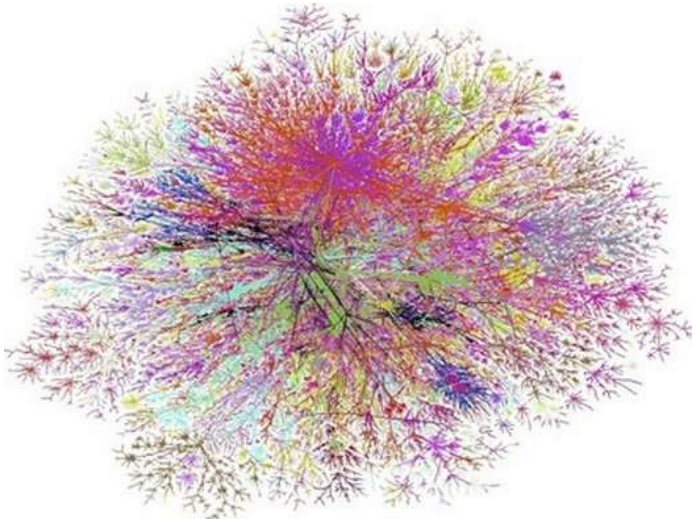
Profiling complexity: methodological issues and applications to L2 morphology

Gabriele Pallotti
University of Modena and Reggio Emilia

Why study complexity?

NOT

because it's trendy, fuzzy, mystic

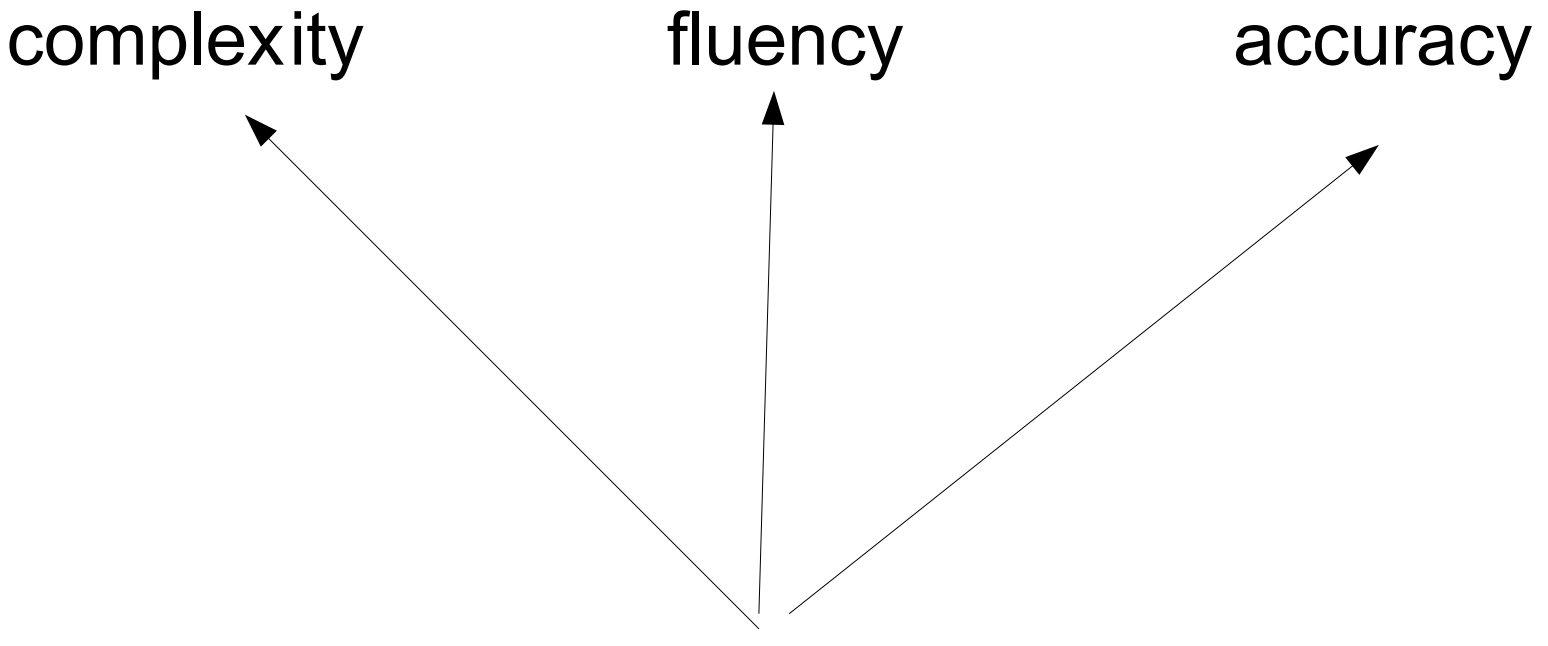


BUT

because

- it is a relevant property of linguistic structures, texts, systems
- it can be objectively measured
- it can be used as a dependent and independent variable to investigate several areas (e.g. L1, L2, typical/atypical language development, task effects, interplay with other dimensions)

Complexity, accuracy, fluency (CAF)



1990-today…
Learning an additional language means building a more complex system, becoming more fluent and more accurate
(Housen, Kuiken & Vedder 2012)

Complexity : interlanguage as such

Fluency : interlanguage use

what is there

Accuracy: interlanguage compared with another language (the target language)

what is missing

Complexity, and what it isn't

Theoretical definition of the construct

Three basic meanings of ‘complexity’

1. **Structural complexity**, a formal property of texts and linguistic systems having to do with the number of their elements and their relational patterns (= *complexity*)

2. **Cognitive complexity**, having to do with the processing costs associated with linguistic structures (= *difficulty*)

3. **Developmental complexity**, the order in which linguistic structures emerge and are mastered in L1 and L2 acquisition (= *development*)

Problems with polysemy

complex$_1$ structures are often more complex$_2$ and complex$_3$

=

complex structures are often more difficult and acquired late

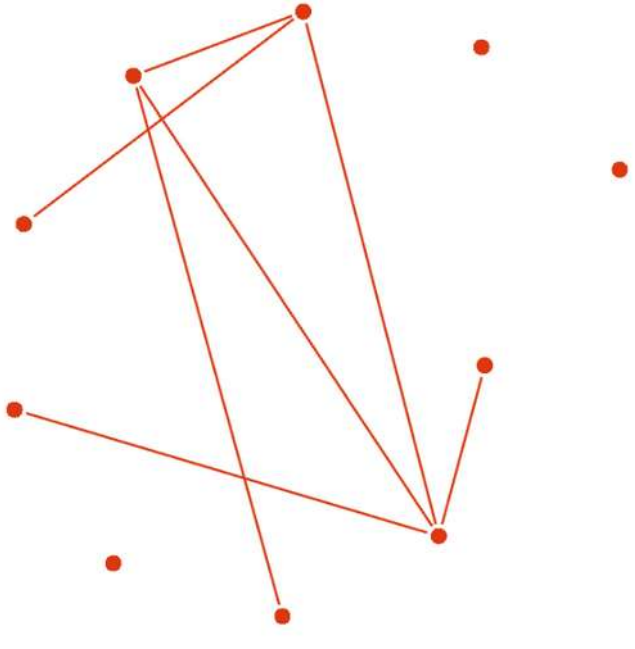
this structure is complex$_3$ because it is complex$_1$ and complex$_2$

=

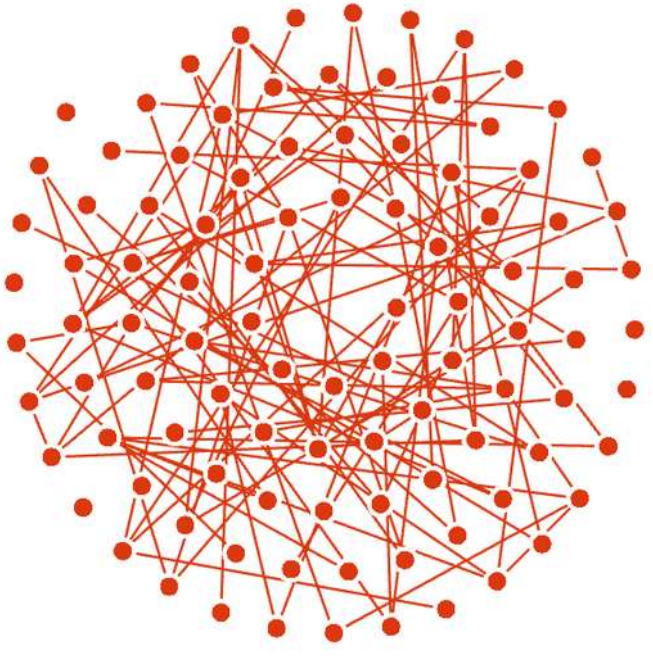
this structure is acquired late because it is complex and difficult

Complexity: a structural definition

"the number and variety of an item's constituent elements and the elaborateness of their interrelational structure". (Rescher 1998: 1)



10 nodes



100 nodes

(ElgraphoCluster)

(Structural) complexity in linguistics

We define (structural) complexity as the quantity and variety of constituents and relationships between constituents (cf. Rescher 1998). These constituents are linguistic forms, resulting from linguistic description or analysis. (Bulté, Housen & Pallotti, in preparation)

➔ Complexity may be computed on linguistic structures, systems, texts without looking at human beings.

What is often called complexity, but it isn’t

Complexity and difficulty

'agent-related complexity', that is, 'difficulty, cost, demandingness' (Dahl 2004) = 'relative complexity' (Miestamo 2008)

“cognitive difficulty reflects rather than creates complexity” (Rescher 1998: 17)
[In SLA] “structural complexity can contribute to psycholinguistic complexity or difficulty, but does not coincide with it” (Housen 2020: 391).

Structural complexity and cognitive difficulty may often be correlated in practice, but this is one more reason for using different terms for the cause (complexity) and the effect (difficulty).

Complexity → Difficulty

Difficulty/sophistication (not complexity)

complexity = “the range of forms that surface in language production and the degree of sophistication of such forms” (Ortega, 2003: 492)

“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: 124)

What is a ‘sophisticated’ form? Rare, well-chosen, hard to master, learned later, structurally complex?

Acquisitional difficulty/frequency (not complexity)

The word *tar* is not more complex than the word *car*. However, it may be more difficult to acquire, because it is less frequent.

A text containing many rare words is not more complex, but it may be more difficult to produce and understand, and thus may be produced/understood only at later developmental stages.

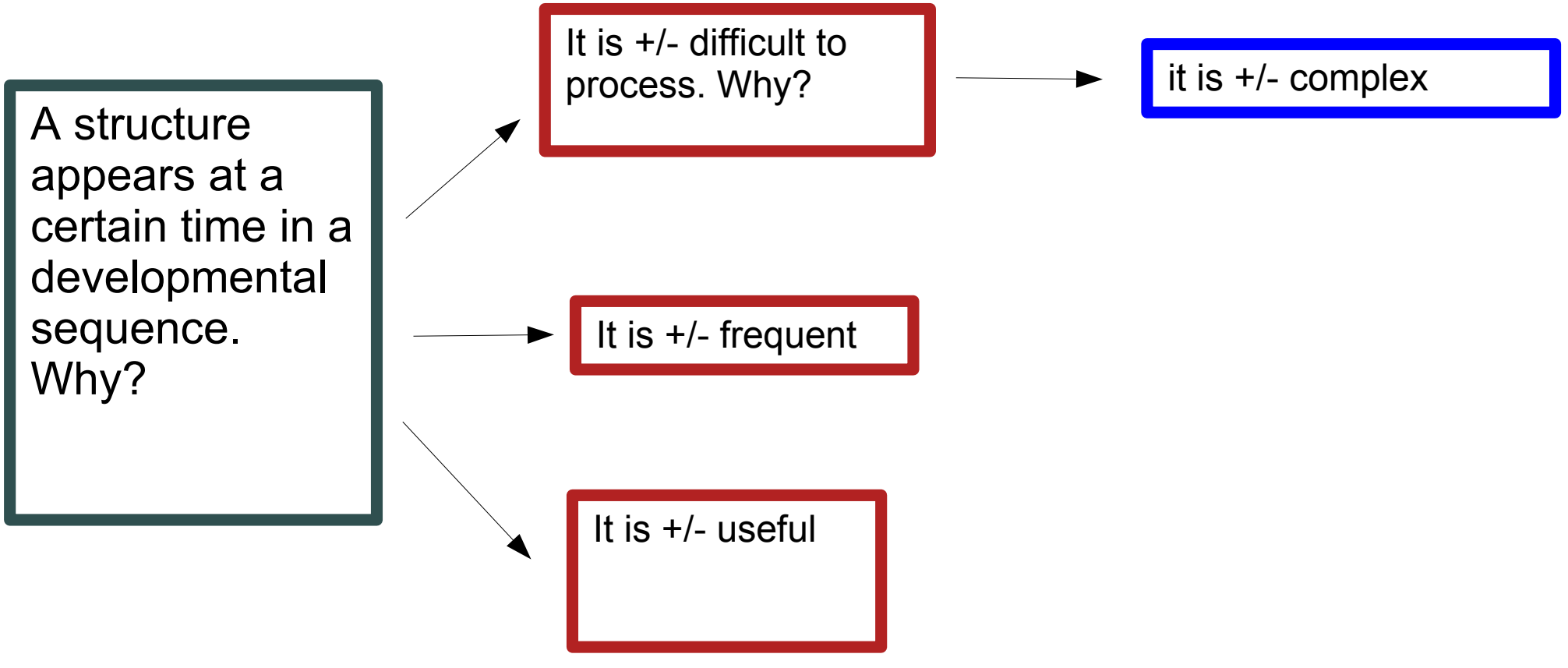
Development (not complexity)

complexity = 'the capacity to use more advanced language” (Ellis 2009)

Complexity as 'L2 acquisition difficulty' (Szmrecsanyi & Kortmann 2009), 'outsider complexity' (Kusters 2003; Trudgill 2001)

If complexity = advanced, then 'complexity grows over time' is no longer a finding, but part of the definition of complexity

All this applied to profiling



Assessing measures: conceptual validity or pragmatic utility?

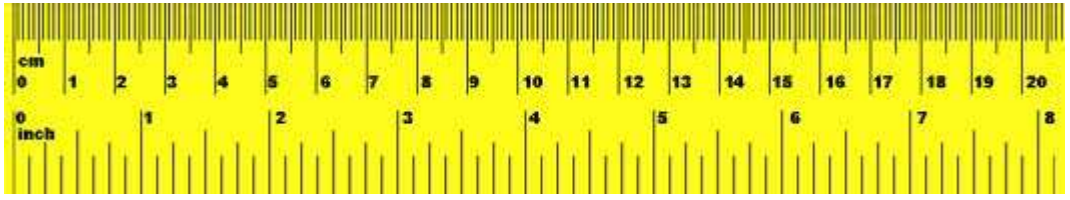
The validity of complexity resides in whether or not it correlates with a psychological reality. Hence ... while syntactic complexity can be judged through examining the degree of subordination, its **validity is corroborated by the reality that children's syntactic development follows a progression** path from less to more subordination, **as they cognitively mature**. (Han & Lew 2012: 194)

The construct of interlanguage complexity, its definition and operationalizations, and its actual measurement would be greatly refined **if what is known by now about acquisitional timing of individual second language grammars were incorporated into systematic validation programs**. (Ortega 2012:134-5)

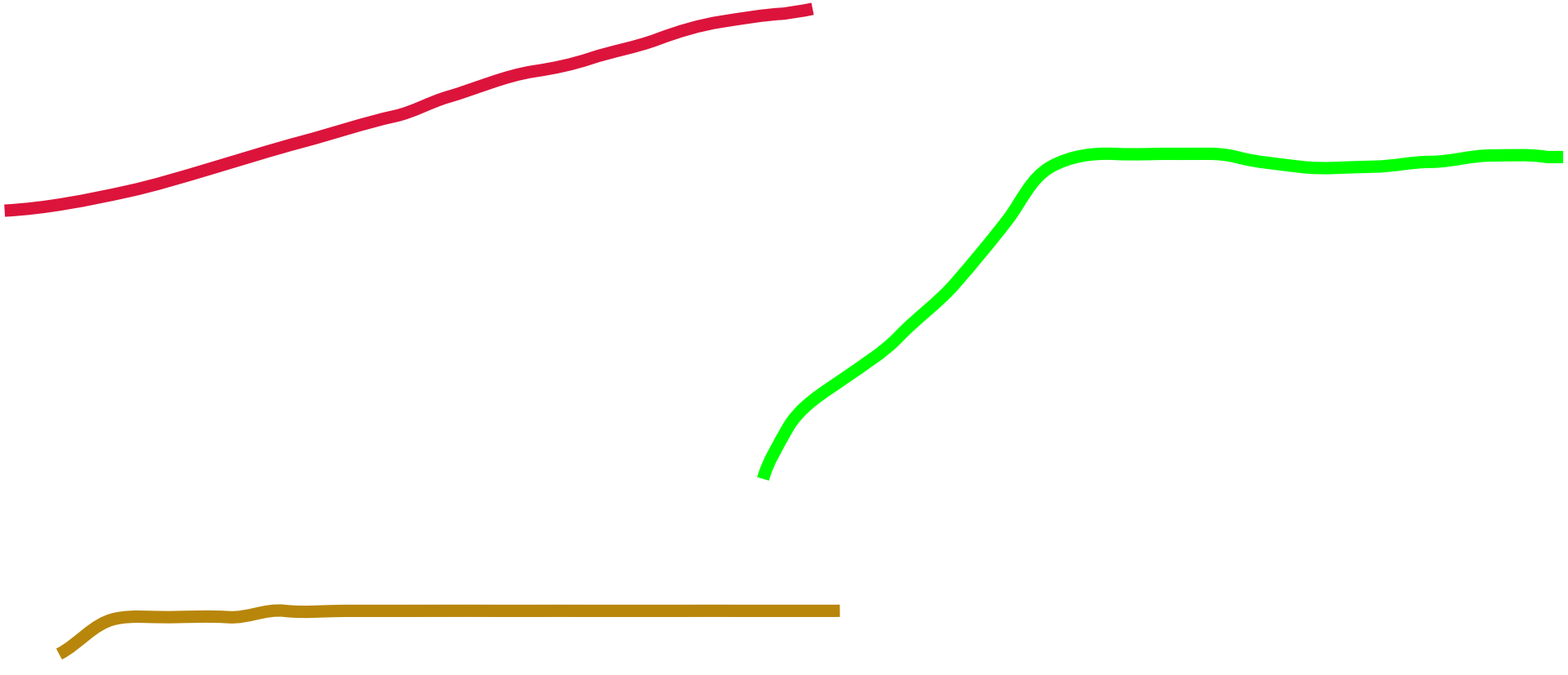
Establishing validity

(Structural) complexity is a purely descriptive notion, an ‘observable attribute’, not a ‘theoretical construct’ (Kane 2001). Its validity has to do with internal consistency and reliability, not with how it correlates with other constructs like development or cognitive processing (which are excluded from its definition). If a complexity measure correlates with development or cognitive processing, this validates the theory postulating such a correlation, not the measure itself.

(how do you validate ‘length’?)



Many measures, many profiles



Measuring morphological complexity

A simple approach to calculating a text's morphological complexity

1. LINGUISTIC ANALYSIS
- Compute the number of different inflectional forms in subsamples of N verbs

2. MATHEMATICAL ANALYSIS
- Compute variety within and across subsamples

Purely structural definition of complexity, independent of difficulty or development

Structural complexity as diversity: parallels between lexicon and morphology

Structural complexity = high diversity of types with low repetition of tokens

Lexical complexity

talk, write, drink > talk, talk, talk (or talk, talking, talks)

Morphological complexity

talk, talking, talks > talking, talking, talking (or talking, writing and drinking).

Lexical complexity: count lexemes

Morphological complexity: count morphemes

Not so easy...

Death of the morpheme?
Anderson (1992). *A-morphous morphology*.

Or alive and kicking?
In all languages, or virtually all, it is appropriate to analytically break words down into component pieces, called morphs, and then to bundle morphs back into the functional units we call morphemes (Goldschmidt 2010, in the *Handbook of Computational Linguistics*)

Morphological processes

base + exponence (process) = inflected word **form**

book → *book-s* (concatenative process)

buch → *bücher : buch* + *er* + *umlaut* (concatentative + non-concatenative process)

kitab → *kutub* (non-concatenative process)

'exponence' (Matthews 1974); 'inflectemes' (Sagot & Walter 2011)

Operationalizing ‘inflection’

Inflection = a formal process affecting a lexical base to express grammatical meaning

Problems:

1. Identifying the base

2. Describing the inflectional process (‘what happens to the base’)

Identifying the base to describe its modifications

For any verbal lexeme, the default base (DB) is defined as the base that appears in most cells of that lexeme's paradigm.

spreche sprechen sprecht (pres); spreche sprechest etc (subj); sprech sprecht etc (imp); sprechend;	sprich sprichst	sprach, sprachst sprachen sprach	gesprochen

DB = sprech-

Describing inflections

	DS	sample WF(s)	exponence
WF is identical to DS	cut	cut (present or past tense)	X
WF consists in DS + additional graphemes at the end of the DS	cut rise, take talk	cuts risen, taken talked	s n ed
WF consists in DS minus some graphological material at the end of the DS	hide	hid	£e
WF consists in DS minus some graphological material in the middle of the DS	feed lead	fed led	_£e_ _£a_
WF consists in DS + additional graphemes replacing parts of the DS at the end of the DS	buy think	bought thought	uy/ought ink/ought
WF consists in DS + additional graphemes replacing parts of the DS in the middle of the DS	find, grind drive, ride	found, ground, bound drove, rode	_i/ou_ _i/o_
multiple aspects	keep, feel break, steal swear, tear	kept, felt broke, stole sworn, torn	_£e_t _ea/o_e _ea/o_n

cp. ‘edit distance’ (Kruskal 1979)

Written morphology

Independent status of written morphology, at least in principle

However, in assessing MC in written texts, morphological and orthographic complexity shouldn't be mixed up

Possible solution: ignore all variation in written forms due to systematic orthographic processes (orthography-inflated complexity).

Oral morphology

found	/faʊnd/	_aɪ/aʊ_	proc 25
bound	/baʊnd/	_aɪ/aʊ_	proc 25
read	/red/	_i:/e_	proc 26

Forms, not cells in paradigms

Counting different formal processes only

German

*wir frag**en*** (1pl.prs.indic), *sie frag**en*** (3pl.prs.indic), *zu frag**en*** (inf) = 1 exponence

Italian

tu cant**i** (2sg.prs.indic), che lui cant**i** (3sg.prs.subj) = 1 exponence

Diversity of form-function relationships?

Same procedure, but instead of counting exponents (forms), count form-function relationships.
This can be operationalized by looking at strings encoding forms and functions as in standard morphemic transcriptions, e.g.:

- *en:*1pl.pres.indic; *en:*3pl.pres.indic; *en*:inf
- *i*:2sg.pres.indic; *i*:3sg.pres.subj

Problems
- what functional features are to be encoded? E.g. do you encode just 'present' or 'present, habitual, indicative'?
- how can one be sure of the functions of grammatical forms in an interlanguage? E.g. does -*ing* correspond to present, progressive, indicative, or just present?

Mathematical analysis

Computing morphological complexity (MC 10)

For each word-class (e.g. nouns, verbs, adjectives) draw sets of N (e.g 10) tokens

For each set, count the exponents' types (min 1 – max 10); then compute the average set-internal variety. **(6+7)/2 = 6.5**

For each set pair, count exponents that are not shared (min 0 – max 20); then compute the average between-set diversity and divide it by two. **5/2 = 2.5**

Add the set-internal diversity score to the between-set diversity score/2, then subtract 1, to arrive at a global inflectional diversity score (morphological complexity).

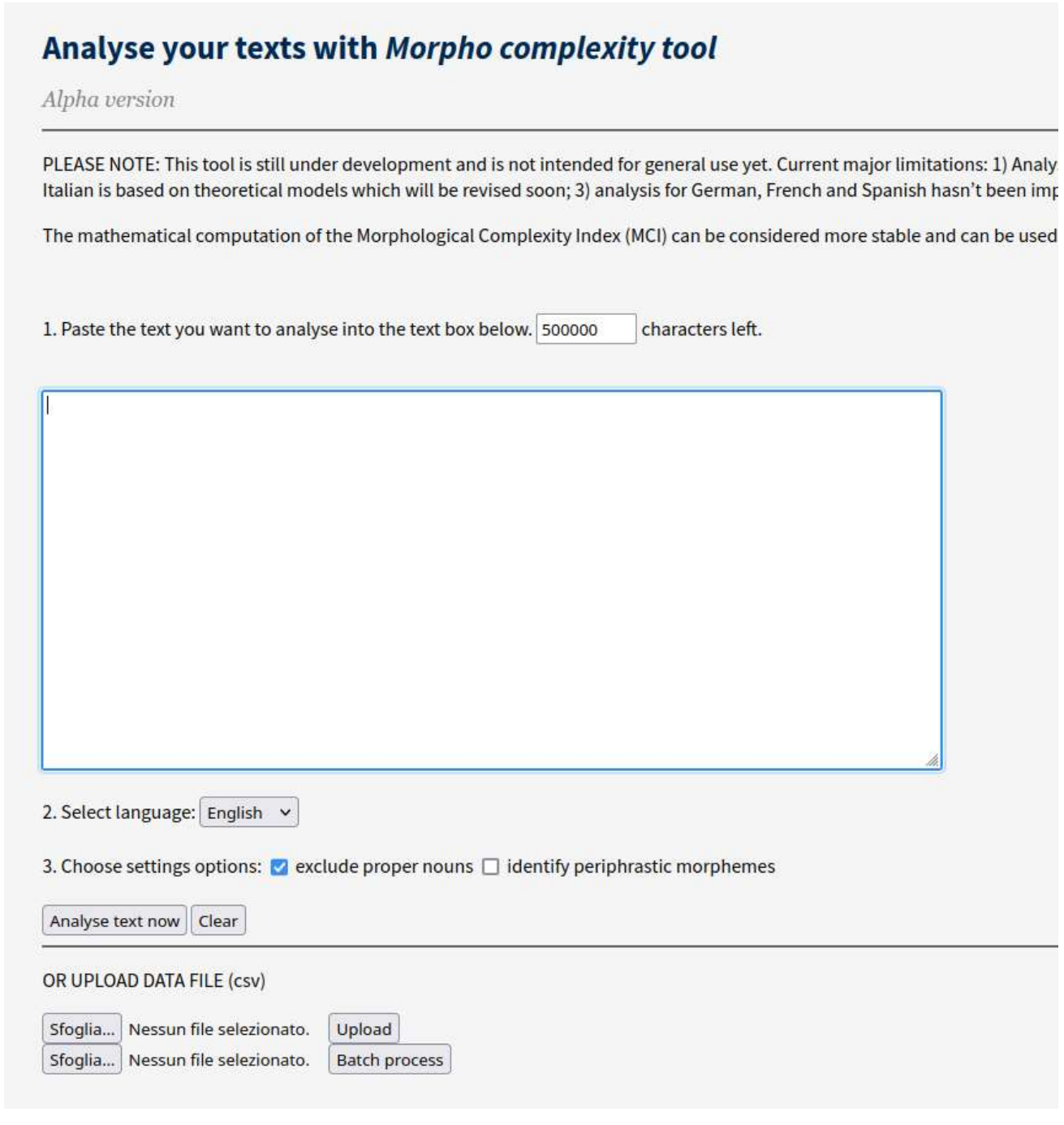
6.5 + 2.5 - 1 = 8.0 (MC10)

ed	ed
ed	**_o/a_**
_i/ou_	**went**
was	was
X	X
X	X
ing	X
are	X
are	are
are	**is**
6	7
_i/ou_, ing, _o/a_, went, is = 5	

Making it even simpler

Compute and average sample-internal variety only: MC10a

Mean Segmental Type/Token Ratio (MSTTR) applied to exponences



English, French, German, Italian, Spanish

Free online

Verb morphology only

(Brezina & Pallotti 2015)

MCI = 4.1

```
ing, ed, is, ing, is, ø, ø, ø, ø, ø, _k/d_, ed, ing, ing, is, ø, is, ø, ø, ø,
ø, ø, ø, ø, s, ing, ø, ø, ø, ø, ing, ing, ø, ø, ø, are, ing, ed, ø, ing, are,
ø, ing, ed, ø, ing, ed, ø, ø, ed, ed, is, ø, ø, ing, ø, ø, ed, ed, ø, ø
```

VERBS AND NOUNS IN TEXT

In my **opinion**, **saying** that in a **world** **dominated** by **science** **technology** and **industrialisation**, there **is** no longer a **place** for **dreaming** and **imagination** **is** false . **Imagination** and **dreams** **belong** to **mankind** and **people** **use** the technological **progress** to **live** **better** their **life** and to **realise** their **dreams**. Most of the scientific and medical **researches**, **made** by the **use** of the **science** **technology**, **realised** the **dream** of **living** in a better **world**: **today**, **using** a personal **computer**, it **is** possible to **discover** the **cause** of a **disease** and , by its **analysis**, it **is** possible to **find** its **cure** and to **save** many human **lives**. **Industralisation** and **science** **technology** after **people** the **tools** to **communicate** as quick as possible even though they **live** away from each other . The **telecomunication** **system** **reduce** **time** and **distances** among **people** of different **countries**. The **use** of an Internet **program**, for **instance**, **make** you **talk** to an unknown **citizen** who **lives** on the other **side** of the **earth**, without **losing** your **dream** or your **imagination**. You **can** always **choose** to **switch** on or off your **PC** and **go** on **living** and **dreaming** in the *traditional* **way**. Obviously , when you **use** a **computer** and **play**, for **instance**, with a virtual **game**, you **have** to **know** you **are** **using** the **imagination** of the **software** **programmer** who **designed** the **game**. But you **can***live* , at the same **time**, the **dream** of **being** into a strange **planet** where there **are** three **suns** and **moons**. The **science** **technology** and its **appliance** in the **industry** **relalized** many **dreams** of the **mankind**, such as the

MCI = 7.8

```
ed, has, brought, ø, ø, ø, ø, ing, ø, _k/d_, ø, ø, ing, ø, ed, _e/o_, _i/ou_,
_fe__t, were, ed, left, ø, ed, was, t, s, ø, ed, are, s, has, _ea/o__en, ø,
ø, is, ing, is, ing, ing, is, was, ed, ø, ø, was, ø, ø, ø, ø, ø, is, ø, ø, ø,
ø, ø, ø, were, s, is, ing, ø, ø, are, ø, ø, ø, ø, ø, ø, ø, ø, _e/t, ing, is,
is, are, n, ø, ø, ø, ø, has, brought, ing, ø, ed, ing
```

The economic **welfare** **reached** by most of the European **Countries** especially after the second **world** **war**, **has** **brought** in our **homes** all **kinds** of **comfort** which **have** **revolutioned** our **customs** and therefore our **mentality**. I **can** **start** by **mentioning** the **television**, the **washing** **machine**, the **dish** **washer**, the **telephone** and **lots** of more **things** which **have** **made** our **life** easy and comfortable and without them we **could** not **live** **today**. **Talking** to my **parents** who **have** **experienced** the **war** and that **got** married during the **sixties**, I **found** out that in those **years** **people** **felt** great **enthusiasm** towards **life** and **were** **optimists** for the **future**. **Reconstruction** and **industrialisation** **created** new **jobs** with the **consequence** that **lots** of **people** **left** the **country** to **move** to **town** which **offered** more **opportunities** especially for the young **generations**. It **was** a golden **period**, where **everybody** **dreamt** about a better future and about a better **society**. It **seems** impossible , but although we **have** now **reached** through **technology** a high **standard** of **life**, we **are** very **pessimists**. It **seems** as **progress** **has** **stolen** our **imagination** and therefore the **love** for small **things**. I **can** **give** few **examples** that such a **fact**: **television** **is** **becoming** lately the **killer** of **conversation** between **parents** and **children**; it **is** almost **disappearing** the **use** of **writing** nice **letters** to **friends**, since there **is** the **telephone**; when I **was** a **child** I **used** to **invent** **games** to **enjoy** with my **friends**. Our **imagination** **was** so vast that we **could** **play** wonderful **games** with simple **stones** or

5164	the	DT	the	
5165	United	NP	United	
5166	States	NPS	States	
5167	was	VBD	be	was
5168	n't	RB	n't	
5169	going	VBG	go	ing
5170	to	TO	to	
5171	collapse	VB	collapse	Ø
5172	if	IN	if	
5173	Chile	NP	Chile	
5174	became	VBD	become	_o/a_
5175	independent	JJ	independent	
5176	.	SENT	.	
5177	[S_END]			
5178	[S_START]			
5179	Why	WRB	why	
5180	were	VBD	be	were
5181	we	PP	we	
5182	so	RB	so	
5183	concerned	JJ	concerned	
5184	about	IN	about	
5185	it	PP	it	
5186	?	SENT	?	

1) Use TreeTagger to identify headword behind inflected word form (IWF)
2) If headword = root (e.g. English *speak* = *speak*, proceed to 4); if not,
3) derive root from headword; e.g. fr. *parler* → *parl-* ; ger. *sprechen* → *sprech-* etc.
4) if IWF can be parsed as root + material concatenated to the right, then exponence = concatenated material
5) if not, search list of irregular cases

if word is or ends in	exponent	examples
iedi	_i_i	siedi
iedo	_i_o	siedo
iedono	_i_ono	siedono
ieda	_i_a	
iedano	_i_ano	
uole	_u_e	vuole, duole, suole
uori	_u_i	muori
uore	_u_e	
asi	d/si	persuasi, dissuasi, evasi
ase	d/se	
asero	d/sero	
aso	d/so	persuaso, evaso
asa	d/sa	persuasa
vide	_e/i_e	vide
vidi	_e/i_i	

Future directions

‘Irregularities’: lexicon or morphology?

How can we draw the line between lexical and morphological complexity in order to compute them separately?

Easy cases

Talk, write, drink = lexical complexity

Talk-ing, talk-s, talk-ed = morphological complexity

Difficult cases

Found, brought, went, was and other ‘irregular verbs’ = lexical, morphological or morpholexical (morphomic) complexity?

/faɪnd/ vs /faʊnd/ (two alternating stems in the lexicon, ‘morphomes’ Aronoff 1994)

OR

_aɪ_ → _aʊ_ (introflective morphological operation, ‘minor rule’ Lakoff 1970)

Stems

Italian *prendere* ‘to take’

Present tense	Simple past	Future	Participle
prend-o	pres-i	prend-erò	pres-o
prend-i	prend-esti	prend-erai	
prend-e	pres-e	prend-erà	
prend-iamo	prend-emmo	prend-eremo	
prend-ete	prend-este	prend-erete	
prend-ono	pres-ero	prend-eranno	

prend-; pres- : 2 stems

prend-o / *cant*-o Lexical complexity
prend-o / prend-i Morphological complexity
prend- / *pres*- Morpholexical (morphomic) complexity

“Les lexèmes ne sont pas nécessairement associés à un radical unique ou privilégié, mais à une collection indexée de radicaux. Dans la tradition française, à la suite de Bonami & Boyé (2003), cette collection a été appelée ESPACE THEMATIQUE”. (Bonami & Boyé 2013:3)

Example
Nous buvons, je bois, il boit, que tu boives, ils boiront

Stem space (= espace thématique) of Italian verbs (Montermini & Bonami 2013)

	Person					
	1	2	3	4	5	6
Future indicative	S6					
Present Conditional						
Present Subjunctive		S2		S4		S2
Present Indicative			S3			
Imperfect Indicative	S1					
Imperfect Subjunctive						
Preterite Indicative	S5		S5			S5
Imperative		S3		S4		
Present Participle	S1					

Thus....

Lexical complexity: breadth of lexical space

Morphological complexity: breadth of morphological space

Morpholexical (morphomic) complexity: breadth of thematic space

Computing MCI for French with a stem-based analysis: De Clercq & Housen (2019)

Comparing different operationalizations of morphological complexity on English and Italian data: Pallotti (2021)

Complexity and accuracy

Complexity = variety of interlanguage forms (NOT: variety of correct forms)

For example

Lexical complexity: *psychologer; rainbrella* (2 lexical items)

Morphological complexity: *I catched it; two childs* (2 morphological exponents)

‘Factorization’ (Pienemann 1998)

Swedish interlanguage

singular	plural
Ø	-a

Standard Swedish

	singular				plural
	attributive		predicative		
	uter	neuter	uter	neuter	
Def.	-a	-a	Ø	-t	-a
Indef.	Ø	-t	Ø	-t	-a

Complexity and accuracy

Complexity should be assessed independently of accuracy, by computing the variety of interlanguage forms (vs variety of target-like forms)

An interlanguage may be rather complex, both lexically and morphologically, and at the same time not very accurate

Otherwise, explicitly state that one is computing ‘the complexity of accurate forms’

Thank you!

gabriele.pallotti@unimore.it