.*
- *They*[they.PRP] *won*[win.VBD] *the*[the.DT] *lottery*[lottery.NN].

2. Parsing and extraction of dependencies



- nsubj(won,they)
- dobj(won,lottery)
- det(lottery,the)
- 49,754,608 dependencies

De Marneffe & Manning (2010)

3. In-house Perl programs

They[they.PRP] *won*[win.VBD]
the[the.DT] *lottery*[lottery.NN]

nsubj(*won*,*they*)
dobj(*won*,*lottery*)
det(*lottery*,*the*)

Lemma + simplified
POS

nsubj(win.VB,they.PRP) 8
dobj(win.VB,lottery.NN) 4
det(lottery.NN,the.DT) 25

Dependencies +
frequencies

A million thanks to Hubert Naets (CENTAL, UCL) for his invaluable help!

4. Association scores

- Assign to each word combination (type) extracted from the VESPA learner corpus an association score computed on the basis of a reference corpus (L2RC)
 - Pointwise mutual information
 - $\text{Freq} > 4$ in reference corpus
- Compute mean PMI scores for each dependency relations in each learner text (cf. Bestgen & Granger, 2014)

L2 research corpus (L2RC)

- 16 major journals in L2 research (1980-2014)
 - Applied Linguistics, Applied Language Learning, Applied Psycholinguistics, Bilingualism: Language and Cognition, The Canadian Modern Language Review, Foreign Language Annals, Journal of Second Language Writing, Language Awareness, Language Learning, Language Learning and Technology, Language Teaching Research, The Modern Language Journal, Second Language Research, Studies in Second Language Acquisition, System, TESOL Quarterly
- 7765 texts
- 66,218,913 words (363 Mio)

Thanks to Luke Plonsky (NAU) who gave me access to these files

Learner group comparisons

- R statistical software
 - Shapiro-Wilk normality tests
 - Normal distribution
 - ANOVAs ($p < 0.05$) + Tukey Contrasts
 - Distribution not normal
 - Kruskal-Wallis rank sum test ($p < 0.05$) + pairwise comparisons using Wilcoxon rank sum tests

amod, advmod, dobj

RESULTS

amod: adjectival modifier

- « Sam eats red meat. » → `amod(meat,red)`
NN JJ

$F(2,98) = 5,642$, $p = 0.00484$, eta squared (η^2) = 0,1061

	N	Mean PMI	sd
B2	25	2.423316	0.3331140
C1	62	2.619221	0.4196816
C2	11	2.904127	0.4382207

B2 – C1

C1 – C2 .

B2 – C2 **

Examples of *amod* dependencies

- $\text{pmi} > 6$: *overwhelming majority, derivational affix, prepositional phrase, hasty conclusion, integral part, slight predominance, keen interest, exhaustive list, adverbial particle, wide range, illustrative example, chronological order, wide variety, spontaneous speech, next section, possible explanation, large majority, significant difference, clear preference*
- $\text{pmi} = 1$: *main function, only conclusion, final part, common history, previous example, different field, same number, enough material, theoretical definition, lower class, common word, long word, syntactic characteristic, real power, specific form, common method, certain way, different function, general definition, different result, syntactic element, simple form*

advmod: adverbial modifier

- `advmod(unprecedented+JJ,totally+RB)`
- `advmod(enough+RB,strangely+RB)`
- `advmod(root+VB,firmly+RB)`

$F(2,98) = 6.382$, $p = 0.00251$, eta squared (η^2) = 0,1184

	N	Mean PMI	sd	
B2	25	1.181200	0.2989820	B2 – C1 **
C1	62	1.388658	0.2837951	C1 – C2
C2	11	1.484164	0.1993552	B2 – C2 **

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Examples of *advmod* dependencies

- $\text{pmi} > 7$:
 - `advmod(incorrect+JJ,grammatically+RB),`
`advmod(significant+JJ,statistically+RB),`
`advmod(rightly+RB,quite+RB),`
`advmod(understandable+JJ,perfectly+RB),`
`advmod(distribute+VB,evenly+RB),`
`advmod(evolve+VB,constantly+RB)`
- $\text{pmi} = 1$:
 - `advmod(interesting+JJ,quite+RB),`
`advmod(possible+JJ,also+RB),`
`advmod(puzzling+JJ,more+RB)`

dobj: direct object

- `dobj(make+VB,statement+NN)`

$F(2,98) = 8.636$, $p = 0.000358$, eta squared (η^2) = 0,1538

	N	Mean PMI	sd	
B2	25	1.790844	0.3868192	B2 – C1
C1	62	1.971910	0.3957174	C1 – C2 **
C2	11	2.376491	0.3560944	B2 – C2 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Examples of *dobj* dependencies

- $\text{pmi} > 7$:
 - $\text{dobj}(\text{arouse}+\text{VB}, \text{curiosity}+\text{NN}), \text{dobj}(\text{fill}+\text{VB}, \text{gap}+\text{NN}),$
 $\text{dobj}(\text{serve}+\text{VB}, \text{purpose}+\text{NN}),$
 $\text{dobj}(\text{pay}+\text{VB}, \text{attention}+\text{NN}), \text{dobj}(\text{play}+\text{VB}, \text{role}+\text{NN}),$
 $\text{dobj}(\text{divert}+\text{VB}, \text{attention}+\text{NN}),$
 $\text{dobj}(\text{corroborate}+\text{VB}, \text{finding}+\text{NN}),$
 $\text{dobj}(\text{avoid}+\text{VB}, \text{misunderstand}+\text{NN})$
- $\text{Pmi} = 1$:
 - $\text{dobj}(\text{have}+\text{VB}, \text{function}+\text{NN}),$
 $\text{dobj}(\text{consider}+\text{VB}, \text{characteristic}+\text{NN}),$
 $\text{dobj}(\text{have}+\text{VB}, \text{characteristic}+\text{NN}),$
 $\text{dobj}(\text{classify}+\text{VB}, \text{adjective}+\text{NN}),$
 $\text{dobj}(\text{mention}+\text{VB}, \text{agent}+\text{NN})$

Interim summary

- *amod, advmod, dobj*:
 - Mean PMIs: $B2 > C1 > C2$
- *amod*
 - Most widely used in automatic studies of phraseological competence in learner production
 - Least significant variable: $B2 / C2$
- *advmod*
 - Intermediate vs. advanced: $B2 / C1-C2$
- *dobj*
 - $B2 - C1 / C2$
 - Medium effect size

How do phraseological indices compare with measures of syntactic and lexical complexity?

STUDY 2

Measures of syntactic complexity

C/T	Clauses per T-unit
DC/T	Dependent clauses per T-unit
VP/T	Verb phrases per T-unit
CN/T	Complex nominals per T-unit
DC/C	Dependent clauses per clause
CN/C	Complex nominals per clause
MLC	Mean length of clause

No statistically significant differences between the three learner groups

L2 Syntactic Complexity Analyzer (Lu, 2010)

Mexures of lexical richness

Lexical density and sophistication

LD	Lexical density
LS1	Lexical sophistication-I
LS2	Lexical sophistication-II
VS1	Verb sophistication-I
CVS1	Corrected VSI
VS2	Verb sophistication-II

Lexical Complexity Analyzer (Lu, 2012)

Lexical density (N_{lex}/N)

- $F(2,98) = 4.744$, $p = 0,0109$, eta squared (η^2) = 0,0908

	N	mean LD	sd
B2	25	0.4676	0.0230
C1	62	0.4694	0.0220
C2	11	0.4909	0.0243

- B2 – C1
- B2 – C2 *
- C1 – C2 *

Signif. codes: 0 '****' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Mesures of lexical richness (2)

Lexical variation	
NDW	NDW (first fifty words)
NDW-ER50	NDW (expected random fifty) (mean T of 10 random 50-word samples)
NDW-ES50	NDW (expected sequence fifty) (mean T of 10 random 50-word sequences)
CTTR	Corrected TTR
RTTR	Root TTR
MSTTR-50	Mean segmental TTR (mean TTR of all 50-word segments)
SVV1	Squared verb variation-I
CVV1	Corrected verb variation-I
NV	Noun variation
AdjV	Adjective variation
AdvV	Adverb variation

Lexical Complexity Analyzer (Lu, 2012)

Noun variation ($T_{\text{noun}}/N_{\text{lex}}$)

- Kruskal-Wallis $X^2 = 6,213$, $p = 0,04476$, $df\ 2$, eta squared = 0,064

Level	N	Median
B2	25	0.26
C1	62	0.24
C2	11	0.33

- B2 – C1
- B2 – C2
- C1 – C2 *

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Interim summary

- No measure of syntactic complexity distinguishes between proficiency levels in VESPA
- Lexical density
 - B2 / C1-C2
- Noun variation
 - B2 > C1 < C2
- Mean PMI of relational collocations is a better descriptor of proficiency levels in VESPA than measures of syntactic complexity and lexical richness

DISCUSSION AND CONCLUSIONS

Phraseology

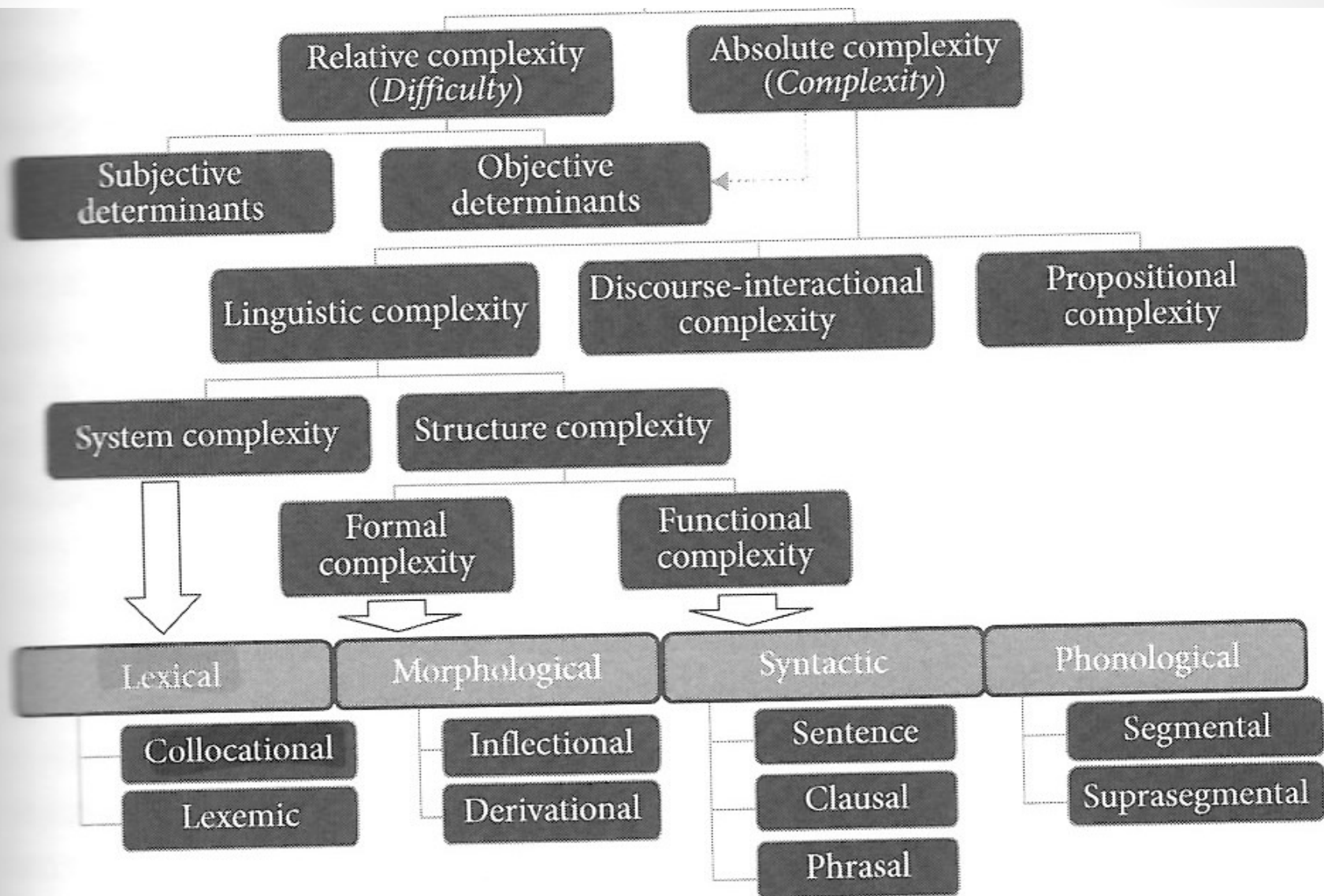
- Dimension of L2 writing quality
- Influence overall perceptions of language proficiency by expert evaluators
 - Adjacent proficiency levels

‘Advancedness’

- “in evaluating the syntactic complexity of compositions written by advanced adult second language learners, T-unit analysis does not seem to reflect accurately the knowledge of the learner” (Bardovi-Harlig, 1992: 391)
- Mean length of multicausal units and amount of subordination “can be used with relative confidence to distinguish broad (but not fine-grained) proficiency differences in second language written data produced at intermediate (but probably not very early or very advanced) levels of linguistic ability.” (Ortega, 2012: 131)
- “The development of (...) lexical productivity and diversity trails off after the first year” (Housen et al., 2008)

Another type of complexity?

- « when considered at the level of the learner's interlanguage system, linguistic complexity has been commonly interpreted as the size, elaboratedness, richness and diversity of the learners' linguistic L2 system » (Bulté & Housen, 2012: 464)



Bulté & Housen (2012: 23)

Phraseological complexity

- *the range of **phraseological** forms that surface in production and the degree of sophistication of such forms. (cf. Ortega 2003)*
- A text with a wide range of phraseological units (= **variability**) and a high proportion of relatively unusual or sophisticated units (= **sophistication**) will be said to be more phraseologically complex than one where the same basic units are often repeated.

From construct properties to measurement

- Variability
 - Measure of Textual Lexical Diversity (MTLD, McCarthy & Jarvis, 2010; Jarvis 2013)
- Sophistication
 - Mean frequency of phraseological units in a reference corpus
 - Mean pointwise mutual information (PMI) score for those units in reference corpus

Context-sensitive measures

- “It is (...) essential that complexity accounts for context” (Rimmer, 2009: 31)
- Register and genre
 - Operationalize the complexity of L2 language by how well it uses the phraseological units and lexico-grammatical characteristics of the norms of its reference genre (cf. Ellis et al, 2013)

Future work

- Construct validity
 - Predictive validity
 - Discriminant validity
- Benchmark linguistic development (Paquot, forthcoming a/b)
 - « any measure that serves as an index of development would probably have to cover the full trajectory of language acquisition, from the lowest level or stage to the highest » (Bulté & Housen, 2012: 37)

Implications for other fields in applied linguistics

- Language assessment and automated essay scoring
 - Many analytic rating scales involve individual assessments of lexis and grammar.
 - Lexis scores predicted by the scores on the grammar scale (Ruegg et al., 2011)
 - Raters often give the same score for the two scales (ibid.)
 - « if one is to accept that the two [lexis/grammar] are inextricably interwoven, the assessment criteria must reflect this by assessing lexicogrammar as a single criterion » (ibid.: 75)
- Readability

Muchas gracias!

Thank you very much!

Check out!

- **The Learner Corpus Association**
 - www.learnercorpusassociation.org
- **The International Journal of Learner Corpus Research**
 - General editors: Marcus Callies & Magali Paquot
 - John Benjamins Publishing
 - First issue just published!

