Segmentation in morphology: wh-en, wh-ere, and how?

Sacha Beniamine Olivier Bonami

Université de Paris, LLF, CNRS

4th AIMM — Stony Brook, May 2019

The place of segmentation in morphology

- Much morphological activity presupposes a morphological segmentation:
  - Formal grammars of all stripes
  - Typology, from Greenberg (1954) to Corbett (2007) and counting
  - Glossing conventions
  - ...
- Surprisingly little explicit contemporary discussion of the issue (Spencer, 2012).
- Not clear that uniform strategies are applied accross languages, analysts, or even subparts of a dataset.
- Intense work on stem alternations since Aronoff (1994) highlighted disagreements on stem-affix boundaries, but did not lead to any resolution.

Why is segmentation hard? I

- In a canonical inflection system (Corbett, 2007):
  - Stems are constant across paradigms.
  - Exponents are constant across lexemes.
- Because of this, canonical systems are easily segmentable:

	1SG	2SG	3SG	1PL	2PL	3PL
FICAR	ˈfiku	ˈfikɐʃ	ˈfikɐ	fiˈkɐmuʃ	fiˈkaiʃ	ˈfikɐ̃ũ
ENTRAR	ˈẽtɾu	ˈẽtɾɐʃ	ˈẽtɾɐ	ẽˈtɾɐmuʃ	ẽˈtɾaiʃ	ˈẽtɾɐ̃ũ
TENTAR	ˈtẽtu	ˈtẽtɐʃ	ˈtẽtɐ	tẽˈtɐmuʃ	tẽˈtaiʃ	ˈtẽtɐ̃ũ

Indicative present of 3 European Portuguese 1st conjugation verbs

  - Stems are longest substrings across rows.
  - (Combinations of) exponents are longest substrings across columns.

Why is segmentation hard? II

- In real systems there is typically a remainder:

	1SG	2SG	3SG	1PL	2PL	3PL
FICAR	ˈfiku	ˈfikɐʃ	ˈfikɐ	fiˈkɐmuʃ	fiˈkaiʃ	ˈfikɐ̃ũ
VIVER	ˈvivu	ˈvivəʃ	ˈvivə	viˈvemuʃ	viˈvɐiʃ	ˈvivɐ̃ĩ
IMPRIMIR	ĩpˈɾimu	ĩpˈɾiməʃ	ĩpˈɾimə	ĩpɾiˈmimuʃ	ĩpɾiˈmiʃ	ĩpˈɾimɐ̃ĩ

Indicative present of 3 European Portuguese fully regular verbs

  - Depending on language and theoretical inclination, the remainder is analyzed as part of a stem allomorph, (part of) a suffix, a thematic vowel/element/affix, etc.

- Note that the remainder is not always peripheral:

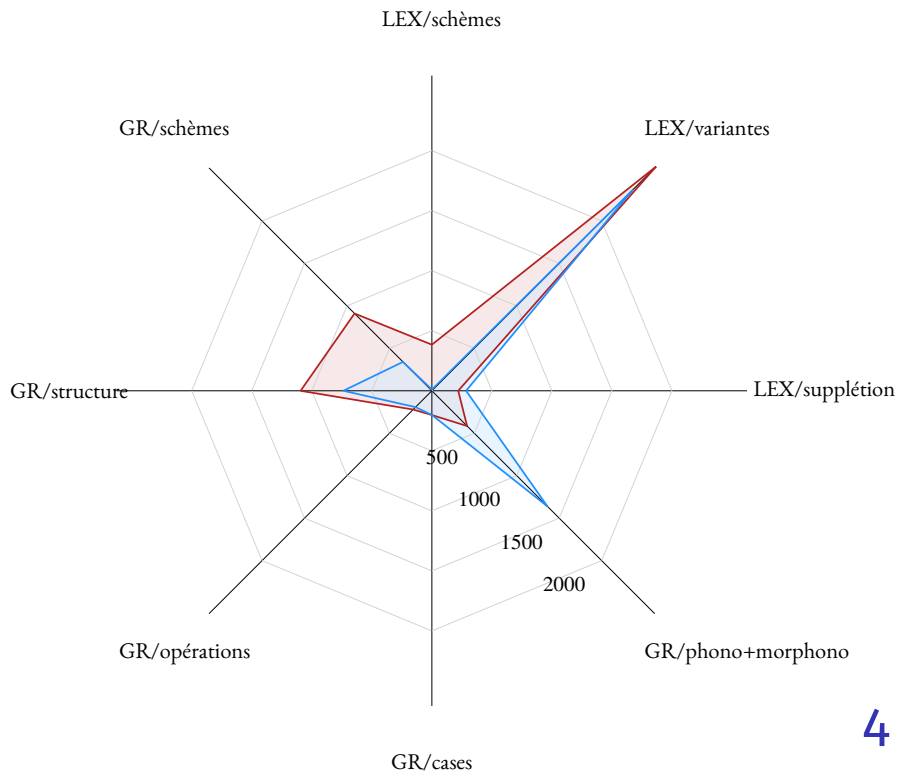
	1SG	2SG	3SG	1PL	2PL	3PL
CHEGAR	ˈʃegu	ˈʃegɐʃ	ˈʃegɐ	ʃəˈgɐmuʃ	ʃəˈgaiʃ	ˈʃegɐ̃ũ
COMEÇAR	kuˈmɛsu	kuˈmɛsɐʃ	kuˈmɛsɐ	kumə̥ˈsɐmuʃ	kumə̥ˈsaiʃ	kuˈmɛsɐ̃ũ

Stress-conditioned vowel alternations

Solution 1: informed choice

- Walther, 2013:
  (see also Walther and Sagot 2011; Sagot and Walther 2013)
  - One just needs to measure which segmentation scheme leads to the most satisfactory description of the system.
  - This can be seen as an empirical question, using information-theoretic measures of complexity and the Minimum Description Length principle.
    - Comparison of different descriptions of the same system within a common formal framework.

- Important lessons:
  - Different segmentation schemes lead to small contrasts in overall description length.
  - Saving a few thousand bits of memory is not a compelling argument.



Solution 2: no segmentation

- Research program: learn inflection without (explicitly) learning word structure
  - Malouf (2017): using RNNs to learn a paradigm function
  - Baayen, Chuang, and Blevins (2018): using two-layered linear networks to learn linear mappings between form and content
  - See also the literature on reinflection in computational linguistics (e.g. Cotterell et al. 2017)
- In line with Blevins's (2006) abstractive approach to morphology, where "morphs are regarded as abstractions over forms, not as the 'building blocks' from which the forms are constructed" (p. 536)
- This is fine as long as we are interested in the learnability (or the actual learning) of morphological systems.

Why segmentation is still relevant I

- However, this does not address directly issues such as the Paradigm Cell Filling Problem (Ackerman, Blevins, and Malouf, 2009, 54, emphasis added):

  > *What licenses reliable inferences about the inflected [...] surface forms of a lexical item?*

  - As Ackerman, Blevins, and Malouf (2009) and later literature emphasize, implicative relations between surface forms play a crucial role in licensing such inferences.
  - Such implicative relations build on the identification of subword sequences that play a predictive role, while others don't—hence they implicitly induce a segmentation.

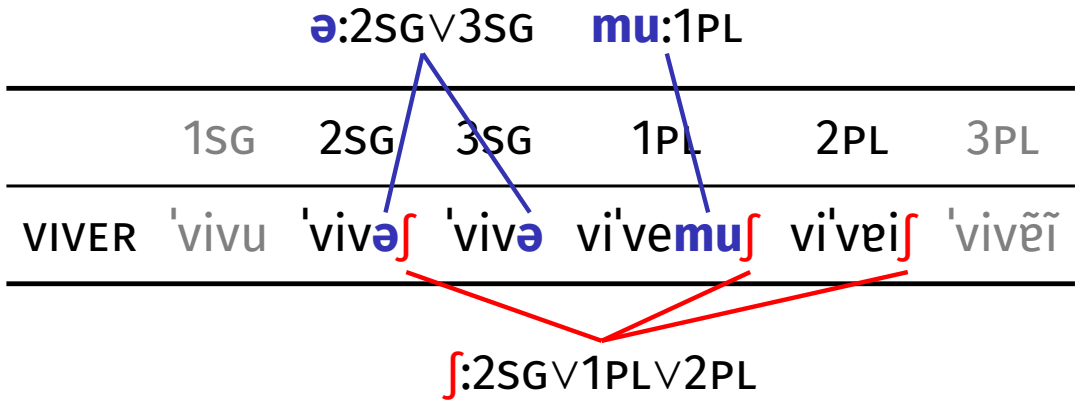
	1SG	2SG	3SG	⇐ 1PL	2PL	3PL
FICAR	ˈfiku	ˈfikɐʃ	ˈfikɐ	⇐ fiˈkɐmuʃ	fiˈkaiʃ	ˈfikɐ̃ũ
VIVER	ˈvivu	ˈvivəʃ	ˈvivə	⇐ viˈvemuʃ	viˈvɐiʃ	ˈvivɐ̃ĩ
IMPRIMIR	ĩpˈrimu	ĩpˈrimәʃ	ĩpˈrimə	⇐ ĩpriˈmimuʃ	ĩpriˈmiʃ	ĩpˈrimɐ̃ĩ

Why segmentation is still relevant II

- Similarly with the Paradigm Cell Recognition Problem (Beniamine, 2018):

  > *What licenses reliable inferences about the morphosyntactic property set expressed by a surface form of a lexical item?*

- Again, various segmentable substrings contribute to the licensing.



- Note that the relevant substrings need not correspond to classical morph(eme)s.
- ...but that is a separate issue.

Solution 3: local surface segmentation

- For purposes of the PCFP and PCRP, we want to identify the distinctive role played by different substrings within wordforms.
- However there is no reason that these should coincide with classical morph(eme)s, stems, or affixes.
  - Morphs are motivated by optimization of the size of the lexicon, a matter we do not worry about within an abstractive approach.
- In the remainder of this talk we will:
  - Present a simple-minded algorithm to classify alternations between surface forms (Beniamine, 2017)
  - Apply it to a collection of large datasets from 7 languages
  - Show how it helps us adress
    1. The PCFP (Bonami and Boyé, 2014; Bonami and Luís, 2014; Bonami and Beniamine, 2016)
    2. Inflectional classification (Beniamine, Bonami, and Sagot, 2017; Beniamine, forthcoming)
    3. The PCRP (new work)

Inferring patterns of alternation

The goal

- We want to design an algorithm that characterizes how pairs of wordforms are related.
- Design goals:
  - Fully deterministic, for reproducibility.
  - As typologically unbiased as possible/practical.
  - Simple enough
    - to be implemented and deployed over large datasets, and
    - for descriptive linguists to criticize.
- We generalize and streamline the rule inference algorithm underlying Albright's (2002) Minimal Generalization Learner.

Main challenge: global decisions

- Many ways of looking at the relation between two strings.
- Local decisions need to be informed by the rest of the lexicon.

SG	PL
ba	baba

$\_ \leftrightharpoons$ ba_ (prefixation)

$\_\_ \leftrightharpoons$ _ab_ (infixation)

$\_ \leftrightharpoons$ _ba (suffixation)

System A

SG	PL
ba	baba
to	bato
ri	bari
su	basu

System B

SG	PL
ba	baba
to	tabo
ri	rabi
su	sabu

System C

SG	PL
ba	baba
to	toba
ri	riba
su	suba

- To relate cells *X* and *Y*, we need a minimal set of patterns such that one pattern relates each *X* to the appropriate *Y*.

A pragmatic solution

- We want to capture the fact that the following pairs instantiate two patterns.

	PRS.1SG	PRS.1PL
GARANTIR	gɐrɐ̃tu	gɐrɐ̃timuʃ
DIVERTIR	divirtu	divərtimuʃ
FERIR	firu	fərimuʃ
PREMIR	primu	prəmimuʃ

Pattern
_ _ ⇋ _im_ʃ
_i_ _ ⇋ _ə_im_ʃ

- How can we derive that?

A pragmatic solution

- We want to capture the fact that the following pairs instantiate two patterns.

	PRS.1SG	PRS.1PL
GARANTIR	gɐrẽtu	gɐrẽtimuʃ
DIVERTIR	divirtu	divərtimuʃ
FERIR	firu	fərimuʃ
PREMIR	primu	prəmimuʃ

Pattern	ED
_ _ ⇋ _im_ʃ	3

- How can we derive that?
  1. Find all patterns that minimize edit distance for some pair.

A pragmatic solution

- We want to capture the fact that the following pairs instantiate two patterns.

	PRS.1SG	PRS.1PL
GARANTIR	gɐrẽtu	gɐrẽtimuʃ
DIVERTIR	divirtu	divərtimuʃ
FERIR	firu	fərimuʃ
PREMIR	primu	prəmimuʃ

Pattern	ED
_ _ ⇋ _im_ʃ	3
_i_ _ ⇋ _ə_im_ʃ	4

- How can we derive that?
  1. Find all patterns that minimize edit distance for some pair.

A pragmatic solution

- We want to capture the fact that the following pairs instantiate two patterns.

	PRS.1SG	PRS.1PL
GARANTIR	gɐrẽtu	gɐrẽtimuʃ
DIVERTIR	divirtu	divərtimuʃ
FERIR	firu	fərimuʃ
PREMIR	primu	prəmimuʃ

Pattern	ED
_ _ ⇋ _im_ʃ	3
_i_ _ ⇋ _ə_im_ʃ	4
_ _ɾ_ ⇋ _ər_m_ʃ	4

- How can we derive that?
  1. Find all patterns that minimize edit distance for some pair.

A pragmatic solution

- We want to capture the fact that the following pairs instantiate two patterns.

	PRS.1SG	PRS.1PL
GARANTIR	gɐrɐ̃tu	gɐrɐ̃timuʃ
DIVERTIR	divirtu	divərtimuʃ
FERIR	firu	fərimuʃ
PREMIR	primu	prəmimuʃ

Pattern	ED
_ _ ⇋ _im_ʃ	3
_i_ _ ⇋ _ə_im_ʃ	4
_ _ɾ_ ⇋ _ər_m_ʃ	4
_ _ ⇋ _əm_ʃ	3

- How can we derive that?
  1. Find all patterns that minimize edit distance for some pair.

A pragmatic solution

- We want to capture the fact that the following pairs instantiate two patterns.

	PRS.1SG	PRS.1PL
GARANTIR	gɐrɐ̃tu	gɐrɐ̃timuʃ
DIVERTIR	divirtu	divərtimuʃ
FERIR	firu	fərimuʃ
PREMIR	primu	prəmimuʃ

Pattern	ED
_ _ ⇋ _im_ʃ	3
_i_ _ ⇋ _ə_im_ʃ	4
_ _ɾ_ ⇋ _ər_m_ʃ	4
_ _ ⇋ _əm_ʃ	3

- How can we derive that?
  1. Find all patterns that minimize edit distance for some pair.
  2. For each pair, determine the subset of compatible patterns, disregarding edit distances.

A pragmatic solution

- We want to capture the fact that the following pairs instantiate two patterns.

	PRS.1SG	PRS.1PL
GARANTIR	gɐrẽtu	gɐrẽtimuʃ
DIVERTIR	divirtu	divərtimuʃ
FERIR	firu	fərimuʃ
PREMIR	primu	prəmimuʃ

Pattern	ED	Coverage
_ _ ⇋ _im_ʃ	3	1
_i_ _ ⇋ _ə_im_ʃ	4	3
_ _ɾ_ ⇋ _ər_m_ʃ	4	1
_ _ ⇋ _əm_ʃ	3	1

- How can we derive that?
  1. Find all patterns that minimize edit distance for some pair.
  2. For each pair, determine the subset of compatible patterns, disregarding edit distances.
  3. Determine the coverage of each pattern.

A pragmatic solution

- We want to capture the fact that the following pairs instantiate two patterns.

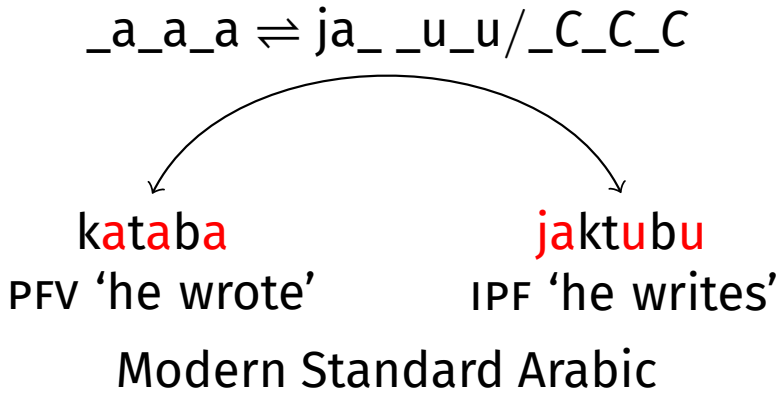
	PRS.1SG	PRS.1PL
GARANTIR	gɐrẽtu	gɐrẽtimuʃ
DIVERTIR	divirtu	divərtimuʃ
FERIR	firu	fərimuʃ
PREMIR	primu	prəmimuʃ

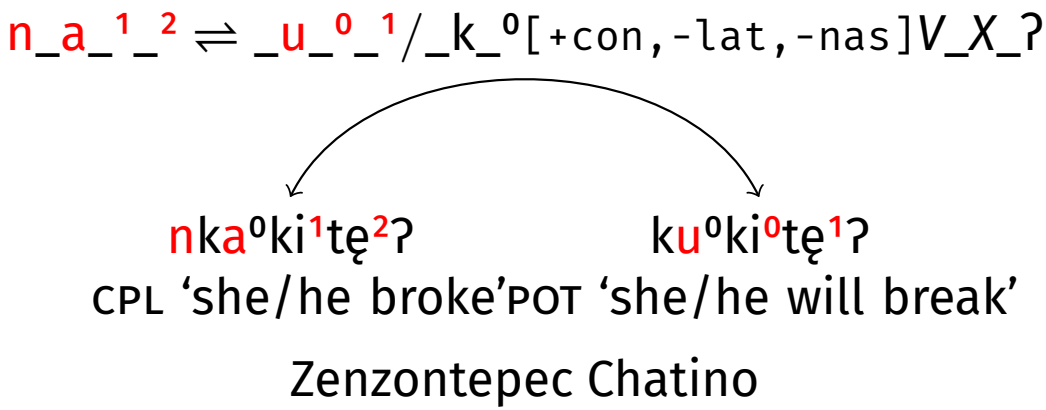
Pattern	ED	Coverage
_ _ ⇋ _im_ʃ	3	1
_i_ _ ⇋ _ə_im_ʃ	4	3
_ _ɾ_ ⇋ _ər_m_ʃ	4	1
_ _ ⇋ _əm_ʃ	3	1

- How can we derive that?
  1. Find all patterns that minimize edit distance for some pair.
  2. For each pair, determine the subset of compatible patterns, disregarding edit distances.
  3. Determine the coverage of each pattern.
  4. For each pair, pick the pattern with maximal coverage.

Qualitative evaluation

- The algorithm finds subtle, relevant patterns





Quantitative evaluation

- The algorithm has less alignment bias than its predecessors
  - Cross-validation shows that patterns extend reasonably well to unseen data

Language	Lexicon size	Align right	Align left	Albright (2002)	Levenshtein distance	Phonological distance
English	6064	.31	.94	.94	.94	.94
M.S. Arabic	1018	.26	.45	.46	.80	.82
French	5249	.24	.95	.94	.94	.94
E. Portuguese	1996	.18	.93	.93	.91	.91
Y. Chatino	324	.30	.29	.33	.36	.36
Z. Chatino	392	.57	.25	.56	.57	.57
Navajo	2157	.32	.25	.37	.42	.42

- This makes it well-suited for cross-linguistic comparison.
- In practical applications we use a more subtle, phonology-aware edit distance inspired by Albright and Hayes (2006).

Interim conclusion

- We have defined a way of characterizing pairwise alternations of forms in inflection systems that
  - is fully deterministic
  - is easily interpretable by descriptive morphologists
  - avoids obvious biases, in particular in terms of directionality of alignment
- In the remainder of the talk, we attempt to show that patterns of alternation substitute usefully for segmentation into stems and exponents in various situations.

Applications to the PCFP

Motivation

- Reformulation of Ackerman, Blevins, and Malouf's 2009 approach to the PCFP.
- A simple example from French:

the PCFP

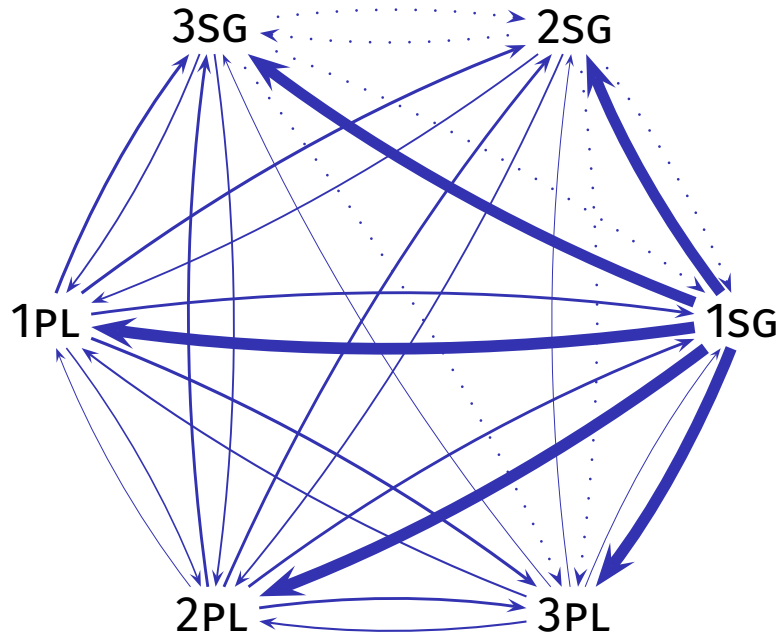
Lemma	SG	PL	Pattern	Patterns compatible with SG form
CHEVAL 'horse'	ʃəval	ʃəvo	_al⇋_o	{_al⇋_o, _ ⇋ _}
JOURNAL 'newspaper'	ʒuʁnal	ʒuʁno	_al⇋_o	{_al⇋_o, _ ⇋ _}
NARVAL 'narwhal'	naʁval	naʁval	_ ⇋ _	{_al⇋_o, _ ⇋ _}
JAGUAR 'jaguar'	ʒagwaʁ	ʒagwaʁ	_ ⇋ _	{_ ⇋ _}

- We want to predict the PL form from the SG form.
- This amounts to
  - predicting the pattern relation SG and PL,
  - on the basis of whatever surface-observable properties of SG might be relevant.
- The relevant properties are those phonological properties that single which patterns could have applied.

Results: descriptive

- New descriptive insights on the inflection systems of
  - Mauritian
    (Bonami, Boyé, and Henri, 2011)
  - French
    (Bonami and Boyé, 2014)
  - European Portuguese
    (Bonami and Luís, 2014)

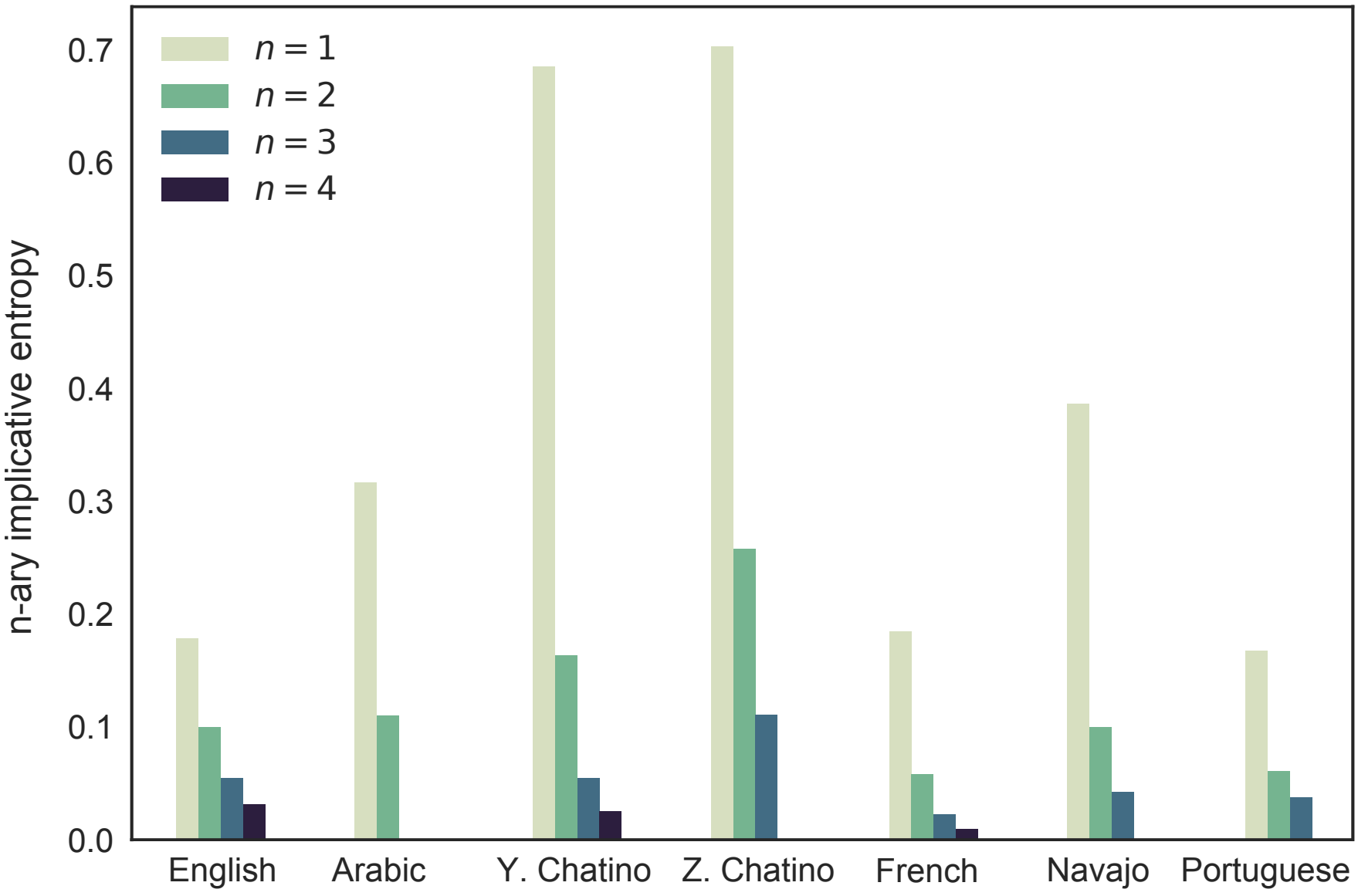
  - - - - - - - - - - - - - -
  - Zenzontepec Chatino
    (Beniamine and Bonami, 2016)
  - Navajo
    (Beniamine, Bonami, and McDonough, 2017)
  - Latin
    (Pellegrini, forthcoming)



Uncertainty in European Portuguese present verbs
(thicker = higher conditional entropy, dotted = zero entropy)

Results: Bonami and Beniamine (2016)

- Extensions of the Low Conditional Entropy Conjecture to simultaneous prediction from multiple paradigm cells.



Interim conclusion

- Paradigms have a fine predictive structure which varies pair of cell by pair of cell.
- Predictability depends on minute properties of surface words.
- Hence examining alternations between surface forms is crucial to uncovering that structure.
- Traditional segmentation does not help here:
  - Both stems and affixes have predictive value, sometimes jointly, sometimes separately
  - Even unsegmentable forms may have predictive force
- Importantly, the relevant properties of surface words arise from a local segmentation into constant and variable substrings, for the purpose of a particular comparison between two forms.

Applications to inflectional classification

Grounding inflectional classification

- Inflectional microclasses (Dressler and Thornton, 1996): sets of lexemes with exactly the same inflectional behaviour

lexeme	PST	PSTP	BSE	3SG	PRSP
DRIVE	drəʊv	drɪvn̩	draɪv	draɪvz	draɪvɪŋ
RIDE	rəʊd	rɪdn̩	raɪd	raɪdz	raɪdɪŋ
HIDE	hɪd	hɪdn̩	haɪd	haɪdz	haɪdɪŋ
FORGET	fəgɒt	fəgɒtn̩	fəgɛt	fəgɛts	fəgɛtɪŋ

- Microclasses instantiate the same vector of pairwise patterns.

lexeme	PST⇋PSTP	PST⇋3SG	3SG⇋BSE	PSTP⇋BSE	PSTP⇋PRSP
DRIVE	_əʊ_ ⇋ _ɪ_n̩	_əʊ_ ⇋ _aɪ_z	_z ⇋ _	_ɪ_n̩ ⇋ _aɪ_	_ɪ_n̩ ⇋ _aɪ_ɪŋ
RIDE	_əʊ_ ⇋ _ɪ_n̩	_əʊ_ ⇋ _aɪ_z	_z ⇋ _	_ɪ_n̩ ⇋ _aɪ_	_ɪ_n̩ ⇋ _aɪ_ɪŋ
HIDE	_⇋ _n̩	_ɪ_ ⇋ _aɪ_z	_z ⇋ _	_ɪ_n̩ ⇋ _aɪ_	_ɪ_n̩ ⇋ _aɪ_ɪŋ
FORGET	_ ⇋ _n̩	_ɒ_ ⇋ _ɛ_s	_s ⇋ _	_ɒ_n̩ ⇋ _ɛ_	_ɒ_n̩ ⇋ _ɛ_ɪŋ

lexeme	PST⇋BSE	PSTP⇋3SG	BSE⇋PRSP	PST⇋PRSP	BSE⇋PRSP
DRIVE	_əʊ_ ⇋ _aɪ_	_ɪ_n̩ ⇋ _aɪ_z	_ ⇋ _ɪŋ	_əʊ_ ⇋ _aɪ_ɪŋ	_z ⇋ _ɪŋ
RIDE	_əʊ ⇋ _aɪ	_ɪ_n̩ ⇋ _aɪ_z	_ ⇋ _ɪŋ	_əʊ_ ⇋ _aɪ_ɪŋ	_z ⇋ _ɪŋ
HIDE	_ɪ_ ⇋ _aɪ_	_ɪ_n̩ ⇋ _aɪ_z	_ ⇋ _ɪŋ	_Vt ⇋ _aɪtVŋ	_z ⇋ _ɪŋ
FORGET	_ɒ_ ⇋ _ɛ_	_ɒ_n̩ ⇋ _ɛ_s	_ ⇋ _ɪŋ	_ɒ_ ⇋ _ɛ_ɪŋ	_s ⇋ _ɪŋ

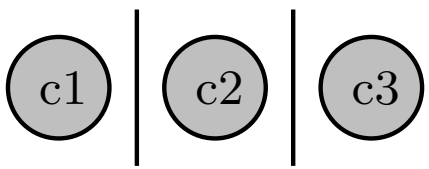
Modes of classification

- In general, we are interested in higher-level groupings, that are based on similarity rather than identity of behavior

lexeme	PST	PSTP	BSE	3SG	PRSP
DRIVE	drəʊv	drɪvn̩	draɪv	draɪvz	draɪvɪŋ
RIDE	rəʊd	rɪdn̩	raɪd	raɪdz	raɪdɪŋ
‘HIDE	hɪd	hɪdn̩	haɪd	haɪdz	haɪdɪŋ
FORGET	fəgɒt	fəgɒtn̩	fəgɛt	fəgɛts	fəgɛtɪŋ

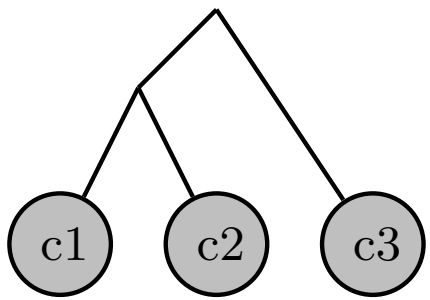
- Three ways of doing this:

Partition



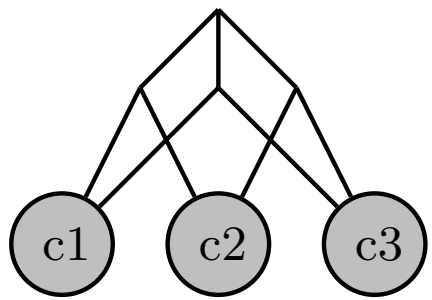
Tratitional

Tree



Corbett and Fraser (1993), Dressler and Thornton (1996),...

Lattice

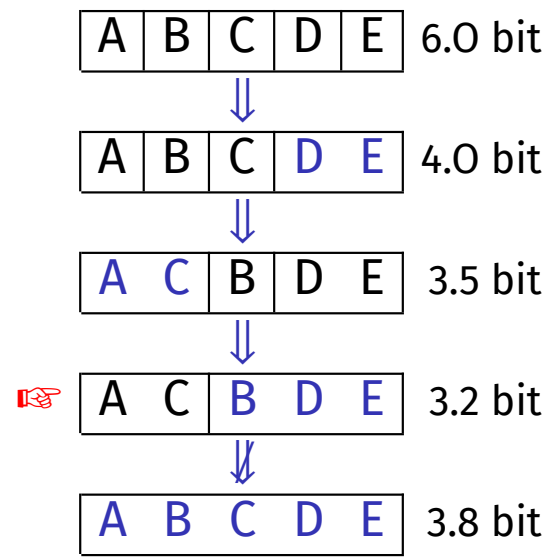


Beniamine (forthcoming), Bonami and Crysmann (2018),...

Inflectional macroclasses

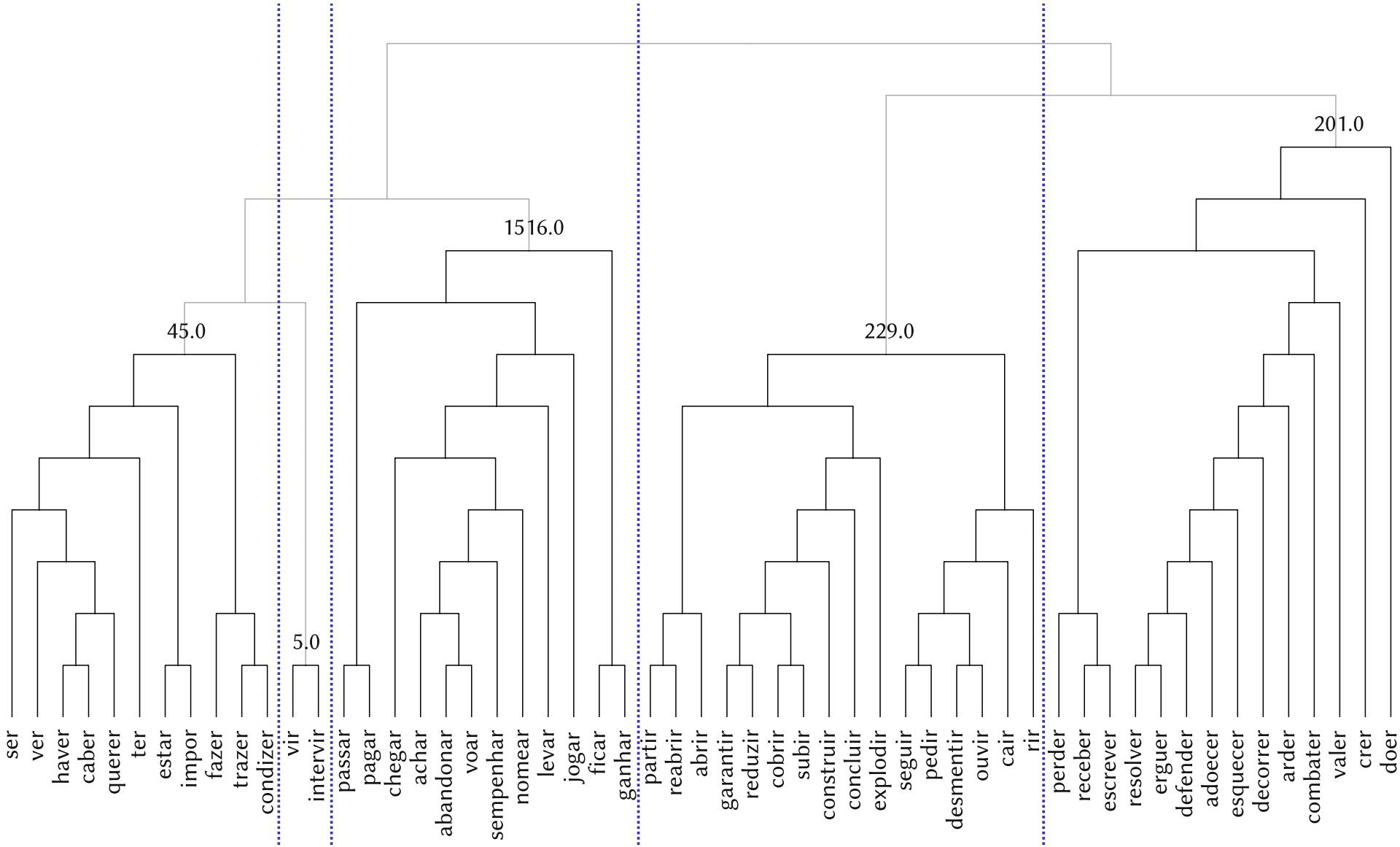
(Beniamine, Bonami, and Sagot, 2017)

- Intuition: macroclasses strike a balance between precision and generality
  - A good macroclass should provide as much information as possible on its members.
  - A good system of macroclass has few classes.
- We formalize this using Minimal Description Length:
  - Microclasses caracterized by a set of patterns
  - Greedy algorithm fuses classes so as to minimize description length
  - Stopping condition: description length increase



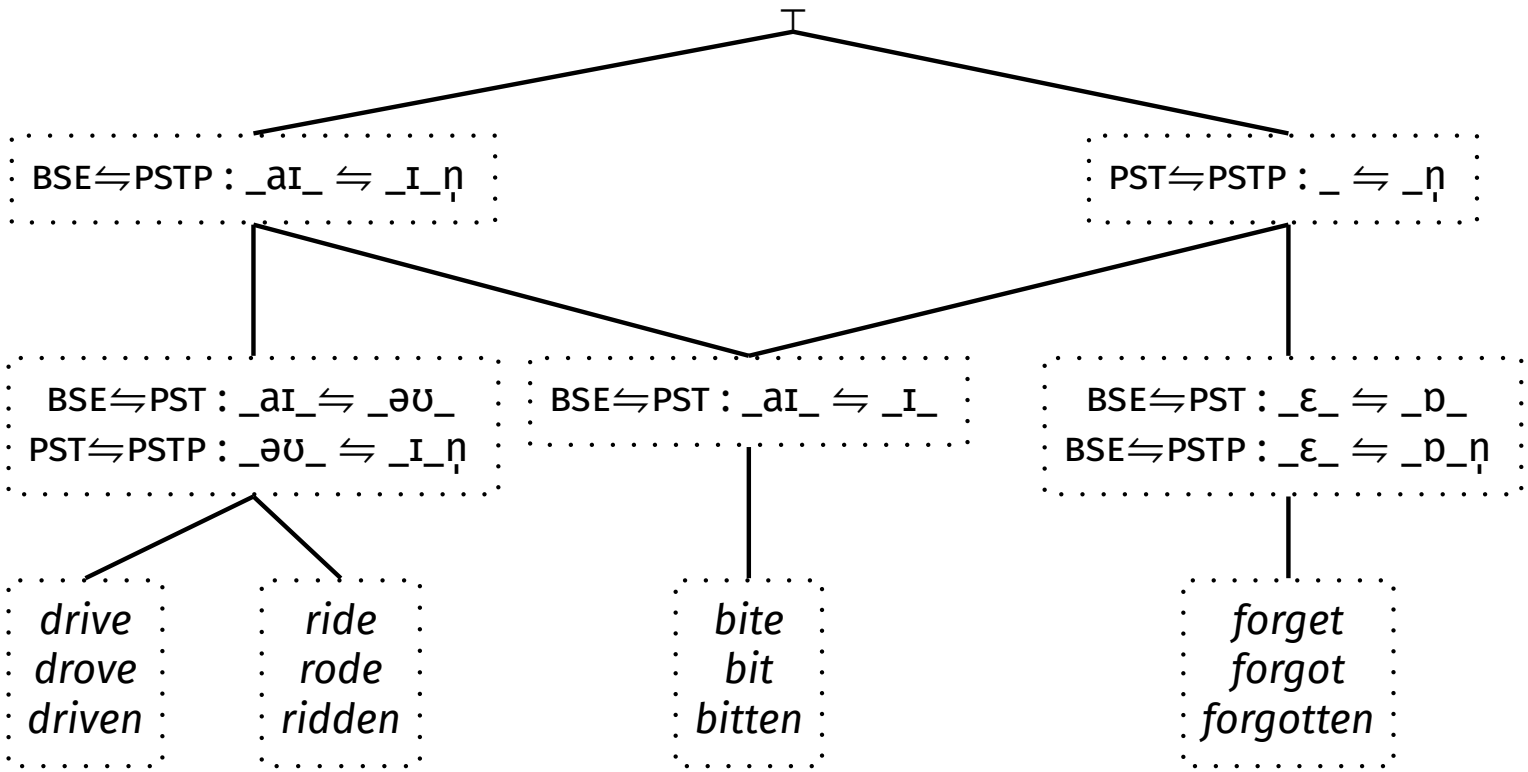
Results

- Portuguese: 3 familiar classes + two more

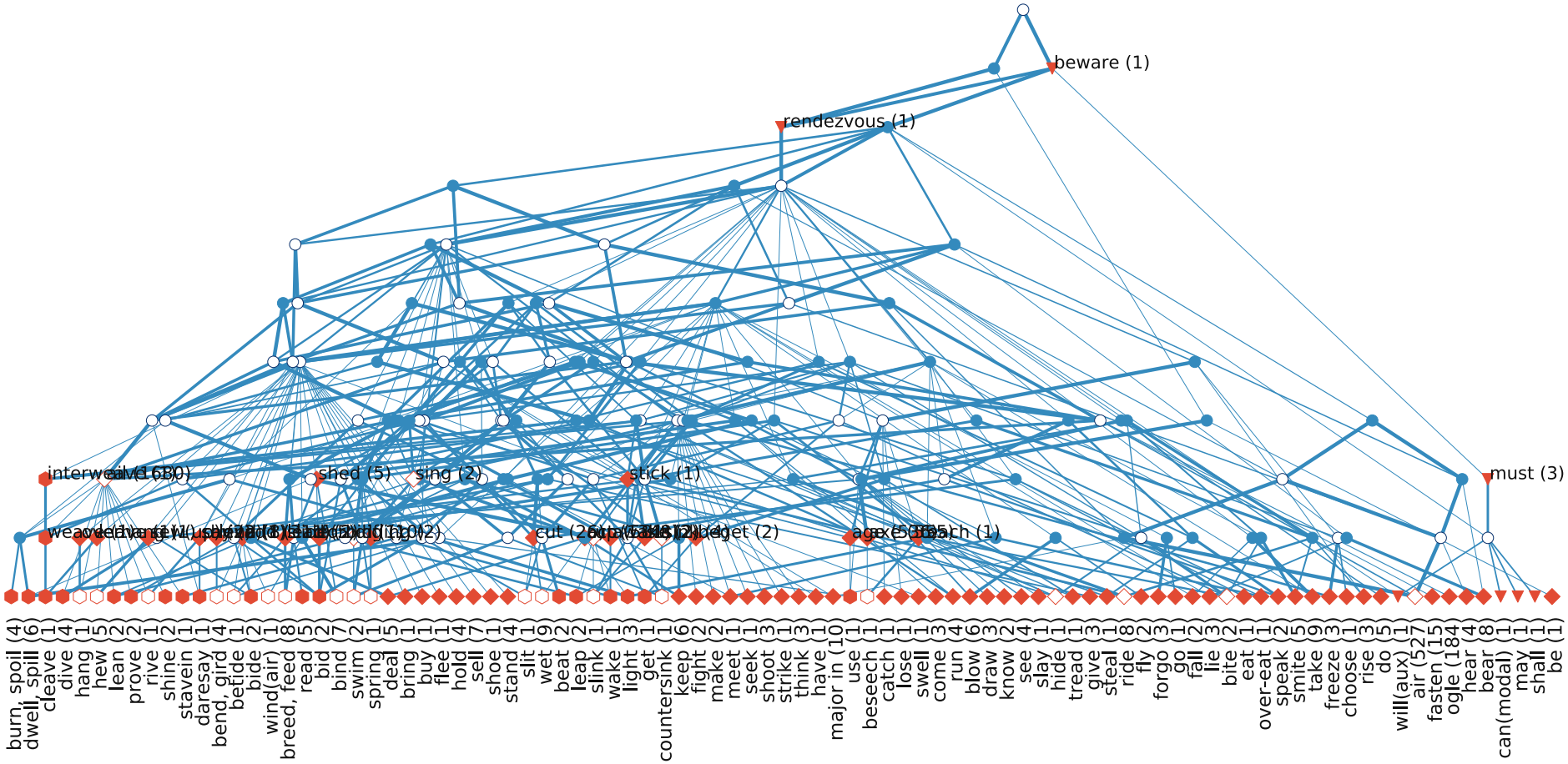


The fine structure of inflectional classification

- Inflection class systems obviously have structure beyond macroclasses.
- Many attempts to address these using trees.
- Because of heteroclisis, lattices are a more appropriate type of structure.



Results



- Inflection class lattices are
  - much more intricate than any hand-designed classification ($254 \leq n \leq 33,199$); but
  - considerably smaller than the full collection of sets of microclasses ($\frac{n}{w} \leq 10^{-10}$).
- High prevalence of heteroclisis: average degree $1.9 < d < 4.5$

Interim conclusion

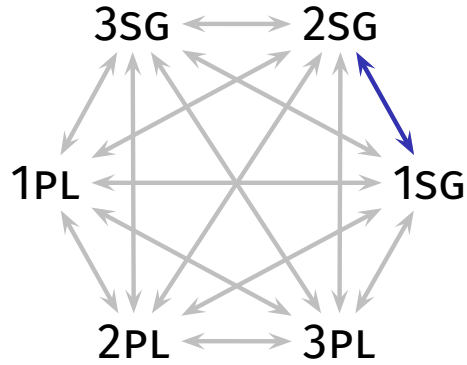
- We have presented two fruitful methods for addressing inflectional classification:
  - Coarse-grained: macro-classes
    - based on Minimal Description Length
  - Fine-grained: inflection class lattices
    - based on Formal Concept Analysis
- In both cases, our algorithms capture generalizations previously made by descriptive linguists, and provide new insights.
- Two interpretations of this:
  - We have shown that we can mimick a classification based on stems and affixes without postulating any such items; or
  - The classification was not based on stems and affixes in the first place, but on observations about interpredictability between surface forms.
- Although detailed historical work would be needed to establish this, the second option is very tempting.

The Paradigm Cell Recognition Problem

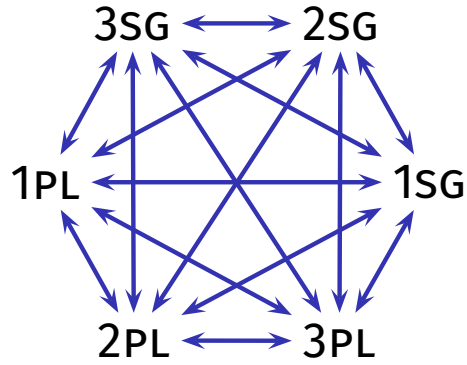
What licenses reliable inferences about the morphosyntactic property set expressed by a surface form of a lexical item?

The next step

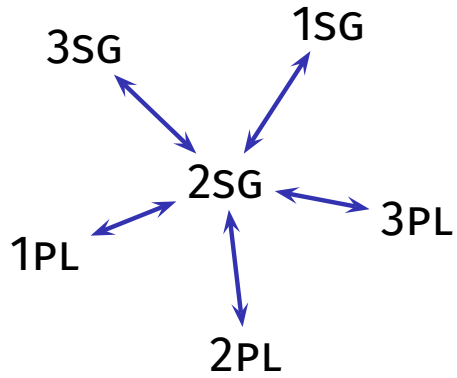
- To address the PCFP, we built on patterns of alternations relating two particular cells in the paradigm.



- For inflectional classification, we compared across lexemes the full set of pairwise alternations.

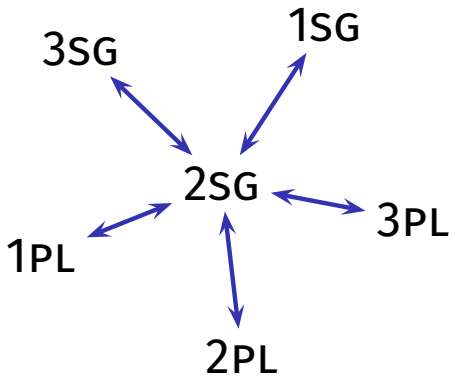


- To address the PCRP, we now want to attend to all alternations between one form and each of the other forms in the paradigm.



Consolidating pairwise alternations

- Starting from a particular wordform filling a particular paradigm cell, we introduce a boundary wherever some pairwise alternation distinguishes constant vs. variable substrings.
- Here for E. Portuguese *festejas* ‘you celebrate’:



Cell	*Alternation*	*Pattern*	*Segmentation*
1SG	fəʃteʒɐʃ ⇌ fəʃteʒu	_ɐʃ ⇌ _u	f ə ʃ t e ʒ+ɐ ʃ
3SG	fəʃteʒɐʃ ⇌ fəʃteʒɐ	_ʃ ⇌ _	f ə ʃ t e ʒ ɐ+ʃ
1PL	fəʃteʒɐʃ ⇌ fəʃtəʒɐmuʃ	_e__ ⇌ _ə_mu_	f ə ʃ t+e+ʒ ɐ+ʃ
2PL	fəʃteʒɐʃ ⇌ fəʃtəʒaiʃ	_e_ɐ_ ⇌ _ə_ai_	f ə ʃ t+e+ʒ+ɐ+ʃ
3PL	fəʃteʒɐʃ ⇌ fəʃteʒɐ̃ũ	_ɐʃ ⇌ _ɐ̃ũ	f ə ʃ t e ʒ+ɐ ʃ
	Final segmentation		f ə ʃ t+e+ʒ+ɐ+ʃ

Morphoids

- We have found the longest substrings that cohere together within the paradigm.
- We call these morphoids.
  - In the canonical case, these correspond exactly to classical stems and inflectional affixes.
  - In situations classically analyzed as instances of allomorphy, they may be smaller than classical morphs.
- What are morphoids exactly?
  - Maximal contiguous strings associated with some content.
  - Basic units of paradigmatic contrast

☞ Morphoids are exactly what one should pay attention to when addressing the PCRP: each morphoid provides a distinct piece of information from those adjacent to it.

Counting morphoids

- Counting the number of morphoids in a word gives us an indication of morphological size: how large a word is in terms of units of paradigmatic contrast.

	Average	Std. dev.	Min.	Median	Max.
French	2.77	1.25	1	3	8
English	1.89	0.79	1	2	6
Portuguese	5.31	1.76	1	5	11
Yaitepec Chatino	4.32	1	2	4	7
Zenzontepec Chatino	4.35	1.59	1	4	10
Modern Standard Arabic	7.12	1.63	1	7	12

- Compare Greenberg's (1954) synthetic index:

$$\frac{\text{Number of tokens of morphemes in the corpus}}{\text{Number of tokens of words in the corpus}}$$

Types of morphoids

- Definitions:
  - Morphoids that are present in all paradigm cells are inert
  - Morphoids that are present in only some paradigm cells are exponential
- Stem fragmentation: number of inert morphoids in a word.
  - Amounts to assessing how far we are from having the canonical situation where stems and exponents are segregated.

	Average	Std. dev.	Min.	Median	Max.
French	1.02	0.14	0	1	2
English	1.02	0.16	0	1	2
Portuguese	1.37	0.48	0	1	3
Yaitepec Chatino	0.95	0.42	0	1	2
Zenzontepec Chatino	1.26	0.52	1	1	3
Modern Standard Arabic	2.31	0.82	0	2	4

Fusion

- Given a structuration of paradigm cells in morhosyntactic property sets, we can quantify the amount of fusion in the system.
- Fusion index: $\frac{\text{number of exponential morphoids}}{\text{number of morphosyntactic properties}}$
  - Zero and cumulative exponence drive the index down
  - Multiple exponence drives the index up

lexeme	cell	form	index
laver	IMP.2.SG	lav	0/3 = 0
laver	SBJV.PRS.2.PL	lav+j+e	2/4 = 0.5
laver	INF	lav+e	1/1 = 1
lever	IND.FUT.ANA.2.PL	l+ɛ+v+ə+ʁ+j+e	5/5 = 1
lever	INF	l+ə+v+e	2/1 = 2
mouvoir	INF	m+u+v+w+a+ʁ	5/1 = 5

	average	std. dev.	min	median	max
Fusion index	0.416	0.300	0	0.4	5

NB At this point it is an open question how to do this in a way that makes crosslinguistic comparisons meaningful.

Exponential purity I

- A morphoid μ is a pure exponent of a morphosyntactic property set σ iff every form expressing σ contains μ and every form containing μ expresses σ.
  - **-m** is a pure exponent of IND.PST.PFV.1PL in French.
- In a canonical system, all exponential morphoids are pure.
- In real systems, few are.
  - -ɔ̃ as an exponent of 1PL in French.

	1SG	2SG	3SG	1PL	2PL	3PL
IND.PRS	lav	lav	lav	lavɔ̃	lave	lav
IND.PST.IPFV	lavɛ	lavɛ	lavɛ	lavjɔ̃	lavje	lavɛ
IND.PST.PFV	lavɛ	lava	lava	lava**m**	lavat	lavɛʁ
IND.FUT	lavəʁɛ	lavəʁa	lavəʁa	lavəʁɔ̃	lavəʁe	lavəʁɔ̃
COND	lavəʁɛ	lavəʁɛ	lavəʁɛ	lavəʁjɔ̃	lavəʁje	lavəʁɛ
SBJV.PRS	lav	lav	lav	lavjɔ̃	lavje	lav
SBJV.PST	lavas	lavas	lava	lavasjɔ̃	lavasje	lavas

Exponential purity II

- When facing impure exponence, our instinct as morphologists is to attempt to reduce it, appealing to:
  - Allomorphy
  - The Elsewhere Condition / Panini's Principle / The subset principle / Defaults / ...
- This is unreasonable if what we are interested in is the PCRP.
- Impure morphoids have predictive value that cannot be captured by seeing it as realizing a single property set.
- ☞ When a French verb ends in ɔ̃:
  - Probability of being in the 1PL: $^{6}/_{7}$
  - Probability of being in the future: $^{2}/_{7}$
  - Probability of being in the simple past: 0
  - If in the future, probability of being in the 1PL: $^{1}/_{2}$
  - If not in the future, Probability to be in the 1PL: 1
  - ...
- We should study the fine predictive structure of morphoids rather than attempt to hide it.

Interim conclusion

- We have proposed a simple way of infering a segmentation of inflected words on the basis of their place in the paradigm.
- Abstractive approach:
  - As close as possible to the surface: no postulation of entities more abstract than strings of phonemes
  - Directly grounded in the PCRP, a slightly idealized version of the problem of recognizing the content of words
- We have shown how it can be deployed to address common concerns:
  - Assessing the morphological size of words for purposes of quantitative typology
  - Reasoning on exponence
- Further natural steps:
  - Grounding a formal typology of exponence (Carroll, submitted).
  - (Semi-)automatic glossing.

Three lessons

1. Segmentation should not be taken for granted
   - There are specific, relevant ways of segmenting words for specific purposes, but
     - this does not entail that a unique segmentation grounds our understanding of what a word is,
     - nor that content should categorically be associated with the segments.
2. For models of morphology need to accomodate the pervasiveness of discontinuous stems and exponents
   - ☞ *m* : *n* rules as first-class citizens in Information-based Morphology (Crysmann and Bonami, 2016).
3. Good typology requires good, reproducible measurement (Round and Corbett, submitted).
   - Designing and distributing reliable instruments is an important goal.
   - ☞ http://drehu.linguist.univ-paris-diderot.fr/qumin/

References I

Ackerman, Farrell, James P. Blevins, and Robert Malouf (2009). "Parts and wholes: implicative patterns in inflectional paradigms." In: *Analogy in Grammar*. Ed. by James P. Blevins and Juliette Blevins. Oxford: Oxford University Press, pp. 54–82 (cit. on pp. 7, 25).

Albright, Adam and Bruce Hayes (2006). "Modeling productivity with the Gradual Learning Algorithm: the problem of accidentally exceptionless generalizations." In: *Gradience in Grammar: Generative Perspectives*. Ed. by Gisbert Fanselow et al. Oxford: Oxford University Press, pp. 185–204 (cit. on p. 22).

Albright, Adam C. (2002). "The Identification of Bases in Morphological Paradigms." PhD thesis. University of California, Los Angeles (cit. on p. 11).

Aronoff, Mark (1994). *Morphology by itself*. Cambridge: MIT Press (cit. on p. 2).

Baayen, R. Harald, Yu-Ying Chuang, and James P. Blevins (2018). "Inflectional morphology with linear mappings." In: *The Mental Lexicon* 13.2, pp. 230–268 (cit. on p. 6).

Beniamine, Sacha (2017). "Une approche universelle pour l'abstraction automatique d'alternances morphophonologiques." In: *Actes de TALN 2017*, pp. 77–85 (cit. on p. 9).

— (2018). "Typologie quantitative des systèmes de classes flexionnelles." PhD thesis. Université Paris Diderot (cit. on p. 8).

References II

Beniamine, Sacha (forthcoming). “One lexeme, many classes: Inflection class systems as lattices.” In: *One-to-many relations in morphology, syntax and semantics*. Ed. by Berthold Crysmann and Manfred Sailer. Language Science Press (cit. on pp. 9, 31).

Beniamine, Sacha and Olivier Bonami (2016). “Generalizing patterns in Instrumented Item-and-Pattern Morphology.” In: *Structural Complexity in Natural Language(s)*. Paris (cit. on p. 26).

Beniamine, Sacha, Olivier Bonami, and Joyce McDonough (2017). “When segmentation helps. Implicative structure and morph boundaries in the Navajo verb.” In: *First International Symposium on Morphology*. Lille (cit. on p. 26).

Beniamine, Sacha, Olivier Bonami, and Benoît Sagot (2017). “Inferring Inflection Classes with Description Length.” In: *Journal of Language Modelling* 5.3, pp. 465–525 (cit. on pp. 9, 32).

Blevins, James P. (2006). “Word-based morphology.” In: *Journal of Linguistics* 42, pp. 531–573 (cit. on p. 6).

Bonami, Olivier and Sarah Beniamine (2016). “Joint predictiveness in inflectional paradigms.” In: *Word Structure* 9.2, pp. 156–182 (cit. on pp. 9, 27).

Bonami, Olivier and Gilles Boyé (2014). “De formes en thèmes.” In: *Foisonnements morphologiques. Etudes en hommage à Françoise Kerleroux*. Ed. by Florence Villoing, Sarah Leroy, and Sophie David. Presses Universitaires de Paris Ouest, pp. 17–45 (cit. on pp. 9, 26).

References III

Bonami, Olivier, Gilles Boyé, and Fabiola Henri (2011). "Measuring inflectional complexity: French and Mauritian." In: *Workshop on Quantitative Measures in Morphology and Morphological Development*. San Diego (cit. on p. 26).

Bonami, Olivier and Berthold Crysmann (2018). "Lexeme and flexeme in a formal theory of grammar." In: *The Lexeme in Descriptive and Theoretical Morphology*. Ed. by Olivier Bonami et al. Berlin: Language Science Press, pp. 175–202 (cit. on p. 31).

Bonami, Olivier and Ana R. Luís (2014). "Sur la morphologie implicative dans la conjugaison du portugais : une étude quantitative." In: *Morphologie flexionnelle et dialectologie romane. Typologie(s) et modélisation(s).* Ed. by Jean-Léonard Léonard. Mémoires de la Société de Linguistique de Paris 22. Leuven: Peeters, pp. 111–151 (cit. on pp. 9, 26).

Carroll, Matthew J. (submitted). "Redundancy in multiple exponence." (Cit. on p. 46).

Corbett, Greville G. (2007). "Canonical typology, suppletion and possible words." In: *Language* 83, pp. 8–42 (cit. on pp. 2, 3).

Corbett, Greville G. and Norman M. Fraser (1993). "Network Morphology: a DATR account of Russian nominal inflection." In: *Journal of Linguistics* 29, pp. 113–142 (cit. on p. 31).

Cotterell, Ryan et al. (2017). "CoNLL-SIGMORPHON 2017 Shared Task: Universal Morphological Reinflection in 52 Languages." In: *Proceedings of the CoNLL SIGMORPHON 2017 Shared Task: Universal Morphological Reinflection*. Vancouver, Canada, pp. 1–30 (cit. on p. 6).

References IV

Crysmann, Berthold and Olivier Bonami (2016). "Variable morphotactics in Information-Based Morphology." In: *Journal of Linguistics* 52.2, pp. 311–374 (cit. on p. 47).

Dressler, Wolfgang U. and Anna M. Thornton (1996). "Italian Nominal Inflection." In: *Wiener Linguistische Gazette* 55-57, pp. 1–26 (cit. on pp. 30, 31).

Greenberg, Joseph H. (1954). "A quantitative approach to the morphological typology of language." In: *Method and Perspective in Anthropology: Papers in Honor of Wilson D. Wallis*. Ed. by Robert F. Spencer. University of Minnesota Press (cit. on pp. 2, 41).

Malouf, Robert (2017). "Abstractive morphological learning with a recurrent neural network." In: *Morphology* 27.4, pp. 431–458 (cit. on p. 6).

Pellegrini, Matteo (forthcoming). "Predictability and implicative relations in Latin inflection." PhD thesis. University of Bergamo (cit. on p. 26).

Round, Erich and Greville G. Corbett (submitted). "Comparability and measurement in typological science: the bright future for linguistics." In: *Linguistic Typology* (cit. on p. 47).

Sagot, Benoît and Géraldine Walther (2013). "Implementing a formal model of inflectional morphology." In: *Proceedings of Systems and Frameworks in Computational Morphology*, pp. 115–134 (cit. on p. 5).

Spencer, Andrew (2012). "Identifying stems." In: *Word Structure* 5, pp. 88–108 (cit. on p. 2).

References V

Walther, Géraldine (2013). “De la canonicité en morphologie: perspective empirique, théorique et computationnelle.” PhD thesis. Université Paris Diderot (cit. on p. 5).

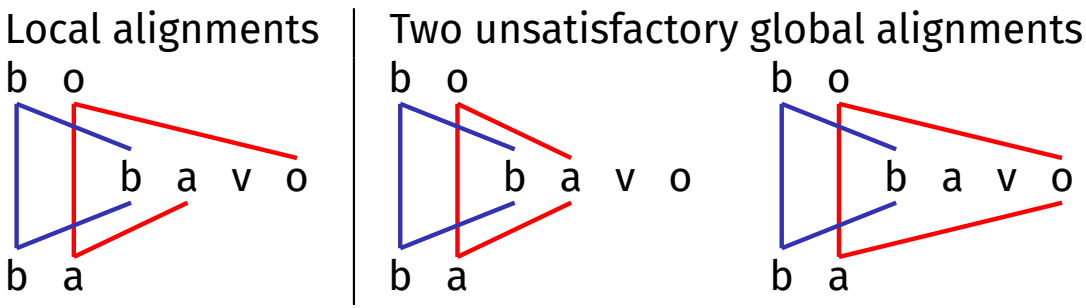
Walther, Géraldine and Benoît Sagot (2011). “Modélisation et implémentation de phénomènes flexionnels non-canoniques.” French. In: *Traitement Automatique des Langues* 52.2, pp. 91–122 (cit. on p. 5).

Why global segmentation fails

- Finding a good global segmentation amounts to finding an optimal global alignment.
- Yet that is not always mathematically possible.
- Consider a system with the following structure:

Lexeme	c_1	c_2	c_3
L_1	bo	ba	bavo
L_2	tu	ta	tavu
L_3	ke	ka	kavu

- Descriptively: c_2 derives from c_1 by vowel deletion and suffixation of -a, c_3 derives from c_1 by infixation of -av-.
- There is no global alignment that preserves local alignments:



- Hence there is no way of choosing a global alignment without disregarding important morphological regularities.
- For this reason, we rely entirely on local alternations and local alignments.