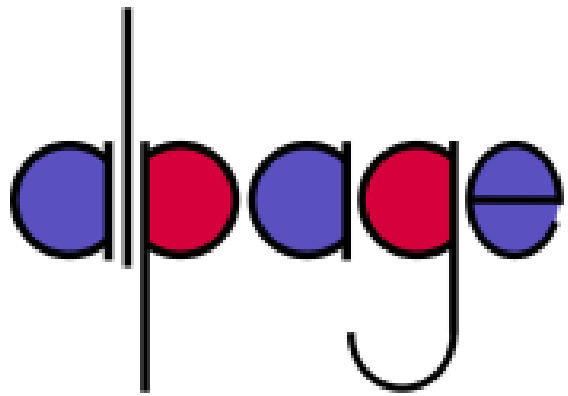
Comparing Complexity Measures

Benoît Sagot

benoit.sagot@inria.fr

Alpage — INRIA & Université Paris–Diderot
Paris, France



Computational Approaches to Morphological Complexity
Friday, February 22, 2013 — Paris, France

Outline

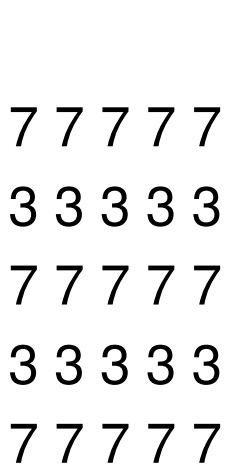
Main topic of this talk: how can we actually measure morphological complexity?

Outline

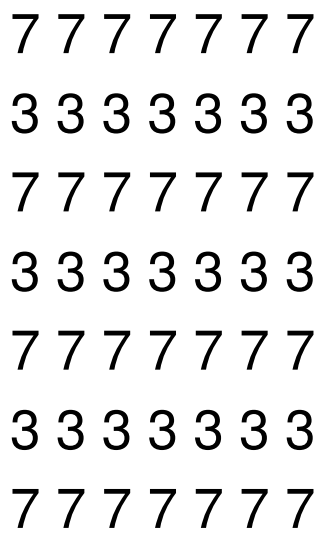
Outline

Complexity as the number of parts of the system

- First intuition, on a non-linguistic example



lower size, lower complexity



larger size, higher complexity

Complexity as the number of parts of the system

- First intuition, on a non-linguistic example

1 2 3 4 5
2 3 4 5 6
3 4 5 6 7
4 5 6 7 8
5 6 7 8 9

lower size, complexity = ?

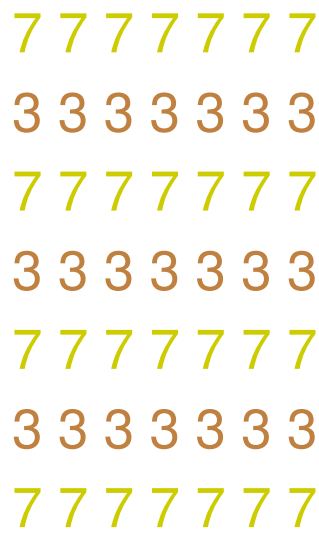
7 7 7 7 7 7 7
3 3 3 3 3 3 3
7 7 7 7 7 7 7
3 3 3 3 3 3 3
7 7 7 7 7 7 7
3 3 3 3 3 3 3
7 7 7 7 7 7 7

larger size, complexity = ?

Complexity lies in data structure

- First intuition, on a non-linguistic example

lower size, some structural complexity



larger size, less structural complexity

Complexity lies in data structure

- Yet another (non-linguistic) example

lower size, some structural complexity

larger size, no structural complexity at all (no structure!)

Complexity as the size of the rules used to generate the data

- On the same (non-linguistic) example

1 2 3 4 5
2 3 4 5 6
3 4 5 6 7
4 5 6 7 8
5 6 7 8 9

not very complex to generate

3 9 0 4 4 6 6 8 2 1 2 5 3 4 9
9 3 1 7 6 9 4 9 6 2 2 5 7 5 3
7 0 7 0 3 9 9 0 7 1 7 5 2 1 3
8 5 8 9 9 7 5 5 7 2 5 1 4 0 4
9 4 9 1 9 2 0 6 1 3 7 5 1 0 4
2 5 4 9 8 0 1 2 3 2 6 9 0 0 4

complex to reproduce exactly, not complex to reproduce if randomness is a primitive concept of the model

Complexity as the minimum size of rules that can generate the data

1 2 3 4 5
2 3 4 5 6
3 4 5 6 7
4 5 6 7 8
5 6 7 8 9

the most compact way to describe that system is more compact than the system

```
(i..i+4)[i=1..4]
```

3 9 0 4 4 6 6 8 2 1 2 5 3 4 9
9 3 1 7 6 9 4 9 6 2 2 5 7 5 3
7 0 7 0 3 9 9 0 7 1 7 5 2 1 3
8 5 8 9 9 7 5 5 7 2 5 1 4 0 4
9 4 9 1 9 2 0 6 1 3 7 5 1 0 4
2 5 4 9 8 0 1 2 3 2 6 9 0 0 4

the most compact way to describe that system is likely to be the sequence itself

Complexity as the minimum size of rules that can generate the data

1 2 3 4 5
2 3 4 5 6
3 4 5 6 7
4 5 6 7 8
5 6 7 8 9

the most compact way to describe that system is more compact than the system

maybe not if the description has to be written in English rather than in some programming language

3 9 0 4 4 6 6 8 2 1 2 5 3 4 9
9 3 1 7 6 9 4 9 6 2 2 5 7 5 3
7 0 7 0 3 9 9 0 7 1 7 5 2 1 3
8 5 8 9 9 7 5 5 7 2 5 1 4 0 4
9 4 9 1 9 2 0 6 1 3 7 5 1 0 4
2 5 4 9 8 0 1 2 3 2 6 9 0 0 4

the most compact way to describe that system is likely to be the sequence itself

Complexity as how predictable is a new instance of the data

1 2 3 4 5

2 3 4 5 6

3 4 5 6 7

4 5 6 7 8

5 6 7 8 9

easy, once the structure is understood

3 9 0 4 4 6 6 8 2 1 2 5 3 4 9

9 3 1 7 6 9 4 9 6 2 2 5 7 5 3

7 0 7 0 3 9 9 0 7 1 7 5 2 1 3

8 5 8 9 9 7 5 5 7 2 5 1 4 0 4

9 4 9 1 9 2 0 6 1 3 7 5 1 0 4

2 5 4 9 8 0 1 2 3 2 6 9 0 0 4

impossible

Outline

Information-theoretic measures

These ideas have been formalized within *Information Theory*

- Minimum size of rules for generating the data (e.g., smallest possible description of the data): **Kolmogorov Complexity** (Solomonoff 1964, Kolmogorov 1965)
  - *how "random" is an individual data instance?*
  - any reasonably expressive formal description language is fine
  - in some cases, there is no way to prove that a given description is the shortest one
  - cannot always be computed exactly: it has to be *approximated*
  - Structure is captured within the description of the data
- If one considers (and model) the system as "emitting" data instances, the amount of uncertainty expected in a new data instance is **Shannon's entropy** (Shannon, 1948)
  - *how random is an entire collection of data instances overall?*
  - modeling the system requires encoding the data as a sequence of independent and identically-distributed random variables according to a certain probabilistic model, which is difficult in practice
  - Structure is captured in the way we model the system

Information-theoretic measures

- There are deep relations between Shannon's entropy and the Kolmogorov complexity
  - in nicely large-scale conditions, they sort of give approximately the same results
- The difference between the two measures can be illustrated by the following example:
  - let us consider a source of data, known to produce only 2 data instances, both being very complex
  - the Kolomogorov complexity (of each instance) is very high
  - the Shannon entropy (of each instance) is at most 1, i.e., very low

Complexity as the minimum size of rules that can generate the data

- As such, the Kolmogorov Complexity is not necessarily a sound estimator of the intuitive notion of complexity

7 9 3 2 3 8 4 6 2 6 4 3 3 8 3
2 7 9 5 0 2 8 8 4 1 9 7 1 6 9
3 9 9 3 7 5 1 0 5 8 2 0 9 7 4
9 4 4 5 9 2 3 0 7 8 1 6 4 0 6
2 8 6 2 0 8 9 9 8 6 2 8 0 3 4
8 2 5 3 4 2 1 1 7 0 6 7 9 8 2

90 decimals of π, ignoring the first 12

low Kolmogorov complexity

3 9 0 4 4 6 6 8 2 1 2 5 3 4 9
9 3 1 7 6 9 4 9 6 2 2 5 7 5 3
7 0 7 0 3 9 9 0 7 1 7 5 2 1 3
8 5 8 9 9 7 5 5 7 2 5 1 4 0 4
9 4 9 1 9 2 0 6 1 3 7 5 1 0 4
2 5 4 9 8 0 1 2 3 2 6 9 0 0 4

the same random sequence as before

huge Kolmogorov complexity

```
a[52514],b,c=52514,d,e,f=1e4,g,h;main()
{for(;b=c-=14;h=printf("%04d",e+d/f))}
for(e=d%=f;g=--b*2;d/=g)d=d*b+f*(h?a[b]:f/5),a[b]=d%--g;}
```

Complexity as the minimum size of rules that can generate the data

- As such, the Kolmogorov Complexity is not necessarily a sound estimator of the intuitive notion of complexity

7 9 3 2 3 8 4 6 2 6 4 3 3 8 3
2 7 9 5 0 2 8 8 4 1 9 7 1 6 9
3 9 9 3 7 5 1 0 5 8 2 0 9 7 4
9 4 4 5 9 2 3 0 7 8 1 6 4 0 6
2 8 6 2 0 8 9 9 8 6 2 8 0 3 4
8 2 5 3 4 2 1 1 7 0 6 7 9 8 2

90 decimals of π, ignoring the first 12

low Kolmogorov complexity

3 9 0 4 4 6 6 8 2 1 2 5 3 4 9
9 3 1 7 6 9 4 9 6 2 2 5 7 5 3
7 0 7 0 3 9 9 0 7 1 7 5 2 1 3
8 5 8 9 9 7 5 5 7 2 5 1 4 0 4
9 4 9 1 9 2 0 6 1 3 7 5 1 0 4
2 5 4 9 8 0 1 2 3 2 6 9 0 0 4

the same random sequence as before

huge Kolmogorov complexity

```
a[52514],b,c=52514,d,e,f=1e4,g,h;main()
{for(;b=c-=14;h=printf("%04d",e+d/f))}
for(e=d%=f;g=--b*2;d/=g)d=d*b+f*(h?a[b]:f/5),a[b]=d%--g;}
```

Complexity as the minimum size of rules that can generate the data

- Our intuition on complexity is related to the amount of structure we *percieve* in the system
  - There is an underlying structure and order in the decimals of π, but we do not percieve it as such
  - the Kolmogorov Complexity catches such unwanted structure
- The solution lies in the way we will approximate the Kolmogorov complexity
  - These approximations have properties that make them more suitable than the Kolmogorov complexity itself

Complexity as the minimum size of rules that can generate the data

There are two standard ways to approximate the Kolmogorov Complexity

- **1. Lossless Compression [LC]**
- **2. Description Length [DL]**

Complexity as the minimum size of rules that can generate the data

- **1. Lossless Compression [LC]**
  - You use it every day. E.g., LZ77, by Lempel and Ziv (1977) is used for `zip` files; LZW, by Lempel, Ziv and Welch (1984), refinment of LZ77, is used in the GIF image format
  - The underlying idea is simple: the Kolmogorov Complexity of your data can be approximated (as closely as required) based on the size of a compressed version of it (it provides an upper bound)
  - *(some of) the structure in the data are captured automatically*, as they are the basis for the compression algorithm

Complexity as the minimum size of rules that can generate the data

- **2. Description Length [DL]**
  - We restrict ourselves to a subset of all possible descriptions, by chosing a *formalism* that licenses only some possible descriptions
  - We define a *code*, that is able to encrypt these descriptions in as optimized a way as possible, *leveraging all possible structural knowledge these descriptions contain*
  - The information content of the description is then estimated as the product of its entropy and its length
  - The Kolmogorov Complexity is approximated by the information content of the description that has the lowest information content (Minimum Description Length paradigm, Rissanen 84)

- In both cases, structural knowledge is exploited: structure within the data, and with DL structure within the description of the data

Complexity as the minimum size of rules that can generate the data

- In such a framework, measuring complexity means measuring how much information is required for describing the data, given a sound inventory of the relevant types of structure we project on the data
  - Given a description (a model), this amount of information is its **compactedness**
  - Compactedness is formalism-dependent: the more relevant structure a formalism can capture, the more relevant the compactedness measure
- The compactedness of the most compact description among all descriptions licenced by the formalism is the **complexity**
  - Complexity is formalism-dependent as well
- We shall denote such complexity measures as **description-based complexity measures**

Complexity as the unpredictability of new instances of the data

- Another way to view complexity is to rely on predictability rather than on compactedness
- Shannon's entropy measures the unpredictability of new instances of the data
  - Less predictability is interpreted as a higher complexity
  - The way we build a probabilistic model of the source is crucial
  - On data generated by a complex system, one component of the system can be isolated if its contribution can be erased
    - One can compare the entropy of the original data with the entropy of a version of the data after the contribution of the component has been removed
- Complexity measures based on Shannon's entropy will be denoted as **entropy-based complexity measures**

Outline

What do we want to compute the complexity of?

- Assessing morphological complexity requires defining explicitely what inflectional morphology exactly covers:
  - do we want to include phonological phenomena within the scope of morphlogical complexity? rather not
  - do we want to include morphonological phenomena? unclear
  - do we want to include periphrastic forms? probably not
  - do we have a clear view about how to distinguish inflectional morphology (to be included) from derivational morphology (not to be included)?

On which data will we try to measure morphological complexity?

We need to define the system and how we represent it: what will our data be?

- data = the “language”
  - we would then measure the intrinsic complexity of a language’s inflectional morphology (the *system*) — this looks difficult
- data = a corpus
  - we would then perform entropy measures on the corpus, e.g., assessing the contribution of morphology to the information conveyed by the corpus
- data = the lexicon
  - we would then perform entropy measures within the lexicon, or complexity measures concerning a particular model of the language’s morphological lexicon

What kind of complexity measures will we use?

Many authors have introduced and/or used morphological complexity measures, that cover all the range of possible approaches sketched above

- Counting-based complexity measures
- Entropy-based complexity measures
  - Data = lexicon (e.g., data instance = form)
  - Data = corpus (e.g., data instance = sentence)
- Description-based complexity measures
  - Data = lexicon (e.g., description = morphological grammar + morphological lexicon)
  - Data = corpus (e.g., description = decomposition of the corpus into sequences of morphs + inventory of these morphs)

Panorama of complexity measures

	Data = corpus	**Data = lexicon**
Counting-based	(McWorther, 2001)	
Entropy-based	(Moscoso del Prado Martín, 2004, 2010) (Pellegrino et al., 2007, 2010)	(Ackerman, Blevins, Maalouf, 2009) (Malouf & Ackerman, 2010) (Bonami et al., 2011)
Description-based	(Juola, 1998)	(Bane, 2008) (Sagot & Walther, 2011)

Panorama of complexity measures

	Data = corpus	**Data = lexicon**
Counting-based	(McWorther, 2001)	
Entropy-based	(Moscoso del Prado Martín, 2004, 2010) (Pellegrino et al., 2007, 2010)	(Ackerman, Blevins, Maalouf, 2009) (Malouf & Ackerman, 2010) (Bonami et al., 2011)
Description-based	(Juola, 1998)	(Bane, 2008) (Sagot & Walther, 2011)

Outline

Counting morphological feature values

- This might be useful but it is hardly a principled way to estimate complexity:
  - are these features well-defined for a given language? (how many cases in Russian? is there a vocative in Slovak? does French have cases?)
  - are these features comparable accross languages? (can one compare the number of cases in Latin and in Hungarian?)
  - how can we compare figures for different features? (how many cases are worth one gender less?)
  - which features should we select? (named “indicators” by Shosted, 2006)
- approach used, e.g., by McWorther (2001), although not explicitely quantitatively, Bickel & Nichols (2005) or Shosted (2006)

Counting morphological feature values

- On two examples (WALS data)

Language	French	Russian
Number of Genders	2	3
Number of Cases	none	6–7

- Given these features, French (nominal) inflectional morphology is less complex than Russian

Counting morphological feature values

- On two examples (WALS data)

Language	Hungarian	Swahili
Number of Genders	none	$\geq$5
Number of Cases	$>$10	none

- The comparison doesn't work any more
- Anyway, we have no estimation of any kind of complexity
  - this would require at least weighting each feature, but how?
- Counting inflection classes or the number of cells in a paradigm is inadequate, for the same reasons

Outline

The Alexina framework and the Le*fff* lexicon

- From now on, we shall focus on French verbal data
- Alexina (Sagot 2010) is a framework for modelling and acquiring lexical information, on both the morphological and the syntactic levels (valence...)
- The Le*fff* (Sagot et al., 2006; Sagot 2010) [Lexique des Formes Fléchies du Français] is a free Alexina lexicon for French
  - It is used in various existing NLP tools (taggers, lemmatizers, parsers...)
  - large-scale: 6,825 unique verbal lemmas (>360,000 infl. forms)
  - includes (among other) a morphological grammar of French inflection, that relies on:
    - parts of the **Parsli morphological formalism (Walther, 2011)**, including many global factorization devices
    - local factorization devices (inflection class sub-typing...)
    - morphographemic rules (if needed)
  - 3 other descriptions (grammar + associated lexicon) have been developed and compared (Sagot & Walther, 2011)

Four different descriptions of French verbal inflection

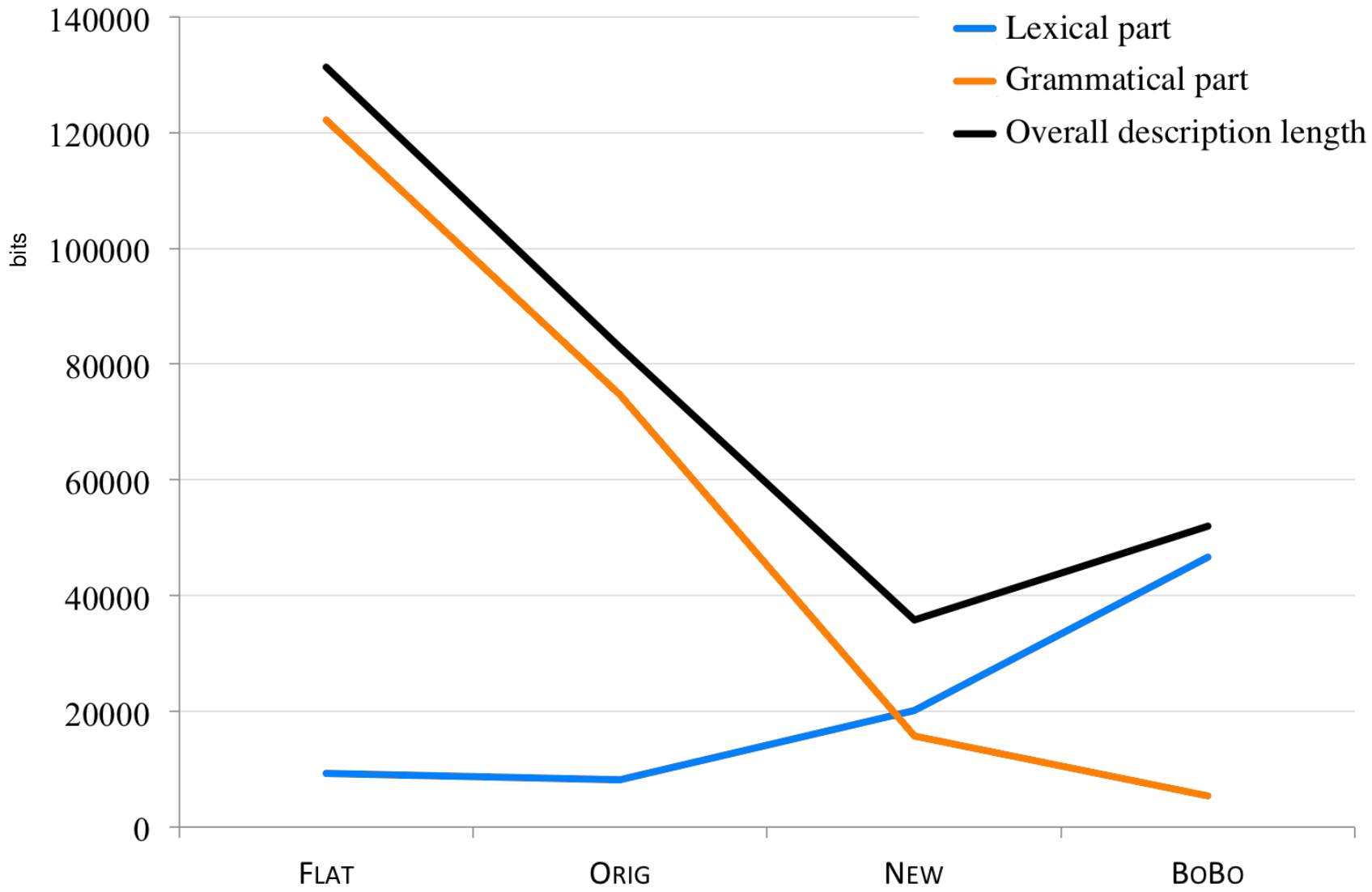
- FLAT: no factorization, no morphographemic rules
- ORIG: the “official” morphological description in the Le*fff*
- NEW: a new description that uses more Parsli concepts
- BoBo: our personal implementation of Bonami & Boyé (2001)’s stem-space-based analysis

Description name	stems	morpho -graphemic rules	factorization devices	Parsli notions	inflection classes	lexical entries
FLAT	1	no	no	no	139	simple
ORIG	1	yes	yes	no	92	simple
NEW	1–12	yes	yes	yes	20	medium
BoBo	12	no	yes	some	1	rich

Description length for structured descriptions

- Standard Description Length approaches encode each description using a set of symbols (an “alphabet”) in a way that takes advantage of the description’s structure
- We decided to breakdown these symbols in several alphabets:
  - we take even more advantage of the description’s structure
  - we can assess the contribution of each alphabet to the overall description length
  - We can separate the contribution of the lexicon from that of the morphological grammar
  - We could spot (and remove) the contribution of morphographemic rules

Complexity of our 4 descriptions of French verbal inflection



Complexity of our 4 descriptions of French verbal inflection

- This way to compute compactedness provides comparative information about the overall compactedness of several descriptions, as well as about the contribution of, e.g., the lexical vs. the grammatical part
- Parsli-based notions allow for capturing more structure within the descriptions, and this improves compactedness
- Still, this measure is an approximation (it depends on the coding, and finding the optimal encoding is hard)
- Moreover, we might want not to depend on a particular way to formalize morphology
  - This is what entropy-based complexity measures on the lexicon have been designed for

Outline

The Paradigm Cell Filling Problem (PCFL)

- Ackerman, Blevins & Maalouf (2009) and Malouf & Ackerman (2010) formulate the complexity of the lexicon using the following task:

  Given exposure to an inflected wordform of a novel lexeme, what licenses reliable inferences about the other wordforms in its inflectional family?

- They measure the reliability of implicative patterns for guessing one cell from another one using **conditional entropy**
  - I.e., how much information is needed for knowing how to fill the target cell, in addition to knowing how the source cell is filled?
- They measure the overall **paradigm entropy** as the average of all such conditional entropies

The Paradigm Cell Filling Problem (PCFL): first issues

- Bonami, Boyé & Henri (2010) identify 4 issues with this approach, corresponding to issues identified earlier (rephrasing is ours)
  1. type frequency is ignored, each inflection class counts just as much as the others
  2. the way we build the inventory of inflection classes affects the results
  3. Ackerman et al. use segmented data, e.g., the boundary between stem and suffix is considered known
  4. phonology and morphonology is embedded in the data, so we do not measure only morphological complexity

Type frequency

- Bonami, Boyé & Henri (2011) modify the approach by taking type frequency in the lexicon into account
  - the weight of each inflection class is not 1, it is the number of lemmas that use this inflection class
- Another (better?) way to measure type frequency is to take corpus data into account
  - the weight of each inflection class is not the number of lemmas that use this inflection class, rather, each lemma contributes as much as its frequency in a corpus

    In our case, the corpus will be the French TreeBank (Abeillé et al. 2002) [FTB]
- Results on Le*fff* verbal data (as for all following tables):

PARADIGM ENTROPY	**All verbs [FR]**	**1st grp verbs only [FR1]**
Manually segmented data		
Type freq. ignored	0.72	0.53
Type freq. from the lexicon	0.17	0.11
Type freq. from the FTB	0.25	0.16

Which inflection classes?

- First modification: ignoring rare inflection classes ($<$ 10 members), i.e., we retain “regular” verbs

PARADIGM ENTROPY	**FR**	**FR1**
Manually segmented data		
Type frequency ignored	0.72	0.53
Type frequency from the lexicon	0.17	0.11
Type frequency from the FTB	0.25	0.16
	“Regular” verbs [FRreg]	**FR1reg**
Manually segmented data		
Type frequency ignored	0.46	0.53
Type frequency from the lexicon	0.14	0.11
Type frequency from the FTB	0.27	0.16

Which inflection classes?

- Second possible modification: change the description, and measure on segmented data
- We can try and build inflection classes automatically
  - ongoing work (applied on Greek), that supports stem alternation
  - here, though, drastic simplification: the longest common initial substring between all forms in a paradigm is considered as the "stem"
  - resulting "suffixes" define inflection classes

PARADIGM ENTROPY	**FR**	**FR1**
Manually segmented data		
Type frequency ignored	0.72	0.53
Type frequency from the lexicon	0.17	0.11
Type frequency from the FTB	0.25	0.16
Automatically segmented data		
Type frequency ignored	0.61	0.13
Type frequency from the lexicon	0.14	0.08
Type frequency from the FTB	0.25	0.08

Non-segmented data

- "Speakers don't see morph boundaries" (Bonami et al. 2011)
  - *sort-ir* > *sort-ais* / *amorti-r* > *amorti-ssais*
- we can compare our measures with those computed by Bonami et al. (2011), who used a machine learning technique for abstracting away rules from any segmentation

  → automatically segmented data

	FR
Manually segmented data (FR)	
Type frequency ignored	0.72
Type frequency from the lexicon	0.17
Automatically data (Bonami et al.)	
Type frequency ignored	0.68
Type frequency from the lexicon	0.74

Shall we ignore morphonology?

- Each of our descriptions includes morphonological rules (rather, morphographemic rules)
- We can apply them or not before measuring complexity

	FR	**FR1**
Manually segmented data, with morphonology		
Type frequency ignored	0.72	0.53
Type frequency from the lexicon	0.17	0.11
Type frequency from the FTB	0.25	0.16
Manually segmented data, without morphonology		
Type frequency ignored	0.65	0.25
Type frequency from the lexicon	0.08	0.03
Type frequency from the FTB	0.16	0.02

Two deeper issues

- We have no idea about which one of these measures is the most sound one
- But there are two deeper issues:
  - Paradigm size has been ignored for now
  - This is not at all a formalism-free measure, as the formalization defines the set of possible operations for guessing one form from another

Conditional entropy depends on the formalization

- The way we model possible transformations from one stem to another do matter
- Imagine a language with a morphology that would have the following properties
  - All stems have the form C_1VC_2
  - Only two forms per paradigm: C_1VC_2 and $C_1VC_2VC_2$
- If the formalism only allows for modeling concatenative affixation, entropy will be very high
- If the formalism allows for identifying what the last syllable is, and includes a reduplication mechanism, entropy will be 0
- Therefore, conditional entropies heavily depend on the way we formalize morphological operations
- Cf. the earlier discussion: we need to be able to capture as much *relevant structure* as possible *before* we can compute complexity metrics
- We need formalisms that capture these relevant structures

Paradigm size does matter

- Let us consider 1st-group verbs [FR1] (productive class, “regular” inflection)
- Let us consider SimpleFrench, an artificial language extracted from French 1st group verbs, retaining only 2 cells:
  - infinitive
  - indicative present 1st person plural
- cf. Mauritian Creole (Bonami, Boyé & Henri 2011)

	FR1	**SimpleFrench**
Manually segmented data, with morphonology		
Type frequency ignored	0.53	0.80

- In other words: our simplified French is more “complex” than FR1 according to this measure!

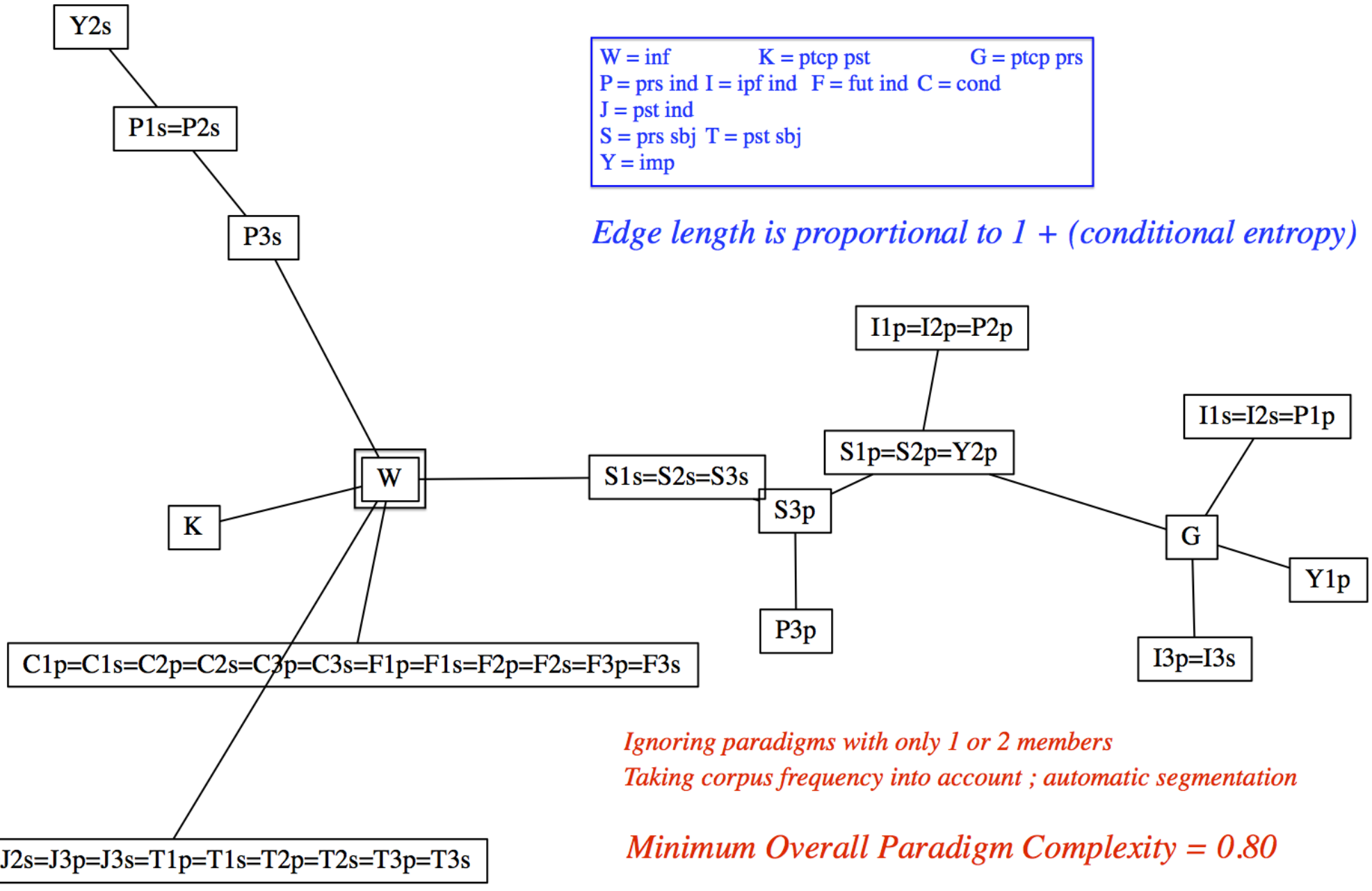
Paradigm size does matter

- Where does this apparent inconsistency come from?
  - more cells to guess, with many easily inter-predictable cells (and others which are not-inter-predictable) lead to a **low** average conditional entropy $\rightarrow$ lower complexity
  - only two cells to guess, but less inter-predictable, lead to a **higher** average conditional entropy $\rightarrow$ higher complexity
- The average of all conditional entropies measures something — arguably something interesting —, but not the complexity we are looking for

Towards a new complexity measure

- Back to the work by Bonami et al. (2011)
  - they find that Maurician Creole verbs (2 cells, low inter-predictability) exhibit a higher complexity than French verbs
  - it might be an artefact from the difference in paradigm size
- We need a way to abstract from paradigm size
  - We have built a complete directed graph that relates each cell to all other cells, each edge being weighted by the conditional probability of the target cell given the source cell
  - We used the Chu-Liu-Edmonds algorithm for extracting the best spanning tree, i.e., the globally optimal internal structure of the set of cells that has the shape of a tree
  - The **sum** (and not take the **average**) of all conditional entropies of the edges retained in the best spanning tree seems to be a more reasonable complexity measure
- This could be performed, among other settings, on non-segmented data or on segmented data prior to applying morphographemic rules
- We call this measure **Minimum Overall Paradigm Complexity**

Towards a new complexity measure



W = inf K = ptcp pst G = ptcp prs
P = prs ind I = ipf ind F = fut ind C = cond
J = pst ind
S = prs sbj T = pst sbj
Y = imp

Edge length is proportional to 1 + (conditional entropy)

Ignoring paradigms with only 1 or 2 members
Taking corpus frequency into account ; automatic segmentation

Minimum Overall Paradigm Complexity = 0.80

Towards a new complexity measure

- Removing sequencially cells that are fully predictable from another (not yet removed) one does not affect MOPC
  - it affects Ackerman et al.'s paradigm complexity

MOPC	**FR1**	**SimpleFrench**
Automatically segmented data, with morphonology		
Type frequency ignored	0.95	0

Towards a new complexity measure

- *Results should be reproduced on non-segmented data*

MOPC	**FR**	**FRreg**
Automatically segmented data, with morphonology		
Type frequency ignored	5.20	1.88
Type frequency from the lexicon	0.48	0.23
Type frequency from the FTB	1.03	0.46

- Using type frequency information from the lexicon, we need on average 0.5 bit of information for knowing its full paradigm (provided we have it segmented. . .)
- When we compared the compactedness of various morphological descriptions using Kolmogorov complexity, we have shown how morphological information has to be balanced between the lexical and grammatical components of a description
- Same here: if more complex word (trans)formation mechanisms are available (in a “grammatical” component), then less information need be associated with each lexical entry

Outline

Conclusion

- One should care a lot about what is measured and how
- Different morphological complexity measures can provide valuable information about different aspects of morphological complexity
- We measure the complexity of a model of morphology, the result therefore depends on the underlying formalization
- Complexity is about capturing structural information, hence again the importance of formalization
- Lexicon-based measures can only rely on large-scale lexicons and ideally on corpus frequencies
- Corpus-based measures are interesting as well
  - They could help understanding how morphological complexity is related to the interaction between morphology and syntax
  - They probably do not measure the same thing: the complexity of (a model of) morphology “by itself” is one thing, the complexity of morphology within the whole language system is another thing

Future work

More work to come, on (among other ideas):

- Finding deterministic (hence automatizable) *and* cross-linguistically operational ways to define paradigms
  - realizational aspects: non-concatenative phenomena, stem alternations...
  - structural aspects: heteroclisis, deponency, suppletion... $\rightarrow$ relationship between canonicity and complexity
- Understanding better the role of morphonology when measuring morphological complexity
- Developing more satisfying morphological complexity measures, leveraging what has been achieved until today
- Using complexity measures to compare various descriptions of a same language, or even to generate the "best" one
- Understanding whether comparing complexity measures accross language actually makes sense, and, if so, under which conditions

Thanks!